Abstract

YU, XIANG. Statistical Methods for the Analysis of Genetic Marker and Microarray Data (under the direction of DR. BRUCE S. WEIR)

With the advent of high-throughput technologies in genomics study, a large volume of data has been accumulated, leaving the challenge for bioinformaticists on how to manage, analyze, and interpret the data. Analysis of genetic marker and microarray data are two important aspects in current bioinformatics studies. In this dissertation work, we tend to explore some statistical issues for such problems. We discuss two extensions of the EM algorithm to infer haplotypes from genotype data, each for a particular sampling scenario. The first one applies to a random sample of both diploid and haploid individuals from the population, in which the haplotype information from the haploid individuals is incorporated into the estimation process. The second one applies to a sample of parent-offspring trios, in which the dependencies between the parental and the offspring genotypes are correctly handled in the analysis. We show that these two modified EM algorithms perform better than the usual one when applied to their corresponding specific samples, respectively. We study the experimental designs in two-color microarray experiments and resolve some of the outstanding issues that are controversial on the use of different experiment designs. We show that the loop and balanced block designs analyzed in a mixed model are more efficient that the reference designs from a statistical point of view. We also provide general guidelines on how to optimize experimental resources to get maximal efficiency using these designs. We present an application of the mixed model to identify transcription factor-gene interactions and to infer transcriptional

regulatory structures in *Sacchromyces cerevisiae* using microarray experiments. We demonstrate the mixed model that pools the observations across all experiments to be a powerful approach.

STATISTICAL METHODS FOR THE ANALYSIS OF GENETIC MARKER AND MICROARRAY DATA

by

XIANG YU

A dissertation submitted to the Graduate Faculty of
NORTH CAROLINA STATE UNIVERSITY
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

BIOINFORMATICS

RALEIGH
2004

APPROVED BY:

Bruce S. Weir Chair of Advisory Committee	Dahlia M. Nielsen Co-chair of Advisory Committee
Gregory Gibson	Russell D. Wolfinger

To my wife, my parents and my parents-in-law

BIOGRAPHY

Xiang Yu was born in Ganzhou, Jiangxi Province in P. R. China on June 6, 1976. He received a B.S. in bioscience from University of Science and Technology of China, Hefei, P. R. China in 1999. In August 1999, he was admitted to Department of Cell Biology, Albert Einstein College of Medicine of Yeshiva University in Bronx, New York where he received one year's PhD training in molecular and cellular biology. In June 2000, he transferred to the bioinformatics program at North Carolina State University and started working towards a PhD in bioinformatics. In May 2003 he received a Master's degree in bioinformatics. From June 2000 to May 2001, he worked at the Bioinformatics Research Center (BRC) at North Carolina State University as a research assistant and software developer in statistical genetics on a joint-internship with Cimarron Software Inc. in Salt Lake City, Utah. Since May 2001, he worked on an internship at SAS Institute Inc. in Cary, NC, focusing on research and development of statistical methods and software in haplotype analysis, design and analysis of microarray as well as proteomics data.

ACKNOWLEDGEMENT

My sincere gratitude goes to my advisor, Dr. Bruce S. Weir for his support, guidance and encouragement throughout my graduate study. His enthusiasm and deep insights in bioinformatics not only provided direction for my dissertation work but also broadened my knowledge in both statistics and bioinformatics. I have been so fortunate to have had his instructions for my dissertation work. I would like to thank the members of my advisory committee, Dr. Dahlia M. Nielsen, Dr. Gregory Gibson and Dr. Russell D. Wolfinger, for all their healthy discussions and encouragement throughout the years. They have been a great resource for me to get ideas and advice on my research work as well as helpful comments on refining my thesis.

Special thanks to Dr. Tzu-Ming Chu and Dr. Shibing Deng at SAS Institute Inc., who spent a lot of time discussing with me on research problems as well as helping me revise my SAS programs.

I am also deeply grateful to many people at BRC, in particular, Ms. Debra Hibbard, Ms. Juliebeth Briseno, Ms. Alex Rogers for their assistance on all the administrative work to keep me on track, and Dr. Kejun Liu and Mr. Jiaye Yu for their helpful discussions and suggestions on my work.

Finally I would like to express my hearted thanks to my wife Junyu Sun for her wonderful support.

Contents

List of Tables Page

List of Figures Page

Chapter One

Introduction

Genetic marker haplotype

Gene-marker association tests have been largely used in mapping genetic traits in which a genetic marker serves as a physical anchor to locate linked genes that affect the traits. A haplotype is the combination of marker alleles from different loci on one chromosome. Haplotypes can bear more information than a single marker in that several linked loci are transmitted together from the parent to the offspring and that several different loci may contribute synergistically to a particular trait. In recent years, much attention has been paid to haplotype-based association tests. A haplotype can be treated as a superior marker and used in regular contingency-table based single-marker tests. Schaid *et al*. (2002) developed a generalized score test between traits and estimated haplotypes and Zaykin *et al*. (2003) proposed a similar regression approach that models a quantitative or categorical trait on estimated probabilities of haplotypes. Tests based on similarities between haplotypes in cases and controls have also been proposed and prove powerful under some conditions (van der Meulen and te Meerman 1997; Tzeng *et al*. 2003). Haplotype-based tests can sometimes be more powerful than a single-marker approach (Martin *et al*. 2000; Fallin *et al*. 2001) whereas some argue that the controversy holds that single-marker analysis performs better than haplotype-based ones (Long and Langley 1999; Kaplan and Morris 2001). Nielsen *et al*. (2004) discussed the effects of two- and three-locus linkage disequilibrium (LD) on the power of these two approaches and find a haplotype-based test to be more powerful when the three-locus LD is strong.

Contemporary genotyping methods can easily type a pair of alleles at each marker locus at a fairly low cost. Which allele comes from which parent at each locus, however,

remains a question without involving a much more complicated haplotyping procedure. Because molecular haplotyping technique is time- and cost-consuming, people tend to infer haplotypes from genotype data using statistical methods. Depending on whether the data are collected from large, collateral families or from randomly sampled individuals from the population, different statistical methods can be applied accordingly. Of those methods, the EM algorithm is the most popular one that works on population data (Hill 1974; Weir and Cockerham 1979; Excoffier and Slatkin 1995). The EM algorithm starts with compiling a list of compatible haplotypes and assigning initial frequency to each one, evaluates relative probabilities of all combinations of possible haplotype pair for each individual and then updates frequencies for each haplotype, and repeats the evaluation and updating steps until convergence is reached. The EM algorithm has proven accurate in practice by a number of researchers either through phase-known data (Tishkoff *et al*. 2000) or from extensive simulation studies (Fallin and Shork 2000). The weakness of the EM algorithm is its exponential complexity since the number of possible haplotypes grows exponentially with the number of marker and each haplotype needs to be evaluated at each expectation-maximization step. An alternative would be the Bayesian method (Stephens *et al*. 2001; Lin *et al*. 2002) that incorporates the coalescent theory in to a Markov Chain Monte Carlo (MCMC) technique. In the Bayesian algorithm, a Gibbs sampler is used to assign a haplotype-pair to each individual and reassignment at each step is calculated based on the current haplotype frequencies as well as the population mutation rate estimated using the coalescent theory. Because it eliminates the need to evaluate every haplotype, the complexity of the Bayesian algorithm is linear and it can work with a large number of markers.

The EM algorithm assumes Hardy-Weinberg equilibrium (HWE) so that the genotype frequency in the population can be calculated as the product of haplotype frequencies. In a random sample of diploid individuals, haplotypes are unknown and must be estimated. When a sample of both diploid and haploid individuals is collected, for instance in genotyping both males and females on the X chromosome, haplotypes are known for some individuals and can thus provide information in estimating the unknowns. The usual EM algorithm works on the diploid individuals and does not take advantage of haploid ones. In chapter 2 we discuss an extension of the usual EM algorithm that can be applied to a mixture of both diploid and haploid samples. We show that the new EM algorithm improves the accuracy of the haplotype frequency estimates by appropriately incorporating the additional information from the known haplotypes. In another scenario, genotypic data from parent-offspring trios are collected and haplotypes are to be estimated. The usual EM algorithm encounters the difficulty that the samples are not necessarily independent. Only the parents can be regarded as randomly sampled from the population and the offspring genotype depends on its parental genotypic configurations. In addition, because the parental genotype determines which haplotype can be possibly transmitted to the offspring, extra information of haplotypes is embedded in the data on the offspring. In chapter 3 we discuss a new EM algorithm to estimate haplotype frequencies that handles and takes advantage of the dependency between the related samples when parent-offspring trios are collected. The new EM algorithm is superior to the conventional EM algorithm that uses only the independent parents in the sample or that uses all samples but ignores the dependency structure. It is also superior to the

conventional EM algorithm applied to a sample of independent individuals with an equivalent size.

DNA microarray experiments

Over the past decade, a number of high-throughput technologies have been developed and applied to the study of genomic science, of which microarray is a most popular one. The microarray technique was first introduced by Brown's lab in 1995 (Schena *et al*. 1995) in the form of spotted cDNA arrays and then as oliogo-nulceotide chips in 1996 (Lockhart *et al*. 1996). By planting the sequences from thousands of genes on a small glass slide followed by hybridization of the extracted RNA to that slide, microarray can measure the transcription level of those genes simultaneously. Despite its recent emergence, microarray technology has brought about great opportunities in functional genomics studies as well as challenges to draw solid statistical inferences from the typically high-dimension yet few-replication microarray data.

One of the basic questions for microarray experiments is to identify which genes are differentially expressed. In statistical term, this is equivalent to testing the null hypothesis that the levels of RNA transcription are equal between the two samples to be compared. Traditionally, biologists evaluate the amount of RNA in a treatment sample and a reference sample and screen out genes that have a 2-fold change in the transcription level. This approach is biologically intuitive but lacks statistical support. A straightforward statistical approach would be to apply the typical two sample statistics such as the t-statistic to test if a gene is differentially expressed. However, we can not rely on the

distribution assumption for *p*-value calculation owing to the factor that the sample size for each gene is usually small and that the normality assumption may not hold for all genes. By pooling all observations from the data and finding sources of variation that contribute to the transcriptional measurements and formulating a hypothesis test for each gene, the ANOVA model has proven a powerful and flexible method to detect differentially expressed genes (Kerr *et al*. 2000; Wolfinger *et al*. 2001). A bonus of the ANOVA model is that it can also control for other covariates that may affect the gene expression level.

As important as data analysis in a microarray study is design of the experiment. Variation in microarray experiments comes from multiple sources including but not limited to spot size, dye-incorporation and hybridization efficiency. These sources of variation should be accounted for when designing a microarray experiment so that the effect of interest is not confounded with ancillary effects as well maximizing the information obtained while minimizing the overall cost. Traditionally, biologists use a reference design in which the RNA samples from a variety of interest and a common reference are hybridized to the same array. The intensity ratio of the variety vs. the reference is then used to assess differential gene expression. Such a reference design "wastes" half of the observations on the reference sample which is typically of no interest. In their study, Kerr and Churchill (2001) discussed the inefficiency of the reference design and proposed a loop design as an alternative that eliminates the need for the reference sample. Dobbin and Simon (2002), however, came up with a somehow controversial conclusion and recommended a reference design for its simplicity despite some loss of efficiency. In chapter 4, we tend to

resolve this controversy. We compare the loop and the reference design from a mixed model perspective and show that the loop design and balanced block design are better than the reference design in terms of statistical efficiency and overall cost. We also provide some general guidelines on how to optimize experimental resources using these designs in practice.

Microarray technology has been widely used but not limited to identify differentially expressed genes or to classify gene clusters based on their expression profiles. In chapter 5, we discuss an experiment designed to identify protein-DNA interactions simultaneously for thousands of genes on a microarray platform (Lee *et al*. 2002). We discuss an application of the log-linear mixed model to draw proper statistical inferences from the so called "genome-wide location analysis" experiment in which thousands of chromatin immunoprecipitations are conducted at the same time on a microarray platform. We also demonstrate how to assemble transcriptional regulatory motifs based on pairs of transcription factors and gene promoters identified by the mixed model to be directly interacting with each other. The results show that a gene-by-gene mixed model that pools observations across all transcription factors to assess each gene for each transcription factor is much more powerful than a t-test that uses only few observations of that gene for that particular transcription factor.

References

Chu TM, Weir BS, Wolfinger RD (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci 176:35-51

Dobbin K, Simon R (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1438-1445

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143-151

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947-959

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229-239

Kaplan N, Morris R (2001) Issues concerning association studies for fine mapping a susceptibility gene for a complex disease. Genet Epidemiol 20:432-457

Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2:183-201

Kerr MK, Martin M, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comp Biol 7:819-837

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM *et al*. (2002) Transcriptional regulatory network in Sacchromyces cerevisiae. Science 298:799-804

Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. Am J Hum Genet 71:1129-1137

Lockhart DJ, Dong H, Byrne MC, Follettie MT, Gallo MV, Chee MS, Mittmann M, Wang C, Kobayashi M, Horton H, Brown EL (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 14:1675-1680

Long AD, Langley CH (1999) The power of association studies to detect the contributions of candidate genetic loci to variation in complex traits. Genome Res 9:720-731

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM (2000) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet 67:383-394

Nielsen DM, Ehm MG, Zaykin D, Weir BS. 2004. Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. Submitted

Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 70:425-434

Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467-470

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statisitical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518-522

Tzeng JY, Devlin B, Wasserman L, Roeder K (2003) On the identification of disease mutations by the analysis of haplotype similarity and goodness of fit. Am J Hum Genet 72:891-902

van de Meulen MA, te Meerman GJ (1997) Haplotype sharing analysis in affected individuals from nuclear families with at least one affected offspring. Genet Epidemiol 14:915-920

Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105-111

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comp Biol 8: 625-637

Zaykin DV, Westfall PH, Young SS, Karnoub MA, Wagner MJ, Ehm MG (2002) Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. Hum Hered 53:79-91

Chapter Two

On the Use of the EM Algorithm for Haplotype Estimation with Samples Consisting of both Diploid and Haploid Individuals

Introduction

Inferring haplotypes from phase unknown genotype data is a common problem in genetic data analysis. The EM algorithm (Hill 1974; Weir and Cokerham 1979; Excoffier and Slatkin 1995) is the most popular method for haplotype inference that utilizes randomly sampled diploid individuals. In some situations we might have haploid and diploid mixture data, for example, when we sample markers on the X chromosome from both females and males. In such a case, the normal EM algorithm can only be applied to females with a pair of X chromosomes, resulting in under-utilization of data and a biased estimate of the population haplotype frequency. One way to utilize all diploid and haploid data might be to first estimate haplotype frequencies using the diploid individuals and then calculate the pseudo counts of each haplotype by first multiplying the inferred haplotype frequency with chromosome counts from the diploid data and then add the real counts of haplotypes from the haploid data. Such a method ignores the information we might have gained from the haploid data while estimating the diploids. Here we propose a modified EM algorithm for estimation of such mixture data that correctly utilizes all information from the data. Simulation studies showed that this new algorithm performs better than one that applies the EM algorithm to the diploid data only and then calculates the pseudo count of each haplotype. Improvement in the accuracy of haplotype estimates is gained from the new algorithm over the normal EM algorithm applied to a diploid sample with equivalent number of chromosomes because of the information gained from the haploid individuals.

Materials and Methods

1. The EM algorithm

Let g_i denote the i^{th} genotype observed (g_i could be either haploid or diploid), i = 1, 2, …, n. Assume randomly mating, the likelihood of observing the data is then

$$L = \prod_{i=1}^{n} p(g_i)$$

We say that a haplotype h is compatible with a diploid genotype g_d if there exists another haplotype h^c such that the gametic union of h and h^c can constitute g_d. Here h^c is the complementary haplotype of h with respect to the diploid genotype g_d. Obviously, if h is compatible with g_d, then h^c is also compatible with g_d. We use $h \otimes g_d$ to denote that h is compatible with g_d and $h \bar{\otimes} g_d$ otherwise. If g_d has missing alleles, then we define $h \otimes g_d$ if there exists another haplotype h' and the union of h and h' differs from g_d only at the missing alleles. Similarly, we can define compatibility of a haplotype h and a haploid genotype g_h as $h \otimes g_h$ if h and g_h have the same allele at all loci except for missing alleles.

Let h_k denote the k^{th} haplotype in the compatible haplotype set H_j of the genotype g_j. If g_j is haploid, then we have $H_j = \{g_j\}$ for a g_j that bears no missing alleles or $H_j = \{h \mid h \text{ and } g_j \text{ have the same allele at all non-missing loci}\}$ when g_j has missing alleles. Under the assumption of random union of gametes, we can construct an EM algorithm as follows:

1a. E-step: the probability of observing each g_j is

$$p(g_j)^{(t)} = \begin{cases} \sum_{k \in H_j} f(h_k)^{(t)} f(h_k^c)^{(t)}, & \text{if } g_j \text{ diploid} \\ \sum_{k \in H_j} f(h_k)^{(t)}, & \text{if } g_j \text{ haploid} \end{cases}$$

where $f(h_k)^{(t)}$ is the frequency of haplotype h_k at the t^{th} iteration.

1b. M-step: let n_d denote the number of diploids and n_h denote the number of haploids in the sample, new haplotype frequency can be obtained from the genotype frequencies.

$$f(h_k)^{(t+1)} = \frac{1}{2n_d + n_h} \left(\sum_{g_j \text{ is diploid}} \frac{m_{jk} f(h_k)^{(t)} f(h_k^c)^{(t)}}{p(g_j)^{(t)}} + \sum_{g_j \text{ is haploid}} \frac{I(h_k \in H_j) f(h_k)^{(t)}}{p(g_j)^{(t)}} \right)$$

where $m_{jk} = 0$ if $h \bar{\otimes} g$; $m_{jk} = 1$ if $h_k \otimes g_j$ and $h_k \neq h_k^c$ (that is, g_j is heterozygous with respect to h_k) ; $m_{jk} = 2$ if $h_k \otimes g_j$ and g_j is homozygous with respect to h_k.

2. Simulations

We performed simulations to verify the adequacy of our EM algorithm. The simulations are based upon empirical haplotype frequency data. Haplotype frequencies at a 20Kb region on the X chromosome in *Drosophila melanogaster* that includes two genes, *zeste* and *tko*, are obtained by direct counting of 27 isogenic X chromosome lines from the Texas population (Aguade *et al*. 1989). The *zeste-tko* region carries 10 restriction site and 4 insertion/deletion polymorphisms. In each sample, we generate n diploid individuals and n haploid ones. We first sampled $3n$ haplotypes based on their frequencies and then randomly paired two gametes from $2n$ of them to generate n dipoid individuals. The sample of $2n$ individuals thus has n diploids with known phase and n haploids. The 14-site haplotype frequencies were then estimated using our EM algorithm for each sample. In the following discussion, we label our EM algorithm EM^m with the superscript m standing for mixture data. We compared the results from EM^m to four other methods. The

first one uses a two-stage approach by first estimating haplotypes from the diploid data and then calculating pseudo counts of haplotypes and adding real counts of haplotypes from the haploids, which we label as EM^{ts}. The second one uses only the diploid data to get haplotype frequencies and we label it EM^{d} and the third one calculates haplotype frequencies by counting only the haploid data and is labeled EM^{h*}. For each approach, we then used the estimated haplotype frequencies to assign the most probable haplotype pair to each diploid individual. EM^{h*} is marked by a * because it does not actually use an EM algorithm to estimate haplotype frequencies. However, to determine the best phase assignment for diploid individuals, the same calculations as the other three methods are performed in the EM^{h*} approach to get the probabilities of a genotype being resolved to a particular haplotype pair based on estimated haplotype frequencies. For the fourth one, we generated an additional data set consisting of diploid individuals but with an equivalent number of chromosomes. For example, if we generate n diploid individuals and n haploid ones for a mixture data set, we will generate another data set with $1.5n$ diploid individuals. We then applied the normal EM algorithm to this data set and labeled it EM^{e}. Three summary statistics, the similarity index (Excoffier and Slatkin 1995), the phase assignment error rate and the switch accuracy (Lin *et al*. 2002) were recorded.

Results

We performed 1,000 independent simulations and recorded the average of the three summary statistics for the four methods in Table 2.1.

The results show an improvement in overall accuracy of haplotype frequency estimates, measured by similarity index, when EM^{m} is used as compared to EM^{ts}. The improvement becomes more significant when the sample size is small, or there exists missing genotypes in the data set. When error rate and switch accuracy are compared, the two methods appear to give the same results since their haplotype frequency estimates tend to be similar in terms of both the range of each individual haplotype frequency and the distribution of frequencies for all haplotypes. Therefore, the probability of a genotype having a particular haplotype pair tends to be close as well, resulting in the same assignment. The results imply that the two-stage EM algorithm might be a good approximation at a first trial in practice though correct inference should be drawn from the EM algorithm that appropriately incorporates the haploid data into analysis.

The improvement of EM^{m} for all three measures over the other two methods, EM^{d} and EM^{h*}, however, turns out to be quite significant because those two methods fail to utilize all information from the data. Note that when there exists missing data, the error of EM^{h*} is greatly inflated because any haplotype bearing one or more missing alleles is excluded from haplotype counting.

For all three measures, we also see a slight improvement in EM^{m} than EM^{e} applied to an equivalent sample of diploid individuals. This is because the haploid data in the sample have provided additional information on haplotypes we want to estimate.

Overall, we found the accuracy of haplotype frequency estimates can be improved by this modified EM algorithm suitable for the special case where both haploid and diploid individuals are sampled and the information from haploid data is appropriately incorporated into the analysis.

