

Design and analysis of genetical genomics studies and their potential applications in livestock research



Alex C. Lam

This thesis is presented for the degree of

Doctor of Philosophy

The University of Edinburgh

2008

Abstract

Quantitative Trait Loci (QTL) mapping has been widely used to identify genetic loci attributable to the variation observed in complex traits. In recent years, gene expression phenotypes have emerged as a new type of quantitative trait for which QTL can be mapped. Locating sequence variation that has an effect on gene expression (eQTL) is thought to be a promising way to elucidate the genetic architecture of quantitative traits. This thesis explores a number of methodological aspects of eQTL mapping (also known as “genetical genomics”) and considers some practical strategies for applying this approach to livestock populations.

One of the exciting prospects of genetical genomics is that the combination of expression studies with fine mapping of functional trait loci can guide the reconstruction of gene networks. The thesis begins with an analysis in which correlations between gene expression and meat quality traits in pigs are investigated in relation to a pork meat quality QTL previously identified. The influence on power due to factors including sample size and records of matched subjects is discussed. An efficient experimental design for two-colour microarrays is then put forward, and it is shown to be an effective use of microarrays for mapping additive eQTL in outbred crosses under simulation. However, designs optimised for detecting both additive and dominance eQTL are found to be less effective.

Data collected from livestock populations usually have a pedigreed structure. Many family-based association mapping methods are rather computationally

intensive, hence are time-consuming when analysing very large numbers of traits. The application of a novel family-based association method is demonstrated; it is shown to be fast, accurate and flexible for genetical genomics. Furthermore, the results show that multiple testing correction alone is not sufficient to control type I errors in genetical genomics and that careful data filtering is essential. While it is important to limit false positives, it is desirable not to miss many true signals. A multi-trait analysis based on grouping of functionally related genes is devised to detect some of the signals overlooked by a univariate analysis. Using an inbred rat dataset, 13 loci are identified with significant linkage to gene sets of various functions defined by Gene Ontology. Applying this method to livestock species is possible, but the current level of annotations is a limiting factor. Finally, the thesis concludes with some current opinions on the development of genetical genomics and its impact on livestock genetics research.

Acknowledgements

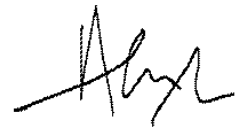
First and foremost, I would like to thank my two supervisors at Roslin, DJ de Koning and Chris Haley, for their guidance and encouragement in the last three years. I am very grateful that I was allowed the freedom and opportunities to explore different aspects of the research topic in the duration of the studentship. I would also like to thank Sara Knott, my university supervisor, for her helpful advice throughout the PhD. I would like to send my gratitude to Gary Evans for his effort in helping me to maintain links with my PhD industrial sponsor, Genus / PIC. I also acknowledge financial support from the Biotechnology and Biological Science Research Council and the Genesis Faraday Partnership.

Thanks to all the students who have worked in the north wing extension in the past three years, (Suzanne, Craig, Claudia, Georgia, Cecile, just to name a few), for making the office a great place to come into everyday. Thanks also go out to the “grown-ups” (Dave Telford, Ross, Ricardo, Wen Hua, Liz, Dave Waddington and many others) for all the laughs they brought and their occasional words of wisdom.

I would like to express my utmost gratitude to my parents, Lam Kam Wah and Yip Yin Siu, as well as to my other half, Rachel Gudger. Without love, patience and encouragement from them, I would not have been able to finish this work.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

A handwritten signature in black ink, appearing to read 'Alex C. Lam', written in a cursive style.

(Alex C. Lam)

List of publications

Refereed:

Lam, A.C., Fu, J., Jansen, R.C., Haley, C.S., and de Koning, D.J. 2008. Optimal design of genetic studies of gene expression with two-colour microarrays in outbred crosses. *Genetics*.180(3): 1691-8.

Lam, A.C., Powell, J., Wei, W.H., de Koning, D.J., and Haley, C.S. 2008. A combined strategy for quantitative trait loci detection by genome-wide association. *BMC Proceedings*. In press.

Lam, A.C., Schouten, M., Aulchenko, Y.S., Haley, C.S., and de Koning, D.J. 2007. Rapid and robust association mapping of expression QTL. *BMC Proceedings* 1 Suppl 1:S144.

Conference abstracts:

Lam, A.C., Sorensen, P., Petretto, E., Aitman, T.J., Haley, C.S., and de Koning, D.J. 2008. Pathway-based approach for genetic analysis of gene expression. *7th Annual Meeting of the Complex Trait Consortium. Abstract A3*.

Lam, A.C., Massault, C., Dornan, S., Wilkinson, J.M., Davey, G., Tilley, R., Blott, S.C., Sargent, Cairns, M., Evans, G., Plastow, G.S., Knott, S.A., Haley, C.S., and de Koning, D.J. 2006. Pathway inference using genetical genomics in pigs. *8th World Congress on Genetics Applied to Livestock Production. Communication 23-14*.

Table of contents

Abstract	2
Acknowledgements	4
Declaration	5
List of publications.....	6
Table of contents	7
List of figures	11
Lists of tables	12
CHAPTER 1	13
General introduction.....	13
CHAPTER 2	23
Integration of eQTL and functional QTL for network inference.....	23
2.1 Introduction	23
2.1.1 Network reconstructions from gene expression.....	23
2.1.2 Integrative Genomics approach.....	26
2.2 Inferring the pathway underlying a pork meat quality QTL.....	28
2.2.1 Meat quality in pigs and calpastatin gene	28
2.2.2 The proposed strategy	30
2.3 Material and methods.....	32
2.3.1 Subjects and meat quality phenotypes	32
2.3.2 Expression profiling	32
2.3.3 Microarray analysis	33
2.3.4 Statistical analysis	33
2.3.5 Pathway inference	33
2.4 Results	36
2.4.1 Marker association to meat quality traits	36
2.4.2 Marker association to expression traits	37
2.4.3 Correlation between significant meat quality traits and gene expression traits	37
2.4.4 Pathway Inference	37
2.5 Discussion	38

2.6 Conclusion	41
Appendix 2.1	43
Appendix 2.2	44
Appendix 2.3	45
Appendix 2.4	46
CHAPTER 3	47
Experimental design for genetical genomics	47
3.1 Introduction	48
3.1.1 Microarrays for gene expression profiling	48
3.1.1.1 One-colour platform	48
3.1.1.2 Two-colour platform	49
3.1.2 Generic microarray designs	49
3.1.2.1 The common reference design	51
3.1.2.2 The loop design	51
3.1.2.3 The block design	52
3.1.3 Microarray experimental designs for genetical genomics	53
3.1.3.1 Selective phenotyping	53
3.1.3.2 Optimal sample allocation	55
3.1.3.3 Distant pair design	56
3.1.4 Distant pair design for outbred crosses	57
3.2 Methods	58
3.2.1 QTL analysis	58
3.2.2 Finding optimal pairs	60
3.2.3 Power assessment via simulations	61
3.2.4 Alternative marker allele frequencies and population sizes	64
3.3 Results	64
3.3.1 Additive effect	64
3.3.2 Additive and dominance effects	70
3.3.3 Fixed number of microarrays with a large F_2 sample size	75
3.4 Discussion	77
3.4.1 Clear benefits in detecting additive effects	78
3.4.2 Complications due to dominance effects	79

3.4.3 Final remarks.....	80
3.5 Conclusion	81
CHAPTER 4	83
Genome-wide association of gene expression	83
4.1 Introduction.....	83
4.1.1 Family-based association	83
4.1.2 Controlling false discovery	87
4.2 Methods.....	90
4.2.1 Data description and pre-processing	90
4.2.2 Filtering on variability of the probesets	91
4.2.3 GRAMMAR procedures	91
4.2.4 Detection of cis-eQTL.....	92
4.2.5 Comparison of GRAMMAR to the full mixed model	92
4.3 Results and Discussion.....	93
4.3.1 Equivalence of GRAMMAR and the full mixed model method	93
4.3.2 Reduction in the number of tests by filtering on expression variability	95
4.3.3 Numerous spurious associations in the initial analysis.....	100
4.3.4 Reduction of spurious associations by filtering on genotype counts	103
4.3.5 Detection of cis-acting loci	104
4.4 Conclusion	105
CHAPTER 5	106
A gene set approach for eQTL mapping.....	106
5.1 Introduction.....	107
5.1.1 What is gene set testing.....	107
5.1.2 A review on gene set testing methodologies.....	111
5.2 The BXH/HXB rat dataset	117
5.3 Methods.....	118
5.3.1 Genotype and microarray data	120
5.3.2 Filtering based on expression and variability.....	120
5.3.3 Linkage analysis.....	121
5.3.4 Filtering based on KEGG.....	121
5.3.5 Gene set testing	122

5.4 Results and discussion	123
5.4.1 Fisher’s Exact Test.....	123
5.4.2 Wilcoxon Test	128
5.4.3 Discussion	135
5.5 Conclusions	138
CHAPTER 6	139
Gene set testing using Gene Ontology	139
6.1 Methods.....	139
6.1.1 Mapping of probesets to GO	139
6.1.2 Gene set testing	140
6.2 Results	141
6.3 Discussion	148
6.4 Conclusion	152
CHAPTER 7	153
General discussion and perspective	153
7.1 Summary	153
7.2 genetical genomics: future directions and pitfalls.....	156
7.3 The use of eQTL in livestock genetics.....	159
Bibliography.....	162

List of figures

Figure 1. 1	16
Figure 2. 1	27
Figure 2. 2	31
Figure 2. 3	35
Figure 3. 1	50
Figure 3. 2	66
Figure 3. 3	71
Figure 3. 4	76
Figure 4. 1	85
Figure 4. 2	94
Figure 4. 3	96
Figure 4. 4	97
Figure 4. 5	99
Figure 4. 6	102
Figure 5. 1	110
Figure 5. 2	113
Figure 5. 3	115
Figure 5. 4	119
Figure 5. 5	127
Figure 5. 6	132
Figure 5. 7	134

Lists of tables

Table 3. 1.....	63
Table 3. 2.....	69
Table 3. 3.....	72
Table 3. 4.....	74
Table 4. 1.....	101
Table 5. 1.....	124
Table 5. 2.....	129
Table 6. 1.....	143
Table 6. 2.....	147

CHAPTER 1

General introduction

The genetic basis of phenotypic variation can be broadly classified into two groups: monogenic and polygenic. Phenotypic variation with a monogenic background results from genetic variation at a single locus, and its inheritance follows a classical Mendelian pattern. On the other hand, the genetics of polygenic traits is often more complex; usually involves multiple genes, sometimes there are interactions between genes, and the environment can have an important role in the manifestation of the final outcome. Such traits are also commonly known as complex traits. Quantitative traits, that are traits with values which exhibit a continuous distribution, such as height and milk yield in dairy cattle, generally have a polygenic basis. With the development of genetic maps of polymorphic markers, it has become easier to conduct analysis to dissect the genetics of quantitative traits (Lander & Botstein 1989). Quantitative Trait Loci (QTL) mapping revealed that, in many cases, a small number of genetic loci contributed a large proportion of the phenotypic variance. QTL mapping has been widely used to identify loci that correlate with quantitative traits relevant to basic biology, inherited diseases and economically relevant traits in livestock and crop for many decades. To date, thousands of QTL have been mapped in various species. However, the rate of success of going from QTL to the characterisation of the loci at a molecular level has been disappointingly low (Flint *et al.* 2005).

The abundance of a gene transcript can be thought of as a proxy measure of gene expression, despite a number of post-transcriptional regulatory mechanisms

such as micro-Ribose Nucleic Acids (miRNA) and messenger RNA (mRNA) degradation that exist in the cell. It has been shown that in the fruit fly *Drosophila melanogaster*, there are a substantial number of genes for which the inter-individual variance in transcriptional abundance is significantly influenced by genotypes, amongst other factors like age and sex (Jin *et al.* 2001). This implies that the gene expression levels can be regarded as heritable quantitative traits. A later study using human lymphoblastoid cells observed familial aggregation of expression phenotype (Cheung *et al.* 2003), i.e. less variability in gene expression amongst individuals with greater relatedness. This supports further the view that there is a genetic contribution to the variation in the level of gene expression.

With the advance in high-throughput technology in genomics, expression profiling of many thousand of gene transcripts can be performed simultaneously using microarrays. As it has become clear that gene expression is a complex quantitative phenotype that is partly under genetic control, analogous to more “traditional” complex traits like blood pressure, QTL mapping methodologies naturally lend themselves to map the genetic loci which regulate transcription. Interestingly, studying the genetics of gene expression, sometimes known as “genetical genomics” (Jansen & Nap 2001), is thought to have enormous potential in dissecting the complex mechanisms underlying complex traits. The key factors that make gene expression and expression QTL (eQTL) potentially a very powerful way of studying the genetics of a biological system are: (a) both the phenotype and the QTL have a genomic location; (b) expression phenotypes are sometimes rich in functional details (e.g. gene annotation and expression pattern) that may be useful for candidate gene selection for related clinical or other functional complex phenotypes;

(c) using a genome scan to survey the transcriptomic variation permits the use of pleiotropic QTL and the correlation between expression phenotypes for inferring genetic pathways. These points will be elaborated further below.

In an eQTL experiment, the genome-wide gene expression for all individuals of a population sample is quantified using microarrays. In a population where there are segregating loci across the genome, the experiment can be thought of as a multi-factorial perturbation to a biological system (Jansen 2003), where we examine the allelic effect on each of the expression phenotypes at each polymorphic locus. This can be seen as a much more efficient way to investigate gene functions in a complex biological system than using targeted single knock-out in model organisms such as mice. An illustration of a hypothetical eQTL experiment is shown in Figure 1.1.

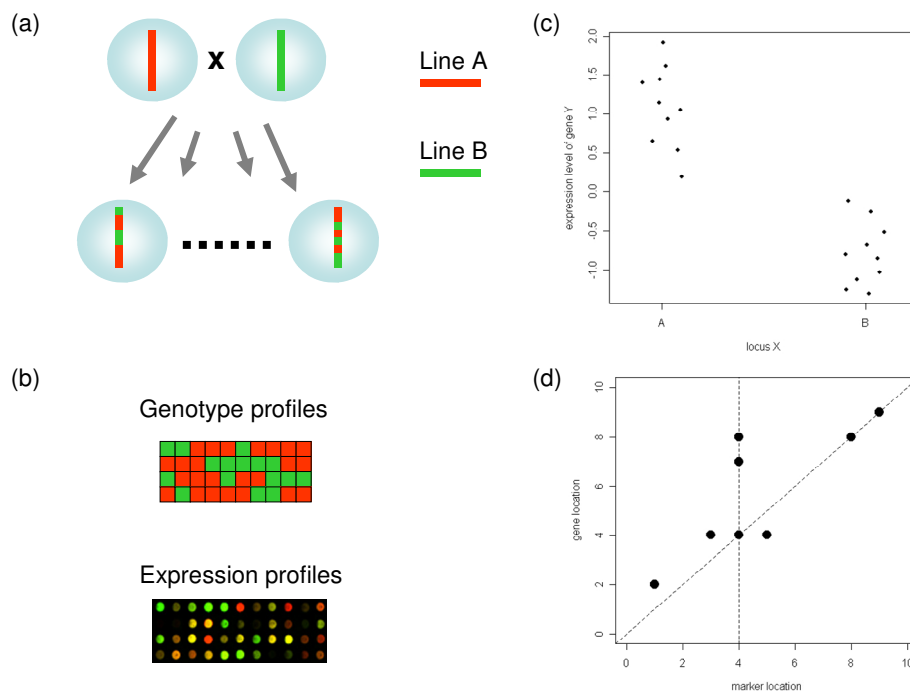


Figure 1. 1

A hypothetical genetic genomics experiment. (a) Design of experiment. An intercross between two inbred lines (A and B) of an organism gives rise to a population of hybrid progeny with a mosaic genome. (b) Genotyping and gene expression profiling. Genotypes of polymorphic markers indicate the inheritance pattern of a DNA segment. Genome-wide transcript abundance is assayed by microarrays. (c) An example of linkage to an expression phenotype. This example shows that the line A allele at locus X is associated with high expression level of gene Y. If there was no linkage at locus X, the expression level of gene Y should be more or less equally distributed amongst individuals with the line A allele and those with the line B allele. (d) eQTL scatter-plot. The result of a genome scan of gene expression phenotypes is often represented by a two-dimensional plot; the genomic location of the markers is shown on the x-axis, and the genomic location of the genes is shown on the y-axis. Significant linkage is indicated by a circle on the plot. See text for further details.

The significance of the linkage can be tested with similar tests to those used for QTL mapping, such as the *t*-test, *F*-test or likelihood ratio test. As shown in Figure 1.1(d), both the eQTL and the linked expression phenotype are features on the genome can be mapped to the physical map. Therefore, eQTL can be classified into two types: if eQTL and gene co-localise, the eQTL is said to be a local eQTL; if eQTL and gene do not co-localise, the eQTL is said to be a distant eQTL (Rockman & Kruglyak 2006). Many local eQTL are likely to be *cis*-eQTL: the regulation of gene expression is due to a direct effect of a polymorphism in close proximity; for example, a single point mutation in the promoter region which affects the initiation of transcription. In the two-dimensional scatter-plot, Figure 1.1(d), the *cis*-eQTL are represented by those plotted along the diagonal line. eQTL which lie distant from the linked gene are likely to exert *trans*-acting regulation; for example, a non-synonymous base substitution in a gene upstream of the linked gene in a signalling pathway. Therefore distant eQTL are often referred to as *trans*-eQTL. In Figure 1.1(d), the eQTL at genomic location 4 is linked the multiple gene transcripts, highlighted by the vertical dotted line. In this case the eQTL is a *cis*-eQTL for one gene and a *trans*-eQTL for two other genes. Pleiotropic eQTL as such are potentially very interesting because they point to the possibility that the eQTL is a master regulator for a number of genes, or that the linked genes belong to the same genetic pathway. For instance, the two *trans*-linked genes are regulated by the gene under *cis*-acting regulation. The multi-factorial nature of eQTL experimentation allows many questions to be asked about a biological system and hypotheses to be generated for modelling the mechanisms underlying complex traits.

One of the earliest applications of whole-genome eQTL mapping was done on yeast (Brem *et al.* 2002). Using a cross between a wild strain and a laboratory strain of *Saccharomyces cerevisiae*, linkage analysis identified 570 expression traits that were linked to one or more loci. A substantial proportion (185 expression traits) was linked to loci in close proximity to the gene itself. As the expected probability of a gene linked to a marker at the same location as itself due to chance is small (Brem *et al.* 2002), this indicates the direct *cis*-acting effects contribute significantly to the genetic control of gene expression variation between individuals. Eight *trans*-acting eQTL “hotspots” were identified, each modulating the expression of a group of 7 - 94 genes of related function, further demonstrating that eQTL mapping is applicable to the study of gene functions and pathways. Transcription factors were thought to be the genetic machinery affected by the polymorphisms represented by the *trans*-eQTL hotspots, hence the large number of genes that fell under their regulation. Surprisingly, it was shown that *trans*-regulatory variation is not enriched in transcription factor coding genes (Yvert *et al.* 2003).

Genetical genomics studies have also been applied to several higher eukaryotic organisms, most notably in mice (Schadt *et al.* 2003; Bystrykh *et al.* 2005; Chesler *et al.* 2005). Take the study on haematopoietic stem cell in mice (Bystrykh *et al.* 2005) as an example, eQTL mapping has led to identification of a number of *cis*-acting genes, carrying allelic polymorphism; some of which have critical roles in haematopoietic stem cell specific function. These genes are therefore good candidate for more in depth functional studies. Reconstruction of putative pathways has also been successfully carried out using the collection of co-regulated transcripts identified through the *trans*-acting hotspots together with the correlation

in their expression profiles with *cis*-regulated transcripts with known function. An alternative approach for pathway reconstruction integrates information from both eQTL and other complex trait QTL in order to systematically identify key genes of which the expression variation is due to the complex trait QTL, and these genes are responsible for driving the variation observed in the complex trait (Schadt *et al.* 2005). Using a BXD F2 intercross of inbred mice, the QTL mapping for an obesity trait was combined with eQTL mapping to identify three novel susceptibility genes for obesity.

Examples of application of genetical genomics in model organisms (Mehrabian *et al.* 2005; Hubner *et al.* 2005; DeCook *et al.* 2006; Li *et al.* 2006) and in humans (Morley *et al.* 2004; Monks *et al.* 2004; Stranger *et al.* 2005) are plentiful in literature. On the other hand, this approach has not yet been widely adopted in livestock species. Kadarmideen and colleagues (2006) discussed the potential uses of eQTL in animal breeding. Since gene expression can be treated as “intermediate” phenotype of complex economically important traits, variation in transcript abundance of relevant genes is conceivably closer to the genetics than the trait specified in the breeding goal. Hence, there is scope for obtaining estimated breeding values (EBVs) for gene expression of animals and incorporating those EBVs directly in selection programme. Also, eQTL can potentially be incorporated in marker assisted selection programme in order to target more directly at the cause of the phenotypic variation. Furthermore, by understanding the basic biology underlying the trait of interest, animal breeders would have a firmer handle on how to devise a more robust breeding programme and minimise the chance to introduce undesirable features to the selected animals. However, some of the major limiting factors related

to the adoption of genetical genomics in farm animal genetics to date are the lack of completed genome sequence for some of the livestock species as well as funding. Even though the chicken (Hillier *et al.* 2004) and cattle genome assemblies are available and have aided international effort in single nucleotide polymorphism discovery (Wong *et al.* 2004; Van Tassell *et al.* 2008), commercial expression microarray manufacturers are only just beginning to produce arrays for livestock species. Also, the current high cost in conducting eQTL experiments limits the accessibility of this approach to some extent to the scientists working on humans or model organisms like mouse in academia or pharmaceuticals, whom in general enjoy greater research budget than scientists in animal health or animal breeding. Nevertheless, as livestock genome projects continue to advance and the experimental cost as well as the uncertainty of the value of genetical genomics diminishes to an acceptable level for the farm animal sector / industry, wider use of expression QTL in livestock genetics may begin to emerge. Indeed, a small scale genetical genomics study focused on a marked body weight QTL in chicken has been outlined in a recent publication (de Koning *et al.* 2007).

As noted above, there is currently a strong interest in utilising the power of expression QTL studies for dissecting the genetics of complex traits. However, as the technology is still at its infancy, much of the methodological aspects are still relatively unexplored. A growing number of studies have focused on statistical issues and pitfalls related to the nature of the data in terms of the high dimensionality and high correlation (Perez-Enciso *et al.* 2003; de Koning & Haley 2005; Carlborg *et al.* 2005; Gibson & Weir 2005; Pastinen *et al.* 2006; Sladek & Hudson 2006). Several other studies are concerned with experimental designs applicable to genetical

genomics for increased efficiency (Jin *et al.* 2004; Piepho 2005; Fu & Jansen 2006; Rosa *et al.* 2006). Clearly, there are urgent needs to advance our understanding of this emerging research field. The objective of this thesis is to further explore a range of issues that are central to genetic analysis of gene expression. The following paragraphs outline the contents and study objectives for each of the subsequent chapters.

Chapter 2 details a candidate gene study combining gene expression data and phenotypic records related to pork meat quality in pig. The aim of the experiment is to identify expression phenotypes that are co-regulated by a meat quality QTL which has been previously characterised. Subsequently, possible roles of these genes related to the QTL and meat quality traits can be explored. The chapter concludes with a critical assessment of the experimental design and reviews the lessons learned from this study from a practical point of view.

Chapter 3 introduces the concept of microarray design and reviews current methods in gene expression profiling which attempt to make efficient use of experimental resources in studying the genetics of gene expression. A new experimental design in two-colour microarrays is put forward as the optimal design for linkage mapping of eQTL, particularly in outbred crosses. Its strengths and weaknesses are examined using simulation.

Chapter 4 showcases a new method called GRAMMAR (Aulchenko *et al.* 2007a) that is amenable to family-based association studies by applying it to a human eQTL dataset with pedigree structure. GRAMMAR is not only capable of attaining similar statistical power as the linear mixed model which can be regarded as the gold standard in association mapping with pedigree data, but is also fast in its

computation; hence it can be shown that the method is adequate for analysing data with extremely high dimensionality such as an eQTL scan. The work also demonstrates the importance of careful data filtering and partitioning in enhancing statistical power and reducing the number of false positives.

Chapter 5 is devoted to a feasibility study of a holistic, pathway-based approach in mapping eQTL. This approach takes an alternative view to the univariate approach which focuses on solely the peak markers and their linked genes; instead it scans for evidence of pleiotropy to genes that belong to a common functional category defined by the KEGG database. Using a published eQTL dataset of a recombinant inbred lines panel of laboratory rats, the usefulness of the approach in revealing biologically meaningful eQTL with relatively small effects is assessed. Insight is provided into the properties of the methods for categorising genes into pathways or functional sets and the test statistics for evaluating the significance of the linkage evidence of the gene sets.

Chapter 6 is an extension of Chapter 5, in which the analysis is repeated using Gene Ontology instead of KEGG. The work confirms that more robust signals can be obtained when there is more extensive coverage over the genes on the microarray by the functional annotations.

Chapter 7 features a final summary and the concluding remarks of the research involved in producing this thesis, a few encouraging as well as cautionary notes on the interpretation of eQTL findings in complex trait research, and perspective on some future research directions regarding the use of gene expression data in livestock genetics.

CHAPTER 2

Integration of eQTL and functional QTL for network inference

A large number of studies attempt to infer gene networks from gene expression data obtained from microarray experiments. Co-expressed genes can be grouped and some form of regulation between those genes is assumed. More recently, several studies have shown that by combining gene expression data with genetics mapping of quantitative phenotypes, gene networks can be inferred with higher accuracy and richer contents compared to using gene expression data alone. In this chapter, a review on the current approaches of gene network inference is first provided. Next, a small investigation using this approach on the genetic mechanism underlying a pork meat quality QTL is described. Finally, I present the results and a critical assessment of the study.

2.1 Introduction

2.1.1 Network reconstructions from gene expression

The functioning of a cell is orchestrated in a number of ways; one of them is through variation in the expression levels of genes. Levels of transcript abundance are often directly related to the levels of gene products, the quantity of which are crucial for the regulation of many cellular processes, such as signal transduction, transcription activation and repression. Therefore, the level of expression by one gene would often have a knock-on effect to the activity of one or more genes, and it is common to consider that genes interact with one another in a network or pathway.

Such networks, or cascades of gene-gene interactions, can result in changes in the physiological state of the cell, organ and the whole organism. Thus, understanding which genes interact and how they interact would help scientists to recognize the complexity of phenotypes, and what to target in order to manipulate the phenotypes, either by molecular biology or selective breeding, to humans' advantage.

As microarray technology is capable of quantifying many thousands of genes simultaneously, it presents an opportunity for researchers to infer gene networks. The basic assumption underlying such analysis is that co-expression of genes hints at co-regulation and a common genetic pathway.

Co-expression refers to genes of which the expression levels are highly correlated. David Bostein and colleagues (Eisen *et al.* 1998) at Stanford University were amongst the first to perform cluster analysis on gene expression profiles and they found that genes of known similar function tended to group together. In their study, clustering of genes in budding yeast is based on levels of expression across a time dimension. A strong tendency of genes sharing common roles in cellular processes occupying a cluster was observed. In a study on expression pattern in the nematode worm *C. elegans* (Kim *et al.* 2001), a search for co-expressed genes was conducted when comparing wildtype versus mutant strains, as well as worms grown under different conditions. Forty-three clusters were identified, many of which are enriched for genes from particular tissues or organs and for genes with related biological roles.

Co-expression of genes provides the first clue as to which set of genes are likely to interact. Co-expression, however, does not reveal the organisation of the gene network; i.e. direct or indirect interaction between two genes, size of the effects,

and causality in the interaction. Various modelling methodologies have been developed to infer gene network through “reverse engineering” (D'haeseleer *et al.* 2000).

To make distinction between direct and indirect interactions, the use of partial correlation coefficients was proposed (de la Fuente *et al.* 2004). A partial correlation coefficient quantifies the correlation between two variables when conditioning on one or several other variables. By calculating high order partial correlation, the structures of gene networks emerge as undirected dependency graphs in which pairs of genes are connected by undirected edges if there is direct dependence between them. Network reconstruction based on a reanalysis of the yeast eQTL dataset (Brem *et al.* 2002) using this method and putative networks were postulated (Bing & Hoeschele 2005). Voy *et al.* (2006) extended the use of correlation graphs from between-genes to between-cliques, groups of interconnected genes, to model relationships in expression between multiple genes.

Yet, the correlation-based graph methods described above do not model causality in the interactions, or the direction of the edges. An alternative approach known as probabilistic graphical models, which include Bayesian networks, was suggested as the state-of-the-art for modelling gene network from expression data (Friedman 2004). In a Bayesian network, the gene expression of a gene is treated as a random variable, and the expression level of a gene is directly influenced by its parents in the model. Bayesian methods can be used to find the model that fits the data with the highest likelihood. Networks inferred by probabilistic graphical models are directed acyclic graphs. These graphs model causality in the interactions, but have the major drawback that feedback loops are not allowed. To model feedback

mechanisms it might be necessary to use multiple network models, each representing a distinct time point where interactions are uni-directional.

