Abstract

HSIEH, WEN-PING. Analysis of Gene Expression Profiles with Linear Mixed Models. (Under the direction of Dr. Greg Gibson and Dr. Russ Wolfinger)

With the emergence of high throughput technology, proper interpretation of data has become critical for many aspects of biomedical research. My dissertation explores two major issues in gene expression profile microarray data analysis. One is quantification of variation across and among species and its effect on biological interpretation. The second part of my work is to develop better statistical estimates that can account for different sources of variation for significant gene detection.

A previously published dataset of oligonucleotide array data for three primate species was analyzed with linear mixed models. By decomposing the variation of expression into different explanatory factors, the differences among species as well as between tissues was revealed at the expression level. Issues of cross-species hybridization and expression divergence compared to mutation-drift equilibrium were addressed.

The power and flexibility of the linear mixed model framework for detection of differentially expressed genes was then explored with a dataset that includes spiked-in controls. The impact of probe-level sequence variation on cross-hybridization was detected through a Gibb's sampling method that highlights potential problems for short oligonucleotide microarray data analysis. A motif as short as fifteen bases can possibly cause significant cross-hybridization.

Finally, a bivariate model using information from both perfect match probes and mismatch probes was proposed as a means to increase the statistical power for detection of significant differences in gene expression. The improved performance of the method was demonstrated through Monte Carlo simulation. The detection power can increase as much as 20% with 5% false positive rate under certain circumstances.

ANALYSIS OF GENE EXPRESSION PROFILES WITH LINEAR MIXED MODELS

By

WEN-PING HSIEH

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, NC
2005

APPROVED BY:

_____ _____

Chair of advisory committee Co-chair of advisory committee

_____ _____

To my parents,

僅以此獻給我親愛的父母謝榮良先生與謝吳麗環女士

Biography

Wen-Ping Hsieh received her Bachelor of Science degree in Mathematics from National Taiwan University at Taipei, Taiwan in 1995. She completed graduate work in Statistics at National Tsing Hua University at Hsinchu, Taiwan, receiving her Masters degree in 1998. In 2001, she commenced doctoral work at North Carolina State University with major in Bioinformatics.

Acknowledgement

I would like to express my sincere appreciation to my advisor Dr. Greg Gibson for the time and guidance he provided. His enthusiasm for research and intelligent ideas inspired me all the time. Even though I can never hope to equal him, the seeds have been planted in my mind. I would also like to thank my co-advisor Dr. Russ Wolfinger for his great input in my research work. I would not have achieved so much in such a short time if it was not for his prompt responses and tons of insightful suggestions.

I would like to thank Dr. Spencer Muse for his encouragement and help. He has been very supportive throughout my studies whether managing the funding, discussing research ideas or simply telling me "you'll be fine". I also benefited a lot from Dr. Dennis Boos' profound knowledge of statistics. He was always very patient with my questions and I am very grateful for his suggestions.

Special thanks go to Dr. Tzu-Ming Chu. He is the one that introduced me to the field of microarray data analysis. I am very grateful for all the assistance he gave me and I really enjoyed the fruitful discussions we had over the past years. He and Fang Chen have been my best support in both my life and my studies.

I thank my beloved Chun-Ming Kuo for his wonderful support.

None of this would be possible if I did not have the love and unconditional support of my family.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

Microarray technology is widely adopted in all levels of biological research. The application extensively covers drug efficacy screening (Jain 2000), prediction of gene function (Zhang et al. 2004), temporal profiling (Cavallaro et al. 2002), and studies of regulatory networks (Lee 2002). It is also an important tool in pharmaceutical and clinical research. While monitoring the expression levels of several thousands of genes at the same time with only limited amount of mRNA, it accelerates the discovery of disease genes and helps us to understand the biological system as a whole. The extraordinarily large amounts of data generated from microarray studies raise considerable challenges for storage and handling of the data. Systematic statistical strategies for correctly retrieving information from the massive datasets have become very critical to the drawing of solid conclusions at the genomic scale.

The first important issue is the experimental design. A lot of suggestions and alternatives have been provided. Yang et al. reviewed the pros and cons of several different methods (Yang et al. 2002b) and evaluated the designs based on their efficiency and purpose. Kerr et al. (2000) provided a detailed discussion of different sources of variation and compared the two most widely used designs, the reference design and loop design. There is also thorough discussion of the number of replicates and proper arrangement of replication in these studies.

Statistical analysis of microarray data includes image processing and signal acquisition at the front end (Yang 2002a). After all the preprocessing steps have been completed, there are different approaches for mining information from the data.

Clustering methods are often used to discover hidden patterns. (Eisen et al. 1998, Tibshirani et al. 1999, Dudoit et al. 2000a); these are called unsupervised learning in information science. Most of the well-developed statistical methods such as hierarchical clustering, k-means clustering, self-organizing maps, and principal component analysis have been applied to seek potential sub-types of tumor samples or to group genes that might be co-regulated (Alizedeh et al. 2000, Eisen et al. 1998). A second category of pattern searching is supervised learning methods, which are also generally known as classification methods. It is often very useful to associate samples with disease types (Ben-Dor et al. 2000, Golub et al. 1999) and to classify genes according to their functional roles (Brown et al. 2000).

1.1 Normalization

Due to the nature of the technology itself, there are several sources of technical variation relating to RNA extraction, reverse transcription, labeling and fluorescence detection. The variation can be systematic, which means that they have a similar effect on many measurements and can be explicitly accounted for. Alternatively, it can also be largely stochastic with the result that it cannot be captured according to current knowledge. The more we know about where the variation comes from, the better we can integrate the sources as systematic effects and account for them.

There are many different ways to correct for systematic noises. The simplest way is to adjust globally for the mean or median (Yang 2002c, Quackenbush 2001). The basic assumption is that most genes are not differentially expressed across different conditions. Hence the mean and median values for all arrays should be approximately

the same. If this assumption is violated, another measure for global noise reduction is to use the information from a set of invariant genes (Li and Wong 2001b). They assume that the rank of probe intensity from non-differentially expressed genes within each array does not change appreciably across arrays, and use this invariant set to determine the normalization relationships among arrays. Another category of normalization is intensity-based, such as Loess transformation (Yang et al. 2002c). The most flexible and comprehensive global normalization method is the ANOVA approach (Kerr et al. 2000, Wolfinger et al. 2001), which can easily integrate all possible sources of noises as effects in the model.

1.2 Significant gene detection

One of the most important goals for microarray experiments is to identify genes that are differentially expressed across different conditions such as treatment and control, different tissues, species, developmental stages, pathological conditions, etc. In two group comparisons, fold change is generally used by biologists to assess the difference between groups (Draghici 2002, DeRisi 1996). It is simple and intuitive. However, the behavior of genes at high levels of expression is different from that observed at low expression levels. There tends to be a lot more variation in low expression levels (Draghici 2002, Rocke and Durbin 2001). A high fold change between groups for a gene might result from randomness and need not have statistical significance.

T-tests take one further step to assess the statistical significance of differences between groups (Dudoit et al. 2000b, Pan 2002, Claverie 1999). This univariate test can also be adjusted for multiple comparisons as described in Dudoit et al (2002b). Another

approach to detect differentially expressed genes is to use model-based clustering methods such as mixture models (Lee et al. 2000). A similar approach is to assay the probability that a given gene is truly differentially expressed between two conditions using empirical Bayes analysis (Efron 2001).

1.3 ANOVA analysis and Two step Linear Mixed Model analysis of microarray data

Kerr et al. were among the first to use an ANOVA approach for microarray data analysis. (Kerr et al. 2001). Their model puts all genes into one comprehensive model and decomposes the variation of expression into different sources such as the dye effect, pin effect and batch effect, etc. It can estimate the variation at the global level as well as the gene level. This approach seems to be promising, but it is computationally prohibitive in most large experiments in which the increased number of genes and conditions increases the number of parameters to be estimated. Two-step analyses (Wolfinger et al. 2001, Wu et al. 2003) have been proposed as alternatives to the full model analysis. Another drawback is that fixed effect ANOVA cannot comprehensively account for all kinds of random factors so an extension proposed by Wolfinger et al (2001) is the use of mixed models.

Mixed models are known for their ability to accommodate correlation through incorporation of random effects. For example, all the estimates from the same array share the same background intensity caused by the manufacturing process and hence are correlated. Fixed effect models treat the effects in the model as if the same categories or levels will be repeated exactly if we do the same experiment again. Some effects, such as slide effects, are better modeled as random effects, since the arrays are not reusable and

the levels will not be exactly repeated the next time we do the same experiment. We can only assume that similar effects will be drawn from the same population (Cui et al. 2002). The method proposed by Wolfinger et al. (2001) uses two interconnected sequential mixed models. The first model accounts for global effects such as the dye effect, slide effect, print tip effect, etc. The residuals from the first model become the response variable in the second model. The second model assesses the gene specific effects. It associates the variation with the varieties we are interested in and makes inference based on *F*-statistics or *t*-statistics to detect genes that are differentially expressed among conditions. Mixed model analysis provides a comprehensive and flexible framework with which to identify all kinds of systematic sources of variance.

Although the two-step mixed model is only equivalent to the one step mixed model under certain circumstances (Yang 2003), it may not make much difference in practice (Wolfinger 2001). While the one step model has a common error model for all genes, the two step model can evaluate the error terms independently for each one and can accommodate more information.

1.4 Basics of Affymetrix arrays

The Affymetrix GeneChip® is currently the most popular commercial type of array. Each gene is represented by 10 to 20 short oligonucleotide sequences, each of which is 25 bases long and complementary to the target sequence (Lockhart 1996). These short sequences are called perfect match (PM) probes. Each PM probe is accompanied by a mismatch (MM) probe, which only differs from its associated PM probe with a single base mismatch in the middle of the sequence. The single base mismatch is assumed to

disable any hybridization to the target sequence and the mismatch probes are designed to capture the comparable level of non-specific binding and systematic noise as perfect match probes.

1.5 How to incorporate mismatch probes

There are many different ways to combine the intensity measure from PM and MM to interrogate target gene expression. Average difference (AvgDiff) is the first approach adopted by Affymetrix as their default analysis in Affymetrix Micro Array Suite (MAS 4.0). This assumes that the difference PM-MM is proportional to the true signal, and takes the weighted average across probes within a probe set as the measurement for the target gene. The fact that PM-MM is not always positive causes problems for log-scale analysis (Hubbell et al. 2002, Irizarry et al. 2003, Naef 2002). Affymetrix therefore updated their algorithm in MAS 5.0 using Tuckey's biweight function (Hubbell et al. 2002). This approach has been criticized for inefficient management of the MM data, among other concerns (Irizarry et al. 2003).

Li and Wong's also devised a method that is based on PM-MM (Li and Wong 2001a). It analyzes the data at the probe level and decomposes the signal into the product of a probe sensitivity index and the gene expression index. Their software dChip has been used in various genomic studies.

Irizarry et al. proposed a background subtraction based on the mode of MM distribution instead of a probe-by-probe subtraction (Irizarry et al. 2003). This RMA method is widely used to assess the array quality. Many empirical studies have indicated that the MM probes capture a lot of true signal, making them problematic for background

noise detection, so several studies have focused solely on PM data (Li and Wong 2001a, Zhou 2002). Chu et. al. adopted a linear mixed model with MM intensity as a covariate, but this approach does not produce much improvement from the PM-only model (Chu et al. 2002). I propose a bivariate model in Chapter 4 following Chu's study, in order to explore the best way of using the MM data under the framework of linear mixed models. This approach is conceptually similar to Wu's study (Wu et al. 2004), who essentially decomposed the intensity of PM into true signal, non-specific binding and optical noise and the intensity of MM into only non-specific binding and optical noise. They modeled the non-specific binding from PM and MM as a bivariate normal distribution. The main difference between our approaches is that my model can dynamically adjust for the correlation between PM and MM based on classical linear mixed models, while they used an constant empirical value of 0.7 in their model.

1.6 Specificity of oligonucleotide arrays

Short probe sequences on the oligonucleotide arrays create cross-hybridization issues. It has been suggested that such problems can be avoided by paying attention to sequence similarity at the probe design stage (Rouillard et al. 2002). However, since most samples in the microarray studies may be from different individual organisms, the complexity of the sequences in the whole genome provides few clues about the specificity problem. A spike-in experiment described in Chapter 3 provides some transcript information that enables close evaluation of the problem. My analysis in Chapter 3 shows that short motifs can provide strong hybridization strength and cause cross-hybridization. Some sequences are more susceptible to non-specific binding than others.

Some recent studies have attempted to assess whether sequences are "sticky" or not using thermodynamic rules (Zhang et al. 2003, Wu et al. 2004, Naef et al. 2003, Mei et al. 2003). The free-energy model proposed by Zhang is based on the nearest-neighbor model with some modifications to compute the gene-specific binding energy and nonspecific binding energy. Naef et al. simplified the model and considered only the sequence composition of the probes. The affinity for non-specific binding was modeled as a function of position with a polynomial of degree 3 for each base. Their approach is arguably successful to this end. Such concepts are also integrated in the probe design of Affymetrix arrays (Mei et al 2003).

Another issue related to sequence specificity is the study of cross-species hybridization. Since the effort needed to make the whole genome array for a new organism is huge, pilot studies are usually carried out on the arrays of a closely related species. Except for newly explored species, comparative genomic studies sometimes also utilize the comparison of different species on the same arrays. Even if it is a preliminary study for screening purpose, it is still desirable to retrieve the most accurate information possible. If we have sequence information for both species that contribute to the arrays and the samples, a straightforward approach would be to delete the probes with mismatches among the species studied. However, we do not have complete sequence information for most newly explored species, and nor do we have information from different strains. An alternative method proposed by Ji et al (2004) uses only probes with significant signals across all samples. This approach can be criticized on the basis that it wastes a great proportion of useful information. My study in Chapter 2 uses a specific feature of Affymetrix arrays, the probe profiles, to control for sequence variation. This

issue was first highlighted in Li and Wong's method (Li and Wong 2001a). Based on the heuristic rules that I propose, "bad" probes that might have polymorphisms or simply do not function are filtered out of the analysis.

1.7 Motivation and research outline

1.7.1. Expression variation of primates

Why human beings look and behave differently from their close relatives, the chimpanzee, gorilla, and orangutan, has been an intriguing question for a long time. Since King et al. (1975) concluded that humans differ from chimpanzees at the regulatory level of gene activities, people have been trying to characterize gene expression differences in different ways. An experiment from Paabo's group was the first study that addressed this question with information from tens of thousands of genes. (Enard et al. 2002) Their experiment included three primate species and two tissues. The conclusion they drew from the analysis of the data using phylogenetic methods favors the idea that gene expression shows accelerated evolution in human brains.