References

Aguade M, Miyashita N, Langley CH (1989) Restriction-map variation at the *zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol Biol Evol 6:123-130

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229-239

Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. Am J Hum Genet 71:1129-1137

Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105-111

Table 2.1: Comparison of four methods for haplotype estimation using diploid-haploid mixture data. Results are from 1,000 simulations.

Method	No. Diploids	No. Haploids	Percent Missing (%)	Similarity Index	Error Rate	Switch Accuracy
EM^{m}	20	20	0.00	0.964	0.0540	0.967
EM^{ts}	20	20	0.00	0.917	0.0540	0.967
EM^{d}	20	20	0.00	0.852	0.0800	0.952
EM^{h*}	20	20	0.00	0.795	0.0130	0.990
EM^{e}	30	0	0.00	0.942	0.0630	0.962
EM^{m}	20	20	5.00	0.934	0.114	0.941
EM^{ts}	20	20	5.00	0.866	0.114	0.941
EM^{d}	20	20	5.00	0.815	0.165	0.913
EM^{h*}	20	20	5.00	0.659	0.743	0.495
EM^{e}	30	0	5.00	0.909	0.125	0.932
EM^{m}	100	100	0.00	0.988	0.0190	0.987
EM^{ts}	100	100	0.00	0.962	0.0190	0.987
EM^{d}	100	100	0.00	0.936	0.0330	0.979
EM^{h*}	100	100	0.00	0.898	0.0150	0.988
EM^{e}	150	0	0.00	0.980	0.0230	0.985
EM^{m}	100	100	5.00	0.979	0.0670	0.969
EM^{ts}	100	100	5.00	0.940	0.0670	0.969
EM^{d}	100	100	5.00	0.928	0.0830	0.959
EM^{h*}	100	100	5.00	0.836	0.714	0.517
EM^{e}	150	0	5.00	0.968	0.0740	0.964

Chapter Three

An EM Algorithm for Haplotype Frequency Estimation

Using Parent-offspring Trios

Abstract

Inferring haplotypes from genotype data is generally accomplished with the EM algorithm for samples of unrelated individuals or by examination of parent-offspring transmissions in pedigrees. The EM algorithm can not be used when there are dependencies among individuals, and pedigree methods are hampered when multiple loci are heterozygous. We propose a modified EM algorithm under the assumption of Hardy-Weinberg equilibrium (HWE) to infer the maximum likelihood estimates of haplotype frequencies using parent-offspring trios, so that we combine the advantages of unrelated samples and family data. By considering only the haplotypes that could possibly have been transmitted from parent to offspring, our EM algorithm can appropriately take into account the family dependencies. This new EM algorithm also handles the common situation of not all loci being typed for all individuals. Simulation studies show the new algorithm to be superior to the usual EM algorithm that uses only independent parents or that uses all individuals but ignores the dependencies.

Introduction

A haplotype, being a natural organization of alleles located at different polymorphic loci on one chromosome, carries more information than single markers that comprise it and may directly reflect inheritance patterns and evolution history (Judson et al. 2000). Inference on haplotypes is always of interest and haplotype analysis has become a particularly important tool in human disease studies. It has been postulated that haplotype-based tests are more powerful than the single-marker association tests in many cases. A haplotype may directly harbor the unobserved disease locus and thus act as a superior marker, or several loci each with a small phenotypic effect could act synergistically to cause the disease. Therefore, it might be more powerful to examine the properties of multi-locus haplotypes in disease gene mapping. The study of association of Alzheimer's disease (AD) and SNPs in the APOE gene region (Martin et al. 2000; Fallin et al. 2001) found that a haplotype-based approach was sensitive in finding disease-predisposing SNPs, whereas standard single marker mapping failed to detect associations for some of the SNPs. Haplotypes have also been shown to provide valuable information in locating disease susceptibility genes in Crohn's disease (Rioux et al. 2001), finding a gene responsible for human lipoprotein lipase (Clark et al. 1998).

Haplotypes can be determined with certain molecular laboratory techniques such as long-range PCR (Michalatos-Beloin et al. 1996), heteroduplex analysis (Chang and Kidd 1997) and sperm-crossover analysis (Jeffreys et al. 2001) etc, but those approaches are usually time- and cost-prohibitive for large scale studies. Statistical methods such as the Expectation-Maximization (EM) algorithm (Hill 1974; Weir and Cockerham 1979) can be used to infer haplotype frequencies with genotypic data collected from the population.

When family data are collected, haplotype configurations can be obtained by tracing parent-offspring gamete transmissions. However, not all haplotypes can be resolved without ambiguity owing to the fact that both the parents and their offspring may be heterozygous at more than one locus. As noted by Hodge et al. (1999), the extent of phase ambiguity increases quickly with the number of loci. Genotyping of additional relatives could reduce the ambiguity, yet it could still be very significant even with a moderate number of loci. To resolve phase with ambiguities, one simple approach might be to first get all unambiguous genotypes and then assign a "best guess" haplotype pair to an ambiguous one based on how frequent these haplotypes are observed from the unambiguous data. The parsimony algorithm proposed by Clark (1990) is an analogous one applied to population data that assigns a single haplotype pair to an ambiguous genotype based upon the currently available haplotype list. Stochastic techniques for haplotype reconstruction have been established for pedigrees that use either simulated annealing with random walk of pedigree sampling (Sobel and Lange 1996) or Hidden Markov Models (HMM) to choose an optimal inheritance vector (Kruglyak et al. 1996). The likelihood based EM algorithm is widely adopted for inferring haplotypes and it is suitable for unrelated individuals where family structures are not incorporated into the analysis. When there are related individuals in the sample, one has to select only the independent ones in order to comply with the conditions of the EM algorithm, resulting in a loss of information. It would be better to combine the population sample with additional information provided by relatives. Boehnke (1991) described a method to estimate allele frequencies that makes use of all available information while correctly incorporating the dependence between relatives. Excoffier and Slatkin (1998) presented

an EM algorithm for solving maximum likelihood estimates for two-locus haplotypes when both siblings and their parents are used in the sample. Rohde and Fuerst (2001) suggested including only the parents in the expectation-maximization steps but their offsprings are used to eliminate unlikely transmissions, assuming the loci are tightly linked and no recombination occurs. Here we extend these studies and propose a modified EM algorithm for haplotype inference that is suitable for parent-child trio analysis. All individuals are used in the EM steps and the assumption of no recombination is relaxed. As an analogy, our algorithm can also be extended to more complex pedigrees and samples consist of both unrelated individuals and related ones.

Materials and methods

Let g_{if}, g_{im}, g_{io} denote the father, mother and offspring observed in the i^{th} trio, i = 1, 2, …, n. Assuming random mating and independent sampling of trios from the population, the likelihood of observing the data, following the general framework of pedigree data analysis (Elston and Stewart 1971) is

$$L = \prod_{i=1}^{n} p(g_{io} \mid g_{if}, g_{im}) p(g_{if}) p(g_{im}) .$$

Method of Excoffier and Slatkin (1998)

For two-locus haplotypes, assuming there is no recombination, Excoffier and Slatkin (1998) considered all family members jointly and wrote the likelihood for each family as

$$L_i = \sum_j g_{fj} g_{mj} s_j ,$$

where the summation is taken over all possible genotype combinations of the members for family i. s_j is the Mendelian segregation probability of the offspring's genotype combination j. Under the assumption of Hardy-Weinberg equilibrium, the genotype frequencies of g_{fj} and g_{mj} can be written as product of haplotype frequencies for each possible haplotype pairs (h_{1f}, h_{2f}) and (h_{1m}, h_{2m}). The segregation probability $s_j = 1/4$ if the offspring's genotype is consistent with one of the four possible gametic compositions $\{h_{1f}h_{1m}, h_{1f}h_{2m}, h_{2f}h_{1m}, h_{2f}h_{2m}\}$ and $s_j = 0$ otherwise. The EM algorithm starts with some initial frequency assignment for all haplotypes and then re-evaluates the haplotype frequencies from the relative probabilities of the combination of parental and offspring genotypes. In his program TRANSMIT, Clayton (1999) uses a similar algorithm to

calculate haplotype frequencies from nuclear families but discards uninformative families (without any affected offspring) and families that have too many possible combinations.

Method of Rhode and Fuerst (2001)

The algorithm described by Rhode and Fuerst (2001) uses only the parents from the sample in the conventional expectation and maximization steps. However, the offspring is used to eliminate impossible combinations of parental genotypes where it fails to inherit one gamete from each parent, assuming no recombination occurs in the loci studied.

The trio EM algorithm

For a genotype g, we define a haplotype h_g to be compatible with g if there exists another haplotype h_g^c such that the gametic union of h_g and h_g^c can constitute g. Here h_g^c is the complementary haplotype of h_g with respect to genotype g. Obviously, if h_g is compatible with g, then h_g^c is also compatible with g. We use $h \otimes g$ to denote that a haplotype h is compatible with g and $h \overline{\otimes} g$ otherwise.

For an offspring, not only should a haplotype h be compatible with the offspring, but also h and its complementary haplotype h_g^c should be transmitted from its parents and therefore be compatible with at least one of these parents. Compatibility of haplotype h with an offspring genotype g and the corresponding parental genotypes g_f and g_m, is then defined as:

$$h \otimes (g, g_f, g_m) \text{ if } h \otimes g \text{ and } \left(\text{either } h \otimes g_f, h_g^c \otimes g_m \text{ or } h \otimes g_m, h_g^c \otimes g_f\right)$$

Under the assumption of random union of gametes, we can construct an EM algorithm as follows:

1. E-step: the probability of observing a genotype g is

$$p(g)^{(t)} = \begin{cases} \sum_{h \otimes g} f(h)^{(t)} f(h_g^c)^{(t)}, & \text{if } g \text{ is a parent} \\ \sum_{h \otimes (g, g_f, g_m)} f(h)^{(t)} f(h_g^c)^{(t)}, & \text{if } g \text{ is an offspring and } g_f \text{ and } g_m \text{ are the parents} \end{cases}$$

where $f(h)^{(t)}$ is the frequency of haplotype h at the t^{th} iteration. For an offspring g, the summation is taken over all haplotype pairs h and h_g^c that can be possibly transmitted from its parents g_f and g_m.

1. M-step: for each haplotype h, new haplotype frequency can be obtained from the genotype frequencies.

$$f(h)^{(t+1)} = \frac{1}{2n} \sum_g \frac{m_h \delta_h f(h)^{(t)} f(h_g^c)^{(t)}}{p(g)^{(t)}}$$

where m_h is the number of times h appears in genotype g. $m_h = 0$ if $h \overline{\otimes} g$; $m_h = 1$ if $h \otimes g$ and $h \neq h_g^c$ (that is, g is heterozygous with respect to h); $m_h = 2$ if $h \otimes g$ and g is homozygous with respect to h. δ_h is an indicator variable defined as

$$\delta_h = \begin{cases} 1, & \text{if } g \text{ is a parent or if } g \text{ is an offspring and } h \otimes (g, g_f, g_m) \\ 0, & \text{otherwise} \end{cases}$$

Partial data

An individual's *k*-locus genotype is called partial if not all alleles at the *k* loci are scored. The EM algorithm maintains a list of all possible haplotypes that can be obtained from the genotype data and these haplotypes are fully scored with no missing alleles. During the iterations, a partial genotype or haplotype is handled by considering all possible configurations of its missing alleles. In the E-step, the frequency of a genotype g with missing alleles will be the summation over all compatible haplotype pairs that can

constitute genotypes differing from g only at the missing alleles. Correspondingly, in the M-step, a haplotype h that is compatible with a partial genotype may bear some missing alleles. Such a haplotype is not considered as a new haplotype, but rather all existing haplotypes with alleles identical to the non-missing ones of h and are compatible with observed parent-offspring structure will get updated.

Simulations

Simulations were performed to verify the adequacy of this EM algorithm. The simulations are based upon empirical haplotype frequency data. Haplotype frequencies at the D4S10 locus in 3 different populations, African American, Finn and Maya, were downloaded from the ALFRED web site (Osier et al. 2001). The D4S10 locus carries 5 restriction site polymorphisms spanning over 15.3Kb. In order to generate N trios, we first sampled $4n$ haplotypes based on their frequencies and randomly paired two gametes to generate $2n$ individuals with phase known genotypes. The $2N$ individuals were then randomly paired and each of the pairs generated one offspring. The N parents-offspring trios were recorded as our input data. We assumed no recombination across the five sites during the mating and generate samples of 25 or 100 trios.

We first use the similarity index, I_f, to measure the overall accuracy of the haplotype frequency estimates (Excoffier and Slatkin 1995).

$$I_f = 1 - \frac{1}{2}\sum_i |\hat{f}_i - f_{0i}|,$$

where $\hat{f}_i$ is the estimated haplotype frequency of the i^{th} haplotype and f_{0i} is the sample frequency. The summation is taken over all true and inferred haplotypes and I_f takes a value between 0, where no haplotype is correctly inferred, and 1, for which the true

haplotype frequencies and the estimates are equal. I_f measures the overall absolute deviations of the estimates from their true frequencies. Similar measurement has been used in several other studies (Stephens et al. 2001; Single et al. 2002).

To assess the accuracy of haplotype reconstruction for each of the individuals, we assigned the most probable haplotype pair to each individual and compared the assignment to its true gamete phase. Similar studies have been adopted by Stephens et al. (2001) and Zhang et al. (2001). Once the EM algorithm has arrived at MLEs for the haplotype frequencies, the probability that a genotype g consists of a particular compatible haplotype h and its complementary haplotype h_g^c is:

$$\Pr(h, h_g^c \mid g) = \frac{2^{m_h \delta_h} h h_g^c}{\Pr(g)}$$

where $m_h = 0$ if $h = h_g^c$ and $m_h = 1$ if $h \neq h_g^c$ indicates if g is heterozygous with respect to h and δ_h is another indicator variable showing if h can be transmitted from either parent. The quantity δ_h is defined as:

$$\delta_h = \begin{cases} 0, \text{ if } g \text{ is an offspring and } \left(\text{either } h \otimes g_m, h \,\overline{\otimes}\, g_f \text{ or } h \otimes g_f, h \,\overline{\otimes}\, g_m\right) \\ 1, \text{ otherwise} \end{cases}$$

For each individual, all compatible haplotype pairs are evaluated and the one with the largest probability is assigned. The error rate for phase identification is then recorded as the proportion of individuals with incorrect phase assignment.

For a large number of markers, the exact linkage phases are less likely to be correctly inferred completely for each individual but the proportion of haplotypes that are correctly reconstructed can be scored. The switch accuracy (SA), measures the number of switches required to recover the original haplotype phases for each individual (Lin et al. 2002):

$$SA = \frac{n-1-sw}{n-1}$$

n is the number of heterozygous sites and sw is the number of crossover events between neighboring heterozygous sites needed to recover the true gamete phases. The switch accuracy we reported is the average over all individuals in the data set.

For each data set, we employed our trio EM algorithm, the usual EM algorithm with only the parents, the usual EM algorithm using all data. We also implemented the algorithm of Rhode and Fuerst (2001) and extended it to be able to handle partial data. We applied this algorithm to our data set, as well. For each simulation, another data set with equal size of independent individuals was generated and the usual EM algorithm was applied on it. Partial data were generated by randomly omitting a given proportion of genotypes uniformly across all markers. We performed 100 independent simulations for each population and recorded the three summary statistics for the five approaches indicated above.

Results

In the following discussion, we use EMT to denote our trio EM algorithm, EMP for the usual EM algorithm that uses only the sample of independent parents, EMA for the usual EM algorithm using all individuals but ignoring parent-child dependency, EMP* for the Rhode and Fuerst (2001) algorithm, and EME for the usual EM algorithm with an equivalent sample of independent individuals.

The similarity index values are summarized in Table 1. It can be seen that, in general, our EMT algorithm improves accuracy for haplotype frequency estimates over all four other methods. The slight improvement of EMT over EMP* is due to a larger sample size that results from using all information. Although EMA performs similarly to EME, it has a slightly larger variance due to inappropriate incorporation of dependent relatives (data not shown). Among the three populations, the accuracy is worst for African American and best for Finnish and that of Mayan falls between those two. This is because African American is the most diversified population. There are 16 different haplotypes for the five markers and five of them have frequencies above 0.05. The most abundant haplotype has a frequency of 0.453. The Finnish sample population has only 10 haplotypes, four of which have frequencies > 0.05 and the most abundant one is 54.2%. The diversity of the Mayan population is between African American and Finns with 14 distinct haplotypes and six of which > 0.05. We can see that the accuracy of haplotype estimates tends to decrease when the true frequency becomes more evenly distributed, as discussed by Fallin and Schork (2000).

Comparisons of phase reconstruction error rates are shown in Table 2. Among the five methods, EMP* is the best to reconstruct haplotypes with the smallest error rate in all

cases. However, it should be noted that EMP* uses information from the offspring to reconstruct the parental phases but the error rate was calculated based solely on parental phases. The error rate is expected to be larger if the offspring's phase were to be assigned thereafter. Comparison of EMT to EMP* shows a higher error rate for EMT, which is due to allowance of recombination in the loci studied but no crossover happens when we generate the data. The discrepancies between EMT and EMP* will be smaller if all individuals were phased accordingly with the EMP* method. For the remaining four methods, EMT gives the best result. EME performs similarly to EMA but the latter has a slightly larger variance (data not shown). The result that EMA performs similarly or even slightly better than EME may be an artifact due to overestimation of information resulted from the sampling process. When trios are sampled, the genotype of the child tends to be similar to those of the parents and thus more weight will be given to the over-represented haplotypes among the trios when their phases are reconstructed. However, the improvement of EMT over EME turns out to be significant. This is due to the employment of related individuals that provide more information on gamete transmissions. The patterns for the three populations are the same as similarity index, that is, African American has the largest error rate and Finnish has the smallest.

The switch accuracies are recorded in Table 3. The results are similar to the error rate. EMT still outperforms EMP, EMA and EME, but slightly worse than EMP*. The patterns across the three populations remain the same, as well.

Discussion

The well known EM algorithm (Excoffier and Slatkin 1995; Hawley and Kidd 1995; Long et al. 1995) for the ML solution of haplotype frequencies was first introduced by Hill (1974) by extending the gene counting procedure (Ceppellini et al. 1955) to multi-locus chromosome counting. Weir and Cockerham (1979) gave analytical solutions of the EM algorithm for two-locus haplotypes. The computation intensity of the EM algorithm grows exponentially when more loci are involved and numerical results are not easy to obtain without fast computers. The EM algorithm, while widely adopted for haplotype inferences, is suitable only for population data. In this study, we have proposed an EM algorithm for reconstruction of haplotypes from parent-offspring trios. The parent-child dependency is taken into consideration by restricting a gamete to be compatible with both parent and child. As a comparison, the usual EM algorithm has to select only the unrelated individuals from each family or treat the samples as if they are unrelated, both being unable to appropriately incorporate all information from the samples.

In this EM algorithm, the evaluation of compatibility allows recombination to occur when parents are forming their gametes. In our simulation study we assumed no recombination when we generate trios using the D4S10 data set. As a result, the most probable phase assignment for the offspring may indicate that a recombination event has occurred for its father or mother or both, which turns out to be an incorrect assignment. However, such errors do not seem to be significant. Our simulation results show that even the condition of no recombination is not satisfied, the trio EM algorithm still gives better results, as compared to methods that use only parents or that ignore dependencies. It also

performs better than the usual EM algorithm with an equivalent sample size of unrelated individuals.

There is no single measurement to evaluate the performance of the EM algorithm. Three different metrics, the similarity index, the haplotype reconstruction error rate and the switch accuracy are used to assess the performance of the algorithm. Overall, we found our trio EM algorithm gives the best results in frequency estimates and performs well in phase assignment, as well. For the three different populations, African American, Finnish and Mayan, the patterns for all three measures are similar, with Finnish yielding the best results due to lowest diversity.

Although the current framework for the EM algorithm is designed for parent-child trios, we can adopt it to more complex family structure by a similar consideration of compatibilities between a haplotype pair and an individual's as well as its parents' genotypes. Another direct extension will be to samples consist of both unrelated individuals and related ones, in which the usual expectation-maximization iterations are taken over unrelated individuals whereas in each EM step, more stringent compatibility assessments are applied to individuals whose parents are also ascertained.

Several previous studies (Excoffier and Slatkin 1998; Clayton 1999; Rhode and Fuerst 2001) have also addressed the problem in estimation of haplotypes when relatives are included in the sample. All of these methods assume no recombination and consider the combination of all parental and offspring genotypes in each family jointly. In so doing, the expectation and maximization steps are very laborious when more markers are included and/or the degree of polymorphisms increases at each marker because the possible combination of parental and offspring genotypes soon becomes very large.

Consider an example trio where both the parents and the offspring are heterozygous at k biallelic loci, the number of possible haplotypes is thus 2^k for each of them. At each expectation and maximization step, the total combination of haplotypes that needs to be considered for this family will be $2^k * 2^k * 4 = 2^{2k+2}$. In our algorithm, the parents are treated the same way as the usual EM algorithm and compatibility assayed twice more for each offspring (with its parents) in each EM step. For the above example, the total number of haplotypes to be traversed for this family is $2^k + 2^k + 2^k = 3 * 2^k$. Note that extra evaluation steps for compatibility are required for the 2^k part of the offspring, but that is not the bottleneck of the computation. Therefore, practically, the complexity remains the same as the usual EM algorithm but the sample size is inflated. Since it is the number of markers (in essence the number of possible haplotypes which grows exponentially) that dominates in the computational intensity rather than the sample size which increases the complexity linearly, our algorithm may be more feasible in practice. As we have demonstrated, relaxation of assumption in recombination does not affect the estimates substantially, even if all the loci are in fact tightly linked.