2.1.2 Integrative Genomics approach

Combining knowledge derived from genetic analysis in segregating populations with gene expression data has been shown to be a powerful way to infer gene networks (Schadt *et al.* 2005). The particular strengths of this approach over the methods above are in its ability to identify the “key-drivers”, or highly connected hubs, of the network as well as to detect causal associations. Genetic loci associated with both expression traits and physiological traits can be used as network priors to help refining models inferred from gene expression alone. In the presence of the co-localization of both eQTL and physiological QTL (pleiotropy), those loci are good candidates for the genetic factors which influence the physiological trait via some regulatory mechanisms on gene expression. The linked expression phenotypes with significant correlation with the physiological trait are even more likely to be involved in the gene network underlying the physiological trait. The relationships of the QTL, the linked expression phenotypes and the physiological traits are inferred using a likelihood-based causality model selection (LCMS) test (Schadt *et al.* 2005). The models tested by LCMS are described in Figure 2.1. Using conditional likelihood, the most probable model that reflects the joint probability distribution for the QTL, expression trait and physiological trait can be selected

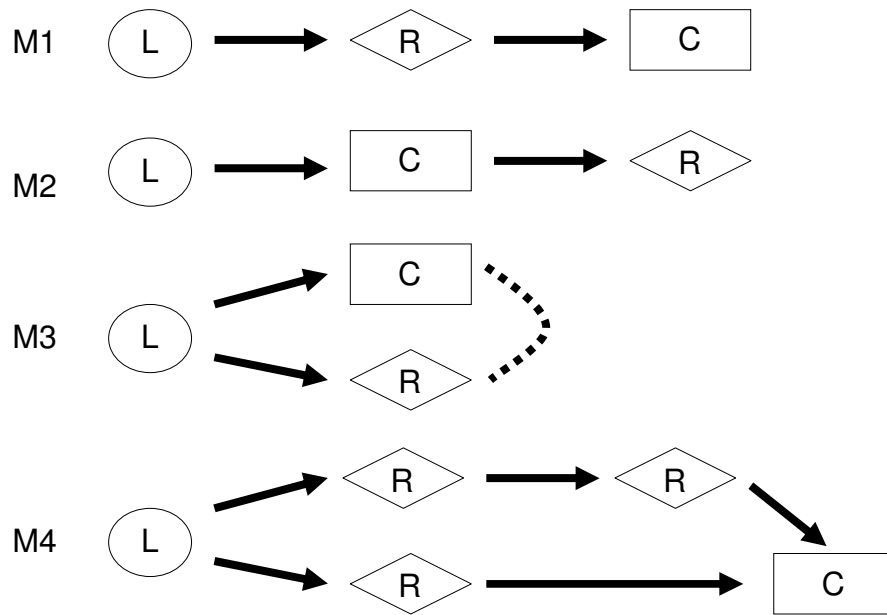


Figure 2. 1

Basic models considered by LCMS where expression levels (R) and complex physiological trait (C) are under the control of a common QTL (L). The QTL is always the head node of all models because genotypes are not dependent on the expression levels or the complex trait. The causal model, denoted by M1, shows the QTL acting on the complex trait simply through the transcript abundance of one gene. The reactive model, M2, shows the variation of transcript abundance as a consequence of the physiological state; i.e. transcript abundance has no influence on the outcome of the complex trait. The independent model, M3, shows that the complex trait and the expression phenotype are under the same genetic control and are correlated, but they do not influence each other. A more complicated causal model is shown as M4, where the complex trait is regulated by multiple genes, including *trans*-regulation through secondary genes. This figure is adapted from Schadt *et al.* (2005).

Genetic information can complement the model selection procedures in Bayesian network models (Zhu *et al.* 2004; Zhu *et al.* 2007). For examples, genes with expression levels under *cis*-regulation are assumed to be more upstream in a genetic regulatory pathway than *trans*-genes. Hence, *cis*-genes are fixed as the parent nodes, and models with *trans*-genes regulating *cis*-genes can be ruled out. Compared with the networks reconstructed from Bayesian network in the absence of genetic data, Zhu *et al.* (2007) showed that genetic information is most helpful in the top layer of the network, and the integrative genomics approach increases the accuracy of network reconstructions.

2.2 Inferring the pathway underlying a pork meat quality QTL

Inspired by the integrative genomic approach outlined in Schadt *et al.* (2005), an eQTL experiment was set up by the pig breeding company PIC and a consortium of academics and industrial partners to investigate the relationship between the natural variation in gene expression and pork meat quality. This section gives a brief introduction on the association of the calpastatin (*CAST*) gene with pork meat quality and outlines the proposed strategy to look for gene networks connecting the QTL in *CAST* and pork meat quality.

2.2.1 Meat quality in pigs and calpastatin gene

Eating quality of meat depends on many characteristics, such as leanness, pH, firmness, and biochemical compositions. It is clearly a complex concept that can be defined in many different ways and phenotyped in a variety of quantitative and

qualitative measures. Nevertheless, improving meat quality in one form or another is an important endeavour for commercial animal breeders, because the companies that supply stocks with more desirable meat quality can differentiate their products from their competitors and increase their market share. Thus, finding genes or genetic markers that associate with certain meat quality traits are helpful in making decisions in selective breeding programmes. A number of meat quality related QTL, for traits such as those relevant to growth performance, body composition, post-stress cortisol levels and glycogen content in skeletal muscle, were discovered (Ciobanu *et al.* 2001; Bidanel *et al.* 2001; Milan *et al.* 2002).

Meat tenderness has been linked to the post-mortem activity of calpastatin (*CAST*), a specific inhibitor of calpain proteases (Sensky *et al.* 1999; Parr *et al.* 1999). Using an F₂ intercross of Berkshire x Yorkshire pigs, Ciobanu *et al.* (2004) discovered QTL for cooking loss and juiciness in the *CAST* region. By sequencing, three missense mutations and five silent mutations in *CAST* were identified. These polymorphisms form haplotypes covering most of the coding region, and it was found that differences in some meat quality traits could be explained by the substitution effects of haplotypes and some of the nonsense mutations. Two of the coding mutations for *CAST*, *Ser66Asn* and *Ser638Arg*, are thought to disrupt the recognition sites for Protein Kinase A. It was discussed that the phosphorylation of *CAST* could affect its inhibitory efficiency and ultimately could have an effect on those meat quality traits (Ciobanu *et al.* 2004).

Because of the evidence supporting *CAST* as an important gene in contributing a significant effect on pork meat quality, there were both scientific and economic values in discovering whether the QTL at the *CAST* region acts on the

meat quality traits through the expression levels of other genes, and if yes, the gene network involved.

2.2.2 The proposed strategy

The overall strategy resembles the strategy outlined by Schadt *et al.* (2005), with a more simplistic model selection step for pathway inference. Details on material and methods are given in the next section. The aim is, given that *CAST* is a meat quality QTL, to find expression traits that are associated with the same locus. The co-localization of QTL and eQTL will then enable the reconstruction of a putative gene network. The general workflow is shown in Figure 2.2.

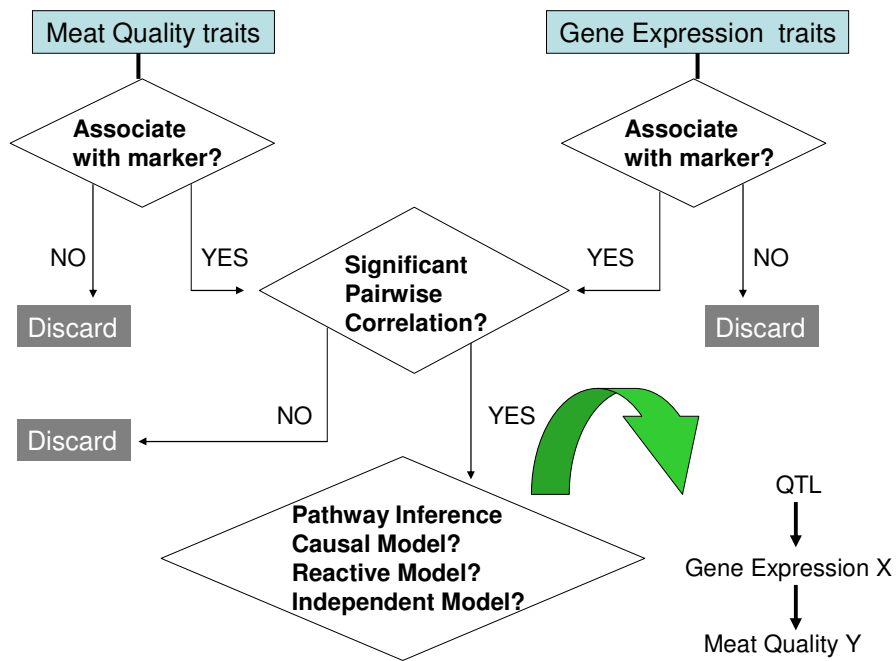


Figure 2. 2

The general workflow for the inference the gene network underlying the *CAST* QTL for meat quality in pigs. The most interesting type of networks would be those which follow the causal model: The QTL in Calpastatin alters expression level of genes which then drive the variation in meat quality traits.

2.3 Material and methods

2.3.1 Subjects and meat quality phenotypes

Five hundred pigs from 5 lines: Landrace, Large White, Duroc, Pietrain, and Meishan Synthetic, were used in the study. The polymorphism genotyped was the non-synonymous SNP resulting in the *Ser638Arg* mutation in *CAST* (Ciobanu *et al.* 2004). The pigs were raised and slaughtered in 22 batches and 361 traits related to biochemical properties, performance, and meat quality were measured. Twenty-three pigs from each of the 5 lines, 115 in total, were used in microarray experiments. Subjects with missing *CAST* genotypes were excluded from the analysis. Thus, 407 samples were available for the meat quality traits - marker association analysis, and 94 samples were available for gene expression traits - marker association analysis.

2.3.2 Expression profiling

RNA was extracted from two muscles (*Longissimus thoracis* and *Semimembranosus*). On cDNA microarrays, a reference design without dye swap was employed whereby each sample was compared against a reference sample which was composed of all 115 samples pooled together. Each array contained 21,168 probes, of which 19,014 were cDNA probes for the pig genome. Within arrays these cDNA probes were replicated 3, 6, or 12 times. Overall, they represent 6,192 non-redundant cDNA clones, whose identities were anonymous. Array hybridisation was performed in three different laboratories. The scanning of arrays was completed in the PIC Cambridge Laboratory.

2.3.3 Microarray analysis

Microarray data were processed and normalised using the R statistical programming language (R Development Core Team 2007) and the print-tip loess method in the Limma package (Smyth 2005). The average of normalised log ratios was taken for each cDNA probeset as the measure of expression level.

2.3.4 Statistical analysis

Marker association to phenotypic traits was assessed using the linear model:

$$Y(\text{trait}) = \mu + L(\text{line}) + B(\text{batch}) + G(\text{genotype}) + e$$

Association to gene expression traits was assessed in a similar way:

$$Y(\text{trait}) = \mu + L(\text{line}) + B(\text{batch}) + P(\text{process_lab}) + G(\text{genotype}) + e$$

All nuisance terms and the *CAST* genotype term were fitted as fixed effect. The threshold of $P_{(CAST)} = 0.05$ for both models were used in this first filtering step. For those phenotypic traits and the gene expression traits with significant marker association, a second filtering step was applied using the Pearson correlation test on pairwise combinations of the two types of traits. A more stringent threshold ($P = 0.01$) was set for the Pearson correlation tests. Linear models were fitted in the statistical package R using the method *lm()*.

2.3.5 Pathway inference

For the combinations of meat quality and gene expression traits that passed both filtering steps, a simple regression procedure was used to determine the best model, from the causal model, the reactive model, and the independent model

illustrated in Figure 2.1. The model selection procedure is described in Figure 2.3 and in the text below.

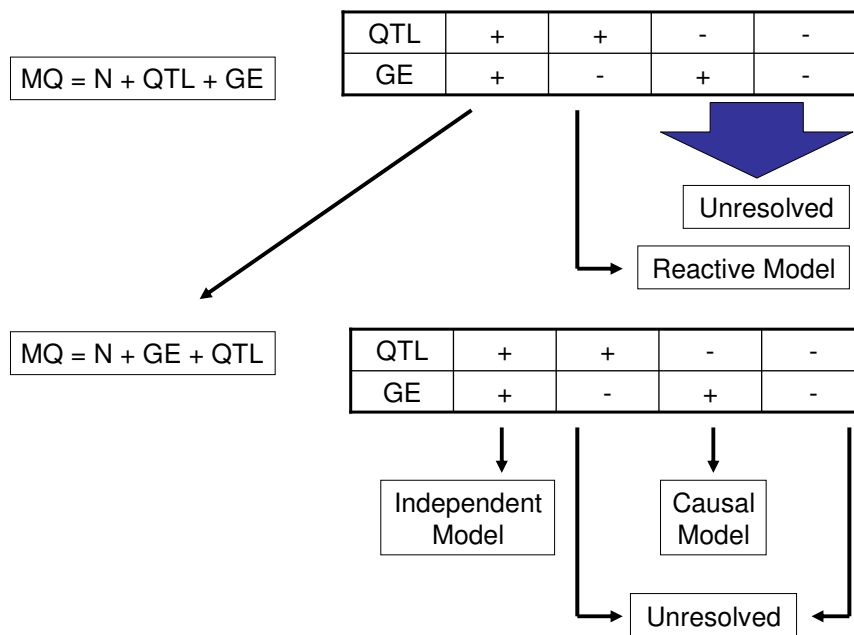


Figure 2. 3

Model selection procedure for pathway inference. The terms used in the regression models are meat quality trait (MQ), nuisance parameters (N), *CAST* genotype (QTL) and gene expression (GE). Gene expression traits are tested one at a time. On the left hand side, the step 1 and step 2 regression models are shown. On the right hand side, the (+) and (-) signs indicates whether the term on that row is significant and not significant, respectively.

The first model in Figure 2.3 fits the QTL ahead of the expression trait. The reactive model can be implicated by the observation that the addition of the expression trait does not explain significantly more of the phenotypic variance than the QTL already does. For those combinations with significant QTL as well as gene expression in the first model, the second model fits the gene expression ahead of the QTL. If adding QTL after gene expression does not explain more of the phenotypic variance, it would imply the causal model where the phenotypic variance is driven by the gene expression trait. On the other hand, if the QTL remains significant, it would imply the independent model where the phenotypic trait and the gene expression where both associated with the QTL and are correlated, but not directly connected. If more than one expression traits fit the causal model, a joint model with multiple gene expression traits would be fitted and a backward selection procedure is used to determine whether these expression traits are connected or act independently. The goodness of fit can be assessed by the Akaike's Information Criterion (AIC). All linear models and model selection procedures were carried out in R.

2.4 Results

2.4.1 Marker association to meat quality traits

Out of the 361 phenotypic traits, the *CAST* genotype was significantly associated with 14 traits (nominal $P < 0.05$). Line and batch were significantly associated with substantially more traits, 258 and 280 respectively. Appendix 2.1 contains a table listing the results for the 14 significant traits.

2.4.2 Marker association to expression traits

Out of 6192 cDNA probes, the *CAST* genotype was significantly associated with 143 of them (nominal $P < 0.05$). Incidentally, many more expression traits were significantly associated with the nuisance parameters; 862, 600, and 3605 expression traits were significantly associated with line, batch, and process lab, respectively.

2.4.3 Correlation between significant meat quality traits and gene expression traits

Pearson correlation tests were performed between the 14 meat quality traits and the 143 expression traits (2002 tests) on all the records. Eleven traits out of 14 showed a significant correlation with at least 1 gene expression trait. Four traits were discarded because the trait records overlap with less than 50% of subjects with gene expression profiled. The remaining 7 traits were found to be correlated with 1 to 7 gene expression traits. Details on the significant correlations are listed in Appendix 2.2.

2.4.4 Pathway Inference

The meat quality traits were analysed with the correlated gene expression traits and *CAST* genotype jointly in the two-step regression approach described in figure 2.3. The P -values for the QTL and the gene expression trait are listed in the table in Appendix 2.3. In the first step, except for C241L, the *CAST* genotype was not linked any of the meat quality traits in the regression model. The integrative genomic approach relies on the co-localisation of the QTL and eQTL. Because the merged data set was smaller than the meat quality trait dataset and was no longer

supportive of *CAST* being the QTL for those traits, it was not possible carry out the inference through to the second step.

For the trait C241L, a trait for the level of a fatty acid found in the *Longissimus thoracis*, both the QTL and the gene expression terms were significant in the first regression model. In the second regression model where the gene expression trait P5251 was fitted ahead of the QTL, the QTL was no longer significant. This indicates that the variance explained by the *CAST* genotype could be instead accounted for by the expression variation of P5251. Hence, the results imply the “causal model”; i.e. *CAST* → P5251 → Fatty Acid level. P5251 represented the clone id “SSH5A B07” on the microarray. Unfortunately this clone has not been sequenced by the company PIC, so the identity of the gene represented by this clone is unknown.

2.5 Discussion

One putative pathway has been suggested by the results. The true identity of gene expression trait P5251 is unknown because the clone has not been sequenced. Previous study on calpastatin discovered association of some of the alleles with cooking loss and juiciness (Ciobanu *et al.* 2004). Association with fatty acid level has not been previously implicated. Whether it is a novel discovery or a false positive can only be verified by separate experiments. However, this project has been terminated by PIC and the putative pathway is not pursued further in this thesis.

Considering the number of phenotypic traits and gene expression traits assayed, the outcome of this study is somewhat disappointing. An important factor to consider is the experimental design of the study. The gene expression experiment

was carried out by PIC as a bigger study with other research interests involved. As a result, many aspects in the experimental design were not favourable for the purpose of the current study. Hence, a critical assessment would be useful to highlight what could be learnt from this exercise.

Firstly, as a candidate gene approach, this study is based on previous findings in which the *CAST* marker is associated with meat quality. However, in the current dataset, the marker is significant for only a small number of meat quality traits. It should be noted that by using a 2-stage filtering method, the significant threshold used here for marker association is already very liberal as it does not correct for testing multiple traits. Even so, associations for cooking loss and juiciness by Ciobanu *et al.* could not be replicated. A possible explanation is that the original studies (Malek *et al.* 2001; Ciobanu *et al.* 2004) were based on a Berkshire x Yorkshire cross, whereas in this studies the genetic heterogeneity in five different lines of pig breeds could lead to substantial loss of power. In addition, it was observed in the results of marker association analysis that the line and batch effects were much more highly significant for many more traits compared to the genotype effect of *CAST*. This indicates that much of the variation in this dataset is simply due to breed and environmental differences. Also, the considerable number of missing records in the phenotype and the genotype data would have further reduced power (see the column “numObs” in Appendix 2.1).

Secondly, the number of microarrays was far fewer than the number of subjects there were phenotypic records for. In a joint analysis for pathway inference, combining the two sets of data led to the many phenotypic records being dropped. This led to dramatic loss of power in pathway inference because there was no

evidence in the reduced dataset to support the QTL. Without a significant QTL, the logic of the whole approach was violated because co-localisation of QTL is the “anchor” of the whole integrative genomic approach for pathway inference. This highlights that the number of animals profiled by microarrays has to match the number of animals phenotyped and genotyped for the integrative approach to be functional.

Thirdly, the lack of gene annotations for the microarray probes made interpretation and verification of results more difficult. For example, I could not follow up whether the marker had any effect on the gene expression levels of calpastatin itself or any of the calpains. There were several reasons for the lack of annotations: (1) the probes were designed from a cDNA library in which the clones had not been annotated; (2) the sequencing project for this array were prematurely terminated by the industrial partner; (3) some of the clones cannot be mapped to known genes; and (4) some probes were known to be badly designed in which they hybridise to multiple genes (C. Sargent, personal communication). In many ways, the fact that the genomic resources in the public sector for farm animals are lagging behind those for model species makes the eQTL analysis in pigs from a bioinformatics prospective more challenging. It has been demonstrated (Kadarmideen & Janss 2007) that a joint approach based on studying model organisms and comparative genomics could be a viable work-around until genomic resources for pigs and other farm animal species become sufficiently advanced for direct eQTL mapping.

Besides focusing on the imperfections in the dataset, it would also be useful to reflect on the assumptions underlying the experiment and consider the outcome of

the study from a different angle. To infer the gene regulatory mechanism behind meat quality from gene expression data, three assumptions were made; (1) that there is a complex gene network underlying the way calpastatin affecting meat quality; (2) it is the change of gene expression levels that drives the variation in meat quality traits; (3) the associations to the gene expression and the physiological traits are constant over time. Could it be that calpastatin and calpains act on meat quality through a basic mechanism without any intermediates? Or perhaps the underlying network is driven by variation at other levels, such as proteins or metabolites, rather than gene expression? What if the pathways related to meat quality were dynamic over time? Wu & Lin (2006) discussed the importance of the time dimension in genetic analysis. Furthermore, the original association reported (Ciobanu *et al.* 2004) were not exclusively due to the *Ser638Arg* mutation. Would a multiple-marker or a haplotype association test be necessary to reveal true associations? All of these open questions point to the fact that no single approach will be applicable to resolve all the complexities in a biological system. Beyond the genetical genomic framework proposed by Jansen and Nap (2001), it is critical to use a wide range of approaches, including both forward and reverse genetics, in deciphering the mechanisms underlying complex traits.

2.6 Conclusion

The current study showed that a sub-optimal experimental design can have a very negative impact on using genetical genomics for pathway inference. Studying large number of a test cross or a single uniform population is better than a pool of individuals from many different populations. Good data quality, large sample size

and matching records are all essential for the integrative genomic approach. Despite the recent successes in reconstructing gene networks in model organisms, there are still significant challenges in transferring the methodology to livestock species. Nonetheless, continual progresses in genomic research on livestock species will gradually reduce the hurdles in interpreting the biological meaning of eQTL and lead to more testable ideas.

Appendix 2.1

MeatQual	Line	Batch	CAST	numObs
BOHAM	9.29E-06	0.008423	0.015761	386
UADRENL2	0.001194	0.284061	0.040886	114
UNORADR2	0.000614	0.903293	0.024059	114
TBAL	0.032219	4.04E-26	0.044082	209
FFAS	0.002864	4.49E-06	0.021464	62
C180F	0.000565	3.92E-24	0.04165	189
C2033F	2.12E-15	6.86E-56	0.004776	189
C241L	0.291763	8.24E-11	0.017043	142
BMINOLTA	5.68E-06	8.21E-22	0.022695	407
AMINOL200	0.00038	0.180577	0.001021	213
BMINOL200	0.000638	1.82E-15	0.040821	213
SMCOHES_MEAN	0.011394	4.10E-09	0.005759	127
SMCOHES_MIN	0.01962	2.41E-08	0.020041	127
SMCOHES_MAX	0.136346	7.26E-05	0.018462	127

Meat quality traits with significant association with the *CAST* genotype. The table shows the *P*-values of line, batch and the *CAST* genotype for each meat quality trait and the number of observations after excluding missing values. Description of the traits can be found in Appendix 2.4.

Appendix 2.2

MQ	GE	numObs	pval	ci1	ci2
BOHAM	P1091	115	0.002778	0.098366	0.437459
BOHAM	P1188	115	0.004189	0.086204	0.427486
BOHAM	P1887	115	0.009173	0.061607	0.407061
BOHAM	P2535	115	0.009453	0.060625	0.406239
TBAL	P1419	115	0.009945	0.058959	0.404841
C180F	P1419	115	0.000627	0.13909	0.470254
C180F	P1576	115	0.007486	-0.412547	-0.068173
C180F	P4646	115	0.002619	0.100081	0.438858
C180F	P4842	115	0.00464	0.083105	0.424932
C180F	P4873	115	0.002456	0.10193	0.440365
C180F	P5240	115	0.00084	0.131476	0.464191
C180F	P5660	115	0.007032	0.070171	0.414212
C241L	P5251	98	0.007824	-0.442163	-0.072612
BMINOLTA	P83	115	0.00553	0.077707	0.42047
BMINOLTA	P1645	115	0.003292	-0.433395	-0.093398
BMINOLTA	P4693	115	0.004347	0.085084	0.426564
BMINOLTA	P4729	115	0.000682	0.136914	0.468524
AMINOL200	P1041	115	0.003605	-0.431186	-0.090704
AMINOL200	P1321	115	0.00393	0.088124	0.429066
AMINOL200	P4804	115	0.002851	0.09762	0.436849
BMINOL200	P814	115	0.00856	-0.408946	-0.06386
BMINOL200	P1321	115	0.001224	0.121399	0.456119
BMINOL200	P1439	115	0.0066	-0.41588	-0.072176
BMINOL200	P1640	115	0.002095	0.106476	0.444062
BMINOL200	P1645	115	0.001242	-0.455789	-0.120989

Meat quality traits and gene expression traits with significant correlation. The columns list the name of the meat quality trait, the identifiers for the gene expression traits, number of observations, the *P*-values of the correlation, and the 95% confidence interval for the correlation coefficient.

Appendix 2.3

Meat Qual	Gene Exp	QTL - 1	GE - 1	QTL - 2	GE - 2	num. obs.
BOHAM	P1091	0.08887	0.01384	0.121503	0.009828	94
BOHAM	P1188	0.08507	0.006079	0.03905	0.01431	94
BOHAM	P1887	0.09283	0.03248	0.16526	0.01691	94
BOHAM	P2535	0.083485	0.004306	0.01577	0.02792	94
TBAL	P1419	0.5058	0.9973	0.8146	0.5198	94
C180F	P1419	0.069824	0.445058	0.091859	0.281971	94
C180F	P1576	0.058251	0.014292	0.275248	0.002759	94
C180F	P4646	0.067237	0.187069	0.117529	0.089355	94
C180F	P4842	0.061776	0.038512	0.158531	0.013207	94
C180F	P4873	0.059338	0.019409	0.109664	0.009811	94
C180F	P5240	0.06361	0.06468	0.205159	0.016648	94
C180F	P5660	0.056914	0.009796	0.108939	0.004872	94
C241L	P5251	0.01485	0.02632	0.129008	0.00246	79
BMINOLTA	P83	0.09111	0.05638	0.37598	0.01141	94
BMINOLTA	P1645	0.09601	0.17603	0.22589	0.08051	94
BMINOLTA	P4693	0.09896	0.38073	0.1954	0.1414	94
BMINOLTA	P4729	0.09234	0.07442	0.13935	0.04528	94
AMINOL200	P1041	0.09095	0.17677	0.11268	0.13241	94
AMINOL200	P1321	0.08348	0.03006	0.290532	0.007636	94
AMINOL200	P4804	0.077475	0.007562	0.344969	0.001631	94
BMINOL200	P814	0.73883	0.06146	0.4636	0.1073	94
BMINOL200	P1321	0.7399	0.0779	0.96183	0.05743	94
BMINOL200	P1439	0.7432	0.166	0.8786	0.1341	94
BMINOL200	P1640	0.729506	0.009232	0.40617	0.01711	94
BMINOL200	P1645	0.721253	0.001939	0.479935	0.00289	94

Results of the 2-step regression model. The columns show the name of the meat quality trait, the identifiers for the gene expression traits, the *P*-values of the *CAST* genotype and gene expression for the first model, the *P*-values of the *CAST* genotype and gene expression for the second model, and the number of observations. In the first model, the QTL is significant ($\alpha = 0.05$) only for the meat quality trait C241L.

Appendix 2.4

MeatQual	Description
BOHAM	Weight of bone in ham
UADRENL2	Adrenaline level in the urine collected after the transportation to the abattoir
UNORADR2	Noradrenaline level in the urine collected after the transportation to the abattoir
TBAL	TBA measured in Longissimus thoracis
FFAS	A trait on lipid fraction
C180F	Composition of a fatty acid
C2033F	Composition of a fatty acid
C241L	Composition of a fatty acid
BMINOLTA	Yellowness measured in the LT muscle at the last rib level
AMINOL200	Redness measured in the LT muscle at the last rib level
BMINOL200	Yellowness measured in the LT muscle at the last rib level with a new machine
SMCOHES_MEAN	Mean Semimembranosus cohesiveness
SMCOHES_MIN	Minimum Semimembranosus cohesiveness
SMCOHES_MAX	Maximum Semimembranosus cohesiveness

Description of traits shown in Appendix 2.1

CHAPTER 3

Experimental design for genetical genomics

A candidate gene approach to map a large number of expression traits assayed by microarrays was presented in chapter 2. One of the main reasons for why the methodology to infer pathways did not work as well as we hoped was due to the small number of subjects with expression data. Unless a study samples a very large number of individuals, most except the largest QTL effect on the expression trait will be missed (Rockman & Kruglyak 2006). Often, small sample size in gene expression research is a direct result of the high cost of microarrays. To reduce the cost or make more efficient use of microarrays can potentially facilitate bigger experiments. Several different microarray designs are routinely used for analysing differential gene expression (Simon *et al.* 2003). However, these designs are not necessarily optimal for mapping eQTL. Since the rise of interest in genetical genomics, several articles proposed new microarray designs relevant to genetic analysis of gene expression (Jin *et al.* 2004; Piepho 2005; Fu & Jansen 2006; Bueno Filho *et al.* 2006). Indeed, experimental design for genetical genomics is an active area of research.

