The design and analysis of microarray experiments using pooled samples for the study of quantitative traits

by

Wuyan Zhang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-majors: Statistics;

Bioinformatics and Computational Biology

Program of Study Committee:
Alicia Carriquiry, Co-major Professor
Jack C. M. Dekkers , Co-major Professor
Dan Nettleton
Kenneth J. Koehler
Susan Lamont
Rantria Maitra

Iowa State University

Ames, Iowa

2007

UMI Number: 3259430

UMI Microform 3259430

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

DEDICATION

To my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

This thesis explores statistical issues associated with the analysis of data arising from microarray experiments and from experiments attempting to identify genes whose expressions are correlated with a phenotypic trait of interest. We refer to the latter as pQTL, or phenotypic quantitative trait loci.

Microarrays are a powerful technique for simultaneously measuring the expression levels of tens of thousands of genes. Microarray experiments can generate a large amount of information which can then be analyzed to identify differentially expressed genes under varying biological or environmental stimuli, to carry out genetic mapping studies, to diagnose disease states and to understand gene regulation and interaction.

While experiments involving mRNA microarrays have become common place in large-scale genomic research, their cost can still be high if a large number of individuals are included in the experiment. Further, the amount of mRNA available from each individual in the study is sometimes not sufficient to be used for analysis by microarray. In these two instances, researchers might consider pooling mRNA from several individuals and obtaining a microarray for a pooled sample rather than for each individual in the experiment. Because pooling mRNA is an effective way to reduce experimental cost or to permit inclusion of individuals even if very small amounts of their mRNA is available, much has been written about the cost of pooling in terms of information loss (Kendziorski et al., 2003; Kendziorski et al., 2005; Shih et al., 2004; Zhang and Gant 2005). It has been argued, for example, that pooling mRNA reduces the power with which we can identify differentially expressed genes (Kendziorski et al., 2003).

The main objective of this dissertation is to investigate the effect of pooling mRNA on different types of inferences drawn from genomic experimentation. We focus on three important types of experiments and on the statistical issues associated with pooling mRNA in each of them. First, we revisit the work of Kendziorski et al. (2003) and of Shih (2004), who have discussed the problem of computing the power of statistical tests when mRNA samples are pooled. We propose a more realistic model for the observed expression level in the pool that more closely represents the processes followed in the laboratory. In particular, our model explicitly recognizes that mRNA is pooled on the original scale, prior to any normalization or transformation. Further, we do not assume that pooling is perfect and construct our pools with potentially different amounts of mRNA contributed by each subject. We derive the appropriate F-test to identify differentially expressed genes under that model, and present formula to analytically calculate the power of different hypothesis testings.

One other important objective of microarray experiments is identification of genes whose transcript abundance is correlated with phenotype for a quantitative trait of interest. Here again we investigate the effect of pooling mRNA on the estimate of the correlation between gene expression and phenotype. We consider two different pooling strategies. In the first approach, individuals are grouped randomly for mRNA pooling. In the second approach, individuals are first stratified by phenotype and pools of phenotypically similar individuals are then constructed. In both cases, we assume that while we can observe the phenotype of every individual in the experiment, we measure gene expression only for pools. We find that when pools are constructed at random, the standard Pearson product-moment correlation estimate is nearly unbiased and, depending on the number of pools, compares well to the estimate obtained from individual data in terms of root mean squared error (RMSE). When individuals are stratified by phenotype, however, the usual product moment correlation estimate is biased and has large RMSE relative to the estimate obtained from individual observations. We therefore

propose a maximum likelihood (ML) estimate of the correlation between gene expression and phenotype. Further, we derive a likelihood ratio testing approach to identify genes potentially in the genetic pathway of the phenotypic trait of interest and compare the performance of the test to a permutation test and to the usual test based on the Fisher transformation of the correlation coefficient.

Finally, we discuss the effect of pooling mRNA on the accuracy with which we can identify genes whose transcriptome abundance is associated with a QTL of interest. To do so, we develop an approach to map eQTL when expression levels are available only from pooled mRNA and called this approach pQTL mapping. We argue that pQTL mapping can dramatically reduce the number of microarrays that must be obtained and also simplify the statistical analyses of the data by focusing only on expression data relevant to the trait(s) of interest. We simulate an intercross F2 population at genome, transcriptome and phenome levels to test the validity of pQTL and transcriptome mapping. We apply the widely used regression method and the composite method, which takes account of linkage disequilibrium effects, to compare the empirical power between pQTL and eQTL transcriptome mapping approaches in identifying trait-related genes.

Our findings suggest that while analyzing data obtained from individuals is always the best approach (at least in terms of power of tests and performance of estimators of correlation coefficients) it is possible to devise mRNA pooling strategies that significantly reduce the cost of experimentation with a modest loss of power or of the accuracy and precision of the estimates of correlation coefficients or locations of eQTLs. In the dissertation, we discuss the advantages and disadvantages of various pooling strategies and propose guidelines to decide between them. We include a review of the relevant literature within each chapter of this dissertation and describe the organization of the thesis at the end of this Introduction chapter. We now briefly revisit some fundamental issues associated with the design and analysis of microarray experiments and with the problem of mapping QTLs.

1.1 Microarray experiments

1.1.1 Microarray technology

Microarray technology has become a widely used tool to investigate biological problem at genetic level. It provides a snap shot of the mRNA activities of thousands of genes in a single experiment, which can be used to study the changes of the whole transcriptome over different experiment conditions or time periods. The different levels of mRNA are produced by a process called the transcription of DNA, which genes are made of. The mRNA then may be translated to active proteins, which in turn may control phenotype traits of interest, cell structure and physiologic state, or the developmental stage of the organism. Therefore, understanding the pattern change of gene mRNA levels will gain us a deeper knowledge in identifying the genes which are controlling the trait of interest.

There are four major steps in performing microarray experiments: RNA preparation, microarray construction, hybridizations and washing procedure, and image analysis (Chen et al., 2004). In the RNA preparation step, the mRNA of the samples is first extracted, and then converted to fluorescent labeled complementary RNA or cRNA. This labeled cRNA samples are more stable then mRNA samples, and are actually placed on the microarray to quantify the gene expression level. In microarray construction step, microarrays can be fabricated using a variety of technologies, including printed filter arrays, printed glass slide arrays, oligonucleotide Affymetrix arrays and Illunima beads arrays (Han et al., 2004). On each microarray, there are millions copies of known DNA sequences (probes), which can match and bind to the RNA from the RNA preparation step. During the hybridization and washing process, the RNA sample is placed on the probes spotted microarray for hybridization. The washing process keeps the complementary binding between probes and RNA, and removes the excess RNA that does not hybridize to any probe. In the last step, an image analysis process is used to measure

the strength of fluoresces dye and quantify the mRNA levels as signal intensity values.

1.1.2 The design of microarray experiments

The design of a microarray experiment involves several steps. First, researchers must decide which microarray platform will be used in the experiment. The number of different microarray technologies continues to increase. Technologies available today include high-density nylon membrane arrays, short oligonucleotide (Affymetix) arrays, bead (Illumina) arrays and others. The choice of other experimental parameters would fall under the umbrella of traditional design of experiments within a statistical framework. Experimental settings, such as the number of microarrays to include in the experiment, the allocation of mRNA samples to microarrays, the number of measurements to be obtained from each microarray, whether to pool or how to pool mRNA from different subjects, depend on the objectives of the experiment, on the number of subjects available, on the amount of mRNA available from each subject and also on cost considerations. Several approaches for the design of microarray experiments have been discussed in the literature. Some of the designs that have been proposed include optimal experimental designs (Kerr and Churchill, 2001), reference designs and direct designs (Speed and Yang, 2002), and factor and time course designs (Glonek, 2004). While it is not possible to recommend one design over others for all possible experimental scenarios, some authors have highlighted the importance of applying general principles of statistical experimental design to microarray experiments (Smyth et al., 2002). We do not discuss the problem of the design of a microarray experiment further in this dissertation except as it refers to the question of pooling mRNA.

1.1.3 Normalization and statistical analysis of microarray experiments

Expression levels measured from a set of arrays are typically "pre-processed" prior to statistical analyses. This pre-processing may include different steps depending on

the type of microarrays used in the experiment, but typically includes a normalization step. The process of normalization consists in (partially) removing any systematic biases which may arise due to variation across microarrays introduced during the experimental process. The idea is to ameliorate the differences between arrays arising from technological rather than from biological differences across the slides. The appropriate normalization approaches vary depending on the microarray platform. For example, Loess curves and quantile normalization are often used to normalize cDNA arrays (Bolstad et al., 2003); MAS 5.0 (Affymetrix, 2001) and robust multiple arrays normalization are recommended approaches for Affymetrix slides (Irizarry et al., 2003); average and rank invariant normalization apply to Illumina arrays (Illumina, 2005), and housekeeping genes and spike-in genes are used for other types of special arrays. Most of these approaches to normalization act globally (on all arrays) and are based on non-linear methods (Lu, 2004; Geller et al., 2003). Chapter 2 of this dissertation discusses in some detail the consequences of not properly accounting for the non-linearity of the normalization step when estimating the power of tests to identify differentially expressed genes using pooled mRNA samples.

Once data have been normalized, the statistical analyses of the data depend on the objectives of the study. Often, interest centers on identifying genes whose expression changes under different environmental conditions, in different biological specimens or during different development stages. Initial methods to identify differentially expressed genes consisted simply of selecting those genes whose observed expressions exhibited a two-fold change between different experiment conditions (Chen et al., 1997). This approach, while simple, has limited usefulness because it does not take into account the variability of expression within genes. At least in terms of usage, procedures based on the standard analysis of variance (ANOVA) to estimate expression differences across "treatments" (e.g., tissues, environmental conditions, time points) are most popular today (see, e.g., Kerr et al., 2000; Chen et al., 2004). Other inferential approaches, including

those based on mixed linear models and mixture models have also been proposed (Pan et al., 2001; Thomas et al., 2001). Several researchers have discussed Bayesian approaches to estimate model parameters in the standard ANOVA-type models or in models relying on more general distributional assumptions (Newton and Kendziorski, 2003; Love and Carriquiry, 2006).

One potential limitation of most statistical approaches proposed for identifying differentially expressed genes is that the statistical analyses is carried out on the normalized data, as if no uncertainties were introduced during the pre-processing step. A more convincing approach would be to jointly model the normalization and the analyses steps since inferences depend on both. Richardson and and Best (2003) had proposed a Bayesian hierarchical modeling approach that permits at least partial accounting of the uncertainties introduced during the pre-processing steps in the final estimates of gene expression differences. However, much remains to be done in that area.

1.2 Mapping quantitative trait loci

A quantitative trait is a continuous trait (such as height or weight), which is controlled by the polymorphism of DNA sequences. Correspondingly, the region of DNA that controls the quantitative trait is called a quantitative trait locus (QTL). The number, effect and location of QTLs explaining variation in the phenotypic trait enable a deeper understanding of the genetic architecture of a trait. QTL mapping is also fundamental to identifying candidate genes underlying a trait. Once a region of DNA is identified as contributing to a phenotype, that region can be sequenced. The DNA sequence of any gene in this region can then be compared to a database of DNA for genes whose function is already known.

QTL mapping, as the name suggests, is the approach by which we attempt to locate a QTL on a chromosome. To do so, we make use of genetic markers with known location

and then test the strength of the association between the markers and the QTL by considering a number of putative locations for the QTL. This approach then serves to identify a particular region of the genome that is likely to contain a gene associated with the trait being assayed or measured. To map QTLs, we first need to develop a resource population, in which genetic marker locus and putative QTL are in linkage disequilibrium. Then, Using the polymorphism of the genetic markers and the variation of the qualitative trait, we fit a statistical model under a set of possible QTL locations to identify the most likely QTL location.

There are two commonly used QTL mapping approaches: single marker analysis and regression interval mapping. For single marker analysis, the association between marker genotype and phenotypic trait is separately evaluated at each marker location. One limitation of this approach is that the effect and location of QTL are confounded in the single marker analysis. The regression interval mapping procedure is based on the work of Haley and Knott (1992). A regression equation is fitted for the effect of a hypothetical QTL at the position of each marker locus and at regular intervals between the marker loci. Then, the resulting regression coefficient(s) are evaluated using a likelihood ratio statistic (LRS) which measures the significance of the coefficient(s). The regression interval mapping approach can not only separately estimate the QTL effect and location but also fits coefficients for both additive and dominance effects for the population.

A trait is often affected by more than one quantitative trait locus. Quantitative trait loci other than the one being mapped are sometimes called "background" loci. These background QTL have two effects. Those which are not linked to the QTL being mapped behave like additional environmental effects (or noise) and reduce the significance of any association. Those which are linked to the QTL being mapped can bias the estimated effect and location of that QTL. Zeng (1994) proposed a composite interval mapping approach to avoid the bias that may result due to the linked QTLs located outside the QTL interval. Composite interval mapping can be viewed as an extension of regression inter-

val mapping, and consists in estimating regression coefficients for a target QTL in one interval while simultaneously estimating partial regression coefficients for "background markers" to account for variance caused by non-target QTLs. Therefore, the composite interval mapping approach is more appropriate when we believe that multiple QTLs may be present in the chromosomal region of interest.

1.3 eQTL transcriptome mapping

The expression of a gene can be viewed as a quantitative trait, and it is controlled by the polymorphisms in that or other genes. Therefore, phenotypic QTL mapping can be applied in gene expression data to identify the QTL which control gene expression levels. Jansen and Nap (2001) proposed integrating global gene expression analysis with QTL mapping in a multi-factorial manner, to allow simultaneous analysis of multiple QTL. In this approach, here called eQTL transcriptome mapping, subjects in a QTL mapping population (e.g. an F2 cross) are individually evaluated for global gene expression and genotyped for markers across the genome. Then, standard QTL mapping methods are used to identify QTL (expression QTL, eQTL) that control variation in the level of expression of individual genes by considering expression of a given gene as a quantitative phenotype. This analysis identifies regions of the genome that harbor genes that control transcription levels of a gene or genes. eQTL transcriptome mapping bridges the gap between genome sequence variation and phenotypic variation by the analysis of transcriptome RNA variation.

When an eQTL encompasses the physical location of the genes for that transcript, it is likely that the causative genetic variation resides within the gene itself (i.e., the transcript is being regulated in cis, and the corresponding QTL is called a cis-QTL). On the other hand, if an eQTL does not encompass the physical location of the gene for that transcript, the transcript is trans-regulated and the corresponding eQTL is called a trans-

QTL. Jansen (2003) provides hypothetical results of an eQTL mapping experiment to explain the relationship between cis and trans eQTLs (Figure 1.1, adapted from Jansen, R.C. 2003 Nature Reviews Genetics). The cis and trans eQTLs either up-regulate (red) or down-regulate (green) the expression level of the corresponding transcripts.

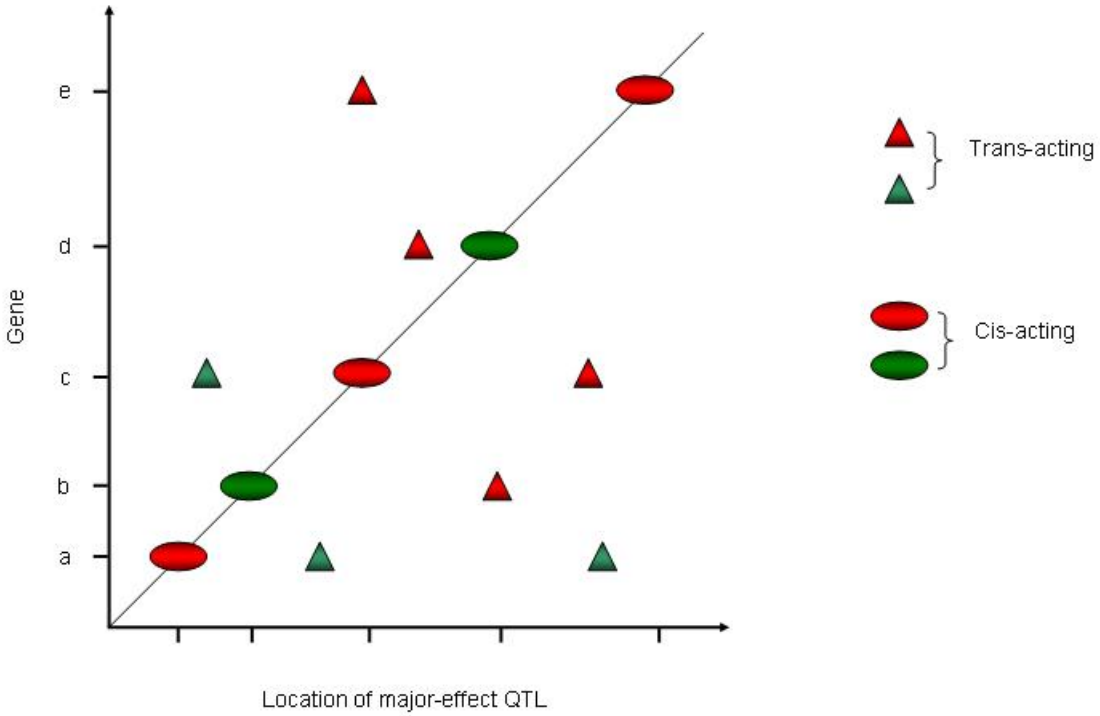

Figure 1.1 The hypothetical relationships between cis eQTL and trans eQTL in eQTL transcriptome mapping

As confirmed in several experiments (Brem et al., 2002; Pomp et al., 2004; Yvert et al., 2004; Gibson and Weir 2005), most trait-associated gene expression differences are trans-acting and not caused by a polymorphism in the gene itself but by polymorphism at an eQTL in other parts of the genome. However, the effects of trans-eQTL are generally much smaller than those of cis-eQTL (Schadt et al., 2003). Differentially expressed genes that are associated with cis-acting eQTLs are important candidate genes for the eQTL

(Doss et al., 2005).

eQTL transcriptome mapping is an effective approach to understand the genetic regulation of a phenotypic trait of interest. However, it is relatively expensive due to the large number of individual microarrays that must be obtained to guarantee sufficient statistical power. In addition, the eQTL transcriptome mapping can point not only to the differentially expressed genes and to the eQTL associated with the trait of interest but also to other eQTLs not related to the phenotypic trait of interest. Thus, statistical methods that permit identifying only genes and eQTLs of interest are needed (Schadt et al., 2003, Drake et al., 2006).

1.4 Thesis organization

This dissertation is organized in the form of three manuscripts, proceeded by an introduction chapter and followed by a general discussion chapter.

Chapter 2 is a manuscript that has been submitted to *Bioinformatics* for possible publication. In the chapter, we propose an improved statistical model and analysis method to identify differentially expressed genes in designs that include mRNA pooling. Several studies have investigated the impact of pooling mRNA on inferences about gene expression, but have typically modeled the process of pooling as if it occurred on the transformed scale (Kendziorski et al., 2003; Kendziorski et al., 2005; Shih et al., 2004; Zhang and Gant 2005), which is unrealistic. We build a statistical model for observed expression in the pool by assuming that mRNA samples are pooled on the original scale. Further, the model takes into account the additional variability that may arise when pools do not include exactly the same mRNA amount for all individuals. We develop the appropriate F statistic to test for differentially expressed genes, and present formulae to calculate the power of various statistical tests under different strategies for pooling mRNA.

Chapter 3 is a manuscript to be submitted, in which we develop a maximum likelihood estimator (MLE) to estimate the true correlation between gene expression levels and trait phenotype in microarray experiments with mRNA pooling. We consider two pooling strategies (at random, or stratifying individuals by phenotype) and evaluate the performance of the standard Pearson product-moment correlation estimate and of the MLE proposed where relative to their performance when microarrays are obtained for each sample individual. Via simulation studies, we compare the Pearson correlation coefficients and MLE in terms of bias and precision for individual, random and stratified pool designs. We also apply the MLE in a likelihood ratio test to determine whether gene expression level is truly correlated with phenotype, and compare the empirical power of the likelihood ratio test to a permutation test and the test based on Fisher's a transformation under both pooling strategies.

Chapter 4 consists of the third manuscript to be submitted. In this Chapter we focus on pQTL transcriptome mapping and evaluate its power relative to the power that can be achieved with eQTL transcriptome mapping for identifying genes and pathways that contribute to variation in the phenotypic trait. eQTL transcriptome mapping is an approach that combines QTL mapping and microarray technology, and has been shown to be promising in dissecting gene regulation networks (Brem et al., 2002). However, the high cost of microarrays tends to limit the power and application of eQTL transcriptome mapping. The pQTL mapping approach is based on eQTL transcriptome mapping and pools of mRNA samples. To carry out pQTL mapping, individuals are first stratified according to the QTL genotype. Individuals in each genotypic group are then randomly allocated to pools. Gene expression levels are measured in the pools rather than individuals. Via simulation studies, we evaluate the efficiency of pQTL transcriptome mapping approach in identifying trait related genes. Further, we compare the empirical power between eQTL and pQTL transcriptome mapping approaches by the regression and composite methods.

Finally, Chapter 5 includes a brief summary of the findings in this work and a general discussion that ties the three manuscripts together. The unifying theme running across the three major chapters in the dissertation is the impact of pooling mRNA on inferences about different quantities of interest in microarray and mapping experiments. In the discussion, we integrate our findings into general recommendations for designing and analyzing microarray experiments that involve pooling mRNA samples.

1.5 References

Affymetrix (2002). Microarray Suite User's Guide, Version 5. Affymetrix, http://www.affymetix.com/support/technical/manuals.affx

Bolstad, B., Irizarry, R., Strand, M. and Speed, T. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185-193.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752-755.

Che, P., Love, T.M., Frame, B.R., Wang, K., Carriquiry, A.L. and Howell, S.H. (2006) Gene expression patterns during somatic embryo development and germination in maize Hi II callus cultures. *Plant Mol Biol*, **62(1-2)**, 1-14.

Chen, J., Delongchamp, R., Tsai, C., Huey-min, H., Sistare, F., Thompson, K.L., Desai, V.G. and Fuscoe, J.C. (2004) Analysis of variance components in gene expression data. *Bioinformatics*, **20**, 1436-1446.

Chen, Y., Dougherty, E.R. and Bitter, M.L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt*, **2**, 364-367.

Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005) Cis-acting expression quantitative trait loci in mice. *Genome Research*, **15**, 681-691.

Drake, T.A., Schadt, E.E. and Lusis, A.J. (2006) Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mamm Genome*, **17**, 466-479.

Gibson, G. and Weir, B. (2005) The quantitative genetics of transcription. *Trends Genet*, **21**, 616-623.

Geller, S., Gregg, J., Hagerman, P. and Rocke, D. (2003) Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, **19**, 1817-1823.

Haley, C.S. and Knott, S.A. (1992) A simple method for mapping quantitative trait loci in line across using flanking markers. *Heredity*, **69**, 315-324.

Han, E., Wu Y., McCarter, R., Nelson, J.F., Richardson, A. and Hilsenbeck S.G. (2004) Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *Journal of Gerontology: Biological Sciences*, **4**, 306-315.

Illumina (2005). Beadstudio User Guide, Rev B.
Illumina, http://http://www.illumina.com/support

Irizarry,R.A., Hobbs,B., Colin, F., Beazer-Barclay,Y.D., Antonellis,K., Scherf,U. and Speed,T.P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249-264.

Newton, M.A. and C.M. Kendziorski. (2003) Parametric Empirical Bayes Methods for Microarrays in The analysis of gene expression data: methods and software. Eds. G. Parmigiani, E.S. Garrett, R. Irizarry and S.L. Zeger, New York: Springer Verlag.

