Linkage disequilibrium based eQTL analysis and comparative evolutionary epigenetic regulation of gene transcription

by

NING JIANG

A thesis submitted to
The University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY

School of Biosciences
The University of Birmingham
November 2012

ABSTRACT

The present thesis contains two independent parts of research, which are summarised as below.

Part I

Genome-Wide Association Study (GWAS) has recently been proposed as a powerful strategy for detecting the many subtle genetic variants that underlie phenotypic variation of complex polygenic traits in population-based samples. One of the main obstacles to successfully using the linkage disequilibrium based methods is knowledge of any underlying population structure. The presence of subgroups within a population can result in spurious association. A robust statistical method is developed to remove the population structure interference in GWAS by incorporating single control marker into testing for significance of genetic association of a polymorphic marker (SNP) with phenotypic variance of a complex trait. The novel approach avoids the need of structure prediction which could be infeasible or inadequate in practice and accounts properly for a varying effect of population stratification on different regions of the genome under study. Both intensive computer simulation study and eQTL analysis in genetically divergent human populations show that the new method confers an improved statistical power for detecting genuine genetic association in subpopulations and an effective control of spurious associations stemmed from population structure.

Part II

Regulation of gene transcription (or expression) plays an essential role for viruses, prokaryotes and eukaryotes when the genetic information is turned into functional products—RNA. It can introduce the variability and adjustability of an organism

through modulating gene expression levels. Gene expression process may be regulated by several aspects, from transcription initiation to methylation-based regulation. In epigenetics, the addition of methyl groups to cytosine bases in DNA sequence is considered as heritable chemical markers. DNA methylation has been confirmed that it plays a fundamental role for regulation of gene expression and is widespread among eukaryotic species, particular in vertebrates. In these analyses, two highly conserved but distinct methylation-based promoter regulatory patterns have been detected in higher vertebrates. These two diverse promoter classifications show distinct features and properties in genomic, methylation, expression and function levels. However, they present a highly conserved performance among several higher vertebrate species.

ACKNOWLEDGEMENTS

The progress of this project owes a great deal to the guidance and encouragement of my supervisor, Professor Zewei Luo. Without his invaluable patience, advice and insight, this project would not be possible to complete. Thank you very much for guiding me in both of my academic research and daily life.

I would also like to thank my second supervisors, Dr Christine Hackett and Dr David Marshall. Your support and encouragement have always been helpful throughout my PhD project. Additionally, I want to express my special thanks to Professor Michael J. Kearsey and Dr Lindsey Leach for their help, suggestion and advice.

I appreciate all the help from my colleagues in University of Birmingham, including Dr Minghui Wang, Dr Tianye Jia, Dr Elena Potokina, Dr Joseph Abraham, Yiyuan Liu and Jing Chen. Thanks for giving me support and sharing your experiences in research. It has been a pleasure to work with all of you.

I express my heartfelt gratitude to my parents Qihe Jiang and Lifang Nie, my grandma Yusheng Sun for helping me a lot in various ways to complete my PhD project. A very special thank you to my love Mrs Hui Jiang whose support and love has been incredibly important to me. It is my great fortune to live with you and always will be.

Lastly but certainly not least, thanks also go to Scottish Crop Research Institute (SCRI) and University of Birmingham. I have been fortunate to receive a joint studentship to make this project possible.

Table of Contents

Part I

Developing the linkage disequilibrium (LD) based association method for eQTL analysis

Chapter I 2

Overall introduction: association mapping, gene transcription process and expression quantitative trait loci (eQTLs) analysis 2

1. 1 Related publications 2

1.2 Overview 2

1.3 Complex traits 4

1.4 Genetic linkage and linkage analysis 6

1.5 Linkage disequilibrium and association studies 8

1.5.1 Definition of LD 8

1.5.2 Genetic association studies 11

1.5.3 Advantages of association studies 12

1.5.4 Population structure: A big challenge in association study 14

1.6 Genetic markers and high-throughput genotyping technologies 17

1.6.1 Genotyping SNP markers using microarray platforms 19

1.6.2 Discovering and genotyping SNPs based on NGS technologies 20

1.6.3 Prediction of genetic polymorphisms from expression microarray dataset 22

1.7 Gene expression process 25

1.8 Gene transcription 28

1.9 Expression quantitative trait loci (eQTLs) analysis 33

1.9.1 A frame work of eQTL analysis 35

1.9.2 Extracting genome wide gene expression profile for eQTL mapping. 36

1.9.3 eQTL Hotspots 40

1.9.4 An important application of eQTL studies 41

1.10 References 44

Chapter II 51

Developing a robust statistical method 51

for association-based 51

expression quantitative trait analysis 51

2.1 Related publication 51

2.2 Overview 51

2.3 Introduction 53

2.3.1 Family-based association studies 54
2.3.2 Genomic control (GC) method and the structure association (SA) analysis .. 55
2.3.3 Principal component analysis (PCA) based method— EIGENSTRAT 58
2.3.4 Correcting the confound effect of population structure using only one genetic marker 59
2.4 Statistical models and methods 62
2.4.1 Method 1: a novel regression analysis with correcting population structure .. 62
2.4.1.1 The novel linear regression model and theoretical analysis 65
2.4.1.2 Significant test of the regression coefficient b_1 68
2.4.1.3 Selection of the control marker 69
2.4.2 Method 2 (Regression analysis without correcting population structure) 70
2.4.3 Method 3 (multiple regression analysis) 71
2.5 Simulation study 73
2.5.1 Simulations of admixed populations 73
2.5.2 Comparisons of three approaches based on simulation data 76
2.5.2.1 Probability of statistical power and false positive inference 76
2.5.2.2 The performance of **Method 3** when the population structure information is unknown 81
2.5.2.3 Performance of the novel method (**Method 1**) using varied control markers 86
2.6 Real data analysis 90
2.6.1 Data resources and pretreatment 90
2.6.1.1 Gene expression data 91
2.6.1.2 Genome-wide genotype datasets for 142 human individuals 92
2.6.2 Validation of population structure 94
2.6.3 Genome-wide association eQTL analysis 97
2.7 Discussion 107
2.8 References 111

Part II

Comparative evolutionary epigenetic regulation of gene transcription

Chapter III 116
General introduction: regulation of gene transcription 116
3.1 Overview 116
3.2 Regulation of gene transcription 117
3.2.1 Regulatory proteins 118
3.2.2 RNA based transcriptional regulation 119

3.3.3 Epigenetic regulation 122
3.4 Reference 123
Chapter IV: 126
Comparative evolutionary epigenetic 126
regulation of gene transcription 126
4.1 Introduction to DNA methylation based transcriptional regulation 126
4.2 Datasets and analytical methods 132
4.2.1 Whole genome sequence datasets and genomic features annotation information 132
4.2.2 Distribution patterns of CpG sites in promoter regions 133
4.2.3 Identification of promoter classes 134
4.2.4 Identification of homologous genes and interspecies conservation analysis 135
4.2.5 Genome-wide DNA methylation data and gene expression data 136
4.2.6 GO annotation datasets and overrepresentation analysis 137
4.3 Analysis and Result 139
4.3.1 Overview of the genome-wide distributions of CpG sites and GC contents in several species 139
4.3.2 Analysis of the pattern of the distribution of CpG sites in the promoter regions 144
4.3.3 Evolutionary conservation of promoters in higher vertebrates 151
4.3.4 Distinct methylation patterns between HCP and LCP promoters across 28 human tissues 155
4.3.5 Distinct expression patterns between HCP and LCP promoters in 107 human tissues 158
4.3.6 Distinct and conserved functions of genes with HCP and LCP promoters .. 161
4.4 Conclusion and Discussion 164
4.5 Reference 168
Part III
Several cooperated researches of the linkage disequilibrium based association study and gene transcription regulation mechanism analysis
1. Inferring Linkage Disequilibrium (LD) in case-control samples 171
1.1 Related publication 171
1.2 Summary 171
2. A powerful statistical method for genetic association studies using case-control samples 173
2.1 Related publication 173
2.2 Summary 173

3. Genetic dissection of agronomic and morphologic traits in highly structured populations of barley cultivars.....174
3.1 Related publication 174
3.2 Summary 175
4. Investigation of gene expression regulatory mechanisms in Saccharomyces cerevisiae (budding yeast) genome 175
4.1 Related publication 175
4.2 Summary 176
Chapter V: Final conclusion and summary of this project.....179
Appendices 183
Appendix I: Druka A, Potokina E, Luo Z, Jiang N, Chen X, Kearsey M, Waugh R (2010) Expression quantitative trait loci analysis in plants. *Plant Biotechnology Journal*, 8(1):10-27. 184
Appendix II: Jiang N, Leach L, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey M, Luo Z (2008) Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, 9(1): 284-293.....199
Appendix III: Jiang N, Wang M, Jia T, Wang L, Leach L, Hackett C, Marshall D, Luo Z. (2011) A robust statistical method for association-based eQTL analysis. *PLoS One*, 6(8): e23192.....208
Appendix IV: Wang M, Jia T, Jiang N, Wang L, Hu X, Luo Z. (2010) Inferring linkage disequilibrium from non-random samples. *BMC Genomics* 11: 328-340. 218
Appendix V: Wang M, Jiang N, Jia T, Leach L, Cockram J, Waugh R, Ramsay L, Thomas B, Luo Z. (2012) Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *TAG Theoretical and Applied Genetics,* 124, 233-246. 219

List of tables

Chapter I

Table I.1: Allele frequencies of QTL and genetic marker 9
Table I.2: Observed and expected haplotype frequencies for QTL and genetic marker... 9
Table I.3: Comparison between linkage and linkage disequilibrium based approaches to explore the causal genes for complex traits..... 14
Table I.4: Main biological and technical characters of genetic markers..... 18
Table I.5: Commercially available high-throughput genotyping platforms..... 20
Table I.6: Commercially available parallel, high-throughput gene expression analysis platforms..... 38

Chapter II

Table II.1: Transmitted and nontransmitted marker alleles M_1 and M_2 from $2n$ parents to n affected offsprings..... 54
Table II.2: genotype distribution of allele M_1 in case-control samples..... 56
Table II.3: Probability distribution of joint genotypes at a test marker and a putative QTL and genotypic values at the QTL..... 64
Table II.4: Parameters defining two subpopulations that are merged to produce admixed 75
Table II.5: Means and standard errors of regression coefficients (b±se) and proportions (..... 79
Table II.6: The values of regression coefficient (*b*) and empirical statistical power estimated from Method 3 when a proportion of individuals *r* was correctly assigned to subpopulations while (1- *r*) **were incorrectly assigned**..... 83
Table II.7: The values of regression coefficient (*b*) and empirical statistical power estimated from Method 3 when a proportion of individuals *r* was correctly assigned to subpopulations while (1- *r*) **were randomly assigned (50% probability to each subpopulation)**..... 84
Table II.8: The values of regression coefficient (*b*) and empirical statistical power estimated from Method 3 when a proportion of individuals *r* was correctly assigned to subpopulations while for (1- *r*) population membership information was **partially known (60%)**..... 85
Table II.9: Investigating the ability of control marker that can remove 'spurious association' between two subpopulations 87
Table II.10: Predicted and observed proportions of significant tests of linkage disequilibrium between a test marker and a putative QTL in different simulation populations without population stratification from Method 1 in which the control marker implemented into the analyses had a constant allele frequency difference of 0.4. 89
Table II.11: Basic features of Affymetrix Human Genome Focus Target Array 92
Table II.12: Genotyping platforms and research institutes in HapMap Project 93

Table II.13: The number of eQTLs detected by three different methods (**Methods 1, 2, 3** or **M1, 2, 3 accordingly**) or detected common between two of these methods from the CEU, CHB+JPT and their mixed samples. 99
Table II.14: The 51 cis-eQTLs predicted by **Method 1** from the mixed sample 103

Chapter IV

Table IV.1: Data resources and information of 10 selected model species 133
Table IV.2: Datasets of the homologous genes for 6 higher vertebrate species..... 136
Table IV.3: Datasets from Gene Ontology database for 6 higher vertebrate species ... 138
Table IV.4: GC content and distribution of CpG sites in vertebrates and invertebrates 141
Table IV.5: Conservation of two classes of promoters across higher vertebrates..... 152
Table IV.6: A list of conserved and overrepresented GO terms for HCP and LCP classes in the 6 higher vertebrates 163

List of figures

Figure I.1: Diagrammatic illustration for the process of gene expression 27
Figure II.1: Comparison of allele frequencies between populations for all SNP markers genotyped in the International HapMap Project. 96
Figure II.2: The first 2 Principal Components from PCA of 142 mixed HapMap Project human samples. 97
Figure II.3: Manhattan plots for the genome-wide eQTL analysis of two genes POMZP3 and HSD17B12; Quantile-quantile (QQ) plots to compare the distributions between expected and observed p-values. 102
Figure II.4: Histograms of coefficient of determination for eQTLs from 142 mixed sample set. 106
Figure IV.1: Addition of a methyl group to the cytosine base in DNA sequence 127
Figure IV.2: Expected and observed proportions of CpGs across the different genomic feature regions of 10 model species. 143
Figure IV.3: Histograms of the CpG sites proportions in the promoters of the 10 model species 146
Figure IV.4: Histograms of GC fractions in the promoters of the 10 model species ... 147
Figure IV.5: Distributions of the CpG sites with respect to transcription start site (TSS) 149
Figure IV.6: Distributions of GC fractions with respect to transcription start site (TSS) 150
Figure IV.7: Cluster analysis of higher vertebrate species 154
Figure IV.8: Distribution of methylation patterns across 28 different human tissues .. 157

List of Abbreviations

BMI	body mass index
cM	centimorgan
d.f.	degrees of freedom
GO	gene ontology
GTF(s)	general transcription factor(s)
GWAS	genome-wide association study
LD	linkage disequilibrium
LOD score	logarithm of odds (the base-10 logarithm of a likelihood ratio)
Myrs	million years
NGS	next generation sequencing
ORF	open reading frame
QTL(s)	quantitative trait locus (loci)
SNPs	single nucleotide polymorphisms
TSS	transcription start site
UTR(s)	untranslated region(s)

Part I

Developing the linkage disequilibrium (LD) based association method for eQTL analysis

Chapter I

Overall introduction: association mapping, gene transcription process and expression quantitative trait loci (eQTLs) analysis

1. 1 Related publications

Druka A, Potokina E, Luo Z, **Jiang N**, Chen X, Kearsey M, Waugh R (2010) Expression quantitative trait loci analysis in plants. Plant Biotechnology Journal, 8(1):10-27.

Please see Appendix Ⅰ for a copy of this article as it appeared in print.

Jiang N, Leach L, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey M, Luo Z (2008) Methods for evaluating gene expression from Affymetrix microarray datasets. BMC Bioinformatics, 9(1)**:** 284-293.

Please see Appendix II for a copy of this article as it appeared in print.

1.2 Overview

The phenomenon of association between two bi-allelic loci has been initially mentioned by Robbins in 1918. He hypothesized that 'association is the constant allele frequencies subject to no pressure except recombination' (Robbins, 1918). Based on this proposition,

association between two loci should ‘continuously debilitate from an initial level to an asymptote at zero’ (Collins, 2007). The first statistical models to calculate haplotype frequencies and association parameters for two bi-allelic markers were developed by Bennett and Hill. In 1965, Bennett introduced a theory for estimating the haplotype frequencies between linked gene-pairs in a randomly mating population. And then, Hill and Robertson developed a method to calculate the coefficient of association (or linkage disequilibrium) in finite population in 1968. During the last 50 years, the genetic association between two loci has become a commonly used tool for dissecting the genetic architecture of complex traits. The phenotypic differences of complex traits among individuals are controlled by multiple genes plus several non-genetic factors. Although comprehensive understanding the genetic basis of complex traits is not an easy task, association studies have successfully predicted a lot of genes with influences on many specific interesting phenotypes, such as drug response, disease susceptibility and crop yield. The principle of association mapping is to localize genes that affect the complex traits through associating nearby genetic polymorphisms, such as Single Nucleotide Polymorphisms (SNPs), Microsatellite polymorphisms and recently developed Single feature polymorphisms (SFPs). Currently, with the development of sequencing and genotyping technologies, the fast growth database of genetic polymorphisms makes development of the linkage disequilibrium based strategy for genome-wide mapping complex traits an interesting and useful research topic in functional genomics in plants, animals and humans.

1.3 Complex traits

Complex traits have been considered as the 'unpolished diamonds' for geneticists and medical researchers. Meanwhile, it is also perhaps the largest challenge coming up against geneticists. The clustering of phenotypic traits or diseases such as height, yield, Body Mass Index (BMI), flower time of plant, type 2 diabetes, Parkinson's disease and cancers, among related individuals show that almost all continuously variable traits (or quantitative traits) and common diseases have complicated genetic architectures, which opposes to Mendelian diseases or monogenic controlled traits (Weir, 2008). The Mendelian trait (or disease), which strictly follows the Mendelian inheritance pattern, typically comes from only single gene's mutation that has extremely high penetrance (Collins, 2007). In contrast, complex traits (or complex diseases) are generally affected through genetic mutations at multiple loci across the whole genome, and each locus only contributes relatively low penetrance. There are also many other factors playing very important roles for complex traits, such as epigenetic effect, environmental regulation, Epistasis (interactions between genetic mutations at different loci), $G \times E$ interactions (interactions between genes and environment). Furthermore, complex traits are very common in our daily life, and have significant impact for the geneticists to improve the medical treatments and increase the economic performance in farming area. Two of the most popular methods (or strategies) for dissecting the genetic architecture of complex traits are linkage analysis and association study. Linkage analysis is the traditional method used to sketchily detect the co-segregation of a small genomic region and a trait of interest in families or pedigrees of known ancestry. On the contrary, association study is a precision and high-resolution method for mapping the casual

genes (or loci) underlying complex traits based on Linkage Disequilibrium (LD) in population(s). During last decades, both linkage mapping and association study have been widely applied for many complex traits (or diseases); and, of course, some historic findings have already been accomplished (Australia and New Zealand Multiple Sclerosis Genetics, 2009, Carlson et al., 2004, Easton et al., 2007, Fung et al., 2006, Giallourakis, 2003).

1.4 Genetic linkage and linkage analysis

Genetic linkage represents the trend of two (or more than two) genes that are closely located with each other on the same chromosome to be co-inherited from one generation to another. For typically diploid species, such as human and most vertebrates, the meiosis reduces the genome to product haploid gametes through two cell division processes. However, the most important events, chromosomal recombinations, can take place during prophase I of meiosis. The recombination events exchange sections or fragments of chromosomes; hence genes on the same chromosome can be separated into different gametes. Therefore, the segregating families or pedigrees can be created. Furthermore, the frequency of recombination between two genes on the same chromosome relies on their genetic distance: if the two genes are separately located with long distance, the greater the chance that a recombination event could take place between them; in contrast, if two genes are closely located with each other on the same chromosome, the recombination may occur at fewer chance between them, and genes are thought to be linked together. The comprehensive understanding of the frequency and distribution of recombination events at the chromosomes is essential for linkage analysis.

Linkage analysis is the traditional method used to explore the causal gene of the phenotypic trait. The basic principle of linkage analysis depends on the co-inheritance of the genetic polymorphisms and phenotypic trait of interest in segregating families or pedigrees. Among individuals from segregating families, a phenotypic trait (or disease) may occur randomly with some genetic polymorphisms, or correlatively with some genetic variants. In the former, the phenotypic trait is considered as be independent of

the genetic polymorphisms. However, the latter is identified as 'genetic linkage'. It indicates that the causal gene underlying this phenotypic trait must be closely located to the linked genetic polymorphisms. The linkage analysis has been immensely successful for identifying the genes that control Mendelian traits (or diseases) (Jimenez-Sanchez et al., 2001). Because the Mendelian trait is, by definition, controlled by a single highly penetrant gene; genetic polymorphisms within 5-10 cM of the causal gene will be highly co-inherited with the phenotypic trait (or disease) status in segregating families (Hirschhorn and Daly, 2005). According to the tremendous success in mapping the susceptible genes of Mendelian diseases, the linkage analysis has also been applied for many complex traits, such as breast and ovarian cancers (Easton et al.), type I diabetes (Nistico, 1996) and heart diseases (Hauser and Boehnke, 1998). But, for most of complex traits (or diseases), linkage studies have only accomplished limited success (Altmuller et al., 2001), the genes which identified by the linkage analysis approaches generally explain only a small proportion of the total heritability of the complex traits. The lack of success can be contributed to several factors. Most importantly, linkage study needs the family-based samples, and the low penetrance of complex disease means that vast amounts of information are required during the analysing. However, the adequate size of family-based samples is not available in reality of many applications. Second, even if the candidate regions can be identified, the resolution of linkage studies is typically on the order of a few cM (usually 5-10 cM), which in terms of the organism genome may correspond to hundreds or thousands of genes. It is very difficult to precisely identify the causal gene from such large of candidate genes' group. Furthermore, the linkage studies are also much less powerful for detecting the

susceptible genes that have modest effects on phenotypic traits (or diseases) (Risch and Merikangas, 1996, Risch, 2000).

1.5 Linkage disequilibrium and association studies

Recently, Linkage Disequilibrium (LD) has been considered as one of the most important topics in population genetics. Linkage disequilibrium generally refers to the 'non-random association of alleles' at two (or more) linked and unlinked genes (or loci) in a population. The estimation of LD coefficient between genetic markers can offer useful information for identifying the alternative evolutionary patterns of genomic variations within or between populations (Lewontin and Kojima, 1960). The current stage of Linkage Disequilibrium (LD) based association mapping focuses on exploration of causal genetic polymorphisms (or mutants) of complex traits (or diseases) in plants, animals and human beings (Easton et al., 2007, Remington et al., 2001, Valdar et al., 2006). The key point of these analyses is the estimation or inference of LD between genetic polymorphisms and functional genes (or loci) that are genetically linked in short distance.

1.5.1 Definition of LD

Linkage Disequilibrium (LD) is not the same as genetic linkage, which represents the combination of two (or more) loci on the chromosome(s). In population genetics, LD represents the observed haplotype frequencies of alleles at two (or more) genetic markers less often or more often than would be expected based on their allele frequencies in a random mating population. Let us consider a simple two bi-allelic loci model: one locus affect a quantitative trait (QTL), and the other one is polymorphic

markers completely lacking of effect on this particular trait. These two alleles are represented by A and a at QTL, by M and m at genetic marker locus. The allelic frequencies for each allele at each locus are listed in table I.1 respectively.

Table I.1: Allele frequencies of QTL and genetic marker

Locus	QTL		Genetic marker	
Alleles	A	a	M	m
Frequencies	f_A	f_a	f_M	f_m

Accordingly, the observed and expected haplotype frequencies of each combined alleles between QTL and genetic marker in a random mating population are shown in table I.2.

Table I.2: Observed and expected haplotype frequencies for QTL and genetic marker

Gamete	MA	Ma	mA	ma
Observed frequencies	f_{MA}	f_{Ma}	f_{mA}	f_{ma}
Expected frequencies	$f_M \cdot f_A$	$f_M \cdot f_a$	$f_m \cdot f_A$	$f_m \cdot f_a$
Discrepancy	$f_{MA} - f_M \cdot f_A$	$f_{Ma} - f_M \cdot f_a$	$f_{mA} - f_m \cdot f_A$	$f_{ma} - f_m f_a$

The discrepancy between the observed and expected frequencies of a haplotype is defined as Linkage Disequilibrium and generally abbreviated to ' D ', i.e. $D = f_{MA} - f_M \cdot f_A$, where f_{MA}, f_M and f_A are frequencies of gamete MA, alleles M and A in this population. This LD measurement indicates that: if the two loci (QTL and marker) are independently segregating in the population, the observed haplotype frequency of MA would equal to $f_M \cdot f_A$ hence $D = 0$; on the contrary, if the alleles at QTL and marker are not independently segregating, the observed frequency of MA would not equal to $f_M \cdot f_A$ hence $D \neq 0$. Although the linkage disequilibrium parameter

D is easy to be estimated, the scale range of D varies depending on the allele frequencies in the population. The theoretical maximum of D can be given by:

$$D_{\max} = \begin{cases} \min(f_M \cdot f_A, (1-f_M)(1-f_A)) & \text{when } D < 0 \\ \min(f_M(1-f_A), (1-f_M)f_A) & \text{when } D > 0 \end{cases}. \qquad \text{(I-5.1)}$$

The equation I-5.1 shows that the range of D can be maximal (=0.25), only when the allele frequencies are 0.5 at both QTL and genetic marker loci. Thus, it is not easy to directly compare the linkage disequilibrium degrees across different loci due to the variation in scale. In 1964, Lewontin introduced a normalization method to divide the standard estimation D by the theoretical maximum $D_{\max}$:

$$D' = \frac{|D|}{D_{\max}} \qquad \text{(I-5.2)}$$

After the normalization, the adjusted D' always varies from 0 to 1 irrespective of the different allele frequencies.

Furthermore, the LD between two genetic loci in a population changes across different generations. In the random mating population without any evolutionary disturbance, such as selection, mutation, migration and genetic drift, the LD between two loci will progressively converge to 0 along the generations at a rate depending on recombination frequency r between these two loci,

$$D_n = (1-r)D_{n-1} \qquad \text{(I-5.3)}$$

where D_n and D_{n-1} are the LD at the n^{th} and $(n-1)^{th}$ generation, r denotes frequency of recombination between the two loci. After a large number of generations, only those closely linked loci, with small magnitude of r, can maintain a significant D. Based on this fact, the Linkage Disequilibrium (LD) is an ideal parameter to detect the genetic

association between markers and casual genes (or loci) of complex traits with high resolution.

1.5.2 Genetic association studies

Genetic association study is a statistical approach of identifying causal genes (or loci) regulating complex traits that employs the historic association (or LD) to link the phenotypic traits to genetic polymorphisms. As mentioned above, the association study is based on the idea that only the causal gene that is physically close to a genetic marker will still be retained in a relatively significant LD after a number of generations. In the last decade, there were many LD based studies attempted to map the genes that underlies complex traits (diseases). All of the mapping methods broadly fall into two classes: 'candidate-gene studies' and 'whole genome studies'.

The 'candidate-gene' based association study is hypothesis-based analysis. The 'candidate genes' are selected for association mapping, either by their location in a genomic region that has been roughly identified via linkage analysis, or on the prior information that they might be related to the complex trait. And then, in a group of population-based samples, one can search and genotype one or several common genetic polymorphism(s) that are located within the candidate genes. Finally, the association study is applied to each of the genetic markers and complex trait. In the simplest form, association mappings contrast the allele frequencies of a particular genetic marker between case and control sampling groups to identify the gene that affects complex trait (Weir, 2008). Since then, many different statistical methods have been developed for association studies, such as Pearson χ^2 test, Armitage's trend test, linear regression

model and likelihood-based strategies for association mapping. Until now, the candidate-gene association mappings have confirmed a lot of genes which are already annotated to contribute to susceptibility of complex traits (diseases) (Hirschhorn and Altshuler, 2002, Lohmueller et al., 2003). But, candidate-gene association studies depend heavily on prior information, usually on particular biological assumptions or the location of the candidate genes within a previously identified region of linkage. So, if the prior information of complex traits (diseases) is unknown, the candidate-gene studies will obviously be unable to explore the genetic architectures of complex traits (diseases).

Alternatively, whole-genome association study, also called as genome-wide association study, is an approach for examining genome-wide genetic polymorphisms in a group of population-based samples to identify whether any genetic marker is associated with a given phenotypic trait. Due to no assumptions have to be made about the location of the candidate genes, this strategy has to survey the whole genome or most of genome for causal genetic variants (or loci). Hence, the genome-wide association study provides a comprehensive option that can be attempted even in the absence of prior information regarding the genomic location of the causal genes.

1.5.3 Advantages of association studies

Compared with traditional linkage analysis, the association studies have several advantages (table I.3). First, linkage studies capture the co-inheritance of causal genes and adjacent genetic polymorphisms within family-based samples or pedigrees of clearly known ancestry. However, association study uses diagnosis of existing natural populations versus the need to generate a segregating population. So, it is feasible to

collect much larger samples for association mapping. Second, population-based samples in association study include historical recombination events and evolutionary segregations at the population level. The ancestry might extend to over thousands of generations. Therefore, the association study is more powerful for identifying causal genes (or loci) that have modest effect on complex trait. Third, in association studies, the mapping resolution is dramatically increased by use of high density genetic markers (e.g. SNPs); association analysis can offer a fine mapping resolution <1cM so that the candidate gene can be directly identified in small DNA sequence regions. It is an enormous improvement over the linkage studies where the candidate gene can hardly be narrowed down to such a resolution. Overall, the LD-based association study is a more precise and higher-resolution approach than traditional linkage analysis in dissecting the genetic architectures of complex traits.

Table I.3: Comparison between linkage and linkage disequilibrium based approaches to explore the causal genes for complex traits

Potential advantages	Candidate-gene studies		Genome-wide studies	
	Linkage analysis	Association mapping	Linkage analysis	Association mapping
No prior information required	×	×	✓	✓
location to small region (high resolution)	×	✓	×	✓
Inexpensive	✓	✓	×/✓	×
Pedigree not required	×	✓	×	✓
Power to detect common alleles of modest effect [a]	×/✓	✓	×/✓	✓
Ability to detect rare allele	✓	×	✓	×
Tools for analysis available	✓	✓	✓	×

[a] Common allele: minor allele frequency >5%

Symbols indicate whether the potential advantage in the left column applies complete (✓), weakly (×/✓), not at all (×).

1.5.4 Population structure: A big challenge in association study

However, the primary hindrance to successfully using association approach is underlying population structure among the collected samples. The significant association may be detected between two genetic markers that have no physical linkage with each other, when samples are collected from different populations.

Population structure, also known as population stratification, represents the systematic variation in allele frequencies across subpopulations, due to different evolutionary histories and varied ancestry backgrounds. In 1988, Chakraborty and Smouse formulated the coefficient of linkage disequilibrium between two loci in admixed populations. Consider a structured population generated from instant admixture of two

genetically divergent random mating populations, the proportion of subpopulation 1 in the mixed population is denoted by *m*. There are two bi-allelic markers with alleles *A, a* and *M*, *m* respectively. In subpopulation *i* (*i*=1 or 2), the allele frequencies are denoted by $f_A^{(i)}$, $f_a^{(i)}$ and $f_M^{(i)}$, $f_m^{(i)}$ respectively. The linkage disequilibrium (LD) in each subpopulation can be expressed as:

$$D^{(1)} = f^{(1)}(AM) - f_A^{(1)} f_M^{(1)} \qquad \text{(I-5.4)}$$

$$D^{(2)} = f^{(2)}(AM) - f_A^{(2)} f_M^{(2)} \qquad \text{(I-5.5)}$$

where $f^{(1)}(AM)$ and $f^{(2)}(AM)$ are the haplotype frequencies of *AM* in subpopulation 1 and 2. However, the LD between two genetic markers in the admixed population is given by:

$$\begin{aligned} D_{admixed} &= f_{admixed}(AM) - f_A f_M \\ &= mD^{(1)} + (1-m)D^{(2)} + m(1-m)\delta_A \delta_M \end{aligned} \qquad \text{(I-5.6)}$$

where $\delta_A = f_A^{(1)} - f_A^{(2)}$ and $\delta_M = f_M^{(1)} - f_M^{(2)}$ (Chakraborty, 1988). Equation I-5.6 shows that the coefficient of LD in admixed population is the summation of (a) a linear combination of the LD between the two loci in each of the subpopulations (i.e. the genuine LD between the two loci in each of the subpopulations), and (b) a nonlinear component of the difference in allele frequencies between the two subpopulations (the spurious LD). When these two genetic markers are actually in linkage equilibrium (LE) within each subpopulation ($D^{(1)} = 0$ and $D^{(2)} = 0$), LD in admixed population may still be observed due to the spurious term. And the magnitude of spurious LD is the product of the subpopulation proportions and the allelic frequency differences. During association studies, the spurious term in equation I-5.6 can seriously lead to both false-positive and false-negative results if not correctly addressed.

To avoid this, one apparent way is to get some non-substructure samples, but unfortunately it is very hard for most of the time, especially when the sample used is fairly large. So how to remove the confound effect of population structure in association mapping is essential for an appropriate setting of GWAS. In chapter II, I will introduce a novel approach to control the confound effect in association studies.

1.6 Genetic markers and high-throughput genotyping technologies

A genetic marker is a short section of DNA sequence with a known genomic location that can be used to describe polymorphisms and segregating individuals at the molecular level. The length of genetic marker can vary from single nucleotide base-pair (such as Single Nucleotide Polymorphisms) to several hundred nucleotide bases (Microsatellite polymorphisms), even as long as over thousand nucleotides for Minisatellite Polymorphic markers. No matter how long it is, the ideal genetic maker should have two characters. First, the genotyping process is very easy to be implemented and have as low an experiment cost as possible. Second, the genetic markers should be high-density and independently distributed across the whole genome. Due to the various molecular biology techniques, and to various biological information can be obtained, a series of genetic markers have been discovered. According to the types of information can be provided, genetic markers can be divided into three main categories: the bi-allelic dominant markers, such as Amplified Fragment Length Polymorphisms (AFLPs); the bi-allelic co-dominant markers, such as Single Nucleotide Polymorphisms (SNPs), Single Stranded Conformation Polymorphisms (SSCPs), Restriction Fragment Length Polymorphisms (RFLPs) and the multi-allelic co-dominant, such as Microsatellite Polymorphisms, Minisatellite Polymorphisms. Here, I briefly listed the biological and technical characters of these different types of genetic markers in table I. 4.

Table I.4: Main biological and technical characters of genetic markers

Genetic marker	Variants	Density	Genotyping accuracy
AFLPs	bi-allelic dominant	Very Low	Medium
RAPDs	bi-allelic dominant	Very Low	Very Low
RFLPs	bi-allelic co-dominant	High	Very High
SSCPs	bi-allelic co-dominant	Medium	Medium
SNPs	bi-allelic co-dominant	Very high	Very High
Microsatellite	multi-allelic co-dominant	Low	High
Minisatellite	multi-allelic co-dominant	Low	High

During last 20 years, the use of genetic markers has played a more and more important part in genetics researches. Among all available genetic markers, SNPs marker is the most popular. The great interest in SNPs has basically come from the recent requirement of very high-density genetic markers for genome-wide association studies to detect the casual variants of complex traits. As indicated by the acronym, a SNPs (Single Nucleotide Polymorphism) marker is only a single nucleotide base-pair mutation (or variant) in DNA sequence. This type of bi-allelic co-dominant markers widely and abundantly spreads across the complete genome. For example, in International HapMap Project, there are over 3.1 million SNPs markers identified in the human genome, when combining Phase I and II releases. In average, one SNP can be genotyped at every 1,000 nucleotide base-pairs. Furthermore, the recent development of

high-throughput SNPs genotyping technologies is another reason for the increasing popularity of SNPs. Compared with traditional genotyping procedure which involves PCR and denatured gel electrophoresis, the high-throughput SNP genotyping technology enables the genotyping procedure to be simultaneously implemented for hundreds of thousands of candidate SNPs, which has made the whole genome-wide association mapping in complex traits to be practised for the first time.

1.6.1 Genotyping SNP markers using microarray platforms

During the last decade, there is a revolution occurring in SNPs genotyping technology: the microarray is able to accurately genotype hundreds of thousands of SNPs in large cohort studies. Currently, this type of high-throughput genotyping technology has been commercially developed by many biotech companies (table I.6). All of the different commercial microarrays can detect as many genetic polymorphisms as can be designed on the array and are easily applied to automatic running (Grant and Hakonarson, 2008). In this section, I focus on the Affymetrix technology and briefly review how the microarray accurately genotypes a large number of SNPs at once (other different commercial microarray genotyping technologies are not discussed here but has been listed in table I.5).

In Affymetrix platform, the input DNA sequences are cut by the restriction enzymes *Sty*I and *Nsp*I and bound to adaptors that recognize the 4-bps overhangs. All DNA segments represent substrates for adaptor ligation; a generic primer that can tie to the adaptor sequence is used to amplify these DNA segments. PCR technique has been used to preferentially amplify fragments in the 200 bps to 1100 bps size range. And then, the

amplified DNA fragments are fluorescently labelled, and hybridized to an Affymetrix DNA genotyping microarray. The chips are washed and scanned automatically; the genotypes are analyzed using standard algorithm GeneChip Genotyping Analysis Software (GTYPE) which developed by the Affymetrix company (http://www.affymetrix.com/support/technical/index.affx). The more detailed information can be learned from www.affymetrix.com.

Table I.5: Commercially available high-throughput genotyping platforms

Company	Detection method	No. SNPs detected simultaneously
Affymetrix	Fluorescence; hybridization to array	10,000~100,000
Illumina	Fluorescence; tags on bead-array	100,000+
ABI	Fluorescence; gel electrophoresis	200,000
Parallele	Fluorescence; tags on array	10,000
Sequenom	Mass spectrometry	1,000
Perlegen	Fluorescence; hybridization to array	100,000+
Third Wave	Fluorescence; plate reader	few

1.6.2 Discovering and genotyping SNPs based on NGS technologies

Over the past decade, several advanced 'high-throughput' sequencing technologies, called Next Generation Sequencing (NGS), have been developed and applied for a variety of applications (Metzker, 2010). These new sequencing methods are able to sequence large and complex genomes much faster, cheaper and more accurate than previous methods, for example Sanger sequencing. A common theme of these NGS platforms is that the whole-genome wide sequences are sheared into small pieces (200-

300 bps), and then the fragments are fixed or attached to a solid support. The fixation of independent fragments into different sites allows more than millions of sequencing reactions to be carried out simultaneously. The resulting data of these new methods consists of millions of short-piece reads (30-200 bps) from unknown locations in the genome. Analysis of these datasets brings an unprecedented computational and informatics challenge, both because of the large amounts of short-reads that the new platforms can generate, and because the length of reads is significantly shorter than previously generated from traditional methods. The statistical methods for assembling (or aligning) short-reads have been developed over the last 4 years, and while some of the short-read mapping problems are now available in form of computer software. Meanwhile, it has seen an accelerating flurry of publications in which the NGS platform is widely applied for a variety of studies. One of the most important applications is the whole genome-wide sequencing, and then identifying mutations or genetic polymorphisms, such as SNPs.

From 2008, a series of programs have been developed for identifying SNPs based on NGS sequencing data, such as SOAPsnp (Li et. al, 2009), MAQ (Li et. al, 2008), SAMtools (the most widely used) (Li et. al, 2009). The most serious challenge for SNP identification lies in inferring the likelihood that a nucleotide position is a homozygous or heterozygous variant given the error ratios of the differential platforms, the probability of incorrect read alignment, and the depth of coverage. The SNP discovery analysis after the alignment generally has a standard procedure in which it filters the mapped short-reads and re-scores the nucleotide base quality values, then followed by a consensus genotype calling step through a Bayesian model. The Bayesian model is to compute the conditional probabilities of the genotype at each nucleotide position:

$P(G|R) = \frac{P(R|G)P(G)}{P(R)}$. This equation shows that one can infer the probability of a particular genotype *G* given the observation data *R* (posterior) given the overall probability of that genotype (prior probability) and the probability of observation data *R* from this genotype. The details of Bayesian approach have been implemented in MAQ consensus genotype calling section (Li et. al, 2008). Although all of the SOAPsnp, MAQ and SAMtools methods use the same Bayesian model to detect SNPs, they are widely diverse in the details and use different interpretations of statistics. For assigning a posterior probability of a certain genotype, MAQ and SAMtools calculate a probability of observing the given read and base quality values for each genotype prior via a binomial distribution; meanwhile, SOAPsnp computes the probability based on various characters of the reads. Finally, the SNPs are identified through comparing the consensus genotype calling results with reference sequence. It is believed that NGS technology will be widely applied to this purpose in the near future.

1.6.3 Prediction of genetic polymorphisms from expression microarray dataset

With the development of the microarray technology, the markers' genotyping and their chromosomal locations can be simultaneously implemented with the expression data in expression microarray experiments. The ability to identify sequence polymorphisms from gene expression microarray data has useful implications in at least two aspects. First, it improves both accuracy and precision in calculating gene expression indices by excluding probes containing genetic polymorphisms, while in turn, improving the statistical power of eQTL analysis (Alberts et al., 2007). Second, it enables the concomitant generation of an abundant collection of reliable genetic markers that can be

used as the framework for subsequent eQTL genetic analysis (Luo et al., 2007, Borevitz et al., 2003). In 2003, a type of marker called 'Single Feature Polymorphism (SFP)' has been derived from Affymetrix expression microarray experiments (Borevitz et al., 2003).

To distinguish hybridization signals associated with any molecular alteration from background, the signals generally need to be collected from the sequences where mutation occurs. The probe-set design of Affymetrix microarrays perfectly meets this need. Statistical methods have been developed to detect polymorphisms between target sequences and probes by testing for non-uniformity of hybridization intensity among every feature in a probe-set for a given gene. The principle for RNA based templates is an extension of that described by Winzeler et al. (1998) who pioneered in the development of a high-throughput genotyping platform by hybridization of labelled total genomic DNA to oligonucleotide arrays.

By attempting to integrate genetic polymorphism screening and gene expression analysis, (Wang et al., 2009)proposed a Bayesian statistical approach for detecting SFP's in transcript sequences and for predicting SFP genotypes when tested in a segregating population derived from genetically divergent parental lines. This was achieved by modelling a perfect match value from Affymetrix cRNA hybridization experiments as a product of the binding affinity between the transcript and probe sequences and the abundance of the transcript. They analyzed two independent microarray datasets (RNA hybridisations from barley and yeast) and demonstrated that their method provided significantly improved robustness and accuracy for predicting SFPs reflecting genuine sequence polymorphism, when compared to five other statistical methods. Their method was appropriate for predicting SFPs from expression

microarray data and from genomic DNA microarray data. By comparing predicted SFPs to those where sequence information was available, they showed that all the methods applied stringent selection criteria to call SFPs and thus only a small fraction of probes were called as SFPs. The approach effectively maintained both false positive and false negative rates at a low level.

1.7 Gene expression process

Gene expression is a complicated biological process in which the heritable information units stored in the genome, called genes, are used to synthesize the functional products (Khaitovich et al., 2006, Hegde and Kang, 2008, Kuznetsov et al., 2002, Schwanhausser et al., 2011, Tomilin, 2008, Xu et al., 2007). In genetics, gene expression process plays an essential role through which the genotype leads to the phenotype: the genetic information conserved in the DNA sequence can be exported through expression process, and then the products of gene expression will guide the phenotype of organisms (Brawand et al., 2011, Robertson, 2010, Tuch et al., 2008). Briefly, gene expression has two major steps (figure I.1): a) from DNA sequence to mature RNA molecule—Transcription process; b) from mature RNA to protein product—Protein Synthesis.

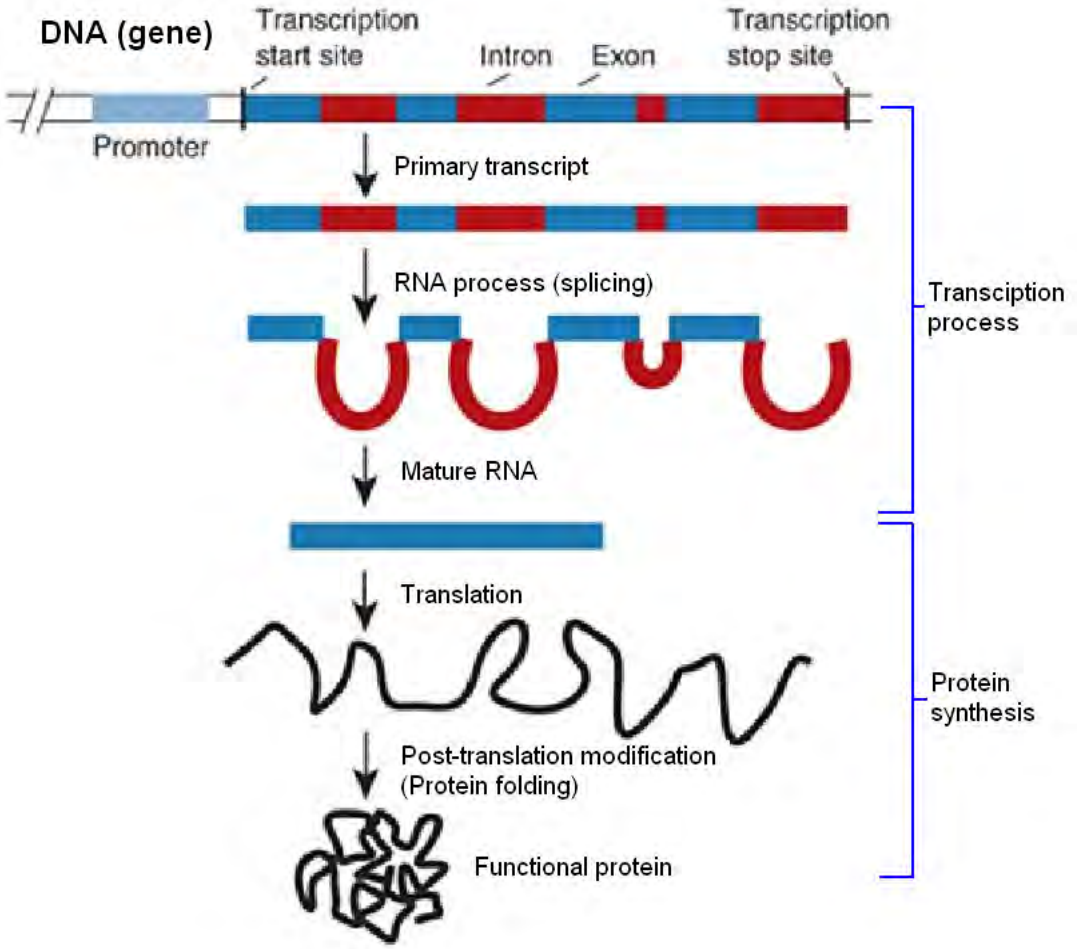
In the transcription process, each gene is separately transcribed to create a primary RNA molecule with essentially the complementary copy of the original DNA sequence (Gnatt et al., 2001, Clancy, 2008a). In most eukaryotes, the gene usually contains both *exons* and *introns*, and only the *exons* store the functional information for further biological processes. Thus, the primary RNA sequences need to be modified by several RNA processes, such as splicing, to remove the *introns* and produce mature RNA molecules or messenger RNAs (mRNAs) which only contain *exons* (Nilsen and Graveley, 2010). The mature transcription product can be directly used as micromolecular machinery for life, such as ribosomal RNA (rRNA), transfer RNA (tRNA), microRNA and siRNA; or, it is considered as intermediate, messenger RNA (mRNA), to be further translated into

protein product (Yusupova et al., 2001, Smith and Shaw, 2008, Clancy and Brown, 2008).

The second step of gene expression is protein biosynthesis. As there is no direct commensurateness (or correspondence) between the nucleotide sequence of RNA molecule and the amino acid sequence of protein, the messenger RNA has to be translated into amino acids (Clancy and Brown, 2008, Laursen et al., 2005). During translation, each three nucleotide bases of an RNA sequence generated by transcription is deciphered by the ribosome to yield a specific amino acid; hence a complete messenger RNA molecule can be decrypted into an amino acid chain, also called a polypeptide chain (Watson, 1963). And then, the amino acids in a linear chain will interact with each other to fold into a three dimensional (3D) structure (Steitz, 2008). The appropriate 3D structure is essential for the protein to preserve its biological function and activity (Panse and Johnson, 2010). Finally, the structured proteins need to be transported to some parts of cells (or tissues, organs) where they are supposed to be (Yusupova et al., 2001, Clancy and Brown, 2008).

The gene expression process is an efficient system for interpreting the genetic information into functional products. Furthermore, the expression of a gene can be regulated in every step, from primary transcription, RNA processing, to translation and post-translation modification (Robertson, 2010, Mattick et al., 2009, Hegde and Kang, 2008). The regulation of gene expression allows organisms to control how the genetic information is used. Overall, gene expression and its regulation are fundamental processes for organisms to maintain their biological behaviours and life (Khaitovich et al., 2006).

Figure I.1: Diagrammatic illustration for the process of gene expression



1.8 Gene transcription

Transcription is the first fundamental step of the gene expression process, which can produce an RNA copy based on the genetic information of a functional gene (Khaitovich et al., 2006, Gnatt et al., 2001, Clancy, 2008a, Clancy, 2008b, Bell and Jackson, 1998). Since both RNA and DNA molecules are nucleic acid sequences, the DNA sequence can be directly read by appropriate RNA polymerases, and then converted into an exactly anti-parallel nucleotide sequence. During transcription, DNA sequence is read and transcribed from 3' to 5' direction; whereas, the RNA sequence is produced from 5' to 3' end (Clancy, 2008a). Although DNA molecule is structured as two anti-strands in double helix format, only one strand, referred to as the 'template strand', can be transcribed. Meanwhile, the other unused strand is referred as the 'coding strand', because its nucleotide codes are the same as the produced RNA sequence (Clancy, 2008c). The transcription process is very highly similar with DNA replication. The most obvious difference is that the nucleotide sequence in RNA molecule uses uracil (U) to substitute all thymine (T) during transcription.

The primary transcription products can be divided into two groups for further processes (Watson, 1963, Clancy, 2008b, Ralston and Kenna, 2008). If genes encode for some non-coding functional RNAs, the primary transcripts need to be converted into mature RNAs via several RNA processes. Alternatively, the majority of genes in the genome are protein-coding genes. The transcripts from these genes will then produce proteins through the translation processes. However, the primary transcription products includes not only the coding sequences that will be decoded into proteins (*exons*) but also some

other sequences, such as the untranslated regions at both 5' and 3'ends (termed as 5'UTR and 3'UTR), *introns* and poly-A end. The untranslated regions at both 5' and 3'ends have been identified as regulatory sequences that do not translate into protein but play important roles for regulating both transcription and protein synthesis (McCarthy, 1998). And the *introns* need to be removed from messenger RNAs via a process of splicing (Nilsen and Graveley, 2010). Therefore, gene transcription is a complex but accurately regulated process. Briefly, the transcription procedure can be summarized in 6 stages: pre-initiation, initiation, promoter clearance, elongation, termination and RNA processing (splicing).

1) Pre-initiation: in both prokaryote and eukaryote organisms, each gene contain a small but essential element— 'promoter'. The promoter region is a stretch of DNA sequence that will facilitate the transcription for this particular gene. Typically, the promoter is located at upstream of the gene which it controls. Furthermore, the core region of a promoter is the minimal section of the promoter required to precisely start the transcription process, which is usually found at around 100 nucleotide bps upstream from the transcription start site of the gene (Kim et al., 1997, Smale and Kadonaga, 2003). At the beginning of transcription, the appropriate RNA polymerase needs to bind with the core promoter region to start transcription process (Kornberg, 2007). However, in eukaryote species, the RNA polymerase cannot directly identify the core promoter regions (Lee and Young, 2000). Alternatively, several proteins, termed as transcription factors, will assist the binding of RNA polymerases. The core promoter regions in eukaryotes usually contain signal boxes, such as TATA box (Maston et al., 2006). The TATA box is the binding site of a particular

transcription factor (TATA-binding protein) (Ouhammouch et al., 2003). The TATA-binding protein is also a medium of another transcription factor (transcription factor IID). After the transcription factor IID binds to TATA box through TATA-binding protein, the appropriate RNA polymerase can recognize and bind with the core promoter to generate a Pre-initiation Complex. While, in prokaryotes, the RNA polymerase can directly recognize and bind with the signal box (Pribnow box) in core promoter regions to initiate the transcription (Schaller et al., 1975).