In the current chapter, I present a review on the general technology behind microarrays, several generic microarray designs, and the more specific designs for genetical genomics. Then, I proceed to describe work on extending the “distant pair design” (Fu & Jansen 2006) for outbred test crosses that are populations typically

used in livestock genetics. The assessment of the effectiveness of this design is illustrated by a simulation study.

3.1 Introduction

3.1.1 Microarrays for gene expression profiling

The original use of microarrays is to quantify the transcript abundance in a collection of cells in a highly multiplex fashion. More recently there have been rapid developments on novel types of microarrays that are manufactured for other uses such as high-throughput genotyping and high resolution copy number variation detection (Fan *et al.* 2006). Here I introduce the basic technology underlying the gene expression arrays. On a microarray there are thousands of immobilised probes that represent genes in the genome. Fluorescent or biotin labelled targets (cDNA or cRNA converted from total RNA sample extracted from the cells) bind to the probes during a hybridisation reaction. The quantities of labelled targets can be estimated by measuring the signal intensity over the probes on the array. For gene expression microarrays there are two main platforms: one-colour and two-colour microarrays.

3.1.1.1 One-colour platform

The most common one-colour microarray is the oligonucleotide array manufactured by the company Affymetrix, known as the GeneChip™ (www.affymetrix.com). On a silicon chip the oligonucleotide probes are lithographically synthesised in parallel. The oligonucleotide probes used are relatively short, 25-mer, and a given transcript is represented by a probe set of multiple probes (commonly 11 or 16 probes in a set). With a one-colour system, a

single sample is hybridised to a single array. Therefore, to assay n samples, n microarrays are required in the experiment. The expression level of a gene of a subject is quantified as the signal intensity of the relevant summarised probeset on the corresponding array.

3.1.1.2 Two-colour platform

Two-colour microarrays are usually made up of glass slides with cDNA probes that are 200 – 1200 bases long spotted by robotic printers, although two-colour oligonucleotide arrays (with probes about 60 bases long) exist, such as the Agilent arrays (www.agilent.com). Two samples that are differentially labelled with the dyes CY5 (red) and CY3 (green) can be co-hybridised on a single array and intensities of each channel are captured by an array scanner. Although it is possible to simply treat each probe intensity of a dye on a single array as the expression level of a gene of a subject, this is not usually advisable because the higher inter-array variability that exist for two-colour platforms (Simon *et al.* 2003). How individual gene expression levels are quantified depends on the choice of microarray design.

3.1.2 Generic microarray designs

Generally, microarray designs regarding sample allocation do not apply to one-colour platforms, except in “selective phenotyping” scenarios which will be described in later sections. However, for two-colour platforms where two samples are co-hybridised on a single array, there are several ways to pair up the samples in an experiment. Three generic designs, the common reference design, the loop design, and the block design, are popular designs for two-colour microarrays.

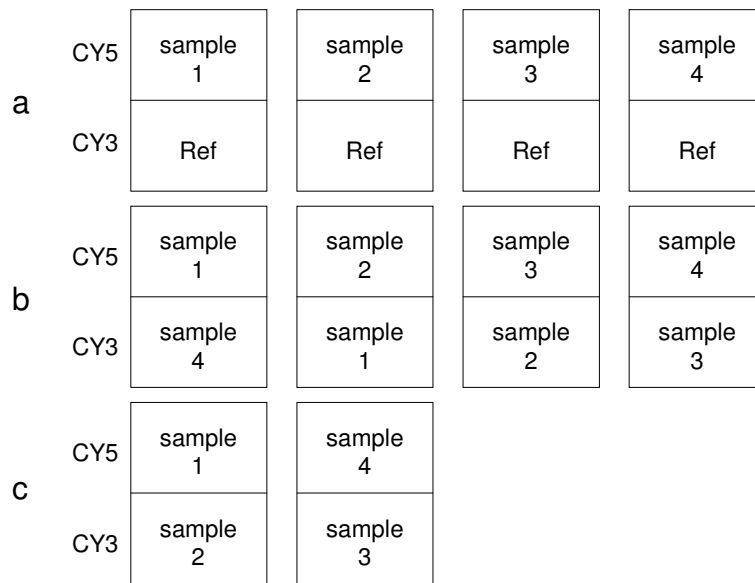


Figure 3. 1

Three generic microarray designs for 4 samples. (a) The Common Reference Design. Each sample is compared with a common reference sample on the same array. Thus, the log intensity ratio of a probe on one array can be directly compared to that on another array. (b) The loop design. Each sample is labelled with CY5 in one array and CY3 in the next array, so all the samples are linked in a loop and measured in both channels. A complex ANOVA analysis is required to compare the expression values of all the samples. (c) The block design. This is a balanced block design with dye-balance built-in the design. In this example, sample 1 and 3 are biological replicates of one treatment group, and sample 2 and 4 are biological replicate of another treatment group. Two arrays are sufficient to accommodate 4 samples.

3.1.2.1 The common reference design

A reference sample is fluorescently labelled with one dye and mixed with another sample labelled with a different dye for co-hybridisation on every array in the common reference design. The intensity of a probe for the sample is measured relative to the intensity of the same probe of the reference sample on the array. The ratio of the two channels transformed in logarithmic scale (log-ratio) is then used as the expression value for the gene the probe designed for. The advantages of this design are that (a) it guards against variation in size and shape of corresponding spots on different arrays; (b) the log-ratio is a standardised value to the common reference sample and can be directly compared across arrays; (c) this design is robust to array failures and flexible enough to accommodate additional conditions to the experiment. On the other hand, the common reference design is not an efficient design, since half of the microarray resource goes into the reference sample which often has no biological relevance. For n microarrays, the common reference design can profile the expression of n samples.

3.1.2.2 The loop design

No reference sample is required with the loop design (Kerr & Churchill 2001); instead the first sample is co-hybridised with the second sample, and the second sample is used again in co-hybridisation with the third sample and so on until a loop is formed. While a sample is labelled with CY5 on one array, on the next array it is labelled with CY3, hence dye balance can be achieved and the effect of dye can be accounted for. The loop design therefore produces twice as much data for any one gene compared to the common reference design with the same number of arrays.

Furthermore, the estimated expression values have smaller standard errors in the loop design than in the common reference design and, hence, use of the loop design increases the precision in the comparison. However, contrasting two samples that are far apart in the loop involves modelling many indirect effects and reduces the margin on higher precision over the common reference design (Dobbin & Simon 2002). Also the loop design is not robust against bad quality arrays as a single sub-standard array can seriously affect the estimation of the levels of gene expression in all samples. Furthermore, the loop design does not improve on the number of subjects profiled using n microarrays compared to the common reference design.

3.1.2.3 The block design

Often in differential expression analysis, subjects from two groups are being compared. In a block design, biologically independent samples, one from each group, are paired for co-hybridisation. If one half of the samples from one group are labelled with CY5 on half of the arrays and the other half with CY3 on the other half of the arrays, and vice versa for the other group, the design would be known as the “balanced block design” (Dobbin & Simon 2002). By balancing the sample assignment to the two dyes, the design attempts to minimise the bias due to dye-specific hybridisation. The block design is the most efficient design as $2n$ samples are profiled using only n microarrays. Comparison of the two groups should be made within-arrays. Comparisons across arrays are subjected to noise resulting from the variation of spots and arrays, although it can be done using a suitable linear mixed model. Other drawbacks include (a) data from this design cannot be easily adapted in more complex experiments where different ways of contrasting different groups are desired; (b) it requires arbitrary pairing and is less effective than the common

reference design when there is large inter-sample variability (Dobbin & Simon 2002).

3.1.3 Microarray experimental designs for genetical genomics

Generic microarray designs for two-colour platforms, particularly the common reference design, are very often chosen for their straight-forwardness in practical implementation. Although the majority of eQTL mapping experiments in the literature to date were carried out using one-colour platforms, there have been examples, (Schadt *et al.* 2003; Monks *et al.* 2004 and others), where the common reference design in two-colour platforms was utilised. One interesting question is whether different ways of utilising the microarray resource exist that are specifically optimised for genetical genomics studies. To answer this question, several groups have put forward experimental design strategies for genetical genomics; some of these strategies are even applicable to one-colour platforms. All of these strategies stem from one motivation: to achieve maximum statistical power using a given number of microarrays. However, the research focus for which these strategies are optimised can be wildly different.

3.1.3.1 Selective phenotyping

This strategy is a generic experimental design which can be applied to general QTL studies, but its benefit is particularly evident in eQTL studies because whole-genome expression profiling is usually far more expensive than traditional phenotyping of physiological traits. The fundamental idea of selective phenotyping is to select the best subset from a larger population for phenotyping, based on the genotypes. This assumes that within a population, such as an intercross, some

progeny are more informative than others. Phenotyping such as a subset of samples can give almost the same performance as phenotyping the full set of samples; hence it represents significant increase in the efficiency of resource utilisation. This is complementary to the idea of selective genotyping, first introduced by Darvasi & Soller (1992), where subjects at both extreme ends of the phenotypic distribution were selected for genotyping because, traditionally, genotyping had been more costly than phenotyping. For genetical genomics, selective phenotyping concerns the choice of subjects for transcript profiling. This general strategy is, therefore, applicable to both types of microarrays to map eQTL.

Different implementations of selective phenotyping exist, and each of these methods has its distinct features. Jannink (2005) argued that the number of recombination breakpoints vary among progeny in any given sample of recombinant progeny. Having genotyped the whole population samples, the progeny with the highest number of recombination events will form the optimal set for phenotyping. This is because most of the power for resolving QTL locations lies with the samples for which the linkage disequilibrium extends over shorter distances. Here, the focus is to optimise the QTL position accuracy.

An alternative view on selective phenotyping is to focus on the power of QTL detection. This can be done by selecting individuals with the highest genotypic dissimilarity (Jin *et al.* 2004). For a given number of subjects to be phenotyped, this method ensures that genotypic contrast is maximised over multiple loci; hence it maximises the power for distinguishing the effects of alternative genotypes at a locus.

More distinct from the two selective phenotyping strategies described above was presented by Nettleton & Wang (2006). Here, their selective strategy is specific for transcription profiling, and the goal of the optimisation is to maximise the power of detecting expression traits that are linked to a previously identified QTL for a traditional phenotype. Hence, the interest of the experiment is to identify pleiotropic QTL; loci that are both QTL for one or more traditional phenotypes and eQTL for one or more expression traits. This strategy not only assumes the genotypes for all markers are available for all individuals, but also the phenotypic values of the traditional quantitative trait. Given the location of the previously identified QTL, individuals with the maximum genotypic dissimilarity as well as the extreme (highest and lowest) phenotypic values are selected for transcriptional profiling.

3.1.3.2 Optimal sample allocation

Contrasting to selective phenotyping, optimal sample allocation is an alternative class of microarray experimental design for genetical genomics which does not emphasize selecting a subset from a larger pool of subjects. Instead, it concentrates on the allocations of samples on two-colour microarrays; i.e. which two subjects to co-hybridise on the same array. Therefore, this class of strategies is specific for eQTL studies and the two-colour platforms of microarrays.

Piepho (2005) discussed the interesting subject of detecting heterosis in gene expression traits. The objective of an experiment to which his strategy applies is to detect over-dominance in gene expression traits rather than mapping locations of eQTL. When the test population consists of two parental inbred strains and the hybrid strain, the optimal allocation will favour parent-hybrid pairs, while parent-parent pairs or hybrid-hybrid pairs will be selected against.

Bueno Filho *et al.* (2006) presented various scenarios using examples based on marker-trait association to a single locus. They argued that for detecting additive effects, the optimal allocation would be to co-hybridise only the homozygous individuals, whereas co-hybridising homozygotes with heterozygotes would be the optimal allocation for detecting dominance effects. Various less structured designs were also presented for the detection epistatic interaction between two loci, and the estimation of heritability and treatment effects with pedigree samples.

3.1.3.3 Distant pair design

The distant pair design (Fu & Jansen 2006) applies selective phenotyping and sample allocation based on genotypic dissimilarity for genome-wide eQTL mapping simultaneously. This design achieves optimality in detection power and efficiency by computing the best overall pairing configuration that maximises the number of genotypic contrast over the whole genome for a given number of arrays. The algorithm behind distant pair design is called simulated annealing (Kirkpatrick *et al.* 1983) which effectively carries out a computer search over a vast number of possible configurations. The implementation outlined in their article is applicable to recombinant inbred lines (RILs); they are inbred strains with a fine mosaic genome of homozygous loci derived from the founder lines. Hence, only additive eQTL can be found with the experimental setup. The optimal design for detecting additive effect using analysis of variance is the design which minimises the variance of the estimated additive effect. This turns out to be one that maximises genotypic dissimilarity within-array and maintains a balance on genotype-to-dye assignment; i.e. a genotype group is assigned to CY5 dye in approximately as many arrays as to

CY3 dye. Over a genome, it was shown that the distant pair design created more genotypic contrast than it would have by the generic balance block design; the power of additive eQTL detection is superior to the loop and the common reference design; and it uses the full capacity that a two-colour microarray system can offer, which is to profile $2n$ samples with n number of microarrays.

3.1.4 Distant pair design for outbred crosses

The distant pair design was shown to be the optimal design for genetical genomics in RILs. To date, with the exception of humans, almost all genetical genomics studies have been carried out in model organisms. However, if the mapping of eQTL was to be applied to livestock species or other non-model species, it would be worthwhile to investigate how the distant pair design might perform in those populations. For researchers studying genetics of outbred species, mapping resources like inbred strains or RILs are often not feasible. By contrast, F_2 intercrosses between two genetically divergent outbred populations are much more readily available. A major complication arising in outbred crosses is due to the fact that there are common sets of alleles segregating in both of the founder populations. Hence, it is often the case that marker genotypes in the F_2 generation would not be fully informative for the origins of lineage at any given locus. This uncertainty obscures how one can define genotypic dissimilarity for the purpose of pair assignment in distant pair design. In addition, F_2 intercrosses (whether inbred or outbred lines) present extra complexity over RILs: the researcher has the option of discovering additive as well as dominance effects. This option can lead to difficulties in defining the optimal pair assignment because a pairing configuration that is optimised for detecting additive effects might be very poor for detecting dominance

effects. Moreover, there is also the issue regarding large genome sizes. It is expected that when genome size increases, finding distant pairs will become more and more difficult. Fu and Jansen (2006) have shown that in RILs a small advantage is achievable with large genomes. However, whether this advantage is also present in an F_2 design remains uncertain. This question is directly relevant to researchers who are interested in studying the genetics of gene expression in livestock species which typically has a large genome. Therefore, the usefulness of the distant pair design for genetical genomic studies in outbred F_2 crosses warrants investigation.

3.2 Methods

3.2.1 QTL analysis

The method for mapping QTL follows the least squares approach (Haley *et al.* 1994). Briefly, the line origins at fixed intervals (e.g. 1 cM) along the genome for the individuals in the F_2 generation are expressed as lineage probabilities, conditional on the marker genotype. This can be done by considering all possible line origin combinations based on the parental and grandparental genotypes, and has been implemented in the online software “QTL Express” (Seaton *et al.* 2002). Assuming that founder lines are fixed for alternative QTL alleles, the lineage probabilities can be used to predict the putative QTL genotypes. Phenotypic values are then regressed onto genetic coefficients calculated for a putative QTL at a fixed position. The genetic coefficients for additive and dominance effects are derived from the conditional probabilities: the additive coefficient (denoted x_a) is the difference of the probabilities for the homozygous line origins, and the dominance coefficient (denoted x_d) is the sum of the probabilities for the heterozygous line origins. An F

ratio test statistic can be used to test the null model (without QTL fitted) against the full model (with QTL fitted) and determine the significance of the presence of QTL. For full details on the derivation of line origin probabilities and regression-based QTL mapping, see (Haley *et al.* 1994).

In the context of a pair design in two-colour microarrays, the gene expression phenotypes can be expressed either in ratios or in signals of the separate channels. In this article ratios over signals are chosen as the phenotypes because the use of ratios can minimise the risk of bias as a result of spot or array effects (Wit & McClure 2004). Fu & Jansen (2006) has argued that there is negligible difference in the final results between ratios and signals, provided that the distributional assumptions for the array and spot effects used in the signal based analysis are correct. The log-ratio of the red channel intensity to the green channel intensity of a probe is equivalent to the difference of the two signal intensities in logarithmic scale. To utilise such phenotypes in the Haley-Knott least squares framework, the linear regression model can be written as:

$$\Delta y_i = \mu + \Delta x_{ai}a + \Delta x_{di}d + e_i \quad (1)$$

where Δy_i is the difference by subtracting the log signal of the green channel from that of the red channel for the i th microarray ($i = 1, \dots, n$); μ is the overall mean; Δx_{ai} is the difference of the additive coefficients by subtracting x_a of the individual assigned to the green channel from x_a of the individual assigned to the red channel for the i th microarray; Δx_{di} is the coefficient difference for dominance x_d ; a and d are the additive and dominance parameters respectively; and e_i is the residual error. In matrix form, the expression can be simplified as $Y = Xb + e$, where $b = (\mu, a, d)^t$.

3.2.2 Finding optimal pairs

The optimal design is defined as the configuration with the minimum for the sum of the variances for the estimated b term, or \hat{b} , in the matrix form of the model above. Following the A-optimality criterion (Wit & McClure 2004), this is equivalent to minimising the trace of $(X'X)^{-1}$. For the regression model in (1), the matrix X consists of a column of 1's for the mean μ , a column of Δx_a coefficients for the additive parameter, and if dominance is included in the model, a third column of Δx_d coefficients. To reach the optimal pairing design over all positions in the genome, I search for the minimum of S , or the sum over all marker loci the trace of $(X'X)^{-1}$. Genetic coefficients at marker loci only are used for optimisation in order to keep the computation tractable.

The simulated annealing technique (Kirkpatrick *et al.* 1983) was used to find the optimal pairing configuration. The procedure was very similar to that used in the original distant pair design (Fu & Jansen 2006). The search was iterative and at any particular iteration step compared the current design with a slightly modified version: samples of two randomly chosen pairs (a, b) and (c, d) in the current design were randomly re-paired in the new design. The new design was accepted if it was better (has a lower value of S) than the current design. It is useful to occasionally accept worse designs with a certain probability to be able to move away from “locally optimal” designs. This probability was $(S_{\text{old}}/S_{\text{new}})^{1/T}$, where T was a tuning parameter that was slowly decreased towards zero during the iterative process. This iterative process was terminated when T became very small, around 1×10^{-40} and 1×10^{-50} . The implementation of finding optimal pairs was accomplished using the R statistical computer program (R Development Core Team 2007).

3.2.3 Power assessment via simulations

Three different genome sizes were studied: 100 cM, 1000 cM and 2000 cM; and for each genome size 100 replicates of F_2 intercrosses were simulated. Firstly, F_1 individuals were generated by randomly mating 20 F_0 sires from founder line one to 80 F_0 dams from founder line two (four dams per sire), each having 5 offspring. Then, another 400 offspring were generated in the F_2 generation by randomly mating 20 F_1 sires to 80 F_1 dams (5 progenies per mating). Marker data were simulated for all samples, with 11 evenly spaced markers per chromosome of 100 cM in length. Four alleles were simulated for every marker segregating at equal frequencies in both founder lines, with marker genotypes in Hardy-Weinberg equilibrium. A single bi-allelic QTL that is fixed for alternative alleles in the founder lines was simulated on the first chromosome at 46 cM. For this QTL, I simulated two alternative settings: (a) an additive QTL without dominance where the homozygous genotypic value $a = 0.5$ and the heterozygous genotypic value $d = 0$; (b) a QTL with complete dominance where $a = 0.5$ and $d = 0.5$. Polygenic background effects were modelled as ten unlinked bi-allelic loci, each with an additive effect of 0.25 and segregating at a frequency of 0.5 in both founder lines, as described in Alfonso & Haley (1998). To mimic the non-genetic factors affecting the gene expression phenotype and technical errors of microarrays, I added an environmental component sampled from a normal distribution with a variance of 0.5 to the simulated phenotype. The narrow-sense heritability (h^2) is 0.47 for the trait and 0.20 for the main QTL on the first chromosome.

To assess the performance of the optimal pair design under the least squares framework, I scanned in 1 cM steps for the most significant p -values obtained in the

marker interval which contains the QTL (between 40 and 50 cM on the first chromosome) under four scenarios. These four scenarios are summarised in Table 3.1 and are described as follows: first, all 400 F_2 subjects and their individual phenotypic measurements were analysed. Conceptually this is equivalent to the common reference design that includes all F_2 individuals. Second, 200 F_2 subjects were randomly selected, together with their individual phenotypic measurements. This scenario also represents the common reference design, but a smaller budget limits the profiling of gene expression to fewer individuals than in the first scenario. Due to the random sampling nature of this scenario, for each simulated population replicate I repeated the random sampling 100 times, and scanned for the most significant p -value in the QTL-containing interval as above. Then the median p -value was selected to represent the performance under this scenario for the given population replicate. Third, I randomly paired up all 400 F_2 subjects and analysed the data with regression model (1). Under this scenario, I also repeated the process 100 times per simulated population replicate and proceeded to obtain the p -value in the same way as in the second scenario. Last, I paired up all 400 F_2 subjects using the optimal pair design. I abbreviate these four scenarios above as “all.data”, “half.data”, “ran.pair” and “opt.pair”, respectively, for reference in the rest of this article. For both “additive only” and “additive and dominance” QTL settings, the data were analysed under those four scenarios.

	abbreviation of the scenarios	description	no. of F₂ subjects profiled	no. of slides required
1	all.data	individual phenotypic values are available for all subjects	400	400
2	half.data	Same as all.data except that 50% of the subjects are selected	200	200
3	ran.pair	pairs are assigned randomly	400	200
4	opt.pair	pairs are assigned according to the outcome of simulated annealing	400	200

Table 3. 1

Summary of the four scenarios investigated in the power study

3.2.4 Alternative marker allele frequencies and population sizes

In the simulations above the marker allele frequencies are equal over all four alleles in both founder lines. This represents a suboptimal scenario in which the marker genotypes in the F_2 generation are expected to have limited information for the line origins. For the genome size of 2000 cM, I also simulated the “best-case scenario” in which each founder line has two unique alleles; i.e. two out of the four alleles are segregating within each founder line, with no common alleles shared by both lines. Such an intercross is equivalent to an F_2 cross between two inbred lines. These two sets of marker allele frequencies would enable us to determine a below average range and the upper bound for the performance of the optimal pair design. In addition, I performed further simulations in which I fixed the number of microarrays being used to 400, and evaluated an F_2 population size of 1000. I compared the performance of the optimal pair design and the common reference design when expression profiling of every individual in the sample population is not possible.

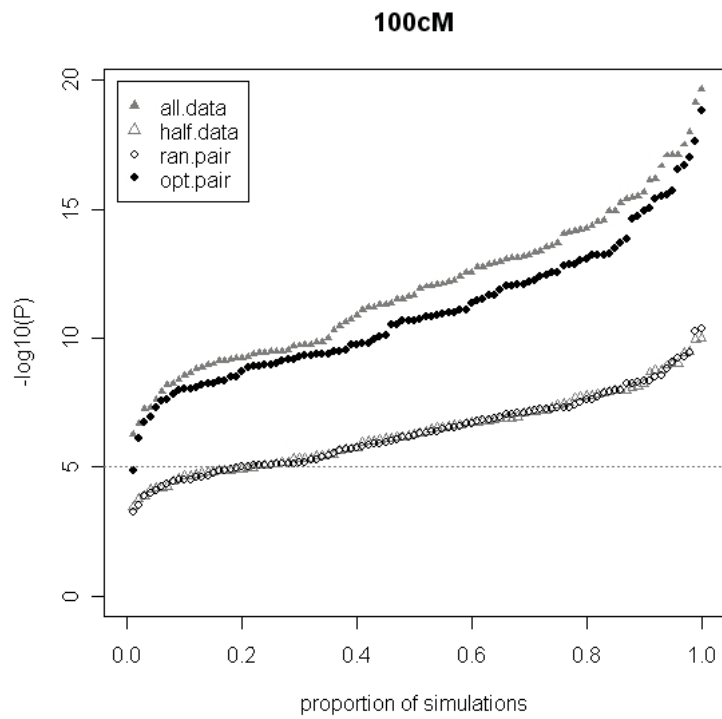
3.3 Results

3.3.1 Additive effect

The power for detecting additive QTL under the four scenarios was investigated. For the results of “opt.pair” presented in this section, I minimised the variance of the additive effect in the regression model by simulated annealing. Figure 3.2 shows the minus log-transformed p -values (sorted in ascending order) for the four scenarios. The scenario with the highest proportion of the largest minus log-transformed p -values can be considered as the most powerful design. For a single

chromosome (Figure 3.2a), the most significant p -values can be found under the “all.data” scenario. But for the “opt.pair” scenario, under which only 200 microarrays would be required, the power to detect the QTL is remarkably close to that under the “all.data” scenario. Under the “half.data” and “ran.pair” scenarios, likewise, only 200 microarrays would be required, but the power is much reduced compared to both “all.data” and “opt.pair”. Incidentally, the performance of “half.data” and “ran.pair” are almost identical, hence most of the data points for these two designs are overlapping on Figure 3.2.

(a)



(b)

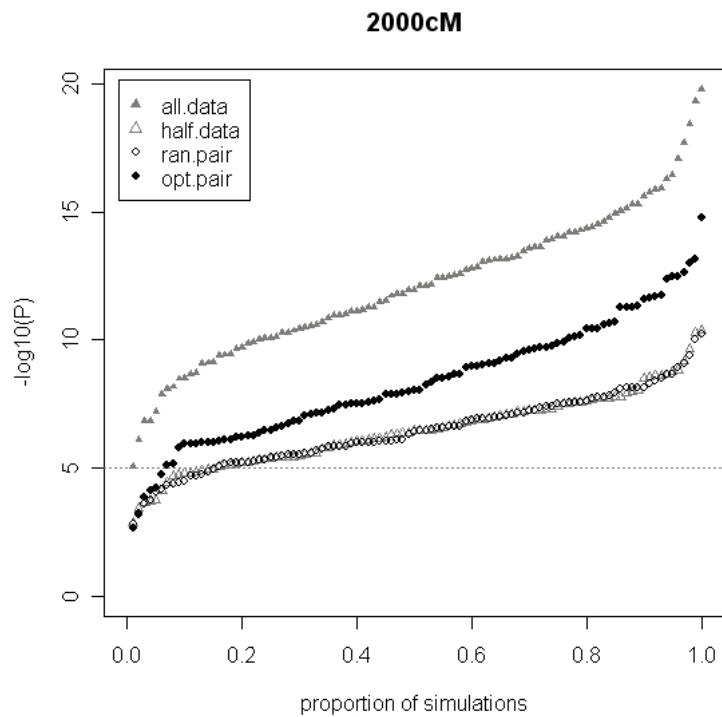


Figure 3. 2

Performance for detecting additive QTL effect under various scenarios. (a) genome size of 100 cM, a single chromosome; (b) genome size of 2000 cM, twenty chromosomes. Horizontal dotted line shows

the significant level of $P = 1 \times 10^{-5}$. The simulations are sorted in ascending order of the $-\log_{10}P$ on the x-axis.

Table 3.2 summarises the performance under the four scenarios by the mean $-\log_{10}P$ and shows the effect of genome size on the power for detecting QTL. The mean $-\log_{10}P$ across different genome sizes under the “all.data”, “half.data” and “ran.pair” scenarios show little deviation. However, the mean $-\log_{10}P$ under the “opt.pair” scenario follows a notable downward trend when the genome size increased. At the genome size of 2000 cM (Figure 3.2b) “all.data” performs best out of the four scenarios. But more importantly, “opt.pair” scenario is the most powerful out of the designs that require 200 microarrays.

Genome	no. of								
size	chr	all.data		half.data		ran.pair		opt.pair	
100 cM	1	11.9	(2.9)	6.4	(1.5)	6.3	(1.5)	11.0	(2.7)
1000 cM	10	12.3	(2.6)	6.6	(1.3)	6.6	(1.4)	9.2	(2.4)
2000 cM	20	12.1	(2.9)	6.5	(1.5)	6.4	(1.5)	8.3	(2.4)
2000 cM *	20	12.9	(2.9)	6.9	(1.5)	6.8	(1.5)	8.9	(2.4)

Table 3. 2

Summary of *P*-values (mean, and standard deviation on -log₁₀ scale) at the main QTL position for additive QTL detection under the four scenarios, where only an additive effect was simulated. Standard deviations are shown in bracket.