Jansen, R.C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet*, **4**, 145-151.

Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet*, **17**, 388-391.

Kendziorski, C.M., Zhang Y., Lan H. and Attie, A.D. (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, **4**, 465-477.

Kendziorski, C.M., Irizarry, R.A., Chen, K.S., Haag, J.D. and Gould, M.N. (2005) On the utility of pooling biological samples in microarray experiments. *PNAS*, **102**, 4252-4257.

Kerr, M.K. and Churchill, G.A. (2001) Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183-201.

Kerr, M.K., Martin, M., and Churchill, G.A. (2000) Analysis of variance for gene expression microarray data. *J Comput Biol*, **7**, 819-837.

Lu, C. (2004) Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays. *BMC Bioinformatics*, **5**, 103-109.

Pan, W., Lin, J. and Le, C. (2001) A mixture model approach to detecting differentially expressed genes with microarray data. *Technical report.* Division of Biostatistics, University of Minnesota.

Pomp, D., Allan, M.F. and Wesolowski, S.R. (2004) Quantitative genomics: exploring the genetic architecture of complex trait predisposition. *J Anim Sci*, **82** E-Suppl, E300-312.

Richardson, S. and Best, N. (2003) Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, **14**, 129-147.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297-302.

Schadt, E.E., Monks, S.A. and Friend, S.H. (2003) A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem Soc Trans*, **31**, 437-443.

Shih, J.H., Michalowska, A.M., Dobbin, K., Ye, Y., Qiu, T.H. and Green, J.E. (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics*, **20**, 3318-3325.

Smyth, G.K. Yang, Y.H. and Speed, T. (2002) Statistical issues in cDAN data analysis. *Function Genomics*, **19**, 256-278.

Speed, T.P. and Yang, Y.H. (2002) Direct versus indirect designs for cDNA microarray experiments. *Technical Report 616*, Department of Statistics, University of California, Berkeley.

Thomas, J.G., Olson, J.M., Tapscott, S.J. and Zhao, L.P. (2001) An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res*, **11**, 1227-1236.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, **35**, 57-64.

Zeng, Z.B. (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457-1468.

Zhang, S. and Gant, T.W. (2005) Effect of pooling samples on the efficiency of comparative studies using microarrays. *Bioinformatics*, **21(24)**, 4378-4383.

CHAPTER 2. POOLING mRNA IN MICROARRAY EXPERIMENTS AND ITS EFFECT ON POWER

A paper submitted to *Bioinformatics*

Wuyan Zhang, Alicia Carriquiry, Dan Nettleton, and Jack C.M. Dekkers

Abstract: Microarrays can simultaneously measure the expression levels of many genes and are widely applied to study complex biological problems at the genetic level. To contain costs, instead of obtaining a microarray on each individual, mRNA from several subjects can be first pooled and then measured with a single array. mRNA pooling is also necessary when there is not enough mRNA from each subject. Several studies have investigated the impact of pooling mRNA on inferences about gene expression, but have typically modeled the process of pooling as if it occurred in some transformed scale. This assumption is unrealistic. We propose modeling the gene expression levels in a pool as a weighted average of mRNA expression of all individuals in the pool on the original measurement scale, where the weights correspond to individual sample contributions to the pool. Based on these improved statistical models, we develop the appropriate F statistics to test for differentially expressed genes. We present formulae to calculate the power of various statistical tests under different strategies for pooling mRNA and compare resulting power estimates to those that would be obtained by following the approach proposed by Kendziorski et al. (2003). We find that the Kendziorski estimate

tends to exceed true power and that the estimate we propose, while somewhat conservative, is less biased. We argue that it is possible to design a study that includes mRNA pooling at a significantly reduced cost but with little loss of information.
KEY WORDS: mRNA pooling; Microarray; Power.

2.1 Introduction

Microarray experiments are widely used to measure the expression levels of tens of thousands of genes simultaneously under different experimental conditions or during different time periods. One of the major interests in microarray experiments is to identify genes which are differentially expressed between conditions or time periods, and enable a deeper knowledge of complex biological problems at the genetic level. However, the unit cost of microarrays continues to be high; even for a moderate number of subjects, cost can be significant. One option to control costs is to pool the mRNA of a group of individuals and then run microarrays on the pools rather than on each individual. Pooling mRNA may also be required when there is not enough mRNA from each subject to hybridize individual microarrays.

The effect and efficiency of pooling mRNA in microarray experiments have been investigated by several researchers. Kendziorski et al. (2003) showed that pooling is most advantageous when biological variability (variability across subjects) in expression level is larger than technical variability (variability introduced in the experimental process). They also derived a formula for the total number of arrays and individuals required in an experiment involving mRNA pools to obtain gene expression estimates and confidence intervals comparable to those that would be obtained when analyzing individual arrays. They concluded that by increasing the total number of individuals in the experiment, it was possible to maintain precision of estimates by pooling, while decreasing the total number of arrays. Shih et al. (2004) also discussed the impact of pooling mRNA on the

power of statistical tests. They derived an expression to carry out power calculations for a given number of arrays and individuals. Further, they used expression data obtained from mice to check the adequacy of the assumption that mRNA expression levels in the pool are close to the average expression levels of individuals in the pool. They showed that the assumption does not hold, especially when the signals are high.

Both Kendziorski et al. (2003) and Shih et al. (2004) derived their results on the transformed scale, that is, after the data were normalized and signal intensity was transformed. Thus, both studies assumed that the mRNA expression in the pool is the average expression of individual samples, and applied the assumption on the transformed scale. This assumption is however not realistic in a biological sense because in the laboratory, the mRNA is extracted from samples and then mixed to form an mRNA pool. Therefore, pooling occurs on the original scale and an assumption that holds on the transformed scale may not hold before transformation.

In this paper, we address the issue of testing for differences in gene expression across treatments. We assume that the expression level in a pool is approximately equal to the average expression of individuals in the pool on the original scale. More precisely, we assume that the expression level in a pool is a weighted average expression of individual samples on the original scale, where weights correspond to the proportional contribution of each individual to the pool. By including the weights in the average, we account for the fact that in the process of combining individual mRNA samples, the mixing proportions may not be identical and thus, that the pool may contain more mRNA from some individuals than from others.

Under the assumptions above, we derive expressions to calculate power under different treatment effect sizes, number of mRNA pools, number of individuals per mRNA pool, and number of repeated measurements per pool. We wish to understand how much power is lost by pooling mRNA. We also wish to find efficient experimental designs for pooling mRNA samples, while keeping costs down.

2.2 Methods

2.2.1 Notation and model

Microarray gene expression measurements tend to be right skewed and thus not normally distributed. Therefore, data are usually transformed and normalized before statistical analysis. The most common transformation is the log transformation (Geller et al., 2003). The transformed data can then be modeled as a linear function of treatment effects and one or more normally distributed random effects (Lu 2003, Shih et al., 2004, Han et al., 2004). The sources of variation in a microarray experiment are multiple and can be generally classified into two groups: biological variation and technical variation (Kendziorski et al., 2003 and Chen et al., 2004). Biological variation is subject-to-subject variation in gene expression and is due to subject-specific genetic or environmental factors. Technical variation arises from the errors that can potentially be introduced at each of multiple steps in a microarray experiment. These include RNA sample preparation, microarray construction, hybridization and washing procedures and signal detection methods. Here, we focus on the two major categories: biological variation and technical variation.

The expression levels of tens of thousands of genes are measured simultaneously in a microarray experiment. For simplicity of notation, a single gene is considered in the following derivation and analysis. The true gene expression level of a gene on the j^{th} individual in the i^{th} treatment is denoted m_{ij} and can be modeled on the log scale as

$$\log(m_{ij}) = \mu_i + \epsilon_{ij}, \tag{2.1}$$

$i = 1, 2, ..., T$, $j = 1, 2, ..., N$. Here, T is the number of treatments (or conditions), N is the number of individuals per treatment, μ_i is the mean gene expression level for the i^{th} treatment , and ϵ_{ij} is biological error which is assumed to be independently, identically distributed as $N(0, \sigma_b^2)$. We use σ_b^2 to denote the biological variance in gene expression.

Then, the observed gene expression level o_{ij} in log scale can be modeled as

$$\log(o_{ij}) = \mu_i + \epsilon_{ij} + \xi_{ij} = \log(m_{ij}) + \xi_{ij}, \tag{2.2}$$

where ξ_{ij} is technical error which is also assumed to be independently, identically distributed as $N(0, \sigma_t^2)$. Biological and technical errors are assumed to be independent.

Suppose now that the mRNA from n subjects from the same treatment is combined to form a pool. We use m_{ij}^p to denote true expression for a gene in the i^{th} treatment group and j^{th} pool. We assume that the true expression level of a gene in the mRNA pool is a weighted average of the true expression levels of the gene in all individuals in the same mRNA pool, so that

$$m_{ij}^p = \sum_{k=1}^{n} (w_{ijk} * m_{ijk}), \tag{2.3}$$

where

$$w_{ijk} = \frac{z_{ijk}}{z_{ij1} + z_{ij2} + ... + z_{ijn}},$$

$i = 1, 2, ..., T$, $j = 1, 2, ..., P$, $k = 1, 2, ..., n$. Here, P is the number of mRNA pools per treatment and n is the number of individuals per mRNA pool. Therefore, $P * n = N$, where N denotes the total number of individuals per treatment group in the experiment. When $n = 1$, the experiment involves no mRNA pooling (a microarray is made for each individual). The random unobservable weight w_{ijk} represents the relative contribution of individual k to pool j in treatment i. We write the weights as functions of the z_{ijk}, which denote the technical deviations from the ideal pool containing equal amounts of mRNA from each individual sample. We assume that the z_{ijk} are independently, identically distributed as $N(1, \sigma_z^2)$, where σ_z^2 denotes the pooling technical variance. If we denote the observed mRNA level in a pool by o_{ijl}^p we can then model it on the log scale as:

$$\log(o_{ijl}^p) = \log(m_{ij}^p) + \xi_{ijl}, \tag{2.4}$$

where ξ_{ijl} is technical error as defined earlier and $l = 1, 2, ...R$. Here, R is the number of replicated array measurements per pool and $R * P = A$, where A is the total number of arrays per treatment group. $P = A$ if each mRNA pool is measured only once. Note that model (2.4) for the transformed observed mRNA level in a pool is similar to model (2.2) formulated for observed mRNA in an individual array on the transformed scale.

2.2.2 Expectation and variance of $\log(m_{ij}^p)$

The distribution of $\log(m_{ij}^p)$ is analytically intractable, but simulations and goodness-of-fit testing show that it can be well approximated by a normal distribution. We will use μ_i^p and σ_b^{p2} to denote the mean and variance of this normal distribution. We can then write

$$\log(m_{ij}^p) \approx \mu_i^p + \tau_{ij} \tag{2.5}$$

and

$$\log(o_{ijl}^p) \approx \mu_i^p + \tau_{ij} + \xi_{ijl}, \tag{2.6}$$

where τ_{ij} is assumed to be independent and identically distributed $N(0, \sigma_b^{p2})$.

We are interested in testing for differences in gene expression across treatments of the form $\mu_i - \mu_j$. In this subsection, we will derive expressions for μ_i^p and σ_b^{p2} and show that $\mu_i - \mu_j = \mu_i^p - \mu_j^p$ so that the tests of interest can be conducted using data from pools.

To derive expressions for $\mu_i^p = E[\log(m_{ij}^p)]$ and $\sigma_b^{p2} = \text{Var}[\log(m_{ij}^p)]$, we expand $\log(m_{ij}^p)$ using a Taylor series to obtain

$$\log(m_{ij}^p) = \log(\nu_i^p) + \sum_{k=1}^{\infty} \frac{(-1)^{k-1}(m_{ij}^p - \nu_i^p)^k}{k(\nu_i^p)^k}, \tag{2.7}$$

where $\nu_i^p = E(m_{ij}^p)$. Then, the second order approximation to μ_i^p is given by

$$\mu_i^p = E[\log(m_{ij}p)] \approx \log(\nu_i^p) - \frac{\sigma_i^{p2}}{2\nu_i^{p2}}, \tag{2.8}$$

where σ_i^{p2} denotes the variance of m_{ij}^p. In the Appendix, we show that

$$\nu_i^p = E(m_{ij}^p) = e^{\mu_i + \sigma_b^2/2} \tag{2.9}$$

and

$$\sigma_i^{p2} = \mathrm{Var}(m_{ij}^p) = \frac{1}{n}(e^{2\mu_i + 2\sigma_b^2} - e^{2\mu_i + \sigma_b^2})(1 + n^2\sigma_w^2). \tag{2.10}$$

where σ_w^2 is the variance of weights w_{ijk}. Using these expressions, it is easily seen that the expectation of the second order term in the Taylor expansion is free of μ_i; i.e.,

$$\frac{\sigma_i^{p2}}{2\nu_i^{p2}} = \frac{1}{2n}(e^{\sigma_b^2} - 1)(1 + n^2\sigma_w^2). \tag{2.11}$$

Similarly, it can be shown that higher order terms are also free of μ_i, which facilitates the comparison of hypothesis testing results between designs that involve pooling and those that do not.

Applying the delta method, we find:

$$\sigma_i^{p2} \approx \frac{1}{n}(e^{2\mu_i + 2\sigma_b^2} - e^{2\mu_i + \sigma_b^2})(1 + \frac{n-1}{n}\sigma_z^2). \tag{2.12}$$

Therefore, substituting expressions (2.9) and (2.12) into expression (2.8), we obtain:

$$\mu_i^p = E[\log(m_{ij}^p)] \approx \mu_i + \frac{\sigma_b^2}{2} - \frac{1 + \frac{n-1}{n}\sigma_z^2}{2n}(e^{\sigma_b^2} - 1). \tag{2.13}$$

If we apply the delta method again, we get

$$\sigma_b^{p2} = \mathrm{Var}[\log(m_{ij}^p)] \approx \frac{n-1}{n^2}\sigma_z^2 + \frac{1}{n}(e^{\sigma_b^2} - 1). \tag{2.14}$$

We now note that the random effect τ_{ij} in model (2.6) is an error term attributable to biological variation in expression level and to the additional variability that is introduced when pooling mRNA samples.

2.2.3 Power in a design that includes pooling mRNA

One interesting finding is that $\mu_i^p - \mu_j^p = \mu_i - \mu_j$, because from expressions (2.7) and (2.13), we see that $\mu_i^p - \mu_i$ is a constant (free of μ_i) across all treatment groups. Therefore, the hypothesis for testing : $\mu_1^p = \mu_2^p = ... = \mu_T^p$ in the design that includes pooling is equivalent to the hypothesis that would be used for testing $\mu_1 = \mu_2 = ... = \mu_T$ in a design that involves individual microarrays. The corresponding F test statistic for the design with pooling is given by

$$F = \frac{\sum_{i=1}^{T} P(\overline{\log(o_{i..}^p)} - \overline{\log(o_{...}^p)})^2}{T-1} \left[\frac{\sum_{i=1}^{T}\sum_{j=1}^{P}(\overline{\log(o_{ij.}^p)} - \overline{\log(o_{i..}^p)})^2}{T*(P-1)} \right]^{-1}. \tag{2.15}$$

The null hypothesis of no treatment differences is rejected at significance level α if the F statistic is larger than $F_{T-1,T*(P-1),\alpha}$, where $F_{df1,df2,\alpha}$ is the $(1-\alpha)*100th$ percentile of a central F-distribution with $df1, df2$ degrees of freedom.

If the type I error is controlled at level α, power of the test is given by

$$1 - \beta = Pr(F_{T-1,T*(P-1)}(\delta^2) > F_{T-1,T*(P-1),\alpha}), \tag{2.16}$$

with noncentrality parameter δ^2, where

$$\delta^2 = \frac{P\sum_{i=1}^{T}(\mu_i^p - \overline{\mu^p})^2}{\sigma_b^{p2} + \frac{1}{R}\sigma_t^2} \approx \frac{P\sum_{i=1}^{T}(\mu_i^p - \overline{\mu^p})^2}{\frac{n-1}{n^2}\sigma_z^2 + \frac{1}{n}(e^{\sigma_b^2} - 1) + \frac{1}{R}\sigma_t^2}, \tag{2.17}$$

with P and R as defined earlier and $\overline{\mu^p}$ is the mean of μ_i^p across treatments.

For a more general test of hypothesis for a linear combination of the means: $C\mu^p = d$ where $\mu^p = (\mu_1^p, \mu_2^p, ..., \mu_T^p)^t$ with C a known matrix of constants with full rank r, power is calculated as

$$1 - \beta = p(F_{r,T*(P-1)}(\delta^2) > F_{r,T*(P-1),\alpha}), \tag{2.18}$$

where the noncentrality parameter is given by

$$\delta^2 = \frac{P}{\sigma_b^{p2} + \frac{1}{R}\sigma_t^2}(C\mu^p - d)^t(CC^t)^{-1}(C\mu^p - d). \tag{2.19}$$

2.3 Results

2.3.1 Comparing estimates of power

Expressions (2.16) and (2.18) to calculate the power of a test of hypothesis for differences between means are based on an assumption of normality and rely on Taylor approximations. To estimate the impact of these two approximations, we simulated data and calculated power numerically and using the analytical expressions derived in Section 2.2.3. We also compared power from simulation with power calculated analytically under the Kendziorski model (Kendziorski et al., 2003).

We simulated individual data under two different scenarios. For the first scenario, we fixed the number of treatment groups at three ($T = 3$) and the number of individuals per treatment group at 100 ($N = 100$). The mean expression difference between adjacent treatment groups on the log scale ($\mu_1 - \mu_2 = \mu_2 - \mu_3$) was assumed to be 0.2, 0.3, 0.4, or 0.5. Biological and technical variances were fixed at 0.75 and 0.25 ($\sigma_b^2 = 0.75, \sigma_t^2 = 0.25$), respectively. Pooled data under our model were simulated as a weighted average of five or three individuals (weight variation was $\sigma_z^2 = 0.05^2$) on the original scale. Therefore, $n = 5$ and we considered 20 pools per treatment ($P = 20$). For the second scenario, we simulated less individuals and less pools with $N = 15$, $n = 3$ and $P = 5$ while keeping all the other parameters the same. For each scenario, a one-way ANOVA model was fitted to the simulated pooled data to test whether the mean expression level was different across treatment groups. We compared the power of the tests at $\alpha = 0.05$. Results are presented in Table 2.1. Power calculated by simulation was based on 10000 replicates of the experiment. The entries in the column labeled "Analytical power" were calculated under two different models: the proposed model (expression 2.6) and the Kendziorski model (Kendziorski et al. 2003).

In both scenarios, the predicted power as computed using the approach proposed by Kendziorski et al. (2003) appears to be overly optimistic in that it consistently ex-

Table 2.1 Power of the test for treatment difference computed numerically by simulation and analytically by the proposed model and the Kendziorski model

Mean expression difference	Power calculated by simulation	Analytical power	
		Proposed model	Kendziorski model
N=100,n=5,p=20			
0.2	0.383	0.341	0.386
0.3	0.684	0.669	0.739
0.4	0.909	0.904	0.955
0.5	0.991	0.985	0.997
N=15,n=3,p=5			
0.2	0.089	0.090	0.101
0.3	0.159	0.145	0.169
0.4	0.252	0.226	0.274
0.5	0.379	0.334	0.405

ceeds power calculated from simulation. This may be because their approach does not account for the additional variance introduced in the pooling step and because they assume that mRNA samples can be pooled on the log scale directly. If we let σ_z^2 equal to zero while keep all the other parameter unchanged as in the first scenario and calculate power again by simulation, we find that the power estimates are 0.383, 0.682, 0.906 and 0.987, which are very close to the power when $\sigma_z^2 = 0.05^2$. Therefore, the additional variance introduced in the pooling step does not affect power much, and the assumption that pooling mRNA can happen on log scale is the leading factor causing the overestimation in Kendziorski model. On the other hand, the power computed using the analytical expression in Section 2.2.3 is conservative because our estimate for the variance is conservative. Therefore, true power is at least high as our predicted power.

2.3.2 The effect of repeated measurements on power

For a given set of experimental conditions, biological, technical and weight variation in the pooled data are often fixed. Therefore, the power of the test for a given set

of conditions depends on the number of pools, the number of repeated measurements per pool and the number of individuals per pool. Consider the following example: suppose that there are three treatment groups ($T = 3$) and 100 individuals per treatment ($N = 100$), and let the mean expression difference between any two adjacent treatment groups on the log scale be 0.5 ($\mu_i^p - \mu_j^p = 0.5$), which represents a 1.65 fold difference on the original scale. Suppose that total variation is equal to 1, biological variance is three times as large as technical variance ($\sigma_b^2 = 0.75$ and $\sigma_t^2 = 0.25$), and technical standard deviation in the pooling step is 5% of the standardized mean ($\sigma_z^2 = 0.05^2$). Then, for a fixed number of arrays per treatment ($A = 5, 10, 15, 20$), the effect of obtaining repeated measurements on each pool on power is shown in Figure 2.1. We computed power analytically using expression (2.16) with any R and P values that match the equation $R * P = A$. Note, however, that R and P will always have integer values in an actual experiment. Power decreases as the number of repeated measurements per sample increases for fixed numbers of individuals and arrays. Therefore, when the number of subjects is fixed and the number of arrays is limited, a more efficient strategy is to create multiple pools and measure each once rather than to create fewer pools and measure each multiple times. This is consistent with findings in Kendziorski et al. (2003). In the remainder, we assume that each pool is measured once ($R = 1, P = A$).

2.3.3 The effect of the number of mRNA pools on power

Figure 2.2 shows power that is computed using expression (2.16) when different numbers of pools are created under various mean expression differences between adjacent treatments ($\mu_i^p - \mu_j^p = 0.2, 0.3, 0.4, 0.5$). For a fixed number of individuals, the power of the test based on individual samples is always higher than when samples are pooled, as would be expected. Power increases as the number of pools increases, and it is maximized when $P = N$, i.e., when we microarray each individual. The rate at which power increases with mean expression difference is relatively high when the number of

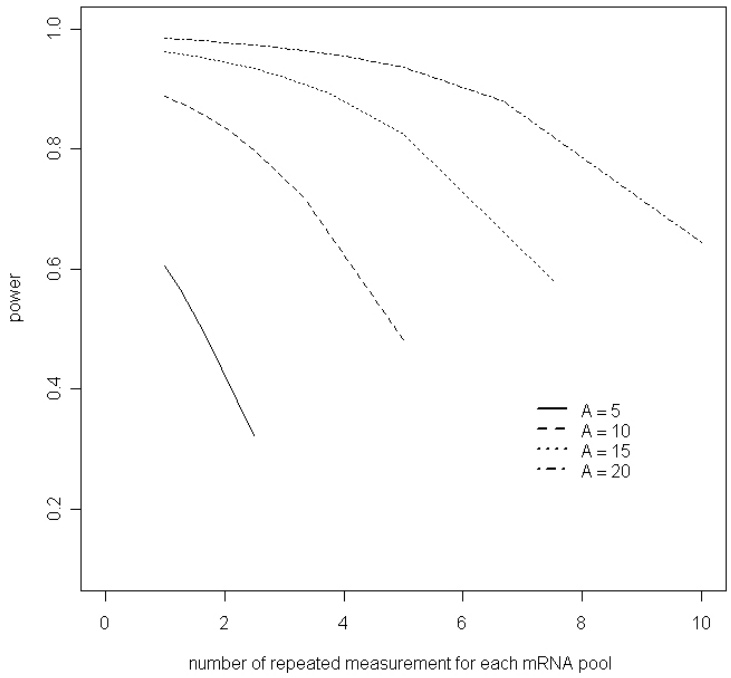

Figure 2.1 The effect of repeated measurement on power for different total numbers of arrays per treatment: $T = 3, N = 100, \mu_i^p - \mu_j^p = 0.5, \sigma_b^2 = 0.75, \sigma_t^2 = 0.25, \sigma_z^2 = 0.05^2$ and $A = 5, 10, 15, 20$.

pools is small, but relatively low when the number of pools is relatively large. When the number of pools is large enough (30 or higher, approximately) we observe no further increase in the power. For example, under $\mu_i^p - \mu_j^p = 0.4$, power increased by 0.2, 0.05 and 0.005 when P increased from 10 to 20, from 20 to 30 and from 50 to 60. The almost flat trend is especially obvious when the mean expression difference is larger ($\mu_i^p - \mu_j^p = 0.4, 0.5$). The slow or almost flat trend in the power curve makes it possible to find a pooling design with power that approaches the power that can be achieved with individual arrays and at the same time control costs. For example, when n changes from 1 to 2 (individual arrays vs. pools of two individuals per sample), power dropped from 0.9999 to 0.9993, from 0.994 to 0.982 and from 0.91 to 0.85 when $\mu_i^p - \mu_j^p = 0.5, 0.4, 0.3$. The higher the power of tests based on individual samples, the higher the number of individuals that can be pooled together without significant loss of information. For example, when $\mu_i^p - \mu_j^p = 0.5$, a design that involves forming $P = 10$ pools with $n = 10$ individuals each has 90% of the power of the design that involves no sample pooling, and yet the cost of arrays is only 10% of the cost of arraying every individual.