2) Initiation: after the RNA polymerase binding to the core promoter region, the transcription process can be initiated immediately (Clancy, 2008c). For eukaryotes, there have three different types of RNA polymerases, and each of them can only bind to the core promoter regions of particular genes: RNA polymerase I only recognizes the core promoter region of ribosomal RNA genes; the RNA polymerase II can bind with both protein–coding genes and some non-coding RNA genes, including snRNA, snoRNA and long non-coding RNA; RNA polymerase III is responsible for 5S rRNA and transfer RNA genes (Lee et al., 2004, Hurwitz, 2005, Grummt and Kivie, 1998, Willis, 1993). However, in prokaryotes, it has only one type of RNA polymerase, which can initiate the transcription process for all types of genes (Ebright, 2000).

3) Promoter clearance: when the first nucleic acid of transcription sequence is synthesized, the RNA polymerase may be cleaved from the promoter region (Pal et al., 2001). At this time, the RNA transcript is possibly to be slipped, hence creates a truncated transcripts. This phenomenon is referred as abortive initiation,

which may take place in both prokaryote and eukaryote species. An enzyme, called σ factor, can keep off the abortive initiation and result in the transcription elongation complex. When the transcript sequence is synthesized over 30 nucleotide base pairs, it will not ever slip and elongation process will continue.

4) Elongation: at this step, the corresponding RNA polymerase covers the template strand of DNA sequence and follows the complementary roles ($A \Rightarrow U, C \Rightarrow G, G \Rightarrow C$ and $T \Rightarrow A$) to produce a copy of RNA sequence (Reines et al., 1999, Nudler, 1999). Unlike DNA replication process, the RNA transcription can be engaged by more than one RNA polymerase on a single template strand, hence several rounds of transcription can occur simultaneously and a large amount of RNA molecule products can be rapidly generated from a particular gene. Furthermore, the elongation also has a 'proof-reading' mechanism which can check and substitute the incorrectly transcribed nucleotide bases (Yan et al., 1999).

5) Termination: the transcription process stops when the RNA polymerases reach the terminator region (Richardson, 2002). The transcription terminator is a stretch of DNA sequence that labels the end of gene for transcription process. The transcription termination is a complex step in eukaryote organisms, and so far little is known about this process. However, at the end of termination, it commonly involves a process called polyadenylation, which adds a polyadenylation signal sequence (5'-AAUAAA-3') at the 3' end of the newly created transcription product (Connelly and Manley, 1988).

6) RNA processing: in prokaryotes, the transcripts produced via above steps are ready for directly being translated into protein. However, the primary transcripts in eukaryotes have to experience a series of RNA processes or modifications to become stable and mature RNA molecules, including splicing, 3' cleavage (Proudfoot et al., 2002, Guhaniyogi and Brewer, 2001). The most important RNA process for primary transcripts is RNA splicing. Nearly all of the primary transcripts in eukaryotes consist of both *exons* and *introns* DNA sections. During RNA splicing process, a protein complex, called as spliceosome, can cut the *introns* and joint the neighbouring *exons* together (Black, 2003, Matlin et al., 2005). In some situations, some *exons* segments can also be removed from transcription sequence. It is referred as 'alternative splicing', which can create different protein products from only a single gene. Thus, the 'alternative splicing' extends the variant and complexity of gene transcriptions. Another RNA process is 3' end cleavage. It takes place when the polyadenylation signal sequence presents at the 3'end. In this process, the primary transcripts are cleaved the polyadenylation signal sequence and then added poly A (around 200 adenines) tails at 3' end to protect the RNA molecules from degradation. Moreover, the poly A tail can be bound by several poly A-binding proteins for assisting the RNA export and translation re-initiation. Finally, the mature RNA molecules have to be transported from the nucleus into cytoplasm for the following protein synthesis.

1.9 Expression quantitative trait loci (eQTLs) analysis

Expression quantitative trait loci (eQTLs) are those genetic loci detected by implementing the traditional quantitative trait loci (QTL) analysis to data on the expression level of genes in samples taken from different individuals (tissues or cell lines) in a random mating population or populations with other different segregation structures. Transcript abundance can be analyzed as a complex trait just like other traditional phenotypes such as weight or height. Usually, gene expression information could be simultaneously extracted for most (or all) of genes over the whole genome, and hence there are normally over thousands of gene expression traits recorded in a typical gene expression profile using microarray or sequencing techniques.

Similar to the traditional phenotypic quantitative trait loci (pQTL) analyses, eQTL analysis needs high-density genetic markers which can be genotyped in all collected samples and constructed the whole genome (or genome wide) genetic variation profile for each sample. These markers' genotyping and their chromosomal locations can be investigated in the selected samples before the eQTL study or can be simultaneously developed with the expression data in expression microarray experiments. High quality genetic marker information is one of the key components of such studies, determining the quality of eQTL analyses, and allows exploring the impact of genetic variation on physiological processes through transcriptional regulation networks (Sieberts and Schadt, 2007). The eQTL studies are ultimately to test for genetic association between a specific genetic markers and the expression level of the interested gene. The significance of the genetic association can be reported as the probability values (*P-value*)

of a null hypothesis test, the Log Odds (*LOD*) score or Likelihood-Ratio (*LR*) of a likelihood function. The values of the test statistic (*P-value, LOD* score or *LR*) need to reach a prior given significance level depending on the sample size, population structure and the proportion of non genetic variation for the trait. The significant eQTLs indicate the genetic polymorphisms in these candidate genetic loci (or genes) have a high confidence to cause the observed variation in gene expression level across samples.

Generally, eQTL can be classified into two groups: *cis*-acting eQTL and *trans*-acting eQTL. In the former, the DNA sequence polymorphism regulating expression level is assumed to be regulated by the genetic polymorphism that locate within or in the close proximity of the gene. In term of classical molecular genetics, such DNA sequence polymorphisms are called *cis*-elements; hence a *cis*-acting eQTL is in harmony with the location of the gene under question. In the case of *trans*-acting eQTL, the location of the identified eQTL does not coincide with the location of the corresponding target gene. Genes underlying *trans*-eQTL are assumed to encode *trans*-acting factors – typically the proteins, by binding to *cis*-elements of other genes, which regulate their mRNA expression levels. Such *trans*-acting eQTL could, for example, represent the location of a transcription factor that regulates the expression of the binding target alone or, potentially, several functionally related genes. In reality, the expression level of a particular gene can be regulated by a combination of both *cis*- and *trans*-acting elements.

It must be punctuated that all eQTL studies rely on the genetic association between markers and gene expression trait in these particular samples under study. Different samples (or populations) may involve different eQTLs; hence the absence of eQTL for a

gene expression trait in a group of collected samples does not indicate that eQTL for that particular gene's expression trait does not exist in other samples (or populations).

1.9.1 A frame work of eQTL analysis

A high quality of eQTL analysis should precisely dissect the genetic architecture of the gene expression variation as a phenotypic trait in a relatively high-resolution. It can have three origins as follows: the gene expression profile accurately extracted for each sample, the high density genetic markers to comprehensively describe the genome-wide genetic polymorphisms for corresponding individuals and appropriate genetic association strategy to integrate these two sources of information into eQTL analysis.

First, gene expression levels are assayed as the steady-state abundance of mRNA transcripts that have been extracted in a specific sample (tissue or cell line) and at a specific time point. The transcript abundance can be measured using a variety of techniques from quantitative Reverse Transcription Polymerase Chain Reaction (RT-PCR) (Czechowski et al., 2004), through gene expression microarrays (Schena et al., 1995) to Next Generation Sequencing (NGS) technology (Wall et al., 2009). When expression levels of a particular gene are identified for all collected samples, the observed variation in expression level for this gene may be recognized as a heritable quantitative trait that can be applied to association mapping for dissecting the genetic architecture at the molecular level. Meanwhile, the accuracy of mapping eQTL also depends on the high density and precisely genotyped genetic polymorphisms. Until now, there are many different types of genetic markers have been developed, such as Simple sequence length polymorphism (SSLP), Restriction fragment length

polymorphism (RFLP), Amplified fragment length polymorphism (AFLP), Microsatellite polymorphism and Single nucleotide polymorphism (SNP). Among these genetic markers, Single Nucleotide Polymorphism (SNP) has obtained the most widespread interest. Because SNPs are highly abundant, and are generally estimated to occur at one out of every 1000 bases in the human genome (Sachidanandam, 2001, Venter, 2001). Currently, with the development of biological techniques, several high-throughput DNA genotyping platforms enable quick and efficient generation of high quality and density SNP markers over the genome under study.

The major challenge of eQTL analysis is how to model and identify the regulators of genome-wide gene expression using such high-dimensionality and complex dataset. Virtually, eQTL analysis shares the same statistical principles and approaches as conventional QTL analysis (Lan et al., 2003).

1.9.2 Extracting genome wide gene expression profile for eQTL mapping.

Rather than focus on a single gene, an eQTL analysis typically involves high-dimensional expression profile of genes in the whole genome (Jansen and Nap, 2001). Although many platforms are suitable for simultaneously measuring the genome wide (or whole genome) gene expression abundances, microarray technology is the most popular choice nowadays. Despite the Next Generation Sequencing technology reveals several improvements for constructing transcriptome, the expensive cost limits the application of NGS technology to small sample size. I tabulated in table I.6 the commercially available expression microarray platforms, where the Affymetrix GeneChip microarray and custom Agilent microarrays are commonly used (Close et al.,

2004). The basic detection units of Affymetrix expression microarrays are 25-base long oligodeoxy-nucleotide probes that are synthesized at specific locations on a coated quartz surface by photolithography (http://www.affymetrix.com/about_affymetrix/outreach/educator/microarray_curricula.affx#1_1). Each unique 25-mer probes' group is called a feature. Over a million features per microarray are usually available for synthesis, allowing multiple (typically from 11 to 22) probe-pairs per gene (the probe set). Generally, a typical Affymetrix GeneChip microarray enables to simultaneously measure ~20, 000 genes' expression levels.

In contrast, the basic detection units of Agilent expression microarrays have 60-base long oligodeoxyribo-nucleotides that are printed on glass slides using Agilent's proprietary SurePrint™ technology. Currently, the Agilent's commercially available chips generally contain either 1 x 244K, 2 x 105K, 4 x 44K or 8 x 15K probes. The Agilent gene expression platform is fully customizable; ready-to-go probes can be ordered to be synthesized on the slide, or alternatively custom sequences can be used to design probes by 'eArray', an online probe design tool (http://www.chem.agilent.com/enUS/products/instruments/dnamicroarrays/pages/gp50660.aspx). An extensive pool of pre-designed probes is also available from the eArray depository. Experimental design using Agilent microarrays can incorporate either a two-dye labelling protocol or single-dye labelling. Currently, the typical Agilent 44K microarray enables measure over 40,000 genes' expression levels at the same time.

Table I.6: Commercially available parallel, high-throughput gene expression analysis platforms.

Company	Detection method	No. of probes per gene	Probe length	No. of channels	No. of genes detected simultaneously
Affymetrix	Fluorescence; hybridization to array	Multiple (10-20 pairs)	25 mer	1	10,000~100,000
Agilent	Fluorescence; hybridization to array	Single	60 mer	1 or 2	15,000~244,000
Nimblegen	Fluorescence; hybridization to array	Multiple (up to 20)	45-60 mer	1 or 2	10,000~100,000
Illumina	Fluorescence; tags on beads	Single	50 mer	1	10,000+
ABI	Fluorescence; gel electrophoresis	Single	60 mer	1	10,000+
Sequenom	Mass spectrometry	Single	60-90 mer	2	Up to 1000

When choosing a microarray platform, one has to consider the genome presentation, analysis performance and experiment at costs. Usually, Affymetrix arrays are in principle more reliable than the Agilent platform because every gene on Affymetrix arrays represented by a multiple probes-set. However, relatively lower genome coverage and higher product cost can counterbalance its detection advantage. Agilent's flexible customization could also, of course, be used to design a cheap multi-probe array. Similar microarray platforms are either commercially available for most model species or can be easily produced for any species with available EST database using the flexibility of the Agilent (or other vendors) approach.

Recently, Next Generation Sequencing platforms, such as Roche/454, Illumina's Solexa, Life/APG SOLiD, Polonator and Helicos/BioSciences, offer an increasingly attractive alternative for digital gene expression measurement (Wall et al., 2009). Despite the relatively expensive cost at the moment, these platforms offer the opportunity to analyze any species – with or without a genome reference sequence – using a high-throughput strategy that enables measure the abundance of all transcripts in a given sample. Although there are only a few eQTL analysis published to date using this approach, it is highly likely that NGS application in RNA level (RNA-seq) will be widely utilized for this purpose in the near future.

If the researchers extract the gene expression profiles from several different platforms, an important question subsequently rises up that how consistently gene expression can be achieved from different technical platforms. It has been documented that commercial microarrays are more technically consistent than non-commercial microarrays (Chen et al., 2007, Coughlan et al., 2004) and that one-dye platforms are typically more

consistent than two-dye platforms (Kuo et al., 2006). I compared the expression of 15,208 barley 'unigenes' profiled using Agilent one-dye and two-dye microarrays with that of the same group of genes from Affymetrix microarrays (unpublished results). I observed a significant discrepancy between the expression levels evaluated from the two types of microarray. Correlation coefficients in estimated expression indices were as low as 65% among the different platforms and the one-dye system generated a higher level of concordance than the two-dye system. The poor performance of consistently gene expression measurement was not only observed across different microarray platforms, but also found in the same platform using different statistical methods to extract expression levels. In 2008, I explored seven main stream methods developed for extracting gene expression levels from Affymetrix microarray datasets in both yeast and barley and found that these methods can be divided into two clusters (Jiang et. al, 2008). The methods within each cluster show correlation coefficients of ≥95%, but the correlation is reduced to ~70% between the two clusters. Furthermore, the number of genes identified to be 'significantly differentially expressed' between the same group of genes varies substantially among the different extraction methods when subject to the same significant threshold or rate of false positives (Storey and Tibshirani, 2003).

1.9.3 eQTL Hotspots

Almost all eQTL analyses reveal that eQTL are not evenly distributed across the whole genome (Breitling et al., 2008, West et al., 2007, Gilad et al., 2008, Wei et al., 2007). When eQTLs are enriched in a specific genomic region more than expected by chance, this genomic region can be considered as 'an eQTL hotspot'. This type of eQTL hotspot is generally of functional interest. A biologically meaningful eQTL hotspot would

represent, for example, the location of a master transcriptional regulator that controls the expression of a group of genes that may link each other in the same biological process or pathway.

When eQTL studies are implemented in different samples using same treatments or tissues (cell lines), the expectation is that consistent eQTL hotspots should be observed. They would reveal the same biological pathways. eQTL studies applied in the same samples but using different treatments or tissues (cell lines) should however yield complementary results, reflecting the dynamic nature of the transcription process. For example, Potokina et al. (2008) observed several regions on chromosomes 2H, 5H and 7H which had many more eQTL than expected by chance alone based on a uniform distribution of genes per cM, using microarray profiles of genome-wide gene expression from germinating embryos of barley species. Interestingly, in the same population using *Puccinia hordei* infected seedling leaves 18 hours post infection, eQTL hotspots on two different chromosomes 1H and 3H, were observed. In the former, some eQTL hotspots did however correlate with the known location of 'malting quality' QTL (a trait expressed and measured in this tissue), while in pathogen challenged tissues at least 2 of the 3 hotspots were enriched for mRNAs related to general 'pathogen responsive genes'. While such observations provide a potential opportunity to unravel the genetic regulation mechanisms of important phenotypic traits, in general, these types of study are currently at a very early stage.

1.9.4 An important application of eQTL studies

The long-term objective of eQTL association mapping is to explore how genotypic variation underlies morphological or physiological consequences by using gene expression levels as intermediate information in molecular level. It means eQTL studies can offer useful information to identify the causal loci for conventional complex traits, particularly when the biologists try to understanding the molecular mechanism of the aetiology and provide new therapeutic targets for complex diseases. For most complex diseases, both morphological and physiological phenotypes which used to represent the characters of diseases in association studies are usually the outcomes of many different genes, which may interact with each other and with environmental factors. On the other hand, there might be many other unanticipated covert phenotypes that are also segregating. But, this wealth of information would be missed without additional specific phenotype assays.

A typical gene expression profile dataset contains the transcription abundance for whole genome or genome-wide genes. The enormous gene expression abundance information can be used as intermediate molecular phenotypes to explore the genetic basis of complex traits in association studies. First, one needs to identify several 'candidate genes' that correctly represent the complex traits under question. Then, these candidate genes would be quickly applied for eQTL studies and easily further analysed for understanding the genetic basis of relative complex traits (diseases).

A straightforward strategy to determining a 'candidate gene' for a given phenotypic trait of interest is to correlate this phenotypic trait values with expression levels of all detectable genes. The analysis constraint is that the same genetically fixed samples have been used to obtain both the trait and gene expression levels. A correlation study returns

a list of correlates (genes) and their respective correlation coefficients. Correlates with the highest absolute correlation coefficient can be potentially used as intermediate molecular traits to dissect the genetic architecture of phenotypic trait. Logically, most highly correlated eQTL should fall into the region containing the phenotypic QTL.

Correlation analysis is of course only the first step, and offers an overview of potential genes related to a trait. Further analysis involves the putative functions of the correlated genes and whether there are multiple coinciding eQTL, or hotspots, that may indicate that the causal gene is a trans-acting 'master regulator'. If such eQTL hotspots predominantly consists of *trans*-eQTL that have annotations from previous studies suggesting some form of functional relatedness, then a master regulatory locus may be inferred. Such a locus has been hypothesized on barley chromosome 2H where a putative regulatory 'master locus' affecting the expression of other genes associated with programmed cell death (PCD) has been proposed (Druka et al., 2008). When making such inferences, particularly in small sample size, it is important to exclude the possibility that chance co-segregation is responsible for the correlation.

1.10 References

ALBERTS, R., TERPSTRA, P., LI, Y., BREITLING, R., NAP, J.-P. & JANSEN, R. C. (2007) Sequence Polymorphisms Cause Many False <italic>cis</italic> eQTLs. *PLoS ONE,* 2, e622.

ALTMULLER, J., PALMER, L. J., FISCHER, G., SCHERB, H. & WJST, M. (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.,* 69, 936-950.

AUSTRALIA & NEW ZEALAND MULTIPLE SCLEROSIS GENETICS, C. (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. *Nature Genetics,* 41, 824 - 28.

BELL, S. D. & JACKSON, S. P. (1998) Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. *Trends in Microbiology,* 6, 222-228.

BLACK, D. L. (2003) MECHANISMS OF ALTERNATIVE PRE-MESSENGER RNA SPLICING. *Annual Review of Biochemistry,* 72, 291-336.

BOREVITZ, J. O., LIANG, D., PLOUFFE, D., CHANG, H.-S., ZHU, T., WEIGEL, D., BERRY, C. C., WINZELER, E. & CHORY, J. (2003) Large-Scale Identification of Single-Feature Polymorphisms in Complex Genomes. *Genome Research,* 13, 513-523.

BRAWAND, D., SOUMILLON, M., NECSULEA, A., JULIEN, P., CSARDI, G., HARRIGAN, P., WEIER, M., LIECHTI, A., AXIMU-PETRI, A., KIRCHER, M., ALBERT, F. W., ZELLER, U., KHAITOVICH, P., GRUTZNER, F., BERGMANN, S., NIELSEN, R., PAABO, S. & KAESSMANN, H. (2011) The evolution of gene expression levels in mammalian organs. *Nature,* 478, 343-348.

BREITLING, R., LI, Y., TESSON, B. M., FU, J., WU, C., WILTSHIRE, T., GERRITS, A., BYSTRYKH, L. V., DE HAAN, G., SU, A. I. & JANSEN, R. C. (2008) Genetical Genomics: Spotlight on QTL Hotspots. *PLoS Genet,* 4, e1000232.

CARLSON, C. S., EBERLE, M. A., KRUGLYAK, L. & NICKERSON, D. A. (2004) Mapping complex disease loci in whole-genome association studies. *Nature,* 429, 446-452.

CHEN, J., AGRAWAL, V., RATTRAY, M., WEST, M., ST CLAIR, D., MICHELMORE, R., COUGHLAN, S. & MEYERS, B. (2007) A comparison of microarray and MPSS technology platforms for expression analysis of Arabidopsis. *BMC Genomics,* 8, 414.

CLANCY, S. (2008a) DNA Transcription. *Nature Education,* 1.

CLANCY, S. (2008b) RNA functions. *Nature Education,* 1.

CLANCY, S. (2008c) RNA transcription by RNA polymerase: prokaryotes vs eukaryotes. *Nature Education,* 1.

CLANCY, S. & BROWN, W. (2008) Translation: DNA to mRNA to Protein. *Nature Education,* 1.

CLOSE, T. J., WANAMAKER, S. I., CALDO, R. A., TURNER, S. M., ASHLOCK, D. A., DICKERSON, J. A., WING, R. A., MUEHLBAUER, G. J., KLEINHOFS, A. & WISE, R. P. (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol. 2004 Mar;134(3):960-8.*

COLLINS, A. R. (2007) *Linkage Disequilibrium and Association Mapping : Analysis and Applications,* Totowa.

CONNELLY, S. & MANLEY, J. L. (1988) A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes & Development,* 2, 440-452.

COUGHLAN, S. J., AGRAWAL, V. & MEYERS, B. (2004) A comparison of global gene expression measurement technologies in Arabidopsis thaliana. *Comparative and Functional Genomics,* 5, 245-252.

CZECHOWSKI, T., BARI, R. P., STITT, M., SCHEIBLE, W.-R. & UDVARDI, M. K. (2004) Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *The Plant Journal,* 38, 366-379.

DRUKA, A., POTOKINA, E., LUO, Z., BONAR, N., DRUKA, I., ZHANG, L., MARSHALL, D., STEFFENSON, B., CLOSE, T., WISE, R., KLEINHOFS, A., WILLIAMS, R., KEARSEY, M. & WAUGH, R. (2008) Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat stem rust pathogen <i>Puccinia graminis</i> f. sp. <i>tritici</i> in barley. *TAG Theoretical and Applied Genetics,* 117, 261-272.

EASTON, D. F., BISHOP, D. T., FORD, D. & CROCKFORD, G. P. (1993) Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. *Am J Hum Genet. 1993 Apr;52(4):678-701.*

EASTON, D. F., POOLEY, K. A., DUNNING, A. M., PHAROAH, P. D. P., THOMPSON, D., BALLINGER, D. G., STRUEWING, J. P., MORRISON, J., FIELD, H., LUBEN, R., WAREHAM, N., AHMED, S., HEALEY, C. S., BOWMAN, R., MEYER, K. B., HAIMAN, C. A., KOLONEL, L. K., HENDERSON, B. E., LE MARCHAND, L., BRENNAN, P., SANGRAJRANG, S., GABORIEAU, V., ODEFREY, F., SHEN, C. Y., WU, P. E., WANG, H. C., ECCLES, D., EVANS, D. G., PETO, J. & FLETCHER, O. (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature,* 447, 1087 - 93.

EBRIGHT, R. H. (2000) RNA Polymerase: Structural Similarities Between Bacterial RNA Polymerase and Eukaryotic RNA Polymerase II. *Journal of Molecular Biology,* 304, 687-698.

FUNG, H. C., SCHOLZ, S., MATARIN, M., SIMON-SANCHEZ, J., HERNANDEZ, D., BRITTON, A., GIBBS, J. R., LANGEFELD, C., STIEGERT, M. L., SCHYMICK, J., OKUN, M. S., MANDEL, R. J., FERNANDEZ, H. H., FOOTE, K. D., RODRIGUEZ, R. L., PECKHAM, E., DE VRIEZE, F. W., GWINN-HARDY, K., HARDY, J. A. & SINGLETON, A. (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls first stage analysis and public release of data. *Lancet Neurobiology,* 5, 911 - 16.

GIALLOURAKIS, C. (2003) IBD5 is a general risk factor for inflammatory bowel disease: replication of association with Crohn disease and identification of a novel association with ulcerative colitis. *Am. J. Hum. Genet.,* 73, 205-211.

GILAD, Y., RIFKIN, S. A. & PRITCHARD, J. K. (2008) Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics,* 24, 408-415.

GNATT, A. L., CRAMER, P., FU, J., BUSHNELL, D. A. & KORNBERG, R. D. (2001) Structural Basis of Transcription: An RNA Polymerase II Elongation Complex at 3.3 Resolution. *Science,* 292, 1876-1882.

GRANT, S. F. A. & HAKONARSON, H. (2008) Microarray Technology and Applications in the Arena of Genome-Wide Association. *Clinical Chemistry,* 54, 1116-1124.

GRUMMT, I. & KIVIE, M. (1998) Regulation of Mammalian Ribosomal Gene Transcription by RNA Polymerase I. *Progress in Nucleic Acid Research and Molecular Biology.* Academic Press.

GUHANIYOGI, J. & BREWER, G. (2001) Regulation of mRNA stability in mammalian cells. *Gene,* 265, 11-23.

HAUSER, E. R. & BOEHNKE, M. (1998) Genetic Linkage Analysis of Complex Genetic Traits by Using Affected Sibling Pairs. *Biometrics,* 54, 1238-1246.

HEGDE, R. S. & KANG, S.-W. (2008) The concept of translocational regulation. *The Journal of Cell Biology,* 182, 225-232.

HIRSCHHORN, J. N. & ALTSHULER, D. (2002) Once and again [mdash] issues surrounding replication in genetic association studies. *J. Clin. Endocrinol. Metab.,* 87, 4438-4441.

HIRSCHHORN, J. N. & DALY, M. J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat Rev Genet,* 6, 95-108.

HURWITZ, J. (2005) The Discovery of RNA Polymerase. *Journal of Biological Chemistry,* 280, 42477-42485.

JANSEN, R. C. & NAP, J.-P. (2001) Genetical genomics: the added value from segregation. *Trends in Genetics,* 17, 388-391.

JIMENEZ-SANCHEZ, G., CHILDS, B. & VALLE, D. (2001) Human disease genes. *Nature,* 409, 853-855.

KHAITOVICH, P., ENARD, W., LACHMANN, M. & PAABO, S. (2006) Evolution of primate gene expression. *Nat Rev Genet,* 7, 693-702.

KIM, T.-K., LAGRANGE, T., WANG, Y.-H., GRIFFITH, J. D., REINBERG, D. & EBRIGHT, R. H. (1997) Trajectory of DNA in the RNA polymerase II transcription preinitiation complex. *Proceedings of the National Academy of Sciences,* 94, 12268-12273.

KORNBERG, R. D. (2007) The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences,* 104, 12955-12961.

KUO, W. P., LIU, F., TRIMARCHI, J., PUNZO, C., LOMBARDI, M., SARANG, J., WHIPPLE, M. E., MAYSURIA, M., SERIKAWA, K., LEE, S. Y., MCCRANN, D.,

KANG, J., SHEARSTONE, J. R., BURKE, J., PARK, D. J., WANG, X., RECTOR, T. L., RICCIARDI-CASTAGNOLI, P., PERRIN, S., CHOI, S., BUMGARNER, R., KIM, J. H., SHORT, G. F., FREEMAN, M. W., SEED, B., JENSEN, R., CHURCH, G. M., HOVIG, E., CEPKO, C. L., PARK, P., OHNO-MACHADO, L. & JENSSEN, T.-K. (2006) A sequence-oriented comparison of gene expression measurements across different hybridization-based technologies. *Nat Biotech,* 24, 832-840.

KUZNETSOV, V. A., KNOTT, G. D. & BONNER, R. F. (2002) General Statistics of Stochastic Process of Gene Expression in Eukaryotic Cells. *Genetics,* 161, 1321-1332.

LAN, H., STOEHR, J. P., NADLER, S. T., SCHUELER, K. L., YANDELL, B. S. & ATTIE, A. D. (2003) Dimension Reduction for Mapping mRNA Abundance as Quantitative Traits. *Genetics,* 164, 1607-1614.

LAURSEN, B. S., S RENSEN, H. P., MORTENSEN, K. K. & SPERLING-PETERSEN, H. U. (2005) Initiation of Protein Synthesis in Bacteria. *Microbiology and Molecular Biology Reviews,* 69, 101-123.

LEE, T. I. & YOUNG, R. A. (2000) TRANSCRIPTION OF EUKARYOTIC PROTEIN-CODING GENES. *Annual Review of Genetics,* 34, 77-137.

LEE, Y., KIM, M., HAN, J., YEOM, K.-H., LEE, S., BAEK, S. H. & KIM, V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J,* 23, 4051-4060.

LEWONTIN, R. C. & KOJIMA, K. I. (1960) The evolutionary dynamics of complex polymorphisms. *Evolution,* 4, 458 - 472.

LOHMUELLER, K. E., PEARCE, C. L., PIKE, M., LANDER, E. S. & HIRSCHHORN, J. N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.,* 33, 177-182.

LUO, Z. W., POTOKINA, E., DRUKA, A., WISE, R., WAUGH, R. & KEARSEY, M. J. (2007) SFP Genotyping From Affymetrix Arrays Is Robust But Largely Detects Cis-acting Expression Regulators. *Genetics,* 176, 789-800.

MASTON, G. A., EVANS, S. K. & GREEN, M. R. (2006) Transcriptional Regulatory Elements in the Human Genome. *Annual Review of Genomics and Human Genetics,* 7, 29-59.

MATLIN, A. J., CLARK, F. & SMITH, C. W. J. (2005) Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol,* 6, 386-398.

MATTICK, J. S., AMARAL, P. P., DINGER, M. E., MERCER, T. R. & MEHLER, M. F. (2009) RNA regulation of epigenetic processes. *BioEssays,* 31, 51-59.

MCCARTHY, J. E. G. (1998) Posttranscriptional Control of Gene Expression in Yeast. *Microbiology and Molecular Biology Reviews,* 62, 1492-1553.

NILSEN, T. W. & GRAVELEY, B. R. (2010) Expansion of the eukaryotic proteome by alternative splicing. *Nature,* 463, 457 - 463.

NISTICO, L. (1996) The CTLA-4 gene region of chromosome 2q33 is linked to, and associated with, type 1 diabetes. *Hum. Mol. Genet.,* 5, 1075-1080.

NUDLER, E. (1999) Transcription elongation: structural basis and mechanisms. *Journal of Molecular Biology,* 288, 1-12.

OUHAMMOUCH, M., DEWHURST, R. E., HAUSNER, W., THOMM, M. & GEIDUSCHEK, E. P. (2003) Activation of archaeal transcription by recruitment of the TATA-binding protein. *Proceedings of the National Academy of Sciences,* 100, 5097-5102.

PAL, M., MCKEAN, D. & LUSE, D. S. (2001) Promoter Clearance by RNA Polymerase II Is an Extended, Multistep Process Strongly Affected by Sequence. *Molecular and Cellular Biology,* 21, 5815-5825.

PANSE, V. G. & JOHNSON, A. W. (2010) Maturation of eukaryotic ribosomes: acquisition of functionality. *Trends in Biochemical Sciences,* 35, 260-266.

PROUDFOOT, N. J., FURGER, A. & DYE, M. J. (2002) Integrating mRNA Processing with Transcription. *Cell,* 108, 501-512.

RALSTON, A. & KENNA, S. M. (2008) mRNA: History of Functional Investigation. *Nature Education,* 1.

REINES, D., CONAWAY, R. C. & CONAWAY, J. W. (1999) Mechanism and regulation of transcriptional elongation by RNA polymerase II. *Current Opinion in Cell Biology,* 11, 342-346.

REMINGTON, D. L., THORNSBERRY, J. M., MATSUOKA, Y., WILSON, L. M., WHITT, S. R., DOEBLAY, J., KRESOVICH, S., GOODMAN, M. M. & BUCKLER, E. S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America,* 98, 11479 - 11484.

RICHARDSON, J. P. (2002) Rho-dependent termination and ATPases in transcript termination. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression,* 1577, 251-260.

RISCH, N. & MERIKANGAS, K. (1996) The future of genetic studies of complex human diseases. *Science,* 273, 1516-1517.

RISCH, N. J. (2000) Searching for genetic determinants in the new millennium. *Nature,* 405, 847-856.

ROBBINS, R. B. (1918) APPLICATIONS OF MATHEMATICS TO BREEDING PROBLEMS II. *Genetics,* 3, 73-92.

ROBERTSON, M. (2010) The evolution of gene regulation, the RNA universe, and the vexed questions of artefact and noise. *BMC Biology,* 8, 97.

SACHIDANANDAM, R. (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature,* 409, 928-933.

SCHALLER, H., GRAY, C. & HERRMANN, K. (1975) Nucleotide sequence of an RNA polymerase binding site from the DNA of bacteriophage fd. *Proceedings of the National Academy of Sciences,* 72, 737-741.

SCHENA, M., SHALON, D., DAVIS, R. W. & BROWN, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science. 1995 Oct 20;270(5235):467-70.*

SCHWANHAUSSER, B., BUSSE, D., LI, N., DITTMAR, G., SCHUCHHARDT, J., WOLF, J., CHEN, W. & SELBACH, M. (2011) Global quantification of mammalian gene expression control. *Nature,* 473, 337-342.

SIEBERTS, S. K. & SCHADT, E. E. (2007) Moving toward a system genetics view of disease. *Mamm Genome. 2007 Jul;18(6-7):389-401. Epub 2007 Jul 26.*

SMALE, S. T. & KADONAGA, J. T. (2003) THE RNA POLYMERASE II CORE PROMOTER. *Annual Review of Biochemistry,* 72, 449-479.

SMITH, A. & SHAW, K. (2008) Discovering the Relationship Between DNA and Protein Production. *Nature Education,* 1.

STEITZ, T. A. (2008) A structural understanding of the dynamic ribosome machine. *Nat Rev Mol Cell Biol,* 9, 242-253.

STOREY, J. D. & TIBSHIRANI, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences,* 100, 9440-9445.

TOMILIN, N. V. (2008) Regulation of mammalian gene expression by retroelements and non-coding tandem repeats. *BioEssays,* 30, 338-348.

TUCH, B. B., LI, H. & JOHNSON, A. D. (2008) Evolution of eukaryotic transcription circuits. *Science,* 319, 1797 - 1799.

VALDAR, W., SOLBERG, L. C., GAUGUIER, D., BURNETT, S., KLENERMAN, P., COOKSON, W., TAYLOR, M. S., RAWLINS, J. N. P., MOTT, R. & FLINT, J. (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nature Genetics,* 38, 879 - 887.

VENTER, J. C. (2001) The sequence of the human genome. *Science,* 291, 1304-1351.

WALL, P. K., LEEBENS-MACK, J., CHANDERBALI, A., BARAKAT, A., WOLCOTT, E., LIANG, H., LANDHERR, L., TOMSHO, L., HU, Y., CARLSON, J., MA, H., SCHUSTER, S., SOLTIS, D., SOLTIS, P., ALTMAN, N. & DEPAMPHILIS, C. (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics,* 10, 347.

WANG, M., HU, X., LI, G., LEACH, L. J., POTOKINA, E., DRUKA, A., WAUGH, R., KEARSEY, M. J. & LUO, Z. (2009) Robust Detection and Genotyping of Single Feature Polymorphisms from Gene Expression Data. *PLoS Comput Biol,* 5, e1000317.

WATSON, J. D. (1963) Involvement of RNA in the synthesis of proteins. *Science,* 140, 17-26.

WEI, S., TIANWEI, Y. & KER-CHAU, L. (2007) Detection of eQTL modules mediated by activity levels of transcription factors. *Bioinformatics,* 23, 2290-2297.

WEIR, B. S. (2008) Linkage Disequilibrium and Association Mapping. *Annual Review of Genomics and Human Genetics,* 9, 129-142.

WEST, M. A. L., KIM, K., KLIEBENSTEIN, D. J., VAN LEEUWEN, H., MICHELMORE, R. W., DOERGE, R. W. & ST. CLAIR, D. A. (2007) Global eQTL Mapping Reveals the Complex Genetic Architecture of Transcript-Level Variation in Arabidopsis. *Genetics,* 175, 1441-1450.

WILLIS, I. M. (1993) RNA polymerase III. *European Journal of Biochemistry,* 212, 1-11.

XU, Y., IKEGAMI, M., WANG, Y., MATSUZAKI, Y. & WHITSETT, J. (2007) Gene expression and biological processes influenced by deletion of Stat3 in pulmonary type II epithelial cells. *BMC Genomics,* 8, 455.

YAN, J., MAGNASCO, M. O. & MARKO, J. F. (1999) A kinetic proofreading mechanism for disentanglement of DNA by topoisomerases. *Nature,* 401, 932-935.

YUSUPOVA, G. Z., YUSUPOV, M. M., CATE, J. H. D. & NOLLER, H. F. (2001) The Path of Messenger RNA through the Ribosome. *Cell,* 106, 233-241.

Chapter II

Developing a robust statistical method for association-based expression quantitative trait analysis

2.1 Related publication

Jiang N, Wang M, Jia T, Wang L, Leach L, Hackett C, Marshall D, Luo Z. (2011) A robust statistical method for association-based eQTL analysis. PLoS One, 6(8): e23192.

Please see Appendix III for a copy of this article as it appeared in print.

2.2 Overview

It has been well established that theoretical kernel for recently surging genome-wide association study (GWAS) is statistical inference of linkage disequilibrium (LD) between a tested genetic polymorphism and a putative locus affecting a complex trait. However, a fundamental problem in such studies is the existence of population structure because it can lead to artificial associations, even within relatively homogeneous populations. Whilst many methods have been proposed to correct for the influence either through predicting the structure parameters or correcting inflation in the test

statistic due to the stratification, these may not be feasible or may impose further statistical problems in practical implementation. I propose here a novel statistical method to control spurious LD in GWAS from population structure by incorporating a control marker into testing for significance of genetic association of a polymorphic marker with phenotypic variation of a complex trait. The method avoids the need of population structure prediction which may be infeasible or inadequate in practice and accounts properly for varying effects of population stratification on different regions of the genome under study. Utility and statistical properties of the new method were tested through an intensive computer simulation study and an association-based genome-wide mapping of expression quantitative trait loci in genetically divergent human populations. The analyses show that the new method confers an improved statistical power for detecting genuine genetic association in subpopulations and an effective control of spurious associations stemmed from population structure when compared with other two popularly implemented methods in the literature of GWAS.

2.3 Introduction

Linkage disequilibrium (LD) based association mapping has received increasing attention in the recent literature (Ardlie et al., 2002, Couzin and Kaiser, 2007, Iles, 2008, McCarthy et al., 2008, Slatkin, 2008, Weiss and Clark, 2002) for its potential power and precision in detecting subtle phenotypic associated genetic variants when compared with traditional family-based linkage studies. Association mapping methods for the genetic dissection of complex traits utilize the decay of LD, the rate of which is determined by genetic distance between loci and the generation time since LD arose (Mackay and Powell, 2007). Over multiple generations of segregation, only loci physically close to the quantitative trait loci (QTL) are likely to be significantly associated with the trait of interest in a randomly mating population, providing great efficiency at distinguishing between small recombination fractions (Remington et al., 2001). Despite this potential power, many reported association studies have not been replicated or have resulted in false positives (Cardon and Bell, 2001, Risch, 2000), commonly caused by 'cryptic' structure in population-based samples. Population structure, or population stratification (Balding, 2006), arises from systematic variation in allele frequencies across subpopulations, which can result in statistical association between a disease phenotype and marker(s) that have no physical linkage to causative loci (Ewens and Spielman, 1995, Lander and Schork, 1994), *i.e.* false positive or spurious associations. This gives rise to an urgent need for methods of adjusting for both population structure and cryptic relatedness occurring due to distant relatedness among samples with no known family relationships.

2.3.1 Family-based association studies

To avoid the problems raised from population stratification, family-based association studies have been proposed, such as the transmission-disequilibrium test (TDT), which is an exact application of McNemar's Test in genetics (Spielman et al., 1993). In this test, the samples are collected from nuclear families which consist of two parents and one affected offspring (family trios). The TDT test is designed to compare the frequencies of marker alleles transmitted from heterozygous parents to affected offspring against those that are not transmitted. For example, there are two alleles M_1 and M_2 at a genetic locus. In total *n* trios' families, there are 2*n* parents and *n* affected offspring. The transmissions and non-transmissions of alleles M_1 and M_2 from parents to offspring can be summarised in a 2×2 table (table II.1):

Table II.1: Transmitted and nontransmitted marker alleles M_1 and M_2 from 2*n* parents to *n* affected offspring

Transmitted allele	Non-transmitted allele		Total
	M_1	M_2	
M_1	**a** $P = M_1M_1 \rightarrow M_1$	**b** $P = M_1M_2 \rightarrow M_1$	**a+b**
M_2	**c** $P = M_1M_2 \rightarrow M_2$	**d** $P = M_2M_2 \rightarrow M_2$	**c+d**
Total	**a+c**	**b+d**	**2*n***

The TDT only uses information from heterozygous parents (*b* and *c* in table II.1). The null hypothesis of TDT test states that the observed ratios of $b/(b+c)$ and $c/(b+c)$ have equal proportions (0.5, 0.5). It means this locus does not genetically

associate with the phenotypic trait. This hypothesis can be inspected via an asymptotically *Chi*-square test with 1 degree of freedom:

$$\chi^2 = \frac{[b-(b+c)/2]^2}{(b+c)/2} + \frac{[c-(b+c)/2]^2}{(b+c)/2} = \frac{(b-c)^2}{b+c} \quad \text{(II-3.1)}$$

In TDT design, the ethnic background of cases and controls needs to be necessarily matched, conferring robustness to the presence of population structure. However, TDT design can detect the presence of linkage between trait and genetic marker only if genetic association (results from linkage-disequilibrium) is present. And, it also requires samples from family trios, which are difficult to obtain compared to population based designs where a large sample is feasibly obtained. Moreover, increased genotyping efforts are required for TDT design to achieve the same power as population based design (Cardon and Palmer, 2003, McGinnis et al., 2002).

2.3.2 Genomic control (GC) method and the structure association (SA) analysis

Numerous methods have been proposed to overcome the problems caused by population structure without the need for family based samples. Among the most widely used approaches are the genomic control (GC) method (Devlin and Roeder, 1999, Bacanu et al., 2002) and the structure association (SA) analysis (Pritchard et al., 2000a, Pritchard et al., 2000b). In the former, the genomic control method was initially designed for case-control studies but has been extended to quantitative traits (Bacanu et al., 2002).

For example, in a case-control study, the genotype distribution of two alleles M_1 and M_2 at a bi-allelic locus can be summarised in a 2×3 table (table II.2):

Table II.2: genotype distribution of allele M_1 in case-control samples

	M_1 alleles			Total
	0	1	2	
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

Devlin and Roeder (1999) employed the Armitage's trend test to examine the association between genetic marker and phenotypic trait:

$$Y^2 = \frac{N(N(r_1 + 2r_2) - R(n_1 + 2n_2))^2}{R(N-R)(N(n_1 + 4n_2) - (n_1 + 2n_2)^2)} \qquad \text{(II-3.2)}$$

where the trend test statistic Y^2 is asymptotic -square distribution with 1 degree of freedom. If there is no population structure in the collected samples, the test statistic can be directly used to evaluate the significant level of genetic association. However, when the population structure is present in samples, the trend test Y^2 will be inflated by a factor λ where the inflation λ depends on the confounding effect of population structure. In genomic control method, Devlin and Roeder assumed that the confounding effects of population stratification can cause a constant inflation factor λ of the test statistic across the whole genome. It uses statistical inference approaches (both of frequentist statistics and Bayesian inference) to estimate this constant inflation factor λ from a group of unlinked genetic markers. And then, the test statistic will be adjusted from the estimate before being applied to infer the significance of association. The genomic control is an efficient method because it does not require knowledge of the

population structure information. However, the genomic control method considers an ideal but unrealistic situation of constant inflation factor λ for all markers, while in reality the influence of population structure on statistical inference of marker-trait association varies over genome locations (Astle and Balding, 2010). It is due to the fact that the divergences of allele frequencies across ancestral populations vary around the whole genome.

The main alternative approach to adjust population structure effect in association studies is the structure association (SA) analysis (Pritchard et al., 2000a, Pritchard et al., 2000b). Briefly, the structure association (SA) method is a Bayesian-based clustering strategy to infer the population structure information using genotypic data. Pritchard assume that there are K subpopulations from which the samples are collected, and each of subpopulation is distinguished by frequencies of a set of characterised alleles. In this method, each sample is assigned into subpopulation based on the genotypes of a group of selected markers. The SA method is widely suitable for many different types of genetic markers, such as Restriction fragment length polymorphisms (RFLPs), Single feature polymorphisms (SFPs) and Single nucleotide polymorphisms (SNPs). However, it requires that these selected markers should be unlinked with each other and also segregated under Hardy-Weinberg equilibrium within subpopulations. In SA method, it does not give a clear guide to select such genetic markers. In practice, it is a challenge to obtain a group of markers matched those requirements without the knowledge of population structure information. Moreover, for the SA method, it is computationally intensive to obtain accurate and reliable values for both the number of subpopulations in real datasets and to assign any individual into a population membership.

2.3.3 Principal component analysis (PCA) based method— EIGENSTRAT

Recently, several methods have been adopted to infer the subpopulation number, including Latent-Class model (Satten et al., 2001), mixture model (Zhu et al., 2002), a Bayesian model — AdmixMap (Hoggart et al., 2003) and a Principal Component Analysis (PCA) based strategy — EIGENSTRAT (Price et al., 2006). These methods share the common assumption that associations among unlinked markers are the result of population structure and subpopulations are allocated to minimize these associations. This step depends critically upon the correct selection of a panel of markers to reflect population structure information, because using different groups of genetic markers can lead to distinct inference results. For example, Price *et al.* (2006) proposed a principal component analysis (PCA) based method, EIGENSTRAT, to explicitly model the ancestral divergences in allele frequency and correct for population stratification by adjusting genotypes through linear regression. Briefly, this simple and efficient approach consists of three steps. At first, principal component analysis (PCA) is applied to genotype information to obtain the continuous axes of genetic variance. In this step, principal component analysis (PCA) successfully reduces the high-throughput genetic information into a few dimensions, keeping as much variance as possible. And then, it adjusts both phenotype and genotype data by the amounts of attribution to ancestry along the axes of variance. Finally, it tests the association statistics based on the ancestry adjusted genotype data and phenotype trait. While EIGENSTRAT provides specific correction for candidate markers, how to choose appropriate markers to infer population structure remains in question. In fact, prediction of the population structure may fail whenever the key assumption behind the structure prediction methods is violated.

2.3.4 Correcting the confound effect of population structure using only one genetic marker

Rather than using a panel of unlinked markers to exploit the cryptic population structure, a single null marker can be used to correct for bias of the test statistic in association studies. Wang *et al.* (2005) suggested using only one well-selected null marker to correct biases from population stratification on association test for a candidate gene. At first, Wang *et al.* (2005) has assessed how the confounding effect of population structure in association studies could be corrected by a single genetic marker. When the selected null marker has greatly varied distribution patterns across subpopulations, spurious association can be partially corrected in a logistic regression model. Furthermore, when the null marker has the same genotype distribution as the candidate gene across subpopulations, the confounding effect of population structure can be completely removed from the logistic regression model. Compared with above models and methods, this method better defines a clear guide to select a type of null maker for correcting the bias (or spurious) due to population structure. However, for this method, they assumed a simplistic situation that the null marker had the same genotypic distribution as the candidate gene, which was unknown in practice. Furthermore, they did not statistically model the power assessment of bias reduction when using null marker distributed un-identically to the candidate gene.

Proposed that the confounding effect of population structure in association studies may be corrected by only one well-selected genetic marker, I develop a novel linear regression model for association-based eQTL analysis. The variation of transcript abundance is widely reported in all organisms' studies to date (Gilad, 2008). It has

already proved that gene expressed variation may be responsible for many kinds of the phenotypic traits in natural populations e.g. branching structure in maize (Clark, 2006), bristle number in fruit flies (McGregor, 2007), beak morphology in Darwin finches (Abzhanov, 2004). Moreover, gene expression variation has been genetically associated with more than hundreds of human complex traits including diverse aspects of behaviour, physiology and disease (Kleinjan, 2005, Wray, 2007). Despite accumulating evidence that transcription variation contributes to many important complex traits, it still know little about the genetic architectures for variation in gene expression levels. Gene expression quantitative trait loci (eQTL) analysis is the main approach to identify which genomic regions control transcription and to explore the effect of variation in these DNA sequence regions. In such studies, the gene transcription abundances are considered as the quantitative traits, and the genetic basis to regulate the variation in transcription level can be identified using genome-wide association study. The expression quantitative trait locus (eQTL) analyses have recently proved that variation in human gene expression levels among individuals and also populations is influenced by polymorphic genetic variants (Campino et al., 2008, Cheung et al., 2003, Spielman et al., 2007). The use of structured populations has meant that to detect the genetic variants accounting for differences in gene expression between subpopulations, Genome-wide association studies (GWAS) had to be carried out separately for each subpopulation and the results subsequently compared. I present here a simple linear regression model of utilizing only one 'control' marker to remove the population structure effect in detecting linkage disequilibrium (LD) between a marker and a putative quantitative trait locus (QTL). I first established the theoretical basis for selection and use of a control marker to correct for population structure and established a regression-based method for

detecting the LD which is integrated with information of the control marker. I investigated the method for its efficiency to test the LD and to reduce false positives stemmed from population structure through intensive computer simulation studies and re-analysis of the gene expression and SNP datasets collected from genetically divergent populations. The new method (**Method 1**) was compared with two alternative methods: single marker regression without population structure correction (**Method 2**) and multiple regression analysis with incorporation of known individual ancestry information (**Method 3**).

2.4 Statistical models and methods

2.4.1 Method 1: a novel regression analysis with correcting population structure

Here, the model is designed to analyze a structured randomly mating population produced through instant admixture of two genetically divergent subpopulations. The proportion of subpopulation 1 in the mixed population is denoted by *m*; meanwhile the proportion of subpopulation 2 can be represented by 1-*m*. Let us consider three bi-allelic loci: one affects a quantitative trait (*Q*) while another two are polymorphic co-dominant markers devoid of direct effect on the phenotypic trait. They are called, for convenience, one of the markers the test marker (*T*) which is used to be tested for genetic association with the QTL, and the other considered as control marker (*C*), assumed to be not associated with both the QTL and the test marker (*i.e.* the linkage disequilibrium *D* equal 0). Two alleles are denoted by *A* and *a* at the putative QTL, *T* and *t* at the test marker, and *C* and *c* at the control marker. Three genotypes at the QTL, *AA*, *Aa* and *aa*, are assumed to affect the quantitative trait by *d*, *h* and *–d* respectively. Trait phenotype of an individual (Y) is assumed to be normally distributed with mean depending on its genotype effect at the QTL and residual variance σ_e^2. Genotypic values at the test marker and control marker are denoted by *X* and *Z,* which are the number of alleles *T* and *C* respectively. In subpopulation *i* (*i*=1 or 2), the allelic frequencies of the QTL, test marker and control marker are denoted by $P_Q^{(i)}$, $P_T^{(i)}$ and $P_C^{(i)}$ respectively, while the coefficients of linkage disequilibrium between any pair of the loci are denoted by $D_{TC}^{(i)}$, $D_{TQ}^{(i)}$ and $D_{CQ}^{(i)}$.