They first generated evolutionary trees based on the expression of 12,600 genes. The distance metrics they used was essentially weighted by the fold change between the expression measurements of individuals. While the topology of their evolutionary tree is confirmed when we reproduced their results, my approach with linear mixed models provides a different perspective on this data set and the conclusions are not necessarily consistent with theirs.

The second issue addressed in this study (Hsieh et al 2003a) was the degree of cross-species hybridization. Since there is difference at the sequence level for the species

studied, the hybridization of great ape target cDNA to the human Affymetrix arrays might not give correct signals. Lower intensity from non-human species might be caused by the lack of hybridization instead of lower transcript abundance. From the observation of consistent probe patterns, we used a heuristic rule to filter probes with potential polymorphisms. The conclusions from the biological analysis regarding relative divergence of gene expression in brain and liver nevertheless remain valid after filtering the probes.

Since I observed significant species variation for a significant proportion of the genes, the third question I asked is whether there are transcripts undergoing some natural selection. Based on the limited information we can have from the small sample, only a few genes that diverged between species greater than expected under mutation-drift balance were detected.

1.7.2. Cross-hybridization detection of Latin Square data

There are many different ways to analyze microarray data. Most of the proposed methods have been demonstrated through experiments on model organisms and have been evaluated by simulation or biological evidence. These only provide indirect support of how well the statistical algorithms fit data to specific scenarios. To enable a fair comparison across different methods, we needed a dataset with known answers. To this end, Affymetrix provided a spike-in experiment that can be used to evaluate various statistical methods (Affymetrix 2001). This experiment was in the form of a standard Latin Square design. There are 14 spike-in transcripts with concentration levels from 0 to 1024 pM. Each of the 14 experiments contains each of the 14 transcripts with each of the

14 concentration levels appearing only once. The Latin Square design allows one to evaluate the main effects of interest using the fewest samples needed, while leaving the higher interaction effects confounded with some of the main effects. Dye-swapping design for cDNA arrays are essentially a Latin Square design with balanced combination of dyes and samples (Kerr et al. 2000).

I demonstrate the accuracy of significant gene detection with linear mixed models. From the analysis I found several unexpected genes showing one of the 14 patterns of the designed transcripts. Through the exploration of sequence similarity with Gibb's sampling, some short motifs were detected that possibly contribute to the inferred cross-hybridization. The work was published in the proceeding of CAMDA 2002 (Hsieh et al, 2003b).

1.7.3. Bivariate model versus PM-only model

Affymetrix arrays are currently the most popular commercial platform used for microarray experiments. Half of the probes on the arrays are designed to catch the real signals and the other half are aimed at catching the background noise. From empirical studies we know the truth somewhat deviates from the assumptions of this design principle. Mismatch probes capture some of the real signal all the time, so cannot be simply subtracted from the perfect match signals. Since mismatch probes do not play a consistent role in background noise detection and vary according to the sequence properties, most recommended current statistical methods ignore the mismatch probes and use only the perfect match probes. I propose a bivariate model in Chapter 4 to incorporate both signals together. The correlation between PM and MM can be

dynamically adjusted and hence the mismatch probes can provide the information for the true signals according to the data itself. Including more data points can also increase the accuracy of parameter estimates, and this increase in power is demonstrated through simulation.

Reference

Affymetrix, (2001), New statistical algorithm for monitoring gene expression on GeneChip probe arrays. http://www.affymetrix.com/support/technical/technotes/statistical_algorithms_technote.pdf

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson J Jr, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO, Staudt LM. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature, 403:503-510.

Ben-Dor, L. Bruhn, I. Nachman, M. Schummer, and Z. Yakhini. (2000) Tissue Classification with Gene Expression Profiles. Journal of Computational Biology, 7:559--584

Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Nat Acad Sci USA 97: 262-267.

Chu TM, Weir BS, and Wolfinger RD (2002). A Systematic Statistical Linear Modeling Approach to Oligonucleotide Array Experiments. Mathematical Biosciences, 176: 35-51.

Cavallaro S, D'Agata V, Manickam P, Dufour F, Alkon DL. (2002) Memory-specific temporal profiles of gene expression in the hippocampus. Proc Natl Acad Sci U S A. 99(25):16279-84.

Claverie, J. (1999) Computational methods for the identification of differential and coordinated gene expression. Hum Mol Genet 8: 1821-1832.

Cui, X, Churchill, G. A. (2003), Statistical tests for differential expression in cDNA microarray experiments. Genome Biology 4, 210.1–210.10.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet. 14(4): 457-60

Draghici S (2002) Statistical intelligence: effective analysis of high-density microarray data. Drug Discovery Today, 7(11):S55-S63

Dudoit S, Fridlyand J, Speed TP (2000a) Comparison of discrimination methods for the classification of tumors using gene expression data. Technical Report, Department of Statistics, UC-Berkeley.

Dudoit S, Yang YH, Callow MJ, Speed TP (2000b) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Technical Report, Department of Statistics, UC-Berkeley.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, Giavalisco P, Nieselt-Struwe K, Muchmore E, Varki A, Ravid R, Doxiadis GM, Bontrop RE, Paabo S. (2002) Intra- and interspecific variation in primate gene expression patterns. Science. 296(5566):340-3.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc. 96 1151–1160.

Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Nat 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.

Hsieh WP, Chu TM, Wolfinger R.D., Gibson G. (2003a) Mixed-model reanalysis of primate data suggests tissue and species biases in oligonucleotide-based gene expression profiles. Genetics. 165(2):747-57.

Hsieh, WP., Chu, TM., and Wolfinger, R.D. (2003b) Who are those strangers in the Latin square? In Methods of Microarray Data Analysis III ed. Johnson, K.F. and Lin, S.M., Kluwer, 199-208.

Hubbell E, Liu WM, Mei R. (2002) Robust estimators for expression analysis. Bioinformatics. 18(12):1585-92.

Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4: 249-264.

Jain KK. (2000). Applications of biochip and microarray systems in pharmacogenomics. Pharmacogenomics 1:289307

Ji, W., Zhou, W., Gregg, K., Yu, N., Davis, S., Davis, S. (2004) A method for cross-species gene expression analysis with high-density oligonucleotide arrays. Nucleic Acids Research; 32(11): e93

Kerr, M.K., Martin, M. and Churchill, G.A. (2000). Analysis of variance for gene expression microarray data. Journal of Computational Biology, 7:819-837.

King MC, Wilson AC. (1975) Evolution at two levels in humans and chimpanzees. Science. 188(4184):107-16.

Lee MLT, Kuo FC, Whitmore GA, Sklar J (2000) Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Nat Acad Sci USA 97: 9834-9839.

Lee, T.I., Rinaldi, N.J., Robert, F., Odom, D.T., Bar-Joseph, Z., Gerber, G.K., Hannett, N.M., Harbison, C.R., Thompson, C.M., Simon I., Zeitlinger J., Jennings, E.G., Murray, H.L., Gordon, D.B., Ren, B., Wyrick, J.J., Tagne, J., Volkert T.L., Fraenkel, E., Gifford D.K., and Young, R.A. (2002) Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science 298: 799-804.

Li, C and Wong, W.H. (2001a) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, Proc. Natl. Acad. Sci. Vol. 98, 31-36.

Li, C and Wong, W.H. (2001b) Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application, Genome Biology 2(8): research0032.1-0032.11

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 Biotech 14: 1675-1680.

Mei R, Hubbell E, Bekiranov S, Mittmann M, Christians FC, Shen MM, Lu G, Fang J, Liu WM, Ryder T, Kaplan P, Kulp D, Webster TA. (2003) Probe selection for high-density oligonucleotide arrays. Proc Natl Acad Sci U. S. A. 100(20):11237-42.

Naef, F; Hacker, CR; Patil, N; Magnasco, M (2002) Empirical characterization of the expression ratio noise structure in high-density oligonucleotide arrays. Genome Biol., 3, research0018.

Naef F. and Magnasco MO., (2003) Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays. Physical Review E, 68:011906.

Pan W. (2002). A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. Bioinformatics 18, 536-545.

Quackenbush J. (2001) Computational analysis of microarray data. Nat Rev Genet. 2(6):418-27.

Rocke DM, Durbin B (2001). A model for measurement error for gene expression arrays. J. Computational Biology. 8:557–569

Rouillard JM., Herbert C and Zuker M. (2002) Oligoarray: Genome-scale oligonucleotide design for microarrays. Bioinformatics (Applications Note), 18:486–487.

Tibshirani R, Hastie T, Eisen M, Ross D, Botstein D, Brown P (1999) Clustering methods for the analysis of DNA microarray data. Technical Report, Department of Statistics, Stanford U.

Wolfinger, R., Gibson, G., Wolfinger, E., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology 8, 625-637.

Wu, H., Kerr, M.K. and Churchill, G.A. (2003). The analysis of gene expression data: methods and software. chapter MAANOVA: A Software Package for the Analysis of Spotted cDNA Microarray Experiments. Springer

Wu, Z, Irizarry, RA, Gentleman, R, Martinez Murillo, F, Spencer, F (2004) A Model Based Background Adjustement for Oligonucleotide Expression Arrays. JASA Vol. 99, Iss. 468; p. 909

Yang, X. (2003) Optimal Design of Single Factor cDNA Microarray Experiments and Mixed Models for Gene Expression Data. Dissertation of Virginia Polytechnic Institute and State University.

Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002a). Comparison of methods for image analysis on cDNA microarray data. Journal of Computational and Graphical Statistics 11, 108-136.

Yang Y. H., Speed T., (2002b) Design issues for cDNA microarray experiments. Nat Rev Genet. 2002 Aug;3(8):579-88

Yang Y. H., Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. (2002c) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30(4):e15.

Zhang, L., Miles, M. F., and Aldape, K. D. (2003) A model of molecular interactions on short oligonucleotide microarrays: implications for probe design and data analysis. Nature Biotechnology, 21:818-821

Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng WT, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR. (2004) The functional landscape of mouse gene expression. J Biol. 3(5):21.

Zhou Y and Abagyan R. (2002) Match-Only Integral Distribution (MOID) Algorithm for high-density oligonucleotide array analysis. BMC Bioinformatics, 3(1):3

Chapter 2 Mixed Model Reanalysis of Primate Data Suggests Tissue and Species Biases in Oligonucleotide-Based Gene Expression Profiles

2.1 Abstract

An emerging issue in evolutionary genetics is whether or not it is possible to use gene expression profiling to identify genes that are associated with morphological, physiological, or behavioral divergence between species, and whether these genes have undergone positive selection. Some of these questions were addressed in a recent study (Enard et al, 2002) of the difference in gene expression between human, chimp and orangutan that suggested an accelerated rate of divergence in gene expression in the human brain relative to liver. Reanalysis of the Affymetrix dataset using analysis of variance methods to quantify the contributions of individuals and species to variation in expression of 12,600 genes indicates that as much as one quarter of the genome shows divergent expression between primate species at the 5% level. The magnitude of fold change ranges from 1.2 –fold up to 8-fold. Similar conclusions apply to reanalysis of Enard et al's (2002) parallel murine dataset. However, biases inherent to short oligonucleotide microarray technology may account for some of the tissue and species effects. At high significance levels, more differences were observed in the liver than the brain in each of the pairwise species comparisons, so it is not clear that expression divergence is accelerated in the human brain. Further, there is an apparent bias toward up regulation of gene expression in the brain in both primates and mice, whereas genes are equally likely to be up- or down-regulated in the liver when these species diverge. A small subset of genes that are candidates for adaptive divergence may be identified on the basis of a high ratio of inter-specific to intra-specific divergence.

2.2 INTRODUCTION

One of the most interesting applications of gene expression profiling in evolutionary genetics is the comparison of transcript abundance between closely related species. Given that studies of yeast, flies and killifish have each suggested that between 10 and 25% of the transcriptome differs significantly in expression level between any two individuals of the same species (CHEUNG and SPIELMAN, 2002; JIN *et al*, 2001; OLEKSIAK *et al*, 2002; CAVALIERI *et al*, 2000), there is an expectation that a similar fraction of the transcriptome may differ between sibling species. Some of these differences will be associated with morphological, physiological, and behavioral diversification, and if causally related to the divergence, may also provide signatures of natural selection. Quantification of transcript abundance within and between species thus has much to contribute to our understanding of the evolutionary forces acting at the level of gene expression.

The first effort to address these questions in relation to human evolution was recently published by Pääbo and coworkers (ENARD *et al*, 2002). The centerpiece of their study was a comparison of 12,600 gene expression profiles of left prefrontal lobe brain samples (Brodmann Area 9) from three humans, three chimpanzees, and an orangutan, each of which had died of natural causes, using Affymetrix U95A oligonucleotide gene chips. They also examined liver samples from the same specimens, and conducted a parallel series of experiments with *Mus musculus*, *M. spretus*, and *M. caroli*, three mouse species that show similar levels of genetic divergence. After

computing a pairwise distance matrix based on the average level of expression for each gene on their arrays, they drew neighbor-joining trees that summarize the overall divergence in transcript abundance for each tissue between the triplets of species. Their major finding was that the branch joining the three human samples to the central node on their brain expression tree was almost twice as long in relative terms as the same branch on the liver tree or in either of the murine trees. The same result was obtained with a smaller experiment using cDNA microarrays. Hence the authors concluded that gene expression had diverged most rapidly in the human brain.

Although it is easy to criticize this study over concerns such as the small sample size, the suitability of senescent individuals, and the validity of extrapolating to general conclusions on the basis of a small section of the brain, it is also the case that the already rich dataset will support further quantitative analyses that may be of interest. In the reanalysis of the data reported here, we sought to address the following questions: how many of the genes on the array are actually significantly more divergent between than within species; what is the mean magnitude of expression divergence between species; why did one of the human samples have an average difference from the other two that was as great as their overall divergence from the chimps; what is the nature of the genes that have diverged in expression; and do the same genes diverge between all three species? Our major finding is that gene expression actually diverges more between human and chimp liver samples than brain.

In the course of our analyses, we also noticed biases in the directionality and significance of changes in expression that led us to question whether the Affymetrix technology is really suitable for interspecific comparisons. We implemented mixed model analysis of variance in SAS (WOLFINGER *et al*, 2001; CHU *et al*, 2002) to tease apart the contributions to transcript abundance of variance among individuals and between species. Fluorescence intensity from each individual perfect match oligonucleotide probe was taken as the measure of expression, rather than the average difference between perfect and mismatch probes. A detailed analysis of a subset of the genes that showed strong species-by-probe interaction effects highlights some of the difficulties associated with the use of oligonucleotide arrays to compare genotypes that diverge at the nucleotide level. Consequently, our results also have implications for the interpretation of Affymetrix data for any comparisons of genetically polymorphic strains.