* inbred line cross

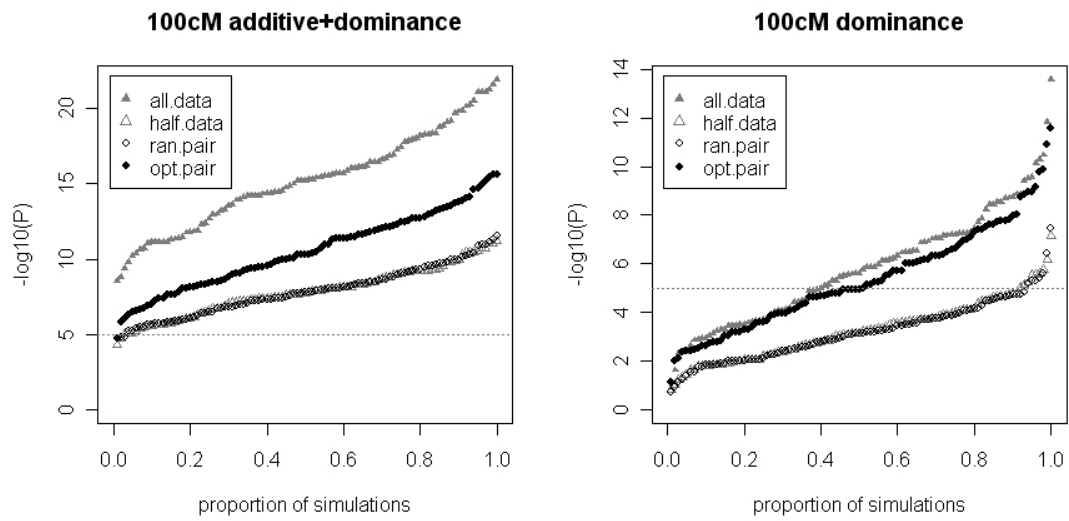
I analysed the simulations of F_2 cross with fully informative markers for the genome size of 2000 cM and found that the power increased slightly under all four scenarios (Table 3.2). The increase in power is expected because line origins can be inferred with certainty. It is important to note that the difference in the power between the suboptimal and the best-case scenario for the marker allele frequencies is small. This indicates that the power assessment using equal marker allele frequencies in the simulations is robust and representative of real outbred F_2 intercrosses, of which the marker allele frequencies in the founder lines are in between those two extremes.

3.3.2 Additive and dominance effects

For the dominant QTL, two levels of analysis were carried out: (a) QTL detection by comparing the full model (additive + dominance) to the null model; and (b) detection of dominance effect by comparing the full model to the reduced model (additive only). In the simulated annealing step of optimal pairing, the dominance coefficients were included as the third column in the matrix X in the linear model (see the methods section).

With a single chromosome (100 cM) genome, the power to detect QTL under the “opt.pair” scenario is clearly lower (Figure 3.3a, left panel) than “all.data”. It can be seen in Table 3.3 that the mean $-\log_{10}P$ under “all.data” is approximately 50% greater than that under “opt.pair”. But “opt.pair” is still more powerful than both “half.data” and “ran.pair”. By contrast, the results (in Figure 3.3a, right panel) show that the “opt.pair” and “all.data” are similarly powerful for detecting dominance effects and superior to both “half.data” and “ran.pair” in a small genome similar to the one simulated here.

(a)



(b)

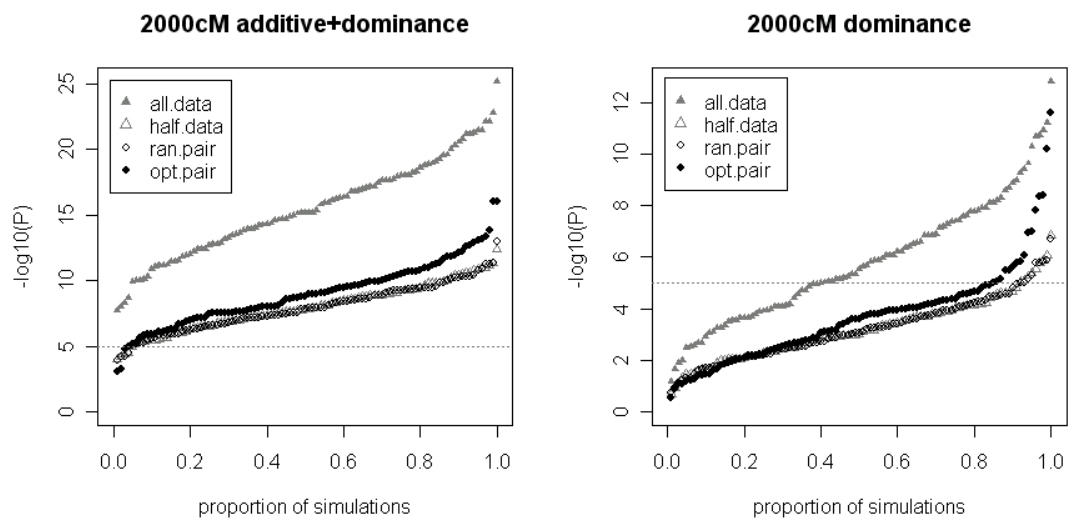


Figure 3.3

Performance for detecting QTL (on left panels) and dominance effects (on right panels) under various scenarios. (A) genome size of 100 cM, a single chromosome; (B) genome size of 2000 cM, twenty chromosomes. Horizontal dotted line shows the significant level of $P = 1 \times 10^{-5}$. The simulations are sorted in ascending order of the $-\log_{10}P$ on the x-axis.

genome									
size	no. of chr	all.data		half.data		ran.pair		opt.pair	
100 cM	1	15.2	(3.2)	7.8	(1.6)	7.8	(1.7)	10.5	(2.6)
1000 cM	10	15.5	(3.3)	8.0	(1.7)	7.9	(1.7)	9.6	(2.3)
2000 cM	20	15.4	(3.7)	7.9	(1.8)	7.9	(1.8)	8.9	(2.5)

Table 3. 3

Summary of *P*-values (mean, and standard deviation in -log₁₀ scale) at the main QTL position for QTL detection (additive + dominance model Vs the null model) under the four scenarios, where both additive and dominance effects were simulated. Standard deviations are shown in bracket.

Table 3.3 and Table 3.4 show that genome sizes again have little effect on power under “all.data”, “half.data” and “ran.pair”. However, the increase in genome size affects optimal pairing more severely here than when no dominance effect has been simulated. At the genome size of 2000 cM, “opt.pair” is only marginally more powerful in detecting the QTL than “half.data” and “ran.pair” (Figure 3.3b, left panel). The power for detecting dominance effect is more drastically affected and “opt.pair” performs similarly to “half.data” and “ran.pair” (Figure 3.3b, right panel). Therefore, in the presence of dominance effects, the advantage in the performance of the optimal pair design in detecting QTL is reduced. Including dominance in the optimisation has a negative impact on the optimal pair design, especially for large genome sizes, when QTL detection is the primary objective.

genome									
size	no. of chr	all.data		half.data		ran.pair		opt.pair	
100 cM	1	5.8	(2.4)	3.3	(1.2)	3.2	(1.2)	5.4	(2.1)
1000 cM	10	5.7	(1.7)	3.1	(0.9)	3.1	(0.9)	3.8	(1.5)
2000 cM	20	5.8	(2.4)	3.2	(1.2)	3.2	(1.2)	3.7	(1.9)

Table 3. 4

Summary of P -values (mean, and standard deviation on $-\log_{10}$ scale) at the main QTL position for dominance detection (additive + dominance model Vs additive model) under the four scenarios, where both additive and dominance effects were simulated. Standard deviations are shown in bracket.

3.3.3 Fixed number of microarrays with a large F_2 sample size

In previous simulations, I observed that “all.data”, which required 400 microarrays, was more powerful in detecting additive QTL effect than using 200 microarrays under the “opt.pair” scenario. Here, I studied the power of these two designs conditioned on a total of 400 microarrays. With F_2 sample size of 1000, neither design can profile all the individuals with 400 microarrays. Under the optimal pair design, 400 pairs were deliberately selected to give the minimum variance for the estimated additive genetic parameter. On the other hand, only 400 individuals (randomly selected from 1000 individuals) could be profiled using the common reference design. Given equal number of microarrays being used, the results in Figure 3.4 show that the optimal pair design outperforms the common reference design.

2000cM, 1000 F2 individuals with 400 arrays

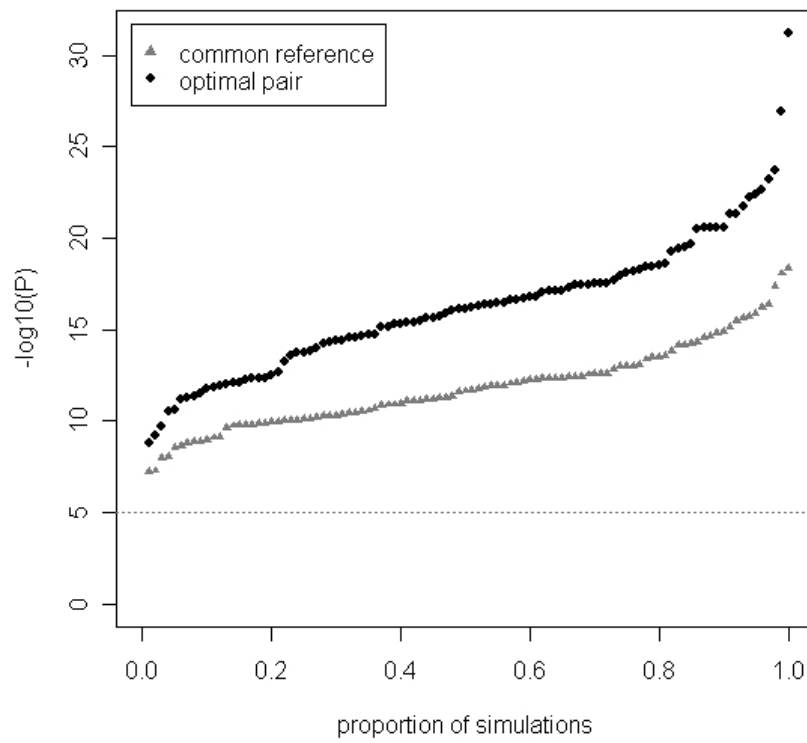


Figure 3. 4

Comparison of the performance for QTL detection under common reference design and optimal pair design when the number of arrays is fixed as 400, and genome size of 2000 cM with the F_2 sample size of 1000. Horizontal dotted line shows the significant level of $P = 1 \times 10^{-5}$. The simulations are sorted in ascending order of the $-\log_{10}P$ on the x-axis.

3.4 Discussion

The distant pair design enables the mapping of eQTL in an efficient and effective manner using recombinant inbred lines. For researchers studying genetics of many outbred species, however, the creation of recombinant inbred lines is impractical. Here I explore whether eQTL studies of natural species would benefit from the same design principles used in “distant pairing”. It is shown that the optimal pair design, an extension of the distant pair design for outbred lines crosses, can indeed improve the efficiency of the use of microarrays and increase the statistical power for detecting eQTL, even for studying organisms with large genome sizes.

Under the linear regression framework, the greatest power is achieved by having the regression coefficients in equal proportions near the top and bottom extremes. For the regression model proposed for the optimal pair design in this article, this would be achieved by pairing up individuals who have large genetic coefficients with opposite signs. However, in a line cross such as the F_2 , it is inevitable that not every pair would result in a regression coefficient that is near one extreme or the other. Furthermore, when the number of independent loci increases (increase in chromosome length and number of chromosomes), the optimal pair assignment for one locus will usually not be optimal for the other loci. The optimal pair assignment over the whole genome is therefore sub-optimal in the perspective of a single locus; i.e. fewer regression coefficients around the extremes. One can therefore expect the performance of distant pairing to degrade to the same level as random pairing eventually as the genome size continues to increase.

3.4.1 Clear benefits in detecting additive effects

It is shown that when there are few loci to consider, such as in a small genome, the power of detecting additive effects with the optimal pair design is similar to using a common reference design that consumes twice the number of microarrays. With near-optimal pairing for individual loci (achievable when there are small number of effectively independent loci), the efficiency of the optimal pair design is very attractive. Moreover, the common reference design with only half the sample size (i.e. the same number of microarrays) performs significantly worse. This highlights the problem of small sample size leading to reduction in power in complex trait analysis.

As expected, the performance of the optimal pair design drops when the genome size increases. Nevertheless, it is very promising that in a large genome the optimal pair design still notably outperforms designs which use the same number of microarray slides. Furthermore, as shown by the excellent performance in smaller genomes, it is evident that the optimal pair design would be beneficial for a focused study of one or more candidate regions within a large genome. The power can be maximised for genomic regions for which the researchers have the most interest, while the power in the rest of the genome would be at least as good as the random pair design. In addition, the results show that with the number of microarrays used being equal, the optimal pair design always gives the highest statistical power of the approaches compared. Therefore, for outbred species that possess large genomes, the optimal pair design can provide both efficient use of microarray resource and good power for the detection of eQTL with additive effect.

3.4.2 Complications due to dominance effects

How does dominance affect the performance of this design? Here, the optimal pair design which optimises for both the additive and dominance effects simultaneously is evaluated; the conclusion is that by including the dominance parameter, the design becomes less optimised for detecting the main (additive) effect. Although over a small genome, the optimal pair design can offer a moderate power advantage for detecting QTL and dominance effects over no optimisation, the performance is affected severely in as much that the power for detecting both the main and the dominance effect degrade to almost the same levels as random pairing with a large genome. The results agree with other studies (Piepho 2005; Bueno Filho *et al.* 2006) that finding a design that is optimal for detecting both additive and dominance effects cannot be achieved. They have shown that optimising for detecting dominance effects would decrease power for detecting additive effects. Therefore, when one has to make a choice between additive and dominance effects for optimisation, the question relates directly to the goal of the experiment. If the goal is to scan across the whole genome for linked loci to gene expression phenotypes, I argue that one could consider focusing on the additive parameter alone for the optimisation. After all, the ultimate interest is to detect QTL. In most cases QTL are expected to have an additive component, even in cases where dominance is present. Optimising for dominance effects should be considered only if there is strong *a priori* evidence for over-dominance in the QTL of interest in a candidate gene study.

3.4.3 Final remarks

It is shown that the extension of the distant pair design, the optimal pair design, can be applied efficiently to outbred line crosses for genetical genomic studies. Having stated that, one has to acknowledge that in experimental design for genetical genomics, there is no “one-size fit all” solution. The most powerful and efficient design will depend on the population structure, marker density, chosen method of analysis, numbers of treatments, and parameter of interest. In human or other natural populations, the Haseman-Elston method (Haseman & Elston 1972) can be applied to sib-pair analysis. In which case, the most effective use of microarray resources to conduct an eQTL linkage analysis would be to profile the expression of a pair of sibs on the same array. It is because the trait squared differences between two sibs are the dependent variable used in this method; these quantities are obtained most accurately when sibs are paired up on the same array.

It is also worth considering the implication of the use of high density Single Nucleotide Polymorphism (SNP) genotyping have on the optimal pair design described in this article. High density SNP genotyping is most widely used in association studies in natural human populations rather than in line crosses of animals discussed above. As linkage disequilibrium spans relatively short distances in human populations, the effective number of independent loci is much higher than what I have modelled in the line cross simulations. This effect is equivalent to increasing the genome size and is likely to have a negative impact on the performance of the optimal pair design than what can be expected in outbred line crosses. Eventually, the distant pairing strategy might become almost equivalent to a pairing strategy based on relationships, in which less related individuals should be

paired for each hybridisation (Rosa *et al.* 2006; Bueno Filho *et al.* 2006). Theoretically, the optimal pair design should always be preferred; since the variance of the estimate of the parameter is minimised, its performance should be at least as good as the common reference design. However, other factors, such as technical simplicity and flexibility in the choice of statistical methods, might shift the balance in favouring the common reference design when the performance advantage in using the optimal pair design becomes less marked. Therefore, it is imperative to consider each experiment and the question of interest on a case-by-case basis. Nevertheless, the results suggest that the efficient design principles outlined by Fu and Jansen (2006) can be applied to a wider context than RILs. With larger eQTL experiments becoming more affordable, one can expect to discover more loci with moderate to small effects. Such attainment will ultimately lead to greater advances to our understanding of the molecular basis of complex traits.

3.5 Conclusion

To better our understanding of the genetic architecture of gene expression phenotypes, there is much needed urgency in performing large experiments and avoiding low-powered studies with small sample sizes. Experimental design should be an integral part of the whole study; with the research goal well defined and an appropriate design formulated, there is hope to achieve the maximum efficiency and effectiveness from the resources available. Until commercial one-colour microarray platforms become commonly available at low costs for livestock species, two-colour microarray platforms will remain the only realistic options for animal scientists inspired by the genetical genomics revolution. The microarray experimental design

strategy presented above encourages and enables them to use the maximum sample size they can afford.

CHAPTER 4

Genome-wide association of gene expression

Chapter 2 outlined a candidate gene eQTL mapping experiment where the main research focus was to identify genes for which the expression levels were associated with a single locus. The real power of genetical genomics, however, lies in the ability to elucidate the genetic basis of global gene expression across a large number of loci in the whole genome. In genome-wide studies of gene expression, it is necessary to consider the challenges related to dealing with extremely large datasets, because the scanning for association to thousands of gene expression traits over thousands of genetic markers results in millions of tests. In this chapter, I will attempt to provide an insight into addressing some of the important issues: (a) using the appropriate method for the data available; (b) conducting the analysis in a computational efficient manner; and (c) accounting for multiple testing to control the level of false positive adequately. To illustrate these practical issues inherent to genetical genomics, an analysis was performed on a real eQTL dataset: the expression of human lymphoblastoid cell lines that has been previously published (Morley *et al.* 2004).

4.1 Introduction

4.1.1 Family-based association

Association mapping has recently become a commonly used method in detecting quantitative trait loci (QTL). This approach is also known as linkage disequilibrium (LD) mapping, which has the simple assumption that a proportion of

the variation in phenotypic values among subjects can be traced to a single allele at one locus (Lynch & Walsh 1998). Since the causal mutation first occurred on an ancestral chromosome, through generations this haplotype gradually reduced in size due to recombinations. Using a dense genetic marker map, it may be possible to capture the population-level disequilibrium to the causal mutation in present-day chromosomes. Risch & Merikangas (1996) argued that association mapping should be more powerful than linkage studies. This is true especially for detecting high-frequency polymorphisms, where the inheritance pattern is sometimes impossible to resolve for linkage analysis because a common allele can often enter a family through multiple founders (Kruglyak 2008). Figure 4.1 illustrates the general idea of population-based association mapping.

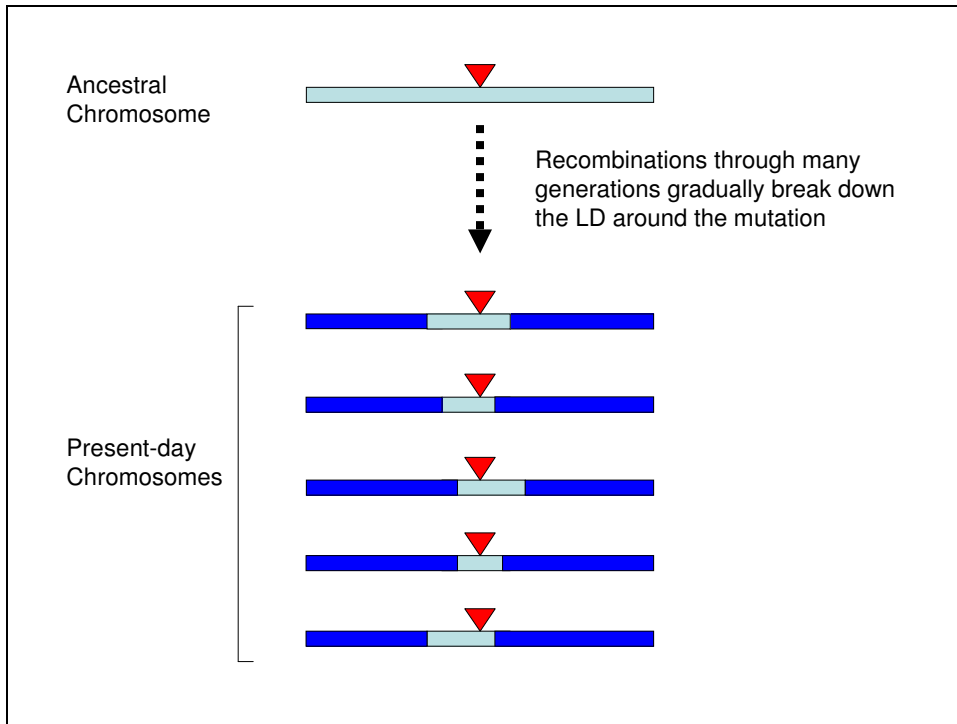


Figure 4. 1

The causal mutation (red triangle) of a quantitative trait arose on the ancestral chromosome, with the chromosomal stretches that are derived from the ancestral chromosome carrying the mutation shown in light blue. The dark blue chromosomal stretches were DNA introduced to the mutation-carrying chromosomes by recombinations. In order to detect the causal mutation, it is necessary to use a dense genetic marker map so that the small stretch of chromosome in LD with the mutation is covered by genetic markers. This figure is adapted from Kruglyak (2008).

In general, population-based designs in QTL mapping are easier to analyse because standard statistical tests such as linear regression and analysis of variance (ANOVA) can be directly applied to test the relationship between trait and marker genotype (Balding 2006). Nonetheless, collection of unrelated subjects might not always be simple. For example, in livestock populations bred by breeding companies, subjects are usually related and data are collected across generations. In other cases, it would be in fact desirable to recruit families for a good experimental design, such as in studies of childhood diseases (Laird & Lange 2006). A family-based design has another attractiveness: most family-based approaches are robust against population admixture (Fulker *et al.* 1999; Laird & Lange 2006), as oppose to population-based studies, which can be strongly affected by hidden substructure within the samples (Marchini *et al.* 2004).

For family-based QTL association analysis, a range of methods and software that utilise information about transmission of alleles, such as the orthogonal test for within-family variation (quantitative trait transmission disequilibrium test, QTDT) (Abecasis *et al.* 2000) and the family-based association test (FBAT) (Laird & Lange 2006) have been developed. By utilising the within-family component of the association alone, these methods are robust in the presence of population stratification. At the same time, study populations are under minimal risk of stratification when the subjects have been carefully selected to remove any genetic “outliers” from the rest. For those populations, the measured genotype approach (Boerwinkle *et al.* 1986) exploits both the variation between- and within-family, and may serve as a powerful tool for QTL analysis. In this approach, a genetic

polymorphism under study is included as a fixed effect or covariate in a mixed model that includes a polygenic component as a random effect.

Although the measure genotype approach can be viewed as the most powerful family-based association method in the absence of population stratification, it is time-consuming and is therefore impractical for genome-wide analysis, particularly for analysing multiple quantitative traits such as global gene expression, due to the need to solve a large number of relatively complex mixed model equations. A fast and simple implementation of the measured genotype approach has recently been proposed (genome-wide rapid association using mixed model and regression, GRAMMAR) (Aulchenko *et al.* 2007a). GRAMMAR first obtains residuals adjusted for family effects and other covariates using a mixed model without fitting the marker fixed effect. Subsequently, the association between the residuals and genetic polymorphisms can be analysed using rapid least-squares regression. Aulchenko *et al.* (2007a) showed that GRAMMAR can yield very similar results to the measured genotype method and the power of GRAMMAR compares favourably to QTDT and FBAT. Moreover, the speed advantage means that GRAMMAR could be a very attractive tool for analysing genetical genomics data. In this chapter, I applied GRAMMAR to re-analyse an eQTL dataset in human lymphoblastoid cell lines (Morley *et al.* 2004) to demonstrate its suitability for genetical genomics when data are collected from families.

4.1.2 Controlling false discovery

The wealth of data generated by large scale genomic studies presents great opportunities to advance our understanding in the mechanisms and interactions of biological molecules which lead to the manifestation of a biological system. At the

same time, it also presents a unique set of challenges related to identification of true positives in the context of a large number of statistical comparisons. With individual tests, the likelihood of rejecting a truly null hypothesis is controlled by the type I error rate (e.g. $\alpha = 0.05$). However, with multiple independent tests, the likelihood of erroneously rejecting at least one null hypothesis increases dramatically. For example, when 100 independent tests are performed at $\alpha = 0.05$, there is a >99% chance of rejecting at least one null hypothesis when they are truly null. In genetical genomics, the number of tests is typically over a million. Therefore, it is essential to deploy appropriate strategies to limit the number of false positives to an acceptable level.

The Bonferroni correction is a simple procedure to adjust the significant threshold according the number of tests conducted. If K tests are performed, each of the K p -values is multiplied by K to obtain the Bonferroni adjusted p -values (Simon *et al.* 2003). One can be 95% confident that all of the positives identified are true when the significant threshold is set to be the Bonferroni adjusted p -value of 0.05. There are two main problems with this method: (1) the Bonferroni adjusted p -value of 0.05, when K is large, corresponds to an extremely small unadjusted p -value. The p -values from parametric tests are usually inaccurate in this low range unless the data follow perfectly the normal distribution or the sample size is very large (Simon *et al.* 2003). (2) The Bonferroni correction is too conservative for many analyses in genetics and genomics because the tests are rarely independent; for example, some correlation often exist between the genotypes of neighbouring markers, or between the expression levels of two or more genes.

Permutation approach is an alternative multiple testing correction method which is robust to departure from parametric assumptions (Churchill & Doerge 1994). To apply in genetic analysis, the idea of this approach is to generate an empirical null distribution by shuffling up the marker genotypes for individual units while retaining their phenotypic values unchanged, and this is repeated over a large number of iterations. Hence, the result of any hypothesis tests from the permuted data is purely due to chance. Ranking the unadjusted p -value alongside the p -values generated from the randomised datasets provides the empirical p -values for the given dataset. Permutation approach is a statistically sound method for estimating the threshold values. However, it can be computationally intensive to run a large number of iterations, depending on the population structure. Furthermore, careful consideration of the design factors is very important, especially in cases where individual units are not exchangeable in a simple manner due to relatedness between subjects (Churchill & Doerge 2008).

False Discovery Rate (FDR) (Storey & Tibshirani 2003) has been proposed as a method which offers a sensible balance in genome-wide studies in keeping the number of false positive low while not being as overly stringent as the Bonferroni approach. Unlike the false positive rate which quantifies the rate that a truly null result is called significant (indicated by p -value), FDR measures the rate at which significant features are truly null. As proposed by Storey & Tibshirani, the significance of each feature in terms of the FDR can be quantified by their q -value. The q -values directly provide a meaningful measure of confidence among the test results called significant. Technically, the q -value for a feature is defined as the minimum FDR that can be attained when calling that feature significant, when all p -

value thresholds have been considered. For a given threshold t , where $0 < t \leq 1$, the FDR is estimated as follows:

$$\widehat{FDR}(t) = \frac{\hat{\pi}_0 m t}{\#\{p_i \leq t\}},$$
 where m is the total number of features, $\#\{p_i \leq t\}$ is

the number of p -values below the threshold t , and $\hat{\pi}_0$ is the estimated proportion of features that are truly null. The quantity of $\hat{\pi}_0$ is estimated from the distribution of the p -values. q -value provides a way to monitor the proportion of false positives amongst all positives. It has been shown that the methodology is less conservative than other false positive control strategies for genome-wide studies and does not lead to substantial loss of power (Storey & Tibshirani 2003).

Apart from applying multiple-testing correction to the results, data quality control is also a crucial step towards reducing the number of false positives. In this chapter, I will demonstrate the importance in careful data filtering, and how q -value can be applied to control type I error in genetical genomics.

4.2 Methods

4.2.1 Data description and pre-processing

The dataset originated from the study by Morley *et al.* (Morley *et al.* 2004). RNA was extracted from lymphoblastoid cells from each individual of 14 CEPH Utah families (3 generations, ~8 offspring per sibship, ~14 subjects per family). The expression levels of ~8,500 transcripts were obtained using the Affymetrix Human Focus arrays. Genotypes of 2,882 SNPs of all subjects were obtained from The SNP Consortium (http://snp.cshl.org/linkage_maps/).

All microarray CEL files were pre-processed by “GCRMA” from the Bioconductor project (www.bioconductor.org) version 1.8.0. From the 2882 SNPs provided, 2,695 were selected as these were polymorphic amongst the individuals genotyped.

4.2.2 Filtering on variability of the probesets

Genes that are not expressed are not relevant to this study. Signal levels for non-expressed genes are typically above zero due to the background signals and other intrinsic systematic noises. Nonetheless, such genes can be detected on the basis that the background variation tends to be much less than real biological variation across samples. The interquartile range (IQR) was adopted as a measure of variability and used IQR of 0.1 as the threshold for this dataset.