Acknowledgements

We thank the Kidd group in Yale University School of Medicine for publishing their data in the ALFRED data base. This research was supported in part by National Institutes of Health grant GM45344.

References

ALFRED website, http://alfred.med.yale.edu/alfred/index.asp

TRANSMIT program, http://www-gene.cimr.cam.ac.uk/clayton/software

Boehnke M (1991) Allele frequency estimation from data on relatives. Am J Hum Genet 48:22-25

Ceppellini R, Siniscalco M, Smith CAB (1955) The estimation of gene frequencies in a random mating population. Ann Eugen 20:97-115

Chang F-M, Kidd K (1997) Rapid molecular haplotyping of the first exon of the human dopamine receptor D4 gene by heteroduplex analysis. Am J Med Genet 74:91-94

Clark AG (1990) Inference of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111-122

Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. Am J Hum Genet 63:595-612

Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. Nature Genet 29:229-232

Excoffier L, Slatkin M (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol Biol Evol 12:921-927

Fallin D, Schork NJ (2000) Accuracy of haplotype frequency estimation for biallelic loci, via the expectation-maximization algorithm for unphased diploid genotype data. Am J Hum Genet 67:947-959

Fallin D, Cohen A, Essioux L, Chumakov I, Blumenfeld M, Cohen D, Schork NJ (2001) Genetic analysis of case/control data using estimated haplotype frequencies: application to APOE locus variation and Alzheimer's disease. Genome Res 11:143-151

Hawley ME, Kidd KK (1995) HAPLO: a program using the EM algorithm to estimate the frequencies of multi-site haplotypes. J Heredit 86:409-411

Hill WG (1974) Estimation of linkage disequilibrium in randomly mating populations. Heredity 33:229-239

Hodge SE, Boehnke M, Spence MA (1999) Loss of information due to ambiguous haplotyping of SNPs. Nature Genet 21:360-361

Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genet 29:217-222

Kruglyak, Daly, Reeve-Daly, Lander ES (1996) Parametric and nonparametric linkage analysis: a unified multipoint approach. Am J Hum Genet 58:1347-1363

Lin S, Cutler DJ, Zwick ME, Chakravarti A (2002) Haplotype inference in random population samples. Am J Hum Genet 71:1129-1137

Long JC, Williams RC, Urbanek M (1995) An E-M algorithm and testing strategy for multiple-locus haplotypes. Am J Hum Genet 56:799-810

Martin ER, Lai EH, Gilbert JR, Rogala AR, Afshari AJ, Riley J, Finch KL, Stevens JF, Livak KJ, Slotterbeck BD, Slifer SH, Warren LL, Conneally PM, Schmechel DE, Purvis I, Pericak-Vance MA, Roses AD, Vance JM (2000) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease. Am J Hum Genet 67:383-394

Michalatos-Beloin S, Tishkoff SA, Bentley KL, Kidd KK, Ruano G (1996) Molecular haplotyping of genetic markers 10 kb apart by allele-specific long-range PCR. Nucleic Acids Res 24:4841-4843

Osier MV, Cheung KH, Kidd JR, Pakistis AJ, Miller PL, Kidd KK (2001) ALFRED: an allele frequency database for diverse populations and DNA polymorphisms-an update. Nucleic Acids Res 29:317-319

Rioux JD, Daly MJ, Silverberg MS, Kinland K, Steinhart H, Cohen Z, Delmonte T et al. (2001) Genetic variation in the 5q31 cytokine gene cluster confers susceptibility in Crohn disease. Nature Genet 29:223-228

Single RM, Meyer D, Hollenbach JA, Nelson MP, Noble JA, Erlich HA, Thomson G (2002) Hapltoype frequency estimation in patient populations: the effect of departures from Hardy-Weinberg proportions and collapsing over a locus n the HLA region. Genet Epidemiol 22:186-195

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58:1323-1337

Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978-989

Tishkoff SA, Pakstis AJ, Ruano G, Kidd KK (2000) The accuracy of statisitical methods for estimation of haplotype frequencies: an example from the CD4 locus. Am J Hum Genet 67:518-522

Weir BS, Cockerham CC (1979) Estimation of linkage disequilibrium in randomly mating populations. Heredity 42:105-111

Zhang S, Pakistis AJ, Kidd KK, Zhao H (2001) Comparisons of two methods for haplotype reconstruction and haplotype frequency estimation from population data. Am J Hum Genet 69:906-912

Table 3.1: Comparisons of similarity index for the five algorithms

Population	*N*	% Missing Genotypes	Similarity Index				
			EMT	EMP	EMP*	EMA	EME
African	25	0	.91	.79	.90	.85	.85
American	25	.05	.88	.76	.86	.82	.81
	100	0	.96	.91	.95	.93	.92
	100	.05	.94	.89	.94	.92	.91
Finn	25	0	.94	.87	.93	.91	.92
	25	.05	.92	.85	.91	.89	.88
	100	0	.98	.94	.96	.96	.96
	100	.05	.96	.93	.95	.95	.95
Maya	25	0	.92	.84	.90	.90	.88
	25	.05	.89	.80	.88	.86	.86
	100	0	.96	.93	.95	.95	.95
	100	.05	.95	.91	.94	.93	.93

Table 3.2: Comparisons of error rate for the five algorithms

Population	*N*	% Missing Genotypes	Error Rate				
			EMT	EMP	EMP*	EMA	EME
African	25	0	.090	.14	.029	.13	.14
American	25	.05	.15	.20	.098	.19	.19
	100	0	.081	.13	.024	.13	.14
	100	.05	.14	.18	.083	.17	.17
Finn	25	0	.046	.083	.015	.070	.070
	25	.05	.091	.13	.058	.11	.11
	100	0	.040	.066	.010	.061	.060
	100	.05	.077	.10	.051	.099	.10
Maya	25	0	.082	.12	.036	.10	.11
	25	.05	.14	.19	.10	.17	.18
	100	0	.071	.11	.032	.11	.11
	100	.05	.13	.17	.096	.16	.16

Table 3.3: Comparisons of switch accuracy for the four algorithms

Population	*N*	% Missing Genotypes	Switch Accuracy				
			EMT	EMP	EMP*	EMA	EME
African	25	0	.90	.84	.97	.85	.84
American	25	.05	.87	.81	.93	.82	.82
	100	0	.90	.85	.97	.85	.84
	100	.05	.88	.83	.94	.83	.83
Finn	25	0	.96	.92	.98	.93	.93
	25	.05	.93	.90	.97	.91	.91
	100	0	.96	.93	.99	.94	.94
	100	.05	.95	.92	.97	.92	.92
Maya	25	0	.95	.92	.98	.93	.92
	25	.05	.92	.89	.95	.90	.90
	100	0	.95	.93	.98	.93	.93
	100	.05	.93	.91	.96	.91	.91

Chapter Four

Why Loop and Balanced Block Designs are Better than Reference Designs for Two-Color Microarray Experiments

Abstract

Experimental design is an important aspect in microarray studies and maximal efficiency can be obtained from a properly designed experiment. Traditionally a cDNA microarray experiment uses a reference design in which the RNA samples from a variety of interest and a common reference are hybridized to the same array. Reference designs are biologically intuitive when fold-change of intensity is used to assess differentially expressed genes. Loop designs and balanced block designs have been recently proposed as more efficient alternatives. In this study we compare the efficiency of these designs under a mixed model framework and provide general guidelines on how to use these designs in practice. The loop design analyzed in a mixed model context is shown to be superior to a reference design in terms of statistical efficiency and power as well as overall cost, regardless of the size of the loop or how far apart two varieties to be compared is in the loop. We illustrate from a previously studied mouse data that the cost efficiency can be 4~5 times greater when using the loop design rather than the reference design.

Introduction

With its capability to assess the expression profiles of thousands of genes simultaneously, microarray technology has become an increasingly important tool in genetical studies (Brown and Botstein, 1999). Microarrays have been used in a wide variety of problems such as to identify whether the gene expression profiles differ among different classes (Jin *et al.* 2001), to discover clusters of genes that are co-regulated (Chu *et al.* 1998; Eisen *et al.* 1998; Spellman *et al*. 1998), and to predict the class of new samples based upon the similarity of their expression profiles with a precompiled multi-gene discriminant (Golub *et al*. 1999).

Microarray studies are factorial experiments in which several different factors contribute to the variation of the fluorescence intensities measured at each spot. These factors should be taken into account accordingly when analyzing the data and an appropriate experimental design can maximize the information obtained from the experiment while minimizing the cost. A traditional microarray experiment applies two RNA samples labeled with different dyes, one from a treatment of interest and the other from a common reference that serves as a baseline, to the same array on which thousands of different genes are spotted. In such a reference design, fold-change in the transcriptional level of the treatment vs. the reference is used to assess whether a gene is differentially expressed. The reference design is biologically intuitive but lacks efficiency since half observations are from the reference that is typically of no interest. Kerr and Churchill (2001) studied experimental designs in microarrays under an ANOVA model (Kerr *et al*. 2000) and pointed out the reference design being inefficient and proposed a loop design as an alternative. Dobbin and Simon (2002) used a similar fixed effects ANOVA model but

allowed gene-to-gene variability to be different and compared the efficiencies of different experimental designs when the purpose of the microarray experiment is to identify differentially expressed genes or to profile genes into clusters according to their expression patterns. In spite of some efficiency loss, they recommend a reference design so some controversy still exists. This paper attempts to resolve some of the outstanding issues.

For a large number of microarray experiments, our main interest is to compare gene expression levels in v varieties and to find out which genes are differentially expressed in which varieties, aka class comparison. Replication of a microarray experiment helps obtain reliable estimates of the expression differences (Lee *et al*. 2000). There are typically three levels of replication that appear in a microarray experiment (Churchill 2003). The first level is biological replication where different samples or different batches within the population of a variety are drawn. The second level is technical replication where each sample is subdivided into different aliquots and each applied to a different array. The third level is subsampling of each gene on the same array. Each level of replication has its own variation and this variation should be taken into account appropriately in design and analysis of a microarray experiment. In this study we follow the gene-by-gene mixed model framework from Wolfinger *et al*. (2001) and Chu *et al*. (2002) for its generality and flexibility in accommodating various types of studies. A major difference of our model from that of Kerr and Churchill (2001) and Dobbin and Simon (2002) is that we treat the array effect to be random.

Results

1. Mixed model for two-color microarray experiments

We consider only balanced cases in which each factor level has the same number of replicates. Suppose we have *v* varieties of interest, for each variety we collect *p* random samples, and on each array, each gene appears *s* times. For each gene *g*, let y_{ijklm} denote the normalized, log base-2 transformed intensity from the i^{th} variety, j^{th} sample, k^{th} array, l^{th} sub-sample on that array, and m^{th} dye (m = 1, 2 for Cy3 and Cy5). Following Wolfinger *et al*. (2002), we consider a simple mixed model for each gene that involves two fixed effects and four random components:

$$y_{ijklm} = \mu + V_i + P_{j(i)} + A_{k(ij)} + S_{l(ijk)} + D_m + \varepsilon_{ijklm} \qquad (1)$$

The variety effect *V* and dye effect *D* effect are considered to be fixed effect and the sample effect *P,* array effect *A*, and subsample effect *S* are considered to be random. It is reasonable to assume that the variation in subsampling is uniform across the whole experiment. Therefore we have $P_{j(i)} \sim N(0,\sigma_i^2)$ for variety *i,* $A \sim N(0,\sigma_A^2)$, $S \sim (0,\sigma_S^2)$ and the stochastic error $\varepsilon \sim N(0,\sigma^2)$. In addition, *P*, *A*, *S*, and ε are mutually independent. Variability at the sampling and the subsampling level are different and normally $\sigma_i^2 >> \sigma_S^2$.

The mixed model specified by (1) can be written in a matrix form:

$$Y = X\beta + Z\gamma + \varepsilon \ ,$$

in which *X* and *Z* are the design matrices for the fixed and random effects, respectively, and β and γ are the parameter vectors of the corresponding fixed and random effects. For calculation ease, we use a cell-means parameterization in model (1):

$$\beta = (\mu_1, ..., \mu_v, d_{\text{cy3}}, d_{\text{cy5}})^T .$$

The variance of the random effects is given by:

$$Var\begin{pmatrix}\gamma \\ \varepsilon\end{pmatrix} = \begin{pmatrix}G & 0 \\ 0 & R\end{pmatrix}.$$

The variance of Y is:

$$V = ZGZ^T + R .$$

The generalized least squares (GLS) estimator of β is given by:

$$\hat{\beta} = (X^T V^{-1} X)^- X^T V^{-1} Y \qquad (2)$$

and its variance is:

$$Var(\hat{\beta}) = (X^T V^{-1} X)^- \qquad (3)$$

Suppose our main interest is to compare the effect between two different levels of variety $v1$ and $v2$, this hypothesis can be written as:

$$H_0 : \theta = c^T \beta = 0 ,$$

and the variance of θ is given by:

$$Var(\theta) = c^T (X^T V^{-1} X)^- c \qquad (4)$$

By the nature of a microarray experiment, each array can only have two different varieties and these two varieties are labeled with two different dyes. Additionally, different sub-samples within the same array are correlated but are independent in different arrays. The variance-covariance structure of the response y_{ijklm} is then given by:

1) for each observation y_{ijklmn} :

$$\sigma^2_{Y,i} = \sigma^2_i + \sigma^2_A + \sigma^2_S + \sigma^2 \qquad (5.1)$$

2) for the same variety, same sample, spotted on the same array in different sub-samples:

$$\sigma_{ss,i} = Cov(y_{ijklm}, y_{ijkl'm}) = \sigma_i^2 + \sigma_A^2, \quad l \neq l' \qquad (5.2)$$

3) for two varieties on the same array (at the same spot) labeled with different dyes:

$$\sigma_{aa} = Cov(r_{ijklm}, r_{i'jklm'}) = \sigma_A^2 + \sigma_S^2, \quad i \neq i' \ \& \ m \neq m' \qquad (5.3)$$

4) for other combinations of *i*, *j*, *k*, *l*, *m* as long as the two observations are on two different arrays:

$$Cov(r_{ijklm}, r_{ijk'lm}) = 0, \quad k \neq k' \qquad (5.4)$$

For class comparison experiments, our primary interest is to compare the differences in expression level for this gene *g* across different varieties, that is, we are interested in the difference $d_{v1,v2} = V_{v1} - V_{v2}$ between varieties *v*1 and *v*2 and the estimator for this difference is denoted $\hat{d}_{v1,v2}$.

2. Reference design

Figure 1A illustrates the layout of reference design of *v* arrays for *v* varieties, in which each variety has only one sample. When more replicates are employed, the same experimental layout is repeated for each full replication. Note that half of the observations are from the reference sample which is of no interest for class comparison between varieties.

For a reference design, each variety is labeled with one dye and a reference sample is labeled with another dye. So the dye effect is completely confounded with variety effect and for the simplest case of *v* arrays for *v* varieties, the model reduces to:

$$y_{ijkl} = \mu + V_i + P_{j(i)} + A_{k(ij)} + S_{l(ijk)} + \varepsilon_{ijkl} \qquad (6)$$

Comparison between two different varieties in a reference design, however, is made within the same dye so the confounding of dye effect does not affect the estimator $\hat{d}_{v1,v2}$. The variance of $\hat{d}_{v1,v2}$, as derived in the appendix, is given by:

$$Var(\hat{d}_{v1,v2}) = \frac{1}{p}\sigma_{v1}^2 + \frac{1}{p}\sigma_{v2}^2 + \frac{2}{p}\sigma_A^2 + \frac{2}{ps}\sigma_S^2 + \frac{2}{ps}\sigma^2 \qquad (7)$$

3. Levels of replication

Table 4.1 summarizes the variance of $\hat{d}_{v1,v2}$ in a reference design under different combinations of biological replication, technical replication, and sub-sampling when the total number of observations per gene is fixed at *pqs*. From table 4.1, we can see that when samples are relatively easy and inexpensive to obtain, increasing the number of biological replicates (and thus the total number of arrays) tends to be more efficient by decreasing the variance $Var(\hat{d}_{v1,v2})$ than increasing the number of technical replicates per sample because both the sample-to-sample variation term (on σ_{v1}^2 and σ_{v2}^2) and the array-to-array one (on σ_A^2) are reduced. This can be particularly important when sample-to-sample variability is large. Applying *s* subsamples per array decreases the residual variance by *s*-fold at the cost of increasing the array preparation effort. In practice, increasing the number of sub-samples per array is generally not necessary. Since the array-to-array variability is usually the dominant random component in a microarray experiment, when samples are relatively harder to collect, adding technical replicates for each sample is beneficial. Although these conclusions are drawn from the reference design, the results will be similar in other designs. In the following sections, we shall

discuss cases without subsampling for simplicity, in which the covariance between two observations at the same spot changes to σ_A^2 in (5.3).

4. Loop design

Figure 1B illustrates a loop design for v varieties and v arrays. The loop design puts variety i and i+1 on array i, for i = 1, …, v-1 and closes the loop with variety v-1 and v on array v. Note it is not necessarily to arrange the v arrays in this particular order as long as the loop is established. The dye labeled on each variety is reversed on two consecutive arrays and note that 2 samples per variety are employed in these v arrays. A simplified diagram for a loop design is shown in Figure 2A in which the v = 6 varieties are connected in a circular fashion, and each connected pair represents an array on which the "from" variety is labeled with dye1 and the "to" variety is labeled with dye 2. When more replicates (each full replication involves 2v samples) are employed, the simplest design is to repeat the same loop but reverse the direction with each full replication. Figure 2B illustrates such a loop design with 2 full replicates (4v samples are used). We consider this scenario where each loop is exactly the same except the dye labeling direction. In this way we implicitly require the number of replicates p for each variety to be even. When the same number of arrays is used in a reference design and a loop design, the number of observations per variety is doubled in a loop design. Let us suppose we double the biological replicates so that a total of 2p samples for each variety are used.

When the values of the random components σ_i^2 (i = 1, …, v), σ_A^2 and σ^2 are known, we can calculate $Var(\hat{d}_{v1,v2})$ using (4) but $Var(\hat{d}_{v1,v2})$ generally does not have a simplified analytical form. However, we can still obtain an upper bound of $Var(\hat{d}_{v1,v2})$ from the

variance of the ordinary least square estimator (OLS), $Var(\tilde{d}_{v1,v2})$ since $Var(\hat{\theta}_{GLS}) \le Var(\hat{\theta}_{OLS})$ (Searle 1971). Define two varieties to be "connected" in a loop design if they appear on a same array and likewise two varieties are "unconnected" if they never appear on the same array. $Var(\tilde{d}_{v1,v2})$ is given by:

$$Var(\tilde{d}_{v1,v2}) = \begin{cases} \frac{1}{2p}\sigma_{v1}^2 + \frac{1}{2p}\sigma_{v2}^2 + \frac{1}{2p}\sigma_A^2 + \frac{1}{p}\sigma^2, & \text{for connected } v1 \text{ and } v2 \\ \frac{1}{2p}\sigma_{v1}^2 + \frac{1}{2p}\sigma_{v2}^2 + \frac{1}{p}\sigma_A^2 + \frac{1}{p}\sigma^2, & \text{for non-connected } v1 \text{ and } v2 \end{cases} \quad (8)$$

5. Balanced block design

In a balanced block design, each array is treated as a block of size two and samples from two different varieties are hybridized to the same array. In addition, balance requires that all varieties appear the same number of times across all arrays in the balanced block design. Randomization is usually performed in assigning varieties to different arrays. Unless there are only two varieties, such a design will be a randomized incomplete block design since not all varieties can appear in the same block. We can view a randomized balanced block design as a shuffled loop design in which 1) the ordering of varieties 1, …, *v* are shuffled and 2) the two varieties placed on each array are also shuffled. Shuffling of the order of varieties is usually granted since in most cases the *v* varieties are treated as *v* distinct nominal categories without specific ordering. There are exceptions, for instance, in a cell-cycle experiment, one might be interested in comparing the gene expression change between two consecutive stages in a time-series fashion. Shuffling of varieties within each array serves as an averaging process and as a result, each variety is expected to have the same distance to all other varieties. Therefore, the expected variance

$Var(\hat{d}_{v1,v2})$ between variety $v1$ and $v2$ in a balanced block design can be viewed as the average of $Var(\hat{d}_{v1,v2})$ over all $v(v-1)/2$ comparisons in a loop design. In the following sections, we shall discuss the results from the loop design and those of a balanced block design will be similar.