2.3 MATERIALS AND METHODS

Mixed Model Analysis of Variance: Variation in gene expression was assessed using a two-step strategy essentially as outlined in CHU *et al* (2002). In the first step, each individual probe measurement was centered relative to the array mean by subtracting the log_2 transformed value of the intensity from the mean log_2 value for the probes on the array. We simply used the Perfect Match (PM) intensities, and ignored the Mismatch (MM) values as we find that these statistically generally just add noise. Data quality was then checked by plotting pairwise scatter plots of the normalized probe intensities for each possible comparison of similar treatments. See Supplementary Figure 2.1 for the six human brain samples. All contrasts show good linear correlations as expected. Values

range from -2 to +4, with the upper limit indicating saturated signal intensity. Almost 0.9% of the probes were at this level, undoubtedly reducing the power of comparisons involving highly expressed genes. As generally observed, transcript abundance is skewed toward a large number of genes showing relatively low abundance, hence the skewed distribution of intensity values about the mean. All pairwise comparisons contrasting individual number 2 with individuals number 1 or 3 have broader scatter plots, reflecting the reported observation that this individual has more divergent expression in the brain than the other two individuals. Similar saturation values were seen for the other tissues and species.

The second modeling step was to fit gene-specific mixed models using PROC MIXED in SAS as follows:

$$\log_2(PM_{ijkl}) = S_i + T_j + P_k + ST_{ij} + TP_{jk} + SP_{ik} + R_{l(ij)} + \varepsilon_{ijkl},$$

PM_{ijkl} denotes the perfect match expression measurement of the *k*th probe of the *l*th individual for *i*th species (human, chimp, or orangutan) in the *j*th tissue (brain or liver). The symbols *S, T, P* represent the fixed effects of species, tissues, and probes respectively. Individual effects within species were specified as random effects, and assumed to be independent and identically distributed according to a normal distribution with mean zero and variance σ_r^2. The ε_{ijkl}'*s* were also specified as independent and identical normal distributions with mean zero and variance σ^2 that are independent of the $R_{l(ij)}$'*s*. For the comparison in Figure 2.6, variance components of species effects were zero for a large fraction of genes, so we instead present results of a simplified general linear model run with PROC GLM in SAS on the reduced data consisting of human and chimp brain arrays only.

Correlation coefficients for filtering probes: Correlation coefficients were calculated between log_2 transformed human brain PM intensity values (a_{ijk}) and corresponding chimp brain values (b_{ijk}) from the *i*th individual, *j*th sample and *k*th probe associated with a particular gene. The expression profile for each species was first calculated as the average probe measure among samples: $a_{..k} = \sum_j \sum_i a_{ijk} /(I * J), b_{..k} = \sum_j \sum_i b_{ijk} /(I * J)$. Subsequently, the correlation coefficient between the expression profiles of those two species was computed as: $(\sum_k a_{..k} b_{..k} - K \sum a_{...} b_{...}) / (\sqrt{(\sum_k a_{..k}^2 - K a_{...}^2)} \sqrt{(\sum_k b_{..k}^2 - K b_{...}^2)})$ where $a_{...}$ is the average of $a_{..k}$ and $b_{...}$ is the average of $b_{..k}$.

Outlier probes were then deleted systematically until the correlation exceeded 0.95. For example, removal of the single inconsistent probes in Figure 2.4C and 2.4D results in a large increase in the overall correlation between human and primate data. The species effect for the remaining probes is consistent, and likely represents a better measure of the true difference in gene expression.

Neighbor-joining trees: Euclidean distance matrices were computed for each pair of arrays on the least squares mean gene expression measures from the mixed model analysis, and rescaled to fit the format required by the package PHYLIP (Felsenstein, 1989). Neighbor-joining trees were generated by the NEIGHBOR option with default settings. This analysis is similar to that of ENARD et al (2002), except that they used average distance measures computed by Affymetrix software.

Crude estimation of the fraction of genes diverging under positive selection:

Following RIFKIN *et al* (2003) and LYNCH and HILL (1986), under mutation-drift equilibrium, the expected squared difference between species for each gene expression level is $\sigma_m^2 t$, where σ_m^2 is the mutational variance and t the time in generations since separation, and the expected level of intraspecific variance, which is assumed to have remained constant in both lineages since divergence, is $2N_e\sigma_m^2$, where N_e is the effective population size. Then the ratio of mean square estimates of the species and individual within species effects $F_{human\text{-}chimp} \sim [MS_{species}/MS_{Ind(Species)}].[\ 2N_e\sigma_m^2/\ \sigma_m^2 t]$. The mutational variances cancel out, so that the relationship between the observed and expected ratio of divergence to polymorphic variance is scaled by the ratio $2N_e/\ t$. Assuming an N_e of 10,000 individuals and one generation every 15 years in the 6-7 million years since divergence between human and chimp, the expected distribution of F ratios is expected to be 20-23 times the standard $F_{1,2}$ distribution (with one degree of freedom for the species comparison, and two degrees of freedom for the three individuals within each species). The outer 2.5% tail for this comparison must exceed an F value of 39, hence under these conservative conditions only ratios greater than (39×20), ~ 800, provide clear evidence for a rate of expression divergence greater than that expected under this simple neutral model. Only 17 genes satisfy this criteria, but relaxation of the population size to 100,000 individuals and number of generations to 100,000 reduces the expected rate of neutral divergence, and almost 500 of the 12,600 genes (4%) would fall into the unexpectedly rapidly divergent class. This analysis serves primarily to highlight the conclusion that even high ratios of between species to among individual variance need not imply the action of positive (diversifying) selection.

2.4 RESULTS

Mixed model analysis of the Affymetrix data

The primate dataset reported by ENARD et al (2002) consists of 28 gene chips, including 14 for each tissue (brain or liver), and two replicates of each of the seven individuals. Each of the approximately 12,600 genes was represented by up to 20 unique probes, although these often overlap as described below. This data was analyzed using mixed model analysis of variance (WOLFINGER *et al*, 2001; CHU *et al*, 2002) as described in the Materials and Methods and briefly here. The data was first centralized simply by taking the logarithm of each probe fluorescence intensity on the base 2 scale, and then subtracting the mean value for the particular gene chip. The relative fluorescence intensity, $log_2(PM_{ijkl})$, thus represents a measure of transcript abundance observed for the *k*th Perfect Match probe for the *l*th individual within the *i*th species (human, chimp or orangutan) sampled for the *j*th tissue (brain or liver). A value of 0 corresponds to a gene expressed at the sample mean, -1 or 1 a gene that is one half or two-fold greater than the sample mean, -2 or 2 a gene that is one quarter or four-fold greater, and so on. If a gene has a mean value on this normalized scale of 2 in one species and –1 in another, we can conclude that there is an eight-fold difference in gene expression between the species. Other methods of normalization have been proposed (QUACKENBUSH, 2002; KERR et al, 2000), but we just consider this log-linear normalization strategy here. We next fit a mixed model with fixed main effects of Species (human, chimp, or orangutan), Tissue (brain or liver) and Probe, and Individual within species as a random effect. Expression

differences and significance were estimated for each effect, as well for each species and tissue comparison.

Several checks of data quality were performed. Figure 2.1 shows a "submarine" scatter plot of standardized residuals (the estimated residuals ε_{ijkl} divided by the square root of the variance of these residuals for each gene model) against predicted value. While there appear to be a large number of outliers, actually just 0.5% of the probes have standardized residuals greater than 3. Many of these can be attributed to data saturation. Testing for the normality of the distribution of residuals for each gene specific model indicated that as many as 39% of the genes did not reach the conservative 0.05 significance level. As discussed below, biases in the data due to probe effects may have a particularly large impact on interpretation of contrasts among species.

Levels of divergence within and between species

Direct visualization of the significance and magnitude of effects in the primate comparisons are provided by volcano plots for each pairwise species contrast and each tissue in Figure 2.2. Note that the main effect estimates are averaged over and adjusted for all of the different oligonucleotide probes for each gene, and significance is assessed in the mixed model taking into account among probe variance. Volcano plots contrast significance on the $-log_{10}(p)$ scale against expression difference on the log_2 scale. Genes toward the left and right on each plot show a large expression difference, and those toward the top have high significance, with values of 2, 3, etc representing p-values of 10^{-2}, 10^{-3}, etc.

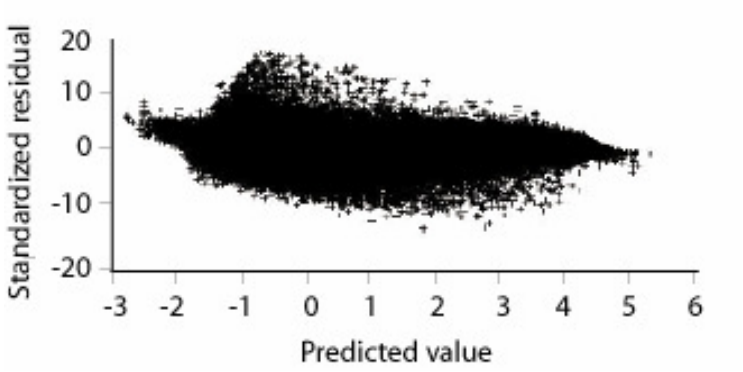


Figure 2.1: Submarine plot of standardized residuals against predicted values for log base 2 transformed signal intensity measurements of each individual oligonucleotide in the primate dataset. The shape of the plot is fairly typical for gene expression data, but asymmetry above and below the horizontal testifies to several percent of probes showing saturation or failure of hybridization.

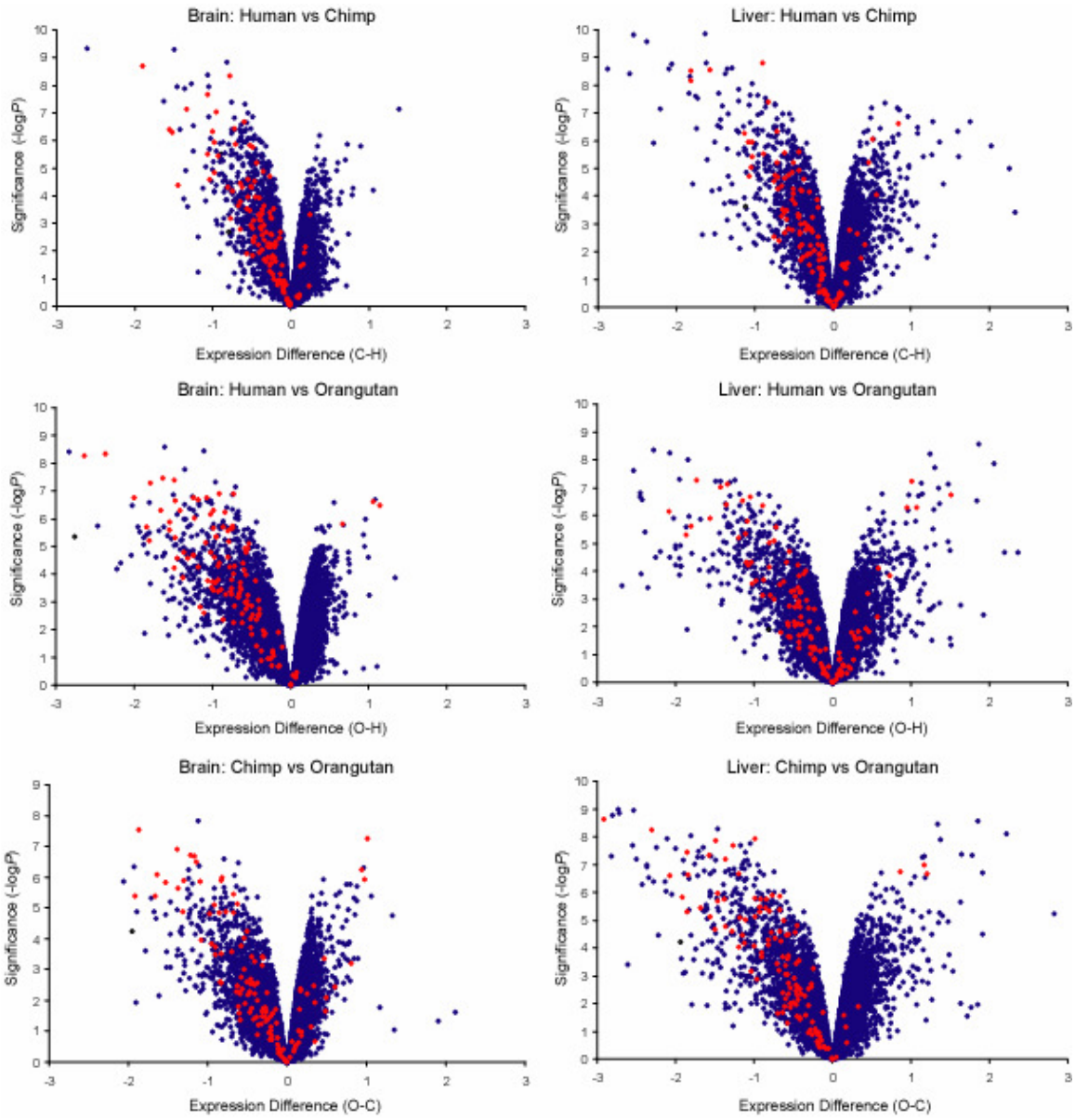


Figure 2.2: Volcano plots of significance against fold change in expression for each primate species comparison in brain (left hand side) and liver (right hand side). Each point represents a single gene analyzed by mixed model ANOVA. Highly significant values toward the top, small expression difference at the center of each plot. Expression difference plotted as difference in the least squares mean of log base 2 normalized expression values for Chimp minus Human (C-H), Orangutan minus Human (O-H) or Orangutan minus Chimp (O-C). The red points are the genes with the most significant (top 1%) species*probe interaction effects: these are clearly asymmetrically distributed in favor of higher apparent expression in the species expected to show the closest sequence homology to the human probes.

Two features of these plots stand out. First, the number of genes toward the top of each plot is greater for the liver than the brain contrasts. For example, comparing human and chimp at the 5% significance level, 25% of the genes show evidence for differential transcript abundance in the brain, and up to 35% in the liver, with a mean of just a 1.2 fold change in either direction. The numbers increase slightly for the contrasts involving the other species. We confirmed this observation on the human-chimp contrast using the analytical approach implemented in dChip software (LI and WONG, 2002), which gave similar results (data not shown). Second, whereas the liver plots are fairly symmetrical, the brain plots are highly asymmetrical: in each case, those genes on the left hand side of each plot are more dispersed across the range of expression differences. Since the plots were drawn with expression difference expressed as chimp minus human, orangutan minus human, and orangutan minus chimp, this means that there are apparently many more genes upregulated in the range of 2 to 4 fold in the human brain relative to the other species than are downregulated. Similarly the chimp shows an apparent bias toward upregulation relative to the orangutan.