4.2.3 GRAMMAR procedures

The full mixed model for detecting marker association can be written as:

$$y = Wa + Xb + Zu + e \quad (1)$$

In expression (1), y is the expression trait values, a , b , u and e are vectors of marker effect, other fixed effects (sex and generation), additive polygenic effect (random) and random residuals respectively. W , X , and Z are incidence matrices related to marker, fixed and polygenic effects respectively.

The fast and robust method proposed by Aulchenko *et al.* (2007a) is composed of 2 steps; the first step accounts for the familial dependence among family members and covariates of nuisance effects, and the second step tests the single SNP effect on the remaining variation by analysis of variance (ANOVA).

Step 1: For the expression values of each probeset I fitted the following mixed model without the marker effect:

$$y = Xb + Zu + e \quad (2)$$

The models were fitted using ASReml (<http://www.vsni.co.uk/products/asreml/>) version 1.0. Narrow-sense heritability (h^2) was estimated for each expression trait using the -P option in ASReml.

Step 2: Using the residuals from step 1 as the new quantitative traits, the marker genotype effect of each SNP on each trait was tested by ANOVA. I used the `lm()` and `anova()` functions in R (www.r-project.org) version 2.3.1. FDR was calculated using the approach proposed by Storey and Tibshirani as implemented in the R package “QVALUE”(Storey & Tibshirani 2003).

4.2.4 Detection of *cis*-eQTL

eQTLs which associate with transcripts within 1 Mb of themselves are considered as *cis*-acting. Besides conducting the analysis at genome-wide level, I isolated a subset of 8462 probable *cis*-acting candidates (expression trait – SNP pairs), which comprised 2066 SNPs and 2797 expression traits, for mapping *cis*-acting eQTL separately. This was a much smaller search space and FDR was applied separately to obtain a new, group-wise significance threshold.

4.2.5 Comparison of GRAMMAR to the full mixed model

10,000 expression trait - SNP combinations were sampled for comparing the performance of the two-step GRAMMAR approach and the full mixed model. Tests using the full mixed model described above were conducted using ASReml.

4.3 Results and Discussion

4.3.1 Equivalence of GRAMMAR and the full mixed model method

The GRAMMAR method produced very similar P values of the marker effect to the full mixed model (Figure 4.2). For the present dataset, I estimated a 6 fold increase in speed with the GRAMMAR approach compared to the full mixed model approach using ASReml with the computing resources available.

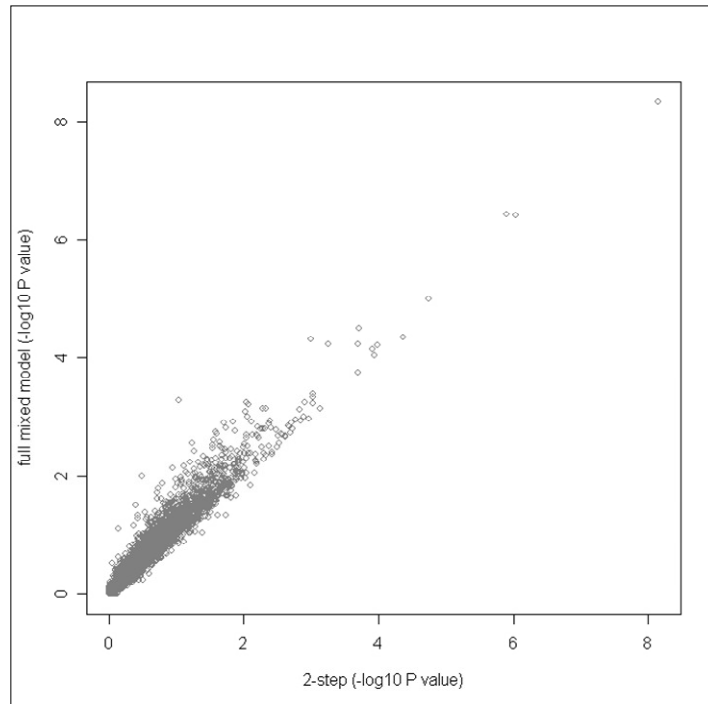


Figure 4. 2

Comparison of the GRAMMAR (2-step) method to the Measured Genotype (full mixed model) approach. The transformed p -values of the genotype effect of 10,000 randomly selected tests from the two methods are shown in this plot. It can be seen from this plots that the two methods produce very similar results.

In fact, because step 1 of the GRAMMAR approach removes the polygenic component in the pedigree dataset, there is now the flexibility to evaluate multiple genetic models in an efficient and simple manner in step 2. This includes the possibility of estimating pairwise epistatic interaction between SNPs, although such analysis would require a more powerful (e.g. larger) study. More recently, a library in R called GenABEL (Aulchenko *et al.* 2007b) has been released which greatly speed up the execution of fitting a large number of linear models. This increases the speed advantage of GRAMMAR over measured genotype and other family-based association method such as QTDT (as shown by Lam *et al.*, accepted) to an even greater extent. In addition, it implies that running permutations to establish genome-wide and experiment-wise threshold with large number of permutation is feasible within a reasonable time period.

4.3.2 Reduction in the number of tests by filtering on expression variability

Figure 4.3 shows that there is a large cluster of expression traits that has very low variability, and figure 4.4 shows a large cluster of expression traits with low log intensity (0 – 4). I used IQR of 0.1 as a cut-off because expression traits below this threshold had low variability as well as low expression level. As a result, the number of probesets was dramatically reduced from 8739 to 4627 (47%).

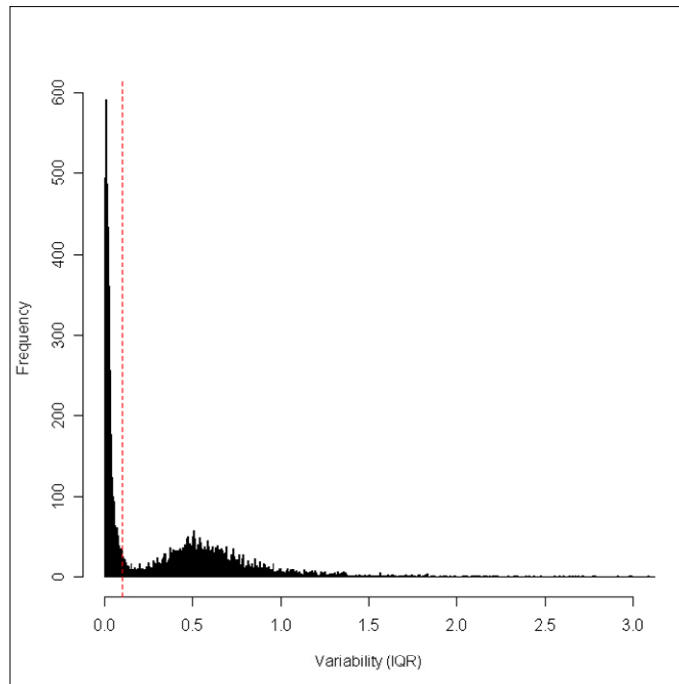


Figure 4. 3
Frequency distribution of the inter-quartile range (IQR) of $\log(\text{intensity})$ of the transcripts. The red dashed line indicates the IQR of 0.1 in $\log(\text{intensity})$.

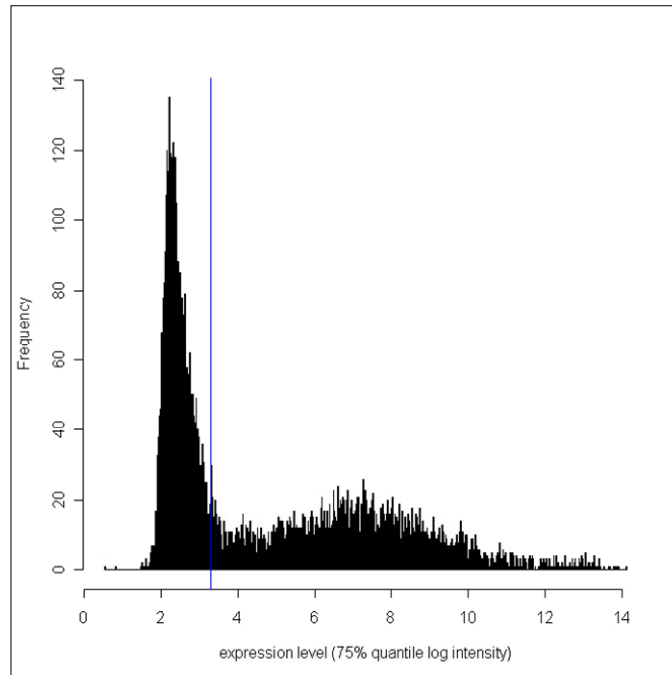


Figure 4. 4

Frequency distribution of the expression level of the transcripts, measured by the 75% quantile of the expression level over all subjects. All transcripts with low variability (below 0.1 IQR) have expression level below by the blue line (the log intensity of 3.3). Therefore, those transcripts with very low variability are also extremely lowly expressed / not expressed at all.

The effect of removing non-expressed genes was roughly mirrored by the heritability distribution. By definition, heritability is a measure of the degree of genetic control of a trait and thus major eQTL detected for traits of low or zero heritability are unlikely to be real. It was reassuring that most expression traits filtered out were of low heritability (Figure 4.5). By removing expression traits that have no biological relevance to the study, this filter substantially reduced multiple-testing and so potentially increased the power to detect real eQTL.

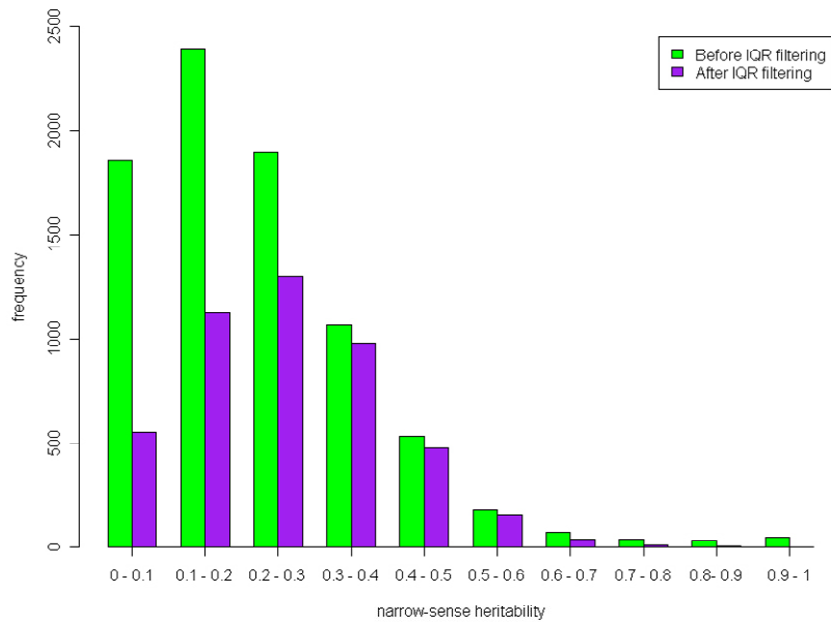


Figure 4.5
Heritability of expression traits. IQR filtering removed mostly the expression traits with low heritability.

4.3.3 Numerous spurious associations in the initial analysis

Using the GRAMMAR method, I detected 2282 associations at 20% FDR (P cut-off = 3.65×10^{-5}). I observed that many significant hits were associated with the same SNPs. Although this phenomenon could be interpreted as some loci being the master regulator for a large number of transcripts, there is evidence that these putative *trans*-acting hotspots are likely to be artefacts. Table 4.1 shows the relationship between the number of significant associations and the sample size in the minor genotype class of a SNP. The SNPs with the most associations (with over 100 transcripts) were those with only 1 or 2 individuals in the minor genotype class. Conversely, I did not find SNPs with higher minor genotype count associated with multiple transcripts to the same extent. As ANOVA compared the phenotypic means of the genotype classes, outliers in the expression traits could have a big effect on the phenotypic mean, especially for SNPs which have genotype classes with a very small number of individuals. Figure 4.6 illustrates an example of such artefacts.

Minor genotype count	No. of SNPs	No. of hits	Max. no. of hits by a single SNP	Avg. no. of hits per SNP
1	103	1054	200	10.23
2	55	333	147	6.05
3 - 6	166	508	48	3.06
7 - 10	56	107	12	1.91
11 - 15	52	85	9	1.63
16 - 20	42	56	4	1.33
21 - 30	45	65	5	1.44
> 30	51	74	6	1.45

Table 4. 1

Relationship between the minor genotype count and number of significant associations without the filtering of SNPs on genotype counts

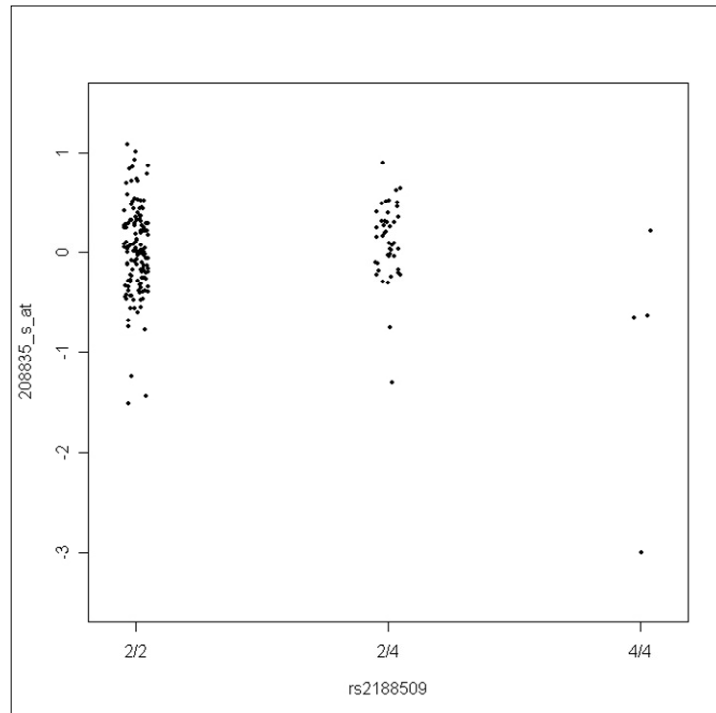


Figure 4. 6
Scatterplot of the expression trait residuals of probeset 208835_s_at after step 1. The x-axis shows the three genotype classes of the SNP rs2188509. The y-axis shows the GRAMMAR-adjusted phenotype of this probeset. Spurious p -value of 2.6×10^{-5} is caused by an outlier in genotype class 4/4.

4.3.4 Reduction of spurious associations by filtering on genotype counts

Subsequently, I employed a screening strategy on the SNP data by excluding any genotype classes with 4 or fewer individuals. 423 SNPs were found to possess at least one such genotype class. When the ANOVA tests were repeated, only 61 associations were detected at 20% FDR (P cut-off = 9.78×10^{-7}). This finding suggests that the vast majority of associations previously detected were due to small sample size in SNP genotype classes, and therefore, unreliable. Note that the *p*-value threshold for the same FDR was much lower after having avoided the detection of many putative artefacts. FDR estimation is strongly influenced by the distribution of the *p*-values. If a large number of spurious effects are present due to violation of the underlying assumptions of the test statistic, excessive detection of false positives will not be prevented by the use of FDR.

This strategy to screen SNPs on the genotype counts is superior to the commonly used filter based on minor allele frequency (typical thresholds used are 3, 5 or 10%). The latter approach is not sensitive to detect SNPs with a small genotype class because rare homozygous genotypes can be observed with minor alleles of moderate frequency under Hardy-Weinberg equilibrium, given the sample size of the current study.

It is also important to note that I only masked out genotype classes with small number of individuals rather than omitting all the data for such SNPs. This has the advantage that the information from the remaining genotype classes could still be used for the tests. For example, having masked out the rare 4/4 (3 individuals)

genotype class from SNP rs1491846, its association with probeset 204133_at was detected at $P = 1.13 \times 10^{-7}$. Hence, this screening method is not only effective in excluding spurious effects, but also preserves genuine effects in the presence of rare genotype classes.

4.3.5 Detection of *cis*-acting loci

Out of the 61 eQTLs detected, 3 eQTLs are within 1Mb of their transcripts (*cis*-acting eQTL). Detecting so few *cis*-acting eQTLs is perhaps not a surprise because the SNP density in this dataset is very low for whole-genome association mapping in humans where it was estimated that ~500,000 SNPs would be required (Kruglyak 2008). Much of the genome would not be in strong linkage disequilibrium with the SNPs used in the genome scan. Effectively, only a small proportion of the genome has been screened. On the other hand, the tests for *cis*-acting eQTL are a tiny proportion of the total number of tests performed genome-wide. Therefore, they are heavily penalised by multiple-testing in the analysis above. Subsequently, I restricted the testing to only the SNPs and transcripts that were less than 1Mb away from each other. This gave rise to 8462 *cis*-acting “candidates” (0.07% of all tests). At 20% FDR (P cut-off = 3.54×10^{-4}), this analysis led to detection of an additional 12 *cis*-acting eQTLs (15 in total). Without laboratory-based validation, it is difficult to conclude whether partitioning the data in this way can increase power of detecting real *cis*-acting eQTL. Nonetheless, this strategy can be considered as a practical way for improving the chance of detecting real *cis*- effects. Because of the technical, statistical limitations and uncertainties in studying *trans*-regulation as described by Pastinen *et al.* (2006), one may wish to dedicate more resources to studying *cis*-acting eQTL over *trans*-acting eQTL. This strategy increases the detection of *cis*-

signals and provides more “prioritised” candidate loci. In the present study, the number of candidates generated is still practically feasible to be followed up in laboratories.

4.4 Conclusion

The two-step approach presented here (GRAMMAR) is simple, fast and efficient for family-based association studies in a mixed model framework. The speed advantage makes this implementation an attractive method for analysing genome-wide association with large number of quantitative phenotypes. Filtering on variability of the probesets dramatically reduces the number of irrelevant expression traits and multiple-testing. The method used here for masking rare genotype classes substantially decreases the number of spurious detection due to phenotypic outliers. Finally, limiting the search to SNPs and transcripts that are in close proximity appears to be a practical approach to avoid the excessive penalty imposed by multiple-testing on *cis*-acting eQTL and to increase the chance of detecting real signals for *cis*-regulation.

CHAPTER 5

A gene set approach for eQTL mapping

This chapter presents an alternative mapping strategy to the eQTL mapping methods used in the previous chapters and by many published studies. The approach is motivated by the idea of gene set testing that has been widely used in microarray analysis for differential expression detection. By assigning genes into groups with common biological functions, based on knowledge derived from bioinformatics resources, the evidence of linkage for a group as a whole can be assessed. Testing gene sets may provide the advantage of increased sensitivity to eQTL of small effects which are sometimes difficult to detect when genes are tested one at a time, because of the strong multiple testing correction imposed on the univariate statistics. Section 5.1 hypothesizes how gene set testing might be useful for finding linkage to pathways and provides a review on the existing methods in gene set testing. Section 5.2 describes the BXH/HXB rat eQTL dataset (Hubner *et al.* 2005). This published dataset was used here to investigate the feasibility of applying the gene set approach to map eQTL. Section 5.3 provides technical details on the methods used to define gene sets and the statistical tests considered. Finally, section 5.4 presents the results and a discussion of the usefulness and pitfalls of the proposed approach in mapping eQTL.

5.1 Introduction

5.1.1 *What is gene set testing*

Gene set testing is a statistical framework for analysing gene expression data using predefined categories. It was originally developed to aid functional interpretation of differential expression (DE) analysis using microarrays (Beissbarth & Speed 2004). The aim is to identify any “unusual phenomena” relating to particular categories or sets of genes. For assessing the significance of genes that have been grouped into functional categories in a DE analysis, in the classical statistical sense, one null hypothesis can be defined as: “the extent of DE is the same across all gene sets”. That is, the selection of DE genes is not expected to introduce a bias for or against any functional categories. When this null hypothesis is rejected, one might postulate that it signifies important functional roles for the genes in the significant gene set which are related to the difference between the treatment groups. In general, there are two scenarios in DE analysis where gene set testing can be particularly beneficial.

The first scenario is when there is a long list of DE genes. With limited resources, it is often the case that only the most significant genes would receive adequate attention in further in-depth investigation. Testing of gene sets could highlight some of the functional groups with interesting biological functions on the list, even when members of the gene sets are not amongst the top few on the list.

The second scenario is related to the multiple testing issues in microarray analysis. After correction for multiple testing, there might be a very short list of DE genes in experiments with low statistical power. True DE genes with moderate

significance are likely to be missed. Although relaxing the threshold would extend the DE gene list and recover some of the real DE genes, it would also inflate the false positive rate. Testing of gene sets provides a post-hoc analysis to distinguish potential true DE genes with moderate significance and non-DE genes with a similar level of significance. While this approach can be thought of as “data-dredging”, gene set testing as a whole is regarded as a desirable data-driven hypothesis generating tool (Allison *et al.* 2006).

Although gene set testing has almost exclusively been applied to DE analysis, it should also be amenable to other genetics studies. For example, it has been applied recently to improve the ability of genome-wide association studies (GWAS) to detect disease mechanisms by considering groups of variants that belong to the same biological pathway (Wang *et al.* 2007). The authors extended the gene set approach from DE genes to SNPs that match to genes by their physical locations. With this approach, in addition to the top 20 or so SNPs detected by GWAS, groups of markers that are less significant yet potentially interesting due to links to interacting genes are also highlighted as candidates in the post-GWAS analysis. Here, I propose to incorporate gene sets into genetical genomics by grouping the gene expression traits.

For an eQTL that links to a gene, that gene might be involved in a gene network which drives certain cellular process. The most common way of analysing eQTL, however, is on a gene-by-gene basis, and only those eQTL that exceed the very stringent threshold, as a result of multiple testing, are considered as true linkage. Because many quantitative traits are expected to have a complex genetic architecture with pathways involving multiple interacting genes, studying the top ranking eQTL only, mostly *cis*-linkage, is unlikely to present us with sufficient information to

understand the transcriptional regulatory network involved. In a highly connected network many components (genes) may have moderate effect sizes to which the signals are too weak to be detected. Figure 5.1 illustrates in a hypothetical pathway how one may fail to detect genuine linkage signals.

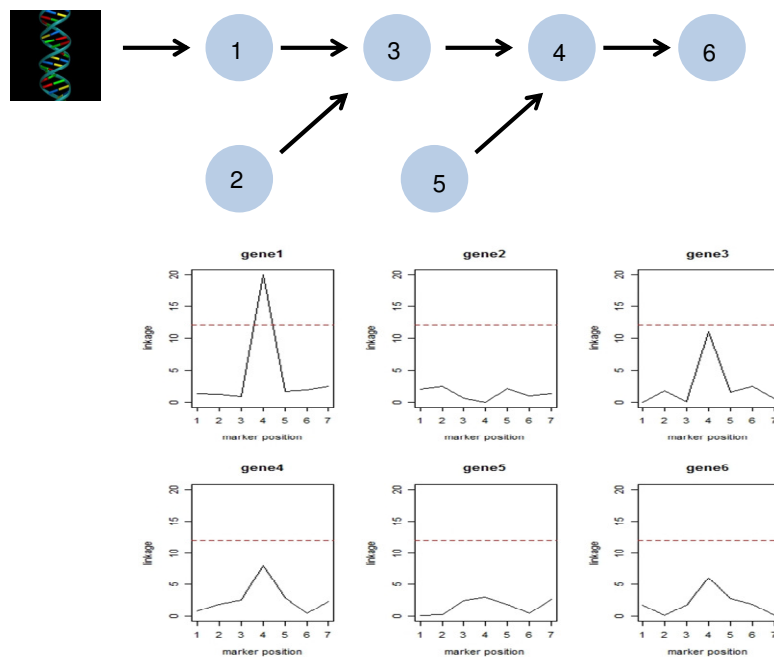


Figure 5. 1

A hypothetical pathway. The blue circles denote genes and the black arrows denote the direction of regulation. A *cis*-eQTL on a chromosome regulates the expression of gene 1. Gene 3 is co-regulated by gene 1 and gene 2. Gene 4 is co-regulated by gene 3 and 5, which together regulates gene 6. The graphs in the lower half of the figure show the linkage profile. The red horizontal lines denote the univariate linkage significance threshold. Linkage is only detected for the expression level of gene 1. *Trans*-acting regulation actually exists for gene 3, 4 and 6, but the linkage signals are too weak due to the influence of other genes and that the regulation due to the eQTL is indirect.

In the hypothetical example in Figure 5.1, the eQTL are indirectly linked to three genes (3, 4 and 6) in the pathway. As the linkage evidence for these *trans*-genes is not very strong, these linkage signals could be rejected as not significant. If we accept that the hypothetical network depicted here is a reasonable model, then a logical next question would be “how can one capture these genuine signals with relatively moderate effect sizes?” Relaxing the significance threshold could prevent some of these genuine signals being rejected, but doing this may risk much noise being wrongly accepted as significant linkages. However, if one looked within gene sets and found many genes with moderate linkage signals, then potentially the pathways represented by the significant gene sets are genuinely linked to the locus. Hence, by incorporate gene set testing with genetical genomics, one can potentially offset the risk in lowering the significant threshold in order to capture pleiotropic eQTL with weak effects.

5.1.2 A review on gene set testing methodologies

There are various implementations of gene set testing, but all have more or less the same underlying principle. First, gene sets are created by grouping all genes that are annotated to the same annotation term according to functional genomics ontological resource like Gene Ontology (GO) (Ashburner *et al.* 2000) or gene pathway resources such as Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa & Goto 2000). Then statistical tests can be used to compare the distribution of the test statistics of genes in a set to a null distribution. Gene set testing in effect shifts the level of analysis of the microarray experiment from single genes to sets of related genes. As previously accumulated biological knowledge is

used to create the gene sets, this approach makes a more biology-driven analysis of microarray data.

Numerous methods (Beissbarth & Speed 2004; Al-Shahrour *et al.* 2004; Boyle *et al.* 2004; Lee *et al.* 2005; Alexa *et al.* 2006; Falcon & Gentleman 2007) make use of a test for independence in a 2 x 2 contingency table with minor variations. This class of methods starts by dividing all the genes into two groups, “significant” and “not significant”, according to the univariate test statistic. Here, the test statistic at the individual gene level is referred to as local statistic. Tests are then carried out to assess whether a gene set is over-represented in the “significant” group. Statistical tests such as Fisher’s Exact Test and Hypergeometric Test (Siegel 1956) are typically used. The test statistic for gene set is referred to as global statistic. Figure 5.2 illustrates the general idea of over-representation of a gene set.

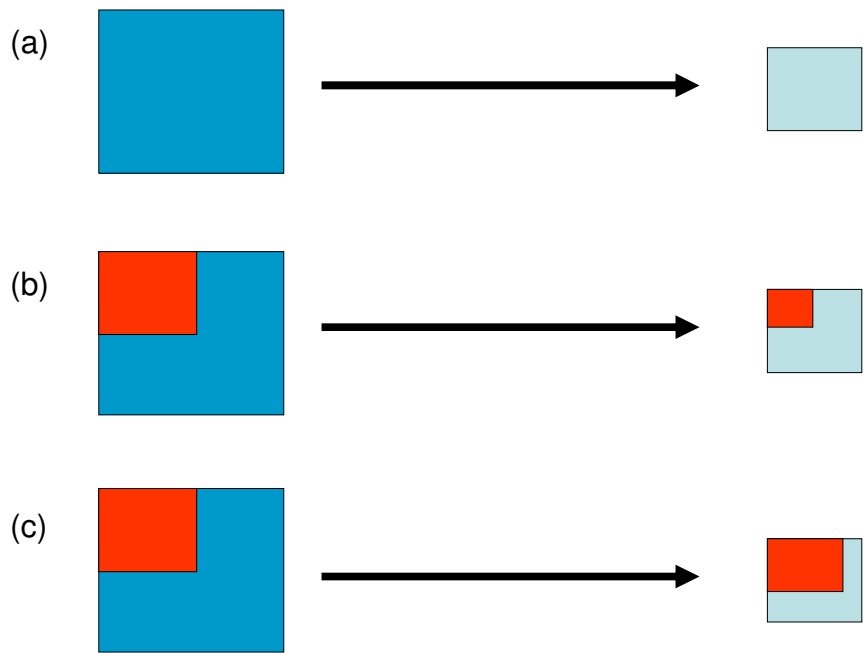


Figure 5. 2

(a) The dark blue box represents all the genes available for testing. Using a cut-off for the local statistic, for instance: $P\text{-value} \leq 0.001$, a subset of the genes are called significant. This subset is represented by the light blue box. (b) Let's suppose a subset of all the genes belong to a pathway / category, represented by the red box. If genes with small P -values are evenly distributed across pathways / categories, the genes represented by the red box should be present in the light blue box with roughly the same proportion as in the dark blue box. (c) If the pathway / category is particularly important, many of its members will have small P -values for their local statistics. Then, the genes from this pathway (red box) will be present in the light blue box in a greater proportion than originally in the dark blue box. There is an "enrichment" of genes from the gene set amongst the significant genes.