6. Comparison of reference and loop design.

From the above discussion, we can see that, for a fixed number of arrays we can employ twice as many samples in a loop design as in a reference design, the relative efficiency of the loop design and the reference design is:

$$RE = \frac{Var_{ref}(\hat{d}_{v1,v2})}{Var_{loop}(\hat{d}_{v1,v2})}.$$

Thus, if a total of pv arrays (p sample per variety are collected) are employed, we have for the reference design:

$$Var_{ref}(\hat{d}_{v1,v2}) = \frac{1}{p}\sigma_{v1}^2 + \frac{1}{p}\sigma_{v2}^2 + \frac{2}{p}\sigma_A^2 + \frac{2}{p}\sigma^2 .$$

And for a loop design using pv arrays, we can recruit $2p$ samples per variety and the variance of $\hat{d}_{v1,v2}$ is given by $Var_{loop}(\hat{d}_{v1,v2}) = c^T(X^T V^{-1} X)^{-} c$. As previously described, the upper bound of $Var_{loop}(\hat{d}_{v1,v2})$ is the variance of the OLS estimator, $Var_{loop}(\tilde{d}_{v1,v2})$. Therefore, the lower bound of the relative efficiency is:

$$RE \geq \frac{Var_{ref}(\hat{d}_{v1,v2})}{Var_{loop}(\tilde{d}_{v1,v2})} = \begin{cases} \dfrac{2\sigma_{v1}^2 + 2\sigma_{v1}^2 + 4\sigma_A^2 + 4\sigma^2}{\sigma_{v1}^2 + \sigma_{v1}^2 + \sigma_A^2 + 2\sigma^2}, \text{ for connected } v1 \text{ and } v2 \\ 2, \text{ for non-connected } v1 \text{ and } v2 \end{cases}.$$

We can see that a loop design is at least twice as efficient as a reference design when our goal is class comparison between different varieties. In essence, in a reference design, half observations provide no information on our parameter of interest.

A general analytical solution of $Var_{loop}(\hat{d}_{v1,v2})$ is usually not available but some analytical results can be obtained in a simple case in which $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_v^2 = \sigma_V^2$. Under such a condition, we assume the sampling variation σ_i^2 is homogeneous across all varieties i = 1, …, v. Table 2 summarizes the analytical forms of $Var(\hat{d}_{v1,v2})$ when the number of varieties (v) ranges from 2 to 7. We noticed that, interestingly, for a given number of v, when σ_V^2, σ_A^2 and σ^2 are fixed, the variance $Var(\hat{d}_{v1,v2})$ depends only on the distance between $v1$ and $v2$ in a loop. This is because the matrix $X^T V^{-1} X$ is symmetrical and so is its inverse. The distance between $v1$ and $v2$ in a loop is calculated as $\min(v2 - v1, v + v1 - v2)$ and therefore the largest variance occurs at the comparison between variety 1 and the mid-point of v ($v/2 + 1$ for an even v and $(v + 1)/2$ for an odd v). In addition, $Var(\hat{d}_{v1,v2})$ is symmetrical from this "central" comparison with the largest distance in the loop, for example, $Var(\hat{d}_{1,3})$ is exactly the same as $Var(\hat{d}_{1,6})$ in an experiment of 7 varieties. Thus, table 2 only lists the variance formula up to the maximal distance d_m:

$$d_m = \begin{cases} \dfrac{v}{2}, & \text{when } v \text{ is even} \\ \dfrac{v-1}{2}, & \text{when } v \text{ is odd} \end{cases}.$$

It is noteworthy that, when there are only two varieties, $Var(\hat{d}_{1,2})$ becomes just $\sigma_V^2 + \sigma^2$ without the array random component σ_A^2 in it. The analytical solution of $Var(\hat{d}_{v1,v2})$, however, becomes more and more complicated as the number of varieties *v* increases and is therefore not pursued for $v > 7$. In the following discussion, we shall proceed with calculations of the relative efficiency using some exemplary values of the three random components σ_V^2, σ_A^2 and σ^2. Cui and Churchill (2003) gave the average values of the three random components in a mouse experiment and the average proportion of σ_V^2, σ_A^2 and σ^2 is roughly 1:10:1. Figure 3 gives the relative efficiencies for *v* = 8, 12, 16 and 20 using this $\sigma_V^2 : \sigma_A^2 : \sigma^2$ ratio. We can see that although the analytical form of $Var(\hat{d}_{v1,v2})$ is hard to obtain, the actual values of $Var(\hat{d}_{v1,v2})$ quickly converge at all distances as *v* becomes larger. The converging limit depends on real values of σ_V^2, σ_A^2 and σ^2.

We then varied the ratio $\sigma_V^2 : \sigma_A^2 : \sigma^2$ and calculated the relative efficiency of the two designs for different numbers of varieties (from 5 to 30) used. The results are shown in Figure 4. We can see that when the array-to-array variability (σ_A^2) is large compared to the sample-to-sample variability (σ_V^2), the loop design gives a much higher efficiency than the reference design. For example, when $\sigma_V^2 : \sigma_A^2 : \sigma^2 = 1:10:1$, we can see that the relative efficiency is always above 3 even if the number of varieties increases to 30, which means the reference design will not be as efficient as the loop design even if the number of arrays is doubled in a reference design so that the two designs have the same number of non-reference samples (the reference design uses twice as many arrays though). The relative efficiency decreases and finally approaches 2 as σ_V^2 becomes larger

and larger. However, as previously stated, RE is never smaller than 2. Since normally in a microarray experiment the array-to-array variation dominates the overall variability, the gain in efficiency will be remarkable if we use a loop design instead of a reference design.

7. Power and sample size calculations

The null hypothesis: $H_0 : c^T \beta = 0$ can be tested with a *t*-statistics given by:

$$t = \frac{\hat{\theta}}{se(\hat{\theta})},$$

in which $\hat{\theta}$ and $\hat{\theta}$ can be obtained using (2) and (4). The null hypothesis is rejected if $|t| > t_{df,1-\alpha/2}$ for a given significance level α and error degree freedom *df*. Under the alternative, the test statistics *t* has a non-central *t*-distribution. For a given effect size δ, the alternative hypothesis can be written as $H_1 : \theta = \delta$ and the non-centrality parameter λ is given by $\lambda = \frac{\delta}{se(\hat{\theta})}$. The power of the *t*-test will thus be:

$$Power = \Pr(|t_{df}(\lambda)| > t_{df,1-\alpha/2} \,|\, H_1) \qquad (9)$$

Using the estimators of the three variance components from Cui and Churchill (2003), we calculated the power for the reference design and the loop design for v = 2 to 7 varieties. The variance of $\hat{\theta}$ is calculated from Table 2 and for a loop design, we used the maximal variance of all comparisons, that is, the variance between the most distant pair. The results are plotted in Figure 5 and we can see that the loop design gives a much higher power in all cases when the same number of arrays is used. The gain in the power results from the reduction of variance by using a loop design. In addition, even if the array

number is doubled in the reference design so that the same number of non-reference samples is used in both designs, the loop design still yields considerably higher power.

For a predefined power level β, the sample size required to detect an effect size δ can also be calculated from (9) by the inverse CDF of the *t*-distribution.

8. Optimization of experimental cost

In planning a microarray experiment, we are always interested in how to maximize the information obtained with a given resource. Cui and Churchill (2003) provided some guidelines on how to optimize the number of biological samples and technical replicates in a mouse experiment. In their study, the experiment uses a nested reference design with treatment effect nested within mouse effect (Prichard *et al.*, 2001) whereas the main interest was to compare treatment differences. Here we extend their study to the two general designs discussed above.

For the reference design, suppose we recruit m samples per variety and apply n technical replicates for each sample. From (7), the variance of $\hat{d}_{v1,v2}$ is:

$$Var_{ref}(\hat{d}_{v1,v2}) = \frac{2}{m}\sigma_V^2 + \frac{2}{mn}\sigma_A^2 + \frac{2}{mn}\sigma^2 \qquad (10)$$

Suppose the cost per sample is C_S and the cost per array is C_A. Then the total cost will be:

$$C = mC_S + mnC_A \qquad (11)$$

There are two types of optimization: (1) to minimize the total cost of the experiment for a predefined power level to detect $\hat{d}_{v1,v2}$ (2) to maximize the power to detect $\hat{d}_{v1,v2}$ under a given budget for the total cost of the experiment. Cui and Churchill (2003) discussed an example of the first case in a mouse experiment.

In scenario one, the variance $Var(\hat{d}_{v1,v2})$ can be calculated using (9) and therefore we can view (10) as fixed. What we need to do is to minimize the total cost under such an experiment requirement, or minimize (11) under the constraint (10). We apply a Lagrange multiplier and solve *m* and *n*. The optimized *n* is given by:

$$n = \sqrt{\frac{C_S(\sigma_A^2 + \sigma^2)}{C_A \sigma_V^2}}.$$

In a loop design, the number of arrays required is reduced by half so the total cost is:

$$C_{loop} = mC_S + \frac{mn}{2}C_A \qquad (12)$$

For a loop design with 2 varieties, since each loop requires 2 samples, the variance is:

$$Var_{loop}^{v=2} = \frac{2\sigma_V^2}{m} + \frac{2\sigma^2}{mn} \qquad (13)$$

Minimizing (12) under the constraint (13) yields:

$$n = \sqrt{\frac{C_S \sigma^2}{C_A \sigma_V^2}}.$$

Results of the reference design and loop design with two varieties are similar to that of Cui and Churchill (2003). For a loop design with 3 varieties, the variance is:

$$Var_{loop}^{v=3} = \frac{4(\sigma_V^2 + 2\sigma_A^2/n + \sigma^2/n)(\sigma_V^2 + \sigma^2/n)}{m(2\sigma_V^2 + 3\sigma_A^2/n + 2\sigma^2/n)} \qquad (14)$$

Minimizing (12) under the constraint (14) results in:

$$\begin{aligned}
&2C_A\sigma_V^6 n^4 + (8C_A\sigma_V^4\sigma_A^2 + 8C_A\sigma_V^4\sigma^2)n^3 \\
&+(C_S\sigma_V^4\sigma_A^2 + 2C_S\sigma_V^4\sigma^2 + 6C_A\sigma_V^2\sigma_A^4 + 10C_A\sigma_V^2\sigma^4 + 22C_A\sigma_V^2\sigma_A^2\sigma^2)n^2 \\
&+(4C_S\sigma_V^2\sigma^4 + 8C_S\sigma_V^2\sigma_A^2\sigma^2 + 12C_A\sigma_A^4\sigma^2 + 14C_A\sigma_A^2\sigma^4 + 4C_A\sigma^6)n \\
&+6C_S\sigma_A^4\sigma^2 + 7C_S\sigma_A^2\sigma^4 2C_S\sigma^6 = 0
\end{aligned}$$

Some examples of optimized replicate number *n* are given in Table 4.3 for different combinations of sample and array cost using the estimated variance components from Cui and Churchill (2003) under a fixed variance. The resulting *n* is round upwards instead to be conservative. We can see from Table 4.3 that, the cost efficiency of the loop design can be 4~5 times greater than the reference design.

In the second scenario, we may want to obtain maximal power under a given budget, or equivalently, to minimize the variance of $\hat{d}_{v1,v2}$ under a fixed cost. Again, the constrained optimization of *m* and *n* can be solved with a Lagrange multiplier. We show in the appendix that, interestingly, the same results are obtained to calculate *n* as in scenario one. In all cases in Table 4.4, the loop design yields a much higher power than the reference design.

Table 4.4 illustrates some numerical examples of optimized *m* and *n* and the associated power under a fixed budget on the whole experiment. The numbers are calculated by plugging in the estimated variance components from Cui and Churchill (2003) and the respective sample and array cost.

Similarly, we can calculate *m* and *n* when there are more varieties are in a loop design and the details are not pursued here.

Discussion

Our model differs from that of Kerr and Churchill (2001) or Dobbin and Simon (2002) in that we treat the array effect to be random. The array effect accounts for the spot-to-spot variability in spotted microarrays and we consider it to be a random variable as if each array is randomly drawn from a large population. Additionally, in a cDNA microarray experiment, each spot is a block of size two. We favor a mixed model over a fixed one since each array serves as a block in which the two varieties share the same experimental characteristics in the whole process from gene spotting, hybridization, washing and drying, to final fluorescence scan. The accumulation of small errors during the experiment is a justification for the assumed normal distribution.

When array effect is considered as fixed, in contrasting the differences between two different varieties far apart in a loop, more arrays are involved in order to "connect" the two varieties to be compared. Therefore, the variance becomes larger as the distance between the two varieties increases and can exceed that of a reference design for $v > 10$ (Kerr and Churchill, 2001). However, using a mixed model, we found that the variance of $\hat{d}_{v1,v2}$ in a reference design is at least twice as much as that of a loop design regardless of the distance between $v1$ and $v2$ or the actual values of the random components σ_V^2, σ_A^2 and σ^2. Without question, we can see this is a result that the number of observations in a loop design is doubled as compared to a reference design and that comparisons made between two connected varieties vs. two non-connected ones only differ in that the former involves both intra- and inter-array variation but the latter involves only inter-array variation. In addition, the relative efficiency gained in a loop design results not only

from doubling the observations using the same number of arrays, but also from blocking, that is, each array serves as a homogenous block and the within-block variability is minimized. This explains why the relative efficiency can exceed 2. Blocking homogenous experimental units has a long history in statistical literature and has proven superior in providing local control over the experimental factors and thus reducing experimental errors. In an extreme case when only two varieties are used, this array-to-array variation even drops from the variance term. Particularly, when this array-to-array variability is large, blocking can improve the efficiency significantly.

The reasons that the reference design is usually practiced in laboratory, as reviewed by Simon *et al*. (2002) are that 1) the relative intensity obtained from the reference design internally controls for spot-to-spot variation caused by spot size or other sample distribution patterns, 2) each pair of varieties can be contrasted without loss of efficiency introduced by involving other varieties in order to connect that pair, 3) the same reference sample provides control over different experiments or across laboratories and it operationally easier than other designs, 4) if one ore more arrays fail, then the loop is "broken" and the researcher is at risk of ruining the experiment.

Note, however, these conclusions were based on a fixed-effects analysis of the designs. By using an appropriate mixed model approach, as well as appropriate planning and housekeeping, we think that the efficiency gain pays off the extra laboratory effort when switching from a reference design to a loop design or a balanced block design, in agreement with Kerr and Churchill (2001). Under a mixed model framework, the loss of

efficiency for two varieties far apart is never as much as that caused by reducing the observations by half. Even when reduced costs are considered for processing the reference design, an appropriately analyzed loop design yields more information per research dollar spent. Another advantage of loop and balanced block designs is the ability to flexibly assign the most important contrasts onto the same arrays, or balance across all comparisons. In practice, it is possible that one or more arrays may fail due to experimental errors. When the array effect is considered to be fixed, a "broken" loop may result in that we may not be able to connect two "isolated" varieties and thus cancel all array effects and obtain an estimate of their contrast. If we treat the array effect as random, a "broken" loop results in missing observations that are deleted in estimating the fixed effects but are preserved in estimating the random effects. As long as there are non-missing observations from the two varieties to be compared, their contrast will be estimable and what changes is the variance of that estimator. In addition, the study of Deng *et al*. (2004) showed that the intensity is superior to intensity-ratio in modeling the two-dye microarray data because the former has a smaller variance. Therefore, introduction of the reference sample does not seem to be necessary just for forming ratios. It should be noted that, the sample size, power and the overall cost calculations in the text are based on the most distant comparison in the loop design in order to be conservative. In real experiments, we are likely to gain more efficiency when we arrange the most interested comparisons to be closer in the loop design or in the balanced block design. How to make optimal arrangement depends on the objective of the experiment, and the details will be pursued in our future work.

Appendix

For the mixed model $Y = X\beta + Z\gamma + \varepsilon$ in which $\beta = (\mu_1, ..., \mu_v, d_{\text{cy3}}, d_{\text{cy5}})^T$, the generalized least square estimator of β is given by:

$$\hat{\beta} = (X^T V^{-1} X)^- X^T V^{-1} Y,$$

and its variance is:

$$Var(\hat{\beta}) = (X^T V^{-1} X)^-,$$

where V is the variance of Y. For the null hypothesis $H_0 : \theta = c^T \beta = 0$, the variance of $\hat{\theta} = c^T \hat{\beta}$ is given by:

$$Var(\hat{\theta}) = c^T (X^T V^{-1} X)^- c.$$

1. Variance calculation of $\hat{d}_{v1,v2}$ for reference design

In a reference design, each array contains one variety and one reference sample so two different varieties are always uncorrelated because they never appear on the same array. The variance V is a block-diagonal matrix:

$$V = \begin{pmatrix} V_1 & & \\ & \ddots & \\ & & V_v \end{pmatrix}_{vps \times vps}$$

where each element V_i is a $ps \times ps$ matrix corresponding to the covariance-matrix of observations from variety i. From 2.1-2.4 we know, for each variety i, the intra-array (between subsample) and inter-array covariance terms are $\sigma_{ss,i}$ and 0, respectively. Therefore, V_i is a block-diagonal matrix of composed of p blocks:

$$V_i = diag\{V_{S,i}\}.$$

The dimension of V_i is $ps \times ps$ and its block element $V_{S,i}$ is an $s \times s$ matrix given by:

$$V_{S,i} = \begin{pmatrix} \sigma^2_{Y,i} & \sigma_{ss,i} & \sigma_{ss,i} & \cdots & \sigma_{ss,i} \\ \sigma_{ss,i} & \sigma^2_{Y,i} & \sigma_{ss,i} & \cdots & \sigma_{ss,i} \\ & & \ddots & & \\ & & & \ddots & \\ \sigma_{ss,i} & \sigma_{ss,i} & \cdots & \sigma_{ss,i} & \sigma^2_{Y,i} \end{pmatrix}_{s \times s}.$$

The inverse of V is:

$$V^{-1} = \begin{pmatrix} V_1^{-1} & & \\ & \ddots & \\ & & V_v^{-1} \end{pmatrix}_{vps \times vps}.$$

Following model (2) for reference design, the design matrix X can be written as:

$$X = \begin{pmatrix} 1_{ps} & & \\ & \ddots & \\ & & 1_{ps} \end{pmatrix}_{vps \times v},$$

Then

$$X^T V^{-1} X = \begin{pmatrix} 1^T_{ps} & & \\ & \ddots & \\ & & 1^T_{ps} \end{pmatrix} \begin{pmatrix} V_1^{-1} & & \\ & \ddots & \\ & & V_v^{-1} \end{pmatrix} \begin{pmatrix} 1_{ps} & & \\ & \ddots & \\ & & 1_{ps} \end{pmatrix} = \begin{pmatrix} 1^T_{ps} V_1^{-1} 1_{ps} & & \\ & \ddots & \\ & & 1^T_{ps} V_v^{-1} 1_{ps} \end{pmatrix}.$$

The difference $\hat{d}_{v1,v2}$ can be rewritten as $\mu_{v1} - \mu_{v2}$ using the parameterization of β and in a matrix notation:

$$\hat{d}_{v1,v2} = c^T \beta,$$

with $c^T = (0, \cdots 1, \cdots, -1, \cdots 0)$ and the two non-zero entries in c^T correspond to μ_{v1} and μ_{v2} in β, respectively.

Therefore, we have

$$Var(\hat{d}_{v1,v2}) = c^T (X^T V^{-1} X)^- c = \left(1_{ps}^T V_{v1}^{-1} 1_{ps}\right)^- + \left(1_{ps}^T V_{v2}^{-1} 1_{ps}\right)^- = \frac{1}{p}\sigma_{v1}^2 + \frac{1}{p}\sigma_{v2}^2 + \frac{2}{p}\sigma_A^2 + \frac{2}{ps}\sigma_S^2 + \frac{2}{ps}\sigma^2 .$$

2. Variance calculation of $\hat{d}_{v1,v2}$ for loop design

Consider a loop design of *v* arrays, in which each variety has 2 samples, and there is no subsampling on each array. Unlike a reference design in which two different varieties never appear on the same array, in a loop design, genes spotted on the same array from two different varieties are correlated and the variance-covariance matrix of *Y* is a block-diagonal matrix:

$$V = \begin{pmatrix} V_1 & & \\ & \ddots & \\ & & V_v \end{pmatrix}_{2v \times 2v}$$

in which each block element V_i is a 2×2 matrix $V_i = \begin{pmatrix} \sigma_{Y,i}^2 & \sigma_{AA} \\ \sigma_{AA} & \sigma_{Y,i+1}^2 \end{pmatrix}$ and note that the subscript $v+1$ refers to variety 1 as in the loop design.

The inverse of *V* is:

$$V^{-1} = \begin{pmatrix} V_1^{-1} & & \\ & \ddots & \\ & & V_v^{-1} \end{pmatrix}_{2v \times 2v},$$

and if we denote the four elements of V_i^{-1} as $w_{i,11}^{-1}$, $w_{i,12}^{-1}$, $w_{i,21}^{-1}$, $w_{i,22}^{-1}$, V^{-1} can be written as:

$$V^{-1}=\begin{pmatrix} w_{i,11}^{-1} & w_{i,12}^{-1} & & & \\ w_{i,21}^{-1} & w_{i,22}^{-1} & & & \\ & & \ddots & & \\ & & & w_{v,11}^{-1} & w_{v,12}^{-1} \\ & & & w_{v,21}^{-1} & w_{v,22}^{-1} \end{pmatrix}_{2v\times 2v}.$$

From model (1) the design matrix X can be written as:

$$X=\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 1 \\ 0 & 1 & 0 & \cdots & 0 & 1 & 0 \\ & & & \ddots & & & \\ 0 & 0 & 0 & \cdots & 1 & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 1 & 0 \end{pmatrix}_{2v\times(v+2)}.$$

$X^TV^{-1}X$ is no longer in a simplified form and neither is its generalized inverse. The variance of $\hat{\theta}=c^T\hat{\beta}$ is still given by $Var(\hat{\theta})=c^T(X^TV^{-1}X)^-c$ but no simplified analytical form can be obtained.