Table II.3 explicitly illustrates probability distribution of joint genotypes at a test marker and a putative QTL in randomly mating populations together with genotypic values at the QTL and the original work for drawing this table was implemented by Luo in 1998. In table II.3, the conditional probabilities Q (or R) are represented the frequency of marker allele T simultaneously carrying allele A (or a) at the QTL, which can be formulated by allele frequencies at marker loci (p) and putative QLT (q) and the coefficient of linkage disequilibrium between the marker and QTL (D). It is clear from Table II.3 that the joint distribution of marker-QTL can be fully characterized by the parameters defining population allele frequencies at the two loci and the coefficient of linkage disequilibrium between them. This provides the theoretical basis for statistical analyses developed below.

Table II.3: Probability distribution of joint genotypes at a test marker and a putative QTL and genotypic values at the QTL

Genotypes at QTL		*AA*			*Aa*			*aa*	
Marker genotypes	*TT*	*Tt*	*tt*	*TT*	*Tt*	*tt*	*TT*	*Tt*	*tt*
Probabilities	$(qQ)^2$	$2q^2Q(1-Q)$	$q^2(1-Q)^2$	$2\,q(1-q)QR$	$2\,q(1-q)(Q+R-2QR)$	$2\,q(1-q)\,(1-Q)(1-R)$	$(1-q)^2R^2$	$2(1-q)^2R(1-R)$	$(1-q)^2(1-R)^2$
Genotypic values at QTL		$\mu + d$			$\mu + h$			$\mu - d$	

where *A* and *a* are segregating alleles at a putative QTL, *T* and *t* are alleles at the test marker locus. Allele frequency of *A* is *q*, allele frequency of *T* is *p*. *Q* and *R* are conditional probabilities of marker allele *T* given QTL allele *A* and *a* respectively, which are formulated as $Q = p + D/q$ and $R = p - D/(1-q)$ where *D* is the coefficient of linkage disequilibrium between the marker and QTL. μ, *d* and *h* are population mean, additive and dominance genic effects at the QTL.

2.4.1.1 The novel linear regression model and theoretical analysis

In 1990, Land and Thompson have introduced the usage of regression of quantitative trait on the number of common alleles of genetic marker as a marker score in marker-assisted selection (MAS) analysis of a quantitative trait. Here, for phenotypic value of a quantitative trait and each of the genetically polymorphic markers, I fit the following form: the genotype X_{ij} of individual i at the given marker locus j may be classified as one of three states: $X_{ij} = 0, 1, or 2$ for homozygous rare, heterozygous and homozygous common alleles, respectively; while Y_i is the phenotypic value for individual $i = 1, \cdots, \cdots, n$. In the present model, I fit a linear regression of the form for each genetic marker:

$$Y_i = b_0 + b_1 X_{ij} + \varepsilon_i \qquad \text{(II-4.1)}$$

where b_0 is the mean of phenotypic value of quantitative trait in total samples, b_1 is the regression coefficient between X and Y, and ε_i is an independent normally distributed random variable with mean 0 and variance σ_e^2. In statistics, the simple linear regression model is to estimate the relationship between Y and one explanatory variable X. In the least-square approach, the best fitted model should minimize the sum of squared residuals ε_i^2 in the regression analysis:

$$\min_{b_0, b_1} Q(b_0, b_1), \ where \ Q(b_0, b_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2 \qquad \text{(II-4.2)}$$

According to the least-square approach, the best fitted regression coefficient b_1 is given by:

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$
$$= \frac{\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i^2 - \frac{1}{n}(\sum_{i=1}^{n} X_i)^2} \quad . \qquad \text{(II-4.3)}$$
$$= \frac{E(XY) - E(X)E(Y)}{E(X^2) - E^2(X)}$$

In association studies, it has been clearly demonstrated that significance of the regression coefficient can be used to infer significance of LD between a polymorphic marker locus and a QTL in a single randomly mating population since the regression coefficient has a form of

$$b_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} = \frac{E(XY) - E(X)E(Y)}{E(X^2) - E^2(X)} = \frac{2D_{TQ}[d + (1-2p_Q)h]}{2p_T(1-p_T)} \qquad \text{(II-4.4)}$$

(Luo, 1998). In equation II-4.4, it is clear that the magnitude of linear regression coefficient b_1 directly depends on the level of association between test marker and putative QTL (D_{TQ}). However, in a structured population, it clearly shows that the LD between a marker and a QTL is given by

$$D_{TQ} = mD_{TQ}^{(1)} + (1-m)D_{TQ}^{(2)} + m(1-m)\delta_T\delta_Q \,, \qquad \text{(II-4.5)}$$

(Chakraborty and Smouse, 1988), where *m* is the proportion of subpopulation 1 in this mixed samples, the superscripts (1) and (2) refer to the subpopulations, $\delta_T = p_T^{(1)} - p_T^{(2)}$

and $\delta_Q = p_Q^{(1)} - p_Q^{(2)}$. The covariance between the QTL and the test marker can be worked out as

$$\begin{aligned}\sigma_{X,Y} &= 2mD_{TQ}^{(1)}(d + h - 2hp_Q^{(1)}) \\ &+2(1-m)D_{TQ}^{(2)}(d + h - 2hp_Q^{(2)}) \\ &+4m(1-m)\delta_T\delta_Q[d + h(1 - p_Q^{(1)} - p_Q^{(2)})]\end{aligned} \quad . \qquad \text{(II-4.6)}$$

Equations II-4.5and II-4.6 show that the association between the QTL and test marker in a mixed population is the summation of (i) a linear combination of the associations between the two loci in each of the subpopulations (i.e. the genuine association due to LD between the two loci in each of the subpopulations), and (ii) a nonlinear component of the differences in allele frequencies between the two subpopulations (i.e. a spurious term of association). The objective of this analysis is to remove the spurious term by using a control marker 'C'.

Consistent with above analysis, the covariance between control marker and QTL (or test marker) in the admixture population are given by

$$\begin{aligned}\sigma_{Y,Z} &= 2mD_{CQ}^{(1)}(d + h - 2hp_Q^{(1)}) \\ &+2(1-m)D_{CQ}^{(2)}(d + h - 2hp_Q^{(2)}) \\ &+4m(1-m)\delta_C\delta_Q[d + h(1 - p_Q^{(1)} - p_Q^{(2)})]\end{aligned} \qquad \text{(II-4.7)}$$

$$\sigma_{X,Z} = 2mD_{TC}^{(1)} + 2(1-m)D_{TC}^{(2)} + 4m(1-m)\delta_T\delta_C \qquad \text{(II-4.8)}$$

where $\delta_C = p_C^{(1)} - p_C^{(2)}$. If the control marker is neither in association with the QTL (i.e. $D_{CQ}^{(1)} = D_{CQ}^{(2)} = 0$) nor with the test marker ($D_{TC}^{(1)} = D_{TC}^{(2)} = 0$), then the covariance between control marker and QTL (or test marker) can be simplified as

$$\sigma_{Y,Z} = 4m(1-m)\delta_C\delta_Q[d + h(1 - p_Q^{(1)} - p_Q^{(2)})] \qquad \text{(II-4.9)}$$

$$\sigma_{X,Z} = 4m(1-m)\delta_T\delta_C \quad \text{(II-4.10)}$$

In an admixed population, the control marker's allelic frequency is $p_C = m p_C^{(1)} + (1-m) p_C^{(2)}$. In a population with allelic frequency p_C at the control marker locus, the expected and observed variances at the control marker are

$$E[\sigma_Z^2] = 2[m p_C^{(1)} + (1-m) p_C^{(2)}][1 - m p_C^{(1)} - (1-m) p_C^{(2)}] = 2 p_C (1 - p_C) \quad \text{(II-4.11)}$$

$$\sigma_Z^2 = 2[mp_C^{(1)} + (1-m) p_C^{(2)}][1 - mp_C^{(1)} - (1-m) p_C^{(2)}] + 2m(1-m)\delta_C^2 \quad \text{(II-4.12)}$$

Thus, the difference between the expected and observed variances at the control marker indicates the existence of population structure,

$$\sigma_Z^2 - E[\sigma_Z^2] = 2m(1-m)\delta_C^2 \quad \text{(II-4.13)}$$

The spurious term in the covariance in equation (II-4.6) can be completely corrected using a single control marker, as follows:

$$\begin{aligned} \tilde{\sigma}_{X,Y} &= \sigma_{X,Y} - \frac{\sigma_{X,Z}\sigma_{Y,Z}}{2\{\sigma_Z^2 - E[\sigma_Z^2]\}} \\ &= 2mD_{TQ}^{(1)}(d + h - 2hp_Q^{(1)}) + 2(1-m)D_{TQ}^{(2)}(d + h - 2hp_Q^{(2)}) \end{aligned} \quad \text{(II-4.14)}$$

Therefore, the regression coefficient calculated from

$$b_1 = \frac{\tilde{\sigma}_{X,Y}}{\sigma_X^2} = \frac{\sigma_{X,Y} - \dfrac{\sigma_{X,Z}\sigma_{Y,Z}}{2\{\sigma_Z^2 - E[\sigma_Z^2]\}}}{\sigma_X^2} \quad \text{(II-4.15)}$$

would reflect correction for the population structure.

2.4.1.2 Significant test of the regression coefficient b_1

In this novel model, the standard Student's *t*-test can be used to test for significance of the regression coefficient b_1. In *t*-test, the statistic has the form

$$t-\text{test}=\frac{Z}{S} \tag{II-4.16}$$

where Z and S represent statistic to be tested and standard error of Z, respectively. In the present setting, Z is the regression coefficient b_1 in the analysis and Standard error of b_1 is given by

$$S_{b_1}=\sqrt{\frac{\sigma_X^2\sigma_Y^2-\tilde{\sigma}_{X,Y}^2}{n\sigma_X^2}} \tag{II-4.17}$$

Given the regression coefficient and the standard deviation, the power of the regression analysis can be predicted from the probability (Johnson and Kotz, 1970)

$$\rho_t=\Pr\{t_v(\delta_t)>t_{\alpha/2;v}\} \tag{II-4.18}$$

where $t_v(\delta_t)$ represents a random variable with non-central t-distribution with v degrees of freedom and non-centrality parameter δ_t and $t_{\alpha/2;v}$ is the upper $\alpha/2$ point of a central t-variable with the same degrees of freedom. The value of v equals n-3 and the non-centrality parameter is given by (Johnson and Kotz, 1970) as

$$\delta_t=\frac{\Gamma[v/2]b_1}{\sqrt{v/2}\,\Gamma[(v-1)/2]S_{b_1}} \tag{II-4.19}$$

where $\Gamma(\bullet)$ stands for a gamma function.

2.4.1.3 Selection of the control marker

In practice, I propose the following procedure to select the control marker for a given test marker. Firstly, any marker but the test marker would be candidate for the control marker if it has or is

- autosomal location
- locates on different chromosomes from the test marker,
- less missing genotype data than a prior given proportion

For each marker passing the above screening, one calculates the expected and observed variances from

$$E[\sigma_Z^2] = 2\,p_C(1-p_C) \qquad \text{(II-4.20)}$$

$$\sigma_Z^2 = \sum_{i=1}^{n}(Z_i-\mu)^2/(n-1) \qquad \text{(II-4.21)}$$

where Z_i is the genotypic value of the candidate control marker (0, 1, 2) for individual $i = 1, \cdots \cdots, n$, and μ and p_C are the mean genotypic value across all individuals ($\sum_{i=1}^{n} Z_i / n$) and the allelic frequency of this marker, respectively. It should be noted that equations (II-4.11) and (II-4.20) are the same and that equation (II-4.21) stands for the sampling variance of the control marker whose expectation is given by equation (II-4.12) in the presence of population structure. The control marker is the one with the maximum difference between observed and expected variances, which has the maximum ability to remove the spurious term in mixed populations and does not introduce bias in single population.

2.4.2 Method 2 (Regression analysis without correcting population structure)

The traditional method fits a simple regression model for detecting LD between the trait phenotype and a test marker as it has been proposed previously (Spielman et al., 2007)

and implemented in a recent population based eQTL analysis in (Luo, 1998), in which the regression coefficient has a form of

$$b_1 = \frac{\sigma^*_{X,Y}}{\sigma^2_X} \qquad \text{(II-4.22)}$$

with a standard error equal to

$$S_{b_1} = \frac{\sigma^2_X \sigma^2_Y - (\sigma^*_{X,Y})^2}{n\sigma^2_X} \qquad \text{(II-4.23)}$$

where $\sigma^*_{X,Y}$ is the non-corrected covariance between test marker locus and the quantitative trait.

2.4.3 Method 3 (multiple regression analysis)

The method regresses the trait phenotype on genotypic value of a test marker (X_{ij} = 0, 1, 2) and the probability of membership to each constituent population P_i (i =1, 2 here) as described in the following multiple regression model

$$Y_i = b_0 + b_1 X_{ij} + b_2 P_i + \varepsilon_i \qquad \text{(II-4.24)}$$

where the $b_2 P_i$ term reflects the population structure effect in mixed populations.

The regression coefficients are given by

$$b_1 = \frac{\sigma^2_P \sigma_{X,Y} - \sigma_{X,P}\sigma_{P,Y}}{\sigma^2_X \sigma^2_P - \sigma^2_{X,P}} \qquad \text{(II-4.25)}$$

$$b_2 = \frac{\sigma^2_X \sigma_{P,Y} - \sigma_{X,P}\sigma_{X,Y}}{\sigma^2_X \sigma^2_P - \sigma^2_{X,P}} \qquad \text{(II-4.26)}$$

where $\sigma_{X,Y}$ is the covariance between test marker and phenotypic trait, $\sigma_{P,Y}$ is the covariance between subpopulation information and phenotypic trait, and $\sigma_{X,P}$ is the covariance between test marker and subpopulation information. Furthermore, the standard errors of the regression coefficients are formulated as

$$S_{b_1} = \sqrt{\frac{\sigma_P^2 \sigma_Y^2}{n\sigma_X^2 \sigma_P^2 - \sigma_{X,P}^2}} \qquad \text{(II-4.27)}$$

$$S_{b_2} = \sqrt{\frac{\sigma_X^2 \sigma_Y^2}{n\sigma_X^2 \sigma_P^2 - \sigma_{X,P}^2}} \qquad \text{(II-4.28)}$$

according to (Snedecor and Cochran, 1967). Significance of association of the test marker with the quantitative trait can be tested through testing for significance of the regression coefficient b_1 by the Student *t*-test.

2.5 Simulation study

The novel method that only uses one control marker to correct the confound effect of population structure is very important for linkage-disequilibrium based association studies. This part presents the efficient performance of the novel method for association studies in admixed populations based on simulated samples. In order to comprehensively explore statistical properties and limitations of the methods described above, I developed and conducted a series of computation simulation studies. All of the simulation studies were implemented in FORTRAN program.

2.5.1 Simulations of admixed populations

A multiple-locus simulation program was developed to simulate diploid population. The simulation program mimics segregation pattern of genes at multiple marker loci and QTL in randomly mating natural populations in terms of simulation parameters defining allele frequencies, linkage disequilibria and population structure as illustrated in Table II.4. The methods were detailed for simulating a population characterized by the joint genotypic distribution at two loci and for sampling individuals from the simulated population. Although the distribution involves only two loci, it is easy to extend to multiple loci because the two locus joint distribution can be easily converted into conditional (or transition) probability distribution of genotypes at one locus on that at another, and genotypes at multiple loci can be simulated as a Markov process governed by the conditional probability distribution. Of course, this will not undermine flexibility to specify any required linkage disequilibrium pattern among any loci. Subpopulations were independently generated and merged to produce the admixed population. For each

subpopulation, the simulation started by randomly sampling genotypes at both QTL and control marker locus for an individual based on the pre-specified allele frequencies, respectively. Given a haplotype genotype at the QTL, the haplotype genotype at the test marker locus was sampled from a probability distribution as given in Table II.3. Phenotype of the trait was generated for each individual according to its genotype at the QTL and prior defined quantitative genetic model and parameters characterizing genetic effects at the QTL (Table II.3 and Table II.4), plus a random number sampled from a normal distribution of mean zero and variance σ_e^2. For simplicity, the QTL genotypic effects were expressed in terms of the QTL heritability defined as in an equilibrium population with QTL allelic frequency q=0.5. The phenotypic variance of the trait was assigned a constant value of 100. In the present study, I was focused on 20 simulated populations defined by simulation parameters listed in Table II.4.

Table II.4: Parameters defining two subpopulations that are merged to produce admixed populations.

Pop.	n	m	h^2	Φ	$p_Q^{(1)}$	$p_T^{(1)}$	$p_C^{(1)}$	$D_{TQ}^{(1)}$	$p_Q^{(2)}$	$p_T^{(2)}$	$p_C^{(2)}$	$D_{TQ}^{(2)}$
1	1000	0.3	0.1	1	0.8	0.4	0.8	0.0	0.5	0.7	0.4	0.0
2	500	0.5	0.1	0	0.8	0.4	0.3	0.0	0.5	0.7	0.6	0.0
3	500	0.5	0.1	0	0.8	0.4	0.8	0.0	0.5	0.7	0.4	0.0
4	500	0.5	0.1	0.5	0.7	0.7	0.3	0.00	0.3	0.3	0.7	0.00
5	500	0.5	0.1	0	0.2	0.8	0.2	0.00	0.8	0.2	0.8	0.00
6	500	0.5	0.2	0.5	0.2	0.8	0.2	0.00	0.8	0.2	0.8	0.00
7	1000	0.7	0.1	0	0.7	0.7	0.2	0.0	0.3	0.3	0.6	0.0
8	500	0.5	0.2	1	0.2	0.8	0.2	0.0	0.8	0.2	0.8	0.0
9	1000	0.5	0.2	1	0.2	0.8	0.2	0.0	0.8	0.2	0.8	0.0
10	500	0.5	0.1	0	0.3	0.7	0.3	0.08	0.7	0.3	0.8	-0.08
11	1000	0.5	0.1	0	0.6	0.6	0.6	0.1	0.4	0.4	0.4	0.1
12	500	0.5	0.1	0	0.6	0.6	0.6	0.1	0.4	0.4	0.4	0.1
13	500	0.7	0.1	0.5	0.4	0.4	0.4	0.05	0.6	0.6	0.6	0.05
14	500	0.5	0.1	0	0.6	0.4	0.6	0.1	0.4	0.4	0.4	0.1
15	500	0.5	0.2	0	0.4	0.5	0.4	0.1	0.6	0.5	0.6	0.1
16	500	0.5	0.1	1	0.4	0.5	0.4	0.05	0.6	0.5	0.6	0.05
17	500	0.4	0.1	0	0.3	0.7	0.5	0.05	0.7	0.4	0.8	0.05
18	500	0.4	0.1	0	0.3	0.7	0.4	0.05	0.7	0.4	0.8	0.05
19	500	0.4	0.1	0	0.3	0.7	0.3	0.05	0.7	0.4	0.8	0.05
20	500	0.4	0.1	0	0.3	0.7	0.2	0.05	0.7	0.3	0.8	0.05

n is the sample size, m is the proportion of subpopulation 1 in the admixture, h^2 is QTL heritability defined in an equilibrium population with QTL allelic frequency p=0.5, Φ is the dominance ratio at the QTL, $p_Q^{(i)}$, $p_T^{(i)}$ and $p_C^{(i)}$ are respectively the allelic frequencies at QTL, test marker and control marker in the i-th subpopulation (i=1,2), and $D_{TQ}^{(i)}$ is the coefficient of linkage disequilibrium between QTL and test marker in i-th subpopulation.

2.5.2 Comparisons of three approaches based on simulation data

The association studies between QTL and test marker in simulation data were implemented by **Method 1** (the novel method), **Method 2** (original simple linear regression) and **Method 3** (multiple regression including known population ancestry). For each method, the simulated observation of the risk/power to predict spurious/genuine association at the test marker was calculated as the frequency of significant statistical tests (significance threshold $\alpha = 0.0001$) over 100 repeated simulations, implying an overall type I error of 1%.

2.5.2.1 Probability of statistical power and false positive inference

I tabulated in Table II.5 means and standard errors of 100 repeated regression coefficients and proportions of significant tests of the regression coefficients. It can be seen that simulations showed good agreement with the theoretical predictions for all three methods.

Listed in Table II.5 were proportions of significant tests of the regression in repeated simulations. In populations 1-9, LD between test marker and QTL was equal to zero in both of two subpopulations, thus there were no genuine LD in admixed populations. Furthermore, in population 10, LD between test marker and QTL had the opposite sign in two subpopulations, and thus completely counterbalanced the genuine LD in admixed populations (genuine LD=0 in admixed populations). It should be stressed that populations 1-10 have no LD between QTL and test marker; the differences in allelic frequency distribution between subpopulations led to spurious prediction of LD in the

structured populations, therefore the 'proportions of statistical tests for significance of the regression coefficients' here is equivalent to the risk of claiming false positives. It is clear that the rate of false positive is properly controlled in association analysis with **Method 1,** and **Method 3**. However, the **Method 3** (multiple regression method) is completely effective in avoiding false positives when the population structure information is fully known. While, **Method 1** predicted the regression coefficients and power to zero at the test marker without the requirement of population structure information. In contrast, **Method 2** did not correct for the population structure, and hence had the highest false positive rates in both simulation and theoretical prediction, performing particularly poorly in populations 4 -10, where the false positive rate reached approximately 100%.

When populations 11-20 have QTL and test marker in true LD, the proportion measures rate provides evaluation of an empirical statistical power for detecting the genetic association. In populations 11-13 the spurious LD from population structure had the same sign as the genuine LD, thus increasing the absolute value of the LD coefficient between QTL and test marker; in this situation, **Method 2** had the highest power to detect the true associations, while **Methods 1** and **3** performed similarly. But both **Methods 1** and **3** could also detect truly existing LD with a high statistical power. In populations 14-16 the test marker had constant frequency between the two subpopulations, so there was no spurious LD between the QTL and test marker. The three methods performed similarly, as expected. In populations 17-20 the spurious LD from population structure had the opposite sign, and thus partially counterbalanced the genuine LD. The observed LD were less than half of the true LD values in the original subpopulations, and **Method 2** completely lost power to predict true associations. By

contrast, both **Methods 1** and **3** were able to correct for population structure, with empirical powers of around 50% and more than 80% respectively.

According to these extensive simulation studies, the **Method 2** is thus inappropriate to be used for genetic association analysis when population structure was present. Although **Method 3** (multiple regression analysis incorporated membership of individuals to constituent populations as a covariate) can precisely correct the spurious association and detect the genuine LD, it seriously depends on the prior population structure information. I have investigated how the knowledge of population structure information could influence the association studies in subsequent analysis. Overall, these results show that the novel method (**Method 1**) provides a powerful test for linkage disequilibrium between polymorphic markers and QTL and an effective control of population structure in the test.

Table II.5: Means and standard errors of regression coefficients (b±se) and proportions (ρ or $\hat{\rho}$) of statistical tests for significance of the regression coefficients from three methods.

Pop	D_{TQ}	D'_{TQ}	Method 1 (the novel method)				Method 2 (simple regression)				Method 3 (Multiple regression)			
			Simulated		Predicted		Simulated		Predicted		Simulated		Predicted	
			b ± se	$\hat{\rho}$	b	ρ	b ± se	$\hat{\rho}$	b	ρ	b ± se[a]	$\hat{\rho}$ [a]	b ± se[b]	$\hat{\rho}$ [b]
1	-0.019	0.00	0.01±0.00	0.01	0.00	0.00	-0.38±0.01	0.13	-0.38	0.19	0.00±0.01	0.00	0.00	0.00
2	-0.022	0.00	0.13±0.02	0.10	0.00	0.00	-0.75±0.01	0.44	-0.75	0.50	-0.00±0.01	0.00	0.00	0.00
3	-0.022	0.00	0.05±0.01	0.03	0.00	0.00	-0.75±0.01	0.43	-0.75	0.50	-0.01±0.01	0.00	0.00	0.00
4	0.04	0.00	-0.087±0.015	0.07	0.00	0.00	1.162±0.006	0.97	1.163	0.98	-0.00±0.007	0.00	0.00	0.00
5	-0.09	0.00	0.015±0.008	0.00	0.00	0.00	-2.371±0.005	1.00	-2.368	1.00	0.006±0.007	0.00	0.00	0.00
6	-0.09	0.00	0.005±0.011	0.00	0.00	0.00	-3.157±0.007	1.00	-3.157	1.00	-0.00±0.009	0.00	0.00	0.00
7	0.034	0.00	0.04±0.01	0.06	0.00	0.00	1.08±0.01	1.00	1.08	1.00	-0.01±0.01	0.00	0.00	0.00
8	-0.090	0.00	0.03±0.01	0.00	0.00	0.00	-2.74±0.01	1.00	-2.74	1.00	0.02±0.01	0.00	0.00	0.00
9	-0.090	0.00	0.01±0.01	0.00	0.00	0.00	-2.74±0.01	1.00	-2.74	1.00	-0.00±0.01	0.00	0.00	0.00
10	-0.04	0.00	0.008±0.009	0.01	0.00	0.00	-1.233±0.006	0.99	-1.234	0.99	-0.00±0.007	0.00	0.00	0.00
11	0.110	0.10	1.61±0.02	0.96	1.72	1.00	2.07±0.01	1.00	2.06	1.00	1.76±0.01	0.99	1.86	1.00
12	0.110	0.10	1.62±0.02	0.95	1.72	1.00	2.06±0.01	1.00	2.06	1.00	1.75±0.01	0.98	1.86	1.00
13	0.058	0.05	0.80±0.03	0.64	0.85	0.61	1.14±0.01	0.94	1.13	0.96	0.92±0.01	0.64	0.91	0.71
14	0.100	0.10	1.78±0.02	0.98	1.86	1.00	1.86±0.01	1.00	1.86	1.00	1.86±0.01	1.00	1.86	1.00
15	0.100	0.10	2.37±0.03	0.97	2.53	1.00	2.52±0.01	1.00	2.53	1.00	2.52±0.01	1.00	2.53	1.00
16	0.050	0.05	0.71±0.02	0.36	0.73	0.39	0.72±0.01	0.33	0.73	0.39	0.72±0.01	0.36	0.73	0.43
17	0.021	0.05	0.97±0.02	0.49	0.83	0.56	-0.12±0.01	0.00	-0.13	0.00	0.99±0.01	0.85	0.98	0.90
18	0.021	0.05	0.88±0.01	0.48	0.83	0.56	-0.13±0.01	0.00	-0.13	0.00	0.97±0.01	0.84	0.98	0.90
19	0.021	0.05	0.84±0.01	0.50	0.83	0.56	-0.11±0.01	0.00	-0.13	0.00	0.99±0.01	0.88	0.98	0.90
20	0.012	0.05	0.79±0.01	0.45	0.78	0.51	-0.42±0.01	0.03	-0.42	0.04	1.06±0.01	0.89	1.07	0.94

D_{rQ} and D'_{rQ} are the coefficients of LD between the marker and QTL in the simulated mixed population before and after correction for population structure respectively. [a]predicted when all individuals were allocated to their correct subpopulations; The predicted values were estimated from theoretical analysis, while the simulated values were estimated from the simulation studies.

2.5.2.2 The performance of **Method 3** when the population structure information is unknown

Method 3 (multiple regression analysis with prior population structure information) apparently had the well performance both in correcting for spurious associations and in predicting genuine associations, but its prerequisite for the correct assignment of individuals to subpopulations is a serious shortcoming. To investigate how this could influence the association studies, I further simulated three scenarios in which only a proportion r of the individuals was correctly assigned to subpopulations, while (1-r) were assigned population membership either incorrectly (scenario 1, Table II.6), randomly (scenario 2, Table II.7), or partially with 60% correctly assigned (scenario 3, Table II.8).

In all scenarios, **Method 3** can correct spurious associations (populations 1–10) and detect genuine associations (populations 11–20) when 95% individuals are correctly assigned to subpopulations (r = 0.95). However, when r is reduced then the power to detect genuine associations and the ability to remove the spurious associations both decrease correspondingly; when r = 0.50 then the false positive rate could rise to 100% and the power declined to zero in some situations. Generally, the **Method 3** could lose its statistical power to detect the truly existing LD (populations 11–20) or make false positive inference of genetic association (populations 1–10) when on average a quarter of individuals under analysis were wrongly allocated to subpopulations. Thus, performance of **Method 3** serious depends on the extent by which individuals are correctly allocated to their belonging populations. In reality, **Method 3** could be

impractical for association studies due to the computational difficulty of correctly assigning individuals to subpopulations. However, **Method 1** that I have developed here addresses this limitation, because it successfully controls the false positive rate to a low level, and has sufficient power to detect genuine associations in admixed populations, while not requiring any individual ancestral information.

Table II.6: The values of regression coefficient (*b*) and empirical statistical power estimated from Method 3 when a proportion of individuals *r* was correctly assigned to subpopulations while (1- *r*) **were incorrectly assigned**.

Pop.	*r* = 0.5		*r* = 0.6		*r* = 0.7		*r* = 0.8		*r* = 0.9		*r* = 0.95	
	b ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power
1	-0.375±0.004	0.15	-0.364±0.004	0.14	-0.330±0.004	0.07	-0.267±0.004	0.03	-0.163±0.004	0.01	-0.089±0.005	0.00
2	-0.748±0.006	0.41	-0.723±0.006	0.36	-0.646±0.006	0.25	-0.508±0.006	0.10	-0.301±0.006	0.01	-0.162±0.006	0.00
3	-0.751±0.006	0.44	-0.726±0.006	0.40	-0.649±0.006	0.26	-0.511±0.007	0.09	-0.301±0.007	0.01	-0.166±0.007	0.00
4	1.236±0.006	0.98	1.201±0.006	0.98	1.086±0.007	0.92	0.879±0.007	0.63	0.542±0.007	0.10	0.303±0.007	0.01
5	-2.362±0.005	1.00	-2.315±0.005	1.00	-2.164±0.006	1.00	-1.866±0.006	1.00	-1.280±0.007	0.98	-0.781±0.008	0.44
6	-3.348±0.007	1.00	-3.284±0.007	1.00	-3.066±0.008	1.00	-2.637±0.009	1.00	-1.817±0.010	0.98	-1.112±0.010	0.49
7	1.084±0.004	1.00	1.056±0.005	1.00	0.967±0.005	1.00	0.798±0.005	0.93	0.508±0.005	0.35	0.295±0.005	0.03
8	-2.744±0.007	1.00	-2.693±0.007	1.00	-2.516±0.007	1.00	-2.182±0.008	1.00	-1.492±0.010	0.87	-0.919±0.010	0.24
9	-2.749±0.005	1.00	-2.696±0.005	1.00	-2.521±0.005	1.00	-2.177±0.006	1.00	-1.503±0.007	1.00	-0.920±0.007	0.68
10	-1.236±0.006	0.99	-1.200±0.006	0.98	-1.083±0.007	0.91	-0.878±0.007	0.62	-0.535±0.007	0.11	-0.299±0.007	0.01
11	2.066±0.004	1.00	2.058±0.004	1.00	2.036±0.004	1.00	1.997±0.004	1.00	1.941±0.004	1.00	1.906±0.004	1.00
12	2.061±0.006	1.00	2.055±0.006	1.00	2.032±0.006	1.00	1.993±0.006	1.00	1.936±0.006	1.00	1.899±0.006	1.00
13	1.143±0.006	0.94	1.136±0.006	0.94	1.114±0.006	0.92	1.076±0.006	0.89	1.020±0.006	0.81	0.976±0.006	0.75
14	1.860±0.007	1.00	1.861±0.007	1.00	1.860±0.007	1.00	1.861±0.007	1.00	1.862±0.006	1.00	1.861±0.006	1.00
15	2.528±0.009	1.00	2.528±0.009	1.00	2.527±0.009	1.00	2.526±0.009	1.00	2.526±0.009	1.00	2.525±0.009	1.00
16	0.728±0.007	0.34	0.728±0.007	0.34	0.726±0.007	0.33	0.727±0.007	0.34	0.728±0.007	0.36	0.727±0.007	0.36
17	-0.133±0.007	0.00	-0.095±0.007	0.00	0.015±0.007	0.00	0.213±0.007	0.00	0.523±0.007	0.08	0.728±0.006	0.36
18	-0.109±0.007	0.00	-0.072±0.007	0.00	0.038±0.007	0.00	0.239±0.007	0.00	0.540±0.007	0.09	0.743±0.007	0.40
19	-0.134±0.007	0.00	-0.099±0.007	0.00	0.014±0.007	0.00	0.212±0.007	0.00	0.520±0.007	0.08	0.727±0.006	0.36
20	-0.432±0.006	0.03	-0.388±0.006	0.02	-0.256±0.007	0.01	-0.012±0.007	0.00	0.392±0.007	0.02	0.690±0.007	0.28

Table II.7: The values of regression coefficient (*b*) and empirical statistical power estimated from Method 3 when a proportion of individuals *r* was correctly assigned to subpopulations while (1- *r*) **were randomly assigned (50% probability to each subpopulation)**.

Pop.	*r* = 0.5		*r* = 0.6		*r* = 0.7		*r* = 0.8		*r* = 0.9		*r* = 0.95	
	b ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power
1	-0.312±0.004	0.05	-0.278±0.004	0.02	-0.229±0.004	0.01	-0.173±0.004	0.00	-0.098±0.004	0.00	-0.055±0.004	0.00
2	-0.601±0.006	0.18	-0.527±0.006	0.09	-0.430±0.006	0.04	-0.316±0.006	0.01	-0.179±0.006	0.00	-0.100±0.006	0.00
3	-0.592±0.006	0.17	-0.516±0.006	0.10	-0.422±0.006	0.04	-0.307±0.006	0.01	-0.167±0.007	0.00	-0.092±0.006	0.00
4	0.997±0.007	0.81	0.885±0.007	0.64	0.731±0.007	0.36	0.544±0.007	0.10	0.308±0.007	0.01	0.160±0.007	0.00
5	-2.038±0.006	1.00	-1.865±0.006	1.00	-1.616±0.007	1.00	-1.282±0.007	0.98	-0.776±0.008	0.43	-0.434±0.008	0.05
6	-2.891±0.008	1.00	-2.647±0.008	1.00	-2.293±0.009	1.00	-1.806±0.010	0.98	-1.101±0.010	0.47	-0.625±0.010	0.05
7	0.889±0.005	0.98	0.789±0.005	0.93	0.665±0.005	0.72	0.503±0.005	0.32	0.282±0.005	0.02	0.148±0.005	0.00
8	-2.362±0.008	1.00	-2.164±0.009	1.00	-1.878±0.009	0.99	-1.486±0.010	0.89	-0.906±0.010	0.23	-0.508±0.011	0.03
9	-2.363±0.006	1.00	-2.164±0.006	1.00	-1.885±0.006	1.00	-1.481±0.007	1.00	-0.912±0.007	0.66	-0.510±0.007	0.08
10	-0.995±0.007	0.80	-0.878±0.007	0.62	-0.721±0.007	0.33	-0.542±0.007	0.11	-0.297±0.007	0.01	-0.160±0.007	0.00
11	2.021±0.004	1.00	2.000±0.004	1.00	1.974±0.004	1.00	1.944±0.004	1.00	1.909±0.004	1.00	1.889±0.004	1.00
12	2.019±0.006	1.00	1.996±0.006	1.00	1.972±0.006	1.00	1.943±0.006	1.00	1.907±0.006	1.00	1.887±0.006	1.00
13	1.084±0.007	0.89	1.062±0.007	0.86	1.035±0.007	0.83	1.004±0.007	0.79	0.959±0.007	0.72	0.938±0.007	0.68
14	1.864±0.006	1.00	1.864±0.006	1.00	1.865±0.006	1.00	1.864±0.006	1.00	1.865±0.006	1.00	1.865±0.006	1.00
15	2.524±0.009	1.00	2.524±0.009	1.00	2.519±0.009	1.00	2.521±0.009	1.00	2.519±0.009	1.00	2.521±0.009	1.00
16	0.729±0.007	0.33	0.728±0.007	0.34	0.728±0.007	0.34	0.729±0.007	0.35	0.729±0.007	0.34	0.728±0.007	0.35
17	0.093±0.007	0.00	0.205±0.007	0.00	0.346±0.006	0.01	0.513±0.006	0.08	0.717±0.006	0.33	0.837±0.006	0.6
18	0.107±0.007	0.00	0.215±0.007	0.00	0.353±0.007	0.02	0.523±0.007	0.09	0.726±0.007	0.36	0.849±0.006	0.61
19	0.098±0.007	0.00	0.212±0.007	0.00	0.349±0.007	0.02	0.512±0.007	0.08	0.723±0.007	0.35	0.847±0.006	0.61
20	-0.130±0.007	0.00	0.015±0.007	0.00	0.196±0.007	0.00	0.421±0.007	0.02	0.701±0.007	0.28	0.875±0.006	0.62

Table II.8: The values of regression coefficient (*b*) and empirical statistical power estimated from Method 3 when a proportion of individuals *r* was correctly assigned to subpopulations while for (1- *r*) population membership information was **partially known (60%)**.

Pop.	*r* = 0.5		*r* = 0.6		*r* = 0.7		*r* = 0.8		*r* = 0.9		*r* = 0.95	
	b ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power	*b* ± se	Power
1	-0.151±0.004	0.00	-0.127±0.004	0.00	-0.097±0.004	0.00	-0.070±0.004	0.00	-0.038±0.004	0.00	-0.022±0.004	0.00
2	-0.277±0.006	0.01	-0.228±0.006	0.00	-0.177±0.006	0.00	-0.124±0.006	0.00	-0.073±0.006	0.00	-0.043±0.006	0.00
3	-0.267±0.006	0.01	-0.220±0.006	0.00	-0.172±0.007	0.00	-0.117±0.006	0.00	-0.061±0.006	0.00	-0.034±0.006	0.00
4	0.472±0.007	0.05	0.396±0.007	0.02	0.304±0.007	0.01	0.210±0.007	0.00	0.109±0.007	0.00	0.057±0.007	0.00
5	-1.143±0.006	0.95	-0.973±0.006	0.78	-0.778±0.007	0.41	-0.556±0.007	0.10	-0.296±0.007	0.01	-0.152±0.007	0.00
6	-1.626±0.009	0.96	-1.389±0.009	0.83	-1.100±0.009	0.45	-0.781±0.009	0.12	-0.420±0.009	0.01	-0.218±0.009	0.01
7	0.439±0.005	0.17	0.363±0.005	0.07	0.281±0.005	0.02	0.194±0.005	0.01	0.096±0.005	0.00	0.043±0.005	0.00
8	-1.328±0.009	0.75	-1.135±0.009	0.49	-0.906±0.010	0.20	-0.647±0.010	0.05	-0.346±0.010	0.00	-0.177±0.010	0.00
9	-1.325±0.006	0.99	-1.137±0.006	0.95	-0.909±0.006	0.67	-0.642±0.007	0.19	-0.347±0.007	0.01	-0.181±0.007	0.00
10	-0.469±0.007	0.06	-0.389±0.007	0.02	-0.297±0.007	0.01	-0.204±0.007	0.00	-0.102±0.007	0.00	-0.051±0.007	0.00
11	1.933±0.004	1.00	1.921±0.004	1.00	1.909±0.004	1.00	1.895±0.004	1.00	1.882±0.004	1.00	1.875±0.004	1.00
12	1.932±0.006	1.00	1.918±0.006	1.00	1.908±0.006	1.00	1.894±0.006	1.00	1.880±0.006	1.00	1.873±0.006	1.00
13	0.991±0.007	0.77	0.977±0.007	0.76	0.961±0.007	0.74	0.947±0.007	0.69	0.927±0.007	0.66	0.920±0.007	0.65
14	1.866±0.006	1.00	1.864±0.006	1.00	1.865±0.006	1.00	1.865±0.006	1.00	1.865±0.006	1.00	1.865±0.006	1.00
15	2.523±0.009	1.00	2.521±0.009	1.00	2.517±0.009	1.00	2.521±0.008	1.00	2.519±0.008	1.00	2.520±0.008	1.00
16	0.730±0.007	0.34	0.729±0.007	0.35	0.728±0.007	0.34	0.728±0.007	0.35	0.728±0.007	0.35	0.728±0.007	0.36
17	0.573±0.006	0.11	0.646±0.006	0.22	0.722±0.006	0.35	0.802±0.006	0.50	0.881±0.006	0.69	0.924±0.006	0.77
18	0.582±0.007	0.14	0.652±0.007	0.23	0.731±0.006	0.37	0.809±0.006	0.52	0.890±0.006	0.70	0.936±0.006	0.78
19	0.582±0.007	0.14	0.651±0.007	0.23	0.727±0.006	0.38	0.806±0.006	0.52	0.889±0.006	0.70	0.934±0.006	0.78
20	0.492±0.006	0.05	0.599±0.006	0.14	0.706±0.006	0.28	0.821±0.006	0.51	0.942±0.006	0.75	1.005±0.006	0.84

2.5.2.3 Performance of the novel method (**Method 1**) using varied control markers

Use of control markers in **Method 1** is the key underpinning for the method to be able to control influence of population structure in the genetic association test. To investigate effect of the control marker on efficiency of the association test, I explored performance of the novel method when different control markers are used in the presence of population structure. Table II.9 shows predicted and observed proportions of significant tests of the disequilibrium between a test marker and a putative QTL in 10 simulation populations with various population structures. Here, all of the 10 simulation populations were generated by same parameter set (simulation population 11 from table II.4), except the allele frequencies of control marker. The proportions were calculated from analyses with **Method 1** by using the control marker with varying allele frequencies. When population structure is present in samples, the method bears a high chance to make a false positive inference and to lose its detecting power if the control marker selected to be implemented in the analysis has a small difference in allele frequency between the subpopulations. However, the risk can be effectively controlled and the reduced power can be recovered when using the control marker with a large allele frequency difference.

Table II.9: Investigating the ability of control marker that can remove 'spurious association' between two subpopulations (Here, the parameters defining the two simulated subpopulations come from simulation population 11.)

Pop	$p_C^{(1)}$	$p_C^{(2)}$	$\left\|p_C^{(1)} - p_C^{(2)}\right\|$	$D_{MQ}^{(1)}$	$D_{MQ}^{(2)}$	D_{MQ}	D'_{MQ}	Method 1 (the novel method)			
								Simulated		Predicted	
								b ± se	Power	b	Power
1	0.50	0.50	0.00	0.00	0.00	-0.03	0.00	-0.924±0.028	**0.71**	0.000	0.00
2	0.60	0.40	0.20	0.00	0.00	-0.03	0.00	0.066±0.035	**0.25**	0.000	0.00
3	0.70	0.30	0.40	0.00	0.00	-0.03	0.00	0.049±0.014	**0.04**	0.000	0.00
4	0.80	0.20	0.60	0.00	0.00	-0.03	0.00	-0.005±0.010	0.00	0.000	0.00
5	0.90	0.10	0.80	0.00	0.00	-0.03	0.00	0.006±0.006	0.00	0.000	0.00
6	0.50	0.50	0.00	0.08	0.08	0.05	0.08	0.384±0.027	**0.10**	1.325	1.00
7	0.60	0.40	0.20	0.08	0.08	0.05	0.08	1.292±0.029	**0.70**	1.325	1.00
8	0.70	0.30	0.40	0.08	0.08	0.05	0.08	1.385±0.013	**0.94**	1.325	1.00
9	0.80	0.20	0.60	0.08	0.08	0.05	0.08	1.325±0.006	1.00	1.325	1.00
10	0.90	0.10	0.80	0.08	0.08	0.05	0.08	1.321±0.005	1.00	1.325	1.00

$p_C^{(i)}$ is respectively the allelic frequencies at control marker in the i-th subpopulation (i=1,2), and $D_{MQ}^{(i)}$ is the coefficient of linkage disequilibrium between QTL and test marker in i-th subpopulation. D_{MQ} and D'_{MQ} are simulated coefficients of LD before and after correction for population structure respectively.

Furthermore, I also explored the performance of **Method 1** (the novel method) when population structure is actually absent. Table II.10 presents the simulated and predicted proportions of significant test of association studies when there is no population structure or the diversity of allelic frequencies at QTL and test marker between the two subpopulations is equal. It demonstrates that the type I error is well controlled and the disequilibrium is efficiently detected by the method using a control marker even when population structure does not actually exist. In addition, all these suggest that implementation of control markers with a non-trial difference in allele frequency will not cause any significant problem of false positive/negative inference when population stratification is actually not existent. In presence of population structure, I propose selection of a marker with largely divergent allele frequencies as the control marker.

Table II.10: Predicted and observed proportions of significant tests of linkage disequilibrium between a test marker and a putative QTL in different simulation populations without population stratification from Method 1 in which the control marker implemented into the analyses had a constant allele frequency difference of 0.4.

Pop	$p_Q^{(1)} = p_Q^{(2)}$	$p_T^{(1)} = p_T^{(2)}$	$\left\|p_C^{(1)} - p_C^{(2)}\right\|$	D'_{TQ}	Simulated		Predicted	
					$\hat{b} \pm se$	$\hat{\rho}$	b	ρ
1	0.50	0.50	0.40	0.00	0.003±0.007	0.00	0.00	0.00
2	0.55	0.50	0.40	0.00	0.001±0.007	0.00	0.00	0.00
3	0.55	0.45	0.40	0.00	-0.001±0.007	0.00	0.00	0.00
4	0.60	0.45	0.40	0.00	0.019±0.008	0.00	0.00	0.00
5	0.60	0.40	0.40	0.00	0.024±0.009	0.01	0.00	0.00
6	0.50	0.50	0.40	0.05	0.894±0.007	0.66	0.894	0.74
7	0.55	0.50	0.40	0.05	0.899±0.007	0.67	0.894	0.74
8	0.55	0.45	0.40	0.05	0.878±0.007	0.65	0.886	0.72
9	0.60	0.45	0.40	0.06	1.058±0.007	0.87	1.052	0.93
10	0.60	0.40	0.40	0.06	1.475±0.011	0.85	1.46	0.91

2.6 Real data analysis

From previous simulation studies, I have demonstrated that the novel method using only one control marker could efficiently and precisely remove the spurious associations and detect the genuine associations. This part will compare the performance of **Method 1** (the novel method) with **Method 2 and 3** in genome wide association studies based on real data analysis — genome wide expression quantitative trait loci (eQTL) analysis.

2.6.1 Data resources and pretreatment

The genome-wide eQTL analysis requires the information of genome-wide genetic variants in a large collected sample group and the whole genome expression levels for each corresponding individual. Due to the expense and labour involved, the genome-wide approach to investigate the genetic architecture of gene expression variation has not been feasible until the last few years (genotyping technique has considerably improved and become much cheaper). The microarray, which considers as the first high-throughput method for genotyping the genetic variants, has moved the genome-wide association mapping from the futuristic to the realistic. Furthermore, microarray may not only be used for genotyping the genome-wide genetic polymorphisms, but also for high-throughput profiling the expression levels of whole genome-wide. Here, two publicly microarray datasets collected from the International HapMap Project (http://www.hapmap.org) and NCBI Gene Expression Omnibus Database (http://www.ncbi.nlm.nih.gov/geo) were used in real data analysis.

2.6.1.1 Gene expression data

The gene expression datasets were collected from Epstein-Barr virus (EBV) transformed lymphoblastoid cell lines of unrelated individuals of European-derived (CEU, 60 Europeans), and Asia-derived (CHB+JPT, 41 Chinese and 41 Japanese). The datasets were originally developed by Spielman et al (2007) to explore population specified gene expression and genetic control of the population specified gene expression, and were downloaded from National Center for Biotechnology Information, Gene Expression Omnibus Database under the accession number GSE5859 (http://www.ncbi.nlm.nih.gov/geo).

For this dataset, the expression profiles were measured by Affymetrix Human Genome Focus Target Array, which provides relatively comprehensive coverage of the human genome. In this commercial expression array, it targets over 8,400 well-characterized human genes. The unique DNA sequences used to represent the genes are selected from GeneBank Database (http://www.ncbi.nlm.nih.gov/genbank/), dbEST (http://www.ncbi.nlm.nih.gov/dbEST/) and RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/). The oligonucleotide probes complementary to each selected DNA sequence are synthesized and fixed on the array. A total 11 pairs of oligonucleotide probe sets are used to present each gene. Table II.11 briefly summarized the basic features of Affymetrix HG Focus array. More technical details of this microarray please refer to the AFFYMETRIX company official website (http://www.affymetrix.com/). The raw hybridized signals from expression microarrays were analyzed using the Affymetrix standard strategy MAS 5.0 and the hybridization intensity was $\log_2$-transformed into expression phenotype. The study focused on 4,197

genes that are expressed at least 80% of the samples in lymphoblastoid cell lines. Of the 4,197 genes, 1,097 were detected significantly differentially expressed between the CEU and CHB+JPT samples (*t*-test, $P < 10^{-5}$; $P_c < 0.05$, Sidak correction) (Westfall and Young, 1993).

Table II.11: Basic features of Affymetrix Human Genome Focus Target Array

Features \ Array	Human Genome Focus Array
Number of probe sets	~8,700
Number of transcripts	~8,500
Number of genes	~8,400
Number of control probe sets	~200
Oligonucleotide probe length	25 bps
Probe pairs	11

2.6.1.2 Genome-wide genotype datasets for 142 human individuals

Genotype data for the corresponding 60 CEU, 41 CHB and 41 JPT samples were obtained from the International HapMap Project (release 19, http://www.hapmap.org). The International HapMap Project, which was officially established in 2002, primarily aimed to provide a genome wide database of common genetic variation in human genome for guiding the design and analysis of clinical studies. Recently, this project has become a crucial resource for geneticists to implement the genetic studies of complex traits in human genome. Furthermore, the genotyping data created by this project is freely available to geneticists all over the world.

The initial version (phase 1) of International HapMap Project has successfully genotyped at least one common SNP marker in every 5 kilobases (KBs) interval across the whole human genome. The genotyping assay was implemented by five different

platforms under 9 research institutes (table II.12). The common marker meant that the project only focused on the SNP markers with minor allele frequency (MAF) equal or greater than 5%. In total, more than 1 million common SNP markers have been genotyped for each individual.