The previous approach requires a strict cut-off on the local statistic to divide the genes into the “significant” and “not significant” groups, which ignores the continuity of the available evidence. An alternative class of methods address this shortcoming by taking into account the quantitative nature of the local statistics. Ranks of the genes belonging to a gene set are compared to ranks of the complement set. Rank-based non-parametric tests such as the Wilcoxon rank sum and the Kolmogorov-Smirnov tests (Siegel 1956) are robust methods to compare two distributions, and there are implementations of these non-parametric approaches in DE analysis (Barry *et al.* 2005; Subramanian *et al.* 2005). Using ranks has the potential benefit of being more sensitive in detecting modest but coordinated directional trends by genes in a gene set. This general idea of the approach is illustrated by Figure 5.3.

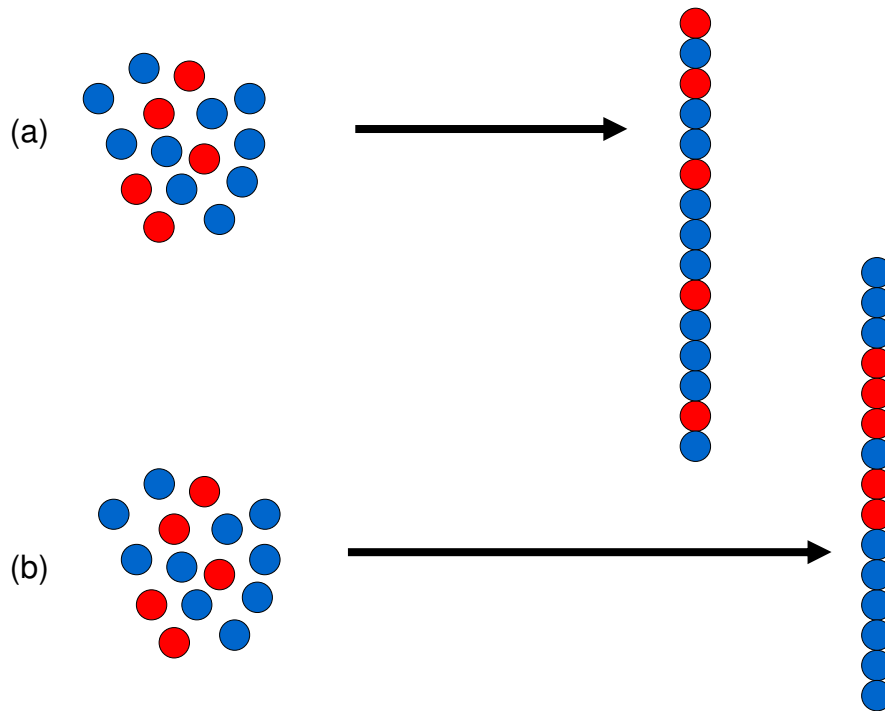


Figure 5. 3

Circles in red represent members of a gene set, and the circles in blue are genes outside the set. The genes are ranked by the local statistic. In (a), the genes in red are from the same distribution as the genes in blue. (b) Even though the local statistics of the red genes do not rank amongst the highest, collectively they rank higher than the blue genes. The genes in red and those in blue probably do not belong to the same distribution.

To assess the significance of the gene set test, the test statistic is compared to the empirical null distribution generated by permutations. One way is to permute the membership of the gene sets. In this case, x number of genes are randomly selected to make up a gene set of size x in each round of permutation. This would generate an empirical distribution where the global statistics are not related to the membership of the gene sets, but to the sizes of the gene sets. This permutation strategy is referred to as “gene sampling” because gene is the unit of sampling under this strategy (Goeman & Buhlmann 2007). An alternative strategy is to permute at the subject level to obtain local statistics. Subsequently these permuted local statistics are used to generate the empirical distribution of the global statistics. In this case, each round of permutation involves using a new set of local statistics generated by the randomisation of subjects. This permutation strategy is referred to as “subject sampling” because the biological subject is the unit of sampling here. As the expression levels of genes within a gene set tend to be more correlated than genes at random, subject sampling retains the structure of gene sets and generates an empirical distribution that better reflects the correlations of genes within a set. Gene sampling, on the other hand, is easier to implement and can be performed much more rapidly. However, it has been argued that the use of gene sampling should be strongly discouraged (Goeman & Buhlmann 2007). The gene sampling approach uses a sample size equal to the number of genes involved in gene set testing instead of the number of biological replicates that is typical in the classical statistical setup. A replication of the experiment under the gene sampling model would therefore involve taking a new sample of genes from the same subjects, which does not make biological sense. Hence it is argued that the P -value produced by gene sampling does

not measure the strength of the evidence based on the biological experiment performed and could be wrongly interpreted as an inflation in power, which is highly misleading. Concerns over the problems of gene sampling permutation strategy were also echoed by Allison *et al.* (2006).

5.2 The BXH/HXB rat dataset

In this chapter, I present a novel use of gene set testing in the context of genetical genomics. The applicability of gene set testing to genetical genomics is demonstrated through a reanalysis of a published eQTL dataset. This section provides a description of the dataset.

The rat has been a model for studying common human diseases for many decades. The spontaneously hypertensive rat (SHR) strain is a widely studied model of human hypertension. Brown Norway (BN) is another strain of rat that has been intensively used for medical research, and is the strain on which the rat reference genome sequence is based. Crossing these two inbred rat strains in a series of sib mating (F_{60}) generates the BXH/HXB panel of recombinant inbred lines (RILs). Animals in RILs have negligible within-line but considerable between-line variation in their genomes which are a fine mosaic of the two founder genomes (Lynch & Walsh 1998). Thirty lines are available in the BXH/HXB panel.

Hubner *et al.* (2005) constructed a linkage map of 1,011 autosomal markers for all chromosomes. Messenger RNA was extracted from fat and kidney tissues from four independent rats from each line; gene expression profiling was performed on each mRNA sample using Affymetrix GeneChip™ Rat230a. Robust multichip average, or RMA, algorithm (Irizarry *et al.* 2003) was used to obtain the summarised

gene expression values. The arithmetic mean was taken over the biological replicates from the same line for all probesets.

Hubner *et al.* (2005) performed linkage analysis on the individual gene expression levels and identified 509 and 761 linkages in fat and kidney, respectively. A large proportion of the most significant eQTL were *cis*-eQTL.

5.3 Methods

A number of steps were carried out to apply gene set testing in the framework of genetical genomics. The general workflow is shown in Figure 5.4.

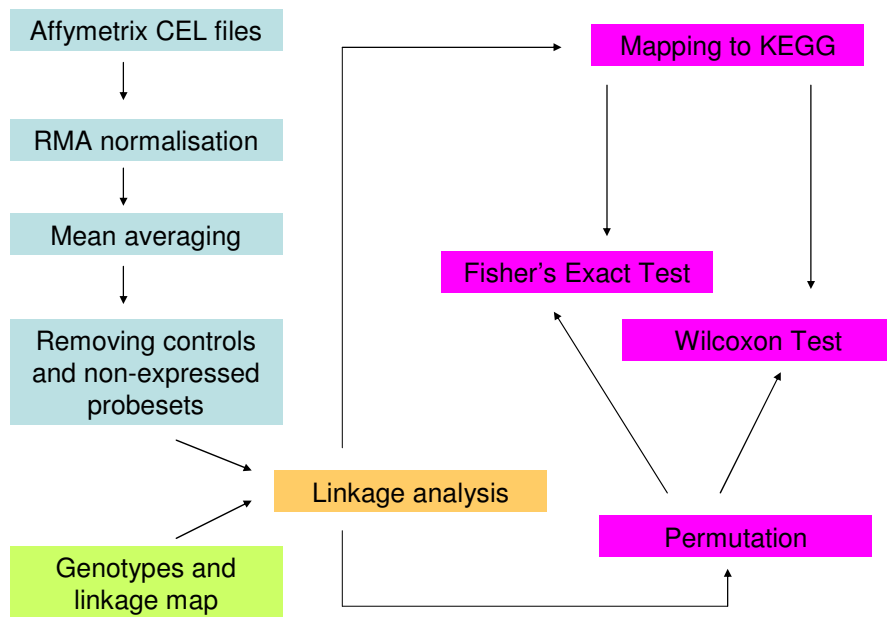


Figure 5. 4

General workflow of the analysis. The steps in light blue are related to microarray data processing.

The steps in purple are related to gene set testing.

5.3.1 Genotype and microarray data

The raw dataset from the Hubner *et al.* study consists of genotype data and gene expression data, courtesy of Prof. Tim Aitman and Dr. Enrico Petretto. Genotypes of 1,011 markers, distinguishing the SHR or the BN line origins at marker loci for the 30 BXH/HXB RILs were available, along with the linkage map. There were 258 Affymetrix CEL files, of which 130 files were gene expression data from fat tissue, and 128 from kidney. For the purpose of studying the methodological aspects of gene set testing in genetical genomics, only the fat tissue dataset is analysed. Gene expression data were processed in the same way as in the original research article (Hubner *et al.* 2005), where the CEL files of the two tissues were processed separately using RMA, and the average gene expression values were taken from the biological replicates. At this stage, there were 15,923 expression traits for the 30 RILs and the 2 progenitor strains.

5.3.2 Filtering based on expression and variability

As discussed in Chapter 4, it is generally a good practice to remove genes that are not expressed in the tissue of interest from downstream analyses. The variability, in terms of standard deviation, and the strength of expression signal, in terms of maximum intensity across samples, was inspected in the two tissues. After visual inspection of the expression data distribution, the variability at the first quartile and the expression signal at the first quartile were set as thresholds. Probesets with variability and expression below both cut-offs were regarded as non-expressed probesets. The two thresholds were 5.88 in log intensity for the maximum variability and 0.10 in log intensity for standard deviation. Together with the control probesets,

non-expressed probesets were discarded, leaving 13,309 probesets for downstream analyses.

5.3.3 Linkage analysis

Linkage analysis was carried out for each of the remaining expression phenotypes using Haley-Knott regression (Haley & Knott 1992). The progenitor strains were excluded. R/QTL (Broman *et al.* 2003) was used to generate the line origin probabilities along the genome in 1 cM intervals. Likelihood Ratio Test (LRT) was carried out to evaluate the model with a single additive QTL versus the model with no QTL along the grid. The LRT statistics were retained as local statistics for gene test testing.

5.3.4 Filtering based on KEGG

For the study described in this chapter, gene sets were defined according to the pathway grouping in the KEGG database. Thus, probesets which did not represent genes present in the KEGG database were redundant. Mapping of probesets to the EntrezGene database (<http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene>) and subsequently to the KEGG database was retrieved using the annotation package “rae230a.db” in Bioconductor (<http://www.bioconductor.org/>). Where multiple probesets mapped to the same EntrezGene entry, the probeset with the highest LRT statistic at the given locus would be chosen for gene set testing. Starting with 13,309 probesets, 11,356 mapped to an EntrezGene entry. After the duplicates were removed the number of non-redundant EntrezGene entries was 9,296. Amongst the genes annotated in KEGG, there were 2,486 entries in EntrezGene, from 187 pathways, that mapped to probesets on the rae230a GeneChip™. After merging the

two sets of genes by the EntrezGene IDs, gene sets with fewer than 5 members present on the microarray were removed. At the end of these series of filtering steps, 2,185 genes from 152 KEGG pathways entered the gene set analyses. This represented 23.5% (2,185 out of 9,296) of the known genes that were probed and expressed in the tissue samples.

5.3.5 Gene set testing

Global test statistics for gene sets were obtained using a one-tailed Fisher's Exact Test and a one-tailed Wilcoxon Test. Both tests were applied along the genome in 1 cM intervals (the same spacing as in the linkage analysis). At every position, the LRT statistics were treated at the local statistics. For each gene set, the local statistics were classified into two groups: members of the gene set and non-members of the gene set.

For the Fisher's Exact Test, the genes were divided into the "significant" and "not significant" groups based on the local statistics. The point-wise P -value of 0.001 was used as the cut-off, which was equivalent to LRT statistic of 10.8 with 1 degree of freedom in the χ^2 distribution. This arbitrary cut-off was chosen because it was reasonably liberal to include eQTL with small effect size, and yet quite stringent so it should not count too many true negatives as positives. For each gene set associated with a KEGG pathway and its complementary non-associated gene set, the number of members in the "significant" and the "not significant" sets were calculated. For each cM along the genome, the global test statistics for each of the pathways were obtained using the "fisher.test" function in R.

For the Wilcoxon Test, the ranking of the local statistics of each KEGG pathway gene set was compared to its complementary non-associated gene set. For

each cM along the genome, the global test statistics for each of the pathways were obtained using the “wilcox.exact” function of the “exactRankTests” package in R.

The significance of the global test statistics were derived using 1000 permutations. “Subject-sampling” strategies were performed, where the linkage analysis was repeated with the RILs shuffled. The local statistics from the 1000 genome scans were stored. Gene set tests were performed on these 1000 sets of local statistics to derive the null distribution of the global statistics. Large data storage and parallel grid computing were provided by the university high performance computing services (<http://www.is.ed.ac.uk/ecdf/>).

5.4 Results and discussion

5.4.1 Fisher’s Exact Test

Gene set testing was carried out along the genome. Using the genome-wise threshold of 0.05, Fisher’s Exact Test identified 8 gene sets showing over-representation in the group of genes with LRT statistics greater than 10.8 in 8 regions in the fat tissue (Table 5.1).

KEGG ID	Chr	cM	Minimum genome-wise <i>P</i> -value	Gene set size	No. of genes with significant local statistic *	KEGG pathway name
05020	1	143	0.013	12	3 (23)	Parkinson's disease
04360	3	210-211	0.015	70	5 (8)	Axon guidance
00630	5	71-72	0.030	7	3 (8)	Glyoxylate and dicarboxylate metabolism
03022	12	18	0.043	15	2 (3)	Basal transcription factors
00260	19	35-36	0.041	24	3 (10)	Glycine, serine and threonine metabolism
04514	20	1-6	0.007	81	9 (14)	Cell adhesion molecules (CAMs)
04612	20	1-5	0.003	46	10 (14)	Antigen processing and presentation
04940	20	2-5	0.003	34	9 (14)	Type I diabetes mellitus

Table 5. 1

Regions in the genome with significant enrichment signal from Fisher's Exact Tests (genome-wise threshold $P \leq 0.05$). * The column shows the number of genes in the gene set with LRT statistic above 10.8. In bracket is the total number of genes with LRT statistic above 10.8 at the position with the maximum global statistic.

Three KEGG pathways (04514, 04612, and 04940) mapped to the same region on chromosome 20. These three signals were also amongst the most highly significant amongst the eight peaks identified. There was considerable overlap in the gene membership for these gene sets: 17 genes are common to all three pathways; and 5 genes are common to two pathways. Almost all of the genes with the LRT statistic above 10.8 were genes common to all three pathways. Indeed, the three pathways have similarities in their functions, being related to the immune system.

The LRT statistics for some of the significant genes from these three pathways were well above 10.8. For example, the LRT statistics for the affymetrix probesets “1369110_x_at”, “1377334_at” and “1371213_at” are all greater than 37.0 at the 5 cM position on chromosome 20 (point-wise $P < 1.2 \times 10^{-9}$). These probesets are probes for some of the *RT-1* genes, also located on the proximal arm of rat chromosome 20, in close proximity of the eQTL. However, the *RT-1* class genes are orthologous to the Major Histocompatibility Complex (MHC) class genes in human and mouse. Strain specific sequence variants have been known to be a major source of *cis*-acting eQTL artefacts (Alberts *et al.* 2005). Since the probes were designed from sequences of BN strain, and the MHC class genes are highly polymorphic, many of the probes are likely to hybridise preferentially to BN transcripts. Therefore, cautious interpretation of these enrichment signals on chromosome 20 would be essential as they are likely to be false positives due to sequence variation on the short oligonucleotide probes.

For the other five signals, the number of genes with local statistic above the threshold of 10.8 ranges from 2 to 5. Those signals with only 3 or fewer significant genes are not treated as real signal of gene set enrichment because the *P*-value of

these signals can be dramatically affected by even the smallest changes in gene set assignment. Clearly, judging the results solely by the P -value of the global statistic can lead to wrong interpretations. The signal for pathway 04360 contains 5 significant genes. The significant genes on this KEGG pathway are Fyn, Rock2, Cxcl12, Cdc42, and Nrp1. These genes are highlighted in the pathway diagram in red in Figure 5.5.

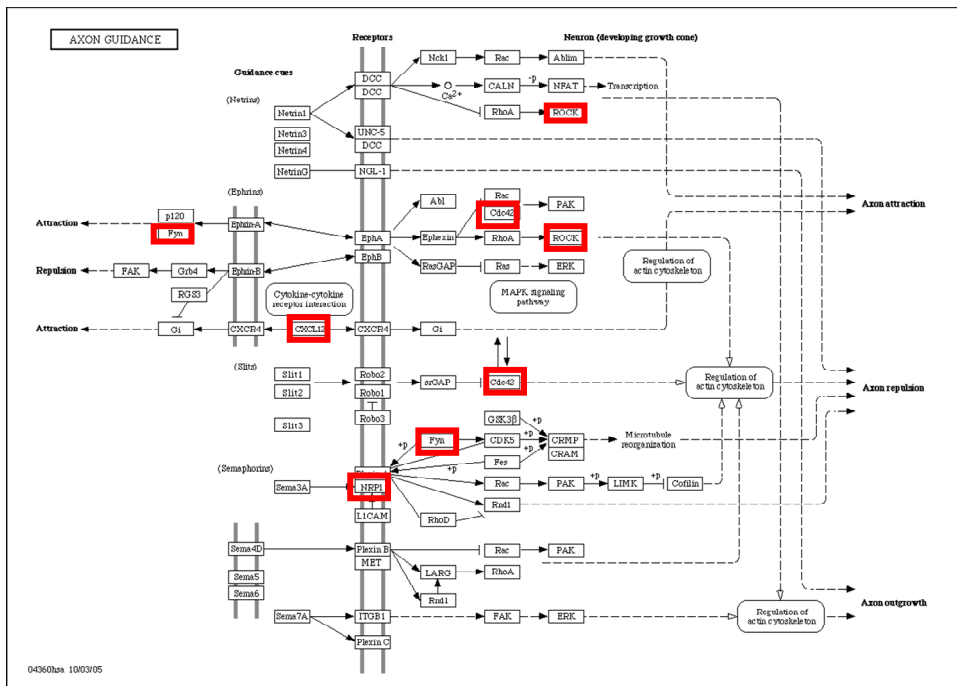


Figure 5. 5

The KEGG pathway for Axon Guidance. The boxes highlighted in red represent the significant genes at the locus of the Fisher's Exact Test signal. The genes Fyn, Cdc42 and Rock2 feature twice on different branches in this pathway diagram.

Those genes with significant linkage are located on several branches on the pathway. Therefore, there is some departure from the model illustrated in Figure 5.1. But interestingly, two of the branches contain two genes (Fyn phosphorylates a complex containing Nrp1, and Rock2 and Cdc42 both interact, directly and indirectly, with Ephexin) which may indicate that the linkage to the expression levels of these genes could be a “knock-on” effect by the linkage to their neighbours on the pathway. However, with the exception of Nrp1, these genes are also involved in many other signalling pathways outside the context of axon guidance. Hence, based on the current information, it is not straight-forward to make any strong inference on the reliability of this pathway signal.

5.4.2 Wilcoxon Test

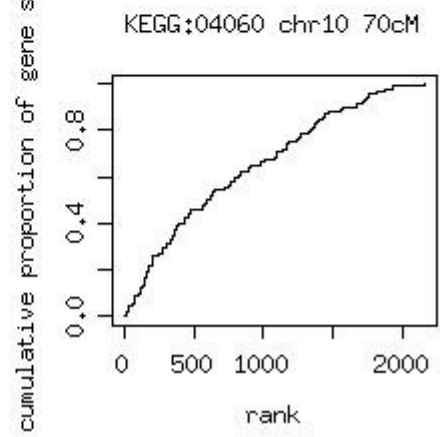
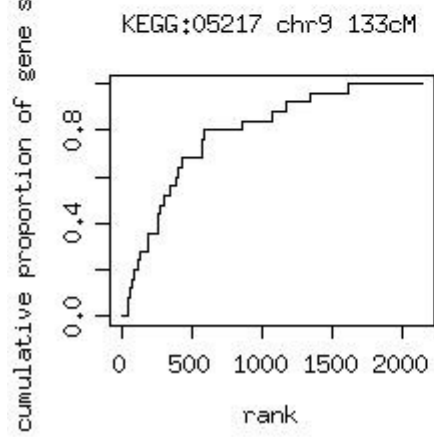
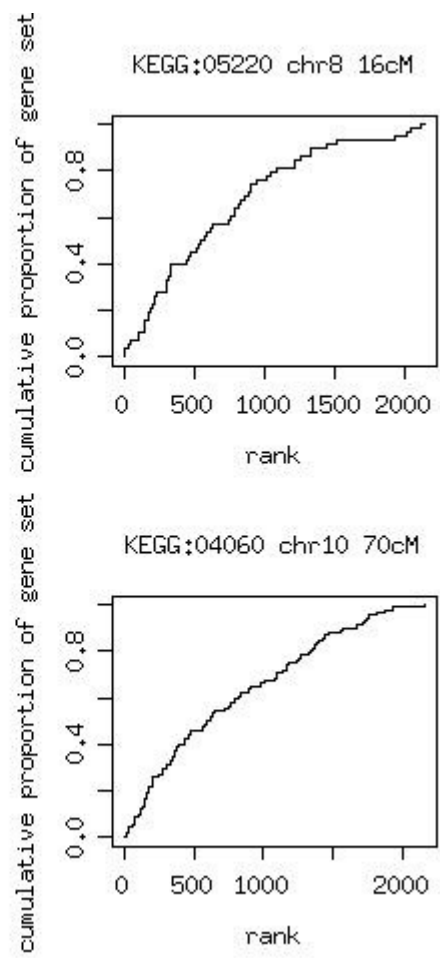
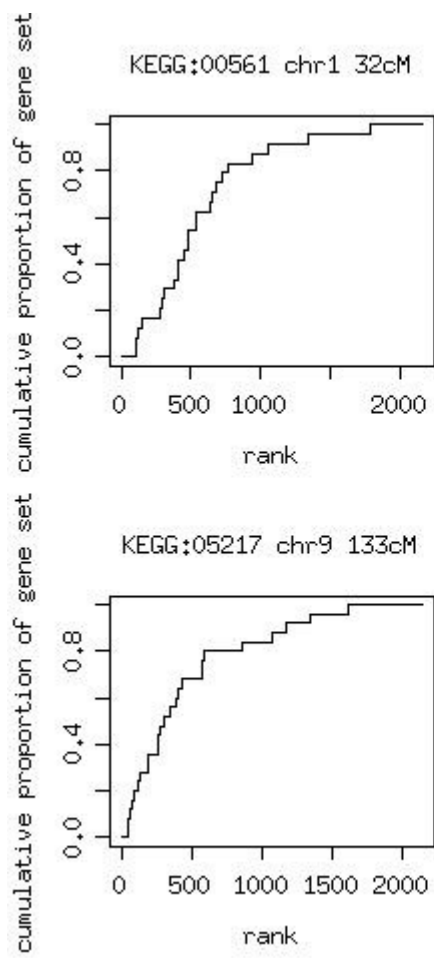
The Wilcoxon Test was also carried out along the genome to provide an alternative flavour to gene set testing for genetical genomics. Using the genome-wise threshold of $P = 0.05$, 9 gene sets over 10 regions with local statistics that rank significantly higher than the rest were identified in fat tissue (table 5.2).

KEGG ID	Chr	cM	Minimum genome-wise <i>P</i> -value	Gene set size	KEGG Pathway name
00561	1	32-33	0.043	24	Glycerolipid metabolism
05220	8	15-18	0.027	58	Chronic myeloid leukemia
05217	9	133	0.002	25	Basal cell carcinoma
04060	10	64-73	0.008	92	Cytokine-cytokine receptor interaction
00272	10	144-155	0.002	10	Cysteine metabolism
04010	11	1-2	0.002	168	MAPK signalling pathway
04010	11	14-16	0.018	168	MAPK signalling pathway
00510	12	21-30	0.005	26	N-Glycan biosynthesis
00030	14	55	0.026	14	Pentose phosphate pathway
04140	20	73-74	0.047	14	Regulation of autophagy

Table 5. 2

Regions in the genome with significant enrichment signal from Wilcoxon Tests (genome-wise threshold $P \leq 0.05$).

None of the signals picked up by the Fisher's Exact Test were reproduced by the Wilcoxon Test. The extent of the upward shift detected by the Wilcoxon Test can be examined by plotting the ranks of the local statistics from the gene sets (Figure 5.6).



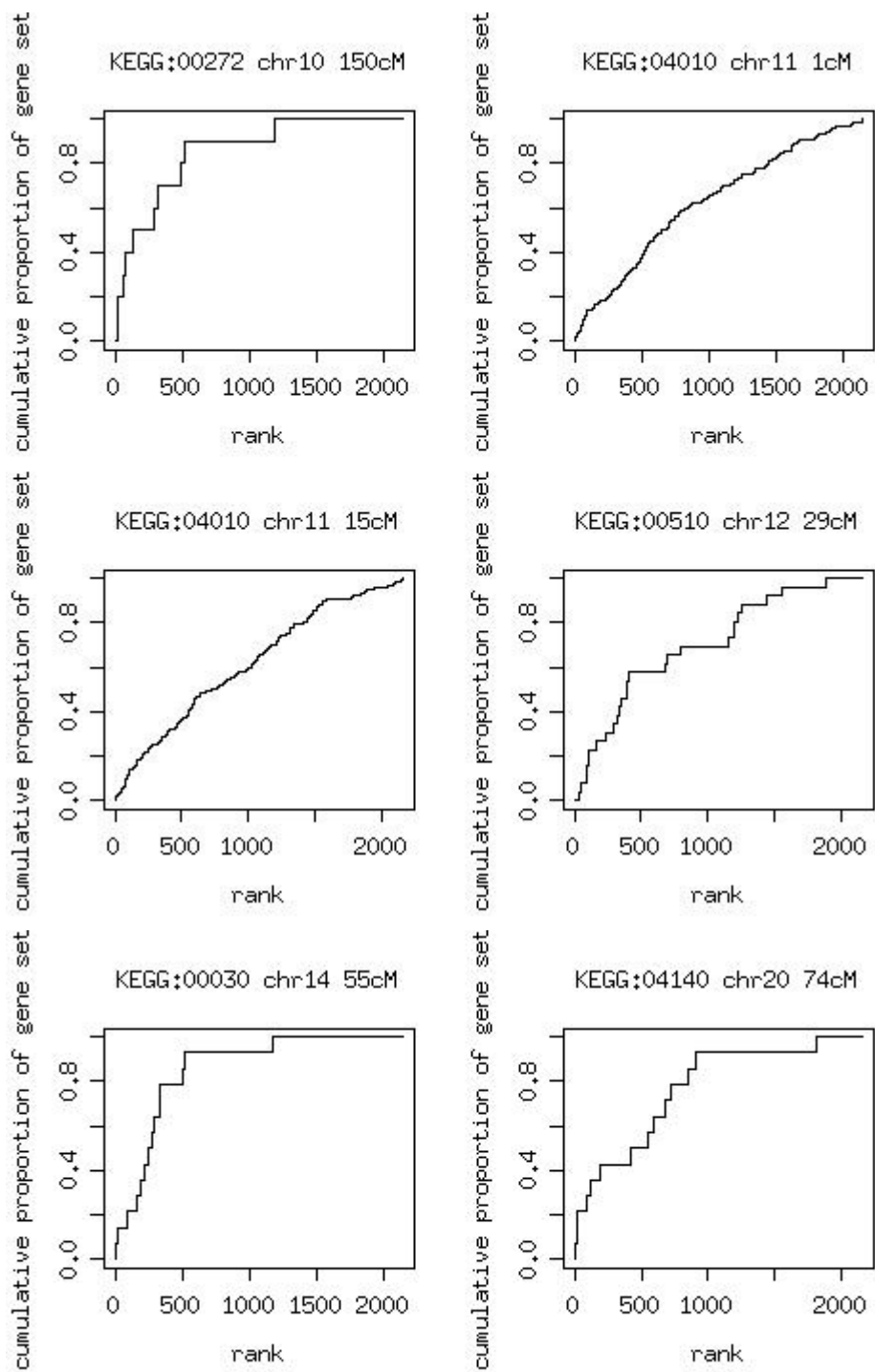


Figure 5. 6

Rank plots for the 10 signals detected by the Wilcoxon test. The ranks of the genes in a KEGG pathway gene set (x-axis) are plotted against the cumulative proportion of genes in the set (y-axis).