When p such loops are employed, we notice that the new design matrix X_p will be the stack of p such X matrices and the new variance matrix V_p will be a block-diagonal matrix of p such V matrices. Therefore each element in the matrix $X^TV^{-1}X$ is multiplied by p and the final variance of θ will be reduced by p-fold.

$$Var(\hat{\theta}_p)=c^T(X_p^TV_p^{-1}X_p)^-c=\frac{1}{p}Var(\hat{\theta}).$$

Upper bound of $Var(\hat{\theta})$ can be obtained from the ordinary least square estimator (OLS) since $Var(\hat{\theta}_{GLS})\le Var(\hat{\theta}_{OLS})$. The OLS estimator of θ is given by:

$$\tilde{d}_{v1,v2}=\bar{y}_{v1\ldots}-\bar{y}_{v2\ldots}=\frac{1}{2}\sum_j\sum_k\sum_m y_{v1jkm}-\frac{1}{2}\sum_{j'}\sum_{k'}\sum_{m'} y_{v2j'k'm'}.$$

$\tilde{d}_{v1,v2}$ can be written in a matrix form $\tilde{d}_{v1,v2} = \lambda^T Y$ in which λ^T is given by:

$$\lambda^T = \left(0,\ \cdots,\ \frac{1}{2},\ \frac{1}{2},\ \cdots,\ -\frac{1}{2},\ -\frac{1}{2},\cdots,\ 0\right),$$

with non-zero entries corresponding to the observations from *v*1 and *v*2. For $v1 = 1$, λ^T is in a form of:

$$\lambda^T = \left(\frac{1}{2},\ 0,\cdots,\ -\frac{1}{2},\ -\frac{1}{2},\cdots,\ 0,\ \frac{1}{2}\right).$$

The variance of $\tilde{d}_{v1,v2}$ is then given by $Var(\tilde{d}_{v1,v2}) = \lambda^T V \lambda$ and we can show that for two connected varieties *v*1 and *v*2:

$$Var(\tilde{d}_{v1,v2}) = \frac{1}{2}\sigma^2_{Y,v1} + \frac{1}{2}\sigma^2_{Y,v1} - \frac{1}{2}\sigma_{AA} = \frac{1}{2}\sigma^2_{v1} + \frac{1}{2}\sigma^2_{v2} + \frac{1}{2}\sigma^2_A + \sigma^2;$$

and for two non-connected ones:

$$Var(\tilde{d}_{v1,v2}) = \frac{1}{2}\sigma^2_{Y,v1} + \frac{1}{2}\sigma^2_{Y,v1} = \frac{1}{2}\sigma^2_{v1} + \frac{1}{2}\sigma^2_{v2} + \sigma^2_A + \sigma^2.$$

When p loops are employed, the upper bound of $Var(\hat{\theta}_p)$ is also given by $\frac{1}{p}Var(\tilde{d}_{v1,v2})$ where $Var(\tilde{d}_{v1,v2})$ is obtained above.

Some analytical solutions of $Var(\hat{\theta})$ can be obtained if we assume the sample-to-sample variance is uniform across all varieties $\sigma^2_1 = \sigma^2_2 = \cdots = \sigma^2_v = \sigma^2_V$. Under such a condition, each block element in the *V* matrix becomes the same 2×2 symmetrical matrix:

$$V_1 = \cdots = V_v = \begin{pmatrix} \sigma^2_V + \sigma^2_A + \sigma^2 & \sigma^2_A \\ \sigma^2_A & \sigma^2_V + \sigma^2_A + \sigma^2 \end{pmatrix}.$$

The inverse of V_1 is:

$$V^{-1}=\begin{pmatrix} a & b \\ b & a \end{pmatrix},$$

in which

$$a=\frac{\sigma_V^2+\sigma_A^2+\sigma^2}{\sigma_V^4+\sigma^4+2\sigma_V^2\sigma_A^2+2\sigma_V^2\sigma^2+2\sigma_A^2\sigma^2},$$

and

$$b=-\frac{\sigma_A^2}{\sigma_V^4+\sigma^4+2\sigma_V^2\sigma_A^2+2\sigma_V^2\sigma^2+2\sigma_A^2\sigma^2}.$$

The dimension of $X^TV^{-1}X$ is $(v+2)\times(v+2)$ and now $X^TV^{-1}X$ has a simpler form.

$$X^TV^{-1}X=\begin{pmatrix} A & B \\ C & D \end{pmatrix}.$$

A is a $v\times v$ matrix corresponding to the v varieties. Its main diagonal elements are all $2a$ and the off diagonal element is either b (corresponding to a connected variety) or 0 (a non-connected one). B is a $v\times 2$ matrix, $B=(a+b)\begin{pmatrix}1_v & 1_v\end{pmatrix}$; and C is a $2\times v$ matrix, $C=B^T$. D is a 2×2 matrix and $D=\begin{pmatrix} va & vb \\ vb & va \end{pmatrix}$.

$$X^TV^{-1}X=\begin{pmatrix}
2a & b & 0 & 0 & \cdots & 0 & b & a+b & a+b \\
b & 2a & b & 0 & \cdots & 0 & 0 & a+b & a+b \\
0 & b & 2a & b & \cdots & 0 & 0 & a+b & a+b \\
\vdots & & \ddots & \vdots & & \vdots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & 2a & b & a+b & a+b \\
b & 0 & 0 & 0 & \cdots & b & 2a & a+b & a+b \\
a+b & a+b & a+b & a+b & & a+b & a+b & va & vb \\
a+b & a+b & a+b & a+b & & a+b & a+b & vb & va
\end{pmatrix}$$

The rank of $X^TV^{-1}X$ is $v+1$ and its generalized inverse can be obtained as:

$$(X^T V^{-1} X)^- = \begin{pmatrix} T^{-1} & 0 \\ 0 & 0 \end{pmatrix}$$

in which T is a $(v+1)\times(v+1)$ matrix:

$$T = \begin{pmatrix} A & (a+b)1_v \\ (a+b)1_v^T & va \end{pmatrix}.$$

The variance of $\hat{\theta} = c^T \hat{\beta}$ can be calculated by $Var(\hat{\theta}) = c^T (X^T V^{-1} X)^- c$ and examples of $Var(\hat{d}_{v1,v2})$ between two varieties $v1$ and $v2$ are given in table 2 when the total number of varieties in the experiments v = 2, …, 7. $Var(\hat{d}_{v1,v2})$ becomes more complicated when v increases so the analytical forms are not given for $v > 7$. It is noteworthy that, because $X^T V^{-1} X$ is symmetrical, its inverse $(X^T V^{-1} X)^-$ is also symmetrical. Therefore, for a given v, $Var(\hat{d}_{v1,v2})$ depends only on the distance between $v1$ and $v2$ in the loop but not on the actual numbering of $v1$ and $v2$.

3. Constrained optimization of m and n

3.1 Minimize total experimental cost for a predefined power

From the text, for the reference design, the optimized m and n minimizes (11) under the constraint (10). We introduce a Lagrange multiplier λ and construct the function $f(m,n,\lambda)$ as following:

$$f(m,n,\lambda) = mC_S + mnC_A + \lambda\left(\frac{2}{m}\sigma_V^2 + \frac{2}{mn}\sigma_A^2 + \frac{2}{mn}\sigma^2 - C_V\right)$$

C_V is some constant representing the fixed variance. To solve m and n, set their respective first partial derivatives to be zero.

$$\begin{cases} \dfrac{df}{dm} = C_S + nC_A - \dfrac{2\lambda}{n}(\sigma_V^2 + \dfrac{\sigma_A^2}{n} + \dfrac{\sigma^2}{n}) = 0 \\ \dfrac{df}{dn} = mC_A - \dfrac{2\lambda}{m}(\dfrac{\sigma_A^2}{n^2} + \dfrac{\sigma^2}{n^2}) = 0 \end{cases}$$

Solving the above equations yields:

$$n = \sqrt{\frac{C_S(\sigma_A^2 + \sigma^2)}{C_A\sigma_V^2}} \tag{A.3.1}$$

For a loop design with 2 varieties, the variance and the cost functions are similar with $\frac{C_A}{2}$ in (12) substituting C_A in (10) and $\frac{2\sigma^2}{mn}$ substituting $\frac{2(\sigma_A^2 + \sigma^2)}{mn}$ in (13). Therefore, the optimized *n* satisfies:

$$n = \sqrt{\frac{2C_S\sigma^2}{C_A\sigma_V^2}} \tag{A.3.2}$$

For a loop design with 3 varieties, $f(m,n,\lambda)$ can be constructed as:

$$f(m,n,\lambda) = mC_S + mnC_A + \lambda\left(\frac{4(\sigma_V^2 + 2\sigma_A^2/n + \sigma^2/n)(\sigma_V^2 + \sigma^2/n)}{m(2\sigma_V^2 + 3\sigma_A^2/n + 2\sigma^2/n)} - C_V\right)$$

The optimized *m* and *n* can be calculated from:

$$\begin{cases} \dfrac{df}{dm} = C_S + nC_A - \dfrac{4\lambda}{m^2}\left(\dfrac{(\sigma_V^2 + 2\sigma_A^2/n + \sigma^2/n)(\sigma_V^2 + \sigma^2/n)}{(2\sigma_V^2 + 3\sigma_A^2/n + 2\sigma^2/n)}\right) = 0 \\ \dfrac{df}{dn} = mC_A - \dfrac{4\lambda\left((\sigma_V^2\sigma_A^2 + 2\sigma_V^2\sigma^2)n^2 + (8\sigma_V^2\sigma_A^2\sigma^2 + 4\sigma_V^2\sigma^2)n + (6\sigma_A^4\sigma^2 + 7\sigma_A^2\sigma^4 + 2\sigma^6)\right)}{mn^2(2\sigma_V^2 n + 3\sigma_A^2 + 2\sigma^2)^2} = 0 \end{cases}$$

Solving the above equations yields:

$$\begin{aligned} &2C_A\sigma_V^6 n^4 + (8C_A\sigma_V^4\sigma_A^2 + 8C_A\sigma_V^4\sigma^2)n^3 \\ &+(C_S\sigma_V^4\sigma_A^2 + 2C_S\sigma_V^4\sigma^2 + 6C_A\sigma_V^2\sigma_A^4 + 10C_A\sigma_V^2\sigma^4 + 22C_A\sigma_V^2\sigma_A^2\sigma^2)n^2 \\ &+(4C_S\sigma_V^2\sigma^4 + 8C_S\sigma_V^2\sigma_A^2\sigma^2 + 12C_A\sigma_A^4\sigma^2 + 14C_A\sigma_A^2\sigma^4 + 4C_A\sigma^6)n \\ &+6C_S\sigma_A^4\sigma^2 + 7C_S\sigma_A^2\sigma^4 2C_S\sigma^6 = 0 \end{aligned} \tag{A.3.3}$$

A3.2 Minimize the variance under a fixed experimental cost

From the text, for the reference design, the optimized *m* and *n* minimizes (11) under the constraint (10).

$$f(m,n,\lambda)=\lambda(mC_S+mnC_A-C)+\frac{2}{m}\sigma_V^2+\frac{2}{mn}\sigma_A^2+\frac{2}{mn}\sigma^2$$

C is some constant representing the fixed cost. To solve *m* and *n*, set their respective partial derivatives to be zero.

$$\begin{cases}\dfrac{df}{dm}=\lambda(C_S+nC_A)-\dfrac{2}{m^2}(\sigma_V^2+\dfrac{\sigma_A^2}{n}+\dfrac{\sigma^2}{n})=0\\ \dfrac{df}{dn}=\lambda mC_A-\dfrac{2}{n^2m}(\sigma_A^2+\sigma^2)=0\end{cases}$$

Solving the above equations yields the same result as (A.3.1). Similarly, for a loop design with 2 and 3 varieties, the optimized *n* is given by (A.3.2) and (A.3.3), respectively. The corresponding *m* can be then calculated according to the cost function.

References

Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nature Genet 21:33-37

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I (1998) The transcriptional program of sporulation in budding yeast. Science 282:699-705

Chu TM, Weir BS, Wolfinger RD (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci 176:35-51

Churchill GA (2002) Fundamentals of experimental design for cDNA microarrays. Nature Genet 32:490-495

Cui X, Churchill GA (2003) How many mice and how many arrays? Replication in mouse cDNA microarray experiment. *Methods of Microarray Data Analysis III*. Ed. Johnson KF and Lin SM, Kluwer Academic Publisher.

Deng S, Chu TM, Wolfinger RD, Truong Y (2004) Should we use intensity or intensity ratio in the differential expression analysis of microarray data? Submitted.

Dobbin K, Simon R (2002) Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1438-1445

Eisen M, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of geneome-wide expression patterns. Proc Natl Acad Sci USA 95:14863-14868

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286:531-537

Jin W, Riley RM, Wolfinger RD, White KP, Passador-Gurgel G, Gibson G (2001) The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nat Genet 2001 29:389-395

Kerr MK, Martin GA, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comp Biol 7:819-837

Kerr MK, Churchill GA (2001) Experimental design for gene expression microarrays. Biostatistics 2:183-201

Lee MT, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 18:9834-9839

Prichard CC, Hsu L, Delrow J, Nelson PS (2001) Project normal: defining normal variance in mouse gene expression. Proc Natl Acad Sci USA 98:13266-13271

Searle SR. (1971) *Linear models*. New York: John Wiley & Sons.

Simon R, Radmacher MD, Dobbin K (2002) Designs of studies using DNA microarrays. Genetic Epidemiol 23:21-36

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9:3273-3297

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comp Biol 8:625-637

Table 4.1: Effects of three different replication levels on the variance of the estimated difference $\hat{d}_{v1,v2}$. The total number of observations per gene is fixed at *pqs* and different combinations of samples per variety, aliquots per sample, and subsampling per array are evaluated and the associated $Var(\hat{d}_{v1,v2})$ are given for the reference design.

Reference design			
Sample/ variety	Tech rep/ sample	Subsample/ array	$Var_{ref}(\hat{d}_{v1,v2})$
p	q	s	$\frac{1}{p}\sigma_{v1}^2+\frac{1}{p}\sigma_{v2}^2+\frac{2}{pq}\sigma_A^2+\frac{2}{pqs}\sigma_S^2+\frac{2}{pqs}\sigma^2$
pq	1	s	$\frac{1}{pq}\sigma_{v1}^2+\frac{1}{pq}\sigma_{v2}^2+\frac{2}{pq}\sigma_A^2+\frac{2}{pqs}\sigma_S^2+\frac{2}{pqs}\sigma^2$
p	qs	1	$\frac{1}{p}\sigma_{v1}^2+\frac{1}{p}\sigma_{v2}^2+\frac{2}{pqs}\sigma_A^2+\frac{2}{pqs}\sigma^2$
pqs	1	1	$\frac{1}{pqs}\sigma_{v1}^2+\frac{1}{pqs}\sigma_{v2}^2+\frac{2}{pqs}\sigma_A^2+\frac{2}{pqs}\sigma^2$

Table 4.2: Analytical solutions of $Var(\hat{d}_{v1,v2})$ in a loop design for v = 2 to 7. The results are from an experiment with one loop (2 samples per variety), and if p such loops are employed, the corresponding variances will be divided by p. The distance between $v1$ and $v2$ in a loop of v varieties is calculated as $\min(v2-v1, v+v1-v2)$. In the formula below, $a = \frac{\sigma_V^2 + \sigma_A^2 + \sigma^2}{\sigma_V^4 + \sigma^4 + 2\sigma_V^2\sigma_A^2 + 2\sigma_V^2\sigma^2 + 2\sigma_A^2\sigma^2}$, and $b = -\frac{\sigma_A^2}{\sigma_V^4 + \sigma^4 + 2\sigma_V^2\sigma_A^2 + 2\sigma_V^2\sigma^2 + 2\sigma_A^2\sigma^2}$.

Number of varieties	Distance	$Var_{ref}(\hat{d}_{v1,v2})$
$v = 2$	1	$\frac{1}{a-b}$
$v = 3$	1	$\frac{2}{2a-b}$
$v = 4$	1	$\frac{2a-b}{2a(a-b)}$
	2	$\frac{1}{a}$
$v = 5$	1	$\frac{4a}{4a^2-2ab-b^2}$
	2	$\frac{2(2a-b)}{4a^2-2ab-b^2}$
$v = 6$	1	$\frac{4a^2-2ab-b^2}{(2a+b)(2a-b)(a-b)}$
	2	$\frac{4a}{(2a+b)(2a-b)}$
	3	$\frac{(2a-b)}{(2a+b)(a-b)}$
$v = 7$	1	$\frac{2(2a+b)(2a-b)}{8a^3-4a^2b-4ab^2+b^3}$
	2	$\frac{2(4a^2-2ab-b^2)}{8a^3-4a^2b-4ab^2+b^3}$
	3	$\frac{4a(2a-b)}{8a^3-4a^2b-4ab^2+b^3}$

Table 4.3: Optimal number of samples and arrays for the reference design and the loop design for a given power. The median of the three random components in the mouse data from Cui and Churchill (2003) are used: $\sigma_V^2 = 0.025$, $\sigma_A^2 = 0.32$ and $\sigma^2 = 0.025$. The number of samples per variety *m* and the number of technical replicates per sample *n* are chosen to ensure a power of at least 0.60 to detect an effect size of 0.6 (corresponding to approximately 50% change in raw intensity if we use the log base-2 scale).

					$v = 2$					
		Reference				Loop				$\frac{Cost_{ref}}{Cost_{loop}}$
C_S	C_A	*m*	*n*	Total Cost*	Power	*m*	*n*	Total Cost	Power	
10	300	12	1	7464	.63	5	1	1600	.64	4.67
20	300	12	1	7728	.63	5	1	1700	.64	4.54
50	300	6	2	7860	.60	5	1	2000	.64	3.93
100	300	6	2	8520	.60	5	1	2500	.64	3.41
					$v = 3$					
		Reference				Loop				$\frac{Cost_{ref}}{Cost_{loop}}$
C_S	C_A	*m*	*n*	Total Cost*	Power	*m*	*n*	Total Cost	Power	
10	300	12	1	10998	.65	4	1	1920	.67	5.73
20	300	12	1	11196	.65	4	1	2040	.67	5.49
50	300	6	2	11790	.62	4	1	2400	.67	4.91
100	300	6	2	12780	.62	4	1	3000	.67	4.26

*The cost of a reference sample is calculated as 1/10 of the non-reference sample cost and summed into the total cost of experiment in the reference design.

Table 4.4: Optimal number of samples and arrays for the reference design and the loop design suppose the budget of the whole experiment is set at \$5000 (with 10% overflow limit). The number of samples per variety *m* and technical replicates per sample *n* are given for different combinations of sample and array cost as well as the associated power to detect an effect size of 0.6 at the log base-2 scale (50% raw intensity change). The median of the three random components in the mouse data from Cui and Churchill (2003) are used: $\sigma_V^2 = 0.025$, $\sigma_A^2 = 0.32$ and $\sigma^2 = 0.025$.

					***v* = 2**				
		Reference				Loop			
C_S	C_A	*m*	*n*	Total Cost*	Power	*m*	*n*	Total Cost	Power
10	300	8	1	4976	.45	8	1	2560	1.00
20	300	8	1	5152	.45	8	1	2720	1.00
50	300	4	2	5240	.42	8	1	3200	1.00
100	300	3	2	4260	.31	8	1	4000	1.00
					***v* = 3**				
		Reference				Loop			
C_S	C_A	*m*	*n*	Total Cost*	Power	*m*	*n*	Total Cost	Power
10	300	6	1	5499	.35	8	1	3840	1.00
20	300	5	1	4665	.30	8	1	4080	1.00
50	300	5	1	5325	.30	8	1	4800	1.00
100	300	2	2	4260	.22	6	1	4500	.98

*The cost of a reference sample is calculated as 1/10 of the non-reference sample cost and summed into the total cost of experiment in the reference design.

Figure Legends

Figure 4.1: Layout of a reference design and a loop design using *v* arrays for *v* varieties. A: reference design. B: loop design.

Figure 4.2: A simplified diagram of a loop design for *v* = 6 varieties. Two consecutive varieties are connected with an arrow. Each connected pair represents an array in which the "from" variety is labeled with dye 1 and the "to" variety is labeled with dye 2. A: loop design when 6 arrays are used. Two samples from each variety are used to construct a full loop. B: a double loop with dye reversal between two connected varieties when 12 arrays are used.

Figure 4.3: Relative efficiency of the loop design vs. the reference design for different number of varieties v = 8, 12, 16 and 20. The proportion of the three random components $\sigma_V^2 : \sigma_A^2 : \sigma^2$ is fixed at 1:10:1.

Figure 4.4: Relative efficiency (RE) of the loop design vs. the reference design for different values of the ratio $\sigma_V^2 : \sigma_A^2 : \sigma^2$ and different numbers of varieties (*v* = 5, 10, 15, 20, 25, 30).

Figure 4.5: Power curve for the loop design vs. the reference design when the total number of arrays is 4*v* (8 samples per variety in the loop design and 4 in the reference design, *v* is the number of varieties). Power curve from another reference design with 8 samples per variety (total array number is 8*v*) is overlaid on the same plot for each *v* = 2

to 7. The median of the three random components in the mouse data from Cui and Churchill (2003) are used: $\sigma_V^2 = 0.025$, $\sigma_A^2 = 0.32$ and $\sigma^2 = 0.025$.