2.3.4 The effect of biological, technical and weight variability on power

From the results presented in Section 2.2.3, we know that power depends on σ_b^{p2} and σ_t^2, as well as on treatment mean differences and on the number of individuals and arrays. When biological variance and the variability introduced by the pooling process are reasonably small, the first term in expression (2.14) is close to 0, and the second term can be approximated by σ_b^2/n because $e^{\sigma_b^2} - 1 \approx \sigma_b^2$. Therefore, σ_b^{p2} can be well approximated by σ_b^2/n. Based on the approximation, the ratio of biological variance to technical variance must be considered. The effect of the ratio of biological to technical variance on power is shown in Figure 2.3. As would be expected, power in designs that involve pooling samples increases as the technical variance gets smaller relative to the biological variance. For example, when the mean expression difference is 0.5 and the

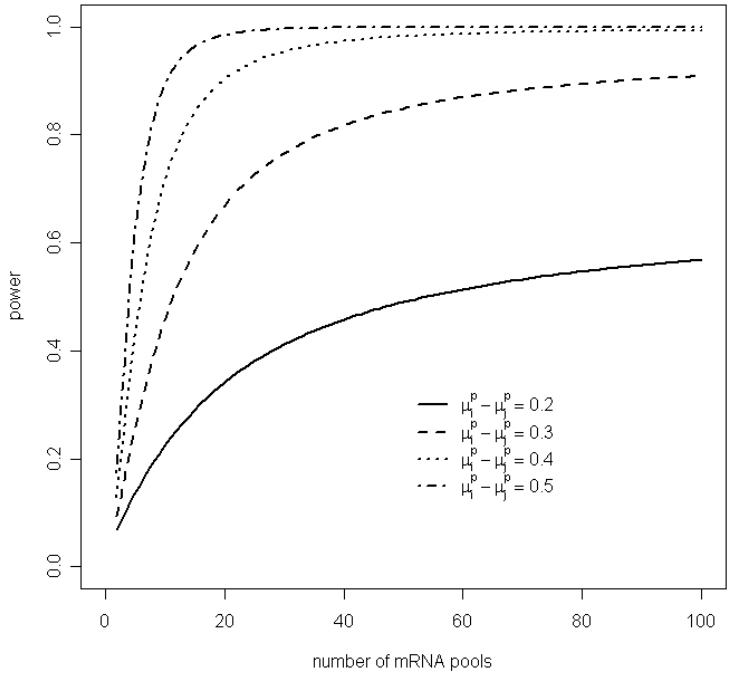

Figure 2.2 Relationship between number of pools and power for different treatment effect sizes $\mu_i^p - \mu_j^p = 0.2, 0.3, 0.4, 0.5$ and for $T = 3, N = 100, \sigma_b^2 = 0.75, \sigma_t^2 = 0.25, \sigma_z^2 = 0.05^2$.

design includes 10 pools of 10 samples each, power increases from 0.72 when $\sigma_b^2 = \sigma_t^2$ to 0.92 when $\sigma_b^2 = 4\sigma_t^2$.

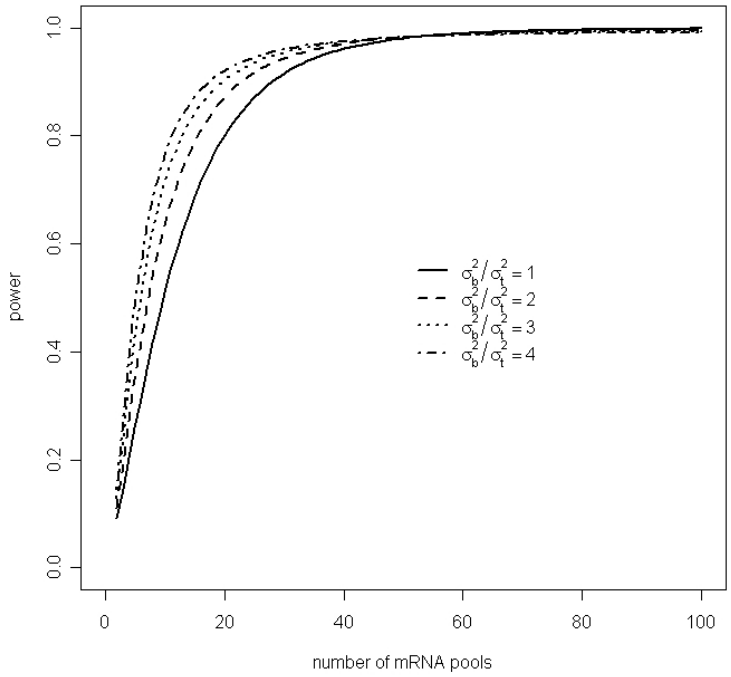

Figure 2.3 Relationship between number of pools and power for different ratios of biological to technical variance $\sigma_b^2/\sigma_t^2 = 1, 2, 3, 4$ and for $T = 3, N = 100, \mu_i^p - \mu_j^p = 0.4$, and $\sigma_z^2 = 0.05^2$.

The additional technical variation introduced in the pooling step does not appear to affect power much (Figure 2.4), even if the pooling technical variance is rather high ($\sigma_z^2 = 0.2^2$). This is because in the denominator of expression (2.17), σ_z^2 is very small compared to $e^{\sigma_b^2} - 1$ and σ_t^2. Also the effect of pooling technical variation is further decreased by the factor $\frac{n-1}{n^2}$. Therefore, the additional technical variation introduced in the pooling step is not a major factor to consider in power calculation.

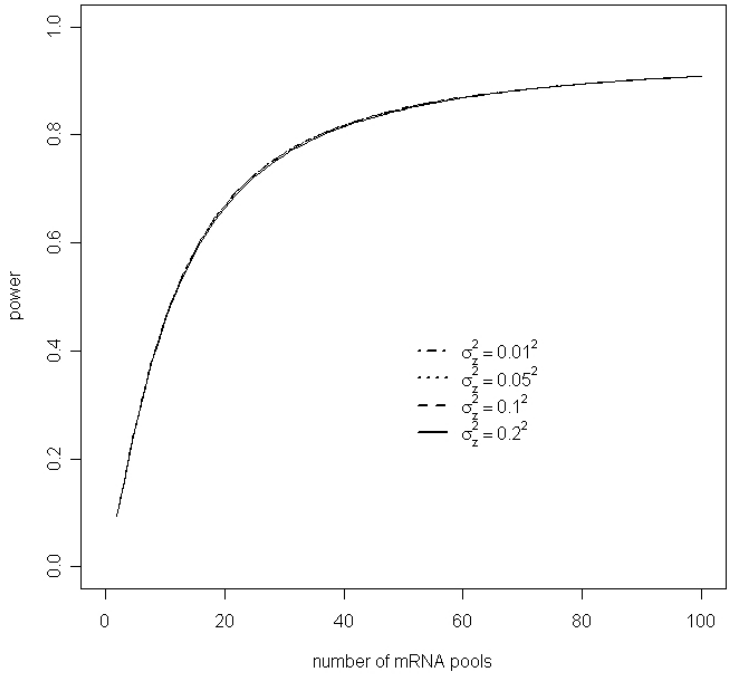

Figure 2.4 Relationship between number of pools and power for different pooling technical variance $\sigma_z^2 = 0.01^2, 0.05^2, 0.1^2, 0.2^2$ and for $T = 3, N = 100, \mu_i^p - \mu_j^p = 0.3, \sigma_b^2 = 0.75, \sigma_t^2 = 0.25$.

2.4 Discussion

Samples of mRNA from individuals are sometimes pooled in microarray experiments, either because the biological material available from each individual is not sufficient to array or to keep costs down. It is to be expected that statistical tests to detect differences in mean gene expression levels across treatments will be affected when they are based on pools of mRNA samples rather than on individual samples, since some information is bound to be lost. In particular, the power of F-tests in the usual ANOVA models is expected to decrease when the experimental design involves pooling of individual samples.

Several authors have investigated the statistical properties of F-tests based on pooled mRNA samples (Kendziorski et al. 2003 and Shih et al. 2004). One limitation in these studies is that the statistical models adopted imply that the mRNA samples are pooled on the log scale, which is unrealistic. We investigated the power of F-tests in ANOVA models when mRNA samples are pooled, but extended the models so that the pooling process is carried out on the original scale. In our formulation, mRNA pools are weighted averages of individual mRNA samples and consider the measurement error that is introduced when pooling potentially different amounts of mRNA from individuals into a pool. We argue that when pooling is assumed to occur on the log scale, the power of the tests is over-estimated and propose an approach to calculate power under the more realistic scenario of pooling on the original scale.

It is not possible to derive an analytical expression for the distribution of pooled gene expression on the log scale. Therefore, we assume that gene expression on the log scale is normally distributed. To check this assumption, we conducted simulation studies and found that, at least for the range common to the microarray data, the normality assumption appears to be reasonable. Our focus is on deriving expressions to calculate the power of F-tests to detect mean gene expression difference across treatments in

designs that involve pooling. Because the F-test is robust to modest departures from normality (Mendes and Pala, 2004), we anticipate that assuming a normal model for the gene log-expression values will not have a noticeable effect on our results. We show that the power estimated using the approach we propose here is conservative in that it tends to slightly under-estimate true power; therefore, true power is at least as high as the estimates resulting from implementing the method we propose.

As might be expected, the power of the tests depends not only on the size of the treatment effect but also on the total number of individuals and pools, the number of pools per treatment, the number of replicated measurements obtained for each pool, and the magnitude of biological and technical variability. For the technical variability, we distinguish the usual variance introduced in the various steps of microarray experiments and the variability that is introduced during the pooling process, resulting from the possibly differential contributions of individual samples to the pool.

We used simulated gene expression data to compare the power of F-tests that can be achieved when analyzing individual mRNA samples and under various pooling designs. We computed power analytically and also via simulation, and compared results to those that would be obtained by implementing the approach proposed by Kendziorski et al. (2003). We found that given a fixed number of individuals and arrays, power tends to be higher when a larger number of pools is measured once than when replicate measurements are obtained on a smaller number of pools. This holds for all values of the biological, technical and pooling variabilities considered in our study. Not surprisingly, we also found that power of tests based on individual samples is always higher than power based on pooled samples. For large enough effect sizes, however, it is possible to design an experiment that involves pooling mRNA samples that almost achieves the power that would be obtained when arraying individual samples, but at a fraction of the cost. Thus, our results suggest that under some conditions, pooling mRNA samples in microarray experiments can be a good strategy if cost is a consideration.

One of the important features of our model is that it attempts to mimic the pooling process as it happens in the lab. We assume that the mRNA pool is a weighted average of expression levels from individual mRNA samples, and let the weights be random variables with zero mean and some variance. Suppose that the mean expression level of individual j in the i^{th} pool on the original scale is μ_i. Because the log is a non-linear transformation, the weighted average of log-transformed individual means will be different than the log of the weighted averages of individual samples. It can be shown that the mean expression in the pool, denoted by μ_i^p is higher than the corresponding weighted average of the log-transformed individual means μ_i. Letting μ_i denote the weighted average of the log-transformed individual means, the difference, which we derive in Section 2.2.2 is approximated by $\mu_i^p - \mu_i \approx \sigma_b^2 - \frac{1+\frac{n-1}{n}\sigma_z^2}{2n}(e^{\sigma_b^2} - 1)$. In the range of microarray experiment data, σ_b^2 is relatively small, and $e^{\sigma_b^2} - 1$ can be approximated as σ_b^2. Therefore, the difference between μ_i^p and μ_i is approximately equal to $(1 - \frac{1+\frac{n-1}{n}\sigma_z^2}{2n})\sigma_b^2$, which is a positive, monotone function of the biological variance. Shih et al. (2004) assumed in their work that $\mu_i^p - \mu_i = 0$ and then tested this assumption using data collected in a microarray experiment on mice. They found that the number of genes with significantly different expression levels across different treatments was higher than would have been expected by chance; this effect was even stronger when expression levels were high. Also, Kendziorski et al. (2005) confirmed further that the pools and the average of individuals were not always in agreement for certain genes and suggested that modeling the pooling process on the transformed scale could be a possible reason. These results can be explained well under our model. Since we show that $\mu_i^p - \mu_i > 0$, the 95% confidence intervals for the difference between mean expression level in the pool and in individual arrays are not centered at zero. Further, since the difference between the two means can be approximated by a positive, increasing function of the biological variance, and the biological variance tends to be positively associated with gene expression levels, we expect that the shifting of the confidence intervals will be more pronounced when the

signal is stronger. In addition, confidence intervals that account for the added variance introduced in the pooling process are somewhat wider. According to our model and to the results obtained by simulation, the proportion of genes that fall outside the 95% confidence intervals discussed by Shih et al. (2004) is 0.077, 0.096, 0.101 and 0.151 when the biological variance is 0.2, 0.4, 0.6 or 0.8, respectively, and the technical variance is held constant at 0.25. These unexpectedly high proportions can be explained under our model, which accommodates pooling in the original (rather than in the log) scale.

One other interesting finding is that after log transformation and assuming of normality, the expected mean expression difference in a design that involves pooling is the same as in a design without pooling, i.e., $\mu_i^p - \mu_j^p = \mu_i - \mu_j$. Thus, a test for the hypothesis that $\mu_i^p = \mu_j^p$ is equivalent to the test $\mu_i = \mu_j$, at least when expression data have been log-transformed. This property might not hold under other transformations, however.

We have focused on power calculation under designs that pool or do not pool mRNA when testing expression differences for a single gene. In microarray experiments, tests involve tens of thousands of genes and the biological variation may differ from gene to gene. Therefore, designs that involve pooling and that permit reaching a certain power may be differ between genes due to differences in biological variation across genes. Thus, finding a single efficient design for pooling mRNA which results in the desired power for all the genes in the microarray experiment might be a challenge.

Whether to pool individuals and how to pool them to minimize the loss of information are important issues in microarray experiments. For a fixed total number of individuals and arrays, a design that includes mRNA pools always leads to smaller power than a design in which each array corresponds to an individual. Under some conditions, however, the loss of power is small, and it is possible to find a low-cost design which almost achieves the power that can be obtained when arraying each individual.

2.5 Acknowledgement

This study was funded by the Center for Integrated Animal Genomics at Iowa State University.

2.6 References

Chen J., Delongchamp R., Tsai C., Huey-min H., Sistare F., Thompson K.L., Desai V.G. and Fuscoe J.C. (2004) Analysis of variance components in gene expression data. *Bioinformatics*. **20**:1436.

Geller S., Gregg J., Hagerman P. and Rocke D. (2003) Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*. **19**:1817.

Han E., Wu Y., McCarter R., Nelson J.F., Richardson A. and Hilsenbeck S.G. (2004) Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *Journal of Gerontology: Biological Sciences*. **4**:306.

Kendziorski C.M., Zhang Y., Lan H. and Attie A.D. (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics*. **4**:465.

Kendziorski C.M., Irizarry R.A., Chen K.S., Haag J.D. and Gould M.N. (2005) On the utility of pooling biological samples in microarray experiments. *PNAS*. **102**:4252.

Lu C. (2004) Improving the scaling normalization for high-density oligonucleotide GeneChip expression microarrays. *BMC Bioinformatics*. **5**:103.

Mendes M. and Pala L. (2004) Evaluation of four tests when normality and homogeneity of variance assumptions are violated. *Journal of Applied Sciences*. **4**:38.

Shih J.H., Michalowska A.M., Dobbin K., Ye Y., Qiu T.H. and Green J.E. (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics*. **20**:3318.

2.7 Appendix

Expression (2.9) and (2.10) are derived as follows,

$$\begin{aligned}
\nu_i^p &= E(m_{ij}^p) \\
&= E[\sum_{k=1}^{n}(w_{ijk} \times m_{ijk})] \\
&= \sum_{k=1}^{n} E(w_{ijk}) \times E(m_{ijk}) \\
&= e^{\mu_i + \frac{\sigma_b^2}{2}} \sum_{k=1}^{n} E(w_{ijk}) \\
&= e^{\mu_i + \frac{\sigma_b^2}{2}},
\end{aligned}$$

$$\begin{aligned}
\sigma_i^{p2} &= \mathrm{Var}\left[m_{ij}^p\right] \\
&= \mathrm{Var}\left[\sum_{k=1}^{n}(w_{ijk}m_{ijk})\right] \\
&= E\left[\left(\sum_{k=1}^{n} w_{ijk}m_{ijk}\right)^2\right] - \left[E\left(\sum_{k=1}^{n} w_{ijk}m_{ijk}\right)\right]^2 \\
&= E\left[\sum_{k=1}^{n}(w_{ijk}m_{ijk})^2\right] + 2\sum_{k=1}^{n}\sum_{l>k}^{n} E\left(w_{ijk}m_{ijk}w_{ijl}m_{ijl}\right) \\
&\quad - \left[\sum_{k=1}^{n} E(w_{ijk})E(m_{ijk})\right]^2 \\
&= \sum_{k=1}^{n}\left[E(w_{ijk})^2 E(m_{ijk})^2\right] + 2e^{2\mu_i+\sigma_b^2}\sum_{k=1}^{n}\sum_{l>k}^{n} E(w_{ijk}w_{ijl}) \\
&\quad - \left[e^{\mu_i+\frac{\sigma_b^2}{2}}\sum_{k=1}^{n} E(w_{ijk})\right]^2 \\
&= e^{2\mu_i+2\sigma_b^2}\sum_{k=1}^{n} E\left(w_{ijk}\right)^2 + e^{2\mu_i+\sigma_b^2} E\left[\left(\sum_{k=1}^{n} w_{ijk}\right)^2 - \sum_{k=1}^{n} w_{ijk}^2\right]
\end{aligned}$$

$$
\begin{aligned}
& -e^{2\mu_i+\sigma_b^2} \\
= \;& ne^{2\mu_i+2\sigma_b^2}E(w_{ijk}^2) + e^{2\mu_i+\sigma_b^2}\left[1 - nE(w_{ijk}^2)\right] - e^{2\mu_i+\sigma_b^2} \\
= \;& nE(w_{ijk})^2\left(e^{2\mu_i+2\sigma_b^2} - e^{2\mu_i+\sigma_b^2}\right) \\
= \;& n\left[(Ew_{ijk})^2 + \sigma_w^2\right]\left(e^{2\mu_i+2\sigma_b^2} - e^{2\mu_i+\sigma_b^2}\right) \\
= \;& n(\frac{1}{n^2} + \sigma_w^2)\left(e^{2\mu_i+2\sigma_b^2} - e^{2\mu_i+\sigma_b^2}\right) \\
= \;& (\frac{1}{n} + n\sigma_w^2)\left(e^{2\mu_i+2\sigma_b^2} - e^{2\mu_i+\sigma_b^2}\right).
\end{aligned}
$$

CHAPTER 3. THE ESTIMATION OF CORRELATIONS BETWEEN PHENOTYPE AND GENE EXPRESSION IN MICROARRAY EXPERIMENTS WITH mRNA POOLING

Wuyan Zhang, Alicia Carriquiry, Dan Nettleton, and Jack C.M. Dekkers

Abstract: Microarrays are used to simultaneously measure the mRNA expression levels of thousands of genes. In such experiments, mRNA samples are sometimes pooled across individuals to reduce cost or to increase mRNA volume. We consider the problem of identifying transcripts whose abundance is correlated with phenotype for a quantitative trait of interest. We assume that the quantitative trait phenotype has been measured on all individuals but that cost considerations require us to measure mRNA expression levels in pools rather than individuals. Therefore, we propose to form disjoint pools of individuals randomly or stratified by phenotype. To assess the impact of the two pooling approaches on the accuracy with which we can estimate the correlation, we use simulated data so that the true correlation between phenotype and genotype is known. We first simulate phenotype and expression level as bivariate normally distributed variables. Then we assume that a pool's measured mRNA expression level is the average of mRNA expression levels of the individuals in a pool. We find that the Pearson correlation coefficient between a pool's trait mean and a pool's measured mRNA expression level overestimates the true correlation between phenotype and expression level when pools are stratified by phenotype. Therefore, we propose obtaining a maximum likelihood estimator (MLE) for the correlation between phenotype and expression that is less

biased. Via simulation studies, we find that this method is effective in both random and stratified pool designs. Also the MLE from stratified pool designs has higher precision than the MLE from random pool designs. Furthermore, our MLE can be used in a likelihood ratio test to determine whether gene expression level is correlated with phenotype. We show that the empirical power in the likelihood ratio test is the same as in a permutation test in both pool designs, and the power in stratified pool designs is generally higher than in random pool designs. Therefore, maximum likelihood estimation for stratified pool designs can provide more precise estimation of the correlation, and has more power to identify trait-related genes. This information is useful to investigate which genes might be involved in the genetic pathway of the phenotypic trait.

KEY WORDS: mRNA Pooling; Correlation; Power.

3.1 Introduction

Microarray experiments are widely used to measure the mRNA expression levels of thousands of genes simultaneously. This technique is a useful tool to understand complex biological problems at the genetic level. One problem of interest is to identify the genes whose transcript abundances are correlated with a phenotypic trait of interest (Caldwell et al., 2001, Booth et al., 2005, Qu and Xu, 2006 and Norholm et al., 2006), because these genes are important candidates to further investigate the genetic regulation and pathways that underly the phenotypic trait.

However, the unit cost of microarrays continues to be high; even for a moderate number of subjects, cost can be significant. Thus, instead of obtaining a microarray for each individual, several authors have proposed designs which involve pooling mRNA samples to reduce the total number of arrays needed (Kerr et al., 2003; Kendziorski et

al., 2003; Shih et al., 2004; and Zhang et al., 2006). The design with mRNA pooling consists of mixing mRNA samples from a group of individuals and then only measuring the expression level on the pools rather than on each individual. Such a design is also convenient when there is not enough mRNA to microarray each individual. In the designs with mRNA pooling, individuals can be either randomly selected to create disjoint pools (random pool design) or pools can be formed based on phenotype (stratified pool design). In the latter case, we assume that the quantitative trait phenotype has been measured on all individuals, and mRNA expression levels are only measured in the pools rather than individuals.