From Figure 5.6 it can be seen that the Wilcoxon test was able to detect loci where a large proportion of the gene set occupied high ranking. The pattern is particularly striking with small gene sets; for example, approximately 80% of the genes in the gene set 00030 are with local statistic ranked within the top 500. Although it might be tempting to interpret the Wilcoxon test signals as enrichment of genes of a pathway linked to the eQTL, on a closer look of the results, it can be noted that the local statistics at the Wilcoxon test signals are not very large. The local statistics of the 10 Wilcoxon test signals are shown as a box-and-whiskers plot in Figure 5.7. From this plot it can be seen that the vast majority of the local statistics underlying all of the signals are below the point-wise 5% significance level.

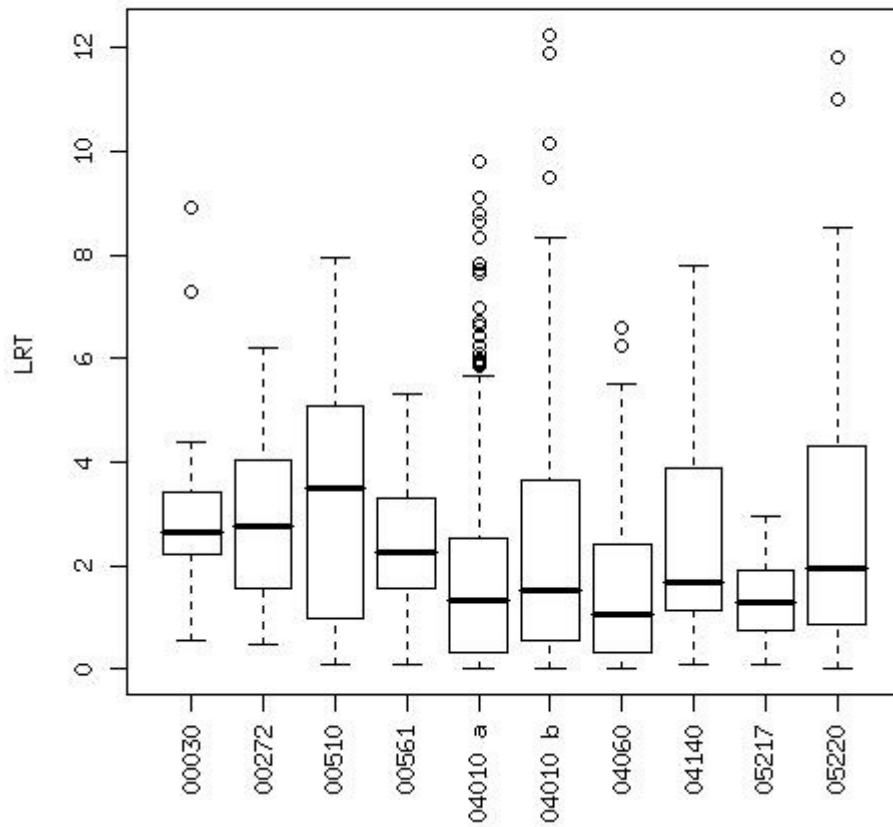


Figure 5. 7

Box-and-whiskers plot showing the LRT statistics of the genes within the gene sets at the significant loci listed in table 5.2. 04010 a represents the hit at 1 cM, and b represents the hit at the 15 cM on chromosome 14. As a guide, the LRT statistic for the point-wise P -value of 0.05 is 3.84. The thin sides of the box indicate the lower quartile and the upper quartile, with the thick line within the box as median. The “whiskers” show the largest / smallest observation that falls within a distance of 1.5 times the box size from the nearest quartile. Data-points beyond are shown as individual circles.

5.4.3 Discussion

Gene set testing was performed on a genome-wide scale. The Fisher's Exact Test and the Wilcoxon Test detected 8 and 10 significant signals respectively. However, on closer inspection, many of the signals appeared unconvincing, despite the small *P*-values of the global statistics obtained.

A number of the signals returned by the Fisher's Exact Test contain gene sets with very few significant genes. As the test focuses on the proportion of genes that is part of a gene set, it is important to keep in mind that extreme *P*-values can be the consequence of small margins in the 2 x 2 contingency table. The most striking result was the signal for pathway 03022; gene set enrichment was implicated when merely 2 significant genes were members of the pathway. This example showed that the Fisher's Exact Test is very sensitive to the number of significant genes at a locus; signals from loci with very few linked genes are not reliable.

Even when the statistics of signals seem more convincing, it is essential to interpret the results carefully. It has been illustrated that looking at the results from a biological angle is extremely crucial. The signals on chromosome 20 seemed exciting at first sight, until it was realised that the genes involved are highly polymorphic and the linkage signals were most likely false positives due to sequence variation on the probe. Locating the significant genes on the 04360 KEGG pathway diagram (Figure 5.5) also revealed that the linkages were not to genes on the same branch of the pathway. However, four genes were found on two distinct branches and the pattern hypothesised in Figure 5.1 could be masked by the fact that some genes were missing from this analysis. Hence, they should not be dismissed as false positive without further consideration. Nevertheless, the branches in the pathways

have quite distinct functions and there would have been no enrichment if those branches were classified as separate pathways in KEGG. Therefore, it is important to bear in mind that there is always a certain level of subjectivity introduced by the database curators. Thus, the validity of any signals should be cautiously examined.

Many of the Wilcoxon Test signals are also somewhat dubious. Although the Wilcoxon Test correctly detected loci where the local statistics of certain gene sets ranked higher than the rest, the linkage evidence for most individual genes at those signals are weak. Similar to the Fisher's Exact Test, the Wilcoxon test is also sensitive to a fairly flat distribution of low local statistic at the locus (i.e. the locus is linked to very few or none of the gene expression traits). Because high rank does not equate to strong linkage evidence, when there is very little evidence of linkage to any genes at a locus, the ranks are meaningless. Consider the signal of pathway 05217 (Figure 5.7), none of the genes were significant, even by ignoring multiple testing. Although the use of a cut-off in Fisher's Exact Test can be criticised for its lack of regard to the continuity of the local statistics, it has the advantage of allowing the user to define what the minimum acceptable level for linkage is. On the other hand, there is no concept of the size of the likelihood ratio statistic in the Wilcoxon Test. If the Wilcoxon Test was conditioned on having even just one or more genes with the local statistic of 10.8 (the cut off used in the Fisher's Exact Test), then all but two signals would have been rejected.

Why do these problems arise with gene set testing in genetical genomics? There are several factors that are likely to be important. Firstly, in genetical genomics, the genotypic effects on the variation of gene expression are tested on a large number of loci. As seen in other published studies, most eQTL are linked to a

small number of genes, except at *trans*-eQTL hotspots. Therefore, there are many loci with few, if any, significant local statistics. Extreme global statistics resulting from these loci can be very misleading. Secondly, only a small fraction of genes on the microarray is mapped to KEGG. The microarray dataset contains probesets that map to 9,296 expressed genes, but only 2,185 genes were used in gene set testing. As over 75% of the genes were not considered by the tests, many eQTL were also dropped from the analysis. Omitting a large number of genes is likely to affect the distribution of the local statistics which could have an effect on the global statistic. Thirdly, many KEGG pathways contain a number of smaller branches. Even when there is signal for enrichment, it is not certain that the genes with a significant local statistic reside on the same branch of the pathway. One potential problem is that the coverage for some pathways is poor on the array, which may explain the failure in detecting linkage to the other genes on the same branch.

Overall, the current study identified only one putative gene set enrichment signal which might be interesting. Yet it is difficult to determine its biological significance. All other signals are likely to be technical artefacts. The small sample size of the current dataset could be attributed to the failure to detect any concrete pathway eQTL signals. Using only 30 RILs, the study might lack the statistical power to find the linkage for the genes downstream of gene pathways. An alternative view is that using only the pathway information from KEGG is too limited. As mentioned above, the annotation of a pathway from one particular database can be very subjective. Curators from different databases may have very different opinion on how inclusive a particular pathway should be, and there can be substantial difference between the same pathways from different databases. Using multiple

pathway resources together may overcome the difficulty in conflicting gene set definition. However, resolving conflicts may also become more difficult using multiple resources if there was a lack of consensus. In addition, the increase in multiple testing is difficult to account for, especially when some gene sets have substantial overlap of genes. Moreover, the permutation step can become more computational intensive.

5.5 Conclusions

In principle, gene set testing should provide extra information on the genetic regulation of pathways. In practice, however, there are still a lot of technical issues to be overcome before it can be widely adopted. Assessing the significance of the gene sets simply by the P -values of both the Fisher's Exact Test and the Wilcoxon Test can be misleading. Further research will be needed to find more suitable methods for testing gene sets with eQTL data. The definition of gene sets, the extent of coverage of the pathways and biological knowledge of the genes involved are all important factors to consider while interpreting the results.

In the next chapter, I present a similar analysis in which the gene sets are defined using Gene Ontology terms. I investigate whether gene set testing would produce a more useable set of results when more genes are included in the analysis.

CHAPTER 6

Gene set testing using Gene Ontology

In Chapter 5 the performance of gene set testing for mapping eQTL was explored, based on the gene set categorisation using the KEGG pathway database (Kanehisa & Goto 2000). A major limiting factor appeared to be that too many genes on the Affymetrix GeneChip™ used in this rat example dataset were not included by KEGG. In this chapter, the Fisher's Exact Test analysis was repeated with Gene Ontology (GO) (Ashburner *et al.* 2000) to define the gene sets. With the gene coverage greater than that of KEGG, it is hoped that some information missed in the univariate analysis can be recovered by jointly considering groups of genes sharing common GO terms.

6.1 Methods

The expression data, the likelihood ratio test statistics (local statistics), and the 1000 sets of local statistics from the null distribution generated by permutations used in Chapter 5 were re-used here. The methodology in this chapter deviated from the last only in the definition of gene sets.

6.1.1 Mapping of probesets to GO

Unlike in KEGG where the entries are self-contained in a flat topology, Gene Ontology terms are organised in a hierarchical structure. These terms describe the functional roles for a group of genes. Each term inherits from one or more parent terms describing the functions in a less specialised way. Naturally, each term can also have one or more child terms to which the functional descriptions are more

specific, and the child GO term inherits a subset of genes from its parents. All GO terms descend from one of the three terms at the top of the hierarchy: Biological Process, Molecular Function and Cellular Component. Genes belonging to the same GO term do not necessarily interact. The description of the GO term applies to all the genes that are members of the term.

To define gene sets, all probesets were first mapped to the EntrezGene database and subsequently to GO terms at all levels using the annotation package “rae230a.db” in Bioconductor (<http://www.bioconductor.org/>). GO terms with very general descriptions are unlikely to be useful for making biological interpretation of their enrichment signals because they tend to encapsulate genes with very diverse functionality. On the other hand, positive signals of GO terms with very specialized descriptions may not be robust because these terms tend to have only very few genes. Therefore, only the gene sets with number of genes between 10 and 100 represented on the microarray were retained. After this step, 5893 genes (63.4% of all known genes found to be expressed in the dataset) from 1676 GO terms entered the gene set analyses.

6.1.2 Gene set testing

Global test statistics for gene sets were obtained using a one-tailed Fisher’s Exact Test along the genome in 1 cM intervals as outlined before in Chapter 5. The Wilcoxon Test was dropped for this analysis. For the Fisher’s Exact Test the point-wise P -value of 0.001 (LRT statistic of 10.8) was used the threshold to divide the local statistics into a “significant” group and a “not significant” group. As before, the “fisher.test” function in R was used to test whether there were over-representations of gene sets amongst the “significant” group.

Gene set testing was applied to the 1000 sets of local statistics generated from linkage analysis on the permuted subjects. As before, global statistics were generated at each cM interval for every gene set. The maximum global statistics for each gene set from every cycle of permutation were collated to derive the null distribution of the global statistics. The global statistics from the un-shuffled dataset were ranked against the null distribution to obtain the genome-wise P -value.

The genome-wise threshold of $P < 0.05$ was used to assess the significance of gene set enrichment at every locus. Signals with fewer than 4 genes exceeding the local statistic threshold of 10.8 were filtered out because those signals were deemed unconvincing in Chapter 5.

An alternative local statistic cut-off (point-wise P -value of 0.005, equivalent to the LRT statistic of 7.8) was used in a repeated analysis. The point of the repeated analysis was not to discover new signals, but to gain an appreciation of how the detected signals could be affected using a different cut-off.

6.2 Results

Using the genome-wise threshold of 0.05 for the global statistic, Fisher's Exact Test identified 40 gene sets showing over-representation in the group of genes with point-wise P -values < 0.001 . Clusters of gene sets were identified as significant for GO terms that were immediately connected; i.e. parent and/or child of significant GO terms were also found to be significant. In those cases the over-representation was conferred by identical genes. The gene set from the most specialised GO description in the cluster was selected to represent the functionality of the significant genes. For example: the GO term representing "neural tube closure" was selected,

whereas the GO terms representing “morphogenesis of epithelium”, “morphogenesis of embryonic epithelium”, “embryonic epithelial tube formation” and “neural tube development” were discarded. After this pruning exercise, signals for 15 gene sets remained and they are listed in Table 6.1.

GO Term	Chr	cM	Minimum genome-wise <i>P</i> -value	Gene set size	No. of genes with significant local statistic *	GO Term description	Ontology #
0004702	1	58 - 60	0.002	36	4 (14)	receptor signaling protein serine/threonine kinase activity	MF
0015370	2	202	0.03	29	5 (71)	solute:sodium symporter activity	MF
0001843	2	206 - 208	0.003	10	4 (48)	neural tube closure	BP
0033014	5	75 - 76	0.009	10	4 (46)	tetrapyrrole biosynthetic process	BP
0016042	5	113 - 114	0.006	60	4 (21)	lipid catabolic process	BP
0001772	5	145	0.041	14	6 (44)	immunological synapse	CC
0030145	7	27 - 29	0.022	56	5 (43)	manganese ion binding	MF
0005681	14	40	0.046	56	8 (74)	spliceosome	CC
0043292	15	6	0.044	59	29 (102)	contractile fiber	CC
0005884	15	11	0.019	24	5 (67)	actin filament	CC
0016712	17	1 - 3	0.008	28	13 (458)	oxidoreductase activity	MF
0005249	17	1 - 2	0.025	61	19 (436)	voltage-gated potassium channel activity	MF
0005179	17	2 - 4	0.028	74	22 (389)	hormone activity	MF
0019882	20	1 - 6	0.004	40	10 (17)	antigen processing and presentation	BP
0042611	20	1 - 5	0.002	24	9 (17)	MHC protein complex	CC

Table 6. 1

Regions in the genome with significant enrichment signal from Fisher's Exact Tests (genome-wide threshold $P \leq 0.05$). * The column shows, at the position with the maximum global statistic, the number of genes in the gene set with LRT statistic above 10.8 (P -value less than 0.001). The total number of genes with significant LRT statistic at the position is quoted in bracket. # The ontology which the GO term stems from. MF = molecular function; BP = biological process; CC = cellular component

Similar to the results for KEGG, the gene sets related to immunity again showed up as significant on chromosome 20 (GO: 0019882 and GO: 0042611) and the RT-1 class genes were again responsible for these *cis*- signals. As discussed in Chapter 5, these signals should be interpreted with extreme caution because of the sequence variation issues on the probed sequences. Other signals were discovered on chromosome 1, 2, 5, 7, 14, 15 and 17 linked to GO terms for all three types of gene ontology: molecular function, biological process and cellular component.

It should be noted that the *P*-values were generated empirically. Take the signals for GO: 0005681 and GO: 0043292 as an example, the nominal *P*-values are actually vastly different (3.6×10^{-7} and 4.3×10^{-37} , respectively). However, if the expression of genes from a set were highly correlated, the global statistic would tend to be large more often simply by chance. This is because when the local statistic of one gene is falsely declared as significant, the other correlated genes in the same set will also likely be falsely declared as significant. The “subject sampling” permutations adjusted for this correlations and produced *P*-values that would truly reflect the significance of the global statistic. The empirical genome-wide *P*-values for these two sets are 0.046 and 0.044, respectively.

Signals of three gene sets were related to ion transport activity. The genes in GO:0015370 are known to be involved in sodium transport. All five significant genes responsible for this signal on chromosome 2 were genes of the solute carrier family: Slc6a20, Slc5a5, Slc6a17, Slc13a3 and Slc20a1. The members of the second gene set, GO:0030145, are known to be involved in manganese ion transport. The significant genes from this gene set, Acvr1c, Tesk1, Galnt1, Bmpr1a and Nudt4 contributed to this signal on chromosome 7. According to gene annotations in the

EntrezGene database, these genes play important roles in the serine / threonine kinase signalling. Interestingly, an enrichment signal for the serine / threonine kinase signalling pathway was detected on chromosome 1. Potassium ion transport is the third ion transport related gene set identified as significant; the signal for GO:0004702 was found on chromosome 17, with 19 genes exceeding the local statistic threshold of 10.8. Most of those genes are members of the potassium channel KCN gene family. Some of the most significant genes include Hcn4, Kcnc1 and Kcnma1, with point-wise *P*-value of 2.9×10^{-5} , 5.9×10^{-6} and 4.7×10^{-6} , respectively.

The chromosome 17 locus where enrichment of linkage for the potassium ion transport genes was detected appeared to be a linkage hotspot: the local statistic for around 400 genes exceeded the point-wise threshold of 10.8 between the 1 cM to the 4 cM positions. The enrichment of oxidoreductase activity genes (GO:0016712) mapped to this region; the signal spanned 1 -3 cM on chromosome 17, with the peak at 1 cM. From this gene set, the local statistics for 13 genes were significant; all of them are members of the cytochrome P450 gene family, such as Cyp4a8, Cyp2d2 and Cyp4b1. The enrichment of hormone activity (GO:0005179) also mapped to this region (2 - 4 cM with the peak at 4 cM), containing genes encoded for various hormones.

Another outstanding signal was on chromosome 5 for the GO term related to muscle fibre. Twenty-nine genes from GO: 0043292 (contractile fibre) were linked to position 6 cM on chromosome 5, including genes encoding for various subunits of skeletal muscle components like actinin, troponin, myosin and tropomyosin. GO:0005884 (actin filament) was also found to be enriched 5 cM further

downstream. However, 4 out of the 5 genes that were significant were part of the 29 significant genes at the upstream signal. Given the overlap in the significant genes and the close proximity of the two signals, it was not clear whether the signal for actin filament was independent to the signal for the contractile fibre.

For the Fisher's Exact Test, the local statistics were dichotomised using an arbitrary cut-off of $P = 0.001$ and subsequently treated as categorical data. An interesting question is to what extent the signals are affected by the choice of the local statistic threshold. Gene set analysis with GO was repeated with a threshold of $P < 0.005$. Out of the 15 gene set eQTL detected using $P < 0.001$, 6 remained significant using the more relaxed threshold for local statistic (shown in Table 6.2).

GO Term	Chr	cM	Minimum genome-wise <i>P</i> -value	Gene set size	No. of genes with significant local statistic *	GO Term description
0001843	2	206	0.047	10	5 (209)	neural tube closure
0006754 **	5	75	0.007	29	10 (178)	ATP biosynthetic process
0005681	14	41	0.033	56	15 (229)	spliceosome
0005179	17	2-5	0.019	74	38 (823)	hormone activity
0019882	20	1 - 7	0.007	40	10 (22)	antigen processing and presentation
0042611	20	1 - 7	0.004	24	9 (22)	MHC protein complex

Table 6. 2

Six remaining enrichment signals from Table 6.2 after lowering the local statistic threshold. * The column shows, at the position with the maximum global statistic, the number of genes in the gene set with LRT statistic above 7.8 (*P*-value less than 0.005). The total number of genes with significant LRT statistic at the position is quoted in bracket. ** The significant genes in GO:0006754 are different to those in GO:0033014 in table 6.1.

Many of the gene set signals disappeared with the change of threshold for dichotomisation, including those with genes encoding for ion transporters and signalling. The signals for categories related to neural tube development, spliceosome and hormone activity remained. At position 75 cM of chromosome 2, the signal for tetrapyrrole biosynthetic process was replaced by ATP biosynthetic process.

6.3 Discussion

In eQTL mapping, the power of the univariate approach, where linkage for a single gene is evaluated one at a time, is heavily penalised by the correction for multiple testing. Evaluating multiple genes offers a post-hoc method to combine knowledge in biology with statistic to explore the data set beyond the breadth of eQTL detection that is capable on a purely statistical basis. The analysis with Gene Ontology partly addressed the major obstacle that was faced in chapter 5; substantially more genes probed by the microarray were annotated in GO than in KEGG. Although common pathways are not implicated by sharing of GO terms, finding pleiotropy for a number of genes from the same gene family can be interesting; for example, it can be treated as a first clue for inferring the mechanisms underlying gene co-regulation.

The SHR rat strain was developed as a model for studying hypertension and metabolic diseases. In the BXH/HXB RIL panel, I identified 13 loci linked to gene set enrichment for various functions; the most interesting ones are related to ions transport activity, lipid metabolism, oxidoreductase activity and hormone activity. While the balance of potassium and sodium ions has been linked to hypertension

(Khaw & Barrett-Connor 1988), the other gene sets contain genes with important roles in metabolism and energy balance. Variants with small contributions to the expression levels in those genes could help to drive the physiological defects in metabolism in hypertensive rats (Aitman *et al.* 1997). There are also a number of enrichment signals for gene sets with less obvious relevance to the hypertension and metabolic diseases. Nonetheless, these signals are constituted of at least 4 (and for some signals many more) functionally related genes, all with very small point-wise P -values in the region between 10^{-3} and 10^{-6} . Hence, these signals are far more convincing than those identified with the KEGG database in chapter 5, either by the Fisher's Exact Test or the Wilcoxon Test.

In this study, I also investigated the effect of altering the local statistic cut-off in the Fisher's Exact Test. I chose to relax the cut-off from $P < 0.001$ to $P < 0.005$ (local statistic from 10.8 to 7.8). I did not increase the stringency of the cut-off because the objective of gene set testing was to detect weak linkage effects. Furthermore, I noted from chapter 5 that the Fisher's Exact Test did not work well when there were very few significant genes. By relaxing the cut-off, an increase in the number of significant genes was observed. One could expect new signals to arise for gene sets with enrichment of local statistics between 7.8 and 10.8. Although signals for those putative eQTL of weak effects might be interesting, I focused on how many of the original signals remained. Signals would remain only if there was an enrichment of genes with local statistic above 7.8 as well as above 10.8.

Many of the original signals disappeared after the lowering of the cut-off. One reason of why those signals were no longer detectable with the lower cut-off could be due to the increase in the amount of false positives in the "significant"

group. In the 2 x 2 table, this would most likely increase the number in the cell for the genes called significant and not in the gene set; hence the signal for gene set enrichment would get diluted. This would be particularly true for gene set with local statistics that could be neatly divided by the cut-off of 10.8; i.e. there was a clear separation of two groups of local statistics (high and low). On the other hand, relaxing the cut-off can strengthen the significance of other signals if the enrichment of high local statistics stretches over either side of the original cut-off, as shown for the spliceosome and hormone activity gene sets. The results suggest that the choice of cut-off can have a dramatic effect on Fisher's Exact Test, and there is no definitive way to decide what the appropriate cut-off should be. The results also indicate that there may be optimal cut-offs for different gene sets. Clearly, this arbitrary is not very satisfactory when testing a large number of gene sets. In addition, it was discussed in chapter 5 that the dichotomisation in Fisher's Exact Test does not use all the information available in the local statistics. In theory, methods for detecting gene set enrichment which do not rely on a rigid cut-off should be preferred. However, more research is required to address the problems with the rank-based methods in eQTL mapping as discussed in chapter 5.

It should be noted that the empirical genome-wide threshold was derived on a per gene set basis. In other words, multiple testing was accounted for by testing each gene set over the entire genome, but not for the number of gene sets being tested. The matter was complicated by the fact that gene sets overlapped with one another to a large extent, hence the tests were not independent. Also, computational difficulties restricted the number of permutation here to 1,000 rounds, which meant the smallest empirical *P*-value possible for the global statistic was 0.001. To use methods such as

Bonferroni or false discovery rate to correct for testing multiple gene sets, there would not be any significant signal at all. However, as this approach was employed for data exploration, I argue that interpretation should not be entirely based on *P*-values. It is demonstrated that gene set testing can be useful in highlighting a substantial number of genes which belong to common functional categories with a fair level of linkage evidence. It should be noted that this approach was never meant to prove that the highlighted genes were genuinely linked; instead, it was intended to help the researchers in prioritising their research efforts after they completed their investigation with the top marker / genes. In this study, even with only a modest number of permutations, it was sufficient to narrow the focus from over a thousand GO terms down to just over a handful. From this point onward, further work should proceed with a stronger emphasis from a biological point of view, combined with other evidence gathered from external sources, to assess the validity of these signals.

Finally, this analysis demonstrated the value of gene annotations in a well managed bioinformatics database. Information in a format that can be data-mined using computational tools is vital to genetics and genomics research. Regarding this aspect, the resources in humans and model organisms are in a much more advanced position compared to livestock species. It is true that comparative genomics allows mapping of orthologous genes from a farm animal species to humans or model organisms and subsequently tapping into the resources in those organisms. However, even in humans and model organisms, a substantial amount of annotations was inferred via *in-silico* methods such as matching of sequence motifs. As a result, misclassifications can occur. Mapping annotations across different species increases the risk of further propagating the erroneous annotations. Efforts are underway to

create bioinformatics resources that are specific for livestock species; for example, gene ontology links to EST sequences in cattle and swine had been established in the past (Harhay & Keele 2003) and more development of manual GO annotation for livestock species, particularly in chicken, is taking shape (McCarthy *et al.* 2007). As the efforts in consolidating livestock genomic resources are gathering pace, it should alleviate some of the challenges in conducting eQTL experiment and the post-analysis in livestock species in the foreseeable future.

6.4 Conclusion

With a wide coverage of genes, gene set testing can be fruitful for identifying putative eQTL with moderate effects. The choice of cut-offs for the Fisher's Exact Test can affect the results and various cut-offs over a small range may be appropriate for different gene sets, depending on the distribution of the local statistics within gene sets. Nevertheless, as a general data exploratory tool, this approach enables a significant reduction in the search space to a manageable size in order for manual data-mining to be carried out by bench biologists.

CHAPTER 7

General discussion and perspective

This chapter features a summary of the key contributions of this thesis. It also provides an outlook on how genetical genomics may continue to develop, and reviews the potential pitfalls that constitute some of the major obstacles in the field other than those already mentioned in earlier chapters. Finally, a perspective on the way genetical genomics may impact on aspects of livestock genetics is given.

7.1 Summary

Investigations on various facets of genetical genomics were presented in chapter 2 - 6. The “take home messages” from these investigations include:

Good experimental design and strict data quality control are absolutely vital for making sense of the final results.

One important conclusion drawn from chapter 2 and 3 is that we should spend substantial efforts in planning prior to conducting an experiment in genetical genomics. Similar to mapping genetic loci for other complex traits, large sample size is required to attain adequate statistical power in eQTL studies. The general rule of thumb is “the greater the sample size the better”. However, I have shown that by devoting efforts to directing the resources most relevant to the question of research interest, an efficient experimental design can make the most from a fixed research budget to maximise sample size and power. Again, power is an important issue when eQTL are combined with functional QTL in an integrated analysis. An experiment

has to be well designed so that the subjects genotyped match those phenotyped and match those expression-profiled; noise from systematic sources such as batch and sample-handling must be minimised; and the potential confounding effect, e.g. using multiple breeds, should be considered before samples are selected. Equally important as having a good experimental design is to have rigorous quality control while data are generated and managed. As many researchers in bioinformatics would quote: “garbage in, garbage out”; the quality of all the data including the mapping annotations needs to be of a dependable standard for the full potential of genetical genomics to be realised.

Multiple testing continues to be problematic for assessing the significance of eQTL.

The multiplicity of eQTL analysis poses considerable challenges; one has to account not only for the multiple loci in a genome-wide search but also for massive number of traits, some of them correlated. I have shown in chapter 4 that by using false discovery rate (FDR) (Storey & Tibshirani 2003) and by reducing the dimension of the dataset (filtering out irrelevant expression traits / extreme genotypes), one can better control the level of false positive discovery and suffer to a lesser extent the loss of power due to test multiplicity. However, the correlation between transcripts is still likely to introduce some bias in estimating FDR. The jury is still out on how to properly account for multiple testing in genetical genomics. Stranger *et al.* (2005) argued that given each expression trait has its own properties of variance and inheritance, it would seem unlikely that genome- and experimental-wise thresholds provide the optimal means for assessing significance. Furthermore, I

made the suggestion that *cis*-eQTL is subjected to a lesser burden of multiple testing than *trans*-eQTL. This view has been echoed by a recent review by Gilad *et al.* (2008). Therefore, partitioning proximal and distal loci for separate analyses should lead to an increase in power. As with any microarray experiments, the golden rule is to always conduct proper validation of the positive findings (Allison *et al.* 2006). Therefore, a pragmatic approach would be to apply a multiple testing correction method that is not overly conservative (e.g., FDR instead of Bonferroni) in the first place and subsequently validate the biological relevance of the positives experimentally. Very recently, a version of FDR which is weighted by expression correlation has been proposed for eQTL mapping (Chen *et al.* 2008a). The method is still to be evaluated independently. However, it is unlikely to be the final answer to multiple testing in the context of eQTL, because there are other factors that will influence detection; for example, heritability of the expression trait and the LD pattern of the genome.