Assessment of the significance of expression differences is complicated by the large number of contrasts that are performed as well as the variable residual variance for each gene. If two genes have the same fold difference between species, but one has higher among individual variance within species than the other, the significance of the species difference will be elevated for the second gene. Further, the more genes that are assessed, the more likely it is that genes exceed a low significance threshold by chance. Consequently, we present the number of genes that are significant and the associated fold increase or decrease in expression between species at three different significance levels in

Table 2.1. These are $4x10^{-6}$ (the conservative Bonferroni-adjusted contrast, calculated as 0.05/12,600, and reflected in a negative log_{10} p-value greater than 5.4), 0.001 ($-log_{10} p > 3.0$) and 0.05 ($-log_{10} p > 1.301$). We also present the average expression difference for both up- and down-regulated genes, the percentage of genes that are apparently upregulated, and the percentage of all genes that are differentially expressed for each contrast.

The murine dataset consisted of 14 Affymetrix GeneChips, each containing up to 20 independent oligonucleotide probes for each of approximately 12,488 genes derived from *Mus musculus* sequences. *M. musculus* and *M. spretus* were each represented by three individuals, with a single hybridization for each of the two tissues (hence 6 arrays each), while *M. caroli* was represented by a single individual (2 arrays). We analyzed the data according to the same model as for the primates. The three data quality checks indicated that the data was slightly more favorable for analysis of variance. Only 0.3% of the datapoints had standardized residuals greater than 3, while 86% of the genes passed the normality test for residuals from the mixed model. However, since there were no replicates of each individual, significance tests are not as powerful as for the primate data. Nevertheless, the overall nature of the analyses is remarkably similar, as documented in Figure 2.3 and Table 2.2. Between the two most closely related species, *M. musculus* and *M. spretus*, approximately 10% of the genes showed significantly different transcript abundance at the 5% significance level, with an average of almost 1.3 fold change in either direction for both brain and liver. The same biases toward greater divergence in the liver, and asymmetric upregulation in the brain favoring *M. musculus*

Table 2.1 Fraction of Genes showing Expression Differences Among Primate Species

Comparison	SigLevel[1]	N(Up)[2]	×Up[3]	N(Dn)[2]	×Dn[3]	×All[3]	%Up[4]	%genes[5]
Raw Data								
Brain: H _ C	Bonferroni	86	1.75	5	1.67	1.74	95	0.7
Brain: H _ C	0.001	522	1.41	173	1.24	1.37	75	5.5
Brain: H _ C	0.05	1520	1.26	1734	1.12	1.18	47	25.8
Liver: H _ C	Bonferroni	126	2.01	41	1.75	1.95	75	1.3
Liver: H _ C	0.001	614	1.49	449	1.33	1.42	58	8.4
Liver: H _ C	0.05	1777	1.27	2664	1.16	1.20	40	35.2
Brain: C _ O	Bonferroni	31	2.17	11	1.71	2.04	74	0.3
Brain: C _ O	0.001	411	1.62	352	1.27	1.44	54	6.1
Brain: C _ O	0.05	1685	1.37	3301	1.16	1.22	34	39.6
Liver: C _ O	Bonferroni	72	2.48	33	2.28	2.41	69	0.8
Liver: C _ O	0.001	528	1.72	369	1.46	1.61	59	7.1
Liver: C _ O	0.05	1586	1.39	2184	1.23	1.29	42	29.9
Brain: H _ O	Bonferroni	91	2.27	13	1.62	2.17	88	0.8
Brain: H _ O	0.001	772	1.71	823	1.23	1.44	48	12.7
Brain: H _ O	0.05	2120	1.44	4466	1.17	1.25	32	52.3
Liver: H _ O	Bonferroni	139	2.68	31	2.19	2.57	82	1.3
Liver: H _ O	0.001	647	1.82	407	1.44	1.66	61	8.4
Liver: H _ O	0.05	1608	1.48	2334	1.21	1.31	41	31.3
Filtered Data								
Brain: H _ C	Bonferroni	37	2.08	5	1.54	2.01	88	0.3
Brain: H _ C	0.001	193	1.48	158	1.24	1.37	55	2.8
Brain: H _ C	0.05	854	1.24	1878	1.12	1.16	31	21.7

Notes

[1] Significance levels: Bonferroni = $-logP > 5.4$; 0.001 = $-logP > 3.0$; 0.05 = $-logP > 1.301$

[2] Number of genes up- or down-regulated at indicated significance level

[3] Magnitude of fold change up (greater in left-hand species) or down (opposite direction) based on the raw (unfiltered) data

[4] Percent of genes that are significantly differentially expressed that are up-regulated

[5] Percent of all genes on the microarrays that are differentially expressed

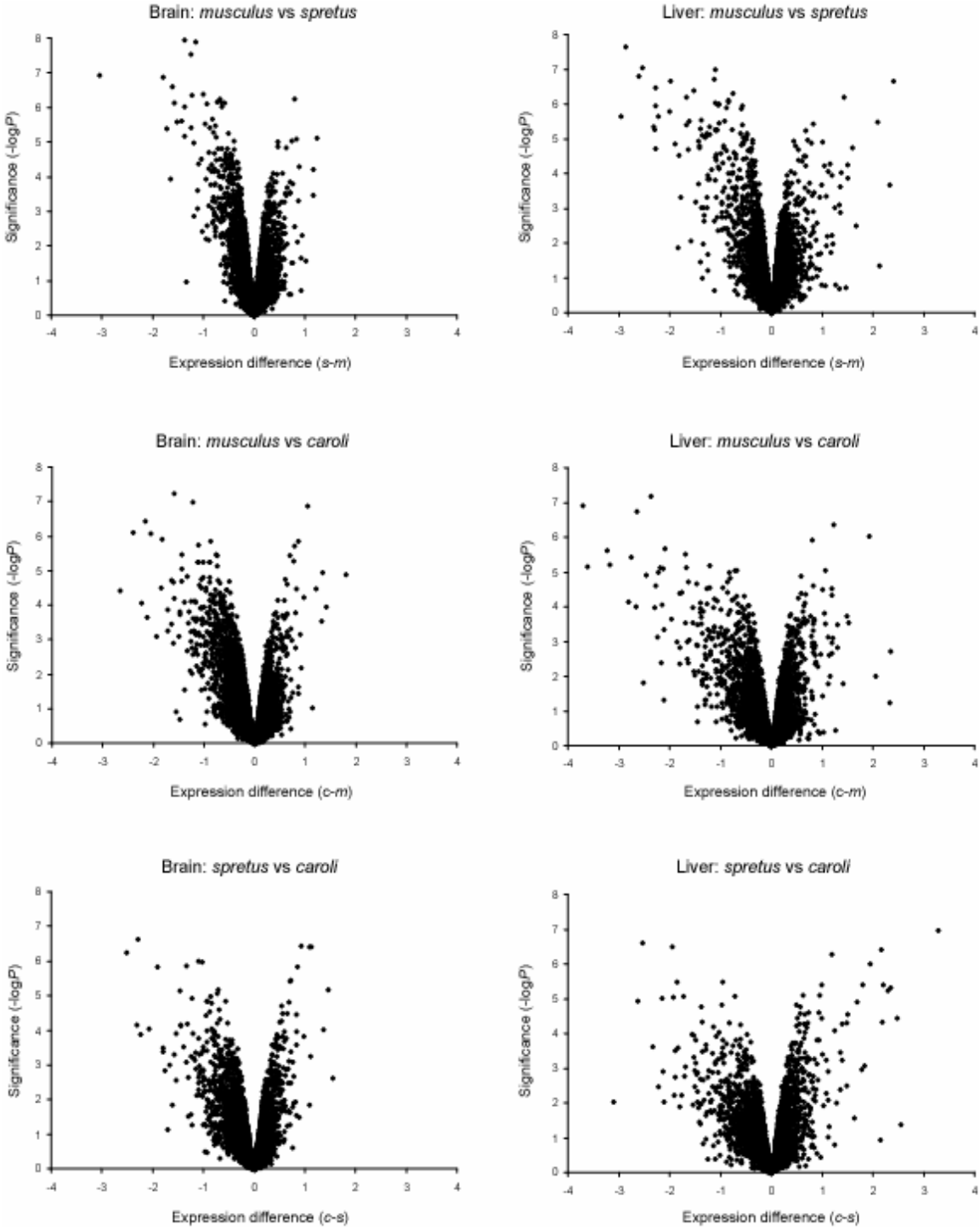


Figure 2.3: Volcano plots of significance against fold change in expression for each murine species comparison in brain and liver. Layout is essentially the same as in Figure 2.2. Species comparisons are Mus spretus minus M. musculus (s-m), M. caroli minus M. musculus (c-m), and M. caroli minus M. spretus (c-s).

Table 2.2 Fraction of Genes showing Expression Differences Among Murine Species

Comparison	SigLevel[1]	N(Up)[2]	×Up[3]	N(Dn)[2]	×Dn[3]	×All[3]	%Up[4]	%genes[5]
	Raw Data							
Brain: M _ S	Bonferroni	23	2.27	1	1.74	2.25	96	0.2
Brain: M _ S	0.001	190	1.53	53	1.41	1.49	78	1.9
Brain: M _ S	0.05	767	1.28	534	1.21	1.25	59	10.4
Liver: M _ S	Bonferroni	27	2.81	4	3.23	2.87	87	0.2
Liver: M _ S	0.001	186	1.80	64	1.72	1.78	74	2.0
Liver: M _ S	0.05	738	1.38	812	1.23	1.30	48	12.4
Brain: S _ C	Bonferroni	6	3.23	6	1.88	2.46	50	0.1
Brain: S _ C	0.001	112	1.82	48	1.55	1.73	70	1.3
Brain: S _ C	0.05	698	1.41	555	1.23	1.33	56	10.0
Liver: S _ C	Bonferroni	5	4.89	5	4.47	4.69	50	0.1
Liver: S _ C	0.001	72	2.08	80	1.84	1.95	47	1.2
Liver: S _ C	0.05	582	1.51	497	1.37	1.44	54	8.6
Brain: M _ C	Bonferroni	12	3.20	4	1.80	2.77	75	0.1
Brain: M _ C	0.001	202	1.83	42	1.61	1.79	83	2.0
Brain: M _ C	0.05	919	1.44	672	1.21	1.34	58	12.7
Liver: M _ C	Bonferroni	7	6.23	3	2.50	4.72	70	0.1
Liver: M _ C	0.001	127	2.19	58	1.64	2.00	69	1.5
Liver: M _ C	0.05	646	1.56	538	1.29	1.43	55	9.5

Notes

[1] Significance levels: Bonferroni = $-logP > 5.4$; 0.001 = $-logP > 3.0$; 0.05 = $-logP > 1.301$

[2] Number of genes up- or down-regulated at indicated significance level

[3] Magnitude of fold change up (greater in left-hand species) or down (opposite direction), based on the raw (unfiltered) data.

[4] Percent of genes that are significantly differentially expressed that are up-regulated

[5] Percent of all genes on the microarrays that are differentially expressed

over *M. spretus* over *M. caroli* are observed, though not as strongly as for the primate data.

From both Tables 2.1 and 2.2, it can be seen that the fraction of genes that appear to be up-regulated (that is, expression in species A is greater than in species B) is consistently reduced as the significance level is relaxed (for example, from 95% to 47% for the human-chimp brain contrast). This implies that the there is a systematic tendency for overestimation of the expression level for genes in the order human > chimp > orangutan (or underestimation in the opposite order). A similar tendency was observed in the murine dataset (*M. musculus* > *M. spretus* > *M. caroli*), and in all cases the consequent apparent bias toward up-regulation is observed in the species genetically closest to *M. musculus*, from which the probe sequences derive.

Probe effects in the context of genetic divergence

This suggests the hypothesis that apparent up-regulation is due to stronger hybridization to individuals of one species over another. At a genome-wide rate of sequence divergence of 1%, if the probes were non-overlapping then only one quarter of them should have any nucleotide differences between species, and only a fraction of these would be near the center of the probe where they are most likely to affect hybridization. Nevertheless, small differences in two or three probes out of 20 could be sufficient to yield an apparent up-regulation of around 1.2-fold. It is also noteworthy that the estimated magnitude of down-regulation is always less than the estimated magnitude of upregulation at the same significance level (hence, the absolute value of the fold-change is always less than the magnitude of up-regulation). This is consistent with the idea that

reduced hybridization to a few probes in the divergent species contributes to apparent up-regulation.

Significance levels are affected by a balance between the fold-change averaged across probes (tending to make more genes appear to be upregulated) and the increase in among-probe variance due to sequence divergence for some probes (tending to reduce the significance of contrasts). We thus asked whether the species-by-probe interaction effect in the mixed model for each gene is more likely to be significant for up-regulated genes. This effect is small in magnitude, but it is significant for more than half of the genes (see Supplementary Information). The red points in the volcano plots in Figure 2.2 indicate the genes with the top 1% of the most significant species-by-probe interaction effects, and these are almost all apparently upregulated. This result is consistent with the hypothesis that the overwhelming bias toward apparent upregulation in the brain in the phylogenetically closest species, which is expected to show the least sequence divergence, might be attributed to loss of hybridization to a subset of probes.