Figure 4.1

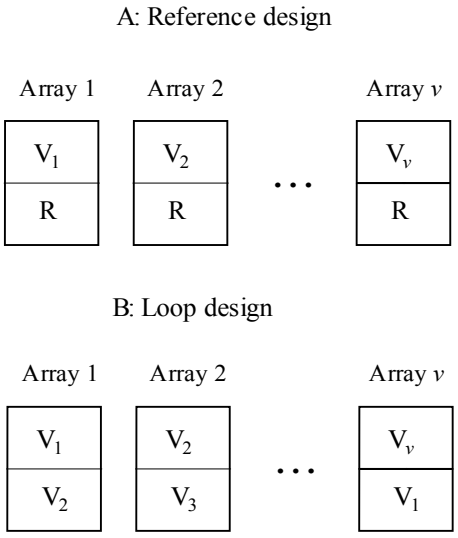
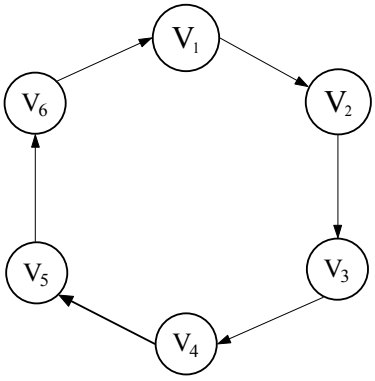

A: Reference design
Array 1
Array 2
Array v
V1
R
V2
R
Vv
R
B: Loop design
Array 1
Array 2
Array v
V1
V2
V2
V3
Vv
V1

Figure 4.2

A: Loop design for 6 arrays



B: Loop design for 12 arrays

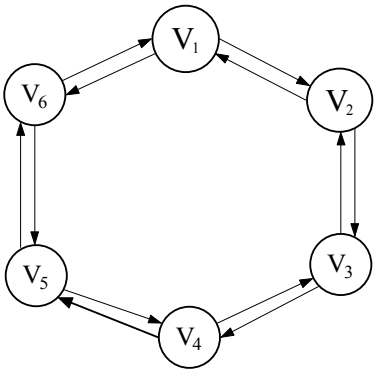


Figure 4.3

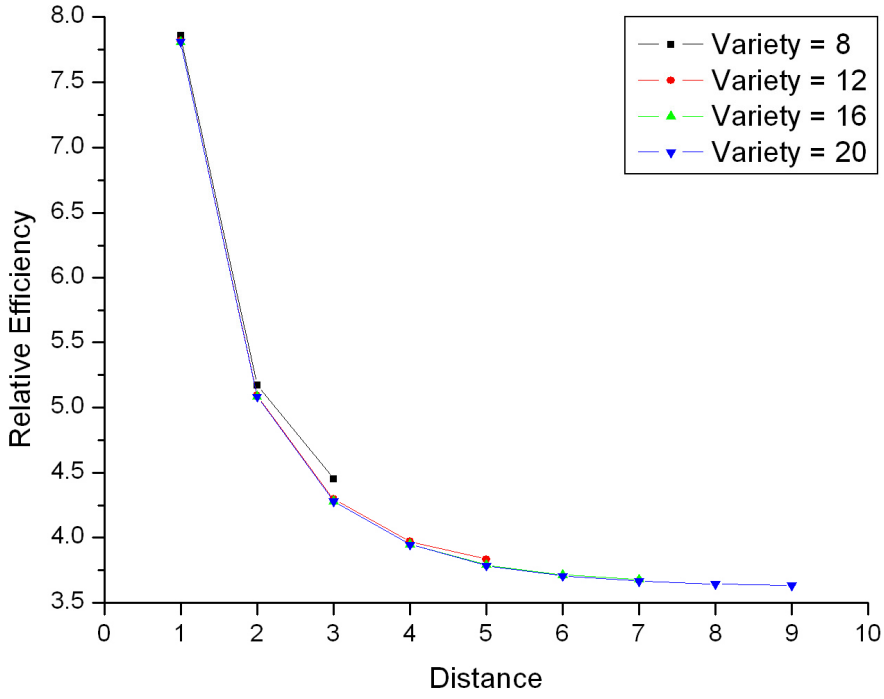

Variety = 8
Variety = 12
Variety = 16
Variety = 20
Relative Efficiency
Distance
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
3.5
0
1
2
3
4
5
6
7
8
9
10

Figure 4.4

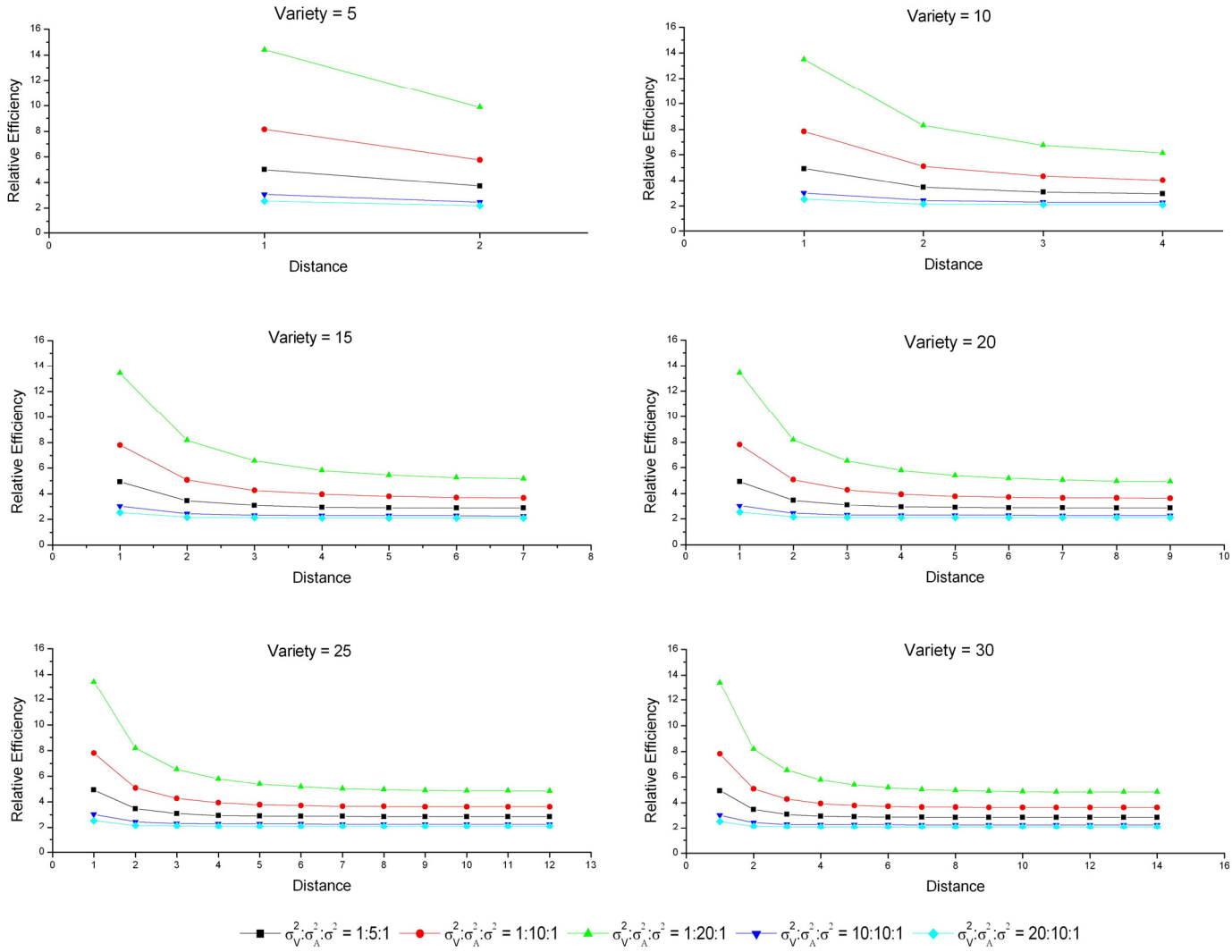
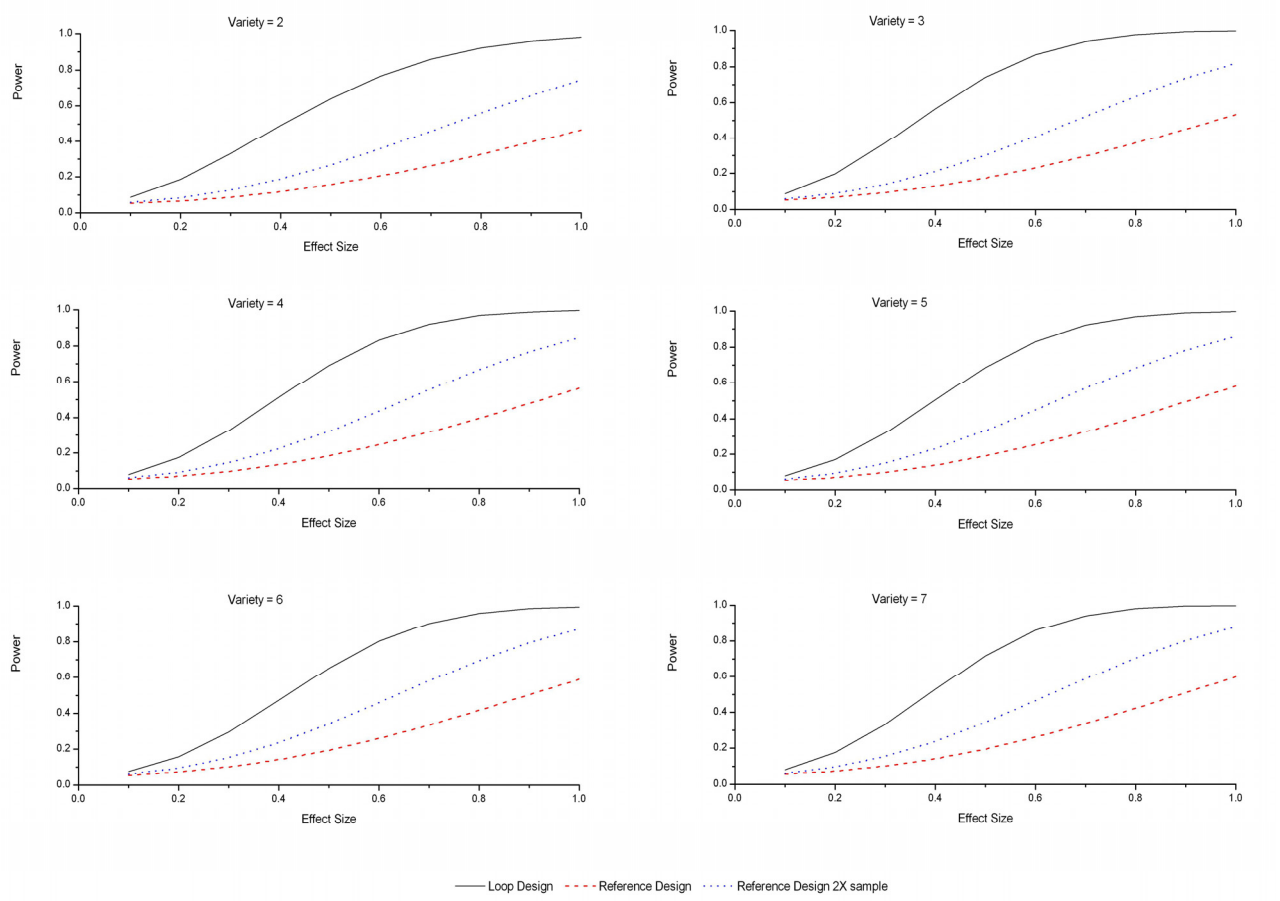

Variety = 5
Variety = 10
Variety = 15
Variety = 20
Variety = 25
Variety = 30
Relative Efficiency
Distance
σ²V:σ²A:σ² = 1:5:1
σ²V:σ²A:σ² = 1:10:1
σ²V:σ²A:σ² = 1:20:1
σ²V:σ²A:σ² = 10:10:1
σ²V:σ²A:σ² = 20:10:1

Figure 4.5


Variety = 2
Variety = 3
Variety = 4
Variety = 5
Variety = 6
Variety = 7
Power
Effect Size
0.0
0.2
0.4
0.6
0.8
1.0
Loop Design
Reference Design
Reference Design 2X sample

Chapter Five

A Mixed Model Approach to Identify Yeast Transcriptional Regulatory Motifs via Microarray Experiments

Abstract

A genome-wide location analysis method has been introduced as a means to simultaneously study protein-DNA binding interactions for a large number of genes on a microarray platform. Identification of interactions between transcription factors (TF) and genes will provide insight into the mechanisms that regulate a variety of cellular responses. Drawing proper inferences from the experimental data is key to finding statistically significant TF-gene binding interactions. We describe how the analysis and interpretation of genome-wide location data can be fit into a mixed model framework that considers the data across all arrays and formulizes appropriate hypothesis tests. The approach is illustrated with data from a yeast transcription factor binding experiment that illustrates how identified TF-gene interactions can enhance initial exploration of transcriptional regulatory networks. Examples of five kinds of transcriptional regulatory structure are also demonstrated.

Introduction

The functionality of a living cell is realized through the concerted activity of a set of genes and gene products. Genome-wide transcriptional analysis, in particular with microarray techniques, provides important information about the expression profiles of clusters of genes that are involved in a variety of cellular processes (DeRisi *et al*. 1997; Cho *et al*. 1998; Gasch *et al*. 2000), yet the how these genes are coordinately regulated remains a topic of intense investigation. Recently, probabilistic models were introduced to build regulation modules using gene expression data from a collection of yeast microarray experiments that measure transcriptional responses to a variety of stress conditions (Segal *et al*. 2001; Pe'er *et al*. 2001; Segal *et al*. 2003). The algorithm developed by Segal *et al*. (2003) takes gene expression data and a large precompiled set of candidate regulatory genes including transcription factors and intermediate signal transduction molecules. Genes are clustered by expression profile based on a regression tree in which each internal node specifies a regulatory input and each leaf node contains a set of genes that respond to a series of regulatory contexts defined by the path from the root node to that leaf node.

With their ability to selectively recognize and bind the promoter of specific genes, transcription factors play a key role in controlling gene expression. Using gene expression profiles, however, whether a transcription factor specifically activates its target genes is usually inferred indirectly from the transcription level of that transcription factors *per se,* as well as the expression patterns of the target genes. Detecting the binding interactions between transcription factors and genes provides extra information that will help to reveal functional features of the genome, and may shed light on the regulatory

mechanisms behind complex cellular process that are coordinated by programmed gene expression involving multiple genes and regulators.

A conventional method to detect protein-DNA interaction *in vivo* is chromatin immunoprecipitation (IP), in which DNA fragments cross-linked to a protein of interest are enriched with a specific antibody to that protein (Orlando 2000). The introduction of microarray technology has enabled investigators to study simultaneous changes in expression across a large number of genes. Although they have been widely used in assessing differentially expressed genes under different conditions, microarray experiments are not limited to expression profiling. By integrating a modified chromatin immunoprecipitation procedure into DNA microarray analysis, Ren *et al*. (2000) developed a genome-wide location analysis method, and demonstrated its capacity to monitor protein-DNA interaction at a genome-wide scale. In their experiments, chromatin immunoprecipitation was used to extract DNA bound by a transcription factor, and IP-enriched DNA fragments (representing promoter sequences bound by that transcription factor) and un-enriched ones were labeled with different fluorophores and hybridized to a microarray containing all known yeast gene promoter sequences. Different fluorescence intensities in the two dyes revealed the presence or absence of binding of that transcription factor to the gene promoter. A recent study of transcriptional regulation in *Sacchromyces cerevisiae* (Lee *et al*. 2002) adopted this genome-wide location analysis method to systematically examine the interactions between 106 known transcription factors and over 6,000 genes. They identified about 4,000 significant binding interactions, which not only helped them to define which transcription factors regulate which genes, but also implied possible structures for pieces of the transcriptional regulatory network.

With a collection of tens of thousands of observations from these microarray experiments, statistical interpretation and inferences from the data becomes the next challenge. Various methods have been developed to infer differentially expressed genes from expression microarray data, and these can also be applied to genome-wide location analysis data. Of these methods, ANOVA-type approaches (Kerr *et al*. 2000; Wolfinger *et al*. 2001; Chu *et al*. 2002) have been advocated for their superiority in partitioning sources of variation, providing reliable estimates of error variance, and for their flexibility in accommodating various experimental designs. Here we emphasize how a mixed model can be appropriately set up to analyze the genome-wide location data from Lee *et al*. (2002). For each gene, a mixed model is fit that uses all observations across all 106 transcription factors for that gene, and a testable hypothesis concerning whether a transcription factor binds to the promoter of that gene is constructed. Significant TF-gene binding is then used to build up transcriptional regulatory motifs that may provide insights as to how transcription factors coordinate in regulation of a set of genes in response to specific cellular processes.

Results

Experimental design and data

In their study, Lee *et al*. (2002) constructed, for each known yeast transcription factor (TF), a strain in which that transcription factor is tagged with a *c-myc* epitope, resulting in 106 tagged strains. Immunoprecipitation was then used to separate DNA fragments representing promoter regions cross-linked to the transcription factor using an anti-Myc antibody. After separation, two pools of DNA, one bound by the transcription factor and the other not, were labeled with different dyes (Cy3 or Cy5) and hybridized to a DNA microarray on which promoter regions of over 6,000 yeast genes were spotted. TF-promoter interactions were then identified by increased fluorescence intensity in the dye coupled to the IP-enriched DNA pool. Three replicates were done for each transcription factor and final data were obtained from 300 arrays.

Data quality inspection

Prior to any analysis, it is important to check data quality and filter aberrant observations. A simple yet effective way is to check the correlations between replicates. We created scatter plots between replicates for each of the 106 transcription factors and found 7 out of 300 arrays that show low correlations (≤ 0.6) with their corresponding replicates. The seven arrays are one of the replicates from seven different experiments with the transcription factors Fkh2, Mcm1, Mot3, Rph1, Swi5, Swi6 and Zap1. A closer look at the dubious array for transcription factor Rph1 showed that over 75% of intensity values are below 11, a number that is normally considered to correspond to background. As a comparison, the 25% percentiles of its two replicates are 54 and 60, and the

corresponding 75% percentiles are 281 and 366, respectively. We decided to discard this array since its quality was too low, possibly due to failure of the hybridization. An important characteristic of a typical cDNA microarray is that the array is prepared in blocks, because the DNA is typically spotted on the slide sequentially with a 4-pin or 16-pin printer. The recorded block identifier makes it possible to track the physical location of each spot and detect local contamination in a small region. The latter can be done using a scatter plot for each block between replicates. We show in Figure 5.1 that, in the array from the problematic transcription factor Zap1, 4 blocks out of 32 exhibit almost no correlation with their replicates while other blocks correlate well. A pseudo image (Figure 5.2) was created that maps the RGB-converted intensities to the physical locations of each spot for this array as well as one of its replicates. The image indicated likely local contamination in a region containing mainly these four blocks. After removal of the observations in these four blocks, the correlation between this array and its two replicates was significantly improved (Figure 5.3). For the other 5 arrays, we did not detect significant block effects and concluded that the low correlations come from experiment-wise variations and kept all those observations in our analysis that follows. The cleaned data set contains 299 arrays with promoters of 6,279 genes each, except for one array from Zap1 in which 900 observations in the four questionable blocks were eliminated.

Mixed model normalization and analysis

Similar to the approach described in Wolfinger *et al*. (2001), we used a mixed model to normalize the data. Let y_{gtda} be the base-2 logarithm of the background-corrected

intensity from gene g (g = 1, …, 6279), transcription factor t (t = 1, …, 106), labeled with dye d (d = 1 for Cy3 and d = 2 for Cy5) in array a (a = 1, …, 299). The normalization model is

$$y_{gtda} = \mu + T_t + D_t + A_a + \xi_{gtda},$$

where μ represents the global mean value, T is the main effect of transcription factor, D is the main effect of dye and A is the random effect for arrays. T and D are assumed to be fixed, A and ξ are assumed to be normally distributed random variables with mean 0 and variance σ_A^2 and σ_e^2, respectively. This normalization corrects effects that are not gene specific. Residuals from the normalization model were then used as the input to our second mixed model, in which we fit each gene individually:

$$r_{tpa} = \mu + T_t + P_p + (TP)_{tp} + A_a + \varepsilon_{tpa}.$$

Here, r_{tpa} is the residual measurement for transcription factor t (t = 1, …, 106), probe p (p = 1 for IP-enriched DNA pool and p = 2 for genomic control DNA) and array a (a = 1, …, 299). μ is the global mean, T is the main effect of transcription factor representing the overall differences in intensity for each yeast strain in which a different TF is tagged, P is the probe main effect representing the overall differences in intensity for two DNA pools hybridized to the array, and TP is the effect of transcription factor by probe interaction representing the intensity differences between the two probes within the same transcription factor. In these gene models, this interaction term $(TP)_{tp}$ accounts for the gene specific differences of binding intensities between transcription factor t and the genomic control. A is the array random effect representing overall variation in fluorescence measurement across arrays and is assumed to be normally distributed with

mean 0 and variance σ_A^2. ε is the error term and is assumed to be normally distributed with mean 0 and variance σ^2 and is independent of A. Typically, a dye effect term that corrects for gene specific intensity differences between the two different dyes needs to be added to the gene model. In this experiment, all IP probes were labeled with one dye (Cy5) and genomic control probes labeled with another (Cy3), except for 35 out of 300 arrays where the dyes are reversed, that is, IP probe was labeled with Cy3 and control probe was labeled with Cy5. However, within each transcription factor, all three replicates were labeled in one direction without dye-flip. Therefore, the dye effect is completely confounded with some combinations of probe and factor effects so it is not estimable.