Gene set testing extracts more information from the data than univariate statistics. Advances in genomics will enhance the value of this approach.

As a complementary approach to tackle multiple testing in genetical genomics, it is shown in chapter 5 and 6 that gene set testing is useful in highlighting co-regulation of functionally related genes. Other researchers have recently begun to apply gene set testing to eQTL analysis. For example, Emilsson *et al.* (2008) identified significant GO enrichment of eQTL genes for inflammatory response and macrophage activations in a cross between two inbred strain of mice, although it is not clear in that case whether potential sequence variation artefacts had been

accounted for. I have shown that gene set testing is a practical method to enable the use of anti-conservative thresholds and yet guard against an unmanageable inflation in false positives. For example, it was possible to highlight the possible links between ion transport activities and metabolic abnormalities in the SHR strain of rats using GO enrichment analysis. A recent study also demonstrated the use of a similar approach in identifying a locus associated with the arachidonic acid metabolic pathway in a rat model for cardiac diseases (Monti *et al.* 2008). At the same time, it is important to note that the results are strongly influenced by the choice of gene set definition and test statistic. Furthermore, users should be aware that the annotations in KEGG or GO can vary in their quality: merely 20% of the rat genes annotated by GO are supported by experimental evidence; the rest are either inferred from electronic annotation or from unknown sources (Rhee *et al.* 2008). More research will be needed to identify the optimal statistics for testing gene sets, and further development in pathway biology and functional genomics is necessary to ensure improvement in the robustness of this method.

7.2 genetical genomics: future directions and pitfalls

The application of genetical genomics to understand complex traits such as human diseases has rapidly gathered pace. Some of the most recent published work reported the use of considerably larger sample size, denser marker map and greater coverage of the transcriptome than studies published merely few years ago. For example, Göring *et al.* (2007) sampled the lymphocyte transcriptional profiles from 1,240 participants in the San Antonio Family heart Study; Dixon *et al.* (2007) genotyped > 400,000 SNPs and assayed the expression of > 54,000 transcripts from

400 children. Stranger *et al.* (2007) has not only study the effect of SNPs on gene expression, but also the effect of structural variations in the genome, including copy number variants (CNVs), on global expression phenotypes. With the decline in cost and increase in throughput for many genomic technologies, it is anticipated that bigger studies with even larger sample size and more comprehensive coverage of the genome will become the “bottom line” for genetical genomics: there will be no place for small and under-powered studies. At the same time, large scale collections of matched phenotypic records, such as clinical traits, are crucial to support the effort in mapping eQTL to enable the reconstruction of molecular networks that cause disease (Emilsson *et al.* 2008; Chen *et al.* 2008b) and other complex traits of interest. We will also need more novel statistical and computational methods to be developed to disentangle the complexity in the data, and ultimately produce detailed networks to be tested on the bench.

As gene expression represents only one level of regulation in a biological system, the future of genetical genomics will also encompass other -omics technologies. A study in the plant *Arabidopsis thaliana* first demonstrated the genetics mapping of variation in metabolomics (Keurentjes *et al.* 2006). At present, there are still issues with considerable technical uncertainty and high cost for QTL mapping in the context of proteomics and metabolomics to be commonplace. Once those issues are overcome in the future, such QTL studies, in conjunction with genetical genomics, will enable researchers to ask a number of questions: what is the genetic mechanism underlying variation in gene products and metabolites? How does the variation in protein expression relate to gene expression? How do protein and metabolic networks drive complex phenotypes? Do proteomic data present a more

accurate picture than networks derived from gene expression data? The genetic mapping of -omics variation at multiple biological regulatory levels, collectively known as “systems genetics” (Threadgill 2006), can potentially revolutionise the approach which geneticists will use to uncover the intricate molecular mechanisms underlying a biological system.

Although we look forward to an exciting future ahead in the field of genetical genomics, it is important not to lose sight on some of the technical pitfalls associated with eQTL mapping. As discussed in this thesis and elsewhere (for example: Alberts *et al.* 2005; Alberts *et al.* 2007), batch and sequence variation effects can introduce serious confounding results unless they are detected and correctly accounted for. Tissue specificity is another important factor which requires particular caution: regulatory networks has been shown to be highly tissue-specific (Hovatta *et al.* 2007). The fine tissue-specificity may invalidate some of the eQTL identified from experiments with RNA extracted from whole or large regions of organs. Even if a specific tissue is used in a study, there is risk of erroneous eQTL due to contamination by cells from neighbouring tissues.

A study on how replicable eQTL are (Peirce *et al.* 2006) found that replicable eQTL were disproportionately *cis*-acting, and few *trans*-acting eQTL were successfully confirmed. Their results suggest that while genetical genomics is effective for identifying *cis*-acting loci which are candidates for major effect QTL, indirect genetic regulation represented by *trans*-acting loci is difficult to detect. It may be that *trans*- effects are generally weak and are of lesser statistical significance, and for those reasons these effects are more sensitive to the technical noise of the experiment. In addition, environmental effects (Gibson 2008) can also significantly

contribute to variation in gene expression; eQTL could appear or disappear, or exhibit opposite direction in its allelic association depending on external stimuli. This phenomenon is known as plasticity. Plasticity (Li *et al.* 2006), as well as sex specificity (Wang *et al.* 2006), in eQTL have been shown empirically to be significant. Ultimately, all of these pitfalls relate back to the necessity of a good experimental design. To produce results that are scientifically sound, researchers need to conduct eQTL mapping using samples of appropriate tissue, age, sex, and exposure to external environmental factors that are relevant to the function trait of interest. However, knowing and accessing the right type of cells at the right time from the right environment is usually not a trivial task (Weiss 2008).

7.3 The use of eQTL in livestock genetics

Typical livestock populations have a number of desirable properties which make them particularly suitable for eQTL discovery using a genetical genomics approach (Haley & de Koning 2006). Livestock species generally have large family sizes and extensive phenotypic records are routinely collected on a large number of animals by the breeding industry for estimating breeding values. Indeed, some of the breeding companies, such as PIC, have an enormous collection of tissue samples and extensive phenotypic records for farm animals with known lineage over large numbers of generations. Environmental conditions in genetic nucleus farms used for artificial selection are also well controlled to allow a fair comparison between subjects. Furthermore, many QTL with known effect size exist in livestock populations and the knowledge is out in the public domain. Therefore, the integrative

approach combining functional QTL and expression QTL could be achieved by profiling the gene expression of matched subjects.

However, genetical genomics has not been taken up by the livestock breeding industry. In Chapter 1, I described the potential use of gene expression QTL for designing breeding programmes in what can be referred to as expression-based marker assisted selection (Kadarmideen *et al.* 2006). Despite the scope of eQTL has in animal breeding, the cost of microarray seems still too high for this approach to be practical. A pure quantitative approach known as genome-wide selection (Meuwissen *et al.* 2001) looks likely to become more routinely used by the industry to design breeding programmes. This approach uses genotypes of a large number of SNPs throughout the genome, without regard to QTL locations and functions, to predict breeding values. Arguably, studying gene functions is currently too expensive when considering the small profit margins on which breeding companies currently operate.

On the contrary, genetical genomics may have wider applications in the animal and human health industry. Understanding the molecular basis of health traits and resistance to pathogens could be useful for disease prevention and discovery of new treatments for animal diseases. In crop science, there are already examples of using genome-wide eQTL mapping as a new tool to find candidate genes related to complex traits, such as resistance to wheat gem rust pathogens in barley (Druka *et al.* 2008). There is no reason why genetical genomics cannot be applied in similar ways in studying disease susceptibility traits in livestock species. The stakes in combating animal diseases are particularly high in the current era of emerging diseases. Diseases such as avian influenza (Yamada *et al.* 2008) and foot and mouth disease

(Haydon *et al.* 2004) are having tremendous impact on the social and economic aspect of today's society. At the same time, farm animals are increasingly being used as human disease models because they are closer related to humans than rodents are (for example: Rogers *et al.* 2008). Rising interests in sustainable agriculture (for example: <http://www.sabre-eu.eu/>) also ensure that there is a place for functional genomics in applied agricultural research. For these reasons, I maintain the view that genetical genomics will be a valuable tool in livestock genetics.

Bibliography

- Abecasis, G. R., L. R. Cardon, and W. O. Cookson, 2000 A general test of association for quantitative traits in nuclear families. *Am.J.Hum.Genet.* **66**: 279-292.
- Aitman, T. J., T. Gotoda, A. L. Evans, H. Imrie, K. E. Heath *et al.* 1997 Quantitative trait loci for cellular defects in glucose and fatty acid metabolism in hypertensive rats. *Nat.Genet.* **16**: 197-201.
- Al-Shahrour, F., R. az-Uriarte, and J. Dopazo, 2004 FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* **20**: 578-580.
- Alberts, R., P. Terpstra, L. V. Bystrykh, G. de Haan, and R. C. Jansen, 2005 A statistical multiprobe model for analyzing cis and trans genes in genetical genomics experiments with short-oligonucleotide arrays. *Genetics* **171**: 1437-1439.
- Alberts, R., P. Terpstra, M. Hardonk, L. V. Bystrykh, G. de Haan *et al.* 2007 A verification protocol for the probe sequences of Affymetrix genome arrays reveals high probe accuracy for studies in mouse, human and rat. *Bmc Bioinformatics* **8**.
- Alexa, A., J. Rahnenfuhrer, and T. Lengauer, 2006 Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics* **22**: 1600-1607.
- Alfonso, L., and C. S. Haley, 1998 Power of different F-2 schemes for QTL detection in livestock. *Animal Science* **66**: 1-8.
- Allison, D. B., X. Cui, G. P. Page, and M. Sabripour, 2006 Microarray data analysis: from disarray to consolidation and consensus. *Nat.Rev.Genet.* **7**: 55-65.
- Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler *et al.* 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat.Genet.* **25**: 25-29.
- Aulchenko, Y. S., D. J. de Koning, and C. Haley, 2007a Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**: 577-585.
- Aulchenko, Y. S., S. Ripke, A. Isaacs, and C. M. van Duijn, 2007b GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**: 1294-1296.
- Balding, D. J., 2006 A tutorial on statistical methods for population association studies. *Nat.Rev.Genet.* **7**: 781-791.

- Barry, W. T., A. B. Nobel, and F. A. Wright, 2005 Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* **21**: 1943-1949.
- Beissbarth, T., and T. P. Speed, 2004 Gostat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**: 1464-1465.
- Bidanel, J. P., D. Milan, N. Iannuccelli, Y. Amigues, M. Y. Boscher *et al.* 2001 Detection of quantitative trait loci for growth and fatness in pigs. *Genetics Selection Evolution* **33**: 289-309.
- Bing, N., and I. Hoeschele, 2005 Genetical genomics analysis of a yeast segregant population for transcription network inference. *Genetics* **170**: 533-542.
- Boerwinkle, E., R. Chakraborty, and C. F. Sing, 1986 The use of measured genotype information in the analysis of quantitative phenotypes in man. I. Models and analytical methods. *Ann.Hum.Genet.* **50**: 181-194.
- Boyle, E. I., S. A. Weng, J. Gollub, H. Jin, D. Botstein *et al.* 2004 GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**: 3710-3715.
- Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak, 2002 Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752-755.
- Broman, K. W., H. Wu, S. Sen, and G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**: 889-890.
- Bueno Filho, J. S., S. G. Gilmour, and G. J. Rosa, 2006 Design of microarray experiments for genetical genomics studies. *Genetics* **174**: 945-957.
- Bystrykh, L., E. Weersing, B. Dontje, S. Sutton, M. T. Pletcher *et al.* 2005 Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat.Genet.* **37**: 225-232.
- Carlborg, O., D. J. de Koning, K. F. Manly, E. Chesler, R. W. Williams *et al.* 2005 Methodological aspects of the genetic dissection of gene expression. *Bioinformatics.* **21**: 2383-2393.
- Chen, L., T. Tong, and H. Zhao, 2008a Considering dependence among genes and markers for false discovery control in eQTL mapping. *Bioinformatics* **24**: 2015-2022.
- Chen, Y. Q., J. Zhu, P. Y. Lum, X. Yang, S. Pinto *et al.* 2008b Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**: 429-435.
- Chesler, E. J., L. Lu, S. Shou, Y. Qu, J. Gu *et al.* 2005 Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat.Genet.* **37**: 233-242.

- Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K. Y. Jen *et al.* 2003 Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat.Genet.* **33**: 422-425.
- Churchill, G. A., and R. W. Doerge, 2008 Naive application of permutation testing leads to inflated type I error rates. *Genetics* **178**: 609-610.
- Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. *Genetics* **138**: 963-971.
- Ciobanu, D., J. Bastiaansen, M. Malek, J. Helm, J. Woollard *et al.* 2001 Evidence for new alleles in the protein kinase adenosine monophosphate-activated gamma(3)-subunit gene associated with low glycogen content in pig skeletal muscle and improved meat quality. *Genetics* **159**: 1151-1162.
- Ciobanu, D. C., J. W. M. Bastiaansen, S. M. Lonergan, H. Thomsen, J. C. M. Dekkers *et al.* 2004 New alleles in calpastatin gene are associated with meat quality traits in pigs. *Journal of Animal Science* **82**: 2829-2839.
- D'haeseleer, P., S. Liang, and R. Somogyi, 2000 Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics.* **16**: 707-726.
- Darvasi, A., and M. Soller, 1992 Selective Genotyping for Determination of Linkage Between A Marker Locus and A Quantitative Trait Locus. *Theoretical and Applied Genetics* **85**: 353-359.
- de Koning, D. J., C. P. Cabrera, and C. S. Haley, 2007 Genetical genomics: combining gene expression with marker genotypes in poultry. *Poult.Sci.* **86**: 1501-1509.
- de Koning, D. J., and C. S. Haley, 2005 Genetical genomics in humans and model organisms. *Trends in Genetics* **21**: 377-381.
- de la Fuente, A., N. Bing, I. Hoeschele, and P. Mendes, 2004 Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics.* **20**: 3565-3574.
- DeCook, R., S. Lall, D. Nettleton, and S. H. Howell, 2006 Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics* **172**: 1155-1164.
- Dixon, A. L., L. Liang, M. F. Moffatt, W. Chen, S. Heath *et al.* 2007 A genome-wide association study of global gene expression. *Nat.Genet.* **39**: 1202-1207.
- Dobbin, K., and R. Simon, 2002 Comparison of microarray designs for class comparison and class discovery. *Bioinformatics.* **18**: 1438-1445.
- Druka, A., E. Potokina, Z. Luo, N. Bonar, I. Druka *et al.* 2008 Exploiting regulatory variation to identify genes underlying quantitative resistance to the wheat

stem rust pathogen *Puccinia graminis* f. sp. *tritici* in barley.
Theor.Appl.Genet. **117**: 261-272.

- Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein, 1998 Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**: 14863-14868.
- Emilsson, V., G. Thorleifsson, B. Zhang, A. S. Leonardson, F. Zink *et al.* 2008 Genetics of gene expression and its effect on disease. *Nature* **452**: 423-4U2.
- Falcon, S., and R. Gentleman, 2007 Using GOstats to test gene lists for GO term association. *Bioinformatics* **23**: 257-258.
- Fan, J. B., M. S. Chee, and K. L. Gunderson, 2006 Highly parallel genomic assays. *Nat.Rev.Genet.* **7**: 632-644.
- Flint, J., W. Valdar, S. Shifman, and R. Mott, 2005 Strategies for mapping and cloning quantitative trait genes in rodents. *Nat.Rev.Genet.* **6**: 271-286.
- Friedman, N., 2004 Inferring cellular networks using probabilistic graphical models. *Science* **303**: 799-805.
- Fu, J. Y., and R. C. Jansen, 2006 Optimal design and analysis of genetic studies on gene expression. *Genetics* **172**: 1993-1999.
- Fulker, D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt, 1999 Combined linkage and association sib-pair analysis for quantitative traits. *Am.J.Hum.Genet.* **64**: 259-267.
- Gibson, G., 2008 The environmental contribution to gene expression profiles. *Nat.Rev.Genet.* **9**: 575-581.
- Gibson, G., and B. Weir, 2005 The quantitative genetics of transcription. *Trends in Genetics* **21**: 616-623.
- Gilad, Y., S. A. Rifkin, and J. K. Pritchard, 2008 Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet.* **24**: 408-415.
- Goeman, J. J., and P. Buhlmann, 2007 Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**: 980-987.
- Goring, H. H., J. E. Curran, M. P. Johnson, T. D. Dyer, J. Charlesworth *et al.* 2007 Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nat.Genet.* **39**: 1208-1216.
- Haley, C., and D. J. de Koning, 2006 Genetical genomics in livestock: potentials and pitfalls. *Anim Genet.* **37 Suppl 1**: 10-12.

- Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**: 315-324.
- Haley, C. S., S. A. Knott, and J. M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**: 1195-1207.
- Harhay, G. P., and J. W. Keele, 2003 Positional candidate gene selection from livestock EST databases using Gene Ontology. *Bioinformatics* **19**: 249-255.
- Haseman, J. K., and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. *Behav. Genet.* **2**: 3-19.
- Haydon, D. T., R. R. Kao, and R. P. Kitching, 2004 The UK foot-and-mouth disease outbreak - the aftermath. *Nat.Rev.Microbiol.* **2**: 675-681.
- Hillier, L. W., W. Miller, E. Birney, W. Warren, R. C. Hardison *et al.* 2004 Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**: 695-716.
- Hovatta, I., M. A. Zapala, R. S. Broide, E. E. Schadt, O. Libiger *et al.* 2007 DNA variation and brain region-specific expression profiles exhibit different relationships between inbred mouse strains: implications for eQTL mapping studies. *Genome Biol.* **8**: R25.
- Hubner, N., C. A. Wallace, H. Zimdahl, E. Petretto, H. Schulz *et al.* 2005 Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat.Genet.* **37**: 243-253.
- Irizarry, R. A., B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs *et al.* 2003 Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**: e15.
- Jannink, J. L., 2005 Selective phenotyping to accurately map quantitative trait loci. *Crop Science* **45**: 901-908.
- Jansen, R. C., 2003 Studying complex biological systems using multifactorial perturbation. *Nature Reviews Genetics* **4**: 145-151.
- Jansen, R. C., and J. P. Nap, 2001 Genetical genomics: the added value from segregation. *Trends in Genetics* **17**: 388-391.
- Jin, C., H. Lan, A. D. Attie, G. A. Churchill, D. Bulutuglo *et al.* 2004 Selective phenotyping for increased efficiency in genetic mapping studies. *Genetics* **168**: 2285-2293.
- Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel *et al.* 2001 The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat.Genet.* **29**: 389-395.

- Kadarmideen, H. N., and L. L. G. Janss, 2007 Population and systems genetics analyses of cortisol in pigs divergently selected for stress. *Physiological Genomics* **29**: 57-65.
- Kadarmideen, H. N., P. von Rohr, and L. L. G. Janss, 2006 From genetical genomics to systems genetics: potential applications in quantitative genomics and animal breeding. *Mammalian Genome* **17**: 548-564.
- Kanehisa, M., and S. Goto, 2000 KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28**: 27-30.
- Kerr, M. K., and G. A. Churchill, 2001 Statistical design and the analysis of gene expression microarray data. *Genet.Res.* **77**: 123-128.
- Keurentjes, J. J., J. Fu, C. H. de Vos, A. Lommen, R. D. Hall *et al.* 2006 The genetics of plant metabolism. *Nat.Genet.* **38**: 842-849.
- Khaw, K. T., and E. Barrett-Connor, 1988 The association between blood pressure, age, and dietary sodium and potassium: a population study. *Circulation* **77**: 53-61.
- Kim, S. K., J. Lund, M. Kiraly, K. Duke, M. Jiang *et al.* 2001 A gene expression map for *Caenorhabditis elegans*. *Science* **293**: 2087-2092.
- Kirkpatrick, S., C. D. Gelatt, Jr., and M. P. Vecchi, 1983 Optimization by Simulated Annealing. *Science* **220**: 671-680.
- Kruglyak, L., 2008 The road to genome-wide association studies. *Nat.Rev.Genet.* **9**: 314-318.
- Laird, N. M., and C. Lange, 2006 Family-based designs in the age of large-scale gene-association studies. *Nat.Rev.Genet.* **7**: 385-394.
- Lander, E. S., and D. Botstein, 1989 Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185-199.
- Lee, H. K., W. Braynen, K. Keshav, and P. Pavlidis, 2005 ErmineJ: Tool for functional analysis of gene expression data sets. *Bmc Bioinformatics* **6**.
- Li, Y., O. A. Alvarez, E. W. Gutteling, M. Tijsterman, J. Fu *et al.* 2006 Mapping determinants of gene expression plasticity by genetical genomics in *C. elegans*. *PLoS.Genet.* **2**: e222.
- Lynch M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc..
- Malek, M., J. C. M. Dekkers, H. K. Lee, T. J. Baas, K. Prusa *et al.* 2001 A molecular genome scan analysis to identify chromosomal regions influencing economic traits in the pig. II. Meat and muscle composition. *Mammalian Genome* **12**: 637-645.

- Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly, 2004 The effects of human population structure on large genetic association studies. *Nat.Genet.* **36**: 512-517.
- McCarthy, F. M., S. M. Bridges, N. Wang, G. B. Magee, W. P. Williams *et al.* 2007 AgBase: a unified resource for functional analysis in agriculture. *Nucleic Acids Res.* **35**: D599-D603.
- Mehrabian, M., H. Allayee, J. Stockton, P. Y. Lum, T. A. Drake *et al.* 2005 Integrating genotypic and expression data in a segregating mouse population to identify 5-lipoxygenase as a susceptibility gene for obesity and bone traits. *Nat.Genet.* **37**: 1224-1233.
- Meuwissen, T. H., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**: 1819-1829.
- Milan, D., J. P. Bidanel, N. Iannuccelli, J. Riquet, Y. Amigues *et al.* 2002 Detection of quantitative trait loci for carcass composition traits in pigs. *Genetics Selection Evolution* **34**: 705-728.
- Monks, S. A., A. Leonardson, H. Zhu, P. Cundiff, P. Pietrusiak *et al.* 2004 Genetic inheritance of gene expression in human cell lines. *Am.J.Hum.Genet.* **75**: 1094-1105.
- Monti, J., J. Fischer, S. Paskas, M. Heinig, H. Schulz *et al.* 2008 Soluble epoxide hydrolase is a susceptibility factor for heart failure in a rat model of human disease. *Nat.Genet.* **40**: 529-537.
- Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens *et al.* 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743-747.
- Nettleton, D., and D. Wang, 2006 Selective transcriptional profiling for trait-based eQTL mapping. *Animal Genetics* **37**: 13-17.
- Parr, T., P. L. Sensky, G. P. Scothern, R. G. Bardsley, P. J. Buttery *et al.* 1999 Relationship between skeletal muscle-specific calpain and tenderness of conditioned porcine longissimus muscle. *Journal of Animal Science* **77**: 661-668.
- Pastinen, T., B. Ge, and T. J. Hudson, 2006 Influence of human genome polymorphism on gene expression. *Hum.Mol.Genet.* **15 Spec No 1**: R9-16.
- Peirce, J. L., H. Li, J. Wang, K. F. Manly, R. J. Hitzemann *et al.* 2006 How replicable are mRNA expression QTL? *Mamm.Genome* **17**: 643-656.
- Perez-Enciso, M., M. A. Toro, M. Tenenhaus, and D. Gianola, 2003 Combining gene expression and molecular marker information for mapping complex trait genes: a simulation study. *Genetics* **164**: 1597-1606.

- Piepho, H. P., 2005 Optimal allocation in designs for assessing heterosis from cDNA gene expression data. *Genetics* **171**: 359-364.
- R Development Core Team, 2007 *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- Rhee, S. Y., V. Wood, K. Dolinski, and S. Draghici, 2008 Use and misuse of the gene ontology annotations. *Nat.Rev.Genet.* **9**: 509-515.
- Risch, N., and K. Merikangas, 1996 The future of genetic studies of complex human diseases. *Science* **273**: 1516-1517.
- Rockman, M. V., and L. Kruglyak, 2006 Genetics of global gene expression. *Nat.Rev.Genet.* **7**: 862-872.
- Rogers, C. S., W. M. Abraham, K. A. Brogden, J. F. Engelhardt, J. T. Fisher *et al.* 2008 The porcine lung as a potential model for cystic fibrosis. *Am.J.Physiol Lung Cell Mol.Physiol* **295**: L240-L263.
- Rosa, G. J., N. de Leon, and A. J. Rosa, 2006 Review of microarray experimental design strategies for genetical genomics studies. *Physiol Genomics* **28**: 15-23.
- Schadt, E. E., J. Lamb, X. Yang, J. Zhu, S. Edwards *et al.* 2005 An integrative genomics approach to infer causal associations between gene expression and disease. *Nat.Genet.* **37**: 710-717.
- Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che *et al.* 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297-302.
- Seaton, G., C. S. Haley, S. A. Knott, M. Kearsey, and P. M. Visscher, 2002 QTL Express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics.* **18**: 339-340.
- Sensky, P. L., T. Parr, A. K. Lockley, R. G. Bardsley, P. J. Buttery *et al.* 1999 Altered calpain levels in longissimus muscle from normal pigs and heterozygotes with the ryanodine receptor mutation. *Journal of Animal Science* **77**: 2956-2964.
- Siegel S., 1956 *Nonparametric statistics for the behavioral sciences*. McGraw-Hill Kogakusha.
- Simon R. M., E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright *et al.* 2003 *Design and Analysis of DNA microarray Investigations*. Springer-Verlag.
- Sladek, R., and T. J. Hudson, 2006 Elucidating cis- and trans-regulatory variation using genetical genomics. *Trends Genet.* **22**: 245-250.
- Smyth, G. K., 2005 Limma: linear models for microarray data, pp. 397-420 in *Bioinformatics and Computational Biology Solutions using R and*

- Bioconductor*, edited by V. C. S. D. R. I. W. H. R. Gentleman. Springer, New York.
- Storey, J. D., and R. Tibshirani, 2003 Statistical significance for genomewide studies. *Proc.Natl.Acad.Sci.U.S.A* **100**: 9440-9445.
- Stranger, B. E., M. S. Forrest, A. G. Clark, M. J. Minichiello, S. Deutsch *et al.* 2005 Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**: e78.
- Stranger, B. E., M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley *et al.* 2007 Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science* **315**: 848-853.
- Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert *et al.* 2005 Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 15545-15550.
- Threadgill, D. W., 2006 Meeting report for the 4th annual Complex Trait Consortium meeting: from QTLs to systems genetics. *Mamm.Genome* **17**: 2-4.
- Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel *et al.* 2008 SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**: 247-252.
- Voy, B. H., J. A. Scharff, A. D. Perkins, A. M. Saxton, B. Borate *et al.* 2006 Extracting gene networks for low-dose radiation using graph theoretical algorithms. *Plos Computational Biology* **2**: 757-768.
- Wang, K., M. Li, and M. Bucan, 2007 Pathway-Based Approaches for Analysis of Genomewide Association Studies. *Am.J.Hum.Genet.* **81**.
- Wang, S., N. Yehya, E. E. Schadt, H. Wang, T. A. Drake *et al.* 2006 Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genet.* **2**: e15.
- Weiss, K. M., 2008 Tilting at quixotic trait loci (QTL): an evolutionary perspective on genetic causation. *Genetics* **179**: 1741-1756.
- Wit E., and J. McClure, 2004 *Statistics for Microarrays: Design, Analysis and Inference*. John Wiley & Sons, Chichester, UK.
- Wong, G. K. S., B. Liu, J. Wang, Y. Zhang, X. Yang *et al.* 2004 A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* **432**: 717-722.
- Wu, R., and M. Lin, 2006 Functional mapping - how to map and study the genetic architecture of dynamic complex traits. *Nat.Rev.Genet.* **7**: 229-237.

- Yamada, T., A. Dautry, and M. Walport, 2008 Ready for avian flu? *Nature* **454**: 162.
- Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss *et al.* 2003 Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat.Genet.* **35**: 57-64.
- Zhu, J., P. Y. Lum, J. Lamb, D. Guhathakurta, S. W. Edwards *et al.* 2004 An integrative genomics approach to the reconstruction of gene networks in segregating populations. *Cytogenet.Genome Res.* **105**: 363-374.
- Zhu, J., M. C. Wiener, C. Zhang, A. Fridman, E. Minch *et al.* 2007 Increasing the power to detect causal associations by combining genotypic and expression data in segregating populations. *Plos Computational Biology* **3**: 692-703.