To further explore whether this is the case, we next examined the actual profiles of fluorescence intensity for representative genes. Figure 2.4 shows plots of relative fluorescence intensity for human and chimp brain arrays for each probe for a set of 6 representative genes. The order and spacing of probes along the abscissa is proportional to the number of bases offset along the gene sequence for each probe. Human intensity values are indicated as large open diamonds, and chimp values as small solid boxes. Gene A is an example of a "well-behaved" probe set: despite absolute differences in intensity for each probe, all probes indicate a similar magnitude of upregulation in the

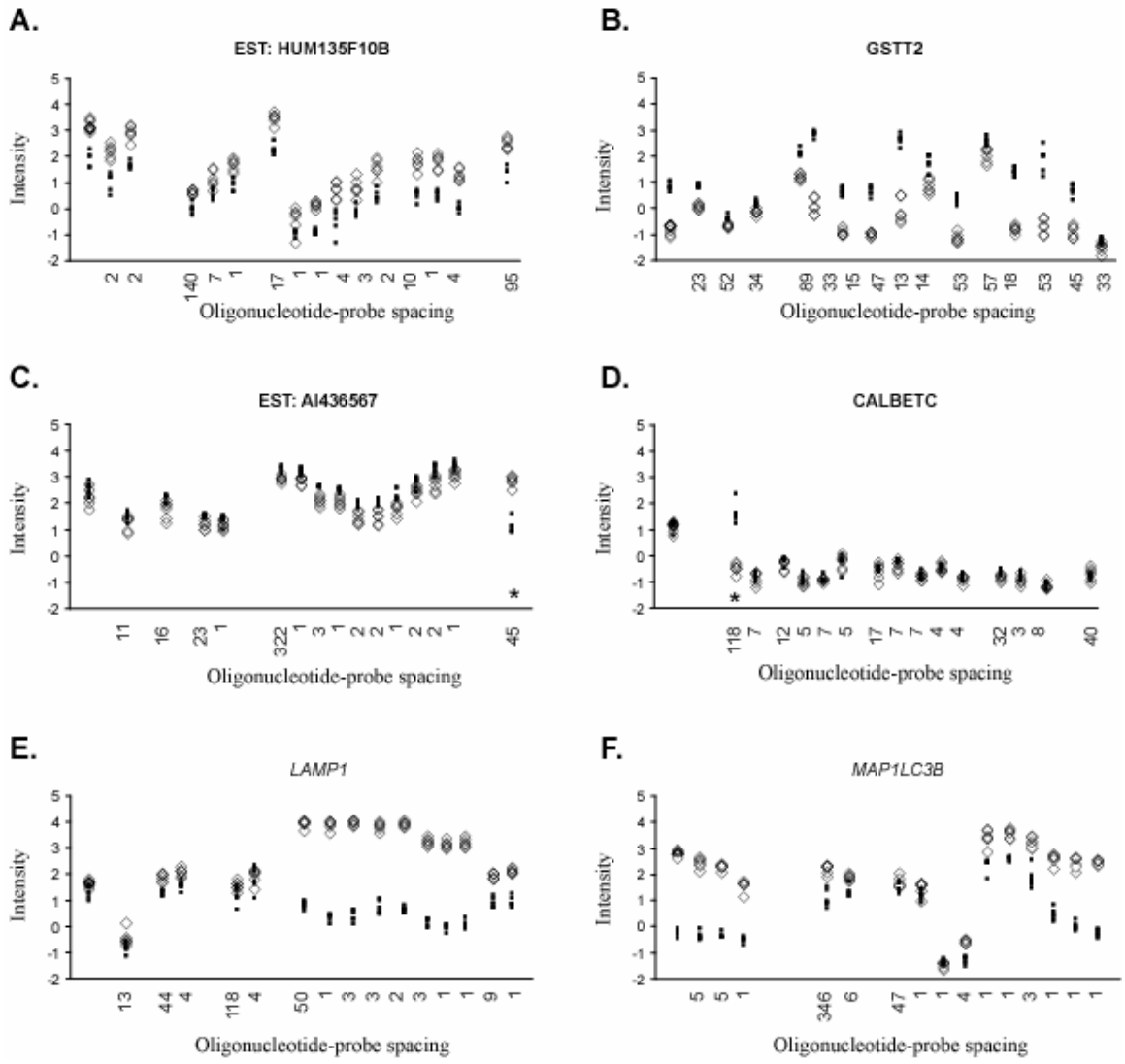


Figure 2.4: Parallel plots of individual oligonucleotide measurements for human and chimp brain samples for six genes. Each of the 16 oligonucleotides are plotted in proportion to the spacing between first nucleotides from 5' to 3': numbers below each plot show the number of nucleotides between these sites. Thus a spacing of 1 represents oligonucleotide probes that overlap by 24 of 25 bases, while a spacing of 45 represents non-overlapping probes. Normalized log base 2 expression level for the perfect match probe on the y-axis: open diamonds human, filled squares chimp. Gene or EST names correspond to GenBank accessions D54318, L38503, AI36567, M92302, J04182, and W28807 for A-F respectively. (A) A "well-behaved" gene with similar differences between species for each probe; (B) "poorly behaved" gene with variable differences; (C) and (D) genes where an overall expression difference is contributed almost entirely by a single probe indicated by the asterisks; (E) and (F) genes where two classes of expression difference, largely but not completely corresponding to overlapping probes, is observed.

human relative to chimp. Gene B by contrast is "poorly-behaved" in so far as each probe predicts a different magnitude for the species difference. Gene C is an example of a locus where a single probe that shows much-reduced hybridization to chimp cDNA (the far right probe) would be sufficient to suggest an overall 1.2-fold upregulation in humans relative to chimps. This situation was also occasionally seen in the reverse direction (one probe gives a stronger chimp signal) as shown for Gene D. However, many of the cases of strong species-by-probe interaction effects involved multiple probes, as seen for Genes E and F. *LAMP1* is apparently upregulated in humans, but only half of the probes showed the difference, and all of these eight probes overlap with their 5' most nucleotides separated by just 14 bases. The next two probes, just 9 and 10 bases further 3', show much reduced species difference. *MAP1LC3B* gave a similar result, except that the species difference was seen in two non-overlapping sets of probes. It is sobering in this case that even probes that overlap by all but one nucleotide give 10-fold differences in signal intensity for both species, and several-fold differences between species.

In an attempt to filter out the probe-by-species interactions, we imposed a constraint that genes should only be included in the analysis if the correlation between human and chimp fluorescence intensity exceeded 0.95. So as to include all genes, we wrote a script to systematically remove outlier probes for each gene until this condition was met. Typically this meant removal of just 2 to 5 probes per gene, but more than half of the genes showed the high correlation without removing any probes. A plot of the expression difference before filtering against after filtering in Figure 2.5A shows many more points below the diagonal than above, indicating that the effect of filtering is typically to reduce the magnitude of the apparent upregulation in human brains, as

expected. However, the volcano plot for the human versus chimp brain comparison in Figure 2.5B remains somewhat asymmetric, and the overall tendency for more genes to be differentially expressed in the liver than the brain when comparing human and chimp is still apparent (see also Table 2.1B).

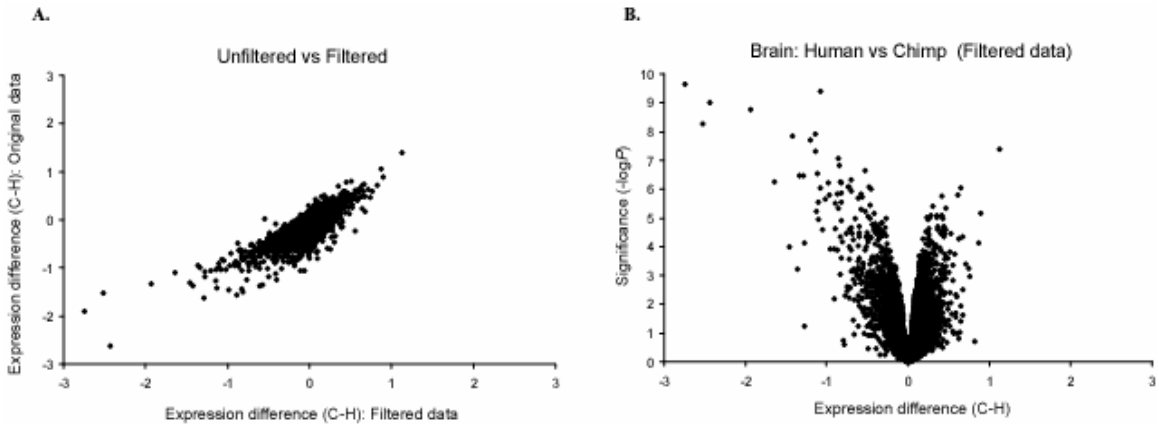


Figure 2.5: Effect of filtering outliers on inference of expression difference between human and chimp brain. (A) Subtraction of human from chimp expression value tends to produce more negative values on the original data than the filtered data: most points lie at or below an imagined diagonal line running through points for which filtering has no effect. (B) Volcano plot after filtering: compared with the top left plot in Figure 2.2, this plot is considerably more symmetric, due to removal of probes that contribute to the large species*interaction effect. Both plots are for just the brain data.

2.5 DISCUSSION

Possible Biases in Oligonucleotide Expression Data

Our mixed model analyses of the primate and murine gene expression data leads to conclusions that are not necessarily consistent with those reported by the original authors (ENARD *et al*, 2002) in so far as there is little evidence for accelerated divergence in gene expression in the human brain. Whichever method of analysis is used, the interpretation should be tempered to some extent by our finding of potential species-specific biases in the magnitude of inferred transcript abundance. Since in all cases more genes were seen to be upregulated in the species that is closest to the one whose sequence was used to generate the probes (that is, *Homo sapiens* or *M. musculus*), the most straight forward explanation is that this bias reflects differential hybridization to loss of perfect sequence matching.

However, three lines of evidence lead us to question this explanation. The first is that detailed analysis of numerous genes that showed a species-by-probe interaction effect (that is, variable differences in transcript abundance among probes within a gene) indicated a complex relationship between sequence and signal. Overlaying the mismatch probe data on the perfect match data does not help at all as it just increases the noisiness of the results (data not shown: many mismatches hybridize as strongly as the match and the difference between match and mismatch also varies greatly by probe within each

gene). Many factors, presumably including amount of cross-hybridization, alternative splicing, and sequence divergence, must contribute to probe effects, and it is not obvious how to deal with these statistically. The fact that Affymetrix' probe selection algorithm tends to choose clusters of sequences that differ by just a few bases also introduces a correlation structure to the data that formally but impractically should be dealt with on a gene by gene basis. We and others (CHU *et al*, 2002; LI and WONG, 2002; SASIK *et al*, 2002) have demonstrated that modeling gene expression profiles by probe within gene is generally much more accurate than using the average difference measure, but it is also clear that genotypic differences can affect the results in ways that are difficult to control.

The second line of evidence arguing against sequence divergence accounting for all of the biases toward upregulation is that the effect appears to be much greater in the brain samples than the liver. This could imply that brain proteins are diverging at a faster rate than liver proteins. Comparative sequence analyses will soon resolve this issue. ENARD *et al* (2002) also provided 2-D gel electrophoresis evidence for divergence in protein sequence and abundance between human and chimp brain, but it is not yet possible to assess whether differential sequence divergence is responsible for at least some of the apparent upregulation of a large number of human genes. The third line of evidence is that the upward bias is only observed for the 10 to 20 percent of genes that show the most significant divergence in gene expression. Below the 5% significance level there are essentially equivalent numbers of genes up and down-regulated in each comparison. Attempts to filter out the largest probe-by-specific effects had little impact on the overall conclusions, arguing that many of the observed differences in gene

expression are real and that there may in fact be a biological basis to the tendency toward increased gene expression in humans over chimps and orangutan. Whether this relates to increased size and/or complexity of the brain remains to be seen.

Divergent Gene Expression among Primates

Our reanalysis of ENARD *et al*'s data (2002) on a gene-by-gene basis is in broad agreement with their analysis based on whole-transcriptome variation in several respects, but also allows quantification of the fraction of genes that contribute to within and between species differences. Both analyses indicate that there are significant differences in gene expression among species that are of a greater magnitude than the differences among individuals within a species, and that there is a general increase in degree of transcriptional divergence as sequence (and hence temporal) divergence increases. As pointed out by ENARD *et al* (2002), conclusions concerning relative rates of divergence are however quite sensitive to the metrics used.

Mixed model analysis provides formal statistical support for 51 genes being differentially expressed between human and chimp Brodmann's area 9 after filtering outlier probes, and just under twice this number if raw data is used. At the less conservative significance threshold of 0.001, 482 genes are differentially expressed with an average almost 1.4 fold change between human and chimp brain, compared with a chance expectation of just 13 genes at this level. Based on the raw data, this number increases to 695 genes, and to 1595 genes when human is compared to orangutan. For the liver, there are 1063 genes

differentially expressed between human and chimp, also with an average 1.4 fold change, and 1054 genes between human and orangutan at a slightly higher mean fold change of 1.6. The chimp-orangutan comparisons are intermediate, with slightly more genes differentially expressed with a larger fold-change in the liver than in the brain.

Most of our comparisons of species and tissue pairs suggest then that there are more genes divergently expressed in the liver than the brain, and that the magnitude of change also tends to be greater in the liver. While it is clear that dramatic cognitive changes have occurred particularly in the human lineage, it also not surprising that transcription has evolved greatly in the liver, given the differences in diet and culture of the primate species. A possible reconciliation of our findings with the inference favored by ENARD et al (2002) that "changes of gene expression in the brain may have been especially pronounced during recent human evolution" is the suggestion that much of that change has occurred on the human-orangutan axis. We also observe a relatively large branch length between all humans and the central node on neighbor-joining trees based on transcriptome-wide average expression differences at each level of significance (see Supplementary Figure 2.2). It is noteworthy though that the relative length of this node, as well as the divergence of the second human individual from the others, is very much a function of the number of genes included in the analysis.

The nature of the differentially expressed genes is also of interest. Those that are significantly divergent between human and chimp brain and liver are tabulated in Supplementary Figure 2.3. A number of neuronal genes such as neurotransmitter

receptors and channels are obvious in the brain list, as are detoxification enzymes such as cytochrome P450s on the liver list. However, the majority of genes have more general potential functions in regulation of cell growth and division, and cell structure: members of most of the major gene ontology categories are represented in both lists.

Finally, we can also ask whether the divergence in gene expression is more likely attributable to drift or diversifying selection. A significantly elevated measure of divergence in expression between species, relative to the observed level of among individual within species variation, is not *prima facie* evidence for selection. Figure 2.6A shows that the most significant genes in the human versus chimp brain comparison both diverge between species and have relatively low levels of intraspecific variance. There are also a large number of genes for which this relationship is reversed. In fact, a histogram of the log ratios of the mean squares for the species and individual within species components is slightly skewed toward low ratios, suggesting that many genes may be more variable within species than expected. RIFKIN et al (2003) have recently proposed, following LYNCH and HILL's approach (1986) for phenotypic traits, that the expected degree of divergence under mutation-drift equilibrium can be formulated by scaling the ratio of mean squares for divergence and polymorphism by the ratio of twice the effective population size over the number of generations since divergence. Assuming a small effective population size for humans of 10,000 individuals and a mean generation time of 15 years over the 6-7 million years since the human and chimp lineages diverged (BRUNET et al, 2002), the expected distribution of ratios for this dataset is 20 to 23 times larger than the $F_{1,2}$ distribution. Consequently, only genes with a divergence to

polymorphism ratio greater than 800, a handful of just 15 to 20 genes in our analysis, clearly lie in the upper 2.5% tails of the expected level of divergence for these population parameters. Relaxation of these conservative assumptions provides suggestive evidence that 5% or more of the genes may be experiencing diversifying selection. Clearly more individuals need to be sampled at different ages and for more targeted tissue samples, but comparison of gene expression divergence, coincident with assessment of nucleotide sequence divergence, is a promising approach to identification of genes that may have contributed to human cognitive evolution.