We investigate whether the two pooling strategies differ in terms of the bias and precision with which we can estimate the correlation between gene expression and phenotype. To do so, we use simulated data, where the true correlation coefficient is known. We first simulate phenotype and expression levels as bivariate normally distributed variables. Then we assume that a pool's mRNA expression level is the average of mRNA expression levels of all the individuals in the same pool. We calculate the widely used Pearson correlation coefficient for both pool designs. We find that in the stratified pooling design, the Pearson correlation overestimates the true correlation between phenotype and gene expression levels. Therefore, we propose an algorithm to obtain a maximum likelihood (ML) estimator of the correlation between phenotype and mRNA expression levels in both pool designs. We also derive a likelihood ratio test to identify the genes whose expression levels are correlated with phenotype. By simulation studies, we find that tge maximum likelihood approach not only provides less biased and more precise estimation in stratified pool design than that in random pool design, but also has higher power to determine trait-related pathway genes. Therefore, the stratified pool design is a preferable choice in terms of bias, precision and power. This manuscript is organized as follows. In Section 3.2, we introduce the two pooling stratifies and describe the ap-

proach for obtaining Pearson product moment and ML correlation estimates. Section 3.3 includes details about the simulation study and the results that were obtained. Finally, a discussion and some conclusions are presented in Section 3.4.

3.2 Methods

3.2.1 Notation and models for random and stratified pool designs

For simplicity of notation, a single gene is considered in the following derivation and analysis. We use y_i to denote the phenotypic measurement and x_i to denote the log scale gene expression level of a single gene for the i^{th} individual. We assume that (y_i, x_i) are jointly normally distributed with mean (μ_y, μ_x), variance (σ_y^2, σ_x^2) and correlation ρ (Wang and Nettleton, 2006), i.e.,

$$\begin{bmatrix} y_i \\ x_i \end{bmatrix} = N\left(\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{bmatrix}\right), \tag{3.1}$$

where $i = 1, ..., N$. Here N is the total number of individuals in the microarray experiment.

In microarray experiments, we cannot directly measure the true mRNA expression. Instead, we measure the observed mRNA expression level, which contains technical error. Technical error can potentially be introduced at each step of a microarray experiment, including sample preparation, array construction, hybridization and washing procedures and signal detection process (Chen et al., 2004 and Churchill, 2002). Therefore, we extend the model to account for technical error. We propose that the observed mRNA expression level be written as the sum of true mRNA expression level and technical error (ϵ_i), i.e.,

$$o_i = x_i + \epsilon_i, \tag{3.2}$$

where ϵ_i is independently, identically distributed as $N(0, \sigma_t^2)$ and independent of x_i and y_i. Here, σ_t^2 is the technical variance introduced in microarray process. Then, based expressions (3.1) and (3.2), y_i and o_i are still jointly distributed as

$$\begin{bmatrix} y_i \\ o_i \end{bmatrix} = N\left(\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 + \sigma_t^2 \end{bmatrix}\right). \tag{3.3}$$

Therefore, the correlation between y_i and o_i is can be derived as follows,

$$Cor(y_i, o_i) = \frac{\rho\sigma_y\sigma_x}{\sqrt{\sigma_y^2(\sigma_x^2 + \sigma_t^2)}} = \rho\sqrt{\frac{\sigma_x^2}{\sigma_x^2 + \sigma_t^2}}. \tag{3.4}$$

In the designs with mRNA pooling, a pool's mRNA expression level is assumed to be the average of mRNA expression levels of all the individuals in the pool. Similarly, a pool's phenotypic value is defined as the average of phenotypic values of all the individuals in the same pool. Let x_i^p and y_i^p be the mRNA expression level and phenotypic value for the i^{th} pool. Then

$$y_i^p = \frac{\sum_{j=1}^{n} y_{ij}}{n}, \tag{3.5}$$

$$x_i^p = \frac{\sum_{j=1}^{n} x_{ij}}{n}, \tag{3.6}$$

where $i = 1, 2, ..., P$ and $j = 1, 2, ..., n$. Here, x_{ij} and y_{ij} are mRNA expression level and phenotypic value of the j^{th} individual for the i^{th} pool, P is the number of pools, and n is the number of individuals per pool, so that $P * n = N$. Again, we model the observed mRNA expression level for a pool as the sum of the true mRNA expression level for a pool and technical error. Then,

$$o_i^p = x_i^p + \epsilon_i. \tag{3.7}$$

In a random pool design, each individual is randomly assigned to a pool. Therefore, based on expressions (3.1), (3.5), (3.6) and (3.7), the observed mRNA expression level and phenotypic value for a pool is still bivariate normal distributed, i.e.,

$$\begin{bmatrix} y_i^p \\ o_i^p \end{bmatrix} = N\left(\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix}, \begin{bmatrix} \sigma_y^2/n & \rho\sigma_y\sigma_x/n \\ \rho\sigma_y\sigma_x/n & \sigma_x^2/n + \sigma_t^2 \end{bmatrix}\right). \tag{3.8}$$

Therefore, the correlation between observed mRNA expression level and phenotypic value for a pool is

$$Cor(y_i^p, o_i^p) = \frac{\rho\sigma_y\sigma_x/n}{\sqrt{(\sigma_y^2/n)(\sigma_x^2/n + \sigma_t^2)}} = \rho\sqrt{\frac{\sigma_x^2}{\sigma_x^2 + n\sigma_t^2}}. \tag{3.9}$$

Notice that the presence of technical error has the effects of attenuating the correlation between y_i^p and o_i^p.

With a stratified pool design, the N individuals are pooled as follows: (1) First, order the N individuals according to the value of y_i, where the ordered y_i is denoted by $y_{(i)}$. Here, $y_{(i)}$ stands for the i^{th} smallest phenotypic value. The x_i corresponding to $y_{(i)}$ is denoted as $x_{[i]}$. We say that $x_{[i]}$ is a concomitant to $y_{(i)}$. (2) Construct P pools of size n by pooling the n samples with the smallest phenotype values, the n samples with the next smallest phenotype values, etc. Then, the average phenotype (y_i^p) and the true and observed mRNA expression levels (x_i^p and o_i^p) for the i^{th} stratified pool are:

$$x_i^p = \frac{\sum_{j=1+n(i-1)}^{i\times n} x_{[j]}}{n}, \tag{3.10}$$

$$y_i^p = \frac{\sum_{j=1+n(i-1)}^{i\times n} y_{(j)}}{n}, \tag{3.11}$$

and

$$o_i^p = \frac{\sum_{j=1+n(i-1)}^{i\times n} o_{[j]}}{n}, \tag{3.12}$$

Where $o_{[j]}$ is $x_{[j]}$ plus measurement error.

Since the pools in the stratified pool design are not created by randomly selecting n individuals from the experimental group of N, the joint distribution of (o_i^p, y_i^p) no longer follows expression (3.8). We now discuss two approaches to estimating the correlation between the true gene expression and phenotypic values.

3.2.2 Pearson product-moment correlation method

The Pearson product-moment correlation method is widely used to detect whether gene expression is correlated with the phenotype for the trait of interest in the experiments in which one microarray is obtained for each individual (Anbahagan et al., 1999, Agrawal et al., 2002, Scherf et al., 2000 and Ueda et al., 2002). Let r denote the Pearson product-moment correlation coefficient estimate between x and y. Then,

$$r = \frac{N\sum_{i=1}^{N} x_i y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{\sqrt{N\sum_{i=1}^{N} x_i^2 - (\sum_{i=1}^{N} x_i)^2}\sqrt{N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2}}. \tag{3.13}$$

When the mRNA of n individuals has been pooled, the Pearson product-moment correlation coefficient between phenotype and gene expression can be, in principle, calculated by replacing x_i and y_i with x_i^p and o_i^p. If we assume σ_x^2 and σ_t^2 are known, given the Pearson product-moment correlation estimator and the relationship between correlation coefficient based on individual and pool data (expression 3.9), we can estimate the true correlation of gene expression and phenotypic value (ρ) for the unobservable individual measurements by

$$\hat{r} = \text{sign}(r)\min\left(\left|r\sqrt{\frac{\sigma_x^2 + n\sigma_t^2}{\sigma_x^2}}\right|, 1\right). \tag{3.14}$$

To investigate whether r is a good estimator of ρ when mRNA has been pooled, we simulated gene expression and phenotypic values as follows. We initially generated 500 pairs (x_i, y_i) (N=500) from model (3.1) with variances for mRNA expression level and phenotype fixed at 1 ($\sigma_x^2 = 1$, $\sigma_y^2 = 1$) and with $\sigma_t^2 = 0$. We varied the true correlation between gene expression and phenotype between -1 and 1 ($\rho = -1.0, -0.9, ..., 1.0$). We then constructed pools of size 5 ($n = 5$) or pools of size 10 ($n = 10$) either randomly or stratifying by phenotypic value. A pool's mRNA expression level and phenotypic value are calculated as in expression (3.5) and (3.7) or expression (3.11) and (3.12). Finally, we calculated the Pearson correlation coefficients to estimate the correlation

between phenotypic value and mRNA expression in the two types of pools. The results are shown in Figure 3.1. Figure 3.1 clearly show that the Pearson product-moment correlation method gives an almost unbiased estimation of ρ in random pool design, but it overestimates the true correlation between mRNA expression levels and phenotype when the pools are constructed by grouping individuals according to phenotype. These results indicate that the Pearson product-moment correlation estimate can be severely biased when pools are stratified by phenotype. We show, in next section, however that a maximum likelihood approach results in an almost unbiased estimation of the correlation in stratified pools with similar or or even higher precision.

3.2.3 Maximum likelihood estimation

We have shown that the standard Pearson correlation estimate is biased when individuals are stratified by phenotype. Here, we propose a ML approach to estimate the correlation between gene expression and phenotype. Because a closed form maximizer of the likelihood is unavailable, we obtain the MLE of ρ ($\hat{\rho}$) numerically.

We now describe the two steps in the derivation of likelihood function (L) i.e.,

$$\begin{aligned} L(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_t^2, \rho | O^p, Y) &= f(O^p, Y | \mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_t^2, \rho) && (3.15) \\ &= f(O^p | Y, \mu_x, \mu_y, \sigma_y^2, \sigma_t^2, \rho) f(Y | \sigma_y^2, \mu_y). && (3.16) \end{aligned}$$

where $Y = (y_1, y_2, ..., y_N)^t$, $O^p = (o_1^p, o_2^p, ..., o_P^p)^t$, fs are the density functions for $O^p|Y$ and Y.

First, based on model (3.1), it is straight forward to write the marginal distribution of Y, i.e.,

$$f(Y) = \frac{\exp[-\frac{1}{2}(Y - \mu_y)^t {\Sigma_y}^{-1}(Y - \mu_y)]}{\sqrt{(2\pi)^N |\Sigma_y|}}. \quad (3.17)$$

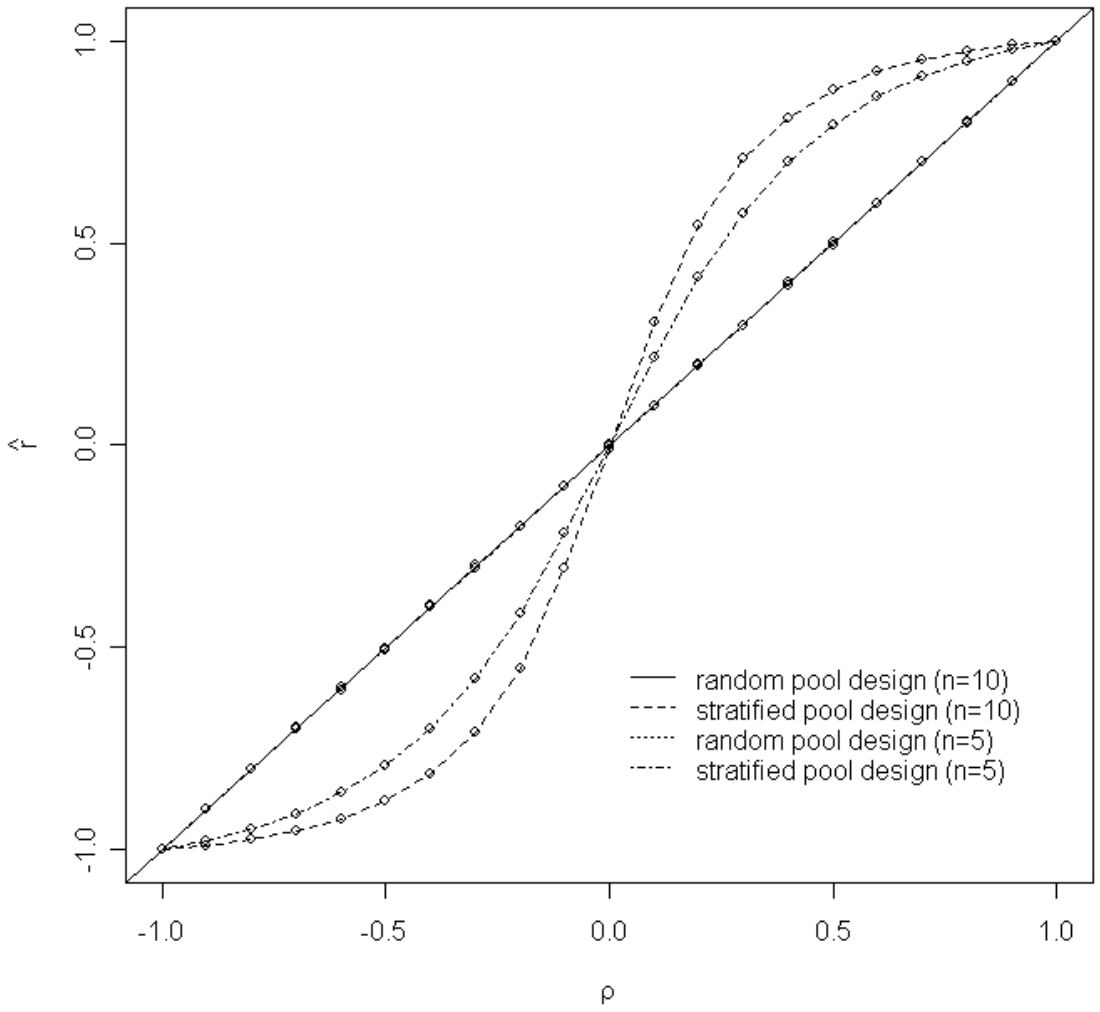

Figure 3.1 Pearson product-moment correlation coefficient estimates in the random pool design and the stratified design ($N = 500, n = 5, P = 100, \sigma_x^2 = 1, \sigma_y^2 = 1$, and $\rho = -1.0, -0.9, ..., 1.0$).

Here, Σ_y is the variance-covariance matrix of Y, and equal to $\sigma_y^2 I_{N\times N}$, where $I_{N\times N}$ is a $N \times N$ identity matrix.

The next step is to derive the distribution of O^p conditional on Y. Recall that for the i^{th} individual, (y_i, x_i) is assumed to be bivariate normally distributed. Then, conditional on phenotype, gene expression is also normally distributed and can be written as

$$x_i|y_i = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y_i - \mu_y) + \epsilon_i, \tag{3.18}$$

where ϵ_i is independent, identically distributed as $N(0, \sigma_x^2(1-\rho^2))$.

Let $X = (x_1, x_2, ..., x_N)^t$, $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_N)^t$, and $\mu_{X|Y} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y - \mu_y)$. We can rewrite expression (3.17) in matrix form, i.e.,

$$X|Y = \mu_{X|Y} + \epsilon, \tag{3.19}$$

where $\mathrm{Var}(\epsilon) = \sigma_{\mathrm{x}}^2(1-\rho^2)\mathrm{I}_{\mathrm{N}\times\mathrm{N}}$.

Let $M_{p\times N}$ denote the design matrix in a pooled mRNA design which assigns each (x_i, y_i) to one of the P pools. Then, conditional on phenotypic values, the average mRNA expression level in the pools is still normally distributed and can be written as

$$X^p|Y = M \times X|Y = M \times \mu_{X|Y} + M \times \epsilon, \tag{3.20}$$

where $X^p = (x_1^p, x_2^p, ..., x_P^p)^t$. In our design with P pools of size n, the $M_{P\times N}$ matrix has a specific format: each row of $M_{P\times N}$ contains P entries equal to $1/n$ and $N - P$ entries equal to 0. Therefore, expression (3.19) can be simplified to

$$X^p|Y = \mu_{X^p|Y^p} + \epsilon^p, \tag{3.21}$$

where $\mu_{X^p|Y^p} = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(Y^p - \mu_y)$, $Y^p = (y_1^p, y_2^p, ..., y_P^p)^t$ and $\epsilon^p = (\epsilon_1^p, \epsilon_2^p, ..., \epsilon_P^p)^t$. Here, ϵ_i^p is independent, identically distributed as $N(0, \sigma_x^2(1-\rho^2)/n)$ and $\mathrm{Var}(\epsilon^{\mathrm{p}}) = \Sigma_{\mathrm{x^p}} = \sigma_{\mathrm{x}}^2(1-\rho^2)/\mathrm{nI}_{\mathrm{P}\times\mathrm{P}}$. Based on the expression (3.7) and (3.20), we can easily derive

that

$$O^p|Y = \mu_{X^p|Y^p} + \varepsilon, \tag{3.22}$$

Where ε is multivariate normally distributed with mean equal to zero and variance (Σ_{O^p}) equal to $(\sigma_t^2 + \sigma_x^2(1-\rho^2)/n)I_{P\times P}$.

Then, the likelihood function can be written as

$$L = \frac{\exp[-\frac{1}{2}(O^p - \mu_{x^p|y^p})^t {\Sigma_{o^p}}^{-1}(O^p - \mu_{x^p|y^p}) - \frac{1}{2}(Y-\mu_y)^t {\Sigma_y}^{-1}(Y-\mu_y)]}{\sqrt{(2\pi)^P|\Sigma_{o^p}|(2\pi)^N|\Sigma_y|}}, \tag{3.23}$$

or equivalently, its log

$$\begin{aligned}\log L \propto & -\frac{P}{2}\log(\frac{\sigma_x^2(1-\rho^2)}{n} + \sigma_t^2) - \frac{n\sum_{i=1}^P (o_i^p - \mu_x - \rho\frac{\sigma_x}{\sigma_y}(y_i^p - \mu_y))^2}{2(\sigma_x^2(1-\rho^2) + n\sigma_t^2)} \\ & -\frac{N}{2}\log(\sigma_y^2) - \frac{\sum_{i=1}^N (y_i - \mu_y)^2}{2\sigma_y^2}.\end{aligned} \tag{3.24}$$

Note that phenotypic values of all individuals and mRNA expression levels of pools are known, so we can calculate MLE for μ_x, μ_y and σ_y^2 as

$$\hat{\mu}_x = \frac{\sum_{i=1}^N o_i}{N} = \frac{\sum_{i=1}^P o_i^p}{P} = \overline{o}_i^p, \tag{3.25}$$

$$\hat{\mu}_y = \frac{\sum_{i=1}^N y_i}{N} = \frac{\sum_{i=1}^P y_i^p}{P} = \overline{y}_i^p, \tag{3.26}$$

$$\hat{\sigma}_y^2 = \sum_{i=1}^N y_i^2 - (\sum_{i=1}^N y_i)^2/N. \tag{3.27}$$

We also assume technical variance can be well estimated from literature or by measuring samples repeatedly to obtain $\hat{\sigma}_x^2$. It is straight forward to show that

$$L(\mu_x, \mu_y, \sigma_y^2, \sigma_x^2, \sigma_t^2, \rho|O^p, Y) \leq L(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_y^2, \hat{\sigma}_t^2, \sigma_x^2, \rho|O^p, Y)$$

for any values in the parameter spaces. Therefore, maximization of the likelihood function (L) with respect to $\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \sigma_t^2$ and ρ can be accomplished by numerically

maximizing $L(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}_y^2, \hat{\sigma}_t^2, \sigma_x^2, \rho | O^p, Y)$ with respect to σ_x^2 and ρ. According to the algorithm proposed by Nelder and Mead (1965), we numerically solved the MLE for σ_x^2 and ρ by the optim function in R.

3.2.4 A test of the hypothesis: $\rho = 0$

In addition to the estimation of the correlation between phenotype and gene expression, another important objective in microarray experiments is to identify the genes whose expression level are significantly related to the phenotype of interest. Therefore, we propose three different approaches to test if $\rho = 0$: the likelihood ratio test based on the MLE, a test based on Fisher's z-transformation for Pearson product-moment correlation estimates, and a permutation test.

We first apply the likelihood ratio test of $H_0 : \rho = 0$. The test statistic is given by $-2\log\lambda$, where

$$\lambda = \frac{L(0, \hat{\sigma}_x^2, \hat{\sigma}_y^2, \hat{\sigma}_t^2, \hat{\mu}_x, \hat{\mu}_y | O^p, Y)}{L(\hat{\rho}, \hat{\sigma}_x^2, \hat{\sigma}_y^2, \hat{\sigma}_t^2, \hat{\mu}_x, \hat{\mu}_y) | O^p, Y)}. \tag{3.28}$$

From expression (3.25), we find that

$$\begin{aligned} -2\log(\lambda) &= P\log(\frac{\hat{\sigma}_x^2(1-\hat{\rho}^2)}{n} + \hat{\sigma}_t^2) + \frac{n}{\hat{\sigma}_x^2(1-\hat{\rho}^2) + n\hat{\sigma}_t^2} \sum_{i=1}^{P} (o_i^p - \hat{\mu}_x - \hat{\rho}\frac{\hat{\sigma}_x}{\hat{\sigma}_y}(y_i^p - \hat{\mu}_y))^2 \\ &\quad - P\log(\frac{\hat{\sigma}_x^2}{n} + \hat{\sigma}_t^2) - \frac{n}{\hat{\sigma}_x^2 + n\hat{\sigma}_t^2} \sum_{i=1}^{P} (o_i^p - \hat{\mu}_x - \hat{\rho}\frac{\hat{\sigma}_x}{\hat{\sigma}_y}(y_i^p - \hat{\mu}_y))^2. \end{aligned} \tag{3.29}$$

The test rejects the null hypothesis that $\rho = 0$ at level α if $-2\log(\lambda) \geq \chi^2_{df,\alpha}$. Here, *df* is the degrees of freedom for the test and is equal to 1 in our application.

An alternative testing approach is to use Fisher's z-transformation of Pearson product-moment correlation estimates to test whether $\rho = 0$ (Dunn and Clark, 1969), i.e.,

$$Z_r = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right). \tag{3.30}$$

Here, r is the Pearson product moment correlation estimate. When $\rho = 0$, Z_r is approximately normally distributed as $N(0, \frac{1}{P-3})$ in the random pool design, where P is the number of pools as defined before.

Besides the two parametric hypothesis tests, we can also consider a nonparametric permutation test to identify the genes which are truly related to phenotype. In this test, we break the pairing relationship between the mean phenotype of a pool and the gene expression of the same pool by randomly shuffling the gene expression of pools while keeping the mean phenotypes unchanged. Then, we calculate Pearson correlation estimate for each permutation and form the sampling distribution of the estimates. Based on the sampling distribution of Pearson correlation estimates, we calculate the p value of the Pearson product-moment correlation for the un-permutated data as

$$p = \frac{1}{M} \sum_{k=1}^{M} 1(|r| \leq |r_k|), \tag{3.31}$$

where r_k is the correlation for the k^{th} of M permutations. Rather than computing r_k for all $P!$ permutation, we obtain a monte carlo approximation to the permutation p value by considering M=1000 randomly selected permutations.