We also performed a simple iterative outlier filtration method based on our mixed model. For each iteration, we first fit our linear mixed model and calculated the standardized residual for each observation, then eliminated observations having a standardized residual greater than three (Chu *et al.* 2002). A more sophisticated method is to perform a statistical test for each potential outlier and filter outliers one-by-one (Chu and Wolfinger 2003) but generally an observation with a standardized residual larger than three can be regarded as an outlier based on the normal distribution assumption of errors. We carried out three iterations and filtered ~0.58% observations as outliers. Data after outlier screening were then fit by our mixed model to obtain the final results. The mean R^2 values of the 6,279 gene models, as shown in Figure 5.3, was about 0.9, indicating that the mixed models offer a good fit to the data and explains most of the observed variations. As described, our goal is to find, for each gene, if a transcription factor *t* that significantly binds the promoter of the gene, as indicated by increased fluorescence intensity in IP-

enriched probe. Therefore, to test the hypothesis whether a transcription factor *t* interacts with the promoter of a gene, we need to contrast the transcription factor by probe interaction term between the IP-enriched probe and the genomic control probe. Our hypothesis test is then formulized as:

$$H_0 : TP_{t1} - TP_{t2} <= 0,$$
$$H_1 : TP_{t1} - TP_{t2} > 0,$$

and a one-sided *t*-test is used to assess the significance for each of the 6,279 genes and 106 transcription factors.

Using a Bonferroni correction for these 6,279×106 (665,574) tests, we found 5,603 significant TF-gene interactions at the 0.05 level, which corresponds to roughly 8 false positives out of 100 million single tests. Of the 6,279 genes, 5,183 have been fully annotated and named with known ORFs from the *Sacchromyces cerevisiae* genome database (SGD, 2003). The 5,603 TF-gene interactions we found involve 101 transcription factors and 2,731 genes. Since Bonferroni adjustment is usually thought to be too conservative, we also performed false discover rate (FDR) adjustment (Benjamini and Hochberg 1995) for the *p*-values obtained from these over 600K tests. For example, the FDR adjusted *p*-value at 0.01 is equivalent to a *p*-value of 4.5×10^{-4} in the original test, which results in 29,533 significant TF-gene interactions. Results from different *p*-value cutoffs for multiple-testing adjustment are summarized in table 1. As a comparison, the error model (Stoughton and Dai 2002) identified 3,985 significant interactions using a single test *p*-value threshold of 0.001. A *p*-value cutoff of 0.05 from the error model would have identified 35,365 significant interactions. Using the same Bonferroni *p*-value cutoff of 7.5×10^{-8}, however, only 427 significant interactions would be retained from the error model. The negative base-10 log *p*-value of the 6,279×106 (665,574) tests from the

mixed model and the error model are plotted against each other in Figure 5. The horizontal dashed line indicates the Bonferroni cutoff (p-value = 7.5×10^{-8}) used in the mixed model and the vertical dashed line indicates the cutoff (p-value = 0.001) Lee *et al.* (2002) used from their error model. The upper right rectangular section contains significant test results found by both models whereas the upper left rectangular section represents significances identified by the mixed model but not by the error model and the lower right section indicates the opposite. The counts in the above three sections are 902, 4701, and 3083, respectively.

Identification of transcriptional regulatory motifs

It has always been of great interest to find out how regulators crosstalk and coordinate in regulating a cellular process. Similar to the method described in Shen-Orr *et al*. (2002) and Lee *et al*. (2003), we can draw inference about simple regulatory motifs based on the interactions identified between a transcription factor and its target genes. A regulatory motif represents a compact, modular form that occurs as a pattern in the transcriptional network and is the basic building block of a complex network. Simple regulatory units include autoregulation, feedforward loops, regulator chains, single-input modules (SIM) and multiple-input modules (MIM). We identified these units from the 5,603 interactions found from the mixed models, which included 101 transcription factors and 2,731 genes.

In an autoregulation motif, a regulator binds to the promoter of its own gene (Figure 5.6). Autoregulation serves to amplify cellular responses, reduce the response time to environmental stimuli, and decreases the biosynthetic cost of regulation (Shen-Orr *et al*. 2002). From the 106 transcription factors, we identified five autoregulation motifs, Rap1,

Ste12, Sum1, Swi4, and Zap1. Zap1p is a transcription factor that directly controls zinc-responsive gene expression in yeast and is found to increase the expression of the *ZAP1* gene itself, presumably through a positive autoregulatory mechanism. Further support for this autoregulation comes from the identification of a Zap1p binding site within the *ZAP1* gene promoter (Zhao *et al*. 1998). Using the error model and a *p*-value cutoff of 0.001, Lee *et al*. (2003) identified 10 autoregulation motifs including the above five as well as Aro80, Nrg1, Rcs1, Smp1, Yap6. They have also confirmed these 10 binding interactions using conventional chromatin immunoprecipitation. Using a false discovery rate of 0.01 for multiple-testing adjustment (which corresponds to 4.5×10^{-4} for a single test), the above five extra factors are included, as well as two additional factors, Dot6 and Rgt1. However, no more transcription factors are found to be auto-regulated as the single test *p*-value threshold continues to be lowered until 0.001, indicating that the proportion of yeast genes encoding transcription factors that are auto-regulated is about 10%, a fact addressed by Lee *et al*. (2002). Since the Bonferroni correction is a conservative approach to multiple-testing problems, we may risk increasing the false negative rate while minimizing false positives. However, it is noteworthy that the sensitivity of our mixed model is much higher as compared to the error model using the same criterion. As discussed in Wolfinger *et al*. (2001) and Chu *et al*. (2004), this conservative Bonferroni adjustment still serves as a good lower bound to screen for the most significant transcription factors and their target genes for further analysis.

A feedforward loop contains a regulator *X* that regulates a second regulator *Y*, such that *X* and *Y* control a common target gene *G* jointly. A feedforward loop is thought to cause a

rapid response to an external signal through amplification of regulators of the target gene. It may also serve as an "AND-gate" control over the target gene *G* when the activation through *X* is transient, and accumulation of both *X* and *Y* signals are required to activate *G* (Shen-Orr *et al*. 2002). We found 58 feedforward loops involving 45 transcription factors, and these potentially control 245 genes. For example, Mcm1p binds to the *ACE2* promoter, and both Mcm1p and Ace2p regulate the expression of *CLN3* (Figure 5.6). During the *S. cerevisiae* cell cycle, the transcriptional regulator Mcm1p is responsible for transcriptional activation of multiple G2-to-M phase specific genes and *ACE2* expression has been shown to depend on Mcm1p (Althoefer *et al*. 1995). Ace2p is functional in late anaphase and activates expression of M-to-G1 specific genes. The target gene *CLN3* encodes a G1 cyclin that allows cells to commit to a new round of division. This forward loop may indicate some kind of temporal control of gene expression that ensures an adequate amount of Cln3 at the start of a new cell cycle. As a comparison, the error model (Lee *et al*. 2002) found Mcm1 and Ace2 form a feedforward loop with each of two target genes *UTH1* and *YKR040C* but not *CLN3*.

A regulator chain consists of a chain of regulators in which regulator X_1 binds to the promoter of a second regulator X_2, X_2 then binds to a third one X_3, and so on. We used a recursive algorithm to find regulator chains. For the i^{th} regulator in a chain, we find all regulator promoters that it binds to, place them at the $(i+1)^{th}$ position in the chain, and continue on to the next position until the cascade ends. There are three possible ways to end a chain: 1) the last regulator does not bind any regulator promoter; 2) the last regulator binds to its own promoter; 3) the last regulator binds to a regulator promoter

earlier in the chain. A special case in 3) is that the last regulator binds to the promoter of the starting regulator of the chain, resulting in a so-called a multi-component loop as discussed in Lee *et al*. (2002). We identified 136 regulator chains from the mixed model results, with lengths varying from 3 to 8. An example of a three-component cascade is Swi4-Sok2-Yap3 in which Swi4p regulates *SOK2* and Sok2p then binds to the promoter of *YAP3* (Figure 5.6). In *S. cerevisiae*, the Yap proteins are known to activate gene expression in response to different environmental stresses and Yap-dependent transcriptional activation is inhibited by PKA (Fernades *et al*. 1997). The Swi4-Sok2-Yap3 chain may be indicative of this inhibition during the PKA-dependent cell cycle progression. Sok2p is a positive regulator of PKA activity and one possible mechanism for PKA-related inhibition of Yap proteins is that the expression of Yap proteins is reduced (Fernandes *et al*. 1997). This indicates that Sok2p may bind to the promoter of *YAP3* and repress its expression. Using the error model (Lee *et al*. 2002), binding of Swi4p to the promoter of *SOK2* was identified but no other transcription factors were found to act downstream of Sok2p in a regulator chain.

Single-input module (SIM) motifs contain a single regulator X that binds a set of genes G_1, …, G_n, presumably under a specific condition, and these genes have no additional transcriptional regulation. To identify a single-input module, we first found genes whose promoter is bound by only one regulator and then grouped these genes by the regulators. We found 59 SIMs from our 5,603 significant binding interactions, and these potentially control 1,033 target genes. Although these results may not represent all bindings under all conditions in a living cell, a single-input motif still serves as a starting point to look at a coordinated discrete unit of biological function. For example, Hap4p, a global regulator

of respiratory gene expression (Forsburg and Guarente 1989), forms a SIM with three genes involved in respiration, *QCR9*, *COX5A* and *COX6*. *QCR9* encodes a subunit of the ubiquinol cytochrome C oxidoreductase complex. Each of Cox5Ap and Cox6p is a subunit of cytochrome C oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain (Figure 5.6). This SIM may also indicate that several genes function stoichiometrically, perhaps because the products assemble together to form a functional protein complex. The results from the error model (Lee *et al*. 2002) indicated that Hap4p forms a SIM with 34 genes including *COX5A* but not *QCR9* or *COX6*.

Multiple-input module (MIM) motifs consist of a set of regulators that bind together to a common set of genes. A MIM implicates potential coordination of gene expression under various conditions, or a possible coordinated cellular process involving multiple regulators, signal molecules and functioning genes. We identified 379 combinations of two or more regulators that control a set of common target genes. For example, Hir1p and Hir2p, known as negative regulators in the transcription of histone genes during the *Saccharomyces cerevisiae* cell cycle, form a MIM with four histone genes, *HHT1*, *HHF1*, *HHT2* and *HHF2* (Figure 5.6). It has been shown that six out of the eight histone genes (*HTA1-HTB1*, *HHT1-HHF1*, and *HHT2-HHF2*) are regulated at the transcriptional level and this transcriptional level regulation plays an important role in the synthesis of the core histones in *Saccharomyces cerevisiae*. Hir1p and Hir2p have been shown to repress the transcription of *HHT1-HHF1* and *HHT2-HHF2*, and the regulation of *HTA1-HTB1* requires another protein Hir3p (Spector *et al*. 1997). These results are consistent with our findings.

Because we used the conservative Bonferroni correction, we expect that the number of true interactions is underestimated, and therefore the number of motifs we identified in each category is also probably an underestimation. For example, if we use an FDR at 0.0001, the number of significant interactions we can identify is 9,158 and the number of motifs in the latter four categories, feedforward loop, regulator chain, single-input modules, multiple-input modules, will be 123, 522, 61 and 675, respectively. The number of motifs we found in each category using different *p*-value cutoffs for multiple-testing adjustment is summarized in Table 5.1.

Discussion

A conventional use of cDNA microarrays is to identify differentially expressed genes. Whether the transcription level of a specific gene is governed by a specific transcription factor is usually inferred indirectly from comparison of certain conditions, for instance, when the transcription factor is deleted or over-expressed. Since transcription typically starts with the association of the transcription factor and co-factors at the promoter site of the target gene, direct study of this protein-DNA interaction has always been of great interest. The introduction of the genome-wide location analysis method (Ren *et al*. 2000) provides a systematic way to infer the binding interactions between a transcription factor and a large number of genes using microarray technology, helping us to understand the mechanisms of transcriptional regulation of these genes.

Analysis of genome-wide location data is similar to that of regular microarrays. Systematic approaches described in Chu *et al*. (2002) provide a general framework for microarray data analysis and can easily be adapted to analyze genome-wide location data. Among the algorithmic steps, data cleaning and outlier detection seems important prior to analysis. In practice, it is not unusual to get low quality arrays since there are many factors that affect the quality of the final image data: quality of clone preparation, uneven DNA printing on the slide, scratches, dust or other artifacts on the array, and non-uniform hybridization, etc. Excluding spots with poor quality in the early stage of microarray analysis is beneficial since the normalization stage usually involves an estimation phase. Several recent studies have addressed performance of quality control during the image-processing phase (Brown *et al*. 2001; Wang *et al*. 2001). When the original images are not available, we can still do some quality control by taking advantage of replicates from

the experiment. We showed here that a correlation check between replicates can lead to discovery of poor quality arrays as well as poor quality regions of spots when this check is performed in a block-by-block fashion.

Power to find significant TF-gene interactions was substantially increased, as compared to a *t*-test based error model approach, by fitting log-linear mixed models for each gene. Pooling all of the data together to compute variance components leads to identification of the sources of variation affecting the response variable, namely the specific mutant strains that are inferred to have altered binding intensity of a transcription factor with the gene promoter. The error model, by taking into consideration the intensity-dependent signal variability that commonly appears in a microarray experiment, estimates the error variance using both a multiplicative and an additive error term, which leads to increased sensitivity and specificity in detecting differential binding interactions. The error model *p*-value for each gene and transcription factor, however, is calculated from only the replicates of that transcription factor, which in this study was only six observations for each gene for each transcription factor. Therefore, the results from the error model are likely to be too conservative and lack power. On the other hand, in our mixed model, all 600 observations for each gene are pooled together, leading to a much more reliable error estimate and increased sensitivity.

An example of the improved resolution of our method is supplied by the report of immunoprecipitations experiments in Iyer *et al.* (2001) that *CWP2*, which encodes a structural protein of the cell wall, is a target gene of Swi6p. Our mixed model identified this binding interaction to be highly significant (p-value = 8.0×10^{-20}) whereas the p-value from the error model is 0.1 (Figure 5.5). As seen from Figure 5.5, the upper left

rectangular section contains significant bindings found by the mixed model but not by the error model. If we set the *p*-value cutoff from 0.001 to 0.0001 in the error model (by moving the vertical reference line from 3 to 4), only about 1,000 significant interactions will be retained in the lower right section that are found by the error model but not the mixed model.

It should be noted that unless we are willing to assume there is no gene-specific dye effect, in which case the global dye effect is removed by the normalization model, we need a dye effect in our mixed model that accounts for it. Gene-specific dye effects could arise from differences in the efficiency of fluorophore incorporation and/or differences in hybridization efficiency of the DNA fragments coupled with different dye molecules. In our gene model, we did not include the dye effect because it is completely confounded as a result of the experimental design. This, however, could somehow bias the estimate of our parameter of interest, that is, the transcription factor by probe interaction. Therefore, even with a conservative Bonferroni correction for multiple testing, we expect that the family-wise false positive rate may exceed 0.05 due to this dye bias. It is noteworthy, however, that for this data set, this confounding issue persists no matter which model is used. In order for the dye effect to be estimable, an appropriate experimental design should be adopted that balances the dyes with respect to the probes within each transcription factor. That is, for each transcription factor, an even number of replicates is required in which dye-reversal is performed for the two probes (IP and control). A practical usage of four replicates is thus advocated for this type of study.

It should also be noted that the algorithms described above for finding transcriptional regulatory motifs are solely based on TF-gene interactions. Although they provide a way

to possibly identify coordinated regulators and genes, many other factors, for example intermediate signal transduction molecules along the cellular process pathway that are not directly controlled by the transcription factors, are not likely to be included in the motifs and eventually in assembling transcriptional regulatory networks. Also, the motifs discovered may contain considerable redundancies, especially for regulator chains and multiple-input modules. For example, two regulator chains may have the same start and end transcription factors but with various lengths, or even differ only at one or two positions where the binding path can be substituted with a different transcription factor. Therefore, results from these motifs should be interpreted with care. Additional information, such as extensive gene expression data that can lead to discovery of co-expression of multiple genes, may assist in creation of a clearer picture of transcriptional regulation. A recent study by Bar-Joseph *et al*. (2003) has demonstrated how to combine the genome-wide location data with gene expression profiles to build up the regulatory networks of genetic modules.

Acknowledgement

We thank R.A. Young's lab for publishing their experimental data online, N.J. Rinaldi and B.S. Weir for discussions and comments on the manuscript.

Reference

Althoefer H, Schleiffer A, Wassmann K, Nordheim A, Ammerer G (1995) Mcm1 is required to coordinate G2-specific transcription in Sacchromyces cerevisiae. Mol Cell Biol 15:5917-5928

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert R, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2003) Computational discovery of gene modules and regulatory networks. Nature Biotech 21:1337-1342

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a new and powerful approach to multiple testing. J Royal Stat Society 57:1289-1300.

Brown CS, Goodwin PC, Sorger PK (2001) Image metrics in the statistical analysis of DNA microarray data. Proc Natl Acad Sci USA 98:8944-8949

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG *et al*. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2:65-73

Chu TM, Weir BS, Wolfinger RD (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Math Biosci 176:35-51

Chu TM, Weir BS, Wolfinger RD (2004) Comparison of Li-Wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays. Bioinformatics 20:500-506

Chu TM, Wolfinger RD (2003) Statistical outlier detection for microarray data. JSM 2003.

de Nadal E, Casadome L, Posas F **(**2003) Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. Mol Cell Biol 23:229-237

DeRisi JL, Lyer VR, Brown PO (1997) Exploring the metabolic and genetic control of gene expression o a genomic scale. Science 278:680-686

Fernandes L, Rodrigues-Pousada C, Struhl K (1997) Yap, a novel family of eight bZIP proteins in Saccharomyces cerevisiae with distinct biological functions. Mol Cell Biol 17:6982-6993

Forsburg SL, Guarente L (1989) Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. Genes Dev 3:1166-1178

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Bosten D, Brown PO (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 11:4241-4257

Iyer VR, Horak CE, Scafe CS, Bostein D, Snyder M, Brown PO (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature 409:533 – 538

Kerr MK, Martin GA, Churchill GA (2000) Analysis of variance for gene expression microarray data. J Comp Biol 7:819-837

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM *et al*. (2002) Transcriptional regulatory network in Sacchromyces cerevisiae. Science 298:799-804

Orlando V (2000) Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. Trends Biochem Sci 25:99-104

Pe'er D, Regev A, Elidan G, Friedman N (2001) Inferring subnetworks from perturbed expression profiles. Bioinformatics 17:S215-S224

Ren B, Robert F, Wyrick JJ, Aparico O, Jennings EG, Simon I, Zeitlinger J, *et al*. (2000) Genome-wide location and function of DNA binding proteins. Science 290:2306-2309

Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD *et al*. (2000) Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science 287:873-880

Segal E, Taskar B, Gasch A, Friedman N, Koller D (2001) Rich probabilistic models for gene expression. Bioinformatics 17:S243-S252

Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genet 34:166-176

Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional reglation of *Escherichia coli*. Nature Genet 31:65-68

Spector MS, Raff A, DeSilva H, Lee K, Osley MA (1997) Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the *Saccharomyces cerevisiae* cell cycle. Mol Cell Biol 17:545-552

Stoughton R, Dai H (2002) Statistical combining of cell expression profiles. US Patent 6351712.

Wang X, Ghosh S, Guo SW (2001) Quantitative quality control in microarray image processing and data acquisition. Nucleic Acids Res 29:E75

Ward MP, Gimeno CJ, Fink GR, Garrett S (1995) SOK2 may regulate cyclic AMP-dependent protein kinase-stimulated growth and pseudohyphal development by repressing transcription. Mol Cell Biol 15:6857-6863

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001) Assessing gene significance from cDNA microarray expression data via mixed models. J Comp Biol 8:625-637

Zhao H, Butler E, Rodgers J, Spizzo T, duesterhoeft S, Eide D (1998) Regulation of Zinc Homeostasis in Yeast by Binding of the ZAP1 Transcriptional Activator to Zinc-responsive Promoter Elements. J Biol Chem 273:28713-28720

Table 5.1: Multiple-testing adjustment and the effect of *p*-value threshold on the number of network motifs detected. Numbers of significant tests and corresponding number of network motifs in each category are listed with different *p*-value thresholds using Bonferroni or false discovery rate (FDR) adjustment. The results are ordered by single test *p*-values from the highest to the lowest for the above methods. The total number of tests is 665,574.

Method	Cutoff	Single test *p*-value	Significances	AR	FFL	RC	SIM	MIM
Error model	.001	.001	3985*	10*	49*	191*, a	90*	295*
FDR	.01	4.5×10^{-4}	29,533	19	501	13,504	49	1,815
FDR	.001	2.4×10^{-5}	15,639	13	237	1,987	58	1,187
FDR	.0001	1.4×10^{-6}	9,158	7	123	522	61	675
Bonferroni	.05	7.5×10^{-8}	5,603	5	58	136	59	379

Note: AR=autoregulation; FFL=feedforward loop; RC=regulator chain; SIM=single-input module; MIM=multiple-input module.

*: result from Lee *et al*. (2002)

a: the number of regulator chain from Lee *et al*. is 188 and they also identified three multi-component loops, which is a special case of regulator chain, so the total number is 191.

Figure legends

Figure 5.1: Scatter plot of block 21, 22, 25 and 26 of array 3997 for transcription factor Zap1 with its two replicates (array 3998 and array 3999). The four panels represent block 21 (top left), 22 (top right), 25 (bottom left), 26 (bottom right), respectively and the eclipse in each cell represents 90% density curve.