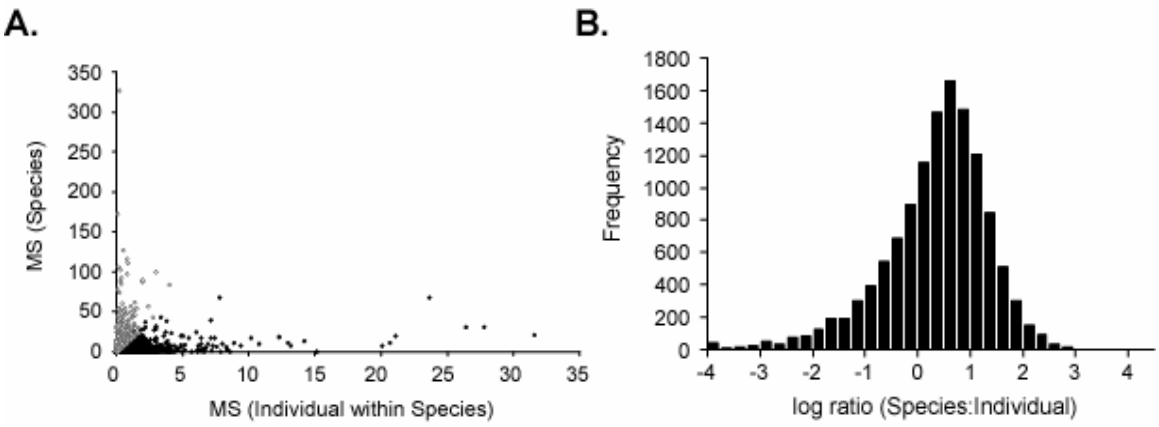
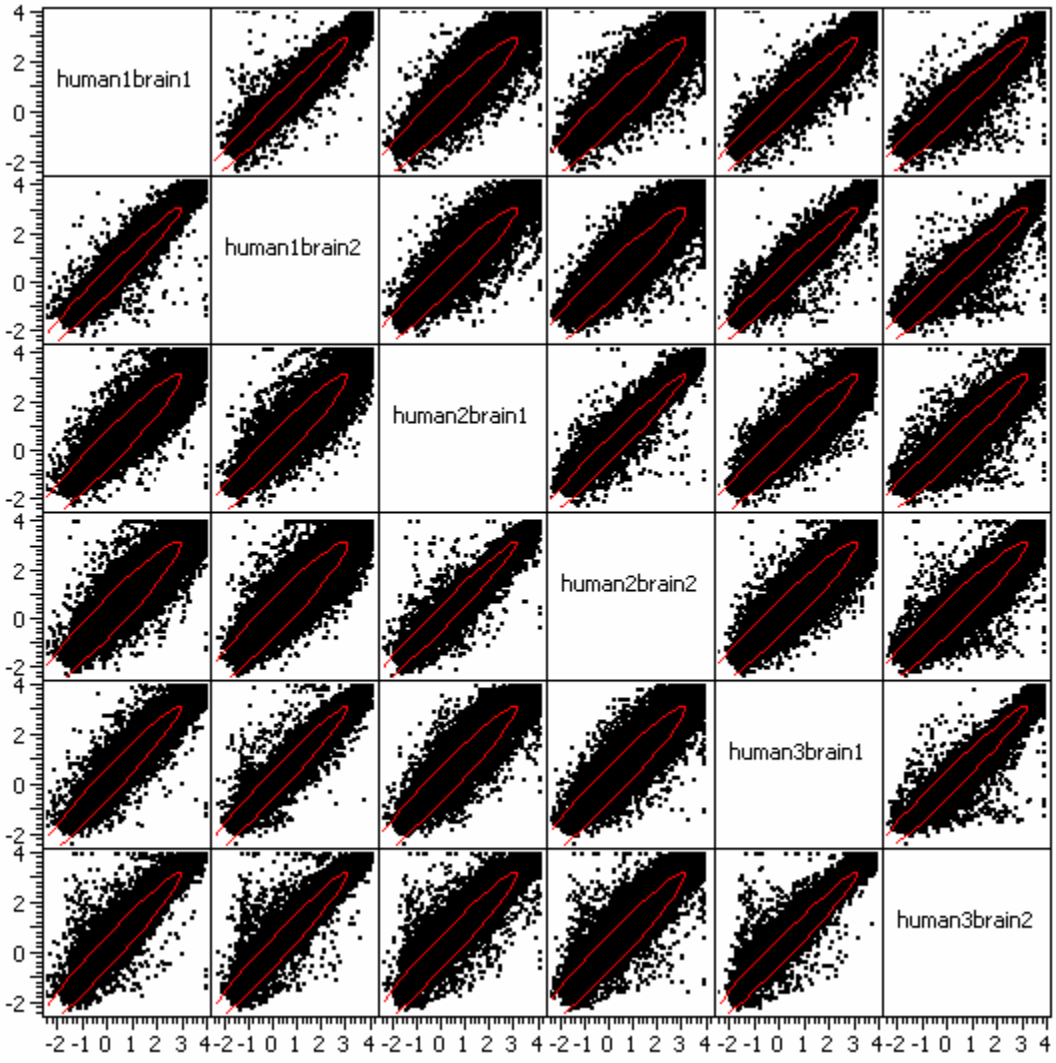
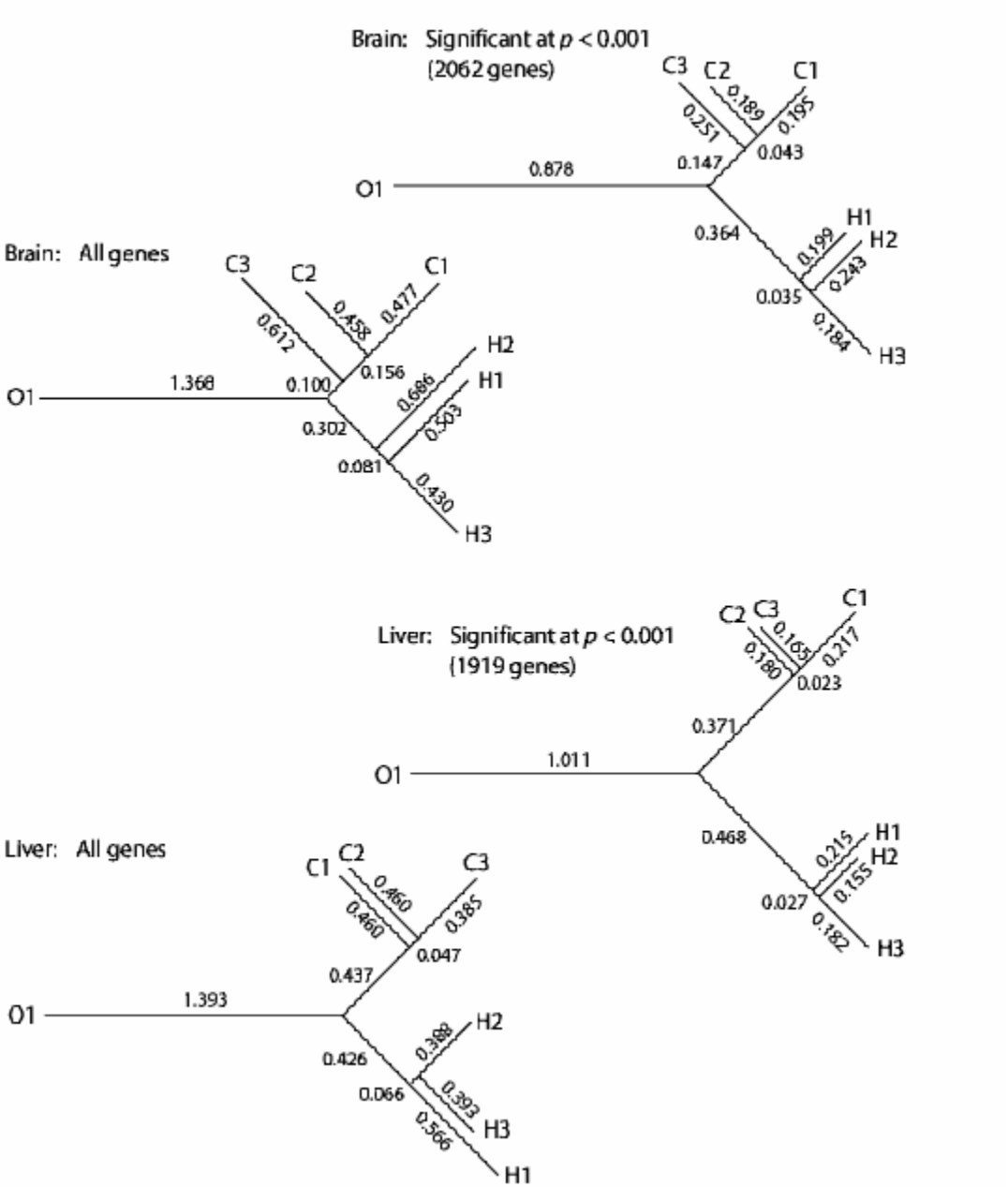


Figure 2.6: Contrast of contributions of species and individual within species to expression variance between human and chimp brain. (A) Plot of mean square from a general linear ANOVA for the species and individual within species terms for each gene. Open diamonds toward the left of the figure show genes with a significant F-ratio, indicating significant divergence between species relative to variation within. Note the large number of genes (filled diamonds, mostly toward right of plot) with much greater variation within than between species. (B) Histogram of frequency of log base 10 ratio of mean square species:mean square individual within species values from (A). Only a few genes have a ratio approaching 1000 (that is, 3 on the log scale), whereas approaching 25% of the genes have a ratio above 10.



Supplementary Figure 2.1: Scatterplot Matrix of log base 2 fluorescence intensity measures for each of the 6 human brain samples (two replicates of each of three individuals) against one another. The breadth of scatter of points is inversely proportional to the correlation between the pair arrays.



Supplementary Figure 2.2: Neighbor-joining trees for primate data. Top to bottom: Brain, 2062 most significant genes ($p < 0.001$) from the combined three-species dataset; Brain, all genes; Liver, 1919 most significant genes ($p < 0.001$); Liver, all genes. In all plots, the branch length is proportional to the distance between individuals (1 orangutan, at tip of left branch, 3 humans and 3 chimps; each averaged over two replicates). As noted by Enard et al, the branch from the central node to the base of the 3 humans is more than twice as long than the corresponding branch to chimps, but only for the brain (not liver) sample. However, the connectivity between individuals within a species is strongly influenced by the number of genes included in the analysis.

Supplementary Figure 2.3: Excel spreadsheet sorted by significance of differentially expressed genes with and without filtering for brain and liver comparisons of human and chimp. See http://statgen.ncsu.edu/ggibson/HCO/PrimateSigGenes.xls

2.6 ACKNOWLEDGEMENTS

We thank Svante Pääbo for correspondence and comments on the analysis of data generated in his laboratory, and Bruce Weir, Zhao-Bang Zeng, Spencer Muse, and Maryellen Ruvolo for discussions. WPH is supported by a Genome Science Graduate Fellowship from the Bioinformatics Program at NC State University, and research in GG's laboratory is supported in part by NIH PO1-GM45344.

2.7 LITERATURE CITED

Brunet M, Guy F, Pilbeam D, Mackaye HT, Likius A, et al, (2002) A new hominid from the Upper Miocene of Chad, Central Africa. Nature **418:** 145-151.

Cavalieri D, Townsend JP, Hartl DL, (2000) Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis. Proc. Natl Acad. Sci. USA **97:** 12369-12374.

Cheung VG, Spielman RS, (2002) The genetics of variation in gene expression. Nature Genetics Suppl. **32:** 522-525.

Chu TM, Weir B, Wolfinger R, (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Mathematical Biosciences **176:** 35-51.

Enard W, Khaitovich P, Klose J, Zollner S, Heissig F, et al, (2002). Intra- and interspecific variation in primate gene expression patterns. Science **296:** 340-343.

Felsenstein, J., 1989 PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics **5:**164-166.

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. Nature Genetics **29:** 389-395.

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

Li C, Wong WH., (2002) Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc. Natl Acad. Sci. USA **98:** 31-36.

Lynch, M., and Hill, W.G., (1986) Phenotypic evolution by neutral mutation. Evolution **40:** 915-935.

Oleksiak MF, Churchill GA, Crawford DL., (2002) Variation in gene expression within and among natural populations. Nature Genetics **32:** 261-266.

Quackenbush, J., (2002). Microarray data normalization and transformation. Nature Genetics Suppl. **32:** 496-501.

Rifkin SA, Kim J, White KP., (2003) Evolution of gene expression in the Drosophila melanogaster subgroup. Nature Genetics **33:** 138-144.

Sasik R, Calvo E, Corbeil J., (2002) Statistical analysis of high-density oligonucleotide arrays: a multiplicative noise model. Bioinformatics **18:** 1633-1640.

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. Compu. Biol. **8:** 625-637.

Chapter 3. WHO ARE THOSE STRANGERS IN THE LATIN SQUARE?

3.1 Abstract

A Latin Square data was provided by Affymetrix to evaluate statistical algorithms for microarray data analysis. I approach it from a classical parametric statistical modeling perspective. The first stage is to formulate a reasonable model for the probe-level data based on extant knowledge of the experimental design and technology. Some options are presented and it is settled on a linear mixed model for the $\log_2$ perfect match data. Upon applying this model to the data for every gene in turn, it is discovered that not only do the fourteen spiked-in genes appear highly significant, but that a few additional, unexpected, genes display profiles remarkably similar to those of the fourteen. Except for probe sets aimed at examining the same genes, it is likely that some short motifs might be the reason for this cross hybridization. Each of these genes was investigated and some details were provided with plausible explanations.

Key words: perfect match versus mismatch probes, mixed model, cross hybridization

3.2 INTRODUCTION

Microarray data analysis is a complex process involving image analysis, normalization, modeling, and clustering. Each step plays an important role in reaching accurate conclusions. To evaluate methods used for analysis, a high quality data set is very useful. The Affymetrix Latin Square data is quite appropriate for this end [Affymetrix technical report, 2002], as it provides not only a good experimental design but also known targets for evaluation. The design will be briefly summarized in the next section.

In this report, the data set was first normalized and then a linear mixed model [Chu et al., 2002] was used to detect probe sets with significant variation between experiments, which were designed to have 14 different patterns of transcript concentration. The hybridization level of each gene is examined for each target profile. It is assumed that only the target probe sets will show the expression profile matching their respective spiked genes. Curiously, this analysis retrieves not only the putative targets that match the spiked genes, but also some unexpected probe sets that show the same profiles. Certain motifs were considered to be the reason for cross hybridization and some examples were discussed in detail.

3.3 THE AFFYMETRIX LATIN SQUARE EXPERIMENT

Affymetrix designed an experiment on the human HG-U95A arrays to test their new statistical algorithm for data analysis. Each sample hybridized to the arrays consists of the same 14 known transcripts but with different concentrations. Experiment A to T represents 14 combinations of transcript concentrations as shown in Table 3.1. Each experiment has two to twelve replicates.

Table 3.1. Latin Square design. Each column represents one combination of concentrations for the 14 transcripts listed in the leftmost column. The concentration is in the unit of pM.