Table II.12: Genotyping platforms and research institutes in HapMap Project

Research institute	Genotyping platforms
Welcome Trust Sanger Institute in UK	*Illumina BeadArray*
McGill University and Genome Quebec innovation center	*Illumina BeadArray*
Illumina Inc.	*Illumina BeadArray*
Chinese HapMap Consortium	*Illumina BeadArray*
Broad Institute of Harvard and MIT in USA	*Illumina BeadArray*
RIKEN in Japan	*Third Wave Invader*
Baylor College of Medicine with ParAllele Bioscience	*ParAelle MIP*
University of California and Washington University in USA	*PerkinElmer Acycolprime-FP*
Perlegene Sciences Inc.	*Affymetrix Microarray*

Here, I downloaded the genotyping dataset (released at the 19th version) for the present genome wide eQTL analysis. Compared with initial (phase 1) dataset, additional million common SNP markers were genotyped for the same samples. Most of the additional genotype data for Release 19 version were generated from the Affymetrix GeneChip Mapping Array 500K set, the Illumina HumanHap100 and HumanHap 300 SNP assays. Overall, there were more than 2.2 million and 2.0 million common SNP markers for the CEU samples and CHB+JPT samples respectively. Comparison between the CEU and CHB+JPT samples provided genotype data of 1,606,182 unique SNP markers among all 142 individuals (60 CUE and 82 CHB+JPT samples).

I selected and re-analysed the gene expression and SNP datasets in the present study for several reasons. Firstly, these samples were collected from the populations whose genetic diversification was well verified (Consortium, 2003, Consortium, 2005, Consortium, 2007), and make a typical example which the method is designed for. Secondly, gene expression phenotype bears a wide spectrum of genetic controls from *cis* to *trans* regulation and different levels of heritability. Some of these quantitative phenotypes show population specified expression or heterogeneity of underlying genetics. These enable the method to be tested under different genetic backgrounds. Finally, re-analysis of the same datasets recently published allows a direct comparison of analysis with the method developed in the present study with that implemented in the published analysis.

2.6.2 Validation of population structure

In 2005, The International HapMap Project reported that the CHB and JPT samples' allele frequencies were generally very similar, but different to the allele frequencies of CEU samples (Figure II.1). I first explored deviation in genotypic distribution at each of nearly 2 million SNP markers from the Hardy-Weinberg equilibrium (HWE) within CEU and CHB+JPT samples separately and in mixed of the two samples by using both Pearson's chi-squared test and Fisher's exact test. To account for the multiple tests, I set the significant different level at $p < 2.5 \times 10^{-8}$ ($P_c = 0.05$ after Sidak correction). The analyses did not detect any of the SNP markers whose genotypic distribution showed significant deviation from HWE in either of the two samples. However, when all CEU and CHB+JPT samples were merged together there were approximately 3,000 markers scattered across all autosomes deviating significantly from the HWE expectation (2911

markers from Pearson's chi-squared test, consistent with 3011 markers from Fisher's exact test). These analyses show that the CEU and CHB+JPT samples can be recognized to be collected from genetically divergent random mating populations and that a mixed of them represents an example of samples from these populations. Population structure in the mixed sample was visualized as a score plot of the first two principal components built on the 2911 SNP markers, which explained a total of 62% of variability of the marker data (Figure II.2).

Figure II.1: Comparison of allele frequencies between populations for all SNP markers genotyped in the International HapMap Project.

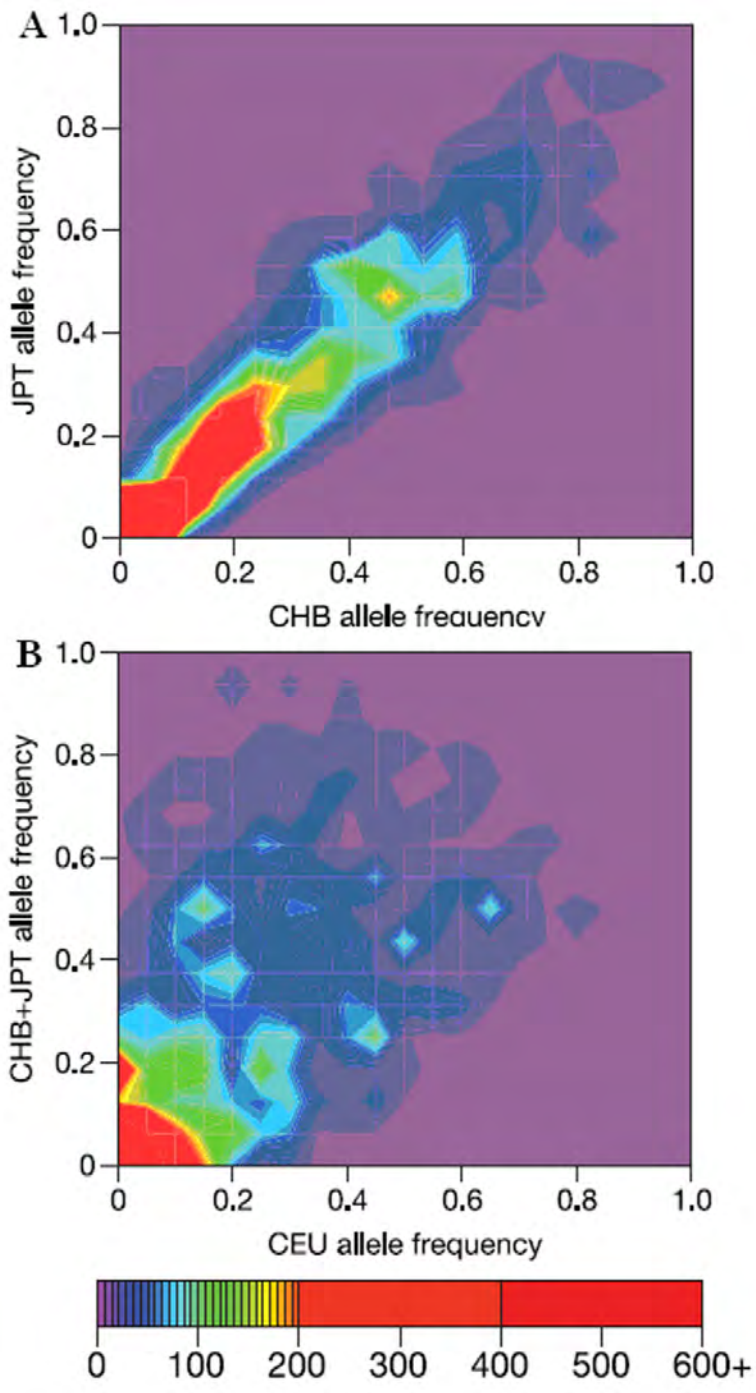
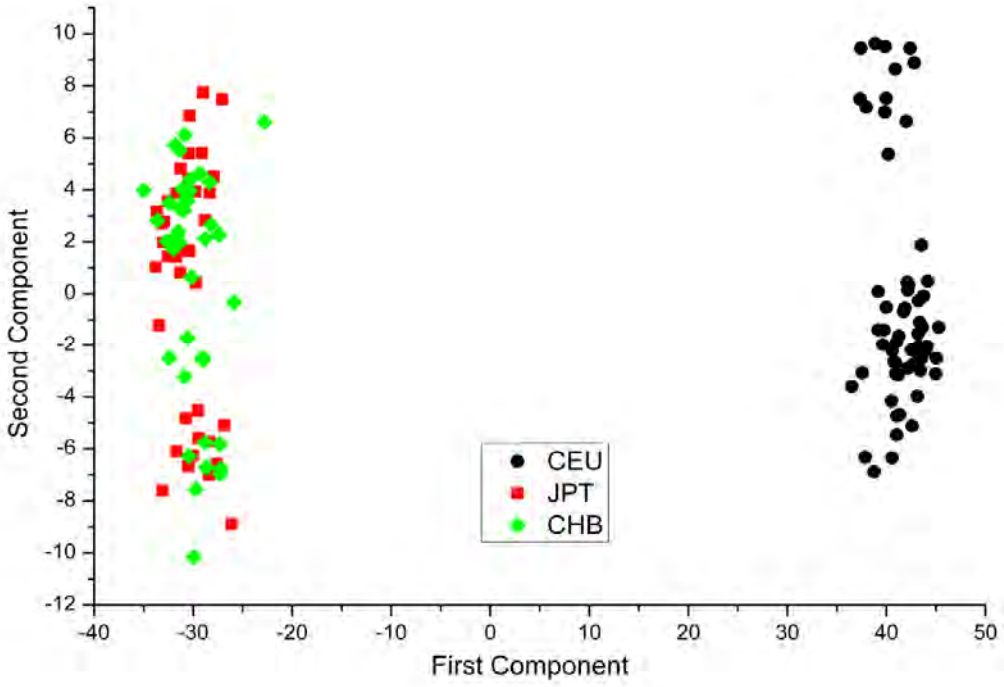


Figure II.2: The first 2 Principal Components from PCA of 142 mixed HapMap Project human samples.



The first and second principal components explained 60.77% and 1.34% of total variability respectively.

2.6.3 Genome-wide association eQTL analysis

I implemented the three methods described above to perform association mapping of expression quantitative trait loci (eQTL) using the gene expression and SNP marker datasets. The analysis was carried out on the CEU and CHB+JPT samples separately or jointly. An eQTL in the present analysis was defined as an independent peak in the p-value profile across a given chromosome. Peaks occurring within 5 Mb of adjacent peaks were taken as a single eQTL peak because of insufficient evidence to declare the existence of multiple eQTL peaks over such narrow intervals (Morley et al., 2004). The eQTL location was defined as the location within the peak with the smallest p-value. To

account for the large number of tests, I set the significance level at nominal $P < 2.5 \times 10^{-8}$ ($P_c < 0.05$ after Sidak correction), a conservative level also used previously (Spielman et al., 2007, Westfall and Young, 1993). A *cis*-regulated eQTL was operationally defined by the presence of significant association with a SNP in the region 500 kb upstream of the start of the transcript to 500 kb downstream of the 3' end; otherwise, the eQTL was classified as *trans*-acting. Table II.13 summarizes the number of eQTL detected by the three methods (**Method 1** developed in the present study, **Method 2** the simple regression analysis employed by Spielman et al in 2007, and **Method 3** the multiple regression analysis) from the Europe derived, Asia derived samples and their mixed respectively.

Table II.13: The number of eQTLs detected by three different methods (**Methods 1, 2, 3** or **M1, 2, 3 accordingly**) or detected common between two of these methods from the CEU, CHB+JPT and their mixed samples.

The number of eQTLs per expression trait	The CEU samples			The CHB+JPT samples			The mixed CEU and CHB+JPT samples				
	M1	**M2**	**M1+2**	**M1**	**M2**	**M1+2**	**M1**	**M3**	**M1+3**	**M3[a]**	**M3+3[a]**
1	280	312	**225**	263	255	**209**	206	251	**145**	398	**89**
2	58	57	**33**	43	41	**25**	16	13	**5**	136	**1**
3	20	21	**10**	13	16	**7**	2	7	**2**	97	**0**
4	10	16	**6**	8	6	**4**	2	2	**1**	72	**0**
5	4	4	**1**	5	6	**2**	0	0	**0**	48	**0**
6	3	1	**1**	1	3	**1**	0	0	**0**	37	**0**
7	3	3	**1**	0	2	**0**	0	0	**0**	22	**0**
8	0	2	**0**	1	0	**0**	1	0	**0**	22	**0**
9	2	1	**1**	0	0	**0**	0	1	**0**	14	**0**
>=10	19	22	**5**	6	7	**1**	2	2	**1**	1,111	**1**
Total eQTLs	1,009	1,149	**912**	633	670	**554**	296	354	**226**	1,975	**240**
cis-eQTLs	21	22	**21**	48	49	**48**	51	58	**51**	618	**53**
trans-eQTLs	988	1127	**891**	585	621	**506**	245	296	**175**	1,339	**187**

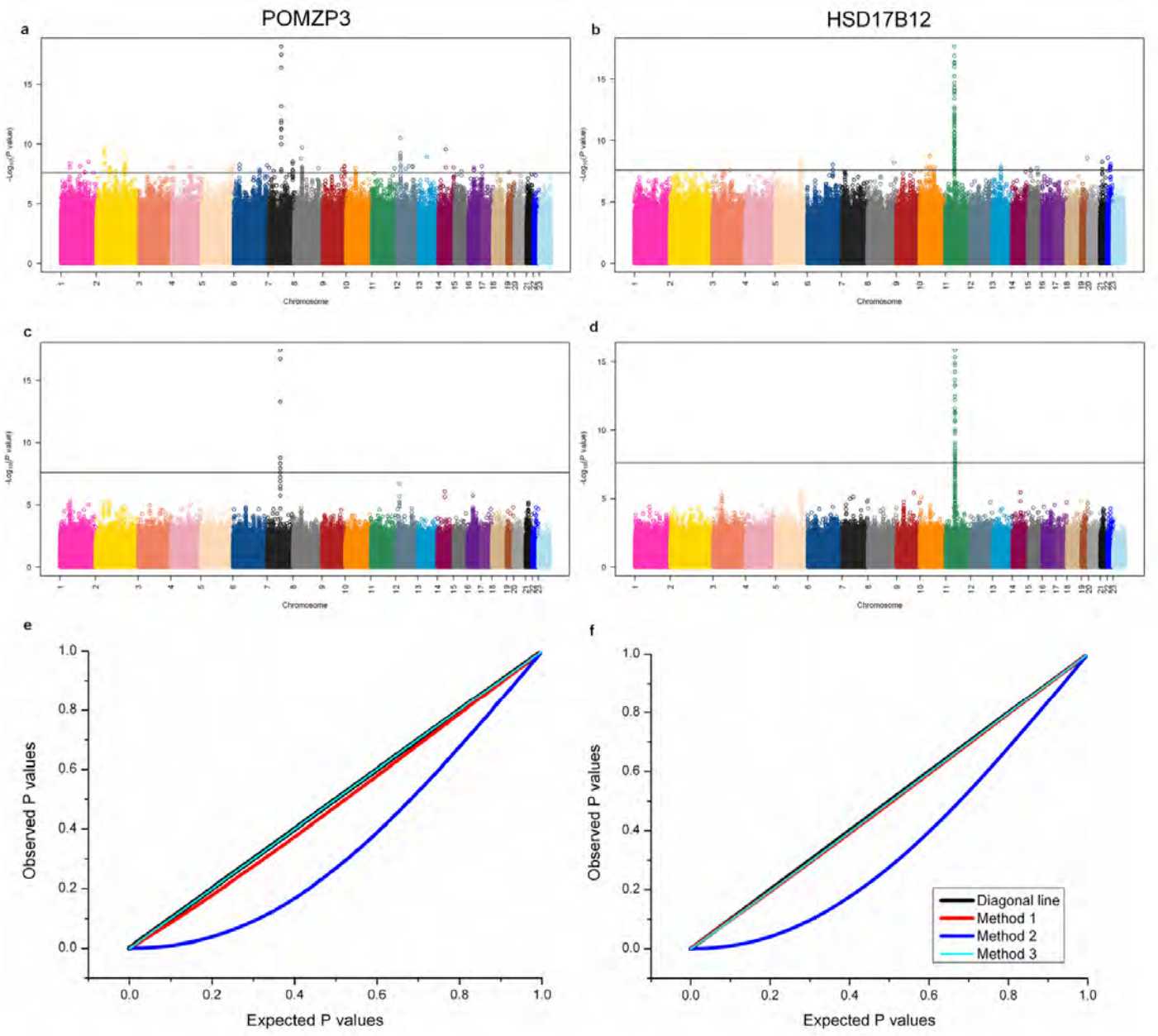
M3[a] is for Method 3 when individuals were randomly assigned to the Europe derived sample (CEU) with probability of 58% or to the Asia derived sample (CHB+JPT) otherwise.

It can be seen that the eQTL analysis results from the CEU and CHB+JPT samples are quite comparable between **Method 1** and **2** in terms of the number of detected eQTLs and their map locations**,** suggesting a comparable predictability of the two methods in the absence of population structure. In the mixed sample, 64% of eQTL detected by the multiple regression analysis (**Method 3**) with use of full population membership information can be recovered by the method developed in the present study (**Method 1**), confirming the predictability of the latter in the presence of the population structure. I explored the predictability of **Method 3** when individuals were randomly assigned to the Europe derived sample (CEU) with probability of 58% or to the Asia derived sample (CHB+JPT) otherwise. The analysis showed that only 12% (240/1,975) of eQTL detected by the method with the partial population membership information was consistent with those detected by the same method with the full membership information, suggesting that the predictability of the method depends heavily on certainty of the membership information and that the method may generate a large proportion of false positives when the information is not complete.

The POMZP3 and HSD17B12 (on the human chromosome 7 q11.23 and chromosome 11 q11.2 respectively) are two well-characterized and *cis*-regulated genes (Campino et al., 2008, Cheung et al., 2005, Morley et al., 2004, Ouyang et al., 2008, Peng et al., 2007, Spielman et al., 2007). Although all the three methods considered here were able to detect the previously identified *cis*-regulators from the three samples, there were a large number of spurious association signals predicted from the simple regression analysis (**Method 2**) with the mixed sample (Figure II.3: A and B, respectively). It is clear that these spurious associations were effectively removed in the analysis with **Method 1**, reflecting the effectiveness of the latter in controlling the false positives

(Figure II.3: C and D, respectively). In the mixed samples, **Method 1** was able to reveal 296 significant eQTL, 51 of which were *cis*-regulators (Table II.13). Firstly, the cis-eQTL predicted here include all the 11 *cis*-acting regulators reported by Spielman et al. (Spielman et al., 2007) who performed the simple regression analysis (**Method 2**) in the CEU and CHB+JPT samples separately. In addition to 16 previously detected *cis*-acting factors, **Method 1** detected 35 novel cis- eQTL and all the eQTL explained 20~70% of variability in expression of the genes regulated (Table II.14). I compared the 245 *trans*-regulators detected by the novel method from the mixed sample against the Gene Ontology (GO) Molecular Function annotation database (http://www.geneontology.org/) and found that 101 (42%) *trans*-eQTLs predicted were mapped into the category of transcriptional factors, 82 (33%) *trans*-regulators played a role in signal pathway activity. In total, 75% *trans*-regulators predicted by the present method were previously known to play a role in gene regulation. All these reveal a significantly improved statistical power of the present method in detecting the true genetic associations.

Figure II.3: Manhattan plots for the genome-wide eQTL analysis of two genes POMZP3 and HSD17B12; Quantile-quantile (QQ) plots to compare the distributions between expected and observed p-values.



Plots show score (-$\log_{10}$ p-value) for all SNPs by physical position for POMZP3 and HSD17B12 respectively based on simple linear regression (**Method 2**, a and b) and corrected linear regression (**Method 1**, c and d) in 142 mixed population samples.

Table II.14: The 51 cis-eQTLs predicted by **Method 1** from the mixed sample

NUM	Gene	cis-SNP ID	SNP position	P-value	R^2	Reference
1	**UGT2B17**	**rs3100645**	**Chr4:69806739**	**2.22E-16**	**0.38**	**[a]**
2	**POMZP3**	**rs2005354**	**Chr7:75856016**	**3.87E-28**	**0.58**	**[a] , [b]**
3	**PEX6**	**rs2395943**	**Chr6:42987528**	**3.13E-17**	**0.40**	**[a]**
4	**PSPHL**	**rs10243293**	**Chr7:55583849**	**2.25E-10**	**0.25**	**[a] , [b]**
5	**CSTB**	**rs2838386**	**Chr21:44080386**	**4.95E-12**	**0.29**	**[a] , [b]**
6	**DNAJD1**	**rs2281778**	**Chr13:41395977**	**2.64E-13**	**0.32**	**[a]**
7	**AP3S2**	**rs4932265**	**Chr15:88153061**	**1.24E-11**	**0.28**	**[a]**
8	**HSD17B12**	**rs1061810**	**Chr11:43842243**	**4.77E-16**	**0.37**	**[a], [b]**
9	**NUBP2**	**rs1065663**	**Chr16:1779024**	**2.02E-09**	**0.22**	**[a]**
10	**B4GALT1**	**rs10124479**	**Chr9:33126233**	**5.94E-10**	**0.24**	**[a]**
11	**TPP2**	**rs1887355**	**Chr13:100933170**	**1.09E-08**	**0.21**	**[a]**
12	**IRF5**	**rs12155080**	**Chr7:128212697**	**7.33E-18**	**0.41**	**[b]**
13	**CHI3L2**	**rs942694**	**Chr1:111082865**	**1.20E-09**	**0.23**	**[b]**
14	**CPNE1**	**rs12480408**	**Chr20:34950229**	**1.56E-19**	**0.44**	**[b]**
15	**CTSH**	**rs10400902**	**Chr15:76947435**	**6.90E-15**	**0.35**	**[b]**
16	**GSTM2**	**rs366631**	**Chr1:109551199**	**8.83E-26**	**0.54**	**[b]**
17	DFNA5	rs12700538	Chr7:24379328	1.35E-09	0.23	-
18	HEBP2	rs2076273	Chr6:138684210	5.80E-09	0.21	-
19	EVI2A	rs2107359	Chr17:29842786	1.29E-08	0.20	-
20	CRYZ	rs10890142	Chr1:74549064	5.40E-09	0.21	-
21	PARP4	rs7317850	Chr13:22792037	1.59E-09	0.23	-
22	RRM1	rs10767857	Chr11:4132198	4.92E-13	0.31	-
23	RPL31	rs12472882	Chr2:101267759	3.70E-23	0.50	-
24	HLA-DPB1	rs9277463	Chr6:33100194	5.46E-30	0.60	-
25	TSG101	rs1395320	Chr11:18512504	4.37E-12	0.29	-
26	DDX42	rs1043127	Chr17:62264581	4.50E-17	0.39	-
27	MEST	rs12672246	Chr7:129671295	6.04E-09	0.21	-
28	AMFR	rs2440468	Chr16:56196080	9.57E-15	0.35	-
29	GSTM3	rs1332018	Chr1:109581699	1.24E-08	0.21	-
30	ECD	rs6480700	Chr10:74446293	7.26E-10	0.24	-
31	MTRR	rs326123	Chr5:7929599	2.17E-13	0.32	-
32	RABGGTA	rs3940231	Chr14:22738491	7.23E-16	0.37	-
33	BACH1	rs733610	Chr21:29576646	2.66E-15	0.36	-
34	GSTM1	rs366631	Chr1:109551199	2.72E-27	0.57	-
35	SLC7A7	rs12884337	Chr14:21263242	1.15E-11	0.28	-
36	TAP2	rs241448	Chr6:32843645	2.11E-16	0.38	-
37	APOBEC3B	rs17000581	Chr22:37608815	2.47E-09	0.22	-

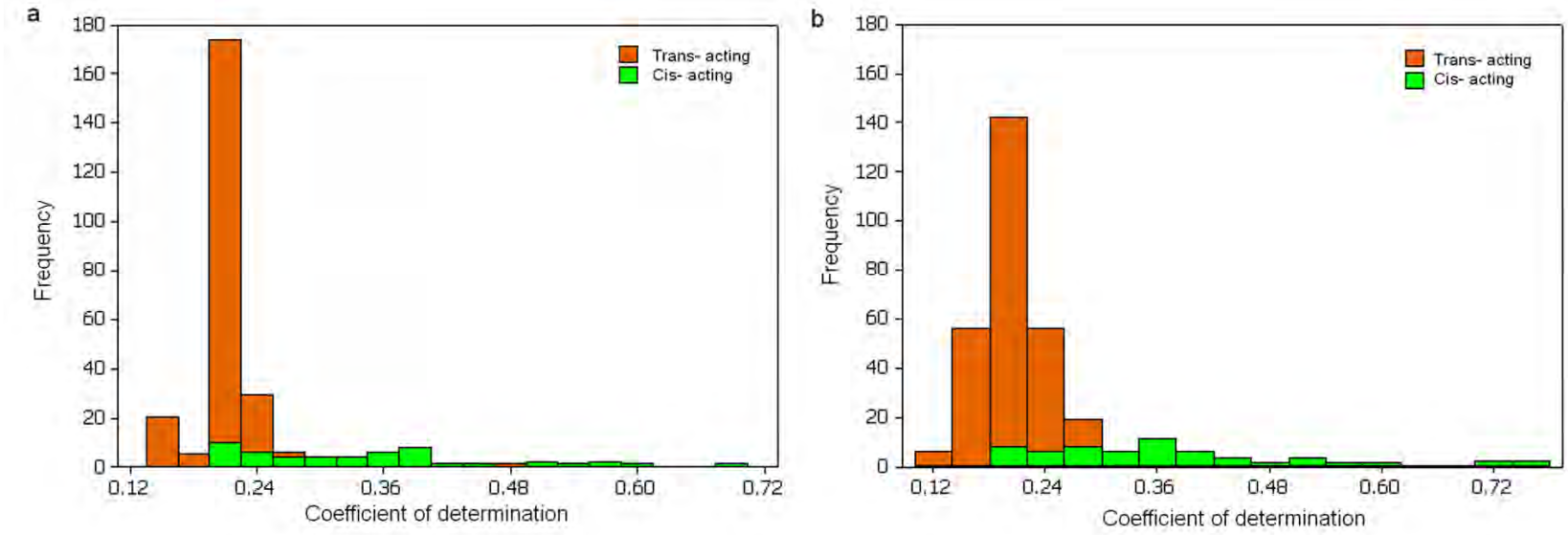
38	NT5C2	rs10883824	Chr10:104477484	4.01E-16	0.38	-
39	HBS1L	rs12663447	Chr6:135297596	2.08E-24	0.52	-
40	HLA-DQB1	rs9275141	Chr6:32697538	5.55E-15	0.35	-
41	BTN3A2	rs9366653	Chr6:26462226	1.06E-16	0.39	-
42	TLR1	rs3924112	Chr4:38692990	1.03E-13	0.32	-
43	MXRA7	rs1014390	Chr17:75310009	1.29E-13	0.32	-
44	CD47	rs6768207	Chr3:109345802	2.88E-11	0.27	-
45	GLT8D1	rs736408	Chr3:52792702	5.38E-09	0.21	-
46	TIMM13	rs3848633	Chr19:2360880	1.47E-17	0.40	-
47	POLR1D	rs9512760	Chr13:25971662	1.23E-08	0.21	-
48	SMUG1	rs3136375	Chr12:52867203	8.83E-13	0.30	-
49	POLR1E	rs10758432	Chr9:37478009	7.49E-11	0.26	-
50	ERAP2	rs2548540	Chr5:96304251	4.95E-39	0.70	-
51	IPP	rs12091503	Chr1:45521162	6.67E-17	0.39	-

a for 'Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, et al. (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics* 39: 226–231.'
b for 'Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, et al. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430: 743–747.'

It is interesting to note that the number of *cis*- eQTL detected from the mixed samples is larger than that from the component samples separately whilst a much larger number of *trans*- eQTL are detected in the component samples than in their mixed. This observation may reflect the fact that an increase in size of the mixed sample has enhanced the statistical power to detect cis- eQTL and thus led to an increased number of cis- eQTL detected. However, if linkage disequilibria between genes regulated and their trans- regulators are in opposite directions between different populations, the LD may be counter-balanced in the merged population, and thus decrease the number of the trans- eQTL to be detected. Despite a relatively small number of *cis* eQTLs detected, the *cis*-regulated effects were generally stronger than those in *trans*, with the most coefficients of determination $R^2 > 50\%$ regulators in *cis* (Figure II.4), consistent with

findings in human and mice (Dixon et al., 2007, Hubner et al., 2005, Morley et al., 2004, Schadt et al., 2003).

Figure II.4: Histograms of coefficient of determination for eQTLs from 142 mixed sample set.



a for Method 1 and b for Method 3

2.7 Discussion

Linkage disequilibrium (LD) based association mapping has been advocated as the method of choice for identifying chromosomal regions containing disease-susceptibility loci or loci affecting other complex quantitative traits of interest (Risch and Merikangas, 1996). However, it is well known that the presence of population structure can result in false positive inference of genetic association between a test marker and trait loci. Various methods have been proposed in the literature to tackle this problem (Pritchard et al., 2000a, Pritchard et al., 2000b, Satten et al., 2001, Zhu et al., 2002, Hoggart et al., 2003, Risch and Merikangas, 1996) and many of them have heavily depended on adequate prediction of the population structure (Pritchard et al., 2000a, Price et al., 2006). Efficiency of the methods is thus largely affected by adequacy of population structure prediction. It has been shown that adequate prediction of population structure is in fact not a feasible task (Yu et al., 2006). On the other hand, it is obvious that effect of the population stratification on association tests may vary across different regions of the genome (Weiss and Clark, 2002, Mackay and Powell, 2007, Remington et al., 2001). Thus, the methods designed to correct for the stratification caused spurious associations through adjusting the test statistic by subtracting a constant inflation in the statistic may not perfectly reflect this observation (Risch, 2000, Wang et al., 2005). To address these problems, I have proposed here a statistical method for correcting for stratification confounding effect in LD-based QTL mapping. The method extends the idea of using control markers to correct for background effect on a statistical test for significance of QTL at any given genome position in linkage-based QTL mapping analysis (Alexander et al., 2009) and enables the effect of population stratification in the LD-based QTL

analysis to be adjusted at a local basis (Wang et al., 2005). I presented here a simple but effective method to determine the control marker and demonstrated that incorporation of control markers would not cause any significant statistical problem even though population structure does actually not exist.

The new method developed in this study is tested and compared with other most popularly implemented methods in the literature of genetic association studies through intensive computer simulation studies and analysis of large scale and high quality gene expression and SNP datasets for mapping expression QTL. These analyses strongly support outperformance of the new method for its significantly improved statistical power to detect genuine LD between any polymorphic markers and putative trait loci and its effectiveness in controlling spurious association due to population stratification. Worthwhile, although the multiple regression analysis based on a mixed linear model does also provide a control of the influence of population stratifications, its efficiency depends heavily on accuracy of prediction of the population structure and on accurate allocation of individuals' membership to the constituent populations. Any bias in the structure prediction and uncertainty in the membership allocation may lead to severe consequence on its analytical efficiency. It has been argued that several factors may substantially influence or even disable the prediction of population structure (Zeng, 1994, Patterson et al., 2006). Therefore, the method virtually avoids the need for sophisticated prediction of population ancestry of individuals and, in turn, effectively controls any bias embedded with the prediction. The method was designed for modelling and analyzing samples collected from different ethnical (or ecological) cohorts (or populations) with or without a clear clue about their genetic diversity. This is a very popular practice in many GWAS analyses, particularly with human samples

(Spielman et al., 2007, Kang et al., 2008, Fung et al., 2006, Satake et al., 2009, Simon-Sanchez et al., 2009, Cockram et al.).

Wang et al has proposed use of a single null marker to correct for population structure in a candidate gene based association analysis using case and control samples (Wang et al., 2005). In their settings, the null marker was fitted as a dichotomous variable in parallel to the test candidate gene in a logistic regression model, and the influence of population structure on the association test at the candidate gene was adjusted by subtracting the regression coefficient associated with the null marker from the coefficient associated with the gene. Question rises to the parallel formulation: which is the major effect to be tested in the model? In contrast, the novel method was developed upon a rigorous population genetics model in which contributions of three different loci (i.e. the test marker, QTL and control marker) to the linkage disequilibrium pattern are properly formulated. The method is thus more appropriate for population based association studies. Although theoretical analysis was built on a single marker test, the idea and principle of the method could be extendable to the haplotype-based association mapping which uses information from multiple marker loci (Schaid et al., 2002, Schaid, 2004). This is because the population confounding term is linearly attached to the main disequilibrium terms in the covariance between the test polymorphism and trait effect (seeing equation II-4.6). My goal is to remove the confounding term from the covariance and, thus form of the main disequilibrium terms either in genotype at an individual marker locus or in haplotypes at multiple marker loci will not affect the way to correct for the confounding term. Although the method was presented for two genetically divergent populations, the overall pattern of LD between any test marker and trait locus in their admixed population may become theoretically more complicated

when the admixture involves more than two populations. Before having invested more theoretical investigation to the problem, I would suggest to merge those genetically less divergent objects together as I did in the present analysis with the Chinese and Japanese samples and to correct for the stratification raised from between the most divergent populations such as the European derived and the Asia derived samples.

2.8 References

ALEXANDER, D. H., NOVEMBRE, J. & LANGE, K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*.

ARDLIE, K. G., KRUGLYAK, L. & SEIELSTAD, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* 3**,** 299-309.

ASTLE, W. & BALDING, D. J. (2010) Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science,* 24**,** 7.

BACANU, S.-A., DEVLIN, B. & ROEDER, K. (2002) Association studies for quantitative traits in structured populations. *Genetic Epidemiology,* 22**,** 78-93.

BALDING, D. J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics* 7**,** 781-791.

CAMPINO, S., FORTON, J., RAJ, S., MOHR, B., AUBURN, S., FRY, A., MANGANO, V. D., VANDIEDONCK, C., RICHARDSON, A., ROCKETT, K., CLARK, T. G. & KWIATKOWSKI, D. P. (2008) Validating discovered cis-acting regulatory genetic variants: application of an Allele Specific Expression approach to HapMap populations. *PLoS One* 3**,** e4105.

CARDON, L. R. & BELL, J. I. (2001) Association study designs for complex diseases. *Nature Review Genetics,* 2**,** 91-99.

CARDON, L. R. & PALMER, L. J. (2003) Population stratification and spurious allelic association. *Lancet,* 361**,** 598-604.

CHAKRABORTY, R. & SMOUSE, P. E. (1988) Recombination of haplotypes leads to biased estimates of admixture proportions in human populations. *Proceedings of the National Academy of Sciences,* 85**,** 3071-3074.

CHEUNG, V. G., CONLIN, L. K., WEBER, T. M., ARCARO, M., JEN, K. Y., MORLEY, M. & SPIELMAN, R. S. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics,* 33**,** 422 - 425.

CHEUNG, V. G., SPIELMAN, R. S., EWENS, K. G., WEBER, T. M., MORLEY, M. & BURDICK, J. T. (2005) Mapping determinants of human gene expression by regional and genome-wide association. *Nature,* 437**,** 1365–1369.

COCKRAM, J., WHITE, J., ZULUAGA, D. L., SMITH, D., COMADRAN, J., MACAULAY, M., LUO, Z., KEARSEY, M. J., WERNER, P., HARRAP, D., TAPSELL, C., LIU, H., HEDLEY, P. E., STEIN, N., SCHULTE, D., STEUERNAGEL, B., MARSHALL, D. F., THOMAS, W. T. B., RAMSAY, L., MACKAY, I., BALDING, D. J., CONSORTIUM, T. A., WAUGH, R. & O'SULLIVAN, D. M. Genome-wide association mapping to candidate polymorphism resolution in the unsequenced barley genome. *Proceedings of the National Academy of Sciences*.

COUZIN, J. & KAISER, J. (2007) Genome-wide association: closing the net on common disease denes. *Science,* 316**,** 820-822.

DEVLIN, B. & ROEDER, K. (1999) Genomic control for association studies. *Biometrics* 55**,** 997-1004.

DIXON, A. L., LIANG, L., MOFFATT, M. F., CHEN, W., HEATH, S., WONG, K. C. C., TAYLOR, J., BURNETT, E., GUT, I., FARRALL, M., LATHROP, G. M., ABECASIS, G. R. & COOKSON, W. O. C. (2007) A genome-wide association study of global gene expression. *Nature Genetics,* 39**,** 1202 - 1207.

EWENS, W. J. & SPIELMAN, R. S. (1995) The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 57**,** 455-464

FUNG, H.-C., SCHOLZ, S., MATARIN, M., SIM -S CHEZ, J., HERNANDEZ, D., BRITTON, A., GIBBS, J. R., LANGEFELD, C., STIEGERT, M. L., SCHYMICK, J., OKUN, M. S., MANDEL, R. J., FERNANDEZ, H. H., FOOTE, K. D., RODR UEZ, R. L., PECKHAM, E., DE VRIEZE, F. W., GWINN-HARDY, K., HARDY, J. A. & SINGLETON, A. (2006) Genome-wide genotyping in Parkinson's disease and neurologically normal controls: first stage analysis and public release of data. *The Lancet Neurology,* 5**,** 911-916.

HOGGART, C. J., PARRA, E. J., SHRIVER, M. D., BONILLA, C., KITTLES, R. A., CLAYTON, D. G. & MCKEIGUE, P. M. (2003) Control of Confounding of Genetic Associations in Stratified Populations. *Am J Hum Genet* 72**,** 1492-1504.

HUBNER, N., WALLACE, C. A., ZIMDAHL, H., PETRETTO, E., SCHULZ, H., MACIVER, F., MUELLER, M., HUMMEL, O., MONTI, J., ZIDEK, V., MUSILOVA, A., KREN, V., CAUSTON, H., GAME, L., BORN, G., SCHMIDT, S., MÜLLER, A., COOK, S. A., KURTZ, T. W., WHITTAKER, J., PRAVENEC, M. & AITMAN, T. J. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nature Genetics,* 37**,** 243 - 253.

ILES, M. M. (2008) What can genome-wide association studies tell Us about the genetics of common disease. *PLoS Genetics* 4**,** e33.

JOHNSON, N. L. & KOTZ, S. (1970) *Distributions in statistics: continuous univariate distributions,* Boston, Houghton Mifflin.

KANG, H. M., ZAITLEN, N. A., WADE, C. M., KIRBY, A., HECKERMAN, D., DALY, M. J. & ESKIN, E. (2008) Efficient Control of Population Structure in Model Organism Association Mapping. *Genetics,* 178**,** 1709-1723.

LANDER, E. S. & SCHORK, N. J. (1994) Genetic dissection of complex traits. *Science,* 265**,** 2037-2048.

LUO, Z. W. (1998) Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. *Heredity,* 80**,** 198-208.

MACKAY, I. & POWELL, W. (2007) Methods for linkage disequilibrium mapping in crops. *Trends in Plant Science,* 12**,** 57-63.

MCCARTHY, M. I., ABECASIS, G. R., CARDON, L. R., GOLDSTEIN, D. B., LITTLE, J., IOANNIDIS, J. P. A. & HIRSCHHORN, J. N. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics,* 9**,** 356-369.

MCGINNIS, R., INFORMATION, C., SHIFMAN, S. & DARVASI, A. (2002) Power and efficiency of the TDT and case-control design for association scans. *Behavior Genetics* 32**,** 135-144.

MORLEY, M., MOLONY, C. M., WEBER, T. M., DEVLIN, J. L., EWENS, K. G., SPIELMAN, R. S. & CHEUNG, V. G. (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430**,** 743–747.

OUYANG, C., SMITH, D. D. & KRONTIRIS, T. G. (2008) Evolutionary signatures of common human cis-regulatory haplotypes. *PLoS One* 3**,** e3362.

PATTERSON, N., PRICE, A. L. & REICH, D. (2006) Population Structure and Eigenanalysis. *PLoS Genet,* 2**,** e190.

PENG, J., WANG, P. & TANG, H. (2007) Controlling for false positive findings of trans-hubs in expression quantitative trait loci mapping. *BMC Proceedings,* 1**,** S157.

PRICE, A. L., PATTERSON, N. J., PLENGE, R. M., WEINBLATT, M. E., SHADICK, N. A. & REICH, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics,* 38**,** 904 - 909.

PRITCHARD, J. K., STEPHENS, M. & DONNELLY, P. (2000a) Inference of population structure using multilocus genotype data. *Genetics,* 155**,** 945-959.

PRITCHARD, J. K., STEPHENS, M., ROSENBERG, N. A. & DONNELLY, P. (2000b) Association Mapping in Structured Populations. *American Journal of Human Genetics* 67**,** 170-181.

REMINGTON, D. L., THORNSBERRY, J. M., MATSUOKA, Y., WILSON, L. M., WHITT, S. R., DOEBLEY, J., KRESOVICH, S., GOODMAN, M. M. & BUCKLER, E. S. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America* 98**,** 11479-11484.

RISCH, N. J. (2000) Searching for genetic determinants in the new millennium. *Nature* 405**,** 847-856

RISCH, N. J. & MERIKANGAS, K. (1996) The future of genetic studies of complex human diseases. *Science* 273**,** 1516–1517.

SATAKE, W., NAKABAYASHI, Y., MIZUTA, I., HIROTA, Y., ITO, C., KUBO, M., KAWAGUCHI, T., TSUNODA, T., WATANABE, M., TAKEDA, A., TOMIYAMA, H., NAKASHIMA, K., HASEGAWA, K., OBATA, F., YOSHIKAWA, T., KAWAKAMI, H., SAKODA, S., YAMAMOTO, M., HATTORI, N., MURATA, M., NAKAMURA, Y. & TODA, T. (2009) Genome-wide association study identifies common variants at four loci as genetic risk factors for Parkinson's disease. *Nat Genet,* 41**,** 1303-1307.

SATTEN, G. A., FLANDERS, W. D. & YANG, Q. H. (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *American Journal of Human Genetics* 68**,** 466-477.

SCHADT, E. E., MONKS, S. A., DRAKE, T. A., LUSIS, A. J., CHE, N., COLINAYO, V., RUFF, T. G., MILLIGAN, S. B., LAMB, J. R., CAVET, G., LINSLEY, P. S., MAO, M., STOUGHTON, R. B. & FRIEND, S. H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422**,** 297–302.

SCHAID, D. J. (2004) Evaluating associations of haplotypes with traits. *Genetic Epidemiology,* 27**,** 348-364.

SCHAID, D. J., ROWLAND, C. M., TINES, D. E., JACOBSON, R. M. & POLAND, G. A. (2002) Score Tests for Association between Traits and Haplotypes when Linkage Phase Is Ambiguous. *The American Journal of Human Genetics,* 70**,** 425-434.

SIMON-SANCHEZ, J., SCHULTE, C., BRAS, J. M., SHARMA, M., GIBBS, J. R., BERG, D., PAISAN-RUIZ, C., LICHTNER, P., SCHOLZ, S. W., HERNANDEZ, D. G., KRUGER, R., FEDEROFF, M., KLEIN, C., GOATE, A., PERLMUTTER, J., BONIN, M., NALLS, M. A., ILLIG, T., GIEGER, C., HOULDEN, H., STEFFENS, M., OKUN, M. S., RACETTE, B. A., COOKSON, M. R., FOOTE, K. D., FERNANDEZ, H. H., TRAYNOR, B. J., SCHREIBER, S., AREPALLI, S., ZONOZI, R., GWINN, K., VAN DER BRUG, M., LOPEZ,

G., CHANOCK, S. J., SCHATZKIN, A., PARK, Y., HOLLENBECK, A., GAO, J., HUANG, X., WOOD, N. W., LORENZ, D., DEUSCHL, G., CHEN, H., RIESS, O., HARDY, J. A., SINGLETON, A. B. & GASSER, T. (2009) Genome-wide association study reveals genetic risk underlying Parkinson's disease. *Nat Genet,* 41**,** 1308-1312.

SLATKIN, M. (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics* 9**,** 477-485.

SNEDECOR, G. W. & COCHRAN, W. G. (1967) *Statistical methods*, The Iowa State University.

SPIELMAN, R. S., BASTONE, L. A., BURDICK, J. T., MORLEY, M., EWENS, W. J. & CHEUNG, V. G. (2007) Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics,* 39**,** 226–231.

SPIELMAN, R. S., MCGINNIS, R. E. & EWENS, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* 52**,** 506-516.

The International HapMap Consortium (2003) The International HapMap Project. *Nature* 426**,** 789-796.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature,* 437**,** 1299-1320.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449**,** 851-862.

WANG, Y. T., LOCALIO, R. & REBBECK, T. R. (2005) Bias correction with a single null marker for population stratification in candidate gene association studies. *Human Heredity* 59**,** 165-175.

WEISS, K. M. & CLARK, A. G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends in Genetics* 18**,** 19-24.

WESTFALL, P. H. & YOUNG, S. S. (1993) *Resampling-based Multiple Testing,* New York, Wiley.

YU, J., PRESSOIR, G., BRIGGS, W. H., BI, I. V., YAMASAKI, M., DOEBLEY, J. F., MCMULLEN, M. D., GAUT, B. S., NIELSEN, D. M., HOLLAND, J. B., KRESOVICH, S. & BUCKLER, E. S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics,* 38**,** 203–208.

ZENG, Z. B. (1994) Precision Mapping of Quantitative Trait Loci. *Genetics,* 136**,** 1457-1468.

ZHU, X., ZHANG, S. L., ZHAO, H. & COOPER, R. S. (2002) Association mapping, using a mixture model for complex traits. *Genetic Epidemiology* 23**,** 181-196.

Part II

Comparative evolutionary epigenetic regulation of gene transcription

Chapter III

General introduction: regulation of gene transcription

3.1 Overview

Transcription is the first essential procedure leading to gene expression. In this step, DNA sequence of a particular gene is copied to create a RNA molecule. The transcription process for gene expression is common to all known organisms, including viruses, prokaryotes and eukaryotes. However, all genes in the genome are not equally transcribed (or expressed). Only a few genes in the genome are transcribed in all tissues and at all of the time. For example, a typical human cell has only about 3% to 5% of its total genes transcribed at any time, which are considered as 'house-keeping' genes to maintain the basic biological process for life (Hsiao et al., 2001). Meanwhile, most genes in the genome are differentially transcribed in different tissues (or cells) and at different development stages. The different transcription levels can be modulated by several regulation mechanisms, such as transcription factor, microRNA, DNA methylation and histone protein binding. In genetics, the regulation of gene transcription can drive organisms to produce the RNA in particular tissues (or cells), at particular times and even at particular abundances; in order to make the organisms to flexibly adapt different stages of development, variable environments and external

signals. Thus, understanding the transcriptional regulation mechanism will be important for both biological and medical research.

3.2 Regulation of gene transcription

As mentioned above, all of genes in the whole genome are not uniformly transcribed and expressed in all cells (or tissues) and at any times. The transcription process must be controlled to allow the cell to produce some particular transcriptions when and where they are needed, and even how much the particular transcription products are needed (Moore, 2005, Chen and Rajewsky, 2007, Levine and Tjian, 2003). Thus, the regulation of gene transcription is vital to give organisms the ability to drive the cell differentiation and morphogenesis processes, to adjust to environmental changes, to response to both internal and external signals, etc (Cornell et al., 2010). Moreover, regulation of transcription may also have the function of a substrate for evolutionary change, since modulation of the location, timing and abundance of gene transcriptions can have an esoteric effect on the roles (or functions) of the gene in different cells or tissues, even in different species (Brawand et al., 2011, Khaitovich et al., 2006, Robertson, 2010, Tuch et al., 2008).

The regulation can take place at two possible stages during synthesizing of the matured RNA (Fenton, 1992). First, any step of gene transcription can be regulated, from pre-initiation to RNA processing. Second, the transcription can be regulated after they are completely produced via post-transcription regulation mechanisms (Halbeisen et al., 2008). For the post-transcription regulation, the stableness of the transcription products mainly participates in the regulation of transcription: the unstable RNA molecules can

be easily degraded; hence it results in low transcription abundances. Briefly, the regulation of transcription is a complicated process, which simultaneously combines multiple mechanisms to dynamically control the transcription of gene. According to the sources of influence, regulation of gene transcription can be divided into three main routes: a) by regulatory proteins, such as transcription factors, repressors and activators; b) by non-coding RNA molecules, including microRNAs, small interfering RNAs (siRNAs) and long non-coding RNAs (lncRNAs); 3) by epigenetic regulations, for example, the structure changes of chromatins and DNA methylation.

3.2.1 Regulatory proteins

Regulatory proteins are a group of proteins that involve in regulating gene transcription process. Generally, the typical regulatory proteins must be bound to particular DNA binding sites to touch off the up or down regulation functions (Johnson and McKnight, 1989). The DNA binding sites are usually located at the promoter region or around the transcription start site (TSS), and can be divided into multiple groups, such as enhancers, operators, insulators and silencers (Berg et al., 1982). The mechanisms of regulatory proteins may also vary, from prohibiting the RNA polymerases binding to core promoters, to encouraging the transcription as activators. Briefly, the regulatory proteins can be divided into at least 4 different mechanisms: 1) general transcription factors, 2) specificity factors, 3) repressors and 4) activators.

1) General Transcription Factors (GTFs), also referred as the basal transcription factors, are a group of proteins which can bind to the specific sites on DNA sequence to initially switch on the transcription process (Orphanides et al., 1996, Lee and Young, 2000). Generally, the GTFs guide and place the

appropriate RNA polymerase at the start site of a coding sequence and then conduct the polymerase to transcribe it into RNA. In bacteria, it has only one general transcription factor, called as sigma factor. In eukaryotes, transcription initiation involves several GTFs, including TFIIA, TFIIB, TFIID, TFIIE and TFIIF (Latchman, 1997).

2) Specificity Factors are a set of proteins which can only bind to a few specific promoters, and make these promoters more or less likely to bind with RNA polymerase (Mitchell and Tjian, 1989).
3) Repressors are a set of proteins that can attach to the operator regions (Herschbach and Johnson, 1993, Rojo, 2001). The operator is a segment of coding sequences which are overlapped with or closed to the promoter region. By binding to the operator, the repressor can physically block the RNA polymerase to bind to the promoter region, and further prevent the transcription of the genes.
4) Activators are also typical DNA-binding proteins, which can obviously encourage the transcription levels of the targeted genes (Busby and Ebright, 1999). The activators are functioned by binding to a specific site of DNA sequence (activator site) located at or very near a promoter and making interactions with the subunits of RNA polymerase.

3.2.2 RNA based transcriptional regulation

Furthermore, the RNA based transcriptional regulation is also widespread in most eukaryotes (Barrandon et al., 2008, Kurokawa et al., 2009). Recently, the biological functions of non-coding RNA molecule in modulating gene transcription have become a

topic of intense interests (Mattick, 2009). A typical RNA based mechanism is often through non-coding RNAs as media to control the transcriptions, including microRNA, small interference RNA and long non-coding RNA.

1) microRNA are a set of short (about 22 nucleotides) non-coding RNA molecules which can bind to their complementary sequences in the 3' end of target mRNAs, and repress the transcription synthesize (Bartel, 2004, Bartel and Chen, 2004). microRNA was first identified during research on development in *C. elegans* regarding the Lin-14 gene (Lee et al., 1993). This study found that the transcription level of Lin-14 gene could be regulated by presence of a short RNA sequence that contains only 22 nucleotides and is completely complementary to the 3' UTR of target gene. After that around 20,000 microRNAs have been identified in over 168 eukaryotic species including human, mouse, rat and Arabidopsis (Lagos-Quintana et al., 2003, Lim et al., 2003). Furthermore, these microRNAs can bind to more than 60% of all annotated genes and result in the negative regulation of the transcription level for target genes (Wightman et al., 1993, Enright, 2003, Ambros, 2004, Lewis et al., 2005, Friedman et al., 2009). It has demonstrated that microRNAs involve in several mechanisms of the post-transcriptional regulation, including sequestering, transcript degradation and translational suppressing (Ambros, 2004, Bartel, 2004, Bartel and Chen, 2004, Zeng et al., 2003). Thus, microRNAs are an essential part of transcription regulation system and can further involve into most biological developments and processes, such as cell cycle, proliferation, metabolism, development and

organogenesis (Brennecke et al., 2003, Cuellar and McManus, 2005, Poy et al., 2004).

2) Small interfering RNAs (siRNAs), also termed as short interfering RNAs or silencing RNAs, are another set of post-transcriptional regulators, which are double-stranded RNA sequence in 20~25 nucleotides length (Hutvagner et al., 2004, Zeng et al., 2003, Doench et al., 2003). siRNAs and their biological function in gene transcriptional regulation has been first found by Hamilton and Baulcombe in 1999. Their research revealed that siRNA could interfere with the transcription synthesize of specific genes containing complementary sequence.