3.3 Results

3.3.1 Pearson product-moment and ML approaches to estimating ρ in the absence of technical error.

We simulated true mRNA expression levels and phenotypes (x_i and y_i) for 100 individuals ($N = 100$) according to expression (3.1) with $\sigma_x^2 = 4$, $\sigma_y^2 = 1$, $\sigma_t^2 = 0$ and $\rho = 0.1, 0.2, ..., 0.9$. Pools were created by either randomly allocating individuals to pools or by first stratifying individuals by phenotypic value. In all cases, pools consisted of 10 individuals ($n = 2$ and $P = 50$). The expression levels and phenotypic values were calculated as the average of all individuals in the same pool.

Table 3.1 Comparison of Pearson correlation coefficients and maximum likelihood estimators in individual, random and stratified pool designs ($N = 100, n = 2, P = 50, \sigma_x^2 = 4, \sigma_y^2 = 1$, and $\sigma_t^2 = 0$). Results are shown as mean (standard deviation) from 1000 replicates.

True correlation	Individual design		Random pool design		Stratified pool design	
	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$
0	0.006 (0.102)	0.006 (0.103)	0.005 (0.143)	0.005 (0.148)	0.009 (0.145)	0.006 (0.104)
0.1	0.094 (0.105)	0.094 (0.106)	0.096 (0.140)	0.098 (0.144)	0.132 (0.148)	0.096 (0.109)
0.2	0.199 (0.094)	0.201 (0.095)	0.199 (0.138)	0.202 (0.140)	0.276 (0.127)	0.202 (0.096)
0.3	0.301 (0.089)	0.303 (0.090)	0.300 (0.130)	0.305 (0.131)	0.409 (0.115)	0.308 (0.093)
0.4	0.398 (0.086)	0.401 (0.086)	0.393 (0.123)	0.400 (0.122)	0.524 (0.103)	0.405 (0.091)
0.5	0.498 (0.077)	0.501 (0.077)	0.497 (0.107)	0.505 (0.105)	0.632 (0.083)	0.506 (0.081)
0.6	0.599 (0.065)	0.600 (0.065)	0.591 (0.096)	0.600 (0.092)	0.728 (0.063)	0.606 (0.070)
0.7	0.699 (0.052)	0.701 (0.052)	0.693 (0.077)	0.702 (0.071)	0.812 (0.046)	0.705 (0.058)
0.8	0.799 (0.036)	0.800 (0.036)	0.798 (0.053)	0.804 (0.047)	0.883 (0.027)	0.803 (0.040)
0.9	0.898 (0.019)	0.901 (0.019)	0.898 (0.027)	0.901 (0.023)	0.946 (0.013)	0.902 (0.022)

We calculated the standard Pearson correlation coefficient and the proposed ML estimate of the correlation coefficient using data simulated under three different conditions: when mRNA samples were analyzed individually (individual design), when mRNA pools were constructed at random (random pool design) and when individuals were first stratified on the basis of phenotypic value and the mRNA of "similar" individuals was then pooled (stratified pool design).

When microarrays were performed on each individual sample, both approaches for estimating the correlation between phenotype and gene expression resulted in nearly unbiased estimates with similar precision (Table 3.1). Similarly, when pools were constructed by randomly grouping individuals, the two approaches also resulted in nearly unbiased estimates with similar precision. However, the precision of correlation estimates was substantially less for the random pool design than the individual design, as we expected. This is because we have fewer observations to estimate the correlation when mRNA samples are pooled.

For the stratified design, the standard Pearson correlation coefficient estimate was biased and tended to over-estimate the absolute magnitude of the true correlation. The MLE on the other hand, even though slightly biased, appeared to estimate the true correlation much more closely. The standard deviation of the MLE was much lower for the stratified pool design than for the random pool design, but still higher than for the individual design.

Results in Table 3.1 were for relatively large numbers of individuals and larger numbers of individuals in each pool (N and n). We now investigate the behavior of the Pearson correlation estimates and the MLE further in small samples (Table 3.2). By comparing Tables 3.1 and 3.2, the behavior of the two correlation estimates is similar regardless of the sample size (N). However, when N and P are relatively small, the

two correlation estimators tended to be a little more biased and the precision also decreased as N and P decreased. As expected, the Pearson correlation estimator and ML estimator perform better when samples are large.

3.3.2 Pearson product-moment ML approaches to estimating ρ in the presence of technical error.

Consider now the more realistic scenario where technical error is not negligible. We again evaluated the performance (in terms of bias and precision) of the two approaches for estimating the correlation between phenotype and gene expression using individual mRNA samples and under the two pooling strategies. We simulated phenotypic and gene expression values as described in Section 3.3.1, but now added a technical error distributed as an iid normal random variable with mean 0 and variance $\sigma_t^2 = 1$.

In order to estimate the association between phenotype and gene expression using the Pearson product-moment approach, we proceeded in two steps. We first estimated the correlation coefficient using the observed gene expression and the observed phenotypic value in the pools (o_i^p and y_i^p). We then used expression (3.9) to obtain an estimate of ρ, the correlation between phenotype and true gene expression at the individual level by assuming the σ_x^2 and σ_t^2 are known. Results are presented in Table 3.3. Results for the individual design were similar to those obtained when technical error was not taken into account in the model. Both estimators were essentially unbiased and had similar standard deviation. For the random pool design, both approaches also tended to give an unbiased estimate with similar precision. For the stratified pool design, however, the MLE appeared to outperform the standard Pearson correlation coefficient. The empirical bias of the MLE was much smaller than that of the Pearson estimate for all values of the true correlation. Surprisingly, the standard deviation was not uniformly

Table 3.2 Comparison of Pearson correlation coefficients and maximum likelihood estimators in individual, random and stratified pool designs ($N = 20, n = 2, P = 10, \sigma_x^2 = 4, \sigma_y^2 = 1$, and $\sigma_t^2 = 0$). The results are shown as mean (standard deviation) calculated from 1000 replicates.

True correlation	Individual design		Random pool design		Stratified pool design	
	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$
0	0.000 (0.227)	0.000 (0.236)	0.003 (0.335)	0.003 (0.370)	-0.001 (0.333)	-0.001 (0.264)
0.1	0.093 (0.231)	0.097 (0.240)	0.104 (0.325)	0.105 (0.361)	0.129 (0.329)	0.103 (0.264)
0.2	0.201 (0.226)	0.208 (0.234)	0.193 (0.342)	0.205 (0.358)	0.280 (0.326)	0.225 (0.263)
0.3	0.291 (0.216)	0.301 (0.222)	0.290 (0.311)	0.287 (0.337)	0.394 (0.294)	0.320 (0.248)
0.4	0.391 (0.194)	0.403 (0.198)	0.387 (0.294)	0.400 (0.312)	0.517 (0.243)	0.427 (0.218)
0.5	0.490 (0.183)	0.500 (0.183)	0.480 (0.269)	0.507 (0.278)	0.620 (0.209)	0.522 (0.198)
0.6	0.590 (0.159)	0.602 (0.158)	0.578 (0.241)	0.607 (0.242)	0.721 (0.167)	0.625 (0.173)
0.7	0.688 (0.129)	0.695 (0.127)	0.670 (0.198)	0.703 (0.187)	0.802 (0.125)	0.715 (0.141)
0.8	0.795 (0.087)	0.804 (0.085)	0.782 (0.145)	0.809 (0.125)	0.887 (0.069)	0.822 (0.092)
0.9	0.896 (0.048)	0.900 (0.046)	0.892 (0.078)	0.909 (0.060)	0.947 (0.033)	0.910 (0.051)

smaller for the MLE. Because the bias of the standard Pearson estimate was so much larger than that associated to the MLE, however, the mean squared error of the MLE is below that of the Pearson estimate for all values of ρ. Notice also that when ρ is high (0.9 and above), Pearson product-moment estimate can be outside of the parameter space. This can occur because the estimator of ρ obtained from expression (3.9) is not bounded and therefore must use expression (3.14) as the estimator of ρ. The MLE, by construction, will always be inside the parameter space. When comparing Table 3.1 and 3.3, the standard deviation for the designs without technical error was always lower than that for the designs with technical error, regardless which of the three designs was applied.

In the presence of technical error, the behavior of the two estimates was again evaluated for the small sample size scenario (Table 3.4). The behavior of the two estimators in the presence of technical error is the same as it is in the absence of technical error. We again observe that in the individual design and in the random design, both approaches results in a slightly biased estimates with similar precision. When pools are stratified by phenotype, the ML approach is superior to the Pearson product-moment correlation approach in terms of bias and standard deviation.

3.3.3 A test of the hypothesis: $\rho = 0$

A test of the hypothesis that the correlation between phenotype and expression level is zero can be carried out using a likelihood ratio test, Fisher's z-transformation or a permutation test as described in Section 3.2.4. In this section, we first investigate whether the test statistic for the likelihood ratio test proposed in expression (3.30) has the anticipated asymptotic distribution. To do so, we use the simulated data described in Section 3.3.1, for the case where the true correlation between phenotype and gene expression is equal to 0. In each of the 1000 replicate data sets, we first obtain $\hat{\rho}$, the MLE of ρ, and then carry out a test of hypothesis:

Table 3.3 Comparison of Pearson correlation coefficients and maximum likelihood estimators in individual, random and stratified pool designs ($N = 100, n = 2, P = 50, \sigma_x^2 = 4, \sigma_y^2 = 1$, and $\sigma_t^2 = 1$). The results are shown as mean (standard deviation) calculated from 1000 replicates.

True correlation	Individual design		Random pool design		Stratified pool design	
	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$
0	0.001 (0.114)	0.000 (0.115)	0.002 (0.177)	0.002 (0.184)	-0.002 (0.175)	-0.002 (0.130)
0.1	0.105 (0.113)	0.106 (0.113)	0.108 (0.168)	0.099 (0.174)	0.135 (0.170)	0.099 (0.127)
0.2	0.203 (0.108)	0.203 (0.108)	0.202 (0.165)	0.207 (0.169)	0.281 (0.163)	0.209 (0.126)
0.3	0.297 (0.104)	0.300 (0.104)	0.297 (0.161)	0.301 (0.164)	0.405 (0.155)	0.308 (0.125)
0.4	0.399 (0.098)	0.402 (0.096)	0.398 (0.156)	0.409 (0.156)	0.524 (0.139)	0.404 (0.119)
0.5	0.498 (0.091)	0.494 (0.092)	0.494 (0.150)	0.505 (0.147)	0.651 (0.123)	0.509 (0.116)
0.6	0.597 (0.083)	0.601 (0.080)	0.597 (0.132)	0.610 (0.128)	0.761 (0.100)	0.610 (0.103)
0.7	0.699 (0.069)	0.702 (0.064)	0.693 (0.122)	0.708 (0.115)	0.856 (0.085)	0.712 (0.096)
0.8	0.798 (0.055)	0.802 (0.050)	0.794 (0.105)	0.811 (0.093)	0.948 (0.068)	0.815 (0.083)
0.9	0.899 (0.040)	0.903 (0.035)	0.894 (0.083)	0.909 (0.070)	0.989 (0.049)	0.912 (0.069)

Table 3.4 Comparison of Pearson correlation coefficients and maximum likelihood estimators in individual, random and stratified pool designs ($N = 20, n = 2, P = 10, \sigma_x^2 = 4, \sigma_y^2 = 1$, and $\sigma_t^2 = 1$). The results are shown as mean (standard deviation) calculated from 1000 replicates.

True correlation	Individual design		Random pool design		Stratified pool design	
	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$	$\hat{r}$	$\hat{\rho}$
0	-0.009 (0.252)	-0.008 (0.264)	0.006 (0.403)	0.009 (0.449)	-0.010 (0.412)	-0.008 (0.400)
0.1	0.092 (0.246)	0.095 (0.258)	0.106 (0.401)	0.114 (0.439)	0.118 (0.410)	0.105 (0.397)
0.2	0.192 (0.241)	0.200 (0.254)	0.202 (0.400)	0.216 (0.429)	0.275 (0.394)	0.218 (0.388)
0.3	0.287 (0.236)	0.299 (0.245)	0.298 (0.380)	0.318 (0.416)	0.415 (0.359)	0.320 (0.352)
0.4	0.390 (0.226)	0.405 (0.234)	0.380 (0.377)	0.411 (0.406)	0.525 (0.326)	0.435 (0.320)
0.5	0.491 (0.209)	0.513 (0.217)	0.462 (0.361)	0.506 (0.401)	0.654 (0.292)	0.550 (0.287)
0.6	0.588 (0.193)	0.609 (0.191)	0.573 (0.338)	0.625 (0.373)	0.751 (0.252)	0.676 (0.253)
0.7	0.695 (0.161)	0.711 (0.154)	0.670 (0.313)	0.723 (0.325)	0.865 (0.204)	0.770 (0.204)
0.8	0.789 (0.131)	0.813 (0.121)	0.769 (0.271)	0.813 (0.267)	0.930 (0.167)	0.845 (0.160)
0.9	0.895 (0.101)	0.915 (0.090)	0.888 (0.212)	0.913 (0.200)	0.989 (0.119)	0.920 (0.110)

$$H_0 : \rho = 0 \; H_a : \rho \neq 0$$

using a likelihood ratio testing approach.

The type I errors at different significant levels under $H_0 : \rho = 0$ (over the 10000 replicates) for the three designs are shown in Table 3.5. The distribution of the test statistics are almost uniformly distributed, which suggested that the test statistic proposed in Section 3.2.4 can in fact be used to draw inference about the true linear association between phenotype and gene expression. Besides the likelihood ratio test, we also studied the behavior of the test based on Fisher's z-transformation (Table 3.5). The results also justified the validity of this test in three different designs.

Furthermore, we compared the performance of the permutation test, the likelihood ratio test, and Fisher's z-transformation test on the basis of power across the three scenarios: individual, randomly or stratified pooled samples (Table 3.6). We found that the permutation test, likelihood ratio test and Fisher's z-transformation test had almost the same power in all three different designs. The power based on individual samples is highest as would be expected given that information is lost in the process of pooling. It is interesting to notice, however, that the loss of power was less when pools are stratified. Therefore, a design that involves stratifying individuals by phenotype before pooling may sometimes be preferred because we might be able to reduce the cost of experimentation without compromising power.

3.4 Discussion

Microarray studies using mRNA samples pooled across individuals are often conducted when there is not enough individual mRNA sample to hybridize or when costs prevent obtaining a separate array for each individual. In such designs, gene expression

Table 3.5 The type I errors of the likelihood ratio test, permutation test and the test based Fisher's z-transformation in the individual, random and stratified pool designs ($N = 100, n = 2, P = 50, \sigma_x^2 = 4, \sigma_y^2 = 1,\ \sigma_t^2 = 0$, and $\rho = 0$). The results are shown as type I errors from 10000 replicates.

Significant level		0.001	0.005	0.01	0.05	0.1	0.3	0.5	0.7	0.9
Individual design	Likelihood ratio test	0.0007	0.0052	0.0105	0.0503	0.1029	0.3065	0.5033	0.7004	0.8971
	Permutation test	0.0014	0.0052	0.0121	0.0529	0.1013	0.2977	0.5014	0.7075	0.9017
	Fisher's z-transformation	0.0007	0.0051	0.0123	0.0494	0.0981	0.2962	0.5007	0.7023	0.8994
Random pool design	Likelihood ratio test	0.0019	0.0058	0.0107	0.0547	0.1104	0.3181	0.5214	0.7061	0.9019
	Permutation test	0.0010	0.0046	0.0096	0.0452	0.0989	0.2949	0.5018	0.6932	0.9022
	Fisher's z-transformation	0.0010	0.0052	0.0097	0.0516	0.0985	0.3056	0.5030	0.6994	0.9039
Stratified pool design	Likelihood ratio test	0.0015	0.0074	0.0134	0.0563	0.1081	0.3118	0.5109	0.7036	0.9007
	Permutation test	0.0013	0.0042	0.0076	0.0464	0.0977	0.2943	0.4988	0.6989	0.9029
	Fisher's z-transformation	0.0013	0.0063	0.0126	0.0509	0.0967	0.2982	0.4967	0.6991	0.9018

Table 3.6 The power comparison between the likelihood ratio test, permutation test and Fisher's z-transformation in individual, random and stratified pool designs ($N = 100, n = 2, P = 50, \sigma_x^2 = 4, \sigma_y^2 = 1$, and $\sigma_t^2 = 0$). The results are shown as the percentage of significance at 0.05 level from 1000 replicates.

ρ		0	0.1	0.2	0.3	0.4	0.5
Individual design	Likelihood ratio test	0.05	0.17	0.50	0.86	0.99	1.00
	Permutation test	0.05	0.17	0.51	0.85	0.98	1.00
	Fisher's z-transformation	0.06	0.17	0.50	0.85	0.99	1.00
Random pool design	Likelihood ratio test	0.05	0.09	0.28	0.55	0.80	0.96
	Permutation test	0.05	0.11	0.29	0.54	0.79	0.96
	Fisher's z-transformation	0.05	0.10	0.29	0.55	0.80	0.95
Stratified pool design	Likelihood ratio test	0.06	0.15	0.48	0.83	0.97	1.00
	Permutation test	0.05	0.14	0.48	0.83	0.97	1.00
	Fisher's z-transformation	0.06	0.17	0.49	0.85	0.98	1.00

is measured on the pools rather than on each individual. While pooling mRNA samples can significantly decrease the cost of the experiment, it can also create some challenges. For example, estimation of the correlation between phenotype and gene expression using the standard approach can lead to severely biased estimates depending on the value of the true underlying correlation, the strategy for constructing pools, and the relative size of the technical error variance.

We show that the widely used Pearson product-moment method for estimating the correlation works well in terms of bias and precision when mRNA pools are created randomly but not when individuals are stratified by phenotype prior to pooling. In the latter case, the Pearson correlation coefficient overestimates the magnitude of the true correlation. As Figures 3.1 suggests, the bias tends to increase as pool size (n) increases. One explanation for the bias of the Pearson correlation in stratified pool designs is that it ignores the specific structure of the pool. In the stratified pools, the ordered mean phenotypic values of pools and their concomitant gene expression levels can no longer be considered to be independent, and identically distributed.

In this manuscript, we propose a maximum likelihood approach to estimating the correlation between phenotype and gene expression in experiments that include mRNA pooling. Via simulation studies, we find this method is effective in terms of bias and precision for both random and stratified pool designs. When pools are made up of randomly selected individuals, both the Pearson and the ML approaches exhibit similar behaviors in terms of bias and precision. When individuals are stratified on the basis of phenotype, the MLE is superior to the Pearson product-moment estimate in terms of bias. Further, the standard deviation of the MLE in stratified pools is lower than the standard deviation of the MLE in random pools. Therefore, if a pooling strategy is needed due to cost constraints or limited amounts of mRNA, a design in which individuals with similar phenotypic values are pooled can be a good choice. In that case, the MLE of the correlation results in estimates which are almost unbiased and have lower variance than estimates obtained from randomly constructed pools.

Clearly, we can think of a design in which no mRNA samples are pooled as the "gold standard". We therefore compared the performance of the two estimation approaches using pooled samples to their performance when using individual samples. The standard deviation of the MLE based on individual samples is lower regardless of pooling strategy. This is to be expected given that some information is lost when pooling individual samples. In the case of the Pearson product-moment correlation approach, the standard deviation of estimates in the stratified pool design is sometimes lower than in the individual design and in the random pooling design when $|\rho|$ is large. The surprising finding can be explained by noticing that the estimates are biased upward but at the same time are bounded within $(-1, 1)$.

Besides the estimation of the correlation between phenotype and gene expression level, we also propose three different tests to identify the genes whose expression level are truly correlated with phenotype. We find that the likelihood ratio test, the permutation test and Fisher's z-transformation test are appropriate for all three designs. While

the power for detecting a non-zero correlation is always higher when the test is based on individual samples, the loss due to pooling samples is smaller when individuals are stratified by phenotype. Therefore, when pooling mRNA is required in the experiment designs, the stratified pools are again preferred in terms of the power to identify the trait related genes. In both the likelihood ratio approach and Fisher's z-transformation test, the technical variance was assumed to be well known, which is not required in the permutation test. Therefore, if the main objective is to pick up the trait related genes, not to estimate the correlation between genes and phenotype, the permutation test is more desirable than likelihood ratio test, since it does not require an estimate on technical variance.

Technical error can be introduced in the multiple steps of a microarray experiment. When technical error is relatively large (as compared to biological variability), it should be taken into account in the statistical models. In this work, we assumed the technical variance is known. Otherwise, we would not be able to separate biological variance from technical variance because we assume each sample is only measured once. In practice, researchers can use an estimate of technical variance from the literature or obtain replicate measures on the same samples. If technical error is negligible, but can not be well estimated, we can still use ML or Pearson correlation approaches in absence of technical error as in Section 3.3.1. By this means, the estimation of correlation will conservative, and the magnitude of true correlation will tend to be larger than estimated. Notice that the power to test if a gene is truly correlated with phenotype is also conservative, and the true power is at least as high as the predicted.

One assumption about mRNA pooling is that the mRNA expression level for a pool is the average of mRNA expression levels of all individuals in the pool. For each individual, we have assumed that the log scale mRNA expression level and phenotypic value can be assumed as bivariate normal distribution. Therefore, based on the pooling assumption,

a pool's mRNA expression level on the log scale should be first calculated as the log of the average of mRNA expression levels on the original scale of all individual in the same pool. However, the distribution of a pool's mRNA level on the log scale derived by this assumption is no longer normally distributed and is analytically intractable. Simulations and goodness-of-fit testing show that it can be well approximated by a normal distribution (Zhang et al. 2006). Therefore, our assumption that the mRNA expression level and phenotypic value for a pool are joint normally distributed is a reasonable choice.

Overall, the proposed ML approach is superior to the Pearson correlation method in designs involving mRNA pooling because it out forms the Pearson product-moment correlation method in the stratified pool design and at least as effective in the individual and the random pool design. Further, the ML method can be used to construct a likelihood ratio test to determine whether gene expression levels is related to phenotype. These genes play an important role in studying genetic regulation and pathways of phenotype trait. Because the ML method is relatively complex and computationally intensive, the standard Pearson correlation method may be preferred when pools are created randomly. When pools are created according to phenotypic value, the Pearson correlation method is no longer valid, and the ML approach results in estimates of ρ with less bias and low standard deviation. The ML approach can be used not only in the three designs, but also in more complex designs of QTL scenario. In such a design, pools can be created by QTL genotypes, and stratified within each QTL genotype. ML approaches can still estimate the true correlation between genotype and gene expression within each genotype and pick up the genes which are truly correlated with phenotype trait of interest.

3.5 Acknowledgement

This study was funded by the Center for Integrated Animal Genomics at Iowa State University.

3.6 References

Agrawal, D., Chen, T., Irby, R., Quackenbush, J., Chambers, A.F., Szabo, M., Cantor, A., Coppola, D. and Yeatman, T.J. (2002) Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. *J Natl Cancer Inst*, **94**, 513-521.

Anbazhagan, R., Tihan, T., Bornman, D.M., Johnston, J.C., Saltz, J.H., Weigering, A., Piantadosi, S. and Gabrielson, E. (1999) Classification of small cell lung cancer and pulmonary carcinoid by gene expression profiles. *Cancer Res*, **59**, 5119-5122.

Booth, E.O., Van Driessche, N., Zhuchenko, O., Kuspa, A. and Shaulsky, G. (2005) Microarray phenotyping in Dictyostelium reveals a regulon of chemotaxis genes. *Bioinformatics*, **21**, 4371-4377.

Caldwell, R., Sapolsky, R., Weyler, W., Maile, R.R., Causey, S.C. and Ferrari, E. (2001) Correlation between Bacillus subtilis scoC phenotype and gene expression determined using microarrays for transcriptome analysis. *J Bacteriol*, **183**, 7329-7340.