Transcript ↓		Experiments →													
		A	B	C	D	E	F	G	H	I	J	K	L	M, N,O, P	Q, R, S, T
37777_at	1	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
684_at	2	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024	0
1597_at	3	0.5	1	2	4	8	16	32	64	128	256	512	1024	0	0.25
38734_at	4	1	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5
39058_at	5	2	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1
36311_at	6	4	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2
36889_at	7	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4
1024_at	8	16	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8
36202_at	9	32	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16
36085_at	10	64	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32
40322_at	11	128	256	512	1024	0	0.25	0.5	1	2	4	8	16	32	64
407_at	12	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512	1024
1091_at	13	512	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256
1708_at	14	1024	0	0.25	0.5	1	2	4	8	16	32	64	128	256	512

3.4 STATISTICAL MODEL SELECTION

For these data, the experimental design is well known, although several options are available regarding which dependent variable to use in terms of the perfect match (PM) and mismatch (MM) intensity measurements. Some choices include models for paired differences [Li and Wong, 2001a, 2001b], PM-only [Chu et al.2002], [Lazaridis et al., 2001], adjusted PM [Efron et al., 2000], [Irizarray et al., 2001], both PM and MM [Lemon et al., 2001], [Teng, 1999]. Some of these are on the original scale and some on a log scale, and even compromises have been recommended [Durbin et al., 2002]. How to decide?

Linear reproducibility is one criterion that has bearing, and Table 3.1 lists the average correlation coefficient of several different intensity measurements of the fourteen spiked genes within each of fourteen experimental groups. The log transformed PM and MM values have the best consistency in this metric for most of the experimental groups. Based on this evidence, and the fact that the raw intensities represent pixel counts ranging heterogeneously over several orders of magnitude, a log transformation is justifiable. For my modeling efforts, log base 2 was used so that a unit difference on this scale can be interpreted as a two-fold change in the original scale. Furthermore, if the amount of cross-hybridization for an individual probe is proportional to the observed signal, and the constant of proportionality remains stable across the experiment, then this constant will cancel out any differences taken on the log scale.

Figure 3.1 plots replicate values of log2(PM) and log2(MM) against one another for the three chips in experiment A. This plot reveals some potential data quality issues

that should be addressed. In particular, outliers that appear well away from the main diagonal represent inconsistent measurements that should be handled carefully and potentially filtered out of the analysis.

Regarding how to handle the mismatch data, Figure 3.2, from a randomly chosen chip (1532e99hpp_av04), shows how strongly log2(MM) is correlated with log2(PM). MM is clearly picking up true signal and is subject to noise, and therefore subtracting it directly from PM is likely not the most optimal way to proceed.

Table 3.2. Average correlation coefficient across replicates within the experiment groups

	Average Correlation Coefficient				
Experiment	Log PM	Log MM	PM	MM	PM-MM
A	.9811	.9876	.9712	.9921	.9213
B	.9940	.9911	.9918	.9905	.9798
C	.9926	.9917	.9888	.9843	.9887
D	.9934	.9918	.9880	.9845	.9823
E	.9944	.9915	.9882	.9878	.9881
F	.9905	.9860	.9776	.9823	.9707
G	.9928	.9902	.9883	.9878	.9861
H	.9934	.9890	.9901	.9825	.9865
I	.9904	.9830	.9864	.9804	.9793
J	.9954	.9939	.9912	.9891	.9874
K	.9938	.9770	.9871	.9788	.9785
L	.9948	.9931	.9215	.9868	.8644
M,N,O,P	.9957	.9912	.9898	.9855	.9862
Q,R,S,T	.9952	.9896	.9903	.9883	.9819
Average	.9927	.9891	.9822	.9858	.9701

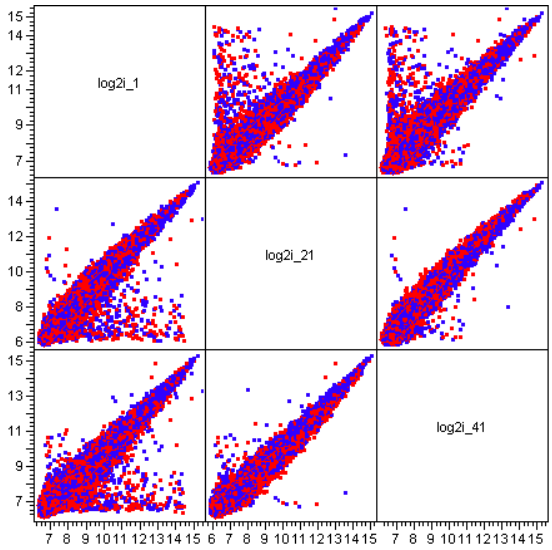


Figure 3.1. Scatter plot matrix of $\log_2$ PM (blue) and $\log_2$ MM (red) for the three replicate chips in experiment A

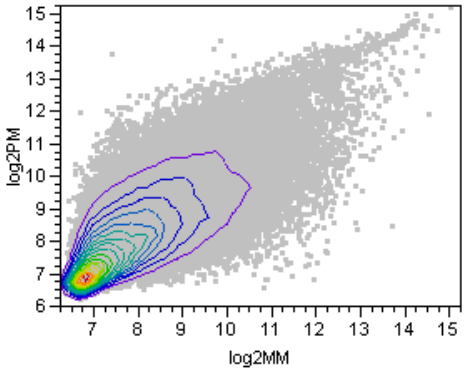


Figure 3.2. Plot of $\log_2$(MM) versus $\log_2$(PM) for chip 1532e99hpp_av04 in Experiment E. Each curve represents a contour of the bivariate density for $\log_2$(MM) and $\log_2$(PM).

Based upon the preceding considerations, I employ $\log_2$(PM) as a dependent variable in a linear modeling context. Research on methods for including MM as an additional dependent variable in a bivariate fashion is reported in Chapter 4. My analysis follows methods reported in [Wolfinger et al., 2001] and [Chu et al., 2002]. In particular, the following two models are employed in turn:

$$\log_2(PM_{ijkg}) = E_i + A_{ij} + \varepsilon_{ijkg} \quad (1)$$
$$R_{ijkg} = E_{ig} + P_{kg} + A_{ijg} + \varepsilon_{ijkg} \quad (2)$$

In Model (1), the symbols *PM, E, A,* and ε represent perfect match probe intensity, global experiment effect, global chip random effect, and stochastic error term, respectively. Here, "global" emphasizes that the corresponding effect applies across all genes. To be more precise, PM_{ijkg} means the intensity of the *j*th replicate of the *i*th experiment for the *k*th perfect match probe of the *g*th gene. In Model (2), the symbols *R, E, P, A,* and ε represent the residual calculated from Model (1), gene-specific experiment effect, gene-specific probe effect, gene-specific chip random effect, and stochastic error term, respectively. Model (1) is fitted once to jointly normalize all of the data, and model (2) is fitted separately for each gene.

3.5. Results

I fit the preceding models to the data for all 12,626 genes using SAS Proc Mixed. After this modeling, outliers for standardized residuals greater than a certain threshold were filtered and then various output statistics were collected. The fourteen spiked genes each had wildly significant results, with overall $-\log_{10}$(p-values) around 300. In Figure 3.3 the

least squares mean of gene-specific experiment effect from model (2) were standardized for each of the 14 target probe sets. The figure displays the profiles for these genes across each of the 14 concentration levels on a standardized scale. The resulting subtle S-shape matches that seen in other analyses.

What surprised us, however, was the significance of several genes in addition to the putative fourteen. These are listed in Table 3.3. Expression profiles of 14 target genes are shown in Figure 3.4 and the top five unexpected genes including the one that was claimed to be missing in the supporting information that Affymetrix provided are displayed in Figure 3.5. A few have obvious explanations, but others do not. One example of the latter is 1032_at, which has an expression profile matching that of the spiked gene 684_at.

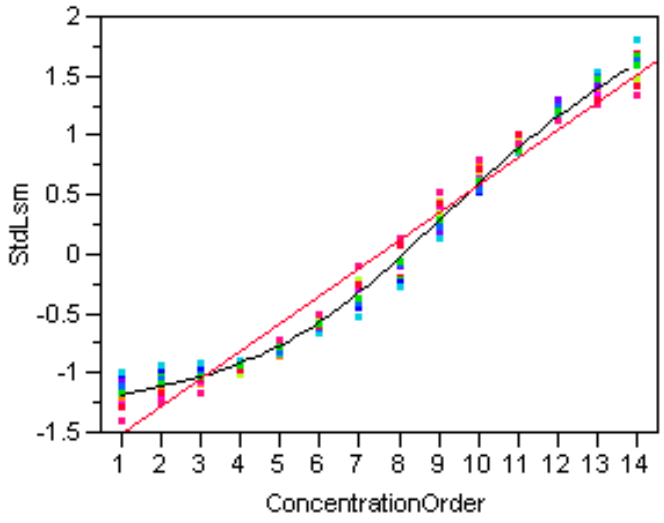


Figure 3.3. Overlay of standardized least squares mean profiles for the 14 spiked genes. Different genes are marked with different colors. The concentration order from 1 to 14 matches to the concentration 0 pM to 1024 pM. The straight line and the curve represent linear and smooth nonparametric fitted lines respectively.

Table 3.3. Unexpected significant genes from the mixed model analysis

Probe Set	Target	Remark
33818_at	AC004472	should be in Latin Square as Transcript #12
546_at	S76965	same profile as 36202_at
1598_g_at	L13720	same profile as 1597_at
37658_at	L13720	same profile as 1597_at
1032_at	U11872	same profile as 684_at

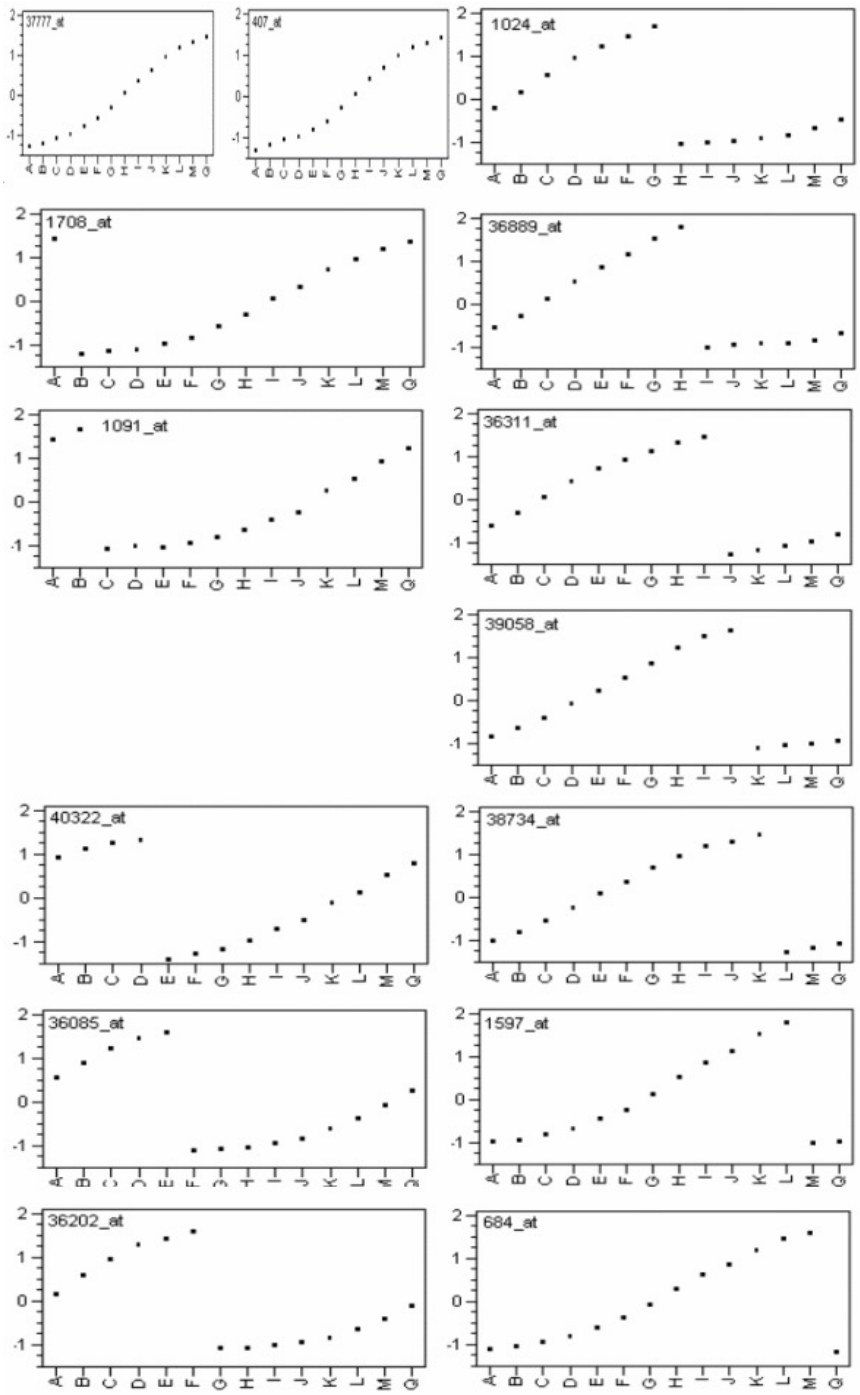


Figure 3.4. Standardized least squares mean profiles of significant genes from the mixed-model analysis.

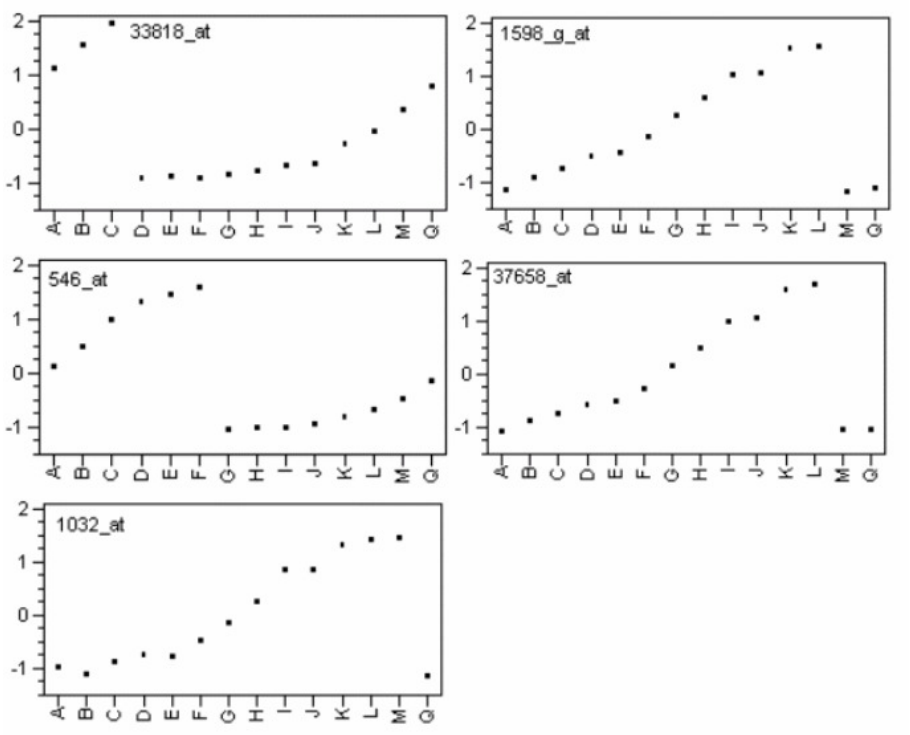


Figure 3.5. Standardized least squares mean profiles of unexpected genes from the mixed-model analysis.