3) Long non-coding RNAs (lncRNAs) are usually defined as non-protein coding RNA sequences that longer than 200 bps (Mercer et al., 2009). During last 10 years, it is found that the lncRNAs played many roles in gene transcription regulation (Rinn and Chang, 2012). At first, lncRNAs can impede the transcription factors binding to the targeted genes, resulting in inducing transcription initiation (Kwek et al., 2002) or repressing transcription elongation (Yang et al., 2001, Yik et al., 2003). In addition, lncRNAs also play multiple roles in post-transcriptional regulation. Similar to microRNAs and siRNAs, they can bind to the complementary mRNAs. The structure of RNA duplexes between lncRNAs and their complementary target RNA products are reported to shroud the essential regions in the mRNAs required to bind trans-acting transcription factors, potentially resulting in post-transcriptional regulations (Chodroff et al., 2011).

3.3.3 Epigenetic regulation

More recently, it has become a common consensus that there is a considerable influence of non-nucleotide-sequence change effects on gene expression, which are termed as 'epigenetic regulation' (Attwood et al., 2002, Mattick et al., 2009, Kurokawa et al., 2009, Jaenisch and Bird, 2003). The Human Epigenome Project has stated that 'Epigenetics is an emerging frontier of science of the study of changes in the regulation of gene activity and expression that are not dependent on gene sequence' (The Human Epigenome Project, 2004). The epigenetics mainly contains two types of chemical markers: DNA methylation and histone modification. For many years, the DNA methylation has been considered to play a crucial role in epigenetic influencing gene expression and widely spread in most eukaryotic species, particular in vertebrates (He et al., 2011, Wu et al., 2011, Poetsch and Plass, 2011). But, the extent of function and mechanism of DNA methylation in gene expression is still unknown. To better uncover many of the unknown of DNA methylation, I comprehensively analyzed the distribution patterns and properties of DNA methylation system across several eukaryotic model species and multiple tissues, attempted to explore the roles of methylation in gene transcriptional regulation. The analyses are fully detailed in chapter IV.

3.4 Reference

AMBROS, V. (2004) The functions of animal microRNAs. *Nature,* 431**,** 350-355.

ATTWOOD, J. T., YUNG, R. L. & RICHARDSON, B. C. (2002) DNA methylation and the regulation of gene transcription. *Cellular and Molecular Life Sciences,* 59**,** 241-257.

BARRANDON, C., SPILUTTINI, B. & BENSAUDE, O. (2008) Non-coding RNAs regulating the transcriptional machinery. *Biology of the Cell,* 100**,** 83-95.

BARTEL, D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* 116**,** 281-297.

BARTEL, D. P. & CHEN, C. Z. (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nature Rev. Genet.,* 5**,** 396-401.

BERG, O. G., WINTER, R. B. & VON HIPPEL, P. H. (1982) How do genome-regulatory proteins locate their DNA target sites? *Trends in Biochemical Sciences,* 7**,** 52-55.

BRAWAND, D., SOUMILLON, M., NECSULEA, A., JULIEN, P., CSARDI, G., HARRIGAN, P., WEIER, M., LIECHTI, A., AXIMU-PETRI, A., KIRCHER, M., ALBERT, F. W., ZELLER, U., KHAITOVICH, P., GRUTZNER, F., BERGMANN, S., NIELSEN, R., PAABO, S. & KAESSMANN, H. (2011) The evolution of gene expression levels in mammalian organs. *Nature,* 478**,** 343-348.

BRENNECKE, J., HIPFNER, D. R., STARK, A., RUSSELL, R. B. & COHEN, S. M. (2003) bantam Encodes a Developmentally Regulated microRNA that Controls Cell Proliferation and Regulates the Proapoptotic Gene hid in Drosophila. *Cell,* 113**,** 25-36.

BUSBY, S. & EBRIGHT, R. H. (1999) Transcription activation by catabolite activator protein (CAP). *Journal of Molecular Biology,* 293**,** 199-213.

CHEN, K. & RAJEWSKY, N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nat Rev Genet,* 8**,** 93-103.

CHODROFF, R., GOODSTADT, L., SIREY, T., OLIVER, P., DAVIES, K., GREEN, E., MOLNAR, Z. & PONTING, C. (2011) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biology,* 11**,** R72.

CORNELL, T. T., WYNN, J., SHANLEY, T. P., WHEELER, D. S. & WONG, H. R. (2010) Mechanisms and Regulation of the Gene-Expression Response to Sepsis. *Pediatrics,* 125**,** 1248-1258.

CUELLAR, T. L. & MCMANUS, M. T. (2005) MicroRNAs and endocrine biology. *Journal of Endocrinology,* 187**,** 327-332.

DOENCH, J. G., PETERSON, C. P. & SHARP, P. A. (2003) siRNAs can function as miRNAs. *Genes Dev.,* 17**,** 438-442.

ENRIGHT, A. J. (2003) MicroRNA targets in Drosophila. *Genome Biol.,* 5**,** R1-R1.

FENTON, M. J. (1992) Review: Transcriptional and post-transcriptional regulation of interleukin 1 gene expression. *International Journal of Immunopharmacology,* 14**,** 401-411.

FRIEDMAN, R. C., FARH, K. K.-H., BURGE, C. B. & BARTEL, D. P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research,* 19**,** 92-105.

HALBEISEN, R., GALGANO, A., SCHERRER, T. & GERBER, A. (2008) Post-transcriptional gene regulation: From genome-wide studies to principles. *Cellular and Molecular Life Sciences,* 65**,** 798-813.

HE, X.-J., CHEN, T. & ZHU, J.-K. (2011) Regulation and function of DNA methylation in plants and animals. *Cell Res,* 21**,** 442-465.

HERSCHBACH, B. M. & JOHNSON, A. D. (1993) Transcriptional Repression in Eukaryotes. *Annual Review of Cell Biology,* 9**,** 479-509.

HSIAO, L.-L., DANGOND, F., YOSHIDA, T., HONG, R., JENSEN, R. V., MISRA, J., DILLON, W., LEE, K. F., CLARK, K. E., HAVERTY, P., WENG, Z., MUTTER, G. L., FROSCH, M. P., MACDONALD, M. E., MILFORD, E. L., CRUM, C. P., BUENO, R., PRATT, R. E., MAHADEVAPPA, M., WARRINGTON, J. A., STEPHANOPOULOS, G., STEPHANOPOULOS, G. & GULLANS, S. R. (2001) A compendium of gene expression in normal human tissues. *Physiological Genomics,* 7**,** 97-104.

HUTVAGNER, G., SIMARD, M. J., MELLO, C. C. & ZAMORE, P. D. (2004) Sequence-specific inhibition of small RNA function. *PLoS Biol.,* 2**,** E98-E98.

JAENISCH, R. & BIRD, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet,* 33**,** 245-254.

JOHNSON, P. F. & MCKNIGHT, S. L. (1989) Eukaryotic Transcriptional Regulatory Proteins. *Annual Review of Biochemistry,* 58**,** 799-839.

KHAITOVICH, P., ENARD, W., LACHMANN, M. & PAABO, S. (2006) Evolution of primate gene expression. *Nat Rev Genet,* 7**,** 693-702.

KUROKAWA, R., ROSENFELD, M. G. & GLASS, C. K. (2009) Transcriptional regulation through noncoding RNAs and epigenetic modifications. *RNA Biology,* 6**,** 233-236.

KWEK, K. Y., MURPHY, S., FURGER, A., THOMAS, B., O'GORMAN, W., KIMURA, H., PROUDFOOT, N. J. & AKOULITCHEV, A. (2002) U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat Struct Mol Biol,* 9**,** 800-805.

LAGOS-QUINTANA, M., RAUHUT, R., MEYER, J., BORKHARDT, A. & TUSCHL, T. (2003) New microRNAs from mouse and human. *RNA,* 9**,** 175-179.

LATCHMAN, D. S. (1997) Transcription factors: An overview. *The International Journal of Biochemistry & Cell Biology,* 29**,** 1305-1312.

LEE, R. C., FEINBAUM, R. L. & AMBROS, V. (1993) The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell,* 75**,** 843-854.

LEE, T. I. & YOUNG, R. A. (2000) TRANSCRIPTION OF EUKARYOTIC PROTEIN-CODING GENES. *Annual Review of Genetics,* 34**,** 77-137.

LEVINE, M. & TJIAN, R. (2003) Transcription regulation and animal diversity. *Nature,* 424**,** 147-151.

LEWIS, B. P., BURGE, C. B. & BARTEL, D. P. (2005) Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell,* 120**,** 15-20.

LIM, L. P., GLASNER, M. E., YEKTA, S., BURGE, C. B. & BARTEL, D. P. (2003) Vertebrate microRNA genes. *Science,* 299**,** 1540-1540.

MATTICK, J. S. (2009) The Genetic Signatures of Noncoding RNAs. *PLoS Genet,* 5**,** e1000459.
MATTICK, J. S., AMARAL, P. P., DINGER, M. E., MERCER, T. R. & MEHLER, M. F. (2009) RNA regulation of epigenetic processes. *BioEssays,* 31**,** 51-59.
MERCER, T. R., DINGER, M. E. & MATTICK, J. S. (2009) Long non-coding RNAs: insights into functions. *Nat Rev Genet,* 10**,** 155-159.
MITCHELL, P. J. & TJIAN, R. (1989) Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. *Science,* 245**,** 371-378.
MOORE, M. J. (2005) From Birth to Death: The Complex Lives of Eukaryotic mRNAs. *Science,* 309**,** 1514-1518.
ORPHANIDES, G., LAGRANGE, T. & REINBERG, D. (1996) The general transcription factors of RNA polymerase II. *Genes & Development,* 10**,** 2657-2683.
POETSCH, A. R. & PLASS, C. (2011) Transcriptional regulation by DNA methylation. *Cancer Treatment Reviews,* 37, Supplement 1**,** S8-S12.
POY, M. N., ELIASSON, L., KRUTZFELDT, J., KUWAJIMA, S., MA, X., MACDONALD, P. E., PFEFFER, S., TUSCHL, T., RAJEWSKY, N., RORSMAN, P. & STOFFEL, M. (2004) A pancreatic islet-specific microRNA regulates insulin secretion. *Nature,* 432**,** 226-230.
RINN, J. L. & CHANG, H. Y. (2012) Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry,* 81**,** 145-166.
ROBERTSON, M. (2010) The evolution of gene regulation, the RNA universe, and the vexed questions of artefact and noise. *BMC Biology,* 8**,** 97.
ROJO, F. (2001) Mechanisms of transcriptional repression. *Current Opinion in Microbiology,* 4**,** 145-151.
TUCH, B. B., LI, H. & JOHNSON, A. D. (2008) Evolution of eukaryotic transcription circuits. *Science,* 319**,** 1797 - 1799.
WIGHTMAN, B., HA, I. & RUVKUN, G. (1993) Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell,* 75**,** 855-862.
WU, H., TAO, J. & SUN, Y. E. (2011) Regulation and function of mammalian DNA methylation patterns: a genomic perspective. *Briefings in Functional Genomics*.
YANG, S., TUTTON, S., PIERCE, E. & YOON, K. (2001) Specific Double-Stranded RNA Interference in Undifferentiated Mouse Embryonic Stem Cells. *Molecular and Cellular Biology,* 21**,** 7807-7816.
YIK, J. H. N., CHEN, R., NISHIMURA, R., JENNINGS, J. L., LINK, A. J. & ZHOU, Q. (2003) Inhibition of P-TEFb (CDK9/Cyclin T) Kinase and RNA Polymerase II Transcription by the Coordinated Actions of HEXIM1 and 7SK snRNA. *Molecular Cell,* 12**,** 971-982.
ZENG, Y., YI, R. & CULLEN, B. R. (2003) MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc. Natl Acad. Sci. USA,* 100**,** 9779-9784.

Chapter IV:

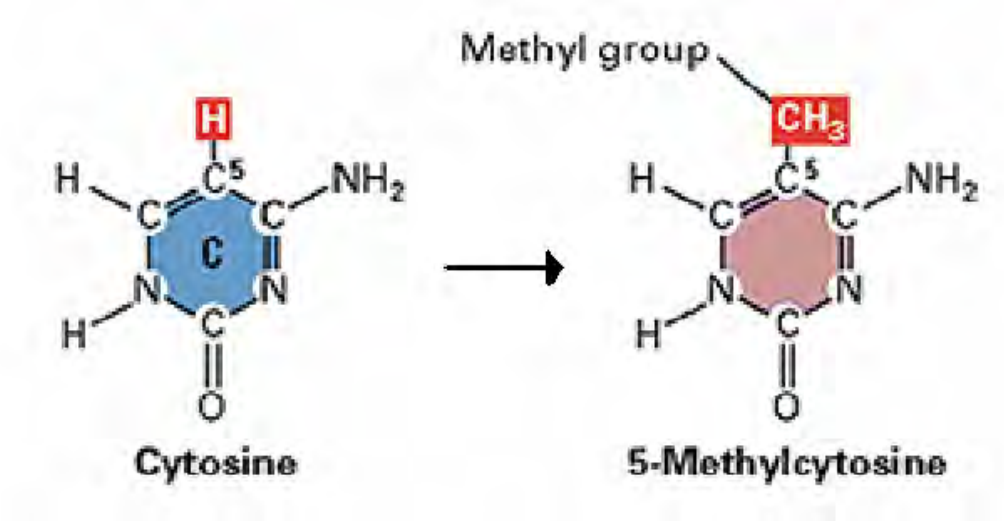
Comparative evolutionary epigenetic regulation of gene transcription

4.1 Introduction to DNA methylation based transcriptional regulation

In epigenetics, the post-replicative addition of methyl groups to the 5-position of the cytosine pyrimidine rings in the DNA sequences, termed as cytosine 'DNA methylation', has long been recognized as heritable chemical modification (Holliday and Pugh, 1975, Riggs, 1975, Day and Sweatt, 2010, Zhu and Reinberg, 2011, Parle-Mcdermott and Harrison, 2011) (figure IV.1). The DNA methylation is highly conserved in most eukaryotic species, including protists, fungi, plants and animals, playing a fundamental role in modulating the biological processes, particularly regulation of transcription (He et al., 2011, Chen and Riggs, 2011, Jaenisch and Bird, 2003, Patra et al., 2008). In previous studies, a prevalent view holds that DNA methylation regulates the gene transcription through two different ways (Geiman and Robertson, 2002, Herman and Baylin, 2003, Fahrner et al., 2002, Attwood et al., 2002, Li, 2002). On one hand, the methylated cytosine bases can physically disrupt the binding of RNA polymerases and transcription factors to the appropriate regions of target genes. On the other, the methylated DNA sequences may be wrapped by multiple

proteins, including methyl-CpG-binding domain proteins (MBDs), histone deacetylases and chromatin remodelling proteins, to form complex structures, which can inactivate the chromatin and hence silence the transcription of the genes coded in the corresponding chromosome region.

Figure IV.1: Addition of a methyl group to the cytosine base in DNA sequence



Although the DNA methylation has been widely accepted to play an essential role in regulating the gene transcriptions in many species, the distribution patterns and levels of DNA methylation appear to vary drastically among different species. Several well-known eukaryotic model organisms, such as *Saccharomyces cerevisiae* (yeast) and *Caenorhabditis elegans* (C. elegans), do not encode any DNA methyltransferase-family genes and hence lack of DNA methylation in their genomes (Bird, 2002, Suzuki and Bird, 2008). Generally, the fungi, plants and invertebrate species, have moderately high methylation levels in many domains of the DNA sequences separated by domains of completely unmethylated genomic regions. This 'mosaic' methylation pattern has also

been discovered in a large number of species, including *Neurospora crassa*, *Arabidopsis thaliana* (Arabidopsis), *Zea mays* (corn), *Oryza sativa* (rice), *Poplulus trichocarpa* (poplar), *Ciona intestinalis* (sea squirt) and *Drosophila melanogaster* (fruit fly) (Chan et al., 2005, Gehring and Henikoff, 2007, Gowher et al., 2000, He et al., 2011, Henderson and Jacobsen, 2007, Zilberman et al., 2007, Montero, 1992, Palmer, 2003). In contrast, the vertebrate species, particularly mammals, typically exhibit 'global' DNA methylation patterns (Chen and Riggs, 2011, Robertson, 2005, Rollins, 2006). In vertebrates, the whole genome candidate methylation sites are completely methylated except those in the promoter regions. The methylation levels in promoter regions of vertebrates are highly varying among tissues and cells in different growth conditions and stages. The difference in methylation patterns among the eukaryotic organisms raises a question of whether it has a similar underlying mechanism at work, or whether the DNA methylation is co-opted to divergent biological functions and roles in different organisms. It has been found that the 'mosaic' methylation in plants and animals mainly targeted to the transposable regions is a crucial transcriptional silencing mechanism depending on the small interfering RNA (siRNA) modulation (Mette et al., 2000, Chan, 2004, Chan et al., 2005). However, there is no evidence supporting such a similar mechanism in vertebrates. Over the last decades, many studies have been carried out to investigate the biological functions that the 'global' methylation patterns played in vertebrates. It has been speculated that the methylation status and the local density of CpG dinucleotides within the promoter regions may act to control the gene transcription in vertebrates (Boyes and Bird, 1992, Hsieh, 1994, Weber, 2007). However, this hypothesis was based on analyses of limited genes and might not be explicable in general situations. The regulated degree of DNA methylation underlies the divergence

of gene expression cross different cell lines (or tissues) in vertebrates remains unknown. Moreover, why are the promoter regions generally, but not always, unmethylated in vertebrates? What is the role or mechanism of the methylation states in promoter regions? These questions are not well explored at present.

Furthermore, the distributions of candidate methylation sites are also varying among eukaryotic species and even among different genome features from the same species. Generally speaking, the substrates for DNA methylation are mainly located at dinucleotide 'CpG' sites in plants and animals, although studies had revealed some cases where cytosine within 'CpHpG' and 'CpHpH' sites were also methylated in plants, where H represents any nucleotide expect guanine (Goll and Bestor, 2005). Noticeably, it was observed that the abundance of CpG sites in human genome was only a quarter of that expected based on the GC content fraction of the human genome sequence (Robinson et al., 2004). It has also been observed that there were many short chromosomal regions which contained more CpG dinucleotides than the rest of the genome, and hence appeared as CpG enriched regions or so-called the CpG islands (Bird, 1986, Antequera and Bird, 1993, Gardiner-Gardner and Frommer, 1987, Glass, 2007). The CpG enriched regions are commonly overlapped with the promoters and many studies have shown that the presence of CpG dinucleotides enrichment in the promoter region is positively correlated with particular gene expression patterns (Ponger et al., 2001, Weber, 2007). Based on the analysis by Saxonov et al. (Saxonov et al., 2006), 72% of the promoters in the human genome had contained CpG enriched regions. Other studies have also revealed that the promoters containing CpG enriched regions in human genome were more frequently associated with 'house-keeping' genes (Robinson et al., 2004, Larsen et al., 1992). Although, so far, there are many significant achievements in

understanding the functional roles of the DNA methylation system in modulating gene transcription of vertebrates, the reason why CpG sites are enriched in promoters and how CpGs in promoters control the gene transcription still remain unanswered. In addition, most of our knowledge about the pattern of CpG sites distribution and the mechanism about the 'global' DNA methylation system involved in the regulation of gene transcription had been derived mainly from the studies of human species. What is the pattern of distribution of CpG sites in other vertebrates? Whether the distribution pattern is conserved across vertebrates? And, whether the regulatory mechanism of DNA methylation is conserved in vertebrates? All of these questions still remain unanswered.

To understand the transcriptional regulatory roles the DNA methylation system functioned in vertebrates' genomes, I firstly examined the distributions and properties of 'CpG' dinucleotides in 10 diverse eukaryotic genomes, including *Homo sapiens* (Human), *Mus musculus* (Mouse), *Rattus norvegicus* (Rat), *Bos taurus* (Cow), *Canis familiaris* (Dog), *Gallus gallus* (Chicken), *Danio rerio* (Zebrafish), *Drosophila melanogaster* (Fruitfly), *Caenorhabditis elegans* (C. elegans), *Arabidopsis thaliana* (Arabidopsis). I found that the distribution pattern of CpG sites is quite similar in the 6 higher vertebrates' genomes. These conserved distribution patterns of CpG sites in vertebrates are an interesting starting point for studying the regulatory mechanisms of DNA methylation in vertebrates. I also analyzed and compared genome-wide transcription and DNA methylation profiles across multiple cell lines (or tissues) of human genome by microarray technology. The results revealed two distinct methylation patterns and regulatory mechanisms in human genomes. Finally, the results of GO analysis provided strong support for the propose that the primary role of the DNA methylation system to control gene transcription was highly conserved in

vertebrates and could be further divided into two distinct classes according to the distribution of CpGs in promoter regions.

4.2 Datasets and analytical methods

With the development of whole genome sequencing technique and wide applications of microarray analysis, a large number of publicly available datasets described below were employed to investigate the DNA methylation mediated transcriptional regulation in vertebrate species.

4.2.1 Whole genome sequence datasets and genomic features annotation information

Whole genome sequence data for each of 10 eukaryotic model organisms was downloaded from the UCSC genome bioinformatics database (http://hgdownload.cse.ucsc.edu/downloads.html) (table IV.1). The genomic annotation information was obtained from the genome annotation database of UCSC GENOME BROWSER (http://genome-archive.cse.ucsc.edu/downloads), the Exon-Intron Database (EID, http://bpg.utoledo.edu/~afedorov/lab/eid.html), The Arabidopsis Information Resource (TAIR) (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp) and Mammalian Promoter Database (MPromDb, http://mpromdb.wistar.upenn.edu/). A FORTRAN program was developed to parse these sequence data and identify the locations of CpG sites, the proportions of CpG dinucleotides, the fractions of GC content in different genomic features of each selected organisms.

Table IV.1: Data resources and information of 10 selected model species

Species		Classification	Genome version	Number of promoters	Data resources
Human	*Homo sapiens*	higher vertebrate	hg18	34,257	UCSCdb,EID, MPromDb
Mouse	*Mus musculus*	higher vertebrate	mm9	38,330	UCSCdb,EID, MPromDb
Rat	*Rattus norvegicus*	higher vertebrate	rn4	12,721	UCSCdb,EID, MPromDb
Cow	*Bos taurus*	higher vertebrate	bosTau6	10,739	UCSCdb,EID
Dog	*Canis familiaris*	higher vertebrate	canFam3	1,481	UCSCdb,EID
Chicken	*Gallus gallus*	higher vertebrate	galGal4	3,640	UCSCdb,EID
Zebrafish	*Danio rerio*	lower vertebrate	danRer7	12,189	UCSCdb,EID
Fruitfly	*Drosophila melanogaster*	Insect	dm3	16,983	UCSCdb,EID
C.elegans	*Caenorhabditis elegans*	Nematode	ce10	8,849	UCSCdb,EID
Arabidopsis	*Arabidopsis thaliana*	Plant	TAIR9	25,516	TAIR

4.2.2 Distribution patterns of CpG sites in promoter regions

In statistics, Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of space if these events independently occur with a known average rate. Thus, Poisson distribution can be used to test randomness in distribution of discrete random events. I inspected whether occurrence of CpG sites in promoter regions could be modelled as the expected Poisson distribution. In the human promoter regions, it had an average of 51 CpG sites for every 1000 bp length. Hence, I could estimate the expected probability of promoters with 0~25 CpG sites in a 1000 bp interval as:

$$\Pr(k = 0 \sim 25; \lambda = 51) = \sum_{k=0}^{25} \frac{\lambda^k}{k!} e^{-\lambda} \qquad \text{(IV-2.1)}$$
$$= \sum_{k=0}^{25} \frac{51^k}{k!} e^{-51}$$
$$= 0.004\%$$

where *e* is the base of the natural logarithm (*e*=2.718), *k* is the number of occurrences of CpGs in 1000 bp sequence, λ is the average number of CpGs in 1000 bp. In the same way, I also calculate the expected Poisson probabilities of promoters with 26~40, 41~50, 51~60, 61~75 and >75 CpG sites in 1000 bp length. And then, I identified the observed proportions of promoters with 0~25, 26~40, 41~50, 51~60, 61~75 and >75 CpG sites in 1000 bp length. Due to variation in length of promoters, I needed to normalize the promoters' length into a fixed value of 1000bp. Pearson's chi-squared test was used to assess the goodness of fit between observed frequency and expected probability:

$$\text{Pearson's } \chi^2 = \sum (O - E)^2 / E \qquad \text{(IV-2.2)}$$

with degrees of freedom $df = 6 - 2 = 4$. I implemented Poisson distribution analysis for each of the 10 selected model species under the studies.

4.2.3 Identification of promoter classes

According to distribution pattern of CpG dinucleotides in the promoter regions from each vertebrate species, the promoters could be grouped into two classes to distinguish High CpG density promoters (HCP) and Low CpG density Promoters (LCP). For every selected vertebrate, I calculated the GC fraction and the ration of observed /expected

(O/E) CpGs in the promoter region of each annotated genes. The O/E ratio was estimated as follows:

$$ratio\ \text{of}\ Obs\ /\ Exp = \frac{number\ \text{of}\ CpG}{number\ \text{of}\ C \times number\ \text{of}\ G} \times N \qquad \text{(IV-2.3)}$$

where N is the length of the promoter. These two classes of promoters was defined using following criteria: the HCP category contains promoters with CpG observed/expected (O/E) ratio above 65% and GC fraction over 55%; the LCP group includes promoters with CpG O/E ratio < 65% and GC fraction < 45%; the rest unclassified promoters are considered as Intermediate CpG density Promoters (ICP) as proposed in (Weber, 2007, Saxonov et al., 2006). The analyse described above was performed through FORTRAN programs.

4.2.4 Identification of homologous genes and interspecies conservation analysis

The homologous genes across 6 higher vertebrate species were downloaded from the NCBI-HomoloGene Database (release 65, http://www.ncbi.nlm.nih.gov/homologene/) (table IV.2). The NCBI-HomoloGene is a comprehensive tool for detection of homologues across the annotated genes of all 6 completely sequenced vertebrate genomes which were recruited in the current analysis. It uses both DNA sequence and protein sequence data to calculate the similarities attributable to descent from a common ancestor, and then identify the homologous gene families. When a gene was confirmed homologue between two vertebrates, I checked whether their promoters were grouped into the same classification in the two species. If the promoter classifications were the

same, I considered the distribution of the CpG sites in the promoter of this homologous gene was conserved between the two species.

Table IV.2: Datasets of the homologous genes for 6 higher vertebrate species

Species	Total number of genes[a]	Number of Homologue genes[b]
Human	19,565	18,631
Mouse	22,566	16,841
Rat	21,943	17,950
Cow	21,121	17,472
Dog	19,176	17,187
Chicken	16,731	13,150

a: the number of annotated genes used in homologue gene analysis
b: the number of homologue genes identified with the other species

4.2.5 Genome-wide DNA methylation data and gene expression data

The methylation data was collected from 28 different human tissue or cell line samples. The genome-wide DNA methylation levels for each of these samples were extracted using the Illumina HumanMethylation27 BeadChip platform, which consists of 27,578 probe units (representing 27,578 CpG sites and covering over 14,000 genes' promoter regions). The methylation microarray raw data were downloaded from NCBI Gene Expression Omnibus (GEO) database under the series accession number GSE17769, GSE20872, GSE24087 and GSE28356 (http://www.ncbi.nlm.nih.gov/geo). The methylation levels for each CpG site and each sample were calculated by the standard Illumina procedure-Genome Bead Studio Software, which provides a quantitative measurement of DNA methylation. The methylation level varied from 0 to 1, reflecting to as completely unmethylated to methylated.

The gene expression data were obtained from 107 different human tissues (or cell lines). The mRNA from each sample was extracted, and then hybridized to Affymetrix U133 human expression microarrays. The Affymetrix U133 human expression microarray GeneChip, which contained over 45,000 probe sets for representing approximately 33,000 well annotated human genes, was an ideal tool to assess whole human genome expression. After RNA hybridization to microarray, raw signal intensities were acquired, and then analysed by the standard Affymetrix algorithm MAS5.0 and normalized by the global median scaling method. Here, all gene expression analyses were implemented by using R scripts. The raw microarray data from the 107 different human tissue or cell line samples was downloaded from NCBI Gene Expression Omnibus (GEO) database under the series accession number GSE7127, GSE17768, GSE24089 and GSE26133.

4.2.6 GO annotation datasets and overrepresentation analysis

The genome-wide GO annotation information for each of the 6 selected vertebrates was downloaded from Gene Ontology database (http://www.geneontology.org/) (table IV.3). To identify GO terms overrepresented in the HCP or LCP grouped genes, the binomial test was employed to compare the number of ORFs in a gene group associated with a GO term of interest to the number of genome-wide ORFs associated with that GO term. For each GO term, a Z statistic is computed as following:

$$Z = \frac{(F_d - F_G)}{\sqrt{\frac{F_G(1-F_G)}{N_d}}} \qquad \text{(IV-2.4)}$$

where F_d was the fraction of the HCP (or LCP) promoter genes annotated into the specific GO term, F_G was the fraction of all annotated genes in that term, and N_d was the total number of genes with the HCP (or LCP) promoters. A GO term was determined to be significantly overrepresented in a group when Z>4.75 ($P < 1.0 \times 10^{-6}$, after the Bonferroni correction for multiple tests).

Table IV.3: Datasets from Gene Ontology database for 6 higher vertebrate species

Species	Number of annotated genes[a]	Number of annotations
Human	45,678	344,494
Mouse	25,503	281,825
Rat	25,106	260,633
Cow	21,582	112,409
Dog	19,891	106,432
Chicken	16,832	96,743

a: the number of genes which are characterized and annotated in GO database

4.3 Analysis and Result

4.3.1 Overview of the genome-wide distributions of CpG sites and GC contents in several species

To properly quantify the extent by which DNA methylation affects gene transcription, one needs knowledge of the distribution patterns of candidate methylation sites in the genome. In animals and plants, methylation primarily occurs at the CpG sites (Goll and Bestor, 2005). Thus, I first overviewed the distributions of the GC content fraction and the CpG concentration in different genome feature regions. Although the distribution of GC content and CpG density had been thoroughly analyzed for humans (Saxonov et al., 2006), little is known about the relevant information in other species. With the fast development of DNA sequencing techniques, numerous species have been fully sequenced. Here, I carried out an investigation of distribution of the GC content and the CpG sites in different genome feature regions of 10 well-studied eukaryotic model species. In these 10 model species, 6 of them (human, mouse, rat, cow, dog and chicken) are higher vertebrates, while the rest species (zebrafish, D. melanogaster, C. elegans and Arabidopsis) are lower vertebrate, invertebrate or plant.

Table IV.4 summarized the GC content fractions and the expected and observed proportions of CpG sites in different genome feature regions of the selected model species. The expected proportions of CpG sites were calculated based on random union of C and G nucleotides. Among the 6 higher vertebrate species, the whole genome and intron regions had the lowest fraction of GC contents (37.95~42.46%), followed by the exon regions which had GC content fractions of 48.88~51.57%. the promoter regions

had distinctly higher GC content (52.21~57.29%). Here, the observation was consistent with previous studies that the functional sequence had a higher GC content level than the entire genome or the non-functional region (Pozzoli et al., 2008). In contrast, for the lower vertebrate, invertebrates and plant I found that while the exon regions had higher GC contents compared with the entire genome or the intron regions, the promoter regions did not show enhanced GC content, but instead showed a similar make up to the entire genome or the intron regions.

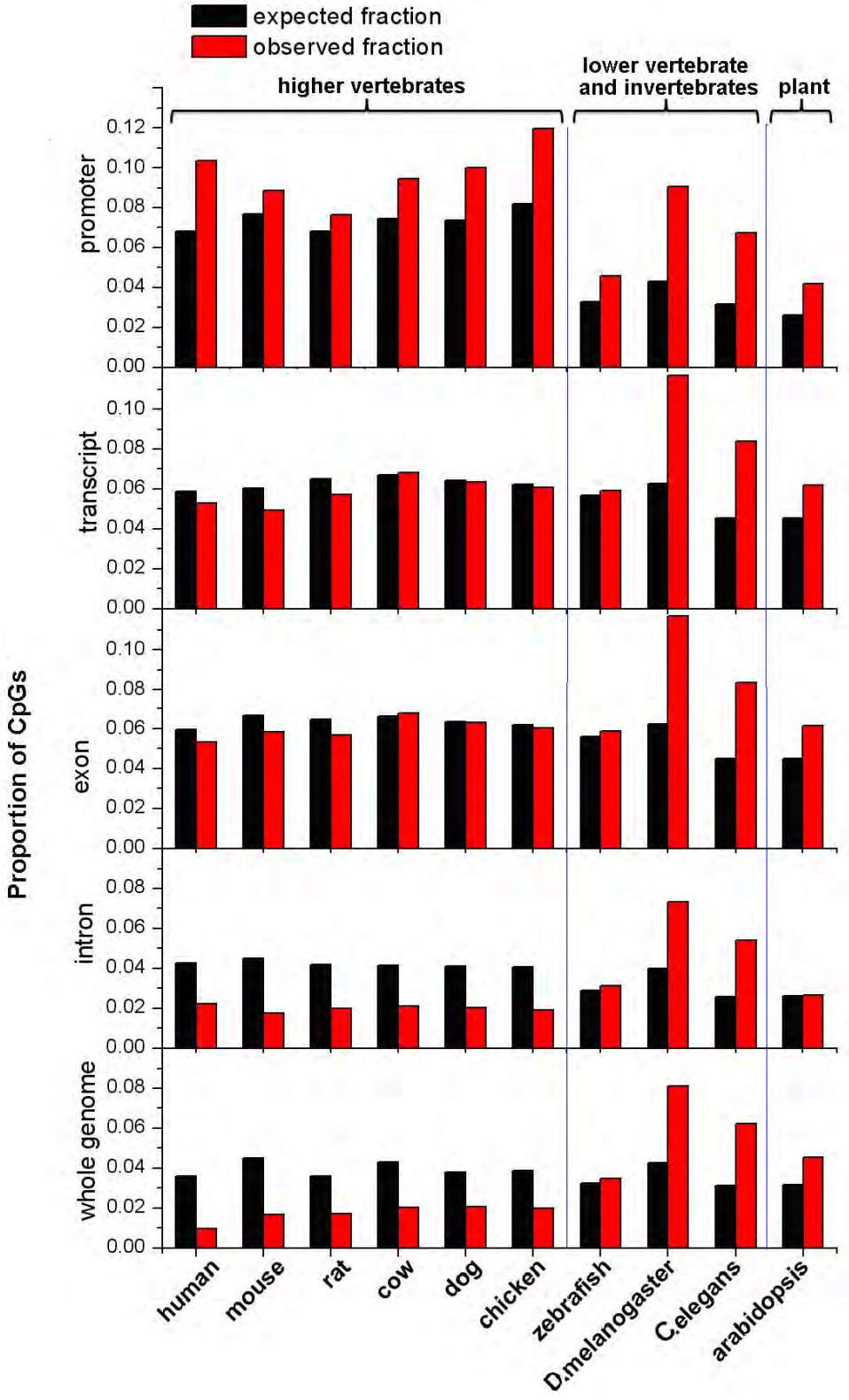
Table IV.4: GC content and distribution of CpG sites in vertebrates, invertebrates and plant

	Higher vertebrates			Lower vertebrate, invertebrates and plant		
	GC %	CpG %[a]	CpG %[b]	GC %	CpG %[a]	CpG %[b]
Genome-wide	37.95~42.39	3.61~4.49	0.95~2.08	35.44~41.24	3.14~4.25	3.48~8.11
Intron	40.37~42.46	4.07~4.50	1.75~2.21	32.14~39.91	2.58~3.98	2.65~7.35
Exon	48.88~51.57	5.97~6.65	5.35~6.78	42.42~50.00	4.50~6.25	5.89~11.69
Transcripts	48.47~51.72	5.87~6.69	4.94~6.82	42.59~50.10	4.54~6.27	5.93~11.71
Promoter	52.21~57.29	6.81~8.21	7.66~11.98	32.42~41.55	2.63~4.32	4.19~9.08

GC% is the proportion of GC content in different genome features. CpG %[a] is the expected fraction of CpG dinucleotides based on the GC content proportion. CpG %[b] is the observed fraction of CpGs for different genome features. I separately detect the proportions of GC content and CpG % for every model organism, and then summarize the fraction ranges in 6 higher vertebrates or 4 lower vertebrate, invertebrates and plant.

With regard to CpG sites, it was found that the CpG dinucleotides were consistently enriched in promoter regions in the 6 higher vertebrates, consistent with the fact that the promoters had higher level of GC contents in these species (Figure IV.2). However, it needs to be stressed that, in the 6 higher vertebrates, the observed CpG sites proportions were higher than expected in the promoters, but lower than expected in the entire genome. In contrast, it was observed a completely different pattern of CpG sites distribution in the other four model species (zebrafish, D. melanogaster, C. elegans and Arabidopsis), in which the observed proportion of CpGs always exceeding those expected and was similar between promoters and the entire genome (Figure IV.2). This discrepancy of CpG sites distribution patterns between higher vertebrates and the other 4 species was significant under the Mann-whitney test (P-value < 0.05).

Figure IV.2: Expected and observed proportions of CpGs across the different genomic feature regions of 10 model species.



4.3.2 Analysis of the pattern of the distribution of CpG sites in the promoter regions

To better understand the role of DNA methylation in regulating gene expression, I focused on the distribution of CpG sites in the promoter regions. The first question was to test whether the CpG sites were randomly distributed in the promoter regions. To address this question, I tested if the occurrence of CpG sites in the promoters followed a Poisson distribution with parameter λ (i.e. mean of the distribution) being the genome-wide average number of CpG sites in a fixed promoter length, here 1000 base pairs (bps). The number of CpG sites occurring per 1000 bp of the promoter was counted and assigned into six categories according to the number of CpG sites: 0~25, 26~40, 41~50, 51~60, 61~75 and >75 CpG sites. Pearson's chi-square test was employed to test whether the observed fractions of CpG sites numbers classified in the six categories were consistent with the expectations based on the Poisson distribution. For the 6 higher vertebrate species, all of the Pearson's chi-square P-values were smaller than 10^{-15}, showing that the distribution of CpG sites in the 6 higher vertebrate genomes did not follow the Poisson distribution, and therefore were not randomly scattered in the promoter regions of these species. However, for the four lower vertebrate, invertebrate and plant species (zebrafish, D. melanogaster, C. elegans and Arabidopsis), the occurrence of CpG sites in the promoters did follow the Poisson distribution and hence they were randomly distributed (Pearson's chi-square P-values all exceeded 5%).

To further characterize the non-random distribution of CpGs in higher vertebrate promoters I looked at the occurrence of 'CpG islands', which are recognized as small dispersed regions of DNA sequence that contain highly dense clusters of CpG

dinucleotides relative to the whole genome. The widely accepted definition of a 'CpG island' is a genomic region with at least 200 bps in length, with the GC content fraction larger than 50% and the observed/expected CpG percentage ratio greater than 60% (Gardiner-Garden and Frommer, 1987). Out of the 34,257 annotated promoters of the human genome, I found 25,248 (74%) promoters containing 'CpG islands' while the rest 9,009 (26%) promoters have only few CpG dinucleotides. For the other 5 higher vertebrate species, CpG islands were detected in over half of their annotated promoters. The density of CpG sites in the promoters of all 6 higher vertebrates showed a bimodal distribution, which has been only found previously in human genome (Figure IV.3) (Saxonov et al., 2006). However, the CpG islands could not be found in the rest of 4 lower vertebrate, invertebrate and plant genomes and the densities of CpG sites in the promoters showed a unimodal distribution in these species (Figure IV.3). Interestingly, a consistent unimodal distribution pattern of GC fraction in both vertebrate and invertebrate species (Figure IV.4). Hence, it is unlikely that the bimodal distribution of CpG proportion in promoters across all higher vertebrates is due to the GC content distribution. I proceeded to explore the functional roles of DNA methylation in regulating gene expression to explain the bimodal distribution pattern of CpGs in the promoters of higher vertebrates.

Figure IV.3: Histograms of the CpG sites proportions in the promoters of the 10 model species (The horizontal axis represents the CpG proportions in promoters, while the vertical axis represents the number of promoters for each model species)

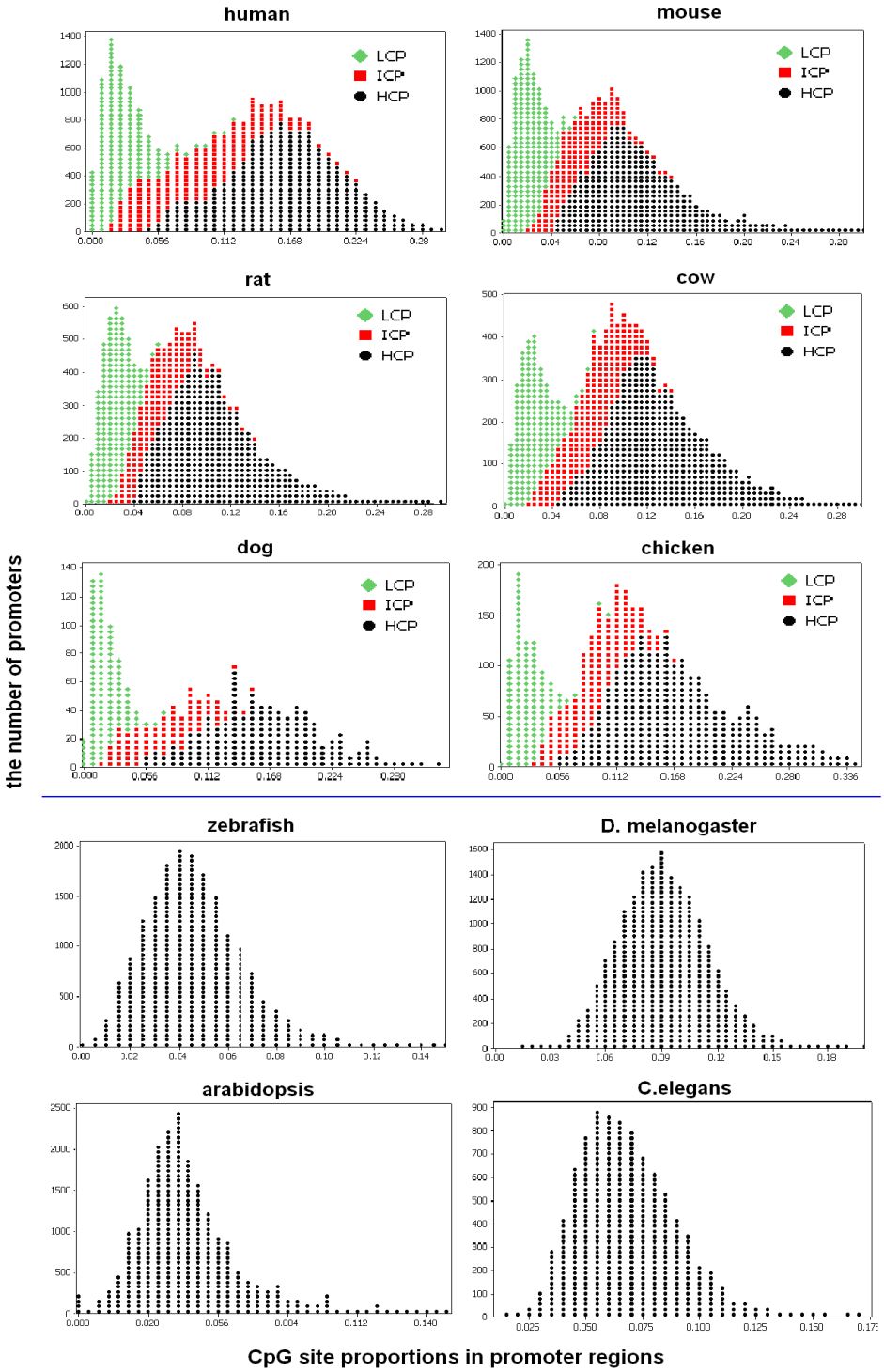
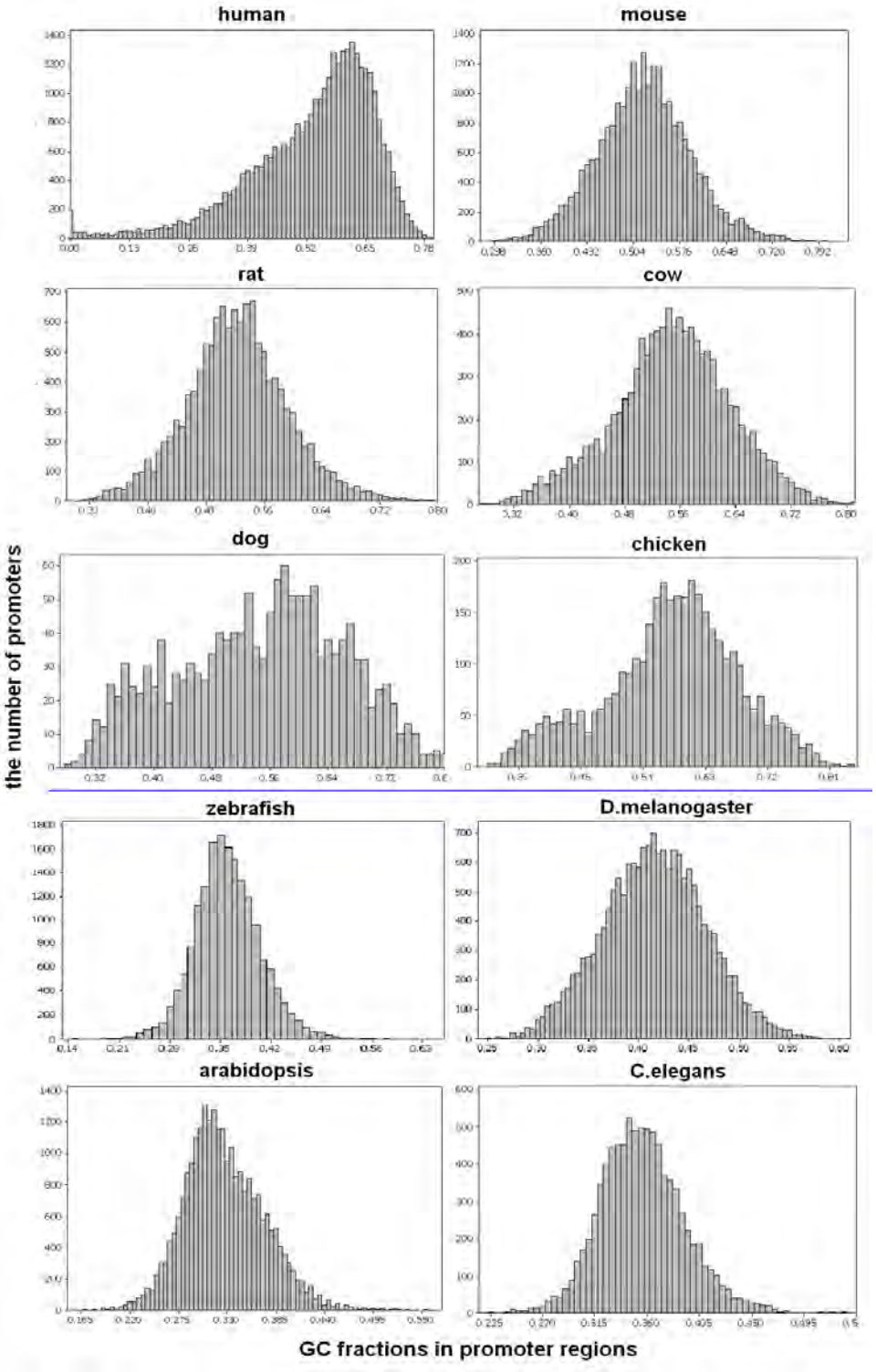


Figure IV.4: Histograms of GC fractions in the promoters of the 10 model species (The horizontal axis represents the GC content fractions in promoters, while the vertical axis represents the number of promoters for each model species)



I classified gene promoters of the higher vertebrate species into two main groups previously defined for human genes (Weber, 2007) according to the GC fraction and observed to expected ratio of CpG sites (O/E). First, High CpG density Promoters (HCP) with GC fraction ≥55% and CpG O/E ≥65%; second, Low CpG density Promoters (LCP) with GC fraction <45% and CpG O/E <65%. The remaining genes were difficult to assign into either group and were grouped as Intermediate CpG density Promoters (ICP) (Weber, 2007, Saxonov et al., 2006). For each of the six higher vertebrates, there were approximately 50%, 25% and 25% of promoters classified as HCP, LCP or ICP respectively. In the following analyses, I focused on the two most divergent classes (HCP and LCP). A striking difference was observed between HCP and LCP promoters both for the GC content faction and the occurrence of CpG sites at varying distances from the transcription start site (TSS) (Figure IV.5 and Figure IV.6). For the HCP promoters in the higher vertebrates, both the proportion of CpG sites and the GC content fraction peaked consistently in the vicinity of the TSS and declined with increasing distance from the TSS. On the other hand, the proportions of CpG sites in LCP promoters were consistently close to zero, despite a peak for the GC content fraction around the TSS. These results indicated a high level of conservation of CpG site distribution among higher vertebrate species, suggestive of a possible link with important biological functions. For the zebrafish, a lower vertebrate, the patterns of GC content fraction and CpG site density at all promoters were similar to those of the HCP promoters of the higher vertebrates. The pattern was obviously different for the invertebrate species, with the GC content fraction and CpG density exhibiting a sharp peak immediately downstream of the TSS, but either a flat curve (*Arabidopsis* and *C. elegans*) or surprisingly a valley (*D. melanogaster*) upstream of the TSS.