Chen, J., Delongchamp, R., Tsai, C., Huey-min, H., Sistare, F., Thompson, K.L., Desai, V.G. and Fuscoe, J.C. (2004) Analysis of variance components in gene expression data. *Bioinformatics*. **20**, 1436-1446.

Churchill, G.A. (2002) Fundamentals of experimental design for cDNA microarrays, *Nat Genet*, **32 Suppl**, 490-495.

Dunn, O,J,, Clark, V. Correlation coefficients measured on the same individuals. (1969) *Am Stat Assoc J* **64** 366-376.

Hargitai, J., Zernant, J., Somfai, G.M., Vamos, R., Farkas, A., Salacz, G. and Allikmets, R. (2005) Correlation of clinical and genetic findings in Hungarian patients with Stargardt disease. *Invest Ophthalmol Vis Sci*, **46**, 4402-4408.

Kendziorski, C.M., Zhang, Y., Lan, H. and Attie, A.D. (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics*. **4**, 465-477.

Kerr, M.K. (2003) Design considerations for efficient and effective microarray studies, *Biometrics*, **59**, 822-828.

Naylor, T.L., Greshock, J., Wang, Y., Colligon, T., Yu, Q.C., Clemmer, V., Zaks, T.Z. and Weber, B.L. (2005) High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res*, **7**, R1186-1198.

Nelder, J. A. and Mead, R. (1965) A simplex algorithm for function minimization. *Computer Journal*, **7**, 308-313.

Norholm, M.H., Nour-Eldin, H.H., Brodersen, P., Mundy, J. and Halkier, B.A. (2006) Expression of the Arabidopsis high-affinity hexose transporter STP13 correlates with programmed cell death, FEBS Lett, 580, 2381-2387. Qu, Y. and Xu, S. (2006) Quantitative trait associated microarray gene expression data analysis. *Mol Biol Evol*, **23**, 1558-1573.

Qu, Y. and Xu, S. (2006) Quantitative trait associated microarray gene expression data analysis. *Mol Biol Evol*, **23**, 1558-1573.

Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Tanabe, L., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B.,

Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O. and Weinstein, J.N. (2000) A gene expression database for the molecular pharmacology of cancer. *Nat Genet*, **24**, 236-244.

Shih, J.H., Michalowska, A.M., Dobbin, K., Ye, Y., Qiu, T.H. and Green, J.E. (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics*. **20**, 3318-3325.

Ueda, M., Terai, Y., Yamashita, Y., Kumagai, K., Ueki, K., Yamaguchi, H., Akise, D., Hung, Y.C. and Ueki, M. (2002) Correlation between vascular endothelial growth factor-C expression and invasion phenotype in cervical carcinomas. *Int J Cancer*, **98**, 335-343.

Wang, D. and Nettleton, D. (2006) Identifying genes associated with a quantitative trait or quantitative trait locus via selective transcriptional profiling. *Biometrics*, **62**, 504-524.

Zhang, W., Carriquiry, A., Nettleton, D. and Dekkers, J. (2006) Pooling mRNA in microarray experiments and its effect on power. submitted to *Bioinformatics*.

CHAPTER 4. pQTL TRANSCRIPTOME MAPPING: A METHOD TO INTEGRATE QTL MAPPING AND GENE EXPRESSION ANALYSIS TO DISCOVER THE GENETIC BASIS OF COMPLEX TRAITS

Wuyan Zhang, Alicia Carriquiry, Dan Nettleton, and Jack C.M. Dekkers

Abstract: eQTL transcriptome mapping blends the power of quantitative trait locus (QTL) mapping with gene expression analysis and enables genome-wide identification of positional candidate genes for QTL and of genes involved in metabolic pathways for the phenotypic trait. Current methods for eQTL transcriptome mapping require conducting individual microarray assays on a large number of individuals to reach adequate statistical power. This is prohibitively expensive for most labs. Therefore an alternative mapping approach (pQTL transcriptome mapping) is proposed, which can dramatically decrease the cost of the experiment while still maintaining sufficient statistical power. This approach essentially consists in implementing eQTL transcriptome mapping using gene expression measures in pooled mRNA samples obtained from a group of individuals. By pooling mRNA samples, it is possible to reduce the cost of experimentation and target the generation of expression data that are relevant to the phenotypic traits of interest. To test the validity of the pQTL transcriptome mapping concept, comprehensive data on an F2 cross population were first simulated at the genome, transcriptome, and phenome levels. Then, we assumed that gene expression levels in mRNA pools are

a weighted average of the gene expression of all individuals in the pool. These averages are taken on the original measurement scale and the weights correspond to individual sample contributions to the pool. Based on simulated data for the mRNA pools, we empirically calculated the power of pQTL transcriptome mapping approaches for finding candidate genes and trait pathway genes using regression and composite mapping methods. We found that pQTL transcriptome mapping using the standard regression method achieved statistical power comparable to the power that can be achieved by eQTL transcriptome mapping. However, when pQTL transcriptome mapping is carried out via composite interval mapping which takes into account linkage disequilibrium effects, there was significant loss of power.
KEY WORDS: mRNA pooling; transcriptome mapping; power.

4.1 Introduction

Many common human diseases or traits of economic importance in livestock are affected by multiple genes. The identification of these genes and the understanding of the underlying pathway are very important. Standard quantitative trait locus (QTL) mapping gives the first insight into the genetic architecture of an organism (Andersson, 2001). In this approach, a resource population that segregates for the traits of interest (e.g. F2) is created, and associations between trait phenotypes and genetic markers across the genome are used to identify chromosomal regions that harbor loci that contribute to genetic variation in phenotypic traits (phenome QTL, pQTL). However, the pQTL region identified in this approach can be large (20-40cM), and further, the genes and their underlying pathway remain unknown. Recently, eQTL transcriptome mapping has been proposed, which integrates the power of quantitative trait locus (QTL) mapping with gene expression analysis and allows us to investigate the genetic regulatory pathways at the mRNA level (Jansen and Nap, 2001).

Microarray technology is a rapidly developing technology, with applications in many species (Gibson and Weir, 2005). This technique permits measuring the expression level of tens of thousands of genes simultaneously under different experimental conditions or over different time periods. Therefore, a microarray experiment can be used to detect genes (DE genes) which are differentially expressed under different conditions. Jansen and Nap (2001) first proposed integrating global gene expression analysis with QTL mapping in a multi-factorial manner, to allow the analysis of multiple QTL. In this approach, here called eQTL transcriptome mapping, individuals in a QTL mapping population (e.g. an F2 cross) are individually evaluated for global gene expression and genotyped for markers across the genome. Then, standard QTL mapping methods are used to identify QTL (expression QTL, eQTL) that control variation in the level of expression of individual genes by considering expression of a given gene as a quantitative phenotype. This analysis identifies genome regions that harbor genes (eQTL) that control transcription levels of a gene or genes. eQTL transcriptome mapping bridges the gap between genome sequence variation and phenotypic variation by the analysis of transcriptome RNA variation (Figure 4.1).

We now discuss further the relationship between pQTL, eQTL and differentially expressed genes. QTL (pQTL or eQTL) are genes/genome regions that harbor one or more sequence polymorphism that affect the trait phenotypic value or gene expression level. eQTL control variation in mRNA expression level while pQTL control variation in a phenotypic trait. Differentially expressed genes are genes whose level of transcription varies in the population. They are not necessarily polymorphic. In addition, not all QTL will themselves be identified as differentially expressed genes but all will result in a structural transcript difference rather than in differences in transcript abundance (Figure 4.1).

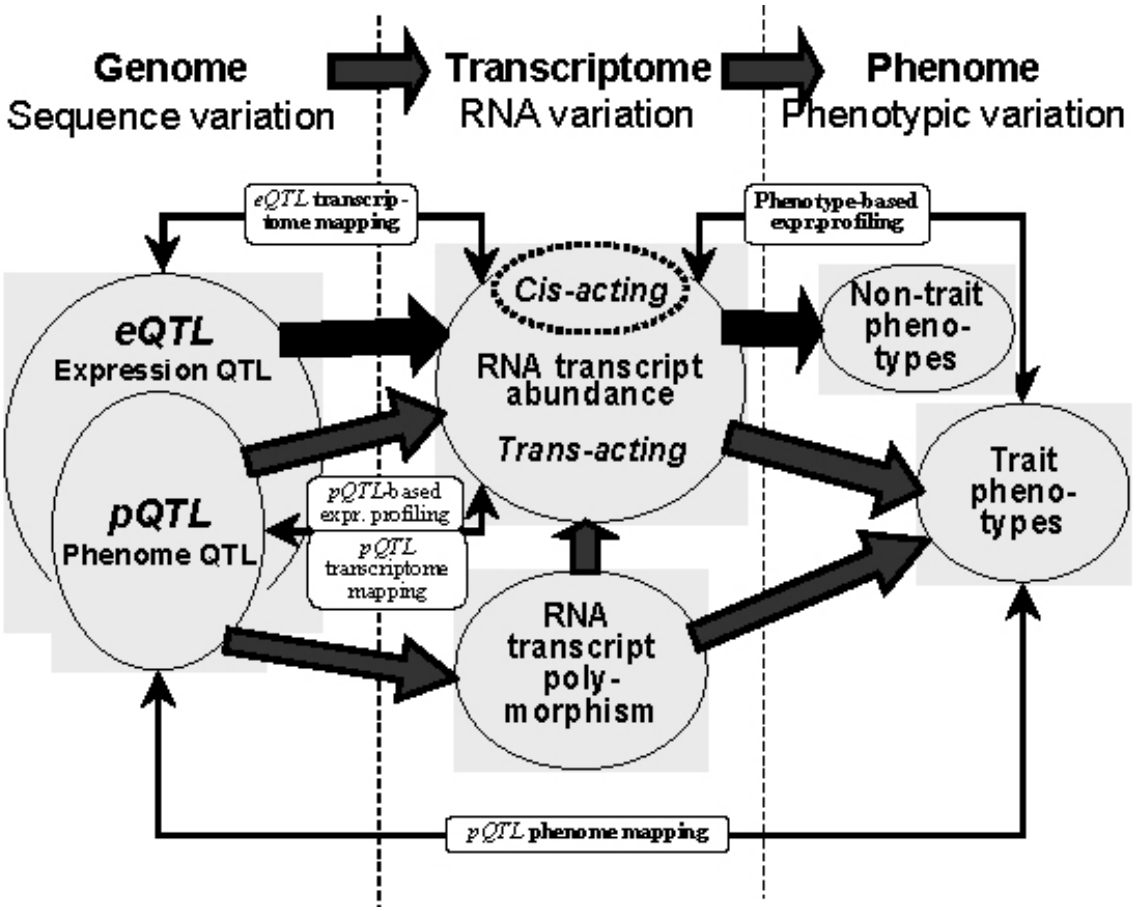

Figure 4.1 Relationships between and methods for analysis of genome, transcriptome, and phenome variation

The eQTL transcriptome mapping approach has been applied to several species (Bao et al., 2006; Brem et al., 2003; Bystrykh et al., 2005; Chesler et al., 2005; Decook et al., 2006; de Koning et al., 2005; Hubner et al., 2005, Kirst et al., 2004; Pomp et al., 2004; Schadt et al., 2003). The relationship between differentially expressed genes and their associated eQTL are summarized into two categories: cis-acting (an eQTL that affects the expression of itself), and/or trans-acting (an eQTL that affects the expression of other distant genes). Yvert et al. (2004) performed eQTL mapping in yeast and found 2294 genes whose expression phenotype was associated with eQTL. Only 25% of 2294 genes were co-localized with the corresponding structural gene (cis-acting) and 75% were transacting. Brem et al. (2002) identified 570 genes in yeast which were associated with one or more eQTL. Sixty four percent of differentially expressed genes were trans-acting and controlled by a small group of eQTL each regulating 7 to 94 genes of related function. In fact, it has been postulated (Pomp, 1999) and supported by evidence (Gibson and Weir 2005; Hubner et al., 2005; Pomp et al., 2004), that about two third of trait-associated gene expression differences are trans-acting and not caused by a polymorphism in the gene itself but by polymorphism at eQTL in other parts of the genome. Although cis-acting association is not as prevalent as trans-acting, the effects of cis-acting eQTL tend to be much stronger (Doss et al., 2005; Schadt et al., 2003). Differentially expressed genes that are associated with cis-acting eQTL have been reported as important candidate genes for diseases or phenotypic trait (Hubner et al. 2005; Pamler et al., 2005; Yaguchi et al., 2005). Therefore eQTL transcriptome mapping helps to identify candidate genes underlying QTL on a global basis and gives us a deeper insight into regulatory pathways and genetic architecture.

In contrast with the rapid development of eQTL mapping methods, little has been written about the power with which we can identify eQTL affecting certain gene expression. No formal approach for calculating power in eQTL studies has been proposed,

perhaps because estimating statistical power in eQTL studies is not straight forward (Kendziorski and Wang, 2006). Difficulties arise because power calculations must account for the power associated with mapping and with gene expression analyses. Recently, de Koning and Haley (2005) proposed methodology to calculate power in various eQTL experimental designs but ignored the challenges that arise because of the inevitable multiple testing in microarray experiments. They found that the power of most eQTL studies to detect loci involved in gene regulation is limited. As a result, they argued that most studies can be expected to fail to detect many loci with moderate effects and even some loci with major effects. What contributes to the typically low power in eQTL studies? Besides small sample sizes, linkage disequilibrium also complicates the identification of a true regulatory eQTL. Pastinen and Hudson (2004) confirmed that one SNP was falsely considered to be involved in gene regulation because of its high linkage disequilibrium with the causative regulatory eQTL.

eQTL transcriptome mapping is relatively expensive due to the large number of microarray experiments that must be performed on each individual if we wish to achieve sufficient statistical power. Also, eQTL transcriptome mapping finds not only the differentially expressed genes and the eQTL related to the trait of interest but also those not related to the trait phenotype. Further analysis is needed to pick out those genes and QTL of interest (Schadt et al., 2003, Drake et al., 2006). Therefore, we discuss an alternative approach denoted pQTL transcriptome mapping (Cabrera et al., 2006) which has the potential to locate trait-related genes at a fraction of the cost of the usual eQTL methodology. We describe the approach in the next section, and later in this manuscript use simulation studies to assess its performance. We argue that the pQTL transcriptome mapping approach can identify genes and pathways that contribute to variation in the phenotypic trait. Even though there is a loss of statistical power, the cost of a pQTL experiment can be dramatically lower than the cost of the corresponding

eQTL study, and this may make the pQTL approach feasible in most laboratories.

4.2 Methods

4.2.1 The concept of pQTL transcriptome mapping

The main objective of the pQTL transcriptome mapping approach is to identify genes whose expression is controlled by a pQTL, by identifying genes that are differentially expressed between individuals with different pQTL genotypes. This approach first pinpoints the pQTL of interest, then finds genes that are differentially expressed in individuals with alternate genotypes at the pQTL. One potential advantage of the proposed pQTL transcriptome mapping approach is that it might not require individual microarray experiments. Instead, we might be able to pool mRNA from individuals with the same pQTL genotype and then do the microarray on the pooled mRNA samples. Thus, the pQTL transcriptome mapping approach can result in substantial savings in experimental costs, a dramatic reduction in the amount of data that must be processed while also help uncover direct associations of QTL with the trait of interest.

The steps that must be followed in a pQTL transcriptome mapping experiment include the following:

1. Using standard procedures, set up a QTL resource population (e.g. F2 cross), evaluate individuals in the population for trait phenotypes and genotype the individuals for genome-wide markers. Collect RNA from tissues of interest.

2. Conduct a standard pQTL scan for the phenotypic traits of interest.

3. For a given pQTL region, group individuals by pQTL genotype based on genotype probabilities derived using marker genotypes from step 1, creating groups that have a given pQTL genotype with, e.g,. at least 90% certainty.

4. Randomly split each pQTL genotype group into P subgroups of equal size and

pool the mRNA of individuals by subgroup. Hybridize mRNA pools to expression arrays and identify genes that are differentially expressed between pQTL genotypes.

4.2.2 Simulated genome structure

To test the power of pQTL transcriptome mapping, comprehensive data on an F2 cross population were simulated at the genome, transcriptome, and phenome levels. We will first describe the genome structure of the F2 population: number of chromosomes and chromosome length, types of QTL, QTL locations and effects, types of DE genes, and location of DE genes. The problem of simulating the whole genome structure is complex and computationally intensive. To illustrate the concept of pQTL transcriptome mapping, we assume that there are two chromosomes (100 cM each), 11 evenly spaced molecular markers per chromosome (10 cM between adjacent markers) and a few QTL and DE genes which are randomly located on each chromosome. The types of QTL are summarized as follows:

pQTL: QTL that do not affect the mRNA expression level but result in a structural transcript difference of the gene that harbor the pQTL. Trait phenotype is directly affected by this type of QTL due to a polymorphism at the QTL itself.

peQTL: an eQTL that controls the mRNA expression of certain genes, which in turn affect the trait phenotype of interest. A peQTL can affect expression of the QTL gene itself (cis-acting) and/or expression of other genes (trans-acting).

npeQTL : an eQTL that controls the expression of the genes that affect non-trait phenotypes. A npeQTL can also be cis-acting and/or trans-acting.

Two types of differentially expressed genes (DE genes) are distinguished :

peGenes: DE genes whose expression level are controlled by the corresponding peQTL. peGenes can co-localize with peQTL or away from peQTL.

npeGenes: DE genes that are controlled by the corresponding peQTL npeQTL. npeGenes can also co-localize with npeQTL or away from npeQTL.

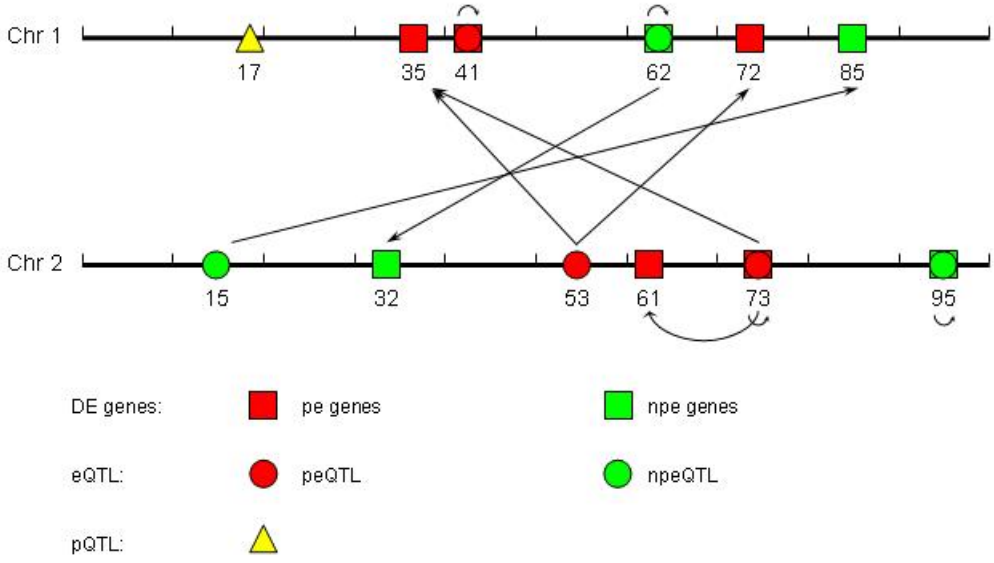

Figure 4.2 The genetic map of different QTL and corresponding DE genes.

The locations of different types of QTL and corresponding DE genes in our simulation are shown in Figure 4.2. Nine DE Genes and five QTL are randomly assigned to different locations on the two chromosomes. The type of DE genes, eQTL and pQTL are represented by different symbols. All the DE genes and associated eQTL that are related to the trait are red or yellow, and those that are not related to the trait are green. The arrows denote the association between DE genes and their corresponding eQTL. Chr 1 has one pQTL, one peQTL and one npeQTL, and chr 2 has two peQTL and two npeQTL. Each eQTL belongs to one of three categories: cis-acting, trans-acting or both. There is a total four cis-acting and six trans-acting genes, which is consistent with the fact that trans-acting genes are more common than cis-acting genes (Jansen and Nap 2004, Pomp et al., 2004, Gibson and Weir 2005). The two QTL at 53cM and 95cM on chr 2 are linked to the pQTL at 73cM on chr 2. They are used to detect how linkage disequilibrium affects false positive results, i.e., the peGene associated with the peQTL at 53cM might be falsely associated with peQTL at 73cM due to the linkage disequilibrium. The peGene at 35cM on chr 1 represents a differentially expressed gene that is associated with multiple eQTL. The peQTL at 73 cM on chr 2 represents an eQTL that controls multiple genes. Based on the genetic map, QTL genotypes and molecular marker genotypes were simulated according to the Haldane mapping function.

4.2.3 Simulation models for mRNA expression levels

The mRNA level of differentially expressed genes is simulated as a linear function of eQTL effects (cis-acting or trans-acting) and biological and technical error (Han et al., 2004, Lu 2003, Kendziorski et al., 2003 and Shih et al., 2004). The true mRNA level of the i^{th} gene for the j^{th} individual is denoted by m_{ij} and modeled on the log scale as

$$\log(m_{ij}) = \mu_i + \sum_{k=1}^{T} a_{ik} + \epsilon_{ij}, \tag{4.1}$$

$i = 1, 2, ..., G$, $j = 1, 2, ..., N$ and $k = 1, 2, ..., T$. Here, G is the number of DE genes, N is the total number of individuals, and T is the total number of eQTL. The quantity μ_i represents the mean gene expression level for the i^{th} gene and a_{ik} is the effect of the k^{th} eQTL on the expression of the i^{th} gene. We assume that all eQTL are additive. Also, we assume that a_{ik} is on the log scale and contains only additive effects; no dominance effects are included. The random error ϵ_{ij} accounts for biological variation, and is assumed to be independently, identically distributed as $N(0, \sigma_b^2)$, where σ_b^2 denotes the biological variance in gene expression. Then, the observed gene expression level o_{ij} on the log scale is modeled as

$$\log(o_{ij}) = \log(m_{ij}) + \xi_{ij}, \tag{4.2}$$

where ξ_{ij} is technical error which is also assumed to be independently, identically distributed as $N(0, \sigma_t^2)$. Here, σ_t^2 is technical variance introduced in the multiple steps of a microarray experiment (Chen et al., 2004). Biological and technical errors are assumed to be independent.

One major step in pQTL transcriptome mapping is to mix the mRNA of randomly selected individuals to form disjoint pools within genotype groups. Suppose now that the mRNA from n subjects is combined to form a pool. Let m_{ij}^p denote true expression for the i^{th} gene in the j^{th} pool. We assume that the true expression level of a gene in the mRNA pool is a weighted average of the true expression levels of the gene in all individuals in the same mRNA pool (Zhang et al. 2006), so that

$$m_{ij}^p = \sum_{k=1}^{n} (w_{ijk} * m_{ijk}), \tag{4.3}$$

where

$$w_{ijk} = \frac{z_{ijk}}{z_{ij1} + z_{ij2} + ... + z_{ijn}},$$

$i = 1, 2, ..., G$, $j = 1, 2, ..., P$, $k = 1, 2, ..., n$, where G is the same as defined before, P is the number of mRNA pools, and n is the number of individuals per mRNA pool such

that, $P*n = N$. When $n = 1$, the experiment involves no mRNA pooling (a microarray is made for each individual). The random unobservable weight w_{ijk} represents the relative contribution of individual k to pool j for gene i and is a function of the z_{ijk}, which denote the technical deviations from the ideal pool containing equal amounts of mRNA from each individual sample. We assume that the z_{ijk} are independently, identically distributed as $N(1, \sigma_z^2)$, where σ_z^2 denotes the pooling technical variance. If we denote the observed mRNA level in a pool by o_{ij}^p we can then model it on the log scale as:

$$\log(o_{ij}^p) = \log(m_{ij}^p) + \xi_{ij}, \tag{4.4}$$

where ξ_{ij} is technical error as defined earlier.