Figure 5.2: Pseudo image of array 3997 and its replicate 3998 from transcription factor Zap1. In a pseudo image, the raw intensity ratios are converted to RGB values from 0 to 255 and the positions of each spot are used as coordinates to plot that spot along with its RGB value.

Figure 5.3: Scatter plot of array 3997 with its two replicates (array 3998 and array 3999) before (left panel) and after (right panel) removal of dubious blocks 21, 22, 25 and 26.

Figure 5.4: Histogram of R^2 values for the 6,279 genes from the mixed model.

Figure 5.5: Comparison of negative $\log_{10} p$-values from the mixed model and the error model. The horizontal dashed reference line indicates the Bonferroni cutoff (p-value = 7.5×10^{-8}) used in the mixed model and the vertical dashed reference line indicates the error model cutoff (p-value = 0.001) used in Lee *et al.* (2002).

Figure 5.6: Examples of regulatory motifs. A transcription factor is represented in an eclipse box and a gene is represented in a rectangular box and an arrow indicates binding of a transcription factor to the promoter of its target gene.

Figure 5.1

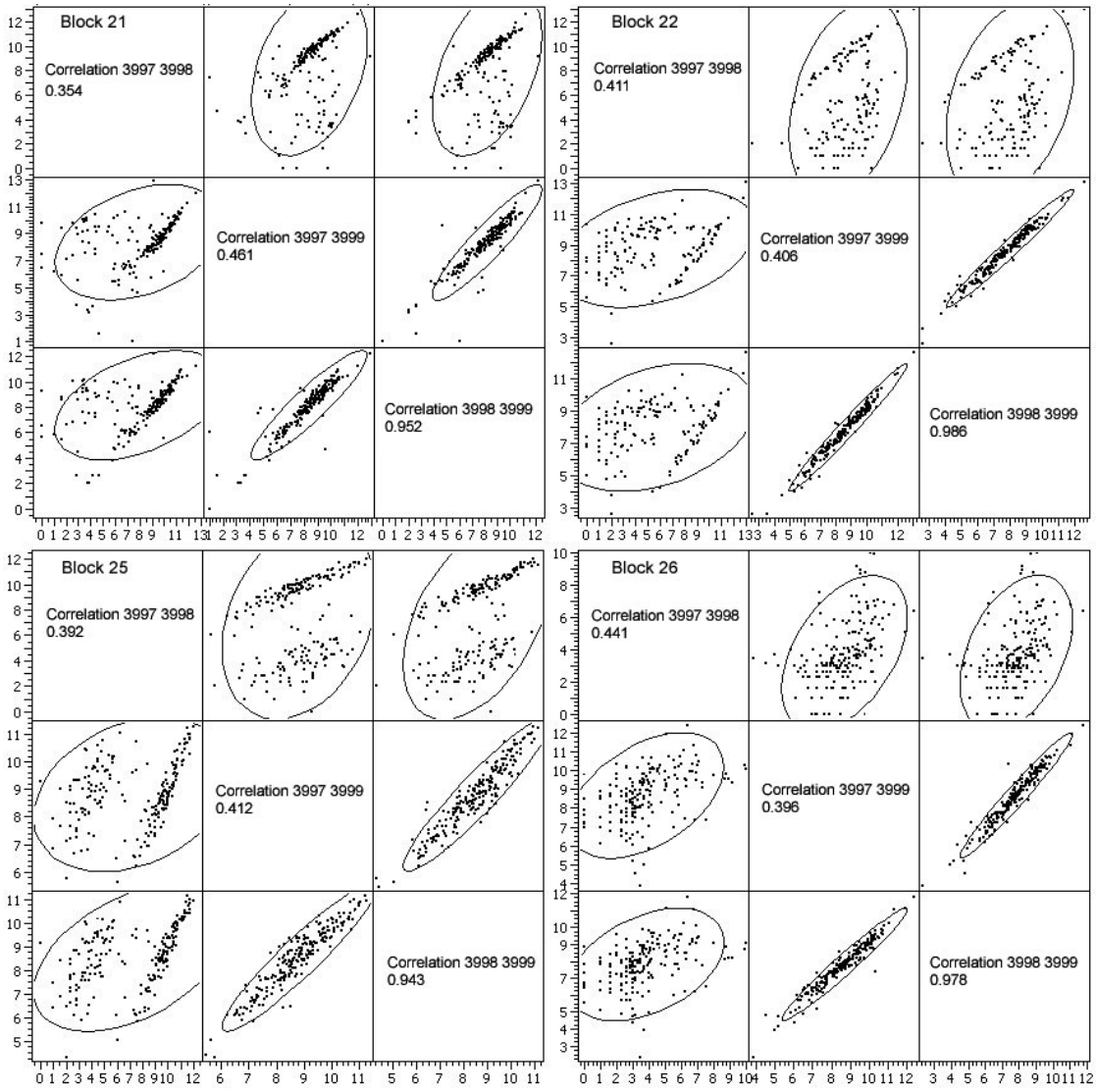

Block 21
Correlation 3997 3998
0.354
Correlation 3997 3999
0.461
Correlation 3998 3999
0.952
Block 22
Correlation 3997 3998
0.411
Correlation 3997 3999
0.406
Correlation 3998 3999
0.986
Block 25
Correlation 3997 3998
0.392
Correlation 3997 3999
0.412
Correlation 3998 3999
0.943
Block 26
Correlation 3997 3998
0.441
Correlation 3997 3999
0.396
Correlation 3998 3999
0.978

Figure 5.2

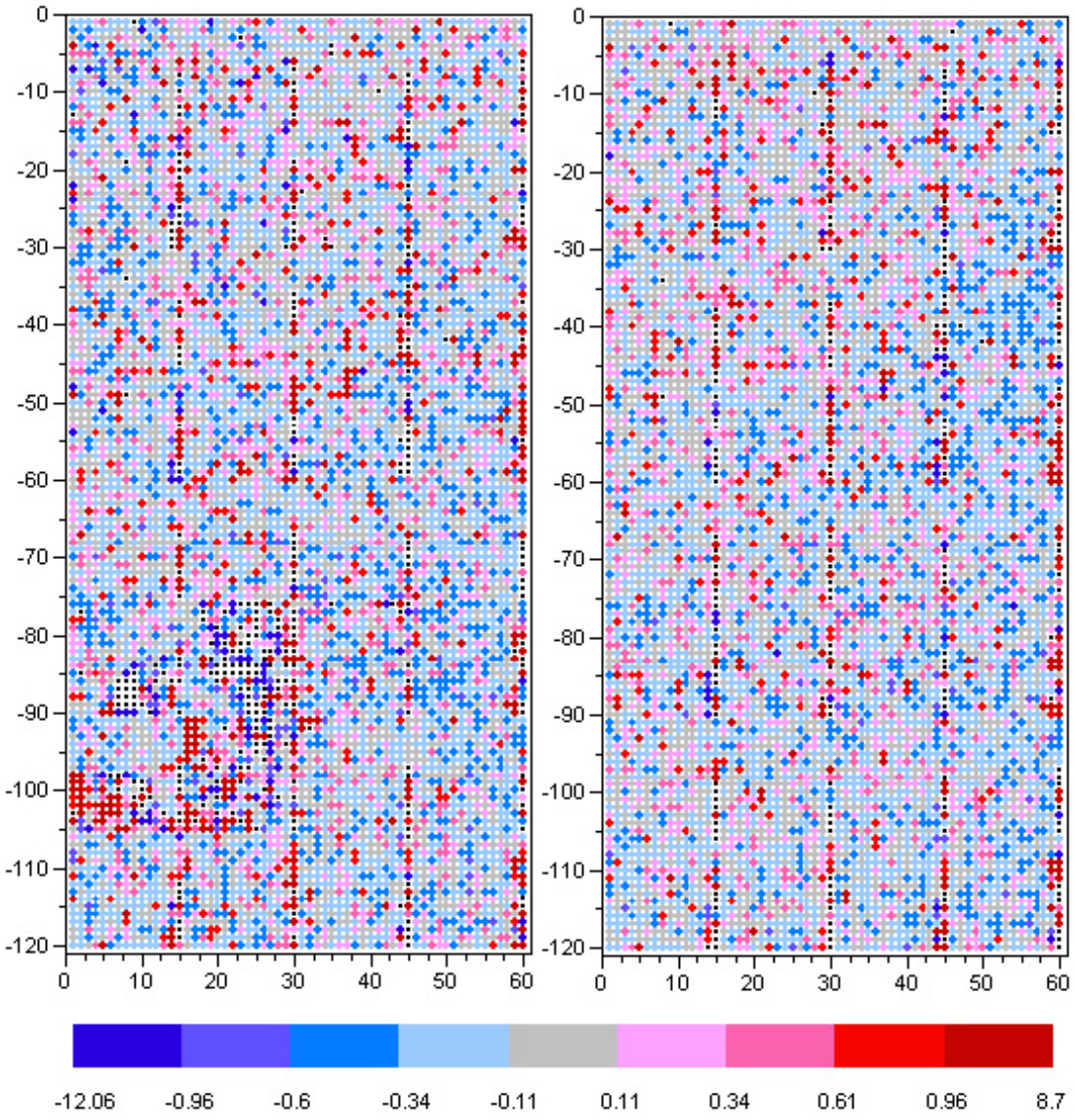

0
-10
-20
-30
-40
-50
-60
-70
-80
-90
-100
-110
-120
0
10
20
30
40
50
60
-12.06
-0.96
-0.6
-0.34
-0.11
0.11
0.34
0.61
0.96
8.7

Figure 5.3

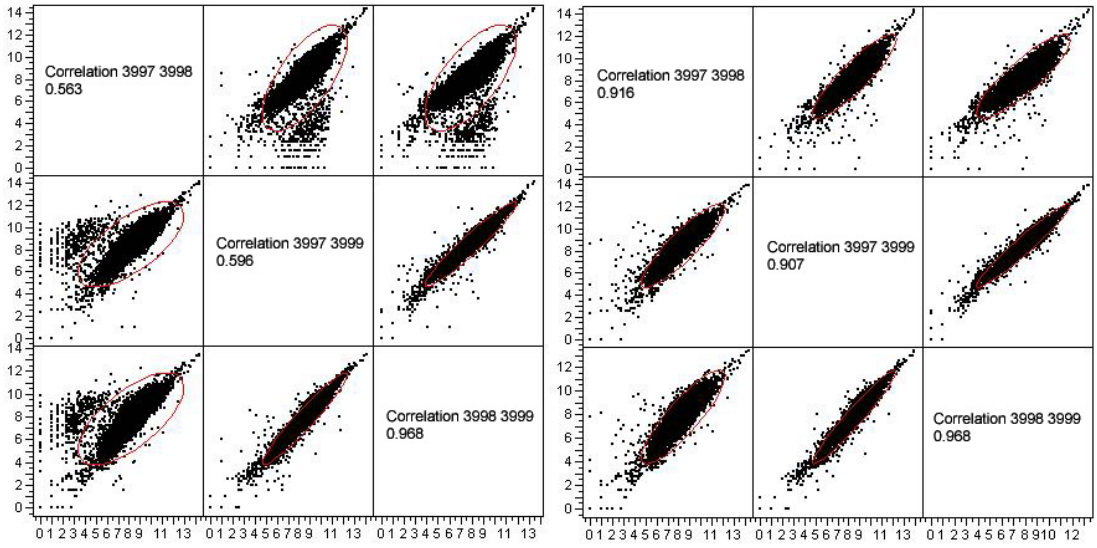

Correlation 3997 3998
0.563
Correlation 3997 3999
0.596
Correlation 3998 3999
0.968
Correlation 3997 3998
0.916
Correlation 3997 3999
0.907
Correlation 3998 3999
0.968

Figure 5.4

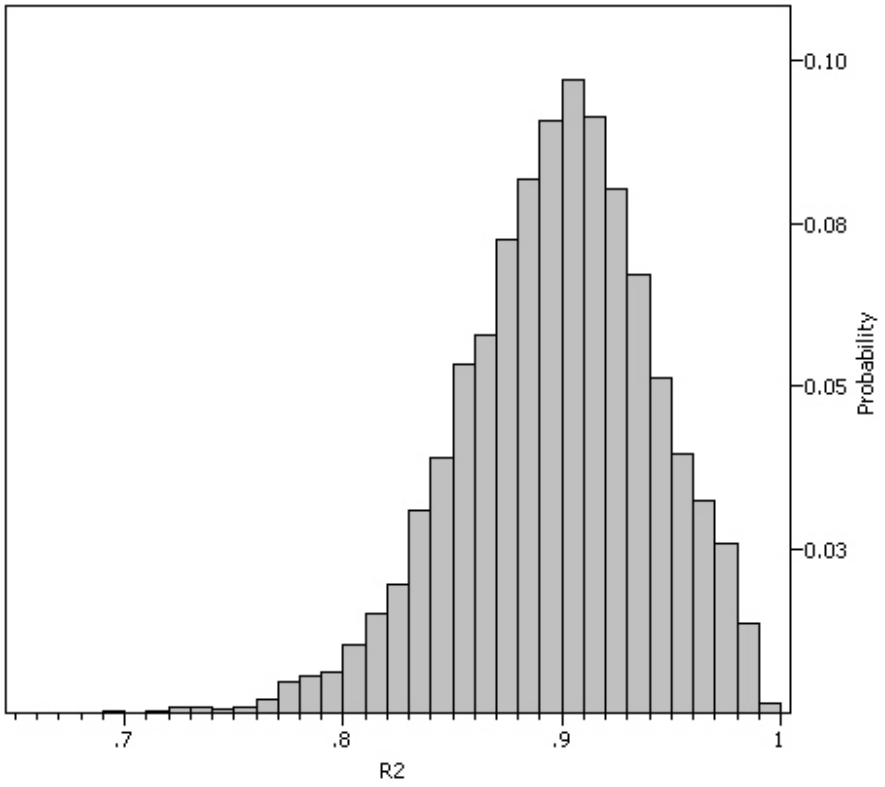


Figure 5.5

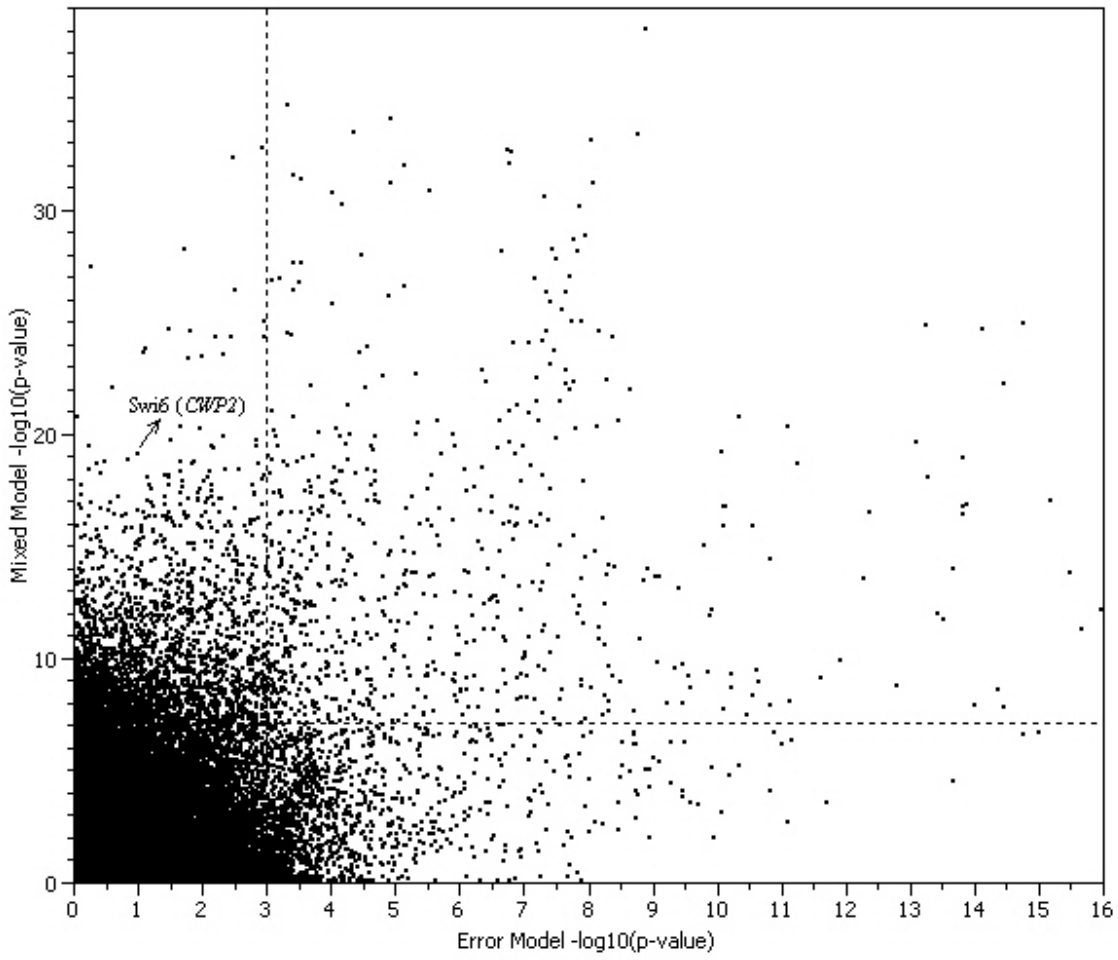
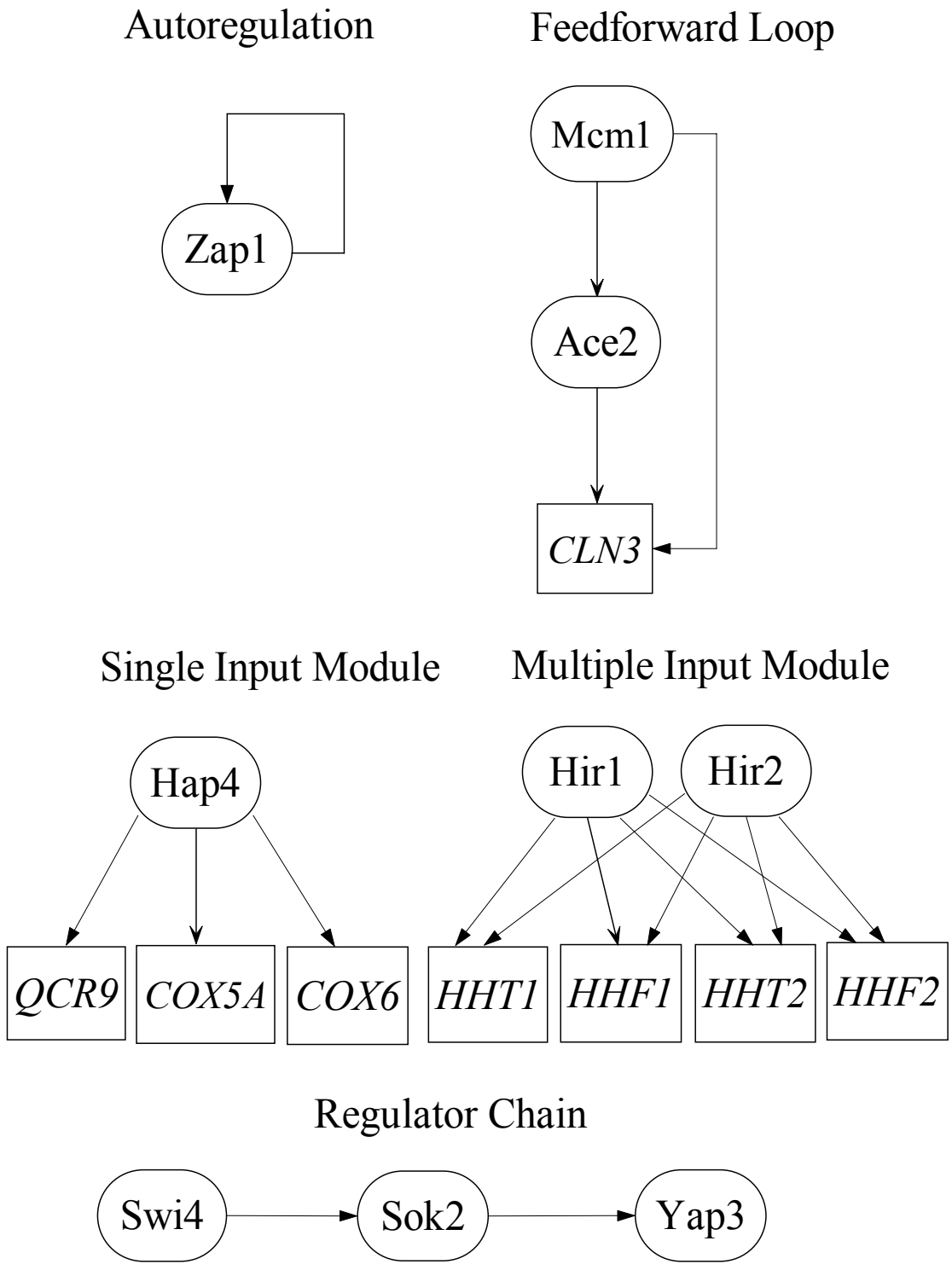

Swi6 (CWP2)
Mixed Model -log10(p-value)
Error Model -log10(p-value)
0
10
20
30
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Figure 5.6


Autoregulation
Zap1
Feedforward Loop
Mcm1
Ace2
CLN3
Single Input Module
Hap4
QCR9
COX5A
COX6
Multiple Input Module
Hir1
Hir2
HHT1
HHF1
HHT2
HHF2
Regulator Chain
Swi4
Sok2
Yap3