Figure IV.5: Distributions of the CpG sites with respect to transcription start site (TSS) (The horizontal axis represents the distance from TSS, while the vertical axis represents average proportion of CpGs in that specific location for each model species)

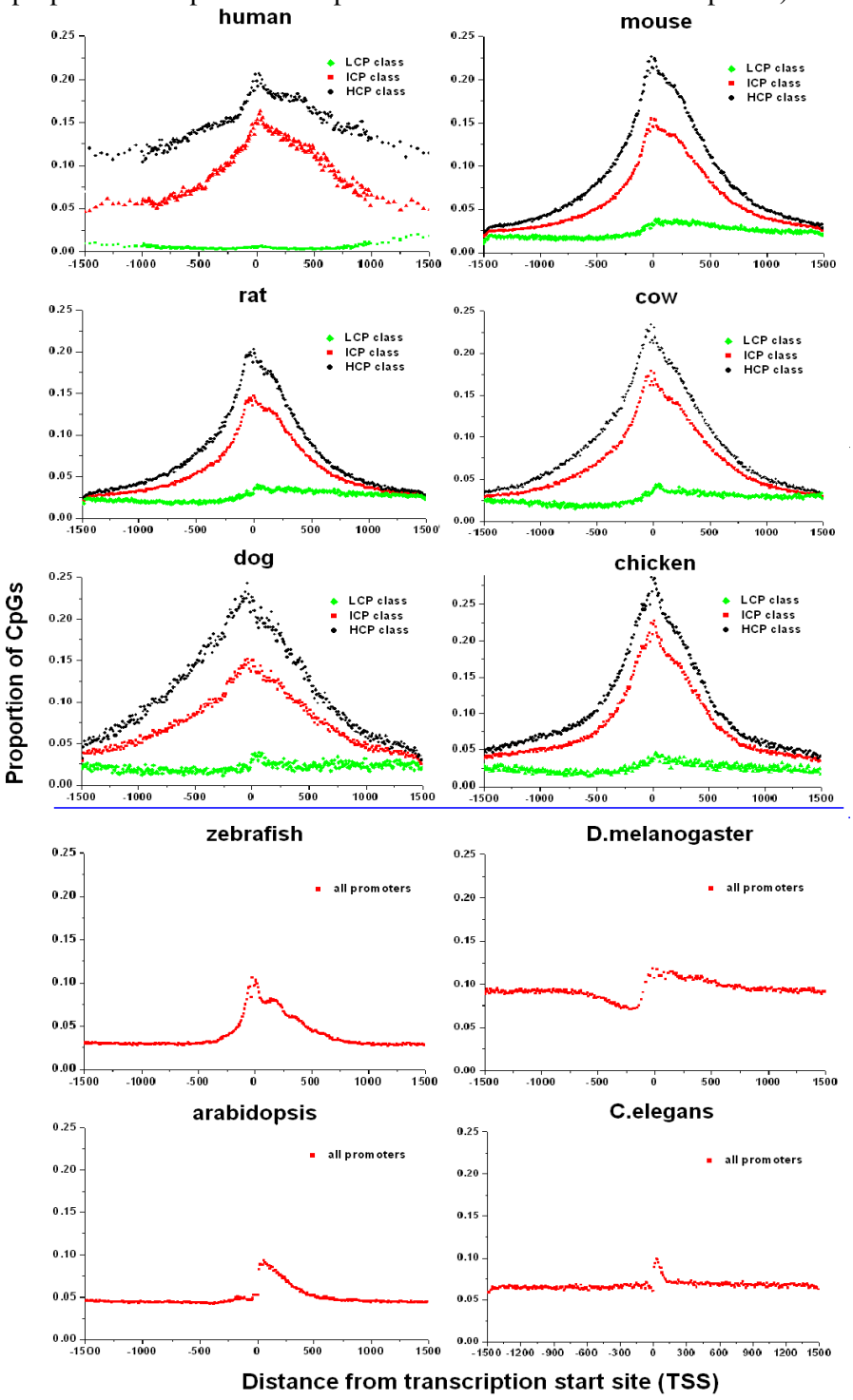
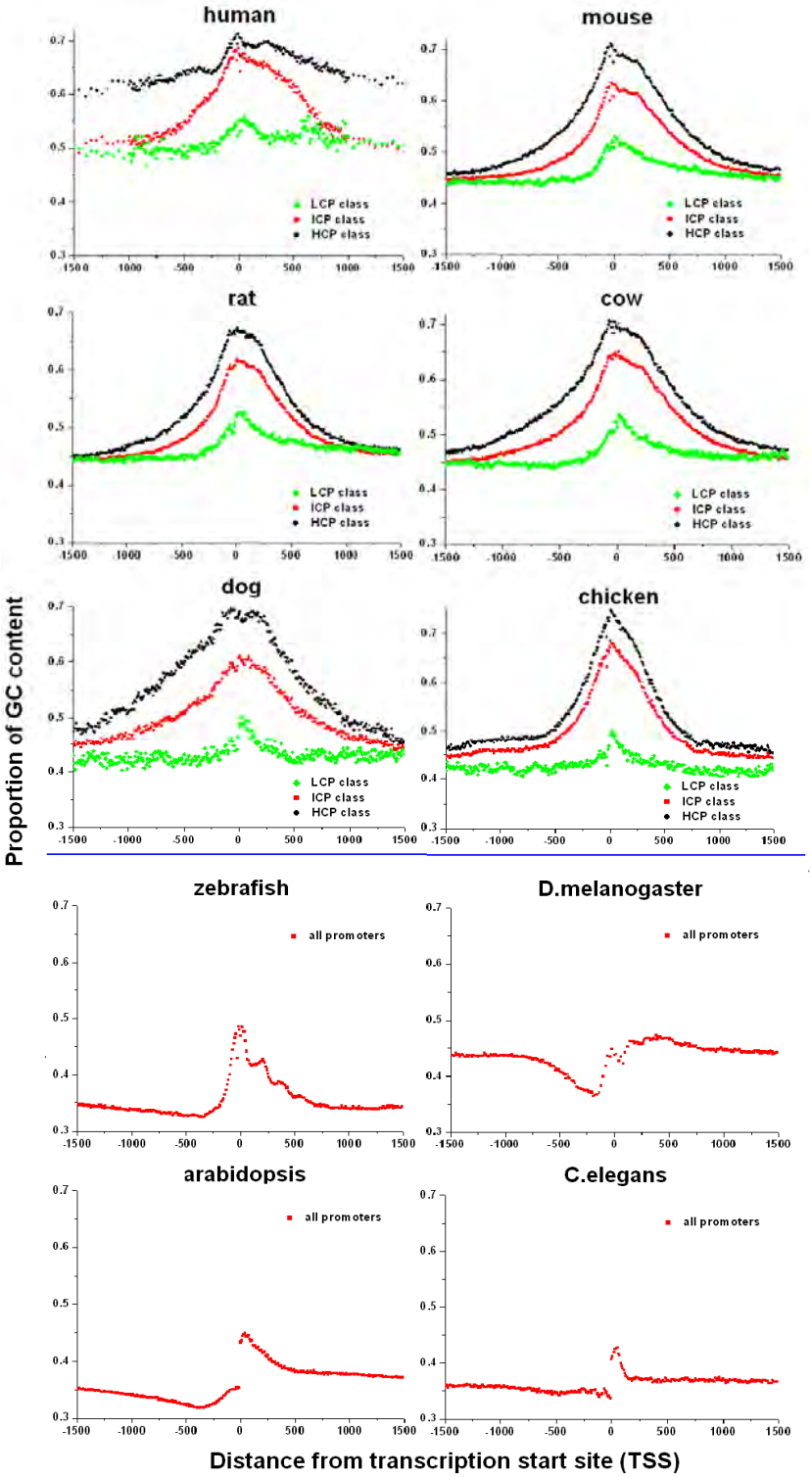


Figure IV.6: Distributions of GC fractions with respect to transcription start site (TSS) (The horizontal axis represents the distance from TSS, while the vertical axis represents the GC fractions in that specific location for each model species)



4.3.3 Evolutionary conservation of promoters in higher vertebrates

In the above analysis, I had observed that the patterns of the distribution of CpG sites in promoters were similar among the 6 higher vertebrate species, suggesting a possible link with biological implications. Here I would like to know whether the CpG sites in the promoters were evolutionarily conserved. For this purpose, I downloaded homologous gene list from the NCBI-HomoloGene database and tested if the promoters from homologous genes of different species were classified into the same promoter category of either the high (HCP) or low (LCP) density of CpG dinucleotides. In this analysis, I only considered the genes which have only one promoter. For a gene homologous between a pair of species among the 6 higher vertebrates, its promoter can be assigned to one of the different promoter classes in either species as described above. If the classifications were identical between a pair of species, I called the CpG sites distribution pattern in the corresponding promoter as being 'conserved'.

The analysis results were tabulated in Table IV.5. Each column of Table IV.5 represents the proportion of HCP or LCP promoters of homologous genes in the column species that were also classified as the same category in the row species. For example, there were 11,224 homologous genes between human and mouse. Among those homologous genes, 93.7% of human genes having HCP promoters were also having HCP promoters in mouse. Meanwhile, 97.6% of the genes with HCP promoters in mouse were also classified as HCP promoters in homologous human genes.

Table IV.5: Conservation of two classes of promoters across higher vertebrates

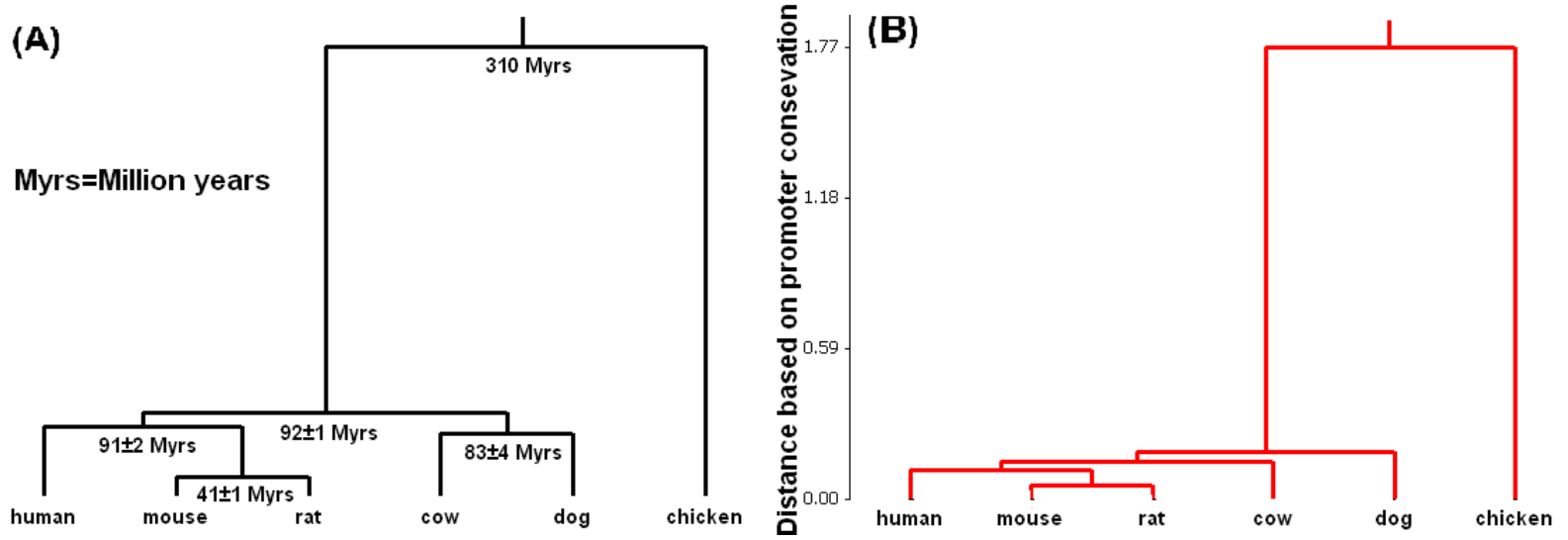
	Proportions of conserved HCP promoters (%)						Proportions of conserved LCP promoters (%)					
	Human	Mouse	Rat	Cow	Dog	Chicken	Human	Mouse	Rat	Cow	Dog	Chicken
Human	7139	97.6	97.4	96.9	88.8	85.7	2895	86.7	87.3	79.5	89.8	42.1
Mouse	93.7	8097	96.7	93.7	85.8	84.7	85.1	4365	89.9	83.2	87.4	48.2
Rat	91.9	92.6	5634	90.4	83.4	89.7	85.0	94.6	2596	86.9	77.9	56.3
Cow	93.6	94.1	95.2	1536	84.5	84.3	90.6	96.0	79.3	577	88.9	54.8
Dog	82.2	80.5	84.9	86.3	435	80.6	81.1	90.6	87.2	97.1	187	40.0
Chicken	89.6	87.0	89.7	84.4	82.5	913	47.3	53.6	51.6	53.7	33.3	251

The diagonal cells show the number of genes with high (HCP) or low (HCP) CpG density promoters in each species. The upper and lower triangles show the percentage of genes in the column species also given the same classification for the row species. For example, for 93.7% of genes with HCP promoters in human also had HCP promoters in mouse, while, 97.6% of genes with HCP promoters in mouse also had HCP promoters in human.

Table IV.5 showed that the HCP promoters were highly conserved among the 6 higher vertebrate species. For each pair of vertebrate species, more than 80% of the HCP promoters of homologous genes in one species were also classified as HCP promoters in the other species. For LCP promoters of homologous genes, the conservation levels were found to be slightly lower than that of the HCP promoters in each pair of vertebrates, particularly in the pairs between chicken and the other 5 mammals. For example, for the homologous genes between dog and chicken, only 33.3% of the LCP promoters in dog were classified as LCP promoters in chicken. This variation may be expected due evolutionary changes that have occurred between bird and mammal lineages.

It is interesting to note that the inter-species conservation of CpG sites in promoters could be used to infer the evolutionary relationships among species. By building a phylogenetic tree using the proportions of HCP/LCP promoters retained in the same category for a pair of species (Table IV.5) as an evolutionary divergence measure, I observed that the 5 mammals were closely linked each other while the chicken was in a separate cluster, which was remarkably similar to the phylogeny derived from DNA and protein sequence data (Hedges, 2002) (Figure IV.7). The only difference between the two phylogenetic trees is how the dog species is linked to the tree. In the tree based on promoter conservation, the dog species diverged prior to all of the other mammals, while for the tree based on DNA and protein sequence data, the dog and cow diverged from the other three mammals around 92 million years ago, before the two separated around 83 million years ago. This discrepancy might be due to the limited number of promoters available for dog species. In fact, a total number of 13,410 genes were annotated for dog species, but only 1,481 promoters were identified.

Figure IV.7: Cluster analysis of higher vertebrate species



(A): the phylogenetic relationships and times of divergence among vertebrates based on DNA and protein sequence data (Hedges, 2002)
(B): the phylogeny of vertebrates constructed through the promoters' conservation information

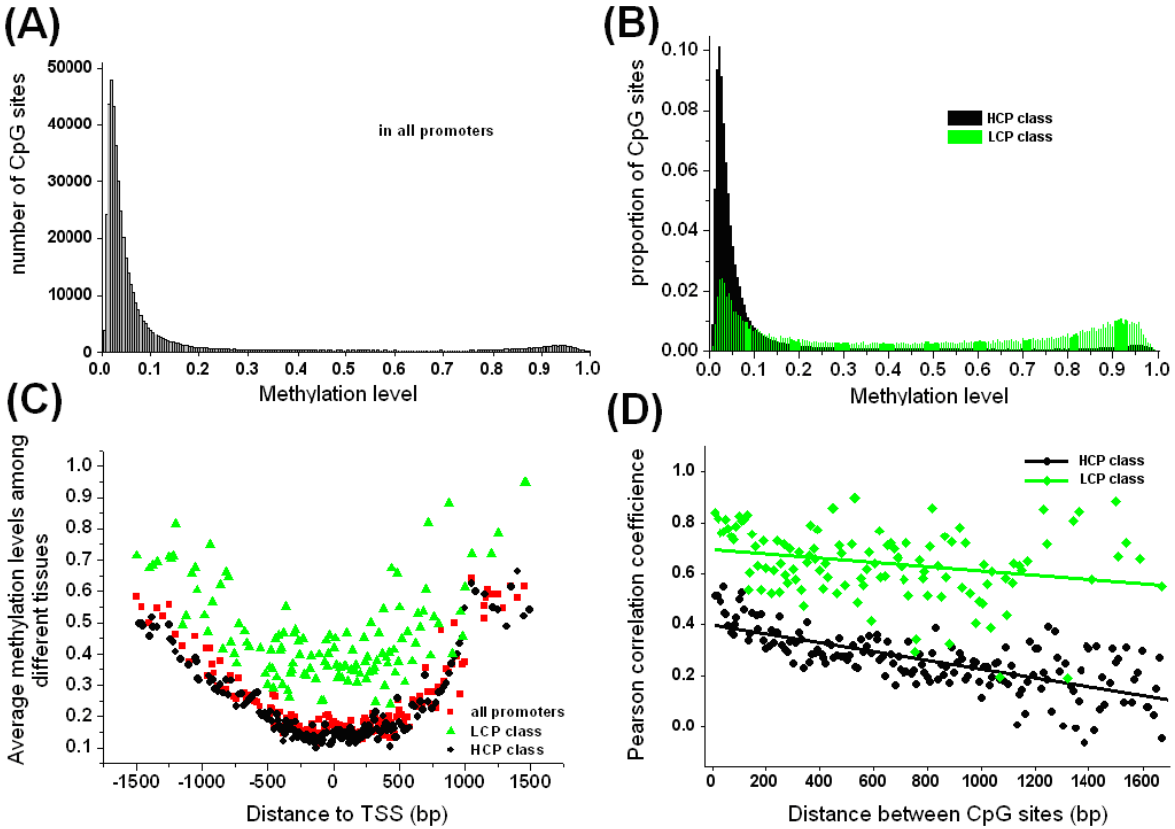
4.3.4 Distinct methylation patterns between HCP and LCP promoters across 28 human tissues

I analysed genome-wide DNA methylation profiles in 28 different human tissues (or cell lines) assayed by the Illumina HumanMethylation27 BeadChip platform (Bonazzi et al., 2011, Loudin et al., 2011, Chari et al., 2011). This BeadChip assessed 27,578 CpG sites located within the promoter regions of 14,475 genes. On average, two CpG sites were interrogated per promoter region. Thus, the BeadChip assay provides an efficient solution for surveying the DNA methylation level in genome-wide promoter regions. Here, the analysis was consistent with previous findings that it had much lower methylation level in promoter region relative to genome-wide genome (Lister et al., 2009). Figure IV.8 shows the distribution of the methylation levels of the CpG sites in the promoters across 28 different human tissues. The majority (72.7%) of the CpG sites in promoter regions were unmethylated (methylation level ≤ 0.1), while 18.5% were identified as semi-methylated (methylation level in 0.1~0.7) and only 8.8% of CpG sites were considered as methylated (methylation level ≥ 0.7). The distribution of methylation levels showed two distinct patterns for LCP compared with HCP promoters (Figure IV.8(B)). HCP promoters showed a unimodal distribution, with 77.1% unmethylated, 16.6% semi-methylated and 6.3% methylated CpG sites, while LCP promoters showed a bimodal distribution, with corresponding proportions of 25.8%, 37.9% and 36.3%. I also investigated the CpG sites' methylation levels with respect to the transcription start site (TSS) in all 28 human tissues. Illustrated in Figure IV.8 (C), each spot represented the average methylation level in an interval of 10 bp surrounding the TSS. I found that the CpG sites in both HCP and LCP promoters had the lowest

methylation levels in the core promoter regions at the vicinity of the TSS. The methylation level increased as their distance to the TSS increased. Furthermore, the CpG sites in the LCP promoters showed much higher methylation levels than those in the HCP promoters.

The Illumina HumanMethylation27 BeadChip interrogated multiple CpG sites for a number of promoters, enabling us to examine whether the methylation levels between two CpG sites located in the same promoters were correlated. For a pair of CpG sites located in the same promoters, the Pearson's correlation coefficient was calculated for their methylation levels across 28 tissues. And then, the average correlation coefficients in 10bps intervals with respect to the distance between these two CpG sites were shown in Figure IV.8 (D). The analysis showed that the methylation levels were positively correlated between CpGs located in the same promoter, particularly when the two CpG sites were close to each other. As the distance between two CpG sites in the same promoter increased, the correlation of methylation levels decreased. Figure IV.8 (D) also shows that the CpG sites within the LCP promoters exhibited greater degrees of correlations in methylation levels than the CpG sites in the HCP promoters.

Figure IV.8: Distribution of methylation patterns across 28 different human tissues



Methylation levels of CpG sites in all promoters **(A)**, and in HCP and LCP promoters **(B)**, across 28 different human tissues. The average methylation levels with respect to the transcription start site, with each point representing the average methylation level in an interval of 10 bp **(C)**. The correlation of methylation levels between all pairwise CpGs sites in the same promoter, with each point showing the average correlation in 10bp intervals according to the distance between CpG sites **(D)**.

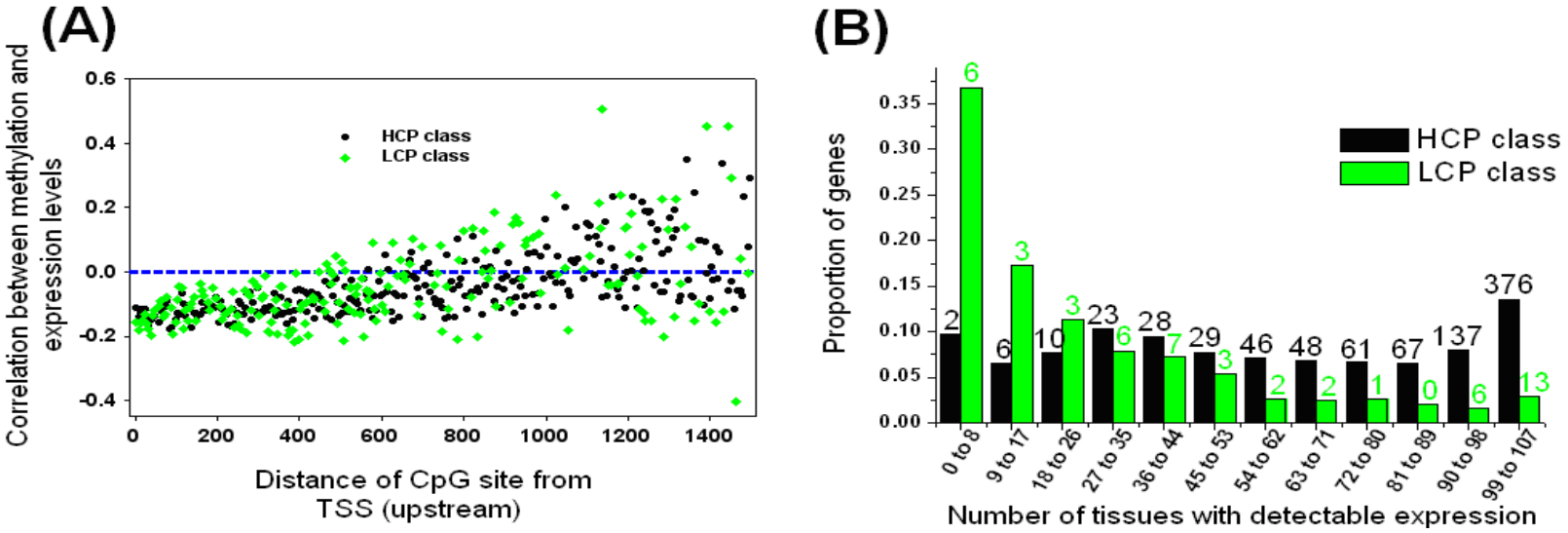
4.3.5 Distinct expression patterns between HCP and LCP promoters in 107 human tissues

I investigated the relationship between promoter DNA methylation and gene expression levels measured across 107 human tissues (including those 28 in the above methylation analysis) by the Affymetrix U133 human expression microarray (Johansson et al., 2007, Chari et al., 2011, Bell et al., 2011). I discovered a clear negative correlation (from -0.05 to -0.18) between the gene expression level and methylation level of each profiled CpG site across the 28 tissues (Figure IV.9(A)). This correlation was confined to CpGs located in the core and proximal promoter regions (0 bp to 250 bp upstream of the TSS). For CpG sites located further upstream (>250bp) from the TSS, the strength of correlation decreased and no obvious relationship with gene expression level was apparent. No differences were observed between LCP and HCP promoters. Methylation of CpG sites in the core and proximal promoter regions must therefore play a crucial role in regulating gene expression levels.

Next, I compared the number of tissues from which each gene was detectably expressed, from a total of 107 tissues (Figure IV.9(B)). The difference between LCP and HCP genes was striking. Genes with LCP promoters tended to be expressed in only a small number of tissues compared to genes with HCP promoters. Over 35% of genes with LCP promoters were expressed in no more than 8 tissues, while only <5% were expressed in 99-107 tissues. On the other hand, genes with HCP promoters showed a reasonably uniform distribution for the number of tissues expressed from 0 to 107, and ~15% of genes were expressed in 99-107 tissues. Genes with LCP promoters were

therefore more likely to be 'tissue-specific', while those with HCP promoters were more likely to be 'housekeeping' genes. In fact, among 885 housekeeping genes identified in the human genome (Zhu et al., 2008) that were also included in the present gene expression dataset, only 5.9% had LCP promoters, while 94.1% had HCP promoters. In particular, 376 (42.4%) of these housekeeping genes had HCP promoters and were expressed in almost all (99-107) tissues.

Figure IV.9: The correlation coefficients between methylation and gene expression level with increasing distance from the transcription start site **(A)**. Distribution of the number of tissues in which HCP and LCP genes are expressed. Each bar is labelled with the corresponding number of house-keeping genes **(B)**.



4.3.6 Distinct and conserved functions of genes with HCP and LCP promoters

In previous gene ontology (GO) studies of human genome, it has been found that the promoters with 'CpG Island' were more likely associated with genes which had basic cellular functions, while the promoters without 'CpG Island' were likely associated with genes with tissues-specific functions (Larsen et al., 1992, Ponger et al., 2001, Saxonov et al., 2006). In this part, I attempted to explore if any GO term was disproportionately overrepresented in HCP or LCP promoter gene group in each of 6 higher vertebrate genomes. In addition, I examined whether the over-represented GO term(s) was shared among the higher vertebrates.

To identify the overrepresentation of GO classes, I analyzed the number of genes in each GO class using the binomial test. A GO term was claimed to be significantly overrepresented when $Z>4.75$ ($P < 10^{-6}$ after Bonferroni correction). For every higher vertebrate, the GO over-representation analysis identified about 100 GO terms significantly overrepresented in HCP or LCP promoter gene group respectively. Briefly, the HCP and the LCP promoter classes were predicted with different GO terms in all 6 selected vertebrates: for LCP promoter genes, the GO terms were particularly presented in immunological function, response to stimulation, and functions characteristic of more differentiated or highly regulated cells, whereas GO terms detected for the HCP promoter genes were involved in the basic cellular processes, such as regulation of transcription, cell cycle, structure, and protein processing.

For those overrepresented GO terms discovered, I tested whether they were shared among the higher vertebrate species. I called those GO terms shared by at least 4

vertebrate species as inter-species 'conserved' terms. Accordingly, 16 and 12 GO terms had been identified to be conserved in HCP and LCP promoter gene groups respectively (Table IV.6). As expected, the conserved GO terms were enriched in 'tissue-specific' functions for the LCP promoter group and enriched in 'house-keeping' functions for the HCP promoter genes.

Table IV.6: A list of conserved and overrepresented GO terms for HCP and LCP classes in the 6 higher vertebrates

GO ID	Conservation[a]	Category[b]	GO term description
Overrepresented in HCP class			
0000122	4	BP	regulation of transcription from RNA polymerase promoter
0003676	4	MF	nucleic acid binding
0003677	4	MF	DNA binding
0003723	4	MF	RNA binding
0004672	4	MF	protein kinase activity
0004930	4	MF	G-protein coupled receptor activity
0005634	4	CC	nucleus
0005730	4	CC	nucleolus
0006915	4	BP	apoptotic process
0016021	4	CC	integral to membrane
0016301	4	MF	kinase activity
0043234	4	CC	protein complex
0043565	5	MF	sequence-specific DNA binding
0044212	4	MF	transcription regulatory region DNA binding
0045892	4	BP	negative regulation of transcription, DNA-dependent
0045893	4	BP	positive regulation of transcription, DNA-dependent
Overrepresented in LCP class			
0004869	4	MF	cysteine-type endopeptidase inhibitor activity
0004984	4	MF	olfactory receptor activity
0006955	4	BP	immune response
0006958	5	BP	complement activation, classical pathway
0006974	4	BP	response to DNA damage stimulus
0007596	4	BP	blood coagulation
0007601	4	BP	visual perception
0008009	4	MF	chemokine activity
0008270	4	MF	zinc ion binding
0009897	4	CC	external side of plasma membrane
0015711	4	BP	organic anion transport
0032729	4	BP	positive regulation of interferon-gamma production

a: the number shown in 'Conservation' represents that the specific overrepresented GO term is conserved in how many vertebrates.
b: the abbreviation shown in 'Category' stands for the three major subontologies comprising GO: CC for 'cellular component', BP for 'biological process', and MF for 'molecular function'.

4.4 Conclusion and Discussion

DNA methylation has an essential role in the modulation of gene transcription in eukaryotic species, particularly in vertebrates (Suzuki and Bird, 2008, Antequera and Bird, 1993a, Attwood et al., 2002, Bennetzen et al., 1994, Zilberman and Henikoff, 2007). While several studies have explored the relationship between regulation of gene transcription by DNA methylation and the CpG content of gene promoters (Boyes and Bird, 1992, Robinson et al., 2004, Hsieh, 1994, Weber, 2007), they have all involved limited datasets or analysis only of the human genome. Moreover, the detailed biological roles of CgG dinucleotides in promoter regions and its impact degree on gene transcription are still unknown. The current study is the first comprehensive assessment of the DNA methylation system and its impact on gene transcription in 10 model eukaryotic (including both higher vertebrate, lower vertebrate, invertebrate and plant) species.

These analyses revealed that the genome-wide distribution patterns of GC content and CpG dinucleotides vary dramatically between higher vertebrates and lower vertebrate/invertebrates/plant. In higher vertebrates, both the GC content and CpG dinucleotides were consistently enriched in functional regions of the genome, particularly in promoter regions, compared with putative 'non-functional' regions including introns and intergenic sequence. This pattern may be explained by the following two observations. First, methylated have a higher probability than unmethylated cytosines to be converted to thymine over evolutionary time (Ehrlich et al., 1982, Kerry Lee, 2001). Second, nearly all CpG sites from non-functional sequences are completely methylated in vertebrate species. Functional constrains within

genic regions would limit the frequent of such mutations. However, I did not detect this pattern for lower vertebrate, invertebrate or plant species; instead, CpG dinucleotides were enriched across all regions of the genome, suggesting that there are similar levels of functional constraint in both 'functional' and putative 'non-functional' regions of the genome.

Focussing on gene promoters I discovered that far from being randomly distributed in gene promoters, CpGs dinucleotides were consistently showed a bimodal distribution pattern in each higher vertebrate species. The previously defined 'CpG rich' promoters (HCP) and 'CpG poor' promoters (LCP) could be observed in all six higher vertebrate species, but not in lower vertebrates, invertebrates or plant. For both groups, CpGs were concentred in core and proximal promoter regions. Furthermore, the classification of genes into HCP or LCP groups was highly conserved among the homologous genes of the six higher vertebrate species. Indeed, the level of conservation of promoter sequences between species could be used to accurately reconstruct the evolutionary relationships between species. All of these observations led us to conclude that the DNA methylation system is highly conserved among higher vertebrate species and to further explore a role for the distribution of CpG dinucleotides within this system.

DNA methylation of CpGs within both HCP and LCP promoters of the human genome is non-random; the level of methylation across the length of the promoter shows a u-shaped distribution, with the lowest levels corresponding with the core promoter regions. This distribution is likely to facilitate transcription initiation, while the increased methylation level in the proximal and distal promoter regions could modulate transcription by impeding the binding of transcription factors. Methylation, specifically

in the core and proximal promoter regions, negatively regulated the gene expression level across multiple human cell lines and tissues. This could be explained by the physical distribution of protein binding sites in promoter regions; the binding sites for RNA polymerase and most essential transcription factors are located in the core and proximal promoter regions, while only a few additional transcription factor binding sites are located in the distal promoter region (>250 bp upstream of the TSS).

I discovered distinct characteristics of HCP and LCP promoters that ultimately relate to their underlying biological functions. The level of CpG methylation was consistently higher within LCP compared with HCP promoters. Methylation levels of CpGs within the same promoter were highly correlated among different cell types or tissues, particularly for two CpGs located in close proximity. However, methylation levels within HCP promoters displayed larger variation compared with the CpG sites within LCP promoters. These differences in the patterns of DNA methylation between the two classes of promoter were reflected in different patterns of gene expression. In particular, I discovered that 94% of annotated housekeeping genes in a comprehensive database contained HCP promoters, confirming previous reports of HCP promoters being more frequently associated with 'housekeeping genes' expressed in a large number of tissues and LCP promoters being associated with 'tissue-specific' genes. For genes with HCP promoters, the DNA methylation system regulates the expression level in a number of tissues, while for genes with LCP promoters, the DNA methylation system provides a functional 'on-off' switch to determined whether the gene is expressed or not. Furthermore, I used genome-wide GO term functional annotations to confirm the functional distinctions between genes with HCP versus LCP promoters reported elsewhere (Ponger et al., 2001, Larsen et al., 1992, Saxonov et al., 2006, Weber, 2007),

but more importantly, I have shown that this relationship is conserved among all 6 model vertebrate species. I can infer that these patterns are determined by the highly conserved regulatory mechanism of DNA methylation.

4.5 Reference

ANTEQUERA, F. & BIRD, A. (1993) Number of CpG islands and genes in human and mouse. *Proc. Natl. Acad. Sci. USA,* 90**,** 11995-11999.

ATTWOOD, J. T., YUNG, R. L. & RICHARDSON, B. C. (2002) DNA methylation and the regulation of gene transcription. *Cellular and Molecular Life Sciences,* 59**,** 241-257.

BIRD, A. (2002) DNA methylation patterns and epigenetic memory. *Genes & Development,* 16**,** 6-21.

BIRD, A. P. (1986) CpG-rich islands and the function of DNA methylation. *Nature,* 321**,** 209-213.

BONAZZI, V. F., NANCARROW, D. J., STARK, M. S., MOSER, R. J., BOYLE, G. M., AOUDE, L. G., SCHMIDT, C. & HAYWARD, N. K. (2011) Cross-Platform Array Screening Identifies COL1A2, THBS1, TNFRSF10D and UCHL1 as Genes Frequently Silenced by Methylation in Melanoma. *PLoS ONE,* 6**,** e26121.

BOYES, J. & BIRD, A. (1992) Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. *EMBO J. ,* 11**,** 327-333.

CHAN, S. W. (2004) RNA silencing genes control de novo DNA methylation. *Science,* 303**,** 1336.

CHAN, S. W., HENDERSON, I. R. & JACOBSEN, S. E. (2005) Gardening the genome: DNA methylation in Arabidopsis thaliana. *Nature Rev. Genet.,* 6**,** 351-360.

CHARI, R., COE, B., VUCIC, E., LOCKWOOD, W. & LAM, W. (2011) An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology,* 4**,** 67.

CHEN, Z.-X. & RIGGS, A. D. (2011) DNA Methylation and Demethylation in Mammals. *Journal of Biological Chemistry,* 286**,** 18347-18353.

DAY, J. J. & SWEATT, J. D. (2010) DNA methylation and memory formation. *Nat Neurosci,* 13**,** 1319-1323.

EHRLICH, M., GAMA-SOSA, M. A., HUANG, L.-H., MIDGETT, R. M., KUO, K. C., MCCUNE, R. A. & GEHRKE, C. (1982) Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. *Nucleic Acids Research,* 10**,** 2709-2721.

FAHRNER, J. A., EGUCHI, S., HERMAN, J. G. & BAYLIN, S. B. (2002) Dependence of Histone Modifications and Gene Expression on DNA Hypermethylation in Cancer. *Cancer Research,* 62**,** 7213-7218.

GARDINER-GARDNER, M. & FROMMER, M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.,* 196**,** 261-282.

GEHRING, M. & HENIKOFF, S. (2007) DNA methylation dynamics in plant genomes. *Biochim. Biophys. Acta,* 1769**,** 276-286.

GEIMAN, T. M. & ROBERTSON, K. D. (2002) Chromatin remodeling, histone modifications, and DNA methylation - How does it all fit together? *Journal of Cellular Biochemistry,* 87**,** 117-125.

GLASS, J. L. (2007) CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res.,* 35**,** 6798-6807.
GOLL, M. G. & BESTOR, T. H. (2005) Eukaryotic cytosine methyltransferases. *Annual Review of Biochemistry.*
GOWHER, H., LEISMANN, O. & JELTSCH, A. (2000) DNA of Drosophila melanogaster contains 5-methylcytosine. *EMBO J,* 19**,** 6918-6923.
HE, X.-J., CHEN, T. & ZHU, J.-K. (2011) Regulation and function of DNA methylation in plants and animals. *Cell Res,* 21**,** 442-465.
HEDGES, S. B. (2002) The origin and evolution of model organisms. *Nat Rev Genet,* 3**,** 838-849.
HENDERSON, I. R. & JACOBSEN, S. E. (2007) Epigenetic inheritance in plants. *Nature,* 447**,** 418-424.
HERMAN, J. G. & BAYLIN, S. B. (2003) Mechanisms of disease: Gene silencing in cancer in association with promoter hypermethylation. *New England Journal of Medicine,* 349**,** 2042-2054.
HOLLIDAY, R. & PUGH, J. E. (1975) DNA Modification Mechanisms and Gene Activity during Development. *Science,* 187**,** 226-232.
HSIEH, C. L. (1994) Dependence of transcriptional repression on CpG methylation density. *Molecular and Cellular Biology,* 14**,** 5487-5494.
JAENISCH, R. & BIRD, A. (2003) Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. *Nat Genet,* 33**,** 245-254.
KERRY LEE, T. (2001) Methylated Cytosine and the Brain: A New Base for Neuroscience. *Neuron,* 30**,** 649-652.
LARSEN, F., GUNDERSEN, G., LOPEZ, R. & PRYDZ, H. (1992) CPG ISLANDS AS GENE MARKERS IN THE HUMAN GENOME. *Genomics,* 13**,** 1095-1107.
LI, E. (2002) Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet,* 3**,** 662-673.
LISTER, R., PELIZZOLA, M., DOWEN, R. H., HAWKINS, R. D., HON, G., TONTI-FILIPPINI, J., NERY, J. R., LEE, L., YE, Z., NGO, Q.-M., EDSALL, L., ANTOSIEWICZ-BOURGET, J., STEWART, R., RUOTTI, V., MILLAR, A. H., THOMSON, J. A., REN, B. & ECKER, J. R. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature,* 462**,** 315-322.
LOUDIN, M. G., WANG, J., EASTWOOD LEUNG, H. C., GURUSIDDAPPA, S., MEYER, J., CONDOS, G., MORRISON, D., TSIMELZON, A., DEVIDAS, M., HEEREMA, N. A., CARROLL, A. J., PLON, S. E., HUNGER, S. P., BASSO, G., PESSION, A., BHOJWANI, D., CARROLL, W. L. & RABIN, K. R. (2011) Genomic profiling in Down syndrome acute lymphoblastic leukemia identifies histone gene deletions associated with altered methylation profiles. *Leukemia,* 25**,** 1555-1563.
METTE, M. F., AUFSATZ, W., VAN DER WINDEN, J., MATZKE, M. A. & MATZKE, A. J. (2000) Transcriptional silencing and promoter methylation triggered by double-stranded RNA. *EMBO J.,* 19**,** 5194-5201.
MONTERO, L. M. (1992) The distribution of 5-methylcytosine in the nuclear genome of plants. *Nucleic Acids Res.,* 20**,** 3207-3210.
PALMER, L. E. (2003) Maize genome sequencing by methylation filtration. *Science,* 302**,** 2115-2117.

PARLE-MCDERMOTT, A. & HARRISON, A. (2011) DNA methylation: a timeline of methods and applications. *Frontiers in Genetics,* 2.

PATRA, S., PATRA, A., RIZZI, F., GHOSH, T. & BETTUZZI, S. (2008) Demethylation of (Cytosine-5-C-methyl) DNA and regulation of transcription in the epigenetic pathways of cancer development. *Cancer and Metastasis Reviews,* 27**,** 315-334.

PONGER, L., DURET, L. & MOUCHIROUD, D. (2001) Determinants of CpG Islands: Expression in Early Embryo and Isochore Structure. *Genome Research,* 11**,** 1854-1860.

RIGGS, A. D. (1975) X-INACTIVATION, DIFFERENTIATION, AND DNA METHYLATION. *Cytogenetics and Cell Genetics,* 14**,** 9-25.

ROBERTSON, K. D. (2005) DNA methylation and human disease. *Nature Rev. Genet.,* 6**,** 597-610.

ROBINSON, P. N., B 枚 HME, U., LOPEZ, R., MUNDLOS, S. & N 眉 RNBERG, P. (2004) Gene-Ontology analysis reveals association of tissue-specific 5 鈥?CpG-island genes with development and embryogenesis. *Human Molecular Genetics,* 13**,** 1969-1978.

ROLLINS, R. A. (2006) Large-scale structure of genomic methylation patterns. *Genome Res.,* 16**,** 157-163.

SAXONOV, S., BERG, P. & BRUTLAG, D. L. (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 1412-1417.

SUZUKI, M. M. & BIRD, A. (2008) DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet,* 9**,** 465-476.

WEBER, M. (2007) Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genet.,* 39**,** 457-466.

ZHU, B. & REINBERG, D. (2011) Epigenetic inheritance: Uncontested? *Cell Res,* 21**,** 435-441.

ZHU, J., HE, F., SONG, S., WANG, J. & YU, J. (2008) How many human genes can be defined as housekeeping with current expression data? *BMC Genomics,* 9**,** 172.

ZILBERMAN, D., GEHRING, M., TRAN, R. K., BALLINGER, T. & HENIKOFF, S. (2007) Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet,* 39**,** 61-69.

Part III

Several cooperated researches of the linkage disequilibrium based association study and gene transcription regulation mechanism analysis

This part presents several other research projects which I have involved during the last 4 years. In these studies, I am a co-author and my contributions to the studies vary as reflected by my authorship. This chapter summarizes these projects and my contribution to them.

1. Inferring Linkage Disequilibrium (LD) in case-control samples

1.1 Related publication

Wang M, Jia T, **Jiang N**, Wang L, Hu X, Luo Z. (2010) Inferring linkage disequilibrium from non-random samples. *BMC Genomics* **11**: 328-340.

Please see Appendix IV for a copy of this article as it appeared in print.

1.2 Summary

Fast advancement in DNA sequencing techniques has greatly facilitated a new tidal wave of genome wide association study (GWAS) to detect subtle genetic polymorphisms that underlie phenotypic variation of complex polygenic traits in humans, plants and animals. Obviously, adequate inference of the LD is the vital basis of efficient and reliable prediction of genetic association. Approaches widely implemented in LD estimations require samples to be randomly selected, which, however, are often ignored and thus raise the general question to the LD community of how the non-random collected samples affect statistical inference of the coefficient of Linkage Disequilibrium.

We proposed a new likelihood-based method for estimating LD using a sample un-randomly selected from the segregating population. Simulation study was conducted to mimic generation of samples with various degrees of non-randomness from the simulated populations. Our approach outperformed its rivals in adequately estimating the disequilibrium parameters in such sampling schemes. In analyzing a 'case and control' sample with β-thalassemia, this novel approach revealed robustness to non-random sampling in contrast to two commonly used methods.

Through a comprehensive simulation study and analysis of a real dataset, we have demonstrated the robustness of the proposed approach to non-randomness in sampling schemes and the significant improvement of the method to provide accurate estimates of the disequilibrium parameter. This approach provided a route to improve statistical reliability in association mapping.

The author assisted with the FORTRAN computer program and data analysis.

2. A powerful statistical method for genetic association studies using case-control samples

2.1 Related publication

This paper has been written up and submitted to BMC Genomics.

2.2 Summary

Genome wide association studies (GWAS) offer an unprecedented opportunity to detect genetic polymorphisms that underlie phenotypic variation of complex traits in humans, plants and animals. The theoretical basis of GWAS analyses is virtually estimation of linkage disequilibrium between any polymorphic locus and a putative trait locus. However, most methods widely implemented for such analyses are vulnerable to several key demographic factors and deliver a poor statistical power for testing for genuine associations but a high rate of false positives.

In this project, we present a novel model to formulate genotypic distribution in terms of LD and other population genetic parameters in any non-random population based samples including cases and controls. A likelihood-based statistical approach is developed to infer these parameters from such samples. The method was explored thoroughly through intensive computer study and analysis of recently published large case and control datasets of Parkinson's disease. The new method, to our best knowledge, is the first parametric statistical framework suitable to model and to infer LD between DNA marker and trait loci using the samples which are non-randomly collected from the segregating population under study. It provides adequate prediction

of LD and other model parameters, an ease integration of samples from multiple genetically divergent populations and the flexibility to incorporate various covariates into the analysis.

Both real data analysis and intensive computer simulation study illuminate that the newly developed method confers significantly improved statistical power for detecting the associations and robustness to difficulties stemmed from non-randomly sampling and genetic structures when compared to the nonparametric trend tests. The new method detected 44 SNPs within 25 chromosomal regions of size < 1Mb in significant association with the disease trait but only 6 SNPs in two of these regions were detected by the trend tests. It discovered two additional SNPs located 1.18 Mb and 0.18 Mb from the PD candidates, FGF20 and PARK8, without invoking risk of false positive.

The author assisted with the FORTRAN computer program and data analysis.

3. Genetic dissection of agronomic and morphologic traits in highly structured populations of barley cultivars

3.1 Related publication

Wang M, **Jiang N**, Jia T, Leach L, Cockram J, Waugh R, Ramsay L, Thomas B, Luo Z. (2012) Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *TAG Theoretical and Applied Genetics,* 124**,** 233-246.

Please see Appendix V for a copy of this article as it appeared in print.

3.2 Summary

Whole genome association mapping has recently become a powerful strategy for exploring traits of agricultural importance in higher plants, especially crops. In this project, we implement a series of association studies for 32 morphologic and 10 agronomic traits in a collection of 615 barley cultivars genotyped by genome-wide polymorphisms from a recently developed barley oligonucleotide pool assay. Strong population stratification effect related to mixed sampling based on seasonal growth habit and ear row number is present in this barley collection. Comparison of seven statistical methods in a genome-wide scan for significant associations with or without correction for confounding by population stratification, revealed that in reducing false positive rates while maintaining statistical power, a mixed linear model solution outperforms genomic control, structured association, stepwise regression control and principal components adjustment. Our analysis reports significant associations for sixteen morphologic and nine agronomic traits and demonstrates the power and feasibility of applying GWAS to explore complex traits in highly structured plant samples.

The author assisted with the data analysis and presentation of the manuscript.

4. Investigation of gene expression regulatory mechanisms in *Saccharomyces cerevisiae* (budding yeast) genome

4.1 Related publication

The paper based on this study has been written up and submitted to the journal of 'molecular biology and evolution'. In this project, the author analyzed both of the RNA-

seq data and expression microarray data, and performed the statistical analyses. The author also contributed to write the manuscript.

4.2 Summary

Sense and anti-sense transcripts (S/AS) are the RNAs that are complementary each other to a various extent and can be either protein-coding or non-protein-coding. The presence of S/AS has been identified as one of common and key structure features in the genomes of both prokaryotes and eukaryotes such as human, mouse, *Drosophila*, *C. elegans*, chicken, rat, cow, zebrafish, yeast, *Arabidopsis*, rice etc. According to their transcription orientation and extent of sequence overlap, S/AS can be classified into three major groups: convergent S/AS where the gene pairs overlap at the 3' ends, divergent S/AS where the gene pairs overlap at the 5' end, and consistent S/AS where the gene pairs overlap and transcript at the same direction. Distribution of these different types of S/AS shows the species specific, for example, the convergent S/AS is prevalent in *Drosophila* and *C. elegans* but not in Human and mouse.

It has been widely supported that the structural characteristics of the S/AS gene pairs confers a natural and important mechanism of regulation of their expression. In 2009, Guell et al. analyzed one of the smallest self-replicating organisms, *Mycoplasma pneumonia*, and found highly frequent antisense transcripts and clusters of operons into transcriptional units. In 2006, Hongay et al. discovered that meiotic entry in *Saccharomyces cerevisiae* is controlled by antisense regulation of *IME4*, a gene required for initiating meiosis. The data reported reveals a transcription interference mechanism which is in sharp contrast to either RNAi in other eukaryotes and the interference between of the sense and antisense transcripts in this system acts only in *cis*

but not in *trans.* By combining a global transcriptome analysis and expression profile analysis in mouse, Katayama et al. presented experimental evidence that RNAi mediated perturbation of an antisense RNA can alter the expression of sense messenger RNA in *cis,* in particular convergent pairs of S/AS show anti-regulation in expression in 2005. Several models have been proposed to explain the interference in expression between S/AS RNAs. The most prominent is the collision model in which RNA synthesis from one DNA strand might clash with transcription from the other strand. According to this model, transcription occurs in only one direction at a given time, and active antisense transcription would suppress sense RNA transcription. This model was proposed using the atomic force microscopy data in *E. coli* and was tested on a pair of convergent genes in the budding yeast.

All the studies aforementioned and reported so far on the S/AS transcriptional interference have been focused on only a few particular gene pairs and limited to their cis regulation. Recently popularizing RNA sequencing techniques enable to explore significance of the transcriptional interference at the genome scale. Several key questions remain unclear in regard to the structure dependent regulation of gene expression. For example, to what extent the transcriptional interference can be mediated in co-regulating expression of the S/AS pairs? What are the functional sequences that govern the co-transcriptional regulation? Whether the co-expression can be in *trans* in addition to *cis*?

To tackle these questions, we focused here on the convergent gene pairs with overlapping 3'UTRs in the *Saccharomyces cerevisiae* genome to check the change pattern in cell-cycle and responds for different nutrient environments. All of the results confirmed the bioinformatics conclusion that the 3'UTR convergent overlapped genes

have a negative regulation *in cis*. Furthermore, we found overlapped 3'UTRs cause a noticeable negative effect from other locus. It proved that *S. cerevisiae* did conserve a regulation machine between convergent genes with overlapped 3'UTR and this machine could affect in both *cis* and *trans*, which was similar with the ancient RNA-depended regulation described in virus and bacteria. Comparing the transcriptome with other organisms, we further revealed the organisms had a trend to share more overlapping UTRs when they were more primeval. It indicated that UTR overlapping could take a regulation function rather than a strategy to narrow the genome.

Upon the evidence above, we can suppose the regulation function in 3' UTR overlapped units of yeast genes maybe an ancient regulation machine conserved from more ancient organisms, which was thought to be a vital and evolutionarily ancient component of genetic regulation as the early stage of small RNA molecule forms.

Chapter V: Final conclusion and summary of this project

With the development of high-throughput technologies (such as microarray), the characterization and property of multi-levels 'omics' information has several important consequences in human and model organisms. First, microarray helps to accurately estimate gene transcription abundances by removing the influence of weak hybridization introduced by a mismatch between target and probe. Furthermore, it can also quickly and conveniently extract the genome-wide information in protein and DNA methylation levels for almost all model species. This is very valuable because it allows better interpretation of the fundamental, natural process through which organisms maintain their biological behaviours and life. Third, it enables one to investigate and identify the genetic variance among different samples. And, the genome-wide characterization of genetic diversity permits identification of many genetic markers that can be used in the physical mapping, linkage analysis and population studies. Additionally, analysis of genetic variability is of considerable significance for identification of genes in complex diseases, drug discovery and pharmacogenetics. Throughout the whole thesis, I comprehensively utilize the 'omics' information from different types of microarray; and mainly focus on the statistical methods for association analysis and the exploration of gene transcriptional regulation mechanism. I have divided this thesis into two independent parts: part I is a methodology research for genome-wide association analysis of complex traits; part II is to explore the DNA methylation mediated regulation of gene transcription in vertebrates.