4.2.4 Simulation model for phenotypic values

As discussed earlier in Section 4.2.2, phenotypic values of an individual can be affected not only by pQTL, which results in structural transcript difference, but also by peQTL, which control the mRNA expression of trait related genes. Therefore, we model the phenotypic values for trait as a linear function of pQTL effect and the true mRNA expression level of trait-related genes on the log scale. The phenotypic value for the i^{th} individual, y_i, can be modeled as

$$y_i = \nu + \sum_{k=1}^{S} a_{pQTLk} + \sum_{j=1}^{G} \beta_j * log(m_{ij}) + \varepsilon_i, \tag{4.5}$$

where $i = 1, 2, ..., N$. Here, ν is the mean phenotypic value across all the individuals; S is the total number of pQTL which affect the trait; a_{pQTLk} is the effect of the k^{th} pQTL on the phenotypic value and we again assume no dominance and no epitasis; β_j is the regression coefficient of phenotypic value on the j^{th} gene's log expression level, i.e., the effect of one unit increase in log of gene expression level on phenotype. ε_i

is the environmental error and assumed to be independently, identically distributed as $N(0, \sigma_p^2)$, where σ_p^2 is the environmental variance.

4.2.5 Statistical models to detect differentially expressed genes

A main objective in transcriptome mapping approaches is to identify the DE genes which are controlled by the QTL that contributes most to phenotype variation. The differentially expressed genes that are detected in eQTL transcriptome mapping can be either trait related or non-trait related. Therefore, we propose implementing a regression method in the context of an eQTL transcriptome mapping experiment to identify the DE genes that are trait related and controlled by the most significant phenotype QTL. The phenotype and marker genotypes are generally known in QTL mapping experiments and using this marker information we can then predict the location of the most significant phenotype QTL. In turn, using information on the most significant QTL location and the nearby marker genotypes, we can predict for each individual the probability of each genotype at the most significant QTL position (Haley and Knott, 1992). Here, we propose an F test to find DE genes which are associated with the most significant QTL. Let p_i^{QQ}, p_i^{Qq} p_i^{qq} be the predicted probabilities of each genotype QQ, Qq and qq for the i^{th} individual. Then, observed gene expression can be modeled as

$$\log(o_{ij}) = \mu_i + \beta^{QQ} * p_i^{QQ} + \beta^{Qq} * p_i^{Qq} + \beta^{qq} * p_i^{qq} + \varepsilon_i. \tag{4.6}$$

Here, β^{QQ}, β^{Qq} and β^{qq} are regression coefficients to predict the QTL effect of QQ, Qq and qq genotypes for the most significant QTL; $\log(o_{ij})$ is the observed j^{th} gene expression level on the i^{th} individual; ε_i is the random error. Therefore, the test of $\beta^{QQ} = \beta^{Qq} = \beta^{qq}$ can be used to find which genes are control by the significant QTL. In pQTL transcriptome mapping, we can only observe the expression level for a pool instead of each individual. Therefore, for the F test in pQTL transcriptome mapping, $\log(o_{ij})$ and p_i represent the observed expression level for a pool and the average genotype

probabilities of all the individuals in the same pool.

The method just described does not only identifies the DE genes controlled by the most significant QTL but also picks up the DE genes controlled by other QTL due to linkage disequilibrium. Therefore, also propose an appropriate F-test to implement in the context of composite interval mapping (Zeng 1994). Composite interval mapping, in contrast to the standard regression approach, can account for the effect of linkage disequilibrium. To carry out composite interval mapping, we include not only the QTL effects in the model, but also the the genotype probabilities of the closest left and right markers bracketing the interval where the QTL is potentially located. In our simulated genetic map, these markers are located at 60 and 90 cM on chr 2. Let $p_{i,l}^{MM}$, $p_{i,l}^{Mm}$, $p_{i,l}^{mm}$, $p_{i,r}^{MM}$, $p_{i,r}^{Mm}$ and $p_{i,r}^{mm}$ be the left and right marker genotype probabilities for the i^{th} individual, respectively. Then the model can be modified as follows

$$\begin{aligned}\log(o_{ij}) &= \beta^{QQ} * p_i^{QQ} + \beta^{Qq} * p_i^{Qq} + \beta^{qq} * p_i^{qq} + \beta_l^{MM} * p_{i,l}^{MM} + \beta_l^{Mm} * p_{i,l}^{Mm} \\ &\quad + \beta_l^{mm} * p_{i,l}^{mm} + \beta_r^{MM} * p_{i,r}^{MM} + \beta_r^{Mm} * p_{i,r}^{Mm} + \beta_r^{mm} * p_{i,r}^{mm} + \varepsilon_i. \qquad (4.7)\end{aligned}$$

Here, β_l^{MM}, β_l^{Mm}, β_l^{mm}, β_r^{MM}, β_r^{Mm}, β_r^{mm} are the QTL effects for the specific left and right markers. Again, because we assume that mRNA from several individuals is pooled, we use the $p_i's$ and $\log(o_{ij})$ to denote the average probabilities and the average expression level of all the individuals in a pool.

4.3 Results

4.3.1 Simulation results to find the eQTL for differentially expressed gene in eQTL mapping

An F2 cross population of 500 individuals was simulated at the genome, transcriptome, and phenome levels as described in Sections 4.2.2, 4.2.3 and 4.2.4. QTL genotypes

and molecular marker genotypes were first simulated based on the genome structure described in Figure 4.2. Then, the true and observed mRNA levels on the log scale of nine DE genes for 500 individuals were simulated according to expressions (4.2) with $\sigma_b^2 = 0.75^2$ and $\sigma_t^2 = 0.25^2$. The phenotypic values were also simulated according to expression (4.5) with $\sigma_p^2 = 2$. The associations between DE genes and the corresponding QTL and parameter values for expressions (4.2) and (4.5) are listed in Table 4.1. In table 4.1, we simulated data in three different scenarios: QTL being in coupling phase and small eQTL effect scenario (I), QTL being in coupling phase and large eQTL effect scenario (II) and QTL being in repulsion phase and small eQTL effect scenario (III). Note that in Table 4.1, the cis-acting eQTL effects are generally stronger than the trans-acting eQTL effects as Schadt et al. (2003) suggested. These eQTL account for about 1-10% of total variation at gene expression level. Also note that in the column labeled "β_j" of Table 4.1, the expression of DE genes can either increase (positive β) or decrease (negative β) the phenotypic trait of interest.

Using individual mRNA expression levels as phenotype, standard interval mapping was applied to find the corresponding eQTL for the nine DE genes. The results of Scenario I are shown as an example in Figure 4.3. Note that all the eQTL were correctly identified for the nine DE genes. Therefore, eQTL transcriptome mapping is effective to identify the eQTL which control gene differential expression.

4.3.2 Simulation results for pQTL transcriptome mapping

To validate the concept of pQTL transcriptome mapping, we used the same simulated 500 F2 individuals at the genome, transcriptome, and phenome level as described in Section 4.2.1. We first applied the standard QTL mapping method to the phenotypic values to find the most significant QTL. Figure 4.4 shows that the phenome QTL is correctly identified at 73cM on the 2^{nd} chr. Then, for the given pQTL region, We allocated individuals into three genotype groups (QQ Qq qq) by the most significant

Table 4.1 Summary of the association between DE genes and the corresponding QTL at gene expression level and phenome level.

Gene	Position	Type	Associated eQTL type and location	eQTL effects			percentage variation at gene expression level			β_j	percentage variation at phenome level		
				I	II	III	I	II	III		I	II	III
1	35 cM chr 1	peGene	peQTL 53cM chr 2	0.3	0.3	-0.3	4.1	4.0	4.1	-0.2	0.9	0.9	0.9
1	35 cM chr 1	peGene	peQTL 73cM chr 2	0.3	0.4	0.3	4.1	7.1	4.1	-0.2	0.9	0.9	0.9
2	41 cM chr 1	peGene	peQTL 41cM chr 1	0	0	0	0	0	0	0.3	3.5	3.4	3.5
3	62 cM chr 1	npeGene	npeQTL 62cM chr 1	0.5	0.5	0.5	11.1	11.1	11.1	0	0	0	0
4	72 cM chr 1	peGene	peQTL 53cM chr 1	0.3	0.3	0.3	4.3	4.3	4.3	-0.1	0.4	0.4	0.4
5	85 cM chr 1	npeGene	npeQTL 15cM chr 2	0.3	0.3	0.3	4.3	4.3	4.3	0	0	0	0
6	32 cM chr 2	npeGene	npeQTL 62cM chr 1	0.3	0.3	0.3	4.3	4.3	4.3	0	0	0	0
7	61 cM chr 2	peGene	peQTL 73cM chr 1	0.3	0.6	0.3	4.3	15.3	4.3	0.4	6.7	7.1	6.4
8	73 cM chr 2	peGene	peQTL 73cM chr 2	0.5	0.7	0.5	11.1	19.7	11.1	0.5	11.1	11.7	10.8
9	95 cM chr 2	npeGene	npeQTL 95cM chr 2	0.5	0.5	0.5	11.1	11.1	11.1	0	0	0	0

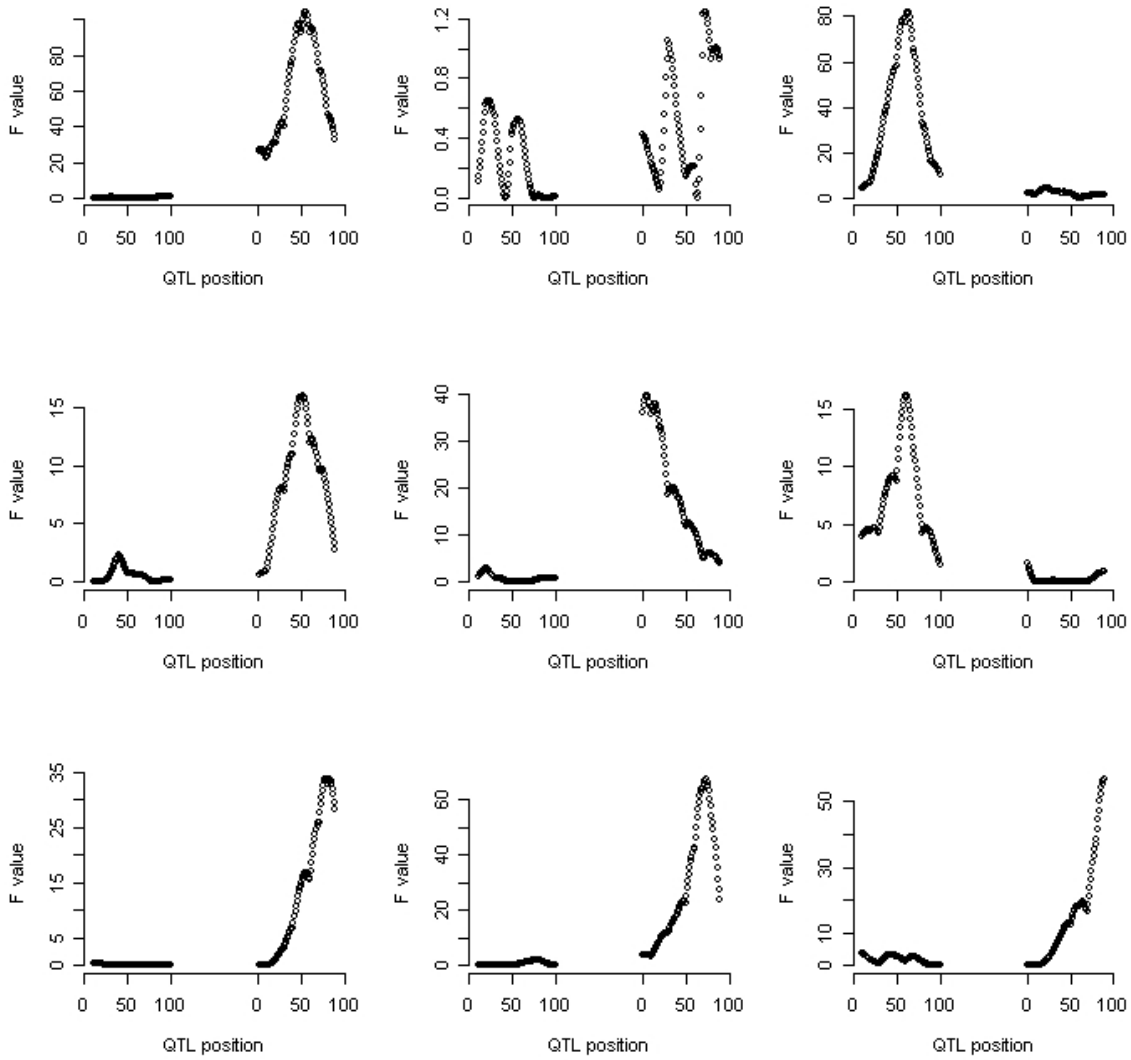

Figure 4.3 eQTL transcriptome mapping of nine simulated differentially expressed genes using the standard mapping method (Scenario I).

QTL genotype based on genotypic probabilities derived using the marker genotypes. Each genotype group contained only those individuals for whom we could estimate the pQTL genotype with at least 90% certainty. In that way, we are guaranteed almost uniform genotypes within each group. The 90% threshold is clearly arbitrary and other thresholds might also be considered. In our case, about 85% of the individuals satisfied the inclusion criterion and were included on one of the genotypic groups.

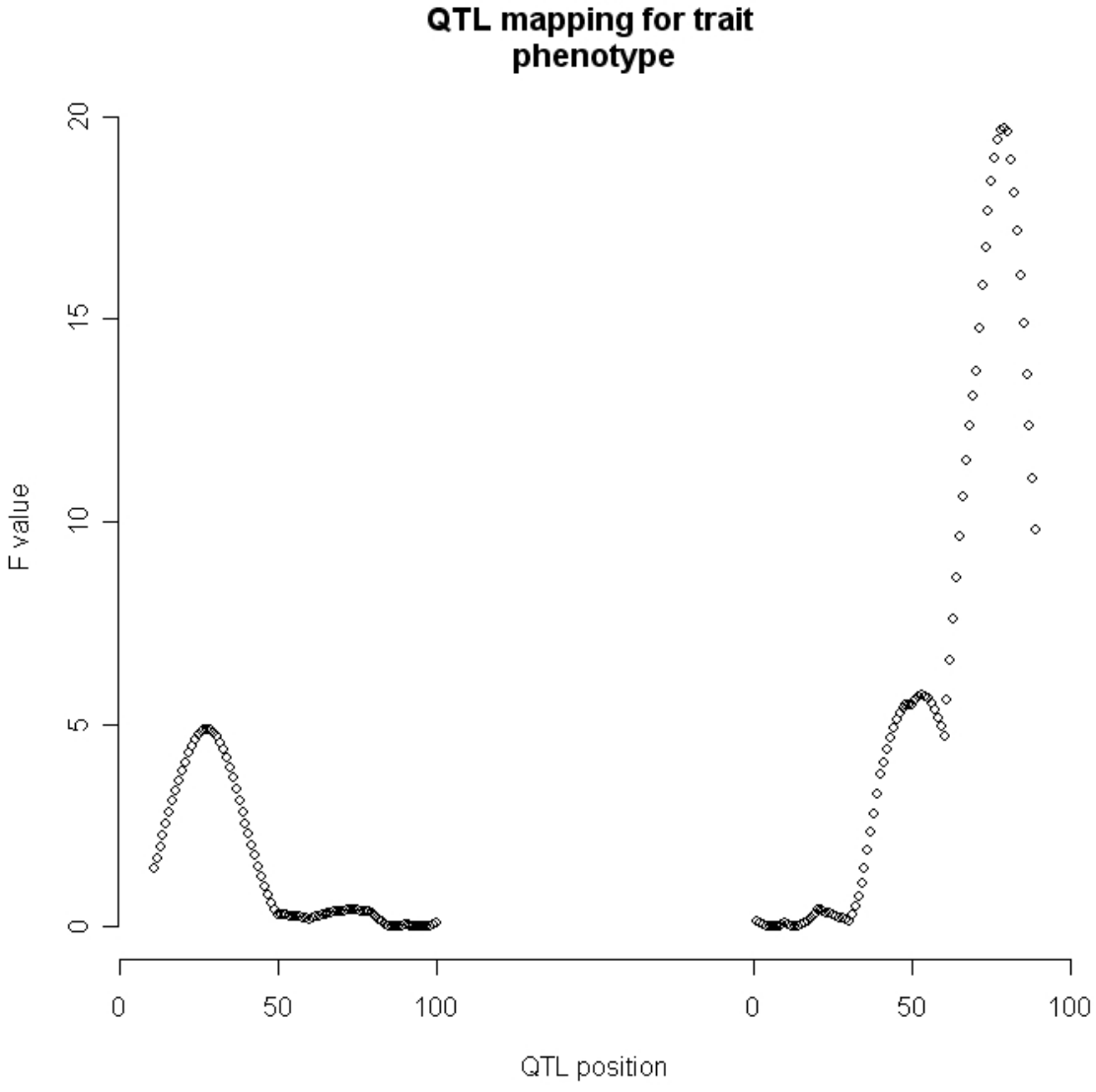

Figure 4.4 Standard QTL mapping to find the most significant phenome QTL (Scenario I).

Within these QTL genotype groups, we randomly picked five individuals ($n = 5$) to form an mRNA pool. We simulated the true and observed mRNA expression level on the log scale for all the mRNA pools according to expressions (4.3) and (4.4) with $\sigma_t^2 = 0.25^2$ and $\sigma_z^2 = 0.05^2$. Then we used the standard regression method to find the DE genes whose expression levels were associated with the most significant QTL located at 73cM on chr 2. We found that four DE genes were significantly associated with the most significant QTL (0.05 level). Among these four significant genes, three DE genes (35cM on chr 1, 61cM on chr 2 and 73cM chr 2) were truly associated with the QTL at 73cM on the 2nd chr and we were expecting to find all three. The fourth DE gene, however, was a false positive due to linkage disequilibrium, i.e., the DE gene at 72cM on chr 1 is controlled by the QTL at 53cM on chr 2, which is located close to the most significant QTL and therefore has a similar QTL genotype.

4.3.3 Empirical power: comparison between pQTL and eQTL transcriptome mapping for identifying trait-related genes

The method of pQTL transcriptome mapping can directly detect trait-related genes. In contrast, although eQTL transcriptome mapping can successfully detect associations between differentially expressed genes and the corresponding eQTL, further analyses need to be performed to pick up the trait-related differentially expressed genes. Therefore, we used the regression method described in Section 3.2.4 to identify the genes that were controlled by the most significant QTL in the context of eQTL transcriptome mapping and compared the empirical power of that procedure to the power that would be achieved if pQTL mapping had been applied. Results are shown on the left section of Table 4.2. We found that in both mapping approaches, the standard regression method correctly identifies genes 1, 7 and 8, but also frequently identifies genes 4, 5 and 9 which are not controlled the most significant QTL. These false positive results are due to linkage disequilibrium. We also noticed that in addition to increasing the chances of false

positive results, linkage disequilibrium can affect the power of the test within the truly associated DE genes. For example, in the positively linked (coupling phase) scenario I and II, the power for gene 1 is very high because this gene is not only controlled by the most significant QTL, but also is controlled by a QTL at 53cM on chr 2, which is positively linked to the most significant QTL. However, in scenario III (repulsion phase), the QTL at 53cM on chr 2 is negatively linked to the most significant QTL and in this case the power with which we can identify gene 1 drops dramatically. The power that can be achieved in the pQTL transcriptome mapping approach is generally smaller than in the eQTL transcriptome mapping approach as we had anticipated. Because gene expression levels are measured in pooled mRNA samples rather than in individual samples, we have less information to draw statistical inferences. Note that the power loss is small relative to the dramatic reduction in the cost of the experiment.

We also estimated the additive effects in the regression approach and the results corresponding to scenario I are shown for illustration in Table . We found that estimates of additive effects were biased upwards when the QTL are in a coupling phase (positively linked) and biased downwards when the QTL are in a repulsive phase (negatively linked).

As can be clearly seen from Table 4.2, the commonly used regression method did not account for the effect the linkage disequilibrium, which results in biased estimation of both power and QTL effects. Therefore, we applied the composite mapping method (Section 4.2.5) to compare the empirical power between eQTL and pQTL transcriptome mapping for finding the trait-related genes. Results under the three different scenarios are presented in the right section of Table 4.2. Power is calculated as the percentage of significant tests at the 0.01 level when testing the hypothesis of gene differential expression under the most significant QTL (at 73cM chr 2). Within each scenario, we found that both approaches can correctly identify the DE genes (genes 1, 7 and 8) associated with the most significant QTL. The stronger the QTL effect, the higher the power to detect the association between QTL and their corresponding DE genes.

Table 4.2 Comparison of empirical power between eQTL and pQTL transcriptome mapping by the regression method and composite mapping method. Power is calculated as the percentage of significant results at the 0.01 level over 1000 replicates.

	Regression method						Composite method					
	I		II		III		I		II		III	
Gene	eQTL	pQTL	eQTL	pQTL	eQTL	pQTL	eQTL	pQTL	eQTL	pQTL	eQTL	pQTL
1	0.97	0.97	1.00	1.00	0.07	0.05	0.25	0.06	0.53	0.08	0.26	0.04
2	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
3	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.01	0.01	0.02	0.01
4	0.67	0.46	0.72	0.47	0.76	0.53	0.01	0.01	0.01	0.01	0.02	0.01
5	0.15	0.09	0.13	0.08	0.15	0.09	0.02	0.02	0.01	0.01	0.01	0.01
6	0.02	0.02	0.01	0.01	0.01	0.01	0.02	0.02	0.02	0.01	0.01	0.02
7	0.93	0.81	1.00	1.00	0.95	0.81	0.29	0.05	0.93	0.14	0.31	0.06
8	0.97	0.97	1.00	1.00	0.99	0.99	0.71	0.09	0.98	0.21	0.80	0.14
9	0.96	0.91	1.00	0.94	0.97	0.89	0.01	0.02	0.01	0.02	0.02	0.03

Table 4.3 Comparison of additive effect estimates between eQTL and pQTL transcriptome mapping by the regression method and composite method. Entries are the mean (standard deviation) of estimates over 1000 replicates.

		Regression method		Composite method	
Gene	True effect	eQTL mapping	pQTL mapping	eQTL mapping	pQTL mapping
1	0.3	0.48 (0.07)	0.48 (0.08)	0.32 (0.17)	0.29 (0.37)
2	0	0.00 (0.07)	0.00 (0.08)	-0.01 (0.17)	0.01 (0.37)
3	0	0.01 (0.07)	0.00 (0.09)	0.00 (0.18)	-0.02 (0.39)
4	0	0.20 (0.07)	0.20 (0.08)	0.01 (0.17)	0.01 (0.38)
5	0	0.10 (0.07)	0.10 (0.08)	0.01 (0.17)	0.00 (0.38)
6	0	0.00 (0.07)	0.00 (0.08)	0.01 (0.17)	0.01 (0.38)
7	0.3	0.29 (0.07)	0.29 (0.08)	0.31 (0.17)	0.28 (0.37)
8	0.5	0.47 (0.07)	0.48 (0.08)	0.51 (0.17)	0.47 (0.37)
9	0	0.35 (0.07)	0.35 (0.08)	0.00 (0.17)	0.03 (0.38)

Note that the power for pQTL transcriptome mapping is much lower than the power for eQTL transcript mapping. Further, the power in the composite mapping method is lower than in the regression method. This might be due to the fact that for composite mapping we must include the genotypic probabilities of two markers in the model. In terms of the estimation of QTL effects, we find that the collinearity between the marker genotype and the predicted QTL genotypic probabilities can be strong in mRNA pools, which can then increase the standard error of the estimator of QTL effect and thus mask potentially significant results. We compared the performance of the estimator of additive QTL effect (Table 4.3) across the two transcriptome mapping approaches and found that the point estimates of additive effect were similar and unbiased. However, the precision with which we can estimate additive QTL effects is higher in the eQTL transcriptome mapping approach.