The research in part I of this thesis introduces a linkage-disequilibrium based statistical algorithm for the genetic dissection of complex traits, particularly in structured populations. Currently, dissecting polygenic variation of complex traits at molecular level is one of the most challenging tasks in the era of functional genomics. The past two decades has witnessed a rapid development in both biological and computational technology, including genome sequences, high density marker maps, high-throughput genotyping methods and new statistical models, all of which greatly facilitate the progress of the reconstruction of genetic architecture for complex traits. Capitalizing on the achievements in both conventional experimental designs and the power of statistical methods, association study is capable of high precision and resolution for illuminating the complex pattern of relationship between genotypes and phenotypes and for understanding the role of complex genetics. However, a fundamental problem in association study is the existence of population structure because it can lead to 'spurious' associations, even within relatively homogeneous populations. Thus, I develop a new approach (**Method 1**) to control the spurious associations from structured populations by incorporating a control marker into the linear regression model. This new approach (**method 1**) has been compared to the original linear regression model which was designed for association study (**method 2**) and multiple regression model (**method 3**). I have explored the different methods from several aspects in both simulation study and real data analysis. Both **method 1** and **method 3** can successfully remove the spurious LD and also maintain the statistical power to identify the real associations at a relatively high level. But, the performance of the multiple regression model (**method 3**) heavily depends on the accuracy of prediction of the population structure and on accurate allocation of individuals' membership to the constituent

populations. Any bias in the structure prediction and uncertainty in the membership allocation can lead to severe consequences on its analytical efficiency. However, the new approach avoids the need for population structure prediction. Therefore, it is a convenient and powerful association analysis method for structured populations.

Here, a number of possibilities still exist for extending the present LD based method of association analysis to increase its general applicability and dissection power. For example, this new method presented here is appropriate for dissecting association for no more than two admixed sub-populations. In the future, it will be useful to extend this method for the admixed population which involves more than two genetically divergent populations. But, this improvement requires the haplotype frequency information between test marker and control marker, which is not available right now. With the development of next-generation sequencing technology, the haplotype information will be available very soon and the present method will be extended at that time.

In part II of my thesis, I utilize several levels of 'omics' information, included genomics, transcriptomics, genome-wide DNA methylation and annotation profiles, to explore the roles of methylation in gene transcriptional regulation, particularly in vertebrate species. DNA methylation in the genome plays a fundamental role in the regulation of gene expression levels and is widespread among eukaryotic species, particularly in vertebrates. Previous studies have revealed a consistent 'global' methylation pattern in higher vertebrate species, in which the CpG sites are completely methylated genome-wide, except in promoter regions. Meanwhile, another striking feature of the CpG dinucleotides in the human genome is that the CpG dinucleotides are obviously depleted genome-wide, but relatively enriched in promoter regions. Thus, one of the key open

questions in understanding the regulatory mechanisms of the DNA methylation system in vertebrates must be related to the distribution and roles of CpG sites in promoter regions and the variety of DNA methylation levels of CpGs in the promoters. In this work, I comprehensively examined and compared the distribution of CpG sites within 10 well-studied eukaryotic model organisms. And then, I investigated the relationships between genome wide DNA methylation profiles and gene expression profiles across multiple human tissues. Third, I used the genome-wide gene ontology information to analyze the conserved association between GO terms and the CpG distribution in promoter regions among 6 higher vertebrates. The analysis revealed two distinct methylation patterns for human gene promoters, involving genes with distinct distributions of CpG dinucleotides within the promoter region. Comparative analysis with five other higher vertebrates revealed that the primary regulatory role of DNA methylation system is highly conserved in higher vertebrates. Ideally, the next step of this analysis would be to investigate the new strategy that how to efficiently integrate the multi-levels of 'omics' information together. All of these analyses will bring us closer to a comprehensive understanding of the mechanisms of gene transcription regulation, which is essential for both biological and medical research.

Along with acquiring more comprehensive genetic knowledge through the above described research, I have gained plenty of sophisticated training and practice in both statistics and computational programming. In addition to using common statistical tools, e.g. minitab, SPSS and STATA, I am also professional in programming statistical methods and simulating statistical ideas in Fortran and R languages. These experiences strengthen my ability to develop new statistical methods and handle various kinds of data in real data analysis during my further career.

Appendices

Appendix I: Druka A, Potokina E, Luo Z, Jiang N, Chen X, Kearsey M, Waugh R (2010) Expression quantitative trait loci analysis in plants. *Plant Biotechnology Journal*, 8(1):10-27.

Plant Biotechnology Journal

Plant Biotechnology Journal (2010) **8**, pp. 10–27 doi: 10.1111/j.1467-7652.2009.00460.x

Review article

Expression quantitative trait loci analysis in plants

Arnis Druka[1], Elena Potokina[2], Zewei Luo[3], Ning Jiang[3], Xinwei Chen[1], Mike Kearsey[3] and Robbie Waugh[1,*]

[1]*Genetics, Scottish Crop Research Institute, Invergowrie, Dundee, UK*
[2]*Vavilov All-Russian Institute of Plant Industry, St Petersburg, Russia*
[3]*School of Biosciences, The University of Birmingham, Birmingham, UK*

Received 27 August 2009;
revised 24 September 2009;
accepted 29 September 2009.
**Correspondence* (fax 44 1382 568587;
e-mail robbie.waugh@scri.ac.uk)

Summary

An expression Quantitative Trait Locus or eQTL is a chromosomal region that accounts for a proportion of the variation in abundance of a mRNA transcript observed between individuals in a genetic mapping population. A single gene can have one or multiple eQTLs. Large scale mRNA profiling technologies advanced genome-wide eQTL mapping in a diverse range of organisms allowing thousands of eQTLs to be detected in a single experiment. When combined with classical or trait QTLs, correlation analyses can directly suggest candidates for genes underlying these traits. Furthermore, eQTL mapping data enables genetic regulatory networks to be modelled and potentially provide a better understanding of the underlying phenotypic variation. The mRNA profiling data sets can also be used to infer the chromosomal positions of thousands of genes, an outcome that is particularly valuable for species with unsequenced genomes where the chromosomal location of the majority of genes remains unknown. In this review we focus on eQTL studies in plants, addressing conceptual and technical aspects that include experimental design, genetic polymorphism prediction and candidate gene identification.

Keywords: expression quantitative trait loci, transcript-level variation, genetical genomics, transcript-derived markers.

Introduction

Phenotypic differences among individuals are partly the result of sequence polymorphisms that produce altered (or absent) proteins and partly the result of qualitative and quantitative differences in gene expression that generate varying amounts of protein in a cell or tissue. While variation within coding sequences is largely immune to environmental stimuli, gene expression at the transcriptional level is frequently considered the opposite, providing variation in the location, timing and/or abundance of individual mRNA. However, such variation in mRNA abundance is not solely determined by environment—a large number of studies have shown that genotypic variation in regulatory sequences can have a profound effect on comparative levels of gene expression among alleles. In the majority of cases, levels of gene expression are equated with the steady-state abundance of individual mRNA transcripts that have been determined in a specific sample at a given point in time. Abundance information can be captured using a variety of techniques and at a range of scales ranging from quantitative reverse transcription polymerase chain reaction (RT-PCR) (Czechowski *et al.*, 2004), through DNA microarrays (Schena *et al.*, 1995) to massively parallel signature sequencing (MPSS) (Brenner *et al.*, 2000) currently facilitated by next generation sequencing (NGS) (Wall *et al.*, 2009). If transcript levels are measured across a population of plants, the recorded variation in mRNA transcript abundance for each gene may be treated as a heritable trait that can be subjected to statistical genetic analyses. This can, in turn, locate and identify the underlying genetic factors that control the observed variation. The terms genetical genomics (GG) or expression quantitative trait loci (eQTLs) (Jansen and Nap, 2001) have been coined to describe this type of analysis.

The first published large scale eQTL mapping experiments (in yeast and mouse) involved small experimental populations and mostly described the results of genetic mapping (Brem *et al.*, 2002; Schadt *et al.*, 2003). These were soon followed by more focused studies on specific

Appendix II: Jiang N, Leach L, Hu X, Potokina E, Jia T, Druka A, Waugh R, Kearsey M, Luo Z (2008) Methods for evaluating gene expression from Affymetrix microarray datasets. *BMC Bioinformatics*, 9(1): 284-293.

BMC Bioinformatics

BioMed Central

Research article

Open Access

Methods for evaluating gene expression from Affymetrix microarray datasets

Ning Jiang[†1], Lindsey J Leach[†1], Xiaohua Hu[3], Elena Potokina[1], Tianye Jia[1], Arnis Druka[2], Robbie Waugh[2], Michael J Kearsey[1] and Zewei W Luo[*1,3]

Address: [1]School of Biosciences, The University of Birmingham, Edgbaston Birmingham B15 2TT, England, UK, [2]Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK and [3]Institute of Biostatistics, Fudan University, Shanghai 200433, PR China

Email: Ning Jiang - nxj677@bham.ac.uk; Lindsey J Leach - ljl193@bham.ac.uk; Xiaohua Hu - xhhu@fudan.edu.cn; Elena Potokina - e.potokina@bham.ac.uk; Tianye Jia - txj663@bham.ac.uk; Arnis Druka - arnis.druka@scri.ac.uk; Robbie Waugh - robbie.waugh@scri.ac.uk; Michael J Kearsey - m.j.kearsey@bham.ac.uk; Zewei W Luo* - z.luo@bham.ac.uk

* Corresponding author †Equal contributors

Published: 17 June 2008

BMC Bioinformatics 2008, **9**:284 doi:10.1186/1471-2105-9-284

Received: 29 January 2008
Accepted: 17 June 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/284

Abstract

Background: Affymetrix high density oligonucleotide expression arrays are widely used across all fields of biological research for measuring genome-wide gene expression. An important step in processing oligonucleotide microarray data is to produce a single value for the gene expression level of an RNA transcript using one of a growing number of statistical methods. The challenge for the researcher is to decide on the most appropriate method to use to address a specific biological question with a given dataset. Although several research efforts have focused on assessing performance of a few methods in evaluating gene expression from RNA hybridization experiments with different datasets, the relative merits of the methods currently available in the literature for evaluating genome-wide gene expression from Affymetrix microarray data collected from real biological experiments remain actively debated.

Results: The present study reports a comprehensive survey of the performance of all seven commonly used methods in evaluating genome-wide gene expression from a well-designed experiment using Affymetrix microarrays. The experiment profiled eight genetically divergent barley cultivars each with three biological replicates. The dataset so obtained confers a balanced and idealized structure for the present analysis. The methods were evaluated on their sensitivity for detecting differentially expressed genes, reproducibility of expression values across replicates, and consistency in calling differentially expressed genes. The number of genes detected as differentially expressed among methods differed by a factor of two or more at a given false discovery rate (FDR) level. Moreover, we propose the use of genes containing single feature polymorphisms (SFPs) as an empirical test for comparison among methods for the ability to detect true differential gene expression on the basis that SFPs largely correspond to *cis*-acting expression regulators. The PDNN method demonstrated superiority over all other methods in every comparison, whilst the default Affymetrix MAS5.0 method was clearly inferior.

Conclusion: A comprehensive assessment of seven commonly used data extraction methods based on an extensive barley Affymetrix gene expression dataset has shown that the PDNN method has superior performance for the detection of differentially expressed genes.

Background

Affymetrix GeneChip microarrays are the most popular high density oligonucleotide gene expression arrays and have become an invaluable tool in genomics studies worldwide. Each gene on an Affymetrix microarray GeneChip is typically represented by a probe set consisting of 11 different pairs of 25-bp oligos covering features of the transcribed region of that gene. Each pair consists of a perfect match (PM) and a mismatch (MM) oligonucleotide. The PM probe exactly matches the sequence of a particular standard genotype, often one parent of a cross, while the MM differs in a single substitution in the central, 13th base. The MM probe is designed to distinguish noise caused by non-specific hybridization from the specific hybridization signal.

Affymetrix microarrays inevitably introduce many sources of variation [1]. Normalization procedures are essential to "correct" for systematic sources of variation of non-biological origin. Affymetrix microarray data are normalized in three steps: background correction, to adjust for hybridization effects unrelated to the interaction between probes and target DNA; normalization, to remove systematic errors and biases thereby allowing data to be compared from one array to another; summarization, combining the multiple probe intensities from a probe set to yield a single value for each gene that best represents the expression level of the RNA transcript. Numerous data extraction methods have been proposed in the literature to perform these crucial steps in processing Affymetrix oligonucleotide microarray data.

The first data extraction method provided as the Affymetrix default was the Average Difference (AD), a linear scale measure that relied upon the difference measure PM-MM to correct for non-specific binding. This measurement was superseded by the current standard MAS5.0 [2], which uses the more appropriate log scale and a robust Tukey Biweight averaging method. It was shown subsequently that one third of probe pairs consistently yield negative signals, showing that use of MM probes for detection of non-specific binding is unreliable [3,4]. In this respect, Irizarry et al. [5] developed the robust multi-array average (RMA) method based solely on PM values. Li and Wong [6] developed a statistical model for probe level data and their model based expression index (MBEI) has been developed into dChip, one of the most popular software approaches used today. Physical energy-based models have also been developed as an attempt to model the formation of DNA-RNA duplexes on oligonucleotide microarrays [7], most notably the positional dependent nearest neighbour (PDNN) model of Zhang et al. [8]. Following this idea, Wu et al. [9] developed the GCRMA method that attempts to combine the strengths of stochastic model based algorithms such as RMA with physical modelling of sequence information. The number of methods available continues to grow, yet there is no consensus as to which is the most appropriate and reliable method for a given application.

Calibration datasets derived from mixture experiments [10], spike-in studies and dilution series [3,5,11-14] have been an invaluable resource to develop and assess data extraction methodology. The advantage of these benchmarking datasets is that the expected outcome of expression analysis is known in advance and so alternative expression measures can be compared in terms of the expected features. This property has been exploited to develop a graphical tool for the evaluation and comparison of expression measures aimed at helping researchers to decipher the multitude of methods available [12,14].

Studies utilizing benchmark datasets have typically observed a large effect of the normalization method on the outcome of the expression analyses [15-17]. However, the performance of 'spike-in' experiments can be affected by sources of systematic variation and it is not clear how this might affect evaluation of different data extraction methods [15]. One alternative strategy involved assessing the gene expression between males and females at Y-chromosome linked genes as a true biological internal control [18]. In this study, the performance of the method was measured by recording how many differentially expressed Y-chromosome linked genes were detected between male and female samples. However, the general applicability of this kind of test is limited.

More recently, Harr and Schlotterer [15] introduced an alternative strategy to evaluate normalization methods by exploiting the existence of bacterial operons in which genes are expected to have highly correlated expression levels. This strategy effectively avoided the systematic biases inherent in the spike-in approach. However, the assumption that expression of operon member genes should be correlated can be violated, for example by internal promoters and/or overlapping regulatory elements [19]. It is increasingly evident that performance analyses using calibration datasets are not necessarily consistent with data from realistic biological studies [16,20], suggesting the need to consider real biological studies in an attempt to evaluate the relative merits of Affymetrix data extraction methods.

In this article we present a comparison of the influence of seven commonly used data extraction methods on the detection of differentially expressed genes using a genome-wide gene expression dataset from eight genetically divergent barley lines. The major challenge arising from the use of this dataset is that one has no *a priori* knowledge of which genes are differentially expressed. To

Table 1: Statistical analyses involved in the seven different methods for calculating gene expression.

Methods	Background Correction	Normalization	Core Statistical Analysis	References
AD	None	Invariant Set	Average difference	Affymetrix [2]
MAS5.0	Spatial effect and MM subtracted	Constant	Robust average (Tukey bi-weight)	Affymetrix [2]
MBEI (PM only)	None	Invariant Set	Multiplicative model	Li and Wong [5]
MBEI (PM-MM)	MM intensities are subtracted	Invariant Set	Multiplicative model	Li and Wong [5]
RMA	Global correction	Quantile	Robust linear model (median polish)	Irizarry et al. [4]
PDNN	Model is fitted accounting for background and specific signal	Quantile	Specific and non-specific binding effects are estimated using free energy model	Zhang et al. [7]
GCRMA	Based on probe sequence	Quantile	Robust linear model (median polish)	Wu et al. [8]

address this challenge we used a novel strategy based on genes in which we detected single feature polymorphisms (SFPs). SFPs are genetic polymorphisms in observed expression within one particular feature (oligonucleotide probe) of a probe set (11 PM and MM probes) on the array [21]. Using two barley 'Genetical Genomics' datasets we have previously shown that SFPs mainly represent expression differences that are the result of polymorphism in *cis*-acting regulators [22]. On this basis we propose that differential expression detected in SFP-containing genes is more likely to reflect true differential expression and so we use this as a criterion to assess the efficacy of the seven methods referred to above in the detection of differential gene expression.

Results

The present study implements seven methods commonly used in the literature to calculate expression indices from Affymetrix microarray gene expression data, which was collected from a well-designed genome-wide microarray hybridization experiment with eight genetically divergent barley cultivars. These methods are summarized in Table 1 and include Average Difference (AD), MAS5.0, MBEI (PM only), MBEI (PM-MM), RMA, PDNN and GCRMA. We explore various statistical properties of the methods in modelling and analyzing the microarray dataset. The findings are compared with those based on an independent dataset of Affymetrix genome-wide gene expression profiled on two divergent yeast strains.

Consistency of gene expression indices calculated from different methods

To explore the consistency of the 22,840 barley gene expression indices estimated from the seven different methods, we calculated Pearson's Product Moment Correlation coefficients in the expression estimates and the correlation analyses are summarized in Table 2. The corresponding results based on the yeast dataset are summarized in Table 4 [see Additional file 1]. The upper triangle in Table 2 contains the means and standard deviations of 24 correlation coefficients, r_{ijk} ($k = 1, 2, ..., 24$). r_{ijk} represents the correlation coefficient between 22,840 corresponding pairs of gene expression indices calculated by methods i and j from the k^{th} microarray sample. The lower triangle shows the overall correlation coefficients between all pairs of 22,840 gene expression indices calculated from methods i and j across all 24 samples (cultivars $m = 1,...,8$ × replicates $n = 1,...,3$). It is clear that the seven methods

Table 2: Pearson's Product Moment Correlation Coefficients among barley gene expression indices calculated from seven different methods.

Method	AD	MAS5.0	MBEI[1]	MBEI[2]	RMA	GCRMA	PDNN
AD	**0.991 ± 0.005**	0.975 ± 0.004	0.988 ± 0.001	0.985 ± 0.001	0.647 ± 0.007	0.791 ± 0.008	0.615 ± 0.007
MAS5.0	0.973	**0.984 ± 0.009**	0.961 ± 0.005	0.965 ± 0.003	0.619 ± 0.026	0.748 ± 0.024	0.583 ± 0.026
MBEI[1]	0.987	0.958	**0.990 ± 0.005**	0.988 ± 0.001	0.664 ± 0.011	0.797 ± 0.009	0.629 ± 0.008
MBEI[2]	0.985	0.963	0.988	**0.990 ± 0.005**	0.643 ± 0.006	0.774 ± 0.006	0.605 ± 0.006
RMA	0.647	0.616	0.662	0.643	**0.993 ± 0.003**	0.914 ± 0.002	0.939 ± 0.008
GCRMA	0.791	0.744	0.797	0.774	0.914	**0.992 ± 0.005**	0.923 ± 0.004
PDNN	0.614	0.581	0.628	0.604	0.940	0.923	**0.992 ± 0.005**

The upper triangle shows the mean and corresponding standard deviation of 24 correlation coefficients, r_{ijk} ($k = 1, 2, ..., 24$). r_{ijk} represents the correlation coefficient between 22,840 corresponding pairs of gene expression indices calculated by methods i and j from the k^{th} microarray sample. The diagonal cells show means and standard deviations of 24 correlation coefficients, r_{imn} ($n = 1, 2, 3$ and $m = 1, 2, ..., 8$). For $n = 1, 2, 3$, r_{imn} corresponds to three correlation coefficients calculated from three possible pairs of replicates for the m^{th} cultivar ($m = 1, 2, ..., 8$) using method i. The lower triangle shows the correlation coefficients between all pairs of 22,840 gene expression indices calculated from methods i and j across all $k = 24$ samples.

[1] MBEI PM only model

[2] MBEI PM-MM model

may be separated into two groups (AD, MAS5.0 and MBEI in one group and RMA, GCRMA and PDNN in the other) according to the correlation coefficients. The coefficient of correlation is greater than 90% within each of the groups but becomes less than 80% between the two groups. The same pattern of correlation in gene expression estimate between these seven methods was also recovered in the analysis of gene expression profiles on two yeast strains. Notably, all of the methods in the first group were based on use of both PM and MM values (with the exception of MBEI PM), while the methods in the second group were based on PM value only. However, the average correlation coefficient between MBEI PM and MBEI PM-MM was as high as 0.988, therefore the division of the seven methods into two groups was unlikely to be caused by using either the PM-MM model or the PM only model.

The diagonal elements in Table 2 represent means and standard deviations of correlation coefficients in gene expression indices between biological replicates. They show that MAS5.0 confers significantly lower correlations between replicates than the other methods (p-value < 10^{-5}, Mann-Whitney U-test), suggesting that the different methods have a profound effect that goes beyond the variance observed across the biological replicates, in support of previous findings [15,17].

We compared the ability of each method to calculate consistent gene expression values between biological replicates of a given barley variety using the intra-class correlation coefficients. The box plot in Figure 1a clearly shows the PDNN method gave a superior performance (largest mean and smallest standard deviation) over all of the other methods across all 22,840 genes (p-value < 0.0001, Mann-Whitney U-test), while the poorest performers were the GCRMA and MAS5.0 methods (p-value < 0.0001, Mann-Whitney U-test). The standard deviation obtained using the PDNN method was significantly lower than all other methods (Levene's test, p-value < 0.0001), except MBEI PM for which a similar standard deviation was obtained. In the analysis of the yeast dataset, the PDNN method also gave a superior performance over several of the other methods (Mann Whitney U-test, p-value < 0.05, with MBEI methods; p-value < 0.0001 with MAS5.0 and GCRMA methods) as shown in Figure 2a [see Additional file 2].

To explain the different performances of the methods illustrated above, we investigated the effect of each step in processing the microarray datasets on estimates of the expression indices in the barley dataset. We tested use of different background correction methods but the same normalization and summarization steps in estimating the genome-wide gene expression indices, and calculated the correlation coefficient for each pair-wise comparison of background correction methods. The correlation coefficients for the MAS5.0 and RMA methods based on different background corrections were greater than 99% and 90%, respectively, which is greater than the correlation between different methods (60%–80%). Therefore the background correction methods did not have a significant effect on the correlation between methods. Similarly, the correlations for the AD and MBEI-PM methods based on different normalization methods were greater than 97% and 99% respectively, showing the normalization methods did not have a detectable effect on the correlation between methods either (Table 5, [see Additional file 3]).

Efficiency in detecting differential gene expression

To compare the ability to detect differentially expressed genes among the barley varieties for the seven data extraction methods, our primary focus is sensitivity, defined as the total number of genes detected with significant differential expression at a given FDR level. Figures 1b and 2b [see Additional file 2] show the number of genes with significant differential expression called by the seven methods across a range of FDR levels, for the barley and yeast datasets respectively. The numbers of genes declared differentially expressed decreased for each method as the FDR level became more stringent; the best performer at every FDR level was the PDNN method and the worst two performers were GCRMA and MAS5.0. Across all FDR levels, there was marked variation among the seven methods in the number of genes detected as differentially expressed. In particular, PDNN detected 70% more differentially expressed genes than MAS5.0 at FDR 0.01 in both the barley and yeast datasets, and over twice as many genes at even more stringent FDR levels in the barley dataset.

The variation in FDR across the seven methods occurs for two reasons; firstly, variation in the number of genes detected significantly differentially expressed among the varieties and secondly, variation in the expected number of genes with detected significant differential expression when there is no real differential expression. Shedden et al. [16] have shown that different methods differ markedly in their tendency to produce outlier expression values and this is reflected in the thresholds required to achieve a specified proportion of false positive calls. Figures 1c and 2c [see Additional File 1] show how the p-value threshold required to achieve a given FDR value differs substantially among the seven methods, for both barley and yeast datasets respectively. Notably, Figures 1b and 1c and also Figures 2b and 2c [see Additional file 2] both illustrate exactly the same order of the seven methods, showing that calibration plays an important role in determining sensitivity in detecting differential gene expression.

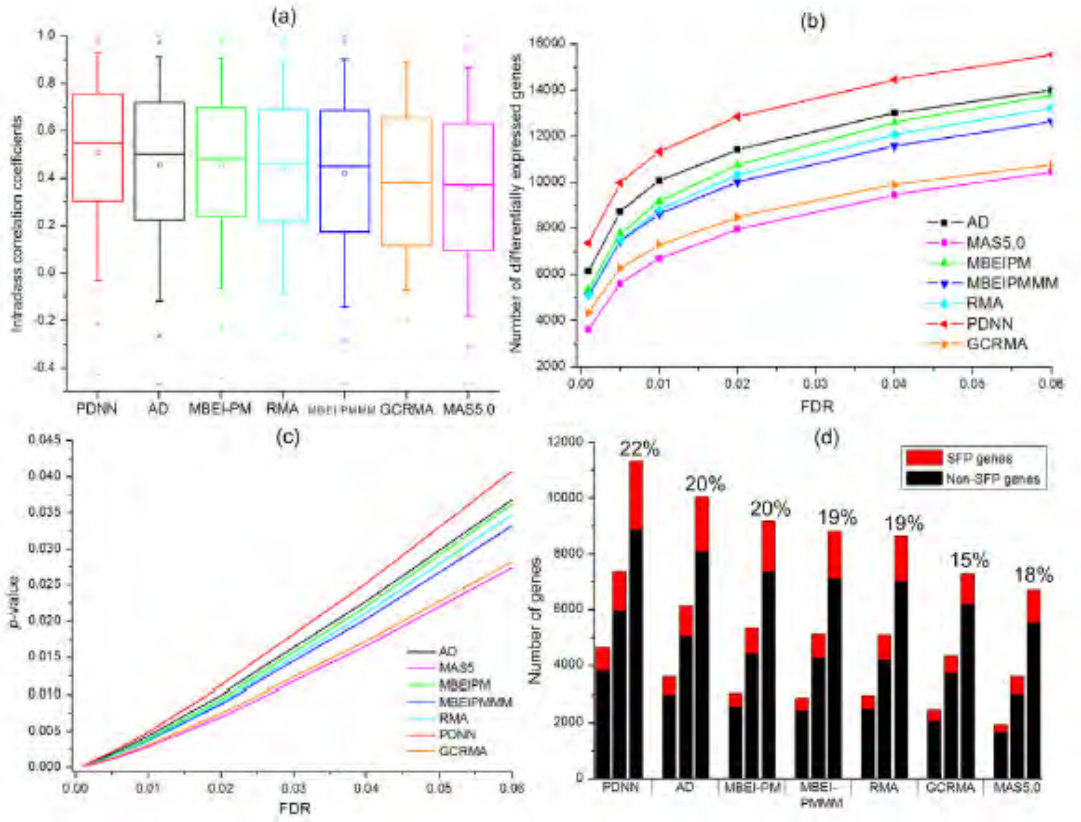


Figure 1
Statistical properties of estimated barley gene expression indices from seven data extraction methods. (a) Intraclass correlation coefficients between biological replicates of the estimated expression indices for 22,840 genes; (b) Sensitivity for detecting differentially expressed genes; (c) Calibration *p*-values across FDR levels; and (d) The number of differentially expressed SFP genes (red segment) and non-SFP genes (black segment). For each method the three columns from left to right correspond to FDR levels 0.0001, 0.001 and 0.01. The proportion of genes declared differentially expressed that showed SFP is illustrated for FDR level 0.01.

Mutual predictability among the seven methods

An important aspect in comparing the different methods would be to compare their ability to detect the same differentially expressed genes, their mutual predictability. Table 3 shows the pair-wise agreement between the methods for the identity of differentially expressed barley genes at FDR = 0.01. The MAS5.0 method shared the fewest calls with the other methods (for example 61 ± 5%), while PDNN had the strongest agreement with the other methods (93 ± 3%); notably, the order of the seven methods for the pair-wise comparison from the strongest to the weakest was PDNN, AD, MBEI (PM only), RMA, MBEI (PM-MM), GCRMA, MAS5.0, consistent with the order shown in Figure 1a. However, all pair-wise comparisons between methods showed that all methods detected differentially expressed genes not detected by the other methods. This suggests that all methods contribute unique but important information on differential gene expression. Interestingly, methods calling similar genes as differentially expressed did not exhibit greater expression similarity. For example, the gene expression index calculated from the MAS5.0 method is highly correlated with the MBEI PM method (r = 0.958), although the MAS5.0 method only detects 57% of the genes called by MBEI PM

Table 3: Mutual predictability of the number of barley genes declared differentially expressed from seven data extraction methods.

Methods	AD	MAS5.0	MBEI[1]	MBEI[2]	RMA	PDNN	GCRMA
AD	10066	5984(89%)	7951(87%)	8030(93%)	7566(86%)	9269(82%)	6610(90%)
MAS5.0	5984(59%)	6716	5257(57%)	5289(61%)	5614(64%)	6243(55%)	5067(69%)
MBEI[1]	7951(79%)	5257(78%)	9185	7674(89%)	6969(79%)	8206(72%)	6048(83%)
MBEI[2]	8030(80%)	5289(79%)	7674(84%)	8650	6744(76%)	7830(69%)	5946(81%)
RMA	7566(75%)	5614(84%)	6969(76%)	6744(78%)	8824	8419(74%)	6736(92%)
PDNN	9269(92%)	6243(93%)	8206(89%)	7830(91%)	8419(91%)	11339	6994(96%)
GCRMA	6610(66%)	5067(75%)	6048(66%)	5946(69%)	6736(76%)	6994(62%)	7310

The diagonal cells show the number of genes declared from each method respectively at FDR = 0.01. The upper and lower triangles show the numbers and percentages (in parentheses) of the genes declared by method *j* (*j* = 1st,...,7th column) and also by method *i* (*i* = 1st,...,7th row, $i \neq j$). For example, the 5984 genes in common to AD and MAS5.0 represent 89% of those detected by MAS5.0 but only 59% of those detected by AD.
[1] MBEI PM only model
[2] MBEI PM-MM model

at FDR = 0.01. On the other hand, the expression index from MAS5.0 has a much lower correlation (r = 0.581) with that from GCRMA even though the MAS5.0 method calls 75% of the genes called by GCRMA. The results of the yeast data analysis (Table 6, [see Additional file 4]) show exactly the same ordering of the seven methods as that obtained from the barley dataset.

An empirical Test for efficiency in predicting true differential gene expression

An important objective was to compare the ability of each method to identify genuine differential expression. To this end, we used a recently identified set of over 7000 barley genes containing single feature polymorphisms that largely represent gene expression markers (GEMs) corresponding to a combination of mainly *cis*-acting expression regulators but also *trans*-acting regulators [22]. On this basis, and in the absence of an expected outcome of the differential expression analysis, we propose that differential expression detected for SFP genes is more likely to reflect true differential expression than for genes that do not contain SFP. Using this criterion we compared each of the seven methods for their ability to detect differential gene expression in the SFP genes (Figure 1d) using the proportion of genes declared differentially expressed that showed SFP. The PDNN method outperformed all other methods (chi-square test, p-value < 0.0001 at FDR = 0.01), while the worst two methods were MAS5.0 and GCRMA (chi-square test, p-value < 0.0001 and p-value < 0.05 respectively at FDR = 0.01; moreover, the performance order from best to worst method matched the orders based on sensitivity, calibration and reproducibility (intra-class correlation) analyses. It should be noted that the SFP analysis does not involve any of the methods under investigation here for quantifying gene expression. Thus, the SFP prediction provides an independent source of information for assessing performance of the methods in detecting differentially expressed genes.

Conclusion

The development of pre-processing methods for Affymetrix oligonucleotide gene expression data has been an area of active research and has led to the availability of a large and growing toolbox of statistical methods for data extraction. This presents a significant challenge for a researcher wanting to identify the most appropriate method to analyze her/his datasets. The present study examined the effect of different data extraction methods on the detection of differentially expressed genes in a barley Affymetrix oligonucleotide microarray dataset. Seven commonly used data extraction methods were used exactly as recommended by their developers, providing a directly relevant comparison of the methods as they will be used in practice by the majority of users of the software, and thus avoiding the well-known over-training problem associated with calibration datasets. The analysis exploits an extensive genome-wide gene expression dataset from eight barley varieties showing extensive variation at phenotypic, transcriptional and genotypic levels. The presence of three replicates for each variety gave a perfectly balanced experimental design and ideal data structure for the main aims of the present research as well as a high power to detect differentially expressed genes by the analysis of variance.

It is clear from the present study that evaluation of the gene expression index is strongly affected by the data extraction method and this in turn has a strong influence on the ability to detect differential gene expression confidently. The seven commonly used methods can be divided into two groups according to the correlation structure in expression indices. Neither the use of different background correction nor normalization procedures could explain the marked variation in expression values estimated from the different methods, as shown previously [15]. Therefore the differences must be caused by the use of different statistical models to estimate the expression values.

Several studies have systematically compared different data extraction methods using tightly controlled calibration datasets, but in doing so, have restricted the comparison to limited amounts of data generated using a limited number of species and platforms [10,12,13]. On the one hand, use of calibration datasets simplifies the data modelling, but on the other hand it avoids the challenges involved in modelling real data involving a larger number of sources of uncontrolled variability. Different studies using Affymetrix spike-in experimental data have tended to produce inconsistent results [9,12,23], possibly due to hidden contaminates. Moreover, the results often conflict with those based on realistic biological datasets. For example, Rajagopalan [11] concluded that it is inadvisable to use the PM only model for microarray data analysis, whereas the current study has shown comparable performance between MBEI PM-MM and MBEI PM only models across all comparisons, and indeed, the PM only model has a superior performance in calculation of consistent gene expression estimates across replicates of a given barley variety (p-value < 0.0001, Mann-Whitney U-test).

The major statistical challenge in using real biological experimental datasets arises from the fact that one cannot know *a priori* whether or not a given gene is truly differentially expressed. Therefore in comparing the sensitivity of each of the seven methods to detect differential gene expression, care and attention must be paid to ensure that detected differences in sensitivity among methods are not due to other factors. The Benjamini and Hochberg [24] false discovery rate (FDR) was used here to control the detection of false positives in a way that was not biased in favour of any particular method.

The seven data extraction methods were explored from several angles, including sensitivity, reproducibility and mutual agreement for the identity of differentially expressed genes. Across a range of FDR levels, the PDNN method had the highest sensitivity to detect differentially expressed genes and this was directly related to the less stringent p-value threshold required by this method to declare differential expression for a given FDR level. This explains the excellent agreement observed for the differentially expressed genes with all of the other methods. The reproducibility of results from microarray experiments is a critical issue for data analysis methods. The seven data extraction methods showed varying sensitivities to the inherent biological variation expected within the system; the PDNN method produced the most consistent results across biological replicates, whilst MAS5.0 and GCRMA produced the poorest results.

In the absence of an expected outcome, detection of differential expression within those genes with single feature polymorphism was used to further assess the ability of each method to detect genuine differential gene expression. The set of differentially expressed genes identified by the PDNN method was significantly enriched for SFP genes compared to all other methods, reflecting the fact that the method incorporated the sequence information into its calculation of expression indices. The PDNN method may have the highest accuracy in detecting genuine differential gene expression compared to the other six data extraction methods. The GCRMA and MAS5.0 methods called only half the fraction of differentially expressed genes called by PDNN; however, their caution is unlikely to reflect improved prediction of genuine differentially expressed genes.

Taken together, all comparisons suggest that the PDNN method is superior to its rivals for the detection of differentially expressed genes in the current dataset. In contrast, Shedden et al. [16] showed using two datasets of gene expression profiled in human tissue samples that no single method could be identified with consistently superior performance. However, both GCRMA and MAS5.0 methods performed consistently poorly in comparison to rival methods, in agreement with the findings presented here. To assess the performance of the PDNN method in smaller and more statistically challenging biological datasets, we conducted the same analyses using a genome-wide Affymetrix dataset of gene expression profiled on two divergent yeast strains, each with four biological replicates. This analysis provided only a single degree of freedom for detecting differential gene expression between yeast strains, therefore we did not expect it to be as powerful as the barley data analysis. However, the results were remarkably similar to those obtained in the barley data analysis, further supporting the superiority of the PDNN method over its rivals in detecting differentially expressed genes.

We have only used a parametric ANOVA to detect differentially expressed genes. However, variation due to the use of different test statistics is smaller than variation due to different processing methods [16,17] so we expect these differences to be robust to the use of different statistical tests. The PDNN method identifies 70% more differentially expressed genes than MAS5.0, and moreover, gave a superior performance in all the analyses. Nevertheless, each and every method is expected to call one or more differentially expressed genes not called by the other methods. Therefore even the less sensitive methods may contribute to our understanding of which genes are differentially expressed.

The reason for superior performance of the PDNN method based on the present dataset may lie in its use of the free energy statistical model to detect both the specific

and non-specific bindings between probes and their corresponding target transcripts, which may accurately model the physical and chemical aspects of probe binding on Affymetrix microarray chips. This may be considered somewhat surprising given findings that positional dependent effects, but not interactions between bases that are physically close, add significant predictive power for specific signal probe effects [25].

The question arising naturally from the present analysis is that of which is the best method for analyzing Affymetrix gene expression data with a view to identifying differentially expressed genes. However, the present study has considered a selection of highly distinguished approaches for data extraction as applied to a barley genome-wide gene expression dataset and recognizes that a greater number of datasets from both controlled experiments and calibration data will be necessary to answer this question. The method chosen will depend on the particular scientific question the study is designed to address and the priorities involved. For example, given the high number of differentially expressed genes detected in a typical microarray experiment, specificity may be a higher priority than sensitivity and influence the method(s) chosen to analyse the results.

Methods

Barley RNA microarray data

The microarray data consisted of three biological replicates (n = 1,..., 3) from each of eight genetically divergent barley varieties (m = 1,...,8) known as Barke, Golden Promise, Haruna Nijo, Morex, Optic, OWB_D, OWB_R and Steptoe (k = 24 samples in total). Total mRNA was extracted from the plant leaves, and then hybridized to a Barley 1.0 Affymetrix microarray GeneChip, which consists of 22,840 probe sets (representing 22,840 genes or ORFs), at the Iowa State University transcriptomics facility. A distributed probe set format array was used to prevent potential local image contamination from completely destroying the data of an entire probe set (PM and MM).

The two yeast strains, whose gene expression data was analysed here, were the two haploid parental lines reported in our previous experimental analysis for the genetic dissection of quantitative trait loci affecting ethanol tolerance in budding yeast [26]. The strains were phenotypically divergent for major fermentation traits and cellular morphology characters. The genome-wide gene expression of the strains was profiled at a steady log-growth stage by using Affymetrix yeast 2.0 GeneChips, consisting of 5,814 probe sets. The microarray data consisted of four biological replicates (n = 1,...,4) from each of the two yeast strains (m = 1,2), a high performance strain designated *PHO*, and a low performance strain designated *PLO* (k = 8 samples in total).

Analysis methods

The raw signal intensities for each probe set (contained in the CEL files) were analysed by the seven most commonly used data extraction methods (AD, MAS5.0, MBEI PM-MM, MBEI PM only, RMA, PDNN and GCRMA) implemented in the R statistical environment. The relevant software was downloaded from the Bioconductor website [27] to produce the genome-wide gene expression indices.

Comparing the correlation coefficients between methods and between replicates

For the between method comparison k correlation coefficients, r_{ijk}, (k = 1, 2,..., 24 for the barley data and k = 1, 2,...,8 for the yeast data) were calculated between 22,840 (barley) or 5,814 (yeast) corresponding pairs of gene expression indices calculated by methods i and j from the k^{th} microarray sample. The within method correlation was calculated as $r_{imn'}$ corresponding to correlation coefficients calculated from each possible pairing of the n = 3 (barley) or n = 4 (yeast) replicates for the m^{th} barley cultivar (m = 1,...,8) or yeast strain (m = 1,2) using method i. The correlation coefficients (between methods and within method) were compared using the Mann-Whitney U-test.

Correlation between methods across all samples

Pearson's correlation coefficient was calculated between all pairs of 22,840 (barley) or 5,814 (yeast) gene expression indices calculated from methods i and j across all k = 24 (barley) or k = 8 (yeast) samples (m cultivars or strains × n biological replicates). The genome-wide gene expression indices and the correlation coefficients were then computed under the scenarios of changing a single step in the three-step normalization procedure (background correction or normalization) whilst maintaining the other two steps the same.

One-way ANOVA to detect differentially expressed genes

Each of 22,840 barley genes (5,814 yeast genes) was tested for differential expression by one-way analysis of variance by partitioning the total variation in gene expression level into variation between groups (8 barley varieties or 2 yeast strains), denoted by s_b^2 and the variation within groups (between replicates), denoted by s_w^2. The F value can then be calculated according to

$$F = \frac{S_b^2}{S_w^2} \quad (1)$$

and the associated p-value obtained. The false discovery rate (FDR) was controlled according to the standard method of Benjamini and Hochberg [24].

Intraclass correlation coefficient

For each of the 22,840 barley genes (5,814 yeast genes), we calculated the intraclass correlation coefficient (r) according to the standard method of Snedecor and Cochran [28] as

$$r = \left(s_b^2 - s_w^2\right) / \left\{s_b^2 + (n-1)s_w^2\right\} \quad (2)$$

for $n = 3$ (barley) or 4 (yeast) replicates.

Genes recording Single Feature Polymorphisms (SFPs)

We implemented the method proposed in our previous paper [22] to detect single feature polymorphisms (SFPs) that segregate between the 8 barley genotypes. We identified a total of 7340 gene specific SFPs in the barley genome, which are referred to here as the 'SFP genes'.

Authors' contributions

ZWL conceived of and designed the study. NJ analyzed the data and performed the statistical analyses. LJL contributed to the data analysis and wrote the manuscript. MJK and TJ participated in the design of the study and data analysis. XH provided the yeast dataset. EP, RW and AD were involved in generating the barley microarray data. All authors read and approved of the final manuscript.

Additional material

Additional file 1

Pearson's Product Moment Correlation Coefficients among yeast gene expression indices calculated from seven different methods.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-284-S1.doc]

Additional file 2

Statistical properties of estimated yeast gene expression indices from seven data extraction methods. (a) Intraclass correlation coefficients between biological replicates of the estimated expression indices for 5,814 genes; (b) Sensitivity for detecting differentially expressed genes; and (c) Calibration p-values across FDR levels.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-284-S2.pdf]

Additional file 3

Pair-wise Pearson correlation coefficients between all pairs of 24 sets (8 barley cultivars × 3 replicates) of 22,840 gene expression indices calculated from the MAS5.0 and RMA methods with different background correction steps and for the AD and MBEI methods with different normalization steps.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-284-S3.doc]

Additional file 4

Mutual predictability of the number of yeast genes declared differentially expressed from seven data extraction methods.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-284-S4.doc]

Acknowledgements

This study is supported by research grants from the BBSRC (RRAD11354) of the United Kingdom. ZWL is also supported by the National Natural Science Foundation (30430380) and Basic Research Program of China (2004CB518605). We would like to thank and acknowledge Andreas Graner, Alan Schulman, Peter Langridge, Kaz Sato, Pat Hayes and Tim Close for providing access to unpublished 'barley genomics community' datasets used in this manuscript.

References

1. Yang YH, Speed T: **Design issues for cDNA microarray experiments.** *Nat Rev Genet* 2002, **3**:579-588.
2. Affymetrix: **Affymetrix Statistical Algorithms Description Document.** [http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf].
3. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay Y, Antonellis K, Scherf U, Speed T: **Exploration, normalization and summaries of high-density oligonucleotide array probe level data.** *Biostatistics* 2003, **4**:249-264.
4. Naef F, Hacker CR, Patil N, Magnasco M: **Empirical characterization of the expression noise ratio structure in high-density oligonucleotide arrays.** *Genome Biol* 2002, **3(4)**:research0018.
5. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP: **Summaries of Affymetrix GeneChip probe level data.** *Nucleic Acids Res* 2003, **31(4)**:e15.
6. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98(1)**:31-36.
7. Sugimoto N, Iba H: **Inference of gene regulatory networks by means of dynamic differential bayesian networks and nonparametric regression.** *Genome Informatics* 2004, **15(2)**:121-130.
8. Zhang L, Miles MF, Aldape KD: **A model of molecular interactions on short oligonucleotide microarrays.** *Nat Biotechnol* 2003, **21(7)**:818-821.
9. Wu Z, Irizarry RA: **Preprocessing of oligonucleotide array data.** *Nat Biotechnol* 2004, **22(6)**:656-658.
10. Lemon WJ, Palatini JJT, Krahe R, Wright FA: **Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays.** *Bioinformatics* 2002, **18(11)**:1470-1476.
11. Rajagopalan D: **A comparison of statistical methods for analysis of high density oligonucleotide array data.** *Bioinformatics* 2003, **19(12)**:1469-1476.
12. Cope LM, Irizarry RA, Jaffee HA, Wu Z, Speed TP: **A benchmark for Affymetrix GeneChip expression measures.** *Bioinformatics* 2004, **20(3)**:323-331.
13. Bolstad BM, Irizarry RA, Astrand M, Speed TP: **A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.** *Bioinformatics* 2003, **19(2)**:185-193.
14. Irizarry RA, Wu Z, Jaffee HA: **Comparison of Affymetrix GeneChip expression measures.** *Bioinformatics* 2006, **22(7)**:789-794.
15. Harr B, Schlotterer C: **Comparison of algorithms for the analysis of Affymetrix microarray data as evaluated by co-expression of genes in known operons.** *Nucleic Acids Res* 2006, **34(2)**:e8.
16. Shedden K, Chen W, Kuick R, Ghosh D, Macdonald J, Cho KR, Giordano TJ, Gruber SB, Fearon ER, Taylor JMG, Hanash S: **Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data.** *BMC Bioinformatics* 2005, **6(26)**:.
17. Hoffmann R, Seidl T, Dugas M: **Profound effect of normalization on detection of differentially expressed genes in oligonucle-**

Appendix III: Jiang N, Wang M, Jia T, Wang L, Leach L, Hackett C, Marshall D, Luo Z. (2011) A robust statistical method for association-based eQTL analysis. *PLoS One*, 6(8): e23192.

OPEN ACCESS Freely available online

PLoS one

A Robust Statistical Method for Association-Based eQTL Analysis

Ning Jiang[1,3], Minghui Wang[1], Tianye Jia[1], Lin Wang[2], Lindsey Leach[1], Christine Hackett[3], David Marshall[4], Zewei Luo[1,2]*

1 School of Biosciences, University of Birmingham, Birmingham, United Kingdom, **2** Laboratory of Population and Quantitative Genetics, Institute of Biostatistics, Fudan University, Shanghai, China, **3** BioSS, Invergowrie, Dundee, Scotland, United Kingdom, **4** Scottish Crop Research Institute, Invergowrie, Dundee, Scotland, United Kingdom

Abstract

Background: It has been well established that theoretical kernel for recently surging genome-wide association study (GWAS) is statistical inference of linkage disequilibrium (LD) between a tested genetic marker and a putative locus affecting a disease trait. However, LD analysis is vulnerable to several confounding factors of which population stratification is the most prominent. Whilst many methods have been proposed to correct for the influence either through predicting the structure parameters or correcting inflation in the test statistic due to the stratification, these may not be feasible or may impose further statistical problems in practical implementation.

Methodology: We propose here a novel statistical method to control spurious LD in GWAS from population structure by incorporating a control marker into testing for significance of genetic association of a polymorphic marker with phenotypic variation of a complex trait. The method avoids the need of structure prediction which may be infeasible or inadequate in practice and accounts properly for a varying effect of population stratification on different regions of the genome under study. Utility and statistical properties of the new method were tested through an intensive computer simulation study and an association-based genome-wide mapping of expression quantitative trait loci in genetically divergent human populations.

Results/Conclusions: The analyses show that the new method confers an improved statistical power for detecting genuine genetic association in subpopulations and an effective control of spurious associations stemmed from population structure when compared with other two popularly implemented methods in the literature of GWAS.

Citation: Jiang N, Wang M, Jia T, Wang L, Leach L, et al. (2011) A Robust Statistical Method for Association-Based eQTL Analysis. PLoS ONE 6(8): e23192. doi:10.1371/journal.pone.0023192

Editor: Momiao Xiong, University of Texas School of Public Health, United States of America

Received April 29, 2011; **Accepted** July 7, 2011; **Published** August 9, 2011

Funding: The research was funded by the Biotechnology and Biological Sciences Research Council (RRAD11534) and the Leverhulme Trust (RCEJ14713). NJ was also supported by a joint studentship between the University of Birmingham and Biomathematics and Statistics Scotland (BioSS). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: z.luo@bham.ac.uk

Introduction

Linkage disequilibrium (LD) based association mapping has received increasing attention in the recent literature [1–6] for its potential power and precision in detecting subtle phenotypic associated genetic variants when compared with traditional family-based linkage studies. Association mapping methods for the genetic dissection of complex traits utilize the decay of LD, the rate of which is determined by genetic distance between loci and the generation time since LD arose [7]. Over multiple generations of segregation, only loci physically close to the quantitative trait loci (QTL) are likely to be significantly associated with the trait of interest in a randomly mating population, providing great efficiency at distinguishing between small recombination fractions [8]. Despite this potential, many reported association studies have not been replicated or have resulted in false positives [9–10], commonly caused by 'cryptic' structure in population-based samples. Population structure, or population stratification [11], arises from systematic variation in allele frequencies across subpopulations, which can result in statistical association between a disease phenotype and marker(s) that have no physical linkage to causative loci [12–13], *i.e.* false positive or spurious associations. This gives rise to an urgent need for methods of adjusting for both population structure and cryptic relatedness occurring due to distant relatedness among samples with no known family relationships.

To avoid the problems raised from population stratification, family-based association studies have been proposed, such as the transmission-disequilibrium test (TDT), which compares the frequencies of marker alleles transmitted from heterozygous parents to affected offspring against those that are not transmitted [14]. In this design the ethnic background of cases and controls is necessarily matched, conferring robustness to the presence of population structure. However, TDT design requires samples from family trios, which are difficult to obtain compared to population based designs where a large sample is feasibly obtained. Moreover, increased genotyping efforts are required for TDT design to achieve the same power as population based design [15–16].

Numerous methods have been proposed to overcome the problems caused by population structure without the need for family based samples. Among the most widely used are the

genomic control (GC) [17] and the structure association (SA) analysis [18–19]. In the former, inflation of the test statistic by population structure is estimated as a constant from unlinked markers in the genomic control group and then the test statistic will be adjusted from the estimate before being applied to infer the association. In the latter, unlinked markers are used to estimate the number of subpopulations from which the sample are collected, and then assign sample individuals to subpopulations. The former method considers an ideal but unrealistic situation of constant inflation factor for all markers, while in reality the influence of population structure on statistical inference of marker-trait association varies over genome locations [20]. For the SA method, it is computationally intensive to obtain accurate and reliable values for both the number of subpopulations in real datasets and to assign individual population membership. Alternative methods have been adopted to infer the subpopulation number, including Latent-Class model [21], mixture model [22] and a Bayesian model AdmixMap [23]. These methods share the assumption that associations among unlinked markers are the result of population structure and subpopulations are allocated to minimize these associations. This step depends critically upon the correct selection of a panel of markers to reflect population structure information. Price *et al.* [24] proposed a principal component analysis (PCA) based method, EIGENSTRAT, to model the ancestral difference in allele frequency and correct for population stratification by adjusting genotypes through linear regression on continuous axes of variation. While EIGENSTRAT provides specific correction for candidate markers, how to choose appropriate markers to infer population structure remains in question. In fact, prediction of the population structure may fail whenever the key assumption behind the structure prediction methods is violated.

Rather than using a panel of unlinked markers to exploit the cryptic population structure, a single null marker can be used to correct for bias of the test statistic in association studies. Wang *et al.* [25] suggested using a well-selected null marker to correct biases from population stratification on odds ratio estimation for a candidate gene within a logistic regression framework. They assumed a simplistic situation that the null marker had the same genotypic distribution as the candidate gene, which, however, was unknown in practice.

The expression quantitative trait locus (eQTL) analyses have recently shown that variation in human gene expression levels among individuals and also populations is influenced by polymorphic genetic variants [26–28]. The use of structured populations has meant that to detect the genetic variants accounting for differences in gene expression between subpopulations, GWAS had to be carried out separately for each subpopulation and the results subsequently compared. We present here a simple regression model of utilizing only one 'control' marker to remove the population structure effect in detecting LD between a marker and a putative quantitative trait locus. We first established the theoretical basis for selection and use of a control marker to correct for population structure and established a regression-based method for detecting the LD which is integrated with information of the control marker. We investigated the method for its efficiency to test the LD and to reduce false positives stemmed from population structure through intensive computer simulation studies and re-analysis of the gene expression (or eQTL) datasets collected from genetically divergent populations. The new method (**Method 1**) was compared with two alternative methods: single marker regression without population structure correction (**Method 2**) and multiple regression analysis with incorporation of known individual ancestry information (**Method 3**).

Materials and Methods

Method 1 (Regression analysis with correcting population structure)

The method analyzes a structured randomly mating population produced through instant admixture of two genetically divergent subpopulations. The proportion of subpopulation 1 in the mixed population is denoted by m. Let us consider three bi-allelic loci: one affects a quantitative trait (Q) while another two are polymorphic markers devoid of direct effect on the trait. We call, for convenience, one of the markers the test marker (T) which is to be tested for association with the QTL, and the other as control marker (C), assumed to be not associated with both the QTL and the test marker (*i.e.* the linkage disequilibrium D equal 0). Two alleles are denoted by A and a at the putative QTL, T and t at the test marker, and C and c at the control marker. Three genotypes at the QTL, AA, Aa and aa, are assumed to affect the quantitative trait by d, h and $-d$ respectively. Trait phenotype of an individual (Y) is assumed to be normally distributed with mean depending on its genotype at the QTL and residual variance σ_e^2. Genotypic values at the test marker and control marker are denoted by X and Z which are the number of alleles T and C respectively. In subpopulation i ($i=1$ or 2), the allelic frequencies of the QTL, test marker and control marker are denoted by $p_Q^{(i)}$, $p_T^{(i)}$ and $p_C^{(i)}$ respectively, while the coefficients of linkage disequilibrium between any pair of the loci are denoted by $D_{TC}^{(i)}$, $D_{TQ}^{(i)}$ and $D_{CQ}^{(i)}$. Table 1 illustrates probability distribution of joint genotypes at a test marker and a putative QTL in randomly mating populations together with genotypic values at the QTL and details

Table 1. Probability distribution of joint genotypes at a test marker and a putative QTL and genotypic values at the QTL.

Genotypes at QTL	*AA*			*Aa*			*aa*		
Marker genotypes	*TT*	*Tt*	*tt*	*TT*	*Tt*	*tt*	*TT*	*Tt*	*tt*
Probabilities	$(qQ)^2$	$2q^2Q(1-Q)$	$q^2(1-Q)^2$	$2\,q(1-q)QR$	$2\,q(1-q)$ $(Q+R-2QR)$	$2\,q(1-q)$ $(1-Q)(1-R)$	$(1-q)^2R^2$	$2(1-q)^2$ $R(1-R)$	$(1-q)^2(1-R)^2$
Genotypic values at QTL	$\mu+d$			$\mu+h$			$\mu-d$		

where A and a are segregating alleles at a putative QTL, T and t are alleles at the test marker locus. Allele frequency of A is q, allele frequency of T is p. Q and R are conditional probabilities of marker allele T given QTL allele A and a respectively, which are formulated as $Q=p+D/q$ and $R=p-D/(1-q)$ where D is the coefficient of linkage disequilibrium between the marker and QTL. μ, d and h are population mean, additive and dominance genic effects at the QTL.

doi:10.1371/journal.pone.0023192.t001

for the parameterization can be found in Luo [29]. It is clear from Table 1 that the marker-QTL distribution can be fully characterized by the parameters defining population allele frequencies at the two loci and the coefficient of linkage disequilibrium between them. This provides the theoretical basis for statistical analyses developed below.

Regression analysis correcting effect of population structure. For phenotype of a quantitative trait and each of the test markers, we fitted the following model: the genotype X_{ij} of individual i at the given marker locus j may be classified as one of three states: $X_{ij}=0, 1, or\ 2$ for homozygous rare, heterozygous and homozygous common alleles, respectively. For this model, we fitted a linear regression of the form for each genetic marker:

$$Y_i = b_0 + b_1 X_{ij} + \varepsilon_i \tag{1}$$

where Y_i is phenotype for individual $i=1,\cdots\cdots,n$, and ε_i are independent normally distributed random variables with mean 0 and variance σ_ε^2. We have demonstrated that significance of the regression coefficient can be used to infer significance of LD between a polymorphic marker locus and a QTL in a single randomly mating population since the regression coefficient has a form of

$$b_1 = \frac{\sigma_{X,Y}}{\sigma_X^2} = \frac{E(XY)-E(X)E(Y)}{E(X^2)-E^2(X)} = \frac{2D_{TQ}[d+(1-2p_Q)h]}{2p_T(1-p_T)} \tag{2}$$

[29]. However, in a structured population, we note that the LD between a marker and a QTL is given by

$$D_{TQ} = mD_{TQ}^{(1)} + (1-m)D_{TQ}^{(2)} + m(1-m)\delta_T\delta_Q, \tag{3}$$

[30], where m is the proportion of subpopulation 1 in this mixed samples, the superscripts (1) and (2) refers to the subpopulations, $\delta_T = p_T^{(1)} - p_T^{(2)}$ and $\delta_Q = p_Q^{(1)} - p_Q^{(2)}$. The covariance between the QTL and the test marker can be worked out as

$$\begin{aligned}\sigma_{X,Y} &= 2mD_{TQ}^{(1)}(d+h-2hp_Q^{(1)}) + 2(1-m)D_{TQ}^{(2)}(d+h-2hp_Q^{(2)}) \\ &\quad + 4m(1-m)\delta_T\delta_Q[d+h(1-p_Q^{(1)}-p_Q^{(2)})].\end{aligned} \tag{4}$$

Equations 3 and 4 show that the association between the QTL and test marker in a mixed population is the summation of (i) a linear combination of the associations between the two loci in each of the subpopulations (i.e. the genuine association due to LD between the two loci in each of the subpopulations), and (ii) a nonlinear component of the differences in allele frequencies between the two subpopulations (i.e. a spurious term of association). The objective of our analysis is to remove the spurious term by using a control marker 'C'. If the control marker is neither in association with the QTL (i.e. $D_{CQ}^{(1)} = D_{CQ}^{(2)} = 0$) nor with the test marker ($D_{TC}^{(1)} = D_{TC}^{(2)} = 0$), then the covariance between control marker and QTL (or test marker) can be given by

$$\sigma_{Y,Z} = 4m(1-m)\delta_C\delta_Q[d+h(1-p_Q^{(1)}-p_Q^{(2)})] \tag{5}$$

$$\sigma_{X,Z} = 4m(1-m)\delta_T\delta_C \tag{6}$$

In an admixed population, the control marker's allelic frequency is $p_C = mp_C^{(1)} + (1-m)p_C^{(2)}$. In a population with allelic frequency p_C at the control marker locus, the expected and observed variances at the control marker are

$$E[\sigma_Z^2] = 2[mp_C^{(1)} + (1-m)p_C^{(2)}][1-mp_C^{(1)} - (1-m)p_C^{(2)}] = 2p_C(1-p_C) \tag{7}$$

$$\sigma_Z^2 = 2[mp_C^{(1)} + (1-m)p_C^{(2)}][1-mp_C^{(1)} - (1-m)p_C^{(2)}] + 2m(1-m)\delta_C^2 \tag{8}$$

where $\delta_C = p_C^{(1)} - p_C^{(2)}$. Thus, the difference between the expected and observed variances at the control marker indicates the existence of population structure,

$$\sigma_Z^2 - E[\sigma_Z^2] = 2m(1-m)\delta_C^2 \tag{9}$$

The spurious term in the covariance in equation (4) can be completely corrected using a single control marker, as follows:

$$\begin{aligned}\bar{\sigma}_{X,Y} &= \sigma_{X,Y} - \frac{\sigma_{X,Z}\sigma_{Y,Z}}{2\{\sigma_Z^2 - E[\sigma_Z^2]\}} \\ &= 2mD_{TQ}^{(1)}(d+h-2hp_Q^{(1)}) + 2(1-m)D_{TQ}^{(2)}(d+h-2hp_Q^{(2)})\end{aligned} \tag{10}$$

Therefore, the regression coefficient calculated from

$$b_1 = \frac{\bar{\sigma}_{X,Y}}{\sigma_X^2} = \frac{\sigma_{X,Y} - \frac{\sigma_{X,Z}\sigma_{Y,Z}}{2\{\sigma_Z^2 - E[\sigma_Z^2]\}}}{\sigma_X^2} \tag{11}$$

would reflect correction for the population structure. The students t-test can be used to test for significance of the regression coefficient b_1. Standard error (se) of b_1 is given by

$$S_{b_1} = \sqrt{\frac{\sigma_X^2\sigma_Y^2 - \bar{\sigma}_{X,Y}^2}{n\sigma_X^2}} \tag{12}$$

Given the regression coefficients and their variances, the power of the regression analysis can be predicted from the probability [31]

$$\rho_t = \Pr\{t_\nu(\delta_t) > t_{\alpha/2,\nu}\} \tag{13}$$

where $t_\nu(\delta_t)$ represents a random variable with non-central t-distribution with ν degrees of freedom and non-centrality parameter δ_t and $t_{\alpha/2,\nu}$ is the upper $\alpha/2$ point of a central t-variable with the same degrees of freedom. The value of ν equals $n-3$ and the non-centrality parameter is given by [31] as

$$\delta_t = \frac{\Gamma[\nu/2]b_1}{\sqrt{\nu/2}\Gamma[(\nu-1)/2]S_{b_1}} \tag{14}$$

where $\Gamma(\cdot)$ stands for a gamma function.

Selection of the control marker. In practice, we propose the following procedure to select the control marker for a given test marker. Firstly, any marker but the test marker would be candidate for the control marker if it has or is

- an autosomal location on different chromosomes from the test marker,
- less missing genotype data than a prior given proportion

For each marker passing the above screening, one calculates the expected and observed variances from

$$E[\sigma_Z^2] = 2p_C(1-p_C) \tag{15}$$

$$\sigma_Z^2 = \sum_{i=1}^{n} (Z_i-\mu)^2/(n-1) \tag{16}$$

where Z_i is the genotypic value of the candidate control marker (0, 1, 2) for individual $i=1,\cdots\cdots,n$, and μ and p_C are the mean genotypic value across all individuals ($\sum_{i=1}^{n} Z_i/n$) and the allelic frequency of this marker, respectively. It should be noted that equations (7) and (15) are the same and that equation (16) stands for the sampling variance of the control marker whose expectation is given by equation (8) in the presence of population structure. The control marker is the one with the maximum difference between observed and expected variances, which has the maximum ability to remove the spurious term in mixed populations and does not introduce bias in single population.

Method 2 (Regression analysis without correcting population structure)

The method fits a simple regression model for detecting LD between the trait phenotype and a test marker as we proposed previously [29] and implemented in a recent population based eQTL analysis in [28], in which the regression coefficient has a form of

$$b_1 = \frac{\sigma^*_{X,Y}}{\sigma^2_X} \tag{17}$$

with a standard error equal to

$$S_{b_1} = \frac{\sigma^2_X\sigma^2_Y - (\sigma^*_{X,Y})^2}{n\sigma^2_X} \tag{18}$$

where $\sigma^*_{X,Y}$ is the non-corrected covariance between test marker locus and the quantitative trait.

Method 3 (multiple regression analysis)

The method regresses the trait phenotype on genotypic value of a test marker ($X_{ij}=0$, 1, 2) and the probability of membership to each constituent population P_i ($i=1$, 2 here) as described in the following multiple regression model

$$Y_i = b_0 + b_1 X_{ij} + b_2 P_i + \varepsilon_i \tag{19}$$

where the b_2P_i term reflects the population structure effect in mixed populations.

The regression coefficients are given by

$$b_1 = \frac{\sigma^2_P\sigma_{X,Y} - \sigma_{X,P}\sigma_{P,Y}}{\sigma^2_X\sigma^2_P - \sigma^2_{X,P}} \tag{20}$$

$$b_2 = \frac{\sigma^2_X\sigma_{P,Y} - \sigma_{X,P}\sigma_{X,Y}}{\sigma^2_X\sigma^2_P - \sigma^2_{X,P}} \tag{21}$$

and standard errors of the regression coefficients are formulated as

$$S_{b_1} = \sqrt{\frac{\sigma^2_P\sigma^2_Y}{n\sigma^2_X\sigma^2_P - \sigma^2_{X,P}}} \tag{22}$$

$$S_{b_2} = \sqrt{\frac{\sigma^2_X\sigma^2_Y}{n\sigma^2_X\sigma^2_P - \sigma^2_{X,P}}} \tag{23}$$

according to [32]. Significance of association of the test marker with the quantitative trait can be tested through testing for significance of the regression coefficient b_j by the Student t-test.

Results

Simulation study

To explore statistical properties and limitations of the methods described above, we developed and conducted a series of computation simulation studies. The simulation program mimics segregation pattern of genes at multiple marker loci and QTL in randomly mating natural populations in terms of simulation parameters defining allele frequencies, linkage disequilibria and population structure as illustrated in Table S1. The methods were detailed for simulating a population characterized the joint genotypic distribution at two loci and for sampling individuals from the simulated population [33]. Although the distribution involves only two loci, it is easy to extend to multiple loci because the two locus joint distribution can be easily converted into conditional (or transition) probability distribution of genotypes at one locus on that at another, and genotypes at multiple loci can be simulated as a Markov process governed by the conditional probability distribution. Of course, this will not undermine flexibility to specify any required linkage disequilibrium pattern among any loci. Subpopulations were independently generated and merged to produce the admixed population. In the present study, we were focused on 10 simulated populations defined by simulation parameters listed in Table S1.

Each simulation was repeated 100 times and simulation data was analyzed using the three different methods described above. We tabulated in Table 2 means and standard errors of 100 repeated regression coefficients and proportions of significant tests of the regression coefficients. It can be seen that **Methods 1** and **2** predicted the regression coefficients adequately in all simulated populations, but **Method 3** did so when all individuals were correctly allocated to their correct subpopulations. Listed in Table 2 were also proportions of significant tests of the regression in repeated simulations. It should be stressed that the proportion measures rate of false positive when the test marker and QTL were in linkage equilibrium such as in the first 4 simulated populations whilst it provides evaluation of an empirical statistical power for detecting the genetic association in populations 5 to 10. It is clear that the rate of false positive was properly controlled in association analysis with **Method 1**, and **Method 3** when all individuals were correctly allocated, and that LD between the test marker and QTL in populations 5–9 was tested significant by these methods with a high statistical power. In contrast, the simple regression analysis (**Method 2**) made a high proportion of false positive inference of the marker and QTL association when the LD was actually absent (populations 1–4) but failed to detect truly existing LD between the two loci (populations 5–9). The method is thus inappropriate to be used for genetic association analysis when population structure was present. Performance of **Method 3**,

Table 2. Means and standard errors of regression coefficients ($b \pm se$) and proportions (ρ or $\hat{\rho}$) of statistical tests for significance of the regression coefficients from three methods.

Pop	D_{TQ}	D'_{TQ}	Method 1				Method 2				Method 3					
			Simulated		Predicted		Simulated		Predicted		Simulated				Predicted	
			$b \pm se$	$\hat{\rho}$	b	ρ	$b \pm se$	$\hat{\rho}$	b	ρ	$b \pm se^a$	$\hat{\rho}^a$	$b \pm se^b$	$\hat{\rho}^b$	β^b	ρ^b
1	0.04	0.00	−0.078±0.015	0.06	0.00	0.00	1.293±0.006	0.98	1.278	1.00	0.006±0.007	0.00	1.035±0.006	0.84	0.00	0.00
2	0.04	0.00	−0.087±0.015	0.07	0.00	0.00	1.162±0.006	0.97	1.163	0.98	−0.008±0.007	0.00	0.940±0.007	0.74	0.00	0.00
3	−0.09	0.00	0.015±0.008	0.00	0.00	0.00	−2.371±0.005	1.00	−2.368	1.00	0.006±0.007	0.00	−2.038±0.006	1.00	0.00	0.00
4	−0.09	0.00	0.005±0.011	0.00	0.00	0.00	−3.157±0.007	1.00	−3.157	1.00	−0.007±0.009	0.00	−2.725±0.008	1.00	0.00	0.00
5	0.02	0.05	0.965±0.021	0.48	0.828	0.55	−0.159±0.007	0.00	−0.166	0.00	0.997±0.006	0.85	0.062±0.007	0.00	0.994	0.91
6	0.04	0.07	1.086±0.008	0.86	1.062	0.92	0.130±0.007	0.00	0.125	0.00	1.280±0.006	1.00	0.375±0.007	0.01	1.274	1.00
7	0.05	0.08	1.341±0.008	0.98	1.325	1.00	0.333±0.007	0.01	0.331	0.01	1.593±0.006	1.00	0.597±0.007	0.14	1.59	1.00
8	0.05	0.08	1.260±0.006	0.99	1.249	0.99	0.313±0.007	0.01	0.312	0.01	1.503±0.006	1.00	0.572±0.007	0.13	1.499	1.00
9	0.04	0.08	1.307±0.014	0.92	1.234	0.99	−0.005±0.006	0.00	0.00	0.00	1.698±0.006	1.00	0.333±0.007	0.02	1.704	1.00
10	−0.04	0.00	0.008±0.009	0.01	0.00	0.00	−1.233±0.006	0.99	−1.234	0.99	−0.003±0.007	0.00	−0.995±0.007	0.80	0.00	0.00

D_{TQ} and D'_{TQ} are the coefficients of LD between the marker and QTL in the simulated mixed population before and after correction for population structure respectively.

[a]predicted when all individuals were allocated to their correct subpopulations;

[b]predicted when half of all individuals were correctly allocated to their subpopulations but other half were randomly allocated to either of the two subpopulations. The predicted values were estimated from theoretical analysis, while the simulated values were estimated from the simulation studies.

doi:10.1371/journal.pone.0023192.t002

which incorporates membership of individuals to constituent populations as a covariate in multiple regression analysis, depends on the extent by which individuals are correctly allocated to their belonging populations. For example, the method lost its statistical power to detect the truly existing LD (populations 5–9) or made false positive inference of genetic association when on average a quarter of individuals under analysis were wrongly allocated to subpopulations (populations 1–4). These results show that the present method provides a powerful test for linkage disequilibrium between polymorphic markers and QTL and an effective control of population structure in the test.

Use of control markers in **Method 1** is the key underpinning for the method to be able to control influence of population structure in the genetic association test. To investigate effect of the control marker on efficiency of the association test, we explored performance of the method when population structure is actually absent or when different control markers are used in the presence of population structure. Table S2 shows predicted and observed proportions of significant tests of the disequilibrium between a test marker and a putative QTL in 10 simulation populations with (b) or without (a) population structure. The proportions were calculated from analyses with **Method 1** by using the control marker either with a constant allele frequency between two subpopulations or with varying allele frequencies. It demonstrates that the type I error is well controlled and the disequilibrium is efficiently detected by the method using a control marker even when population structure does not actually exist (a). In addition, when population structure is present (b), the method bears a high chance to make a false positive inference and to lose its detecting power if the control marker selected to be implemented in the analysis has a small difference in allele frequency between the subpopulations. However, the risk can be effectively controlled and the reduced power can be recovered when using the control marker with a large allele frequency difference. All these suggest that implementation of control markers with a non-trial difference in allele frequency will not cause any significant problem of false positive/negative inference when population stratification is actually not existent. In presence of population structure, we propose selection of a marker with largely divergent allele frequencies as the control marker.

Gene expression and genotype datasets

The gene expression and SNP datasets were collected from Epstein-Barr virus (EBV) transformed lymphoblastoid cell lines of unrelated individuals of European-derived (CEU, 60 Europeans), and Asia-derived (CHB+JPT, 41 Chinese and 41 Japanese). The datasets were originally developed by Spielman et al [28] to explore population specified gene expression and genetic control of the population specified gene expression, and were downloaded from http://www.ncbi.nlm.nih.gov/geo (Gene Expression Omnibus: GSM5859). The expression arrays were analyzed using the Affymetrix MAS 5.0 software and the hybridization intensity was $\log_2$-transformed into expression phenotype. The study focused on 4,197 genes that are expressed in lymphoblastoid cell lines. Of the 4,197 genes, 1,097 were detected to be significantly differentially expressed between the CEU and CHB+JPT samples (*t*-test, $P<10^{-5}$; $P_e<0.05$, Sidak correction) [34]. SNP data scored on the 60 CEU, 41 CHB and 41 JPT samples were obtained from the ><International HapMap Project (release 19,). All markers with an allele frequency of ≥5% were included, giving more than 2.2 million and 2.0 million common SNP markers for the CEU samples and CHB+JPT samples respectively. Comparison between the CEU and CHB+JPT samples provided genotype data for 1,606,182 unique SNP markers among all 142 individuals (60 CUE and 82 CHB+JPT samples).

We selected and re-analysed the gene expression and SNP datasets in the present study for several reasons. Firstly, these samples were collected from the populations whose genetical diversification was well verified [35–37], and make a typical example which the method is designed for. Secondly, gene expression phenotype bears a wide spectrum of genetic controls from *cis* to *trans* regulation and different levels of heritability. Some of these quantitative phenotypes show population specified expression or heterogeneity of underlying genetics. These enable the method to be tested under different genetic backgrounds. Finally, re-analysis of the same datasets recently published allows a direct comparison of analysis with the method developed in the present study with that implemented in the published analysis.

Validation of population structure

In 2005, The International HapMap Project reported that the CHB and JPT samples' allele frequencies were generally very similar, but different to the allele frequencies of CEU samples (Figure S1). We first explored deviation in genotypic distribution at each of nearly 2 million SNP markers from the Hardy-Weinberg equilibrium (HWE) within CEU and CHB+JPT samples separately and in mixed of the two samples by using both Pearson's chi-squared test and Fisher's exact test. To account for the multiple tests, we set the significant different level at $P<2.5\times10^{-8}$ ($P_e=0.05$ after Sidak correction). The analyses did not detect any of the SNP markers whose genotypic distribution showed significant deviation from HWE in either of the two samples. However, when all CEU and CHB+JPT samples were merged together there were approximately 3,000 markers scattered across all autosomes deviating significantly from the HWE expectation (2911 markers from Pearson's chi-squared test, consistent with 3011 markers from Fisher's exact test). These analyses show that the CEU and CHB+JPT samples can be recognized to be collected from genetically divergent random mating populations and that a mixed of them represents an example of samples from these populations. Population structure in the mixed sample was visualized as a score plot of the first two principal components built on the 2911 SNP markers, which explained a total of 62% of variability of the marker data (Figure 1).

Genome-wide association eQTL analysis

We implemented the three methods described above to perform association mapping of eQTL using the gene expression and SNP marker datasets. The analysis was carried out on the CEU and CHB+JPT samples separately or jointly. An eQTL in the present analysis was defined as an independent peak in the p-value profile across a given chromosome. Peaks occurring within 5 Mb of adjacent peaks were taken as a single eQTL peak because of insufficient evidence to declare the existence of multiple eQTL peaks over such narrow intervals [38]. The eQTL location was defined as the location within the peak with the smallest p-value. To account for the large number of tests, we set the significance level at nominal $P<2.5\times10^{-8}$ ($P_e<0.05$ after Sidak correction), a conservative level also used previously [28,34]. A *cis*-regulated eQTL was operationally defined by the presence of significant association with a SNP in the region 500 kb upstream of the start of the transcript to 500 kb downstream of the 3′ end; otherwise, the eQTL was classified as *trans*-acting. Table 3 summarizes the number of eQTL detected by the three methods (**Method 1** developed in the present study, **Method 2** the simple regression analysis employed by Spielman et al in [28], and **Method 3** the multiple regression analysis) from the Europe derived, Asia derived

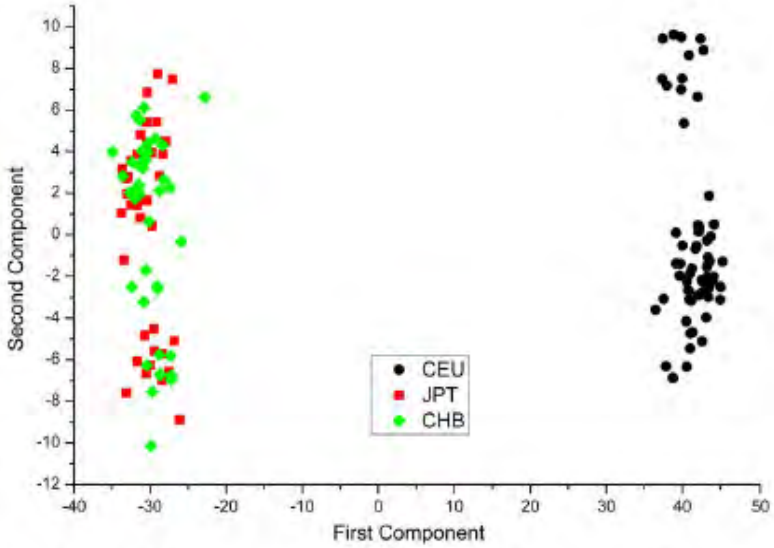


Figure 1. The first 2 Principal Components from PCA of 142 mixed HapMap Project human samples. The first and second principal components explained 60.77% and 1.34% of total variability respectively.
doi:10.1371/journal.pone.0023192.g001

samples and their mixed respectively. It can be seen that the eQTL analysis results from the CEU and CHB+JPT samples are comparable between **Method 1** and **2** in terms of the number of detected eQTLs and estimated locations of these eQTLs, suggesting a comparable predictability of the two methods in the absence of population structure. In the mixed sample, 64% of eQTL detected by the multiple regression analysis (**Method 3**) with use of full population membership information can be recovered by the method developed in the present study (**Method 1**), confirming the predictability of the latter in the presence of the population structure. We explored the predictability of **Method 3** when individuals were randomly assigned to the Europe derived sample (CEU) with probability of 58% or to the Asia derived sample (CHB+JPT) otherwise. The analysis showed that only 12%

Table 3. The number of eQTLs detected by three different methods (**Methods 1, 2, 3** or **M1, 2, 3 accordingly**) or detected common between two of these methods from the CEU, CHB+JPT and their mixed samples.

The number of eQTLs per expression trait	The CEU samples			The CHB+JPT samples			The mixed CEU and CHB+JPT samples				
	M1	M2	M1+2	M1	M2	M1+2	M1	M3	M1+3	M3*	M3+3*
1	280	312	**225**	263	255	**209**	206	251	**145**	398	**89**
2	58	57	**33**	43	41	**25**	16	13	**5**	136	**1**
3	20	21	**10**	13	16	**7**	2	7	**2**	97	**0**
4	10	16	**6**	8	6	**4**	2	2	**1**	72	**0**
5	4	4	**1**	5	6	**2**	0	0	**0**	48	**0**
6	3	1	**1**	1	3	**1**	0	0	**0**	37	**0**
7	3	3	**1**	0	2	**0**	0	0	**0**	22	**0**
8	0	2	**0**	1	0	**0**	1	0	**0**	22	**0**
9	2	1	**1**	0	0	**0**	0	1	**0**	14	**0**
>=10	19	22	**5**	6	7	**1**	2	2	**1**	1,111	**1**
Total eQTLs	1,009	1,149	**912**	633	670	**554**	296	354	**226**	1,975	**240**
cis-eQTLs	21	22	**21**	48	49	**48**	51	58	**51**	618	**53**
trans-eQTLs	988	1127	**891**	585	621	**506**	245	296	**175**	1,339	**187**

M3* is for Method 3 when individuals were randomly assigned to the Europe derived sample (CEU) with probability of 58% or to the Asia derived sample (CHB+JPT) otherwise.
doi:10.1371/journal.pone.0023192.t003

(240/1,975) of eQTL detected by the method with the partial population membership information was consistent with those detected by the same method with the full membership information, suggesting that the predictability of the method depends heavily on certainty of the membership information and that the method may generate a large proportion of false positives when the information is not complete.

The POMZP3 and HSD17B12 (on the human chromosome 7 q11.23 and chromosome 11 q11.2 respectively) are two well-characterized and *cis*-regulated genes [26,28,38–41]. Although all the three methods considered here were able to detect the previously identified *cis*-regulators from the three samples, there were a large number of spurious association signals predicted from the simple regression analysis (**Method 2**) with the mixed sample (Figure 2: a and b, respectively). It is clear that these spurious associations were effectively removed in the analysis with **Method 1**, reflecting the effectiveness of the latter in controlling the false positives (Figure 2: c and d, respectively). In the mixed samples, **Method 1** was able to reveal 296 significant eQTL, 51 of which were *cis*-regulators (Table 3). Firstly, the cis- eQTL predicted here include all the 11 *cis*-acting regulators reported by Spielman et al. [28] who performed the simple regression analysis (**Method 2**) in the CEU and CHB+JPT samples separately. In addition to 16 previously detected *cis*- acting factors, **Method 1** detected 35 novel cis- eQTL and all the eQTL explained 20~70% of variability in expression of the genes regulated (Table S3). We compared the 245 *trans*-regulators detected by our method from the mixed sample against the Gene Ontology (GO) Molecular Function annotation database (http://www.geneontology.org/) and found that 101 (42%) *trans*-eQTLs predicted were mapped

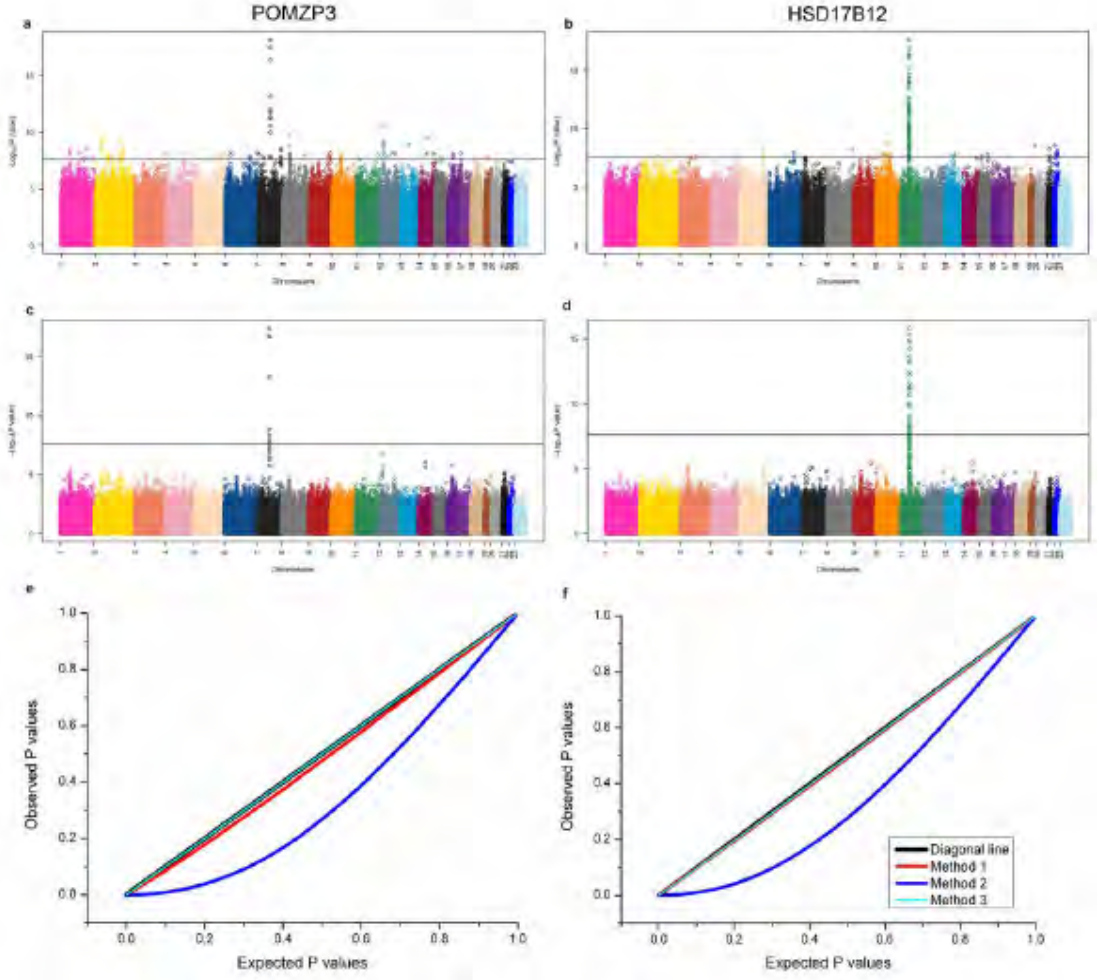


Figure 2. Manhattan plots for the genome-wide eQTL analysis of two genes POMZP3 and HSD17B12; Quantile-quantile (QQ) plots to compare the distributions between expected and observed p-values. Plots show score ($-\log_{10}$ p-value) for all SNPs by physical position for POMZP3 and HSD17B12 respectively based on simple linear regression (**Method 2**, a and b) and corrected linear regression (**Method 1**, c and d) in 142 mixed population samples.
doi:10.1371/journal.pone.0023192.g002

into the category of transcriptional factors, 82 (33%) *trans*-regulators played a role in signal pathway activity. In total, 75% *trans*-regulators predicted by the present method were previously known to play a role in gene regulation. All these reveal a significantly improved statistical power of the present method in detecting the true genetic associations.

It is interesting to note that the number of *cis*- eQTL detected from the mixed samples is larger than that from the component samples separately whilst a much larger number of *trans*- eQTL are detected in the component samples than in their mixed. This observation may reflect the fact that an increase in size of the mixed sample has enhanced the statistical power to detect cis-eQTL and thus led to an increased number of cis- eQTL detected. However, if linkage disequilibria between genes regulated and their trans- regulators are in opposite directions between different populations, the LD may be counter-balanced in the merged population, and thus decrease the number of the trans- eQTL to be detected. Despite a relatively small number of *cis* eQTLs detected, the *cis*-regulated effects were generally stronger than those in *trans*, with about 14% (7/51) *cis*-acting eQTL having a determination coefficient $R^2 > 50\%$ (Figure 3), consistent with findings in human and mice [38,42–44].

Discussion

Linkage disequilibrium based association mapping has been advocated as the method of choice for identifying chromosomal regions containing disease-susceptibility loci or loci affecting other complex quantitative traits of interest [45]. However, it is well known that the presence of population structure can result in false positive inference of genetic association between a test marker and trait loci. Various methods have been proposed in the literature to tackle this problem [19,21–23,46] and many of them have heavily depended on adequate prediction of the population structure [18,24]. Efficiency of the methods is thus largely affected by adequacy of population structure prediction. It has been shown that adequate prediction of population structure is in fact not a feasible task [47]. On the other hand, it is obvious that effect of the population stratification on association tests may vary across different regions of the genome [6–8]. Thus, the methods designed to correct for the stratification caused spurious associations through adjusting the test statistic by subtracting a constant inflation in the statistic may not perfectly reflect this observation [10,25]. To address these problems, we have proposed here a statistical method for correcting for stratification confounding effect in LD-based QTL mapping. The method extends the idea of using control markers to correct for background effect on a statistical test for significance of QTL at any given genome position in linkage-based QTL mapping analysis [48] and enables the effect of population stratification in the LD-based QTL analysis to be adjusted at a local basis. We presented here a simple but effective method to determine the control marker and demonstrated that incorporation of control markers would not cause any significant statistical problem even though population structure does actually not exist.

The new method developed in this study is tested and compared with other most popularly implemented methods in the literature of genetic association studies through intensive computer simulation studies and analysis of large scale and high quality gene expression and SNP datasets for mapping expression QTL. These analyses strongly support outperformance of the new method for its significantly improved statistical power to detect genuine LD between any polymorphic markers and putative trait loci and its effectiveness in controlling spurious association due to population stratification. Worthwhile, although the multiple regression analysis based on a mixed linear model does also provide a control of the influence of population stratifications, its efficiency depends heavily on accuracy of prediction of the population structure and on accurate allocation of individuals' membership to the constituent populations. Any bias in the structure prediction and uncertainty in the membership allocation may lead to severe consequence on its analytical efficiency. It has been argued that several factors may substantially influence or even disable the prediction of population structure [49–50]. Therefore, the method virtually avoids the need for sophisticated prediction of population ancestry of individuals and, in turn, effectively controls any bias embedded with the prediction. The method was designed for modeling and analyzing samples collected from different ethnical (or ecological) cohorts (or populations) with or without a clear clue about their genetic diversity. This is a very popular practice in many GWAS analyses, particularly with human samples [28,51–54].

Wang et al has proposed use of a single null marker to correct for population structure in a candidate gene based association analysis using case and control samples [25]. In their settings, the null marker was fitted as a dichotomous variable in parallel to the test candidate gene in a logistic regression model, and the influence of population structure on the association test at the

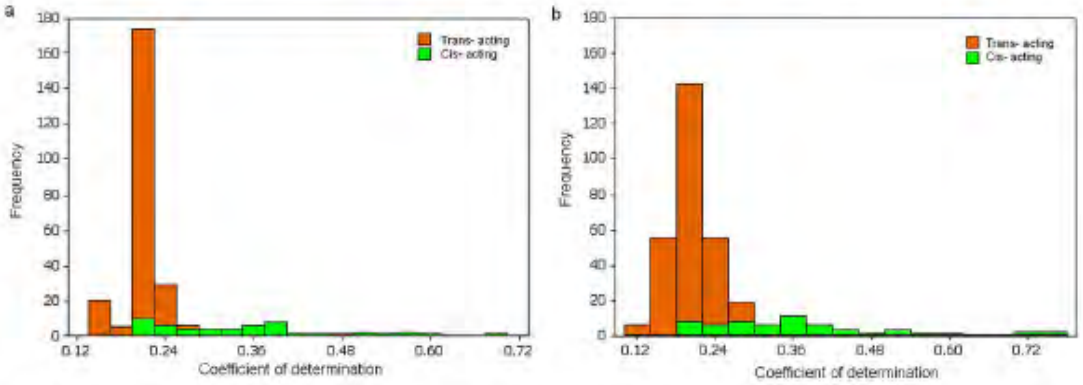


Figure 3. Histograms of coefficient of determination for eQTLs from 142 mixed sample set. a for **Method 1** and b for **Method 3**.
doi:10.1371/journal.pone.0023192.g003

candidate gene was adjusted by subtracting the regression coefficient associated with the null marker from the coefficient associated with the gene. Question rises to the parallel formulation: which is the major effect to be tested in the model? In contrast, our method was developed upon a rigorous population genetics model in which contributions of three different loci (i.e. the test marker, QTL and control marker) to the linkage disequilibrium pattern are properly formulated. The method is thus more appropriate for population based association studies. Although theoretical analysis was built on a single marker test, the idea and principle of the method could be extendable to the haplotype-based association mapping which uses information from multiple marker loci [55–56]. This is because the population confounding term is linearly attached to the main disequilibrium terms in the covariance between the test polymorphism and trait effect (Equation 3). Our goal is to remove the confounding term from the covariance and, thus form of the main disequilibrium terms either in genotype at an individual marker locus or in haplotypes at multiple marker loci will not affect the way to correct for the confounding term. Although the method was presented for two genetically divergent populations, the overall pattern of LD between any test marker and trait locus in their admixed population may become theoretically more complicated when the admixture involves more than two populations. Before having invested more theoretical investigation to the problem, we would suggest to merge those genetically less divergent objects together as we did in the present analysis with the Chinese and Japanese samples and to correct for the stratification raised from between the most divergent populations such as the European derived and the Asia derived samples.

Supporting Information

Figure S1 Comparison of allele frequencies between populations for all SNP markers genotyped in the International HapMap Project. The colour in each bin represents the number of SNPs that display each given set of allele frequencies.
(TIF)

Table S1 Parameters defining two subpopulations that are merged to produce admixed populations.
(DOC)

Table S2 Predicted and observed proportions of significant tests of linkage disequilibrium between a test marker and a putative QTL in different simulation populations from Method 1 in which the control marker implemented into the analyses had either (a) no population structure, and has a constant allele frequency difference of 0.4 at control marker locus or (b) population structure exist, and has varied allele frequency differences at control marker locus.
(DOC)

Table S3 The 51 cis-eQTLs predicted by Method 1 from the mixed sample.
(DOC)

Acknowledgments

We thank Prof. M. J. Kearsey and two anonymous reviewers for their critically constructive comments which have been helpful to improve presentation of the paper.

Author Contributions

Conceived and designed the experiments: ZL. Analyzed the data: NJ MW TJ LW. Contributed reagents/materials/analysis tools: ZL. Wrote the paper: ZL NJ LL CH DM.

References

1. Ardlie KG, Kruglyak L, Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics 3: 299–309.
2. Gauzin J, Kaiser J (2007) Genome-wide association: closing the net on common disease genes. Science 316: 820–822.
3. Ike MM (2008) What can genome-wide association studies tell us about the genetics of common disease. PLoS Genetics 4: e33.
4. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9: 356–369.
5. Slatkin M (2008) Linkage disequilibrium - understanding the evolutionary past and mapping the medical future. Nature Reviews Genetics 9: 477–485.
6. Weiss KM, Clark AG (2002) Linkage disequilibrium and the mapping of complex human traits. Trends in Genetics 18: 19–24.
7. Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends in Plant Science 12: 57–63.
8. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, et al. (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceedings of the National Academy of Sciences of the United States of America 98: 11479–11484.
9. Cardon LR, Bell JI (2001) Association study designs for complex diseases. Nature Review Genetics 2: 91–99.
10. Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847–856.
11. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nature Reviews Genetics 7: 781–791.
12. Ewens WJ, Spielman RS (1995) The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet 57: 455–464.
13. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037–2048.
14. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics 52: 506–516.
15. Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598–604.
16. McGinnis R, Shifman S, Darvasi A (2002) Power and efficiency of the TDT and case-control design for association scans. Behavior Genetics 32: 135–144.
17. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
18. Pritchard JK, Stephens M, Donnelly P (2000a) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
19. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000b) Association Mapping in Structured Populations. American Journal of Human Genetics 67: 170–181.
20. Astle W, Balding DJ (2009) Population structure and cryptic relatedness in genetic association studies. Statistical Science 24: 451–471.
21. Satten GA, Flanders WD, Yang QH (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. American Journal of Human Genetics 68: 466–477.
22. Zhu X, Zhang SL, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genetic Epidemiology 23: 181–196.
23. Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, et al. (2003) Control of Confounding of Genetic Associations in Stratified Populations. Am J Hum Genet 72: 1492–1504.
24. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics 38: 904–909.
25. Wang YT, Localio R, Rebbeck TR (2005) Bias correction with a single null marker for population stratification in candidate gene association studies. Human Heredity 59: 165–175.
26. Campino S, Forton J, Raj S, Mohr B, Auburn S, et al. (2008) Validating discovered cis-acting regulatory genetic variants: application of an Allele Specific Expression approach to HapMap populations. PLoS One 3: e4105.
27. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, et al. (2003) Natural variation in human gene expression assessed in lymphoblastoid cells. Nature Genetics 33: 422–425.
28. Spielman RS, Bastone LA, Burdick JT, Morley M, Ewens WJ, et al. (2007) Common genetic variants account for differences in gene expression among ethnic groups. Nature Genetics 39: 226–231.

Appendix IV: Wang M, Jia T, Jiang N, Wang L, Hu X, Luo Z. (2010) Inferring linkage disequilibrium from non-random samples. *BMC Genomics* 11: 328-340.

Wang et al. BMC Genomics 2010, 11:328
http://www.biomedcentral.com/1471-2164/11/328

BMC Genomics

METHODOLOGY ARTICLE Open Access

Inferring linkage disequilibrium from non-random samples†

Minghui Wang[1], Tianye Jia[1], Ning Jiang[1], Lin Wang[2], Xiaohua Hu[2] and Zewei Luo*[1,2]

Abstract

Background: Linkage disequilibrium (LD) plays a fundamental role in population genetics and in the current surge of studies to screen for subtle genetic variants affecting complex traits. Methods widely implemented in LD analyses require samples to be randomly collected, which, however, are usually ignored and thus raise the general question to the LD community of how the non-random sampling affects statistical inference of genetic association. Here we propose a new approach for inferring LD using a sample un-randomly collected from the population of interest.

Results: Simulation study was conducted to mimic generation of samples with various degrees of non-randomness from the simulated populations of interest. The method developed in the paper outperformed its rivals in adequately estimating the disequilibrium parameters in such sampling schemes. In analyzing a 'case and control' sample with β-thalassemia, the current method presented robustness to non-random sampling in contrast to two commonly used methods.

Conclusions: Through an intensive simulation study and analysis of a real dataset, we demonstrate the robustness of the proposed method to non-randomness in sampling schemes and the significant improvement of the method to provide accurate estimates of the disequilibrium parameter. This method provides a route to improve statistical reliability in association studies.

Background

Linkage disequilibrium (LD) has long been one of the central topics in evolutionary and population genetics. Linkage disequilibrium refers to non-random association of alleles at different linked or unlinked loci in a population. Inference about LD provides useful information for distinguishing between alternative evolutionary models of genetic polymorphisms within or divergence between populations [1]. The current surge of population based association studies has reported identification of causal genetic variants of disease susceptibilities in humans [2] and complex genetic variation in plants and animals [3,4]. The kernel of these studies is inference of LD between the genetic variants and functional loci that are closely genetically linked. Thus, adequate prediction of LD is obviously crucial for reliability and accuracy of these studies.

The coefficient of LD between two biallelic loci is defined as $D = f_{AB} - f_A f_B$ in a randomly mating population, where f_{AB}, f_A and f_B are frequencies of gametes AB, alleles A and B in the population. The genetic parameter has been re-parameterized into different forms for various purposes of LD analysis [5]. Hill proposed the well known "chromosome counting" method to estimate the parameter by using data of genotypes at the two loci from a random sample [6]. It has been widely used in population genetic analyses and studies on linkage disequilibrium based mapping [7,8]. The principle of the analysis has also been employed to develop widely used methods for predicting haplotypes of DNA markers in natural populations [9,10]. In practice, however, it is very rare that the samples for LD analyses are truly randomly collected from the population under study. For example, the samples used in many association studies or population genomics analyses were so collected that the frequencies of some genotypes are artificially inflated to ensure that genotypes involving a rare allele are well represented [11,12]. Weir and Cockerham explored the consequences of implementing the method to estimate LD by using the samples in which some genotypes are missing and stressed that the method should not be used to estimate

* Correspondence: zluo@bham.ac.uk
† School of Biosciences, The University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK
Full list of author information is available at the end of the article

Appendix V: Wang M, Jiang N, Jia T, Leach L, Cockram J, Waugh R, Ramsay L, Thomas B, Luo Z. (2012) Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. *TAG Theoretical and Applied Genetics*, 124, 233-246.

Theor Appl Genet (2012) 124:233–246
DOI 10.1007/s00122-011-1697-2

ORIGINAL PAPER

Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars

Minghui Wang · Ning Jiang · Tianye Jia · Lindsey Leach · James Cockram · Robbie Waugh · Luke Ramsay · Bill Thomas · Zewei Luo

Received: 27 April 2011 / Accepted: 29 August 2011 / Published online: 14 September 2011

Abstract Genome-wide association study (GWAS) has become an obvious general approach for studying traits of agricultural importance in higher plants, especially crops. Here, we present a GWAS of 32 morphologic and 10 agronomic traits in a collection of 615 barley cultivars genotyped by genome-wide polymorphisms from a recently developed barley oligonucleotide pool assay. Strong population structure effect related to mixed sampling based on seasonal growth habit and ear row number is present in this barley collection. Comparison of seven statistical approaches in a genome-wide scan for significant associations with or without correction for confounding by population structure, revealed that in reducing false positive rates while maintaining statistical power, a mixed linear model solution outperforms genomic control, structured association, stepwise regression control and principal components adjustment. The present study reports significant associations for sixteen morphologic and nine agronomic traits and demonstrates the power and feasibility of applying GWAS to explore complex traits in highly structured plant samples.

Communicated by J. Yu.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-011-1697-2) contains supplementary material, which is available to authorized users.

M. Wang · N. Jiang · T. Jia · Z. Luo (✉)
School of Biosciences, The University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
e-mail: z.luo@bham.ac.uk

N. Jiang · R. Waugh · L. Ramsay · B. Thomas
BioSS Unit, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, UK

L. Leach
Department of Plant Sciences, University of Oxford, South Parks Road, Oxford OX1 3RB, UK

J. Cockram
John Bingham Laboratory, National Institute of Agricultural Botany, Cambridge CB3 0LE, UK

Introduction

With the growing availability of genome sequence data and advances in technology for rapid identification and scoring of genetic markers, linkage disequilibrium (LD) based genome-wide association study (GWAS) has gained favour in higher plants, especially crops, for the mapping of genetic factors responsible for complex trait variation (Remington et al. 2001; Gupta et al. 2005; Mackay and Powell 2007; Cockram et al. 2008; Sneller et al. 2009; Atwell et al. 2010). While conventional linkage analysis works on an experimental population derived from a cross of bi-parents divergent for a trait of interest, association mapping applies to collections of samples of a much wider germplasm base. Providing the intrinsic nature of exploiting historical recombination events, association mapping offers increased mapping resolution to polymorphisms at sequence level and should therefore enhance the efficiency of gene discovery and facilitate marker assisted selection (MAS) in plant breeding (Gupta et al. 2005; Moose and Mumm 2008). Plants offer an ease of genetic manipulation allowing production of genetically uniform cultivars through inbreeding, making it possible to conduct replicated assays for many different traits under multiple environmental conditions. Once the plant cultivars are genotyped with high-density markers, association mapping