4.4 Discussion

Through simulation, we have shown that pQTL transcriptome mapping is an effective method to find differentially expressed genes controlled by trait QTL. Although the power of pQTL mapping is lower than the power of eQTL and point estimates of QTL effect sizes are less precise, the lower cost of a pQTL experiment may still justify its use. Therefore, there is a trade-off between power,precision and cost. If the objective of an experiment is to pin-point the most significant trait-related genes and not to identify all the trait-related genes, then pQTL mapping can be an efficient and cost saving alternative. The differentially expressed genes (pQTL DE-genes) identified in the pQTL mapping approach may potentially be genes involved in pathways that are controlled by the pQTL (pQTL pathway genes). Based on map location, they can be grouped into those that are located inside (internal pQTL DE-genes) and outside (external pQTL DE-genes) the pQTL region. Internal pQTL DE-genes can become important candidate genes for the pQTL and might be included in further experimentation.

Many factors can affect the power of pQTL transcriptome mapping approaches. Those include the population size, the marker density, the effect and position of pQTL and eQTL, the interaction between the different QTL, the correlation between expression level and phenotypic value, the biological and technical variances, the pooling technical variation at the transcription level and the environmental variation at the phenome level and the mRNA pooling strategy. The population size, marker density, the effect and position of pQTL and eQTL,the correlation between expression level and phenotypic value and the environmental variability at the phenome level play a main role in the prediction of phenome QTL location (Step 1 and 2 in pQTL transcriptome approach). For a given phenome QTL location, the prediction of genotype probability will depend on marker density and cut-off criteria (Step 3 pQTL transcriptome approach). Based

on the mRNA pools of grouped individuals, the statistical analysis for identifying trait-related genes can be affected by the effect and position of pQTL and eQTL, the effect of the biological and technical variances, the pooling technical variation at the transcription level and the pooling strategy (Step 4 in pQTL transcriptome mapping). In an experiment, the marker density, the effect and position of pQTL and eQTL, interaction between the QTL, the correlation between expression level and phenotypic value, the biological and technical variance, the pooling technical variation at the transcription level and the environmental variation at the phenome level are fixed and cannot be changed by the researcher. Therefore, the mRNA pooling strategy, under the control of the researcher, might be one determinant of the power that can be achieved in a pQTL transcriptome mapping experiment. The strategy for pooling mRNA involves the choice of the number of pools and of the size of each pool. Several researchers (Kendziorski et al., 2003, Shih et al., 2004, Zhang et al., 2006) have investigated the effect of pooling mRNA on power in microarray experiments. They argue that pooling mRNA from a few similar individuals when conducting microarray experiments is feasible and this further supports the concept of the pQTL mapping approach.

We applied two methods to identify trait-related differentially expressed genes in the context of eQTL and pQTL transcriptome mapping approaches. In the commonly used regression method, linkage disequilibrium results in false detection of pQTL pathway genes. The larger the QTL effects and the stronger the linkage disequilibrium, the higher the false positive rate. Further, the regression method for estimating QTL effects can result in biases whose directions depend on whether the QTL are in coupling or repulsive phase. Therefore, we propose that the composite mapping method is a more reliable alternative for identifying DE genes which are associated with a given QTL. To carry out the composite mapping approach, we include QTL genotype and the genotype of the left and right markers bracketing the putative QTL location in the statistical

model. In this way, our analysis accounts for the effect of linkage disequilibrium and avoids the false positive results. Further, it results in unbiased estimates of QTL effects. Note that linkage disequilibrium can also be overcome by using advanced intercross lines as the resource population since these can have less extensive linkage disequilibrium.

One major step in pQTL transcriptome mapping is the prediction of the genotype of all the individuals assuming a QTL location (step 3). The prediction of QTL genotype may depend on the QTL location, the location of nearby markers and the distance of the given QTL location to the nearby markers. If individuals have not been correctly genotyped, this can severely affect the power of the test. In this work, we assigned individuals to a genotype group only if the predicted QTL genotype had probability higher than 90%. Using this criterion, we eliminated about 15% of the individuals from the study. To further investigate the reliability with which we had grouped individuals with equal genotype at the QTL, we compared the predicted QTL genotype with the true QTL genotype and found at least 95% similarity, i.e., we had at least 95% certainty that individuals in the same group had uniform genotype. Therefore, uncertainty about the true genotype at the QTL did not significantly affect the observed power in our simulation study. If the cut-off criterion is lowered, more individual will be kept in the analysis. However, the pools within each genotype group may no longer be uniform and this may result in lower statistical power.

The pQTL transcriptome mapping is proposed under the assumption that there is one major QTL and that the individuals are grouped according to this major QTL genotype. If there are several QTL with moderate effect, different genotypic groups may be created for the multiple pQTL to investigate the main effects and the interactions between pQTL. Further, pQTL transcriptome mapping focuses only on a single phenotypic trait. The same type of analysis can be carried out if other traits are of interest, and would likely

involve different pQTL.

Clearly, eQTL transcriptome mapping can provide a wealth of information on the genetic control of transcriptome variation. However, a large proportion of the differentially expressed genes and their associated eQTL will not be relevant to the phenotypic traits of interest. Instead, they will affect what are referred to as non-trait phenotypes. Identification of relevant expression differences requires additional analyses to connect transcriptome variation to pQTL genotypes. In contrast, the proposed pQTL approach directly pinpoints differentially expressed genes associated with the phenotypic traits of interest.

Overall, the pQTL transcriptome mapping approach is a valid method to explore the genetic architecture of traits of interest. Because an erosion of the power of statistical tests is inevitable, the pQTL transcriptome mapping approach can be a good choice if the main interest is to identify only the most significant differentially expressed genes in the regulatory pathway. The benefits of implementing the pQTL rather than the eQTL approach include a large reduction in the cost of experimentation and also in the amount of data to be analyzed because pQTL only generates expression data relevant to the trait(s) of interest.

4.5 Acknowledgement

This study was funded by the Center for Integrated Animal Genomics at Iowa State University.

4.6 References

Andersson, L. (2001) Genetic dissection of phenotypic diversity in farm animals. *Nat Rev Genet*, **2**, 130-138.

Bao, L., Wei, L., Peirce, J.L., Homayouni, R., Li, H., Zhou, M., Chen, H., Lu, L., Williams, R.W., Pfeffer, L.M., Goldowitz, D. and Cui, Y. (2006) Combining gene expression QTL mapping and phenotypic spectrum analysis to uncover gene regulatory relationships. *Mamm Genome*, **17**, 575-583.

Brem, R.B., Yvert, G., Clinton, R. and Kruglyak, L. (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752-755.

Bystrykh, L., Weersing, E., Dontje, B., Sutton, S., Pletcher, M.T., Wiltshire, T., Su, A.I., Vellenga, E., Wang, J., Manly, K.F., Lu, L., Chesler, E.J., Alberts, R., Jansen, R.C., Williams, R.W., Cooke, M.P. and de Haan, G. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet*, **37**, 225-232.

Cabrera, C.P., Dunn, I., Fell, M., Wilson, P., Burt, D.W., Waddington, D., Talbot, R., Hocking, P.M., Law, A., Haley, C.S., Knott, S. and de Koning D.J. (2006) Application of genetic genomics to a marked QTL in poultry. *8th World Congress on Genetics Applied to Livestock Production.*

Chen, J., Delongchamp, R., Tsai, C., Huey-min, H., Sistare, F., Thompson, K.L., Desai, V.G. and Fuscoe, J.C. (2004) Analysis of variance components in gene expression data. *Bioinformatics.* **20**, 1436-1446.

Chesler, E.J., Lu, L., Shou, S., Qu, Y., Gu, J., Wang, J., Hsu, H.C., Mountz, J.D., Baldwin, N.E., Langston, M.A., Threadgill, D.W., Manly, K.F. and Williams, R.W. (2005) Complex trait analysis of gene expression uncovers polygenic and

pleiotropic networks that modulate nervous system function. *Nat Genet*, **37**, 233-242.

Decook, R., Lall, S., Nettleton, D., and Howell, S.H. (2006) Genetic regulation of gene expression during shoot development in Arabidopsis. *Genetics*, **172**, 1155C1164.

de Koning, D.J., Carlborg, O. and Haley, C.S. (2005) The genetic dissection of immune response using gene-expression studies and genome mapping. *Veterinary immunology and immunopathology*, **105**, 343-352.

de Koning, D.J. and Haley, C.S. (2005) Genetical genomics in humans and model organisms, *Trends Genet*, **21**, 377-381.

Doss, S., Schadt, E.E., Drake, T.A. and Lusis, A.J. (2005) Cis-acting expression quantitative trait loci in mice. *Genome research*, **15**, 681-691.

Drake, T.A., Schadt, E.E. and Lusis, A.J. (2006) Integrating genetic and gene expression data: application to cardiovascular and metabolic traits in mice. *Mamm Genome*, **17**, 466-479.

Gibson, G. and Weir, B. (2005) The quantitative genetics of transcription. *Trends Genet*, **21**, 616-623.

Haley, C.S. and Knott, S.A. (1992) A simple method for mapping quantitative trait loci in line across using flanking markers. *Heredity*, **69**, 315-324.

Han E., Wu Y., McCarter R., Nelson J.F., Richardson A. and Hilsenbeck S.G. (2004) Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. *Journal of Gerontology: Biological Sciences*. **4**:306.

Hubner, N., Wallace, C.A., Zimdahl, H., Petretto, E., Schulz, H., Maciver, F., Mueller, M., Hummel, O., Monti, J., Zidek, V., Musilova, A., Kren, V., Causton, H., Game, L., Born, G., Schmidt, S., Muller, A., Cook, S.A., Kurtz, T.W., Whittaker, J., Pravenec, M. and Aitman, T.J. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet*, **37**, 243-253.

Jansen, R.C. (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet*, **4**, 145-151.

Jansen, R.C. and Nap, J.P. (2001) Genetical genomics: the added value from segregation. *Trends Genet*, **17**, 388-391.

Kendziorski, C. and Wang, P. (2006) A review of statistical methods for expression quantitative trait loci mapping. *Mamm Genome*, **17**, 509-517.

Kendziorski, C.M., Zhang, Y., Lan, H. and Attie, A.D. (2003) The efficiency of pooling mRNA in microarray experiments. *Biostatistics*, **4**, 465-477.

Kirst, M., Basten, C.J., Myburg, A.A., Zeng, Z.B. and Sederoff, R.R. (2005) Genetic architecture of transcript-level variation in differentiating xylem of a eucalyptus hybrid. *Genetics*, **169**, 2295-2303.

Pastinen, T. and Hudson, T.J. (2004) Cis-acting regulatory variation in the human genome, *Science*, **306**, 647-650.

Pomp, D. (1999) Animal models of obesity. *Mol. Med. Today* **5**, 459C460.

Pomp, D., Allan, M.F. and Wesolowski, S.R. (2004) Quantitative genomics: exploring the genetic architecture of complex trait predisposition. *J Anim Sci*, **82** E-Suppl, E300-312.

Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B. and Friend, S.H. (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature*, **422**, 297-302.

Schadt, E.E., Monks, S.A. and Friend, S.H. (2003) A new paradigm for drug discovery: integrating clinical, genetic, genomic and molecular phenotype data to identify drug targets. *Biochem Soc Trans*, **31**, 437-443.

Shih, J.H., Michalowska, A.M., Dobbin, K., Ye, Y., Qiu, T.H. and Green, J.E. (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics*. **20**, 3318-3325.

Yvert, G., Brem, R.B., Whittle, J., Akey, J.M., Foss, E., Smith, E.N., Mackelprang, R. and Kruglyak, L. (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. *Nat Genet*, **35**, 57-64.

Z-B Zeng (1994) Precision mapping of quantitative trait loci. *Genetics*, **136**, 1457-1468.

Zhang, W., Carriquiry, A., Nettleton, D. and Dekkers, J. (2006) Pooling mRNA in microarray experiments and its effect on power. submitted to *Bioinformatics*.

CHAPTER 5. DISCUSSION

Microarrays are powerful tools to simultaneously measure the mRNA expression levels of thousands of genes and are widely applied to study complex biological problems at the genetic level. In such experiments, mRNA from individuals is sometimes pooled and the microarray is then obtained for the pool rather than for each individual. Pooling mRNA is the strategy adopted when the cost of individual microarrays is prohibitive or when the amount of mRNA available for each individual is not sufficient. While pooling mRNA for analyses can be an enabling strategy, questions arise about the loss of information incurred when pooling mRNA from a group of individuals. In this dissertation, we investigate the impact of pooling mRNA on different types of inferences that might be drawn from microarray and QTL mapping experiments.

More precisely we focus on the following questions:

1. What is an appropriate statistical model to represent observed gene expression in a pool of mRNA samples?

2. If we are interested in identifying genes with differential expression under different conditions or treatments, to what extent will pooling mRNA result in a loss of power of the appropriate statistical tests?

3. When designing a microarray experiment that involves pooling mRNA, which pooling strategies result in inferences that better approximate what might have been discovered using individual mRNA samples?

4. Suppose that we have measurements on some phenotypic trait for all individuals in an experiment, but that we have measured gene expression only in pools formed by mixing mRNA from various individuals. How do we estimate the correlation between gene expression and phenotype in that case?

5. What pooling strategy would result in an estimate of the correlation coefficient with smaller bias and root mean squared error?

6. Finally, if the objective of the experiment is to find genes whose expressions are associated with one or more QTL, can we achieve acceptable statistical power if we pool mRNA samples?

We addressed questions 1 - 3 in Chapter 2 of the dissertation. Questions 4 and 5 were the subject of Chapter 3. Finally, in Chapter 4 we discussed the impact of pooling mRNA on power with which we can identify genes whose expressions were associated with one or more QTL.

In Chapter 2 we proposed a realistic statistical model for observed gene expression level in a pool. The model mimicked the mRNA pooling process in that it recognized that mRNA is pooled in the original, untransformed scale and that the mRNA contributions from each individual to the pool might not be exactly equal. Under that model, we developed the appropriate F test to identify differentially expressed genes. Via simulation, we showed that the power derived under our proposed model was conservative and close to the true power. In this sense, our approach performed better than other approaches that had been proposed in the literature. The simulation experiment was designed to permit investigating the effects of the number of pools, the number of individuals per pool, the number of measurements of each pool obtained in the experiment, the effect size, and the relative sizes of the biological and technical variances on the power of the appropriate F-test. We found that obtaining replicate measurements from each pool had negligible effect on power. We also found that for large enough effect sizes,

it was possible to design an experiment based on pooling mRNA samples that almost achieves the power that would be obtained when arraying individual mRNA samples. Thus, we argued that under certain conditions, pooling mRNA samples could be a good strategy if cost is a consideration.

In Chapter 3 we again investigated the effect of pooling mRNA, but this time focused on the accuracy and reliability with which we can identify genes whose transcript abundance is correlated with phenotype for a quantitative trait. Here, we considered two different pooling strategies: mRNA pools are randomly created (random pool design) or individuals are first stratified by phenotype and phenotypically similar individuals are allocated to the same pool (stratified pool design). We found that the commonly used Pearson correlation estimator worked well in the random pool design but was biased in the stratified pool design. Therefore, we proposed a maximum likelihood estimator (MLE) to estimate the correlation between genes transcript abundance and phenotypic trait under the two pooling strategies. Using a simulation study, we showed that the MLE and the Pearson product-moment correlation estimates perform equally well when individuals are randomly allocated to pools. The MLE estimate is superior to the Pearson product-moment correlation estimate (in the sense that it has smaller bias) when individuals are allocated to pools according to their phenotypic value. Furthermore, once an MLE has been obtained, it can be used in a likelihood ratio test to determine whether gene expression level is correlated with phenotype. We showed that the empirical power in the likelihood ratio test is the same as in a permutation test and as in the usual test based on Fisher's z-transformation of the correlation coefficient. We also showed that power is generally higher for the stratified pool design than for the random pool design. Therefore, if mRNA must be pooled, creating the pools by mixing mRNA of individuals with similar phenotype and then estimating the correlation between gene expression and phenotype using ML can be a reasonable strategy.

In Chapter 4, we explored the consequences of pooling mRNA if we wish to identify

genes whose expression levels are controlled by a QTL of interest. Recently, an approach known as eQTL transcriptome mapping was proposed and is being applied today in many species. The eQTL methodology combines quantitative trait locus (QTL) mapping with gene expression analysis and permits investigating the genetic regulatory pathways at the mRNA level. However, the high cost of performing individual microarrays for eQTL transcriptome mapping limits sample sizes and consequently, the power of tests used to identify trait or pathway related genes. Therefore, we investigated whether pooling mRNA might be a viable alternative in this type of experiment. A methodology denoted as pQTL transcriptome mapping was recently proposed (Cabrera et al., 2006), in which individuals are allocated to a pool according to their QTL genotype. Gene expression levels are then measured in the mRNA pools. To assess the performance of the pQTL approach in terms of statistical power, we simulated an intercross F2 population at the genome, transcriptome and phenome levels. We implemented the eQTL and pQTL approaches on the simulated populations and compared the empirical power for the two methods. In the commonly used regression method, the power of pQTL transcriptome mapping is lower but comparable to eQTL transcriptome mapping. However, linkage disequilibrium results in false detection of pQTL pathway genes and biased estimation of QTL effects, whose directions depend on whether the QTL are in coupling or repulsive phase. The composite method accounts for the effect of linkage disequilibrium and results in unbiased estimates of QTL effects. However, pooling mRNA results in a loss of power when compared to eQTL transcriptome mapping and the point estimates of QTL effects are less precise. Although there might be a substantial loss in power, pooling mRNA is clearly advantageous from an economic point of view. We concluded that if the objective of an experiment is to pinpoint only the most significant trait-related genes, pQTL mapping is a cost-effective approach.

Clearly, carrying out microarray experiments using individual mRNA samples is always the best strategy in terms of the accuracy and reliability with which we can

estimate quantities or associations of interest. This is likely due to the effect of larger number of individuals. However, as suggested by the results we have obtained, it is possible to design an experiment that involves mRNA pooling and that can dramatically reduce the cost of experimentation while still achieving acceptable power or precision. It is not possible to recommend a single design that would perform equally well in all situations. Instead, we have tried to describe the factors which contribute to greater power or precision when mRNA must be pooled in the three experimental scenarios on which we have focused in this work. When considering whether to pool the mRNA from various individuals, the research should understand the trade off between power, precision and cost, and be able to choose a suitable design according to the research objectives, the lab resources and cost constraints.

When the number of individuals that can be included in an experiment is fixed, constructing as many pools as possible within the total sample size constraint appears to be the best approach. For example, by pooling individuals in pairs, it is possible to reduce the cost of experimentation by half while minimizing the reduction in degrees of freedom. We showed in Chapter 2 that it is possible to achieve adequate power for any effect size when the pool sizes is small and the number of pools is large. A similar result was obtained in Chapter 3, where we showed, via simulation studies, that in order to estimate the correlation between phenotype and gene expression, an important factor is the number of degrees of freedom (in our case, the number of pools) that can be used in calculations. Thus, in general it appears to be important to include as many pools in the design of the experiment as the experimental budget will allow.

One major assumption in the experiments involving mRNA pooling is that the gene expression level of a mRNA pool is close to the average gene expression of all the individuals in the pool. To investigate whether this is a plausible assumption, we cannot rely on simulated data but must instead carry out a laboratory experiment in which gene expression is measured both at the individual level and also on pools created by

mixing mRNA samples. Several experiments (Kendziorski et al. 2005 and Shih et al. 2004) have been carried out to investigate the validity of the assumption. However, they compared measurements obtained from mRNA pools with individual measurements on a normalized scale, which assumes that pooling mRNA can accomplished on the transformed scale. As we argued in Chapter 2 of this dissertation, this assumption cannot be justified. If the assumption holds on the original scale, then gene expression measurements from the pools would be highly correlated with the average expression measurements obtained from individuals in the pools. We would also expect to observe approximately the same set of differentially expressed genes whether measurements are taken from pools or from individuals. It is important to realize, however, that the degree of correlation between individual and pool measurements will depend on the relative size of the pooling technical error variance, which reflects the accuracy with which the pool is constructed. For example, while we assume that equal amounts of mRNA are contributed by each individual in the pool, in practice this assumption is likely to be violated. If pools are likely to contain very different amounts of mRNA from each member, then the statistical models and analyses should be extended to better reflect the process of pooling mRNA.

In our work, we discussed the pQTL transcriptome mapping approach assuming that the mRNA pools are formed by randomly grouping individuals within a QTL genotype. Under that pooling strategy we implemented both a regression method and a composite mapping method to identify the genes in the pathway of the trait. A next step would be to extend the concept of stratified pools into the pQTL transcriptome mapping methodology. For example, we could think about stratifying individuals according to phenotypic values within each genotype class and then pooling the mRNA from similar individuals. It would be interesting to investigate the effect of stratifying individuals to construct pools on the precision with which we can estimate QTL effects and on the power of the appropriate statistical test. Given that stratifying individuals by phenotypic value

was advantageous when interest centers in estimating the correlation between gene expression and phenotype, we speculate that the appropriate stratification strategy might also result in more reliable inference within the context of pQTL mapping. Investigating whether this is in fact true would permit deciding which pooling strategy is better suited to the estimation of QTL locations and effect sizes and to identifying trait-related genes.

5.1 References

Cabrera, C.P., Dunn, I., Fell, M., Wilson, P., Burt, D.W., Waddington, D., Talbot, R., Hocking, P.M., Law, A., Haley, C.S., Knott, S. and de Koning D.J. (2006) Application of genetic genomics to a marked QTL in poultry. *8th World Congress on Genetics Applied to Livestock Production.*

Kendziorski C.M., Irizarry R.A., Chen K.S., Haag J.D. and Gould M.N. (2005) On the utility of pooling biological samples in microarray experiments. *PNAS.* **102**:4252.

Shih J.H., Michalowska A.M., Dobbin K., Ye Y., Qiu T.H. and Green J.E. (2004) Effects of pooling mRNA in microarray class comparisons. *Bioinformatics.* **20**:3318.

ACKNOWLEDGEMENT

I would like to express my gratitude to my major professors, Dr. Carriquiry and Dr. Dekkers, for their guidance and tremendous efforts afforded to my research projects. I want to thank my committee members, Dr. Nettleton, Dr. Koehler, Dr. Lamont, and Maitra, for their advice and encouragement. Thanks to Dr. Nettleton for his tremendous involvements in my research. Thanks to Dr. Lamont for providing suggestions for Illumina beads array analysis. Also thanks to Dr. Isaacson and various faculty, staff, and students in the Department of Statistics whose help and assistance will be too numerous to acknowledge in detail. Finally, thank my husband, Yong Chen, and my parents, Jianying Zhang and Jianxin Zhang, for their great supports.