I discovered the unexpected genes in Table 3.2 by statistically screening the expression profiles of the target transcripts with that of all probes in the U95A chip. Empirically, the intensity measurement is not quite linearly proportional to the transcript concentration level at high and low concentrations. To reflect the observed intensity measurement patterns, I used least squares mean of the experiment effects in model (2) as queries to retrieve probes with similar expression profiles. I first take averages of expression levels across replicates for each experiment and then calculate the correlation coefficients of those 14 average values with the 14 least squares mean values. The higher the correlation coefficient, the higher the similarity in expression pattern between the target gene and the sequence to which it cross hybridized. Once obtained, the selected matches are obvious by visual comparison of Figures 3.4 and 3.5.

A brief investigation of the five genes in Table 3.2 produced simple explanations for their observed expression profiles, except for that of probe set 1032_at. Five probes of 1032_at have correlation coefficients higher than 0.99 with profile of 684_at while the other 11 probes have correlation coefficients less than 0.5. Since the five highly correlated probe sequences have significant overlap with each other, part of this overlapping sequence was expected to be definitive, and Gibbs Sampling [Lawrence, 1993] was used to identify possible motifs. For each threshold setting on correlation coefficients, the longest common motif of all the probes chosen was identified and manually checked for the most probable candidate sequences. Not all of the probes have similar sequences as those of probe set 1032_at, but two of them were found that supported our conjecture. They are displayed in Table 3.3, and the expression profiles of those probes are shown in Figure 3.6.

The first sequence, GCAGCCGTTT, appears in seven probe sequences of the U95A chip other than those in probe set 1032_at. Only three out of seven match the sequence similarity profile of target 684_at, however they are not so strong as the second motif CCGTTTCTCCTTGGT in probe set 39059_at. This 15-base motif has a similar counterpart in the probe sequences of 1032_at but with an additional base T. If this single base T is allowed to mismatch when hybridizing to the target sequence, then the alignment in Figure 3.7 seems to be a promising reason for the observed cross hybridization.

Table 3.4. Probe sequences with expression profile matching that of target 684_at from netaffy.com. Highlighted motifs can be aligned to the sequence in gene K02215 (target of 684_at) but not the probe sequences of 684_at.

Probe Set	Position	Probe Sequence	Correlation coefficient of expression profile with the target profile 684_at
1032_at	46	agaatatGCAGCCGTTTtctccttc	0.997
1032_at	48	aatatGCAGCCGTTTtctccttcct	0.997
1032_at	49	atatGCAGCCGTTTtctccttcctg	0.998
1032_at	51	atGCAGCCGTTTtctccttcctggg	0.994
1032_at	52	tGCAGCCGTTTtctccttcctgggt	0.994
34404_at	1600	tgGCAGCCGTTTcttaacatgttga	0.875
38729_at	1878	agactcctggGCAGCCGTTTtcctc	0.888
38729_at	1885	tggGCAGCCGTTTtcctcatccttt	0.824
1402_at	2105	gagtGCAGCCGTTTcagaagaaaac	< 0.5
32616_at	2229	acatctgagtGCAGCCGTTTgagaa	< 0.5
32616_at	2238	tGCAGCCGTTTgagaagaaaacatc	< 0.5
40261_at	1338	gacatGCAGCCGTTTcggggtagat	< 0.5
39059_at	2305	gtgcgCCGTTTCTCCTTGGTagcgt	0.996
39059_at	2310	CCGTTTCTCCTTGGTagcgtgcacg	0.991

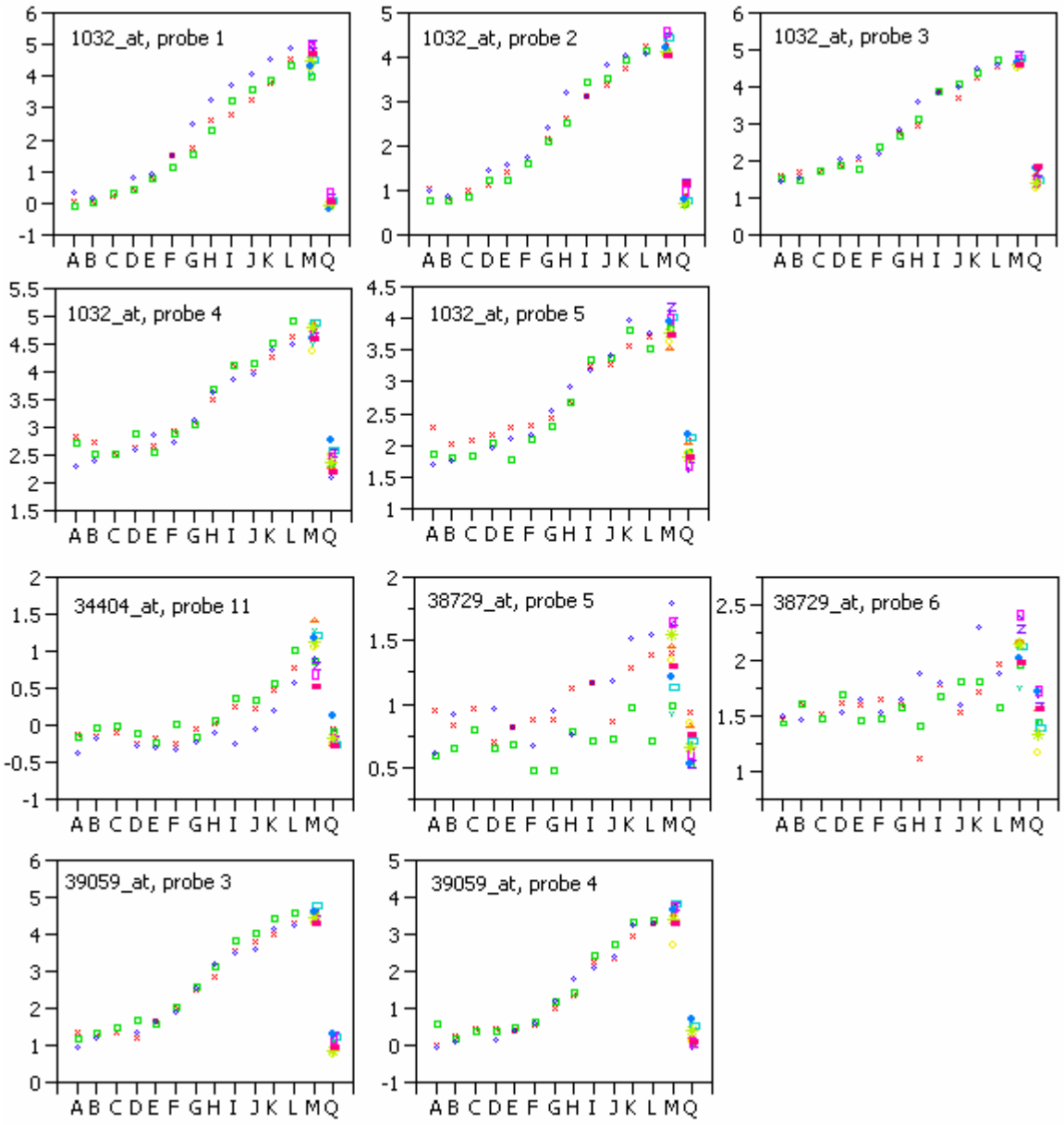


Figure 3.6. Expression profiles of probes listed in Table 3.3. Each plot contains 59 points, which represent 59 chips and there are two to twelve replicates for each experiment.

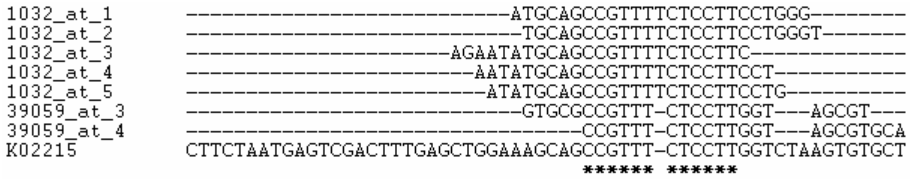


Figure 3.7. Sequence alignment for probes of 1032_at and 39059_at with K02215 (target of 684_at)

3.6 CONCLUSION

A linear mixed model of log2(PM) is a powerful method for assessing significance of gene expression profiles. For the Affymetrix Latin Square data, it detected all fourteen spiked genes with extremely high precision, as well as five additional "strangers". One of the five, 1032_at, did not have an initially obvious explanation, but after a more detailed motif finding exercise, I was able to find a few motifs that likely caused the cross hybridization. Additional spiked-in experiments like this one will be useful for further insights into probe performance and design.

3.7 REFERENCES

Affymetrix, (2001), New statistical algorithm for monitoring gene expression on GeneChip probe arrays. http://www.affymetrix.com/support/technical/technotes/statistical_algorithms_technote.pdf

Chu, T., Weir, B., and Wolfinger, R. (2002). A systematic statistical linear modeling approach to oligonucleotide array experiments. Math. Biosci. 176, 35-51.

Efron, B., Tibshirani, R., Goss, V., and Chu, G. (2000). Microarrays and their use in a comparative experiment, Technical Report, Stanford University.

Durbin, B., Hardin, J., Hawkins, D., and Rocke, D. A variance-stabilizing transformation for gene expression microarray data. Bioinformatics. (2002) 18 Suppl 1:S105-10.

Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U. and Speed, T. (2003). Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 4(2):249-64.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. (1993), Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. Science 262:208-214.

Lazaridis, E. N., Sinibaldi, D., Bloom, G., Mane, S., and Jove, R. (2001). A simple method to improve probe set estimates from oligonucleotide arrays. Math. Biosci. 176, 53-58.

Lemon, W. J., Palatini, J. J. T., Krahe,R., and Wright , F. A. (2001). Theoretical and experimental comparisons of gene expression indexes for oligonucleotide arrays. (in press)

Li, C. and Wong, W. H. (2001a). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proc. Nat. Acad. Sci. USA 98(1), 31-36.

Li, C. and Wong, W. H. (2001b). Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. Genome Biol 2(8), research0032.1-0032.11.

Teng, C-H, Nestorowicz, A., and Reifel-Miller A. (1999). Experimental designs using Affymetrix GeneChips. Nature Genetics 23, 78, DOI: 10.1038/14415 Poster Abstracts

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed model. J. Comp. Bio., 8, 625-637.

Chapter 4. Comparison of Statistical Performance of Univariate and Bivariate Mixed Models for Affymetrix® Probe Level Data

4.1 Abstract

Half of the probes on Affymetrix® microarrays contain a single base mismatch (MM) of a known perfect match (PM) target sequence. While putatively designed to detect non-specific binding, the MM data can also contain true signal, and so debates persist concerning how to best combine PM and MM data for statistical modeling purposes. Most current approaches involve either subtracting some function of MM from PM or ignoring MM altogether. Here a bivariate model is proposed to include both PM and MM based on the mixed linear modeling framework. It directly models the correlation between PM and MM and thereby increases the power of significant gene detection.

It is shown that the bivariate mixed model offers moderate gains in power over a comparable univariate model that ignores the MM data. The gains are more prominent when the number of replicates and the array-to-array variability is small. The models are applied to a small experiment on yeast and these data are used as a basis for a Monte Carlo simulation.

Key words: Affymetrix®, mixed model, bivariate model, perfect match probes, mismatch probes.

4.2 INTRODUCTION

Affymetrix® arrays are currently the most widely adopted commercial platform for high throughput gene expression profiling. Characteristic features of these arrays include the multi-probe representation of each gene and the inclusion of mismatch (MM) probes. MM probes are designed to have the same strength of non-specific binding as their counterpart, perfect match (PM) probes, and so can potentially serve as an adjustment factor for the PM signals (Lipshutz, 1999). However, empirical studies report that typically 30 percent of MM signals are greater than PM signals, and several authors have proposed reasonable methods for handling the difficulties (Hubbell 2002, Irizarry 2003).

PM and MM data occur in pairs, and the two measurements typically have a correlation around 0.8 within any one chip (see Figure 4.1). While a portion of this strong correlation is undoubtedly due to non-specific binding, it is also the case that MM probes detect the true signal as well. The hybridization strength varies from sequence to sequence and is determined by the sequence composition. Promising explanations are available from thermodynamic models (Naef 2002, Zhang 2003). Since the real signal detected by MM varies from sequence to sequence, I propose a bivariate model for each gene under the mixed model framework to directly estimate the correlation between PM and MM. Rather than integrating a biophysical model of nucleic acid hybridization into the analysis, my objective is to provide a general covariance modeling approach to directly incorporate all of the MM data.

These ideas are applied to an experiment conducted by Pharmacia (Teng 2001). The organism under study was yeast, with a control group that was grown in rich media compared to a treated group grown in minimal media. The RNA samples retrieved from

each group were hybridized to three arrays. There are 9335 probe sets included in the Affymetrix® YG-S98 chip. After normalizing the log2 data by mean-centering, gene-specific univariate mixed models were fitted using only the PM probes as the first model (as in Chu et al. 2002) and compared with the results of a bivariate mixed model approach described in Section 2. Comparisons were limited to these two models because the univariate mixed model has already been shown to be very competitive compared to other popular approaches (Chu et al, 2004).

The improvement in power of the bivariate model over the univariate model is demonstrated by Monte Carlo simulation in Section 3. Several factors that contribute to the improvement of power are also discussed, including the correlation between PM and MM, the zero boundary constraint of the array variance component, and the scale of the variance components and residuals.

4.3 DATA ANALYSIS

Preliminary Check of Data and Normalization

The data quality of the six arrays were first checked. From the correlation table of log (base 2) intensity of PM probes in Table 4.1, the arrays within the same treatment are strongly correlated, which means the signals are highly reproducible in this experiment. Probes within a treatment class always show a correlation between arrays greater than 0.95, while the correlation between treatment classes is in the range of 0.88 to 0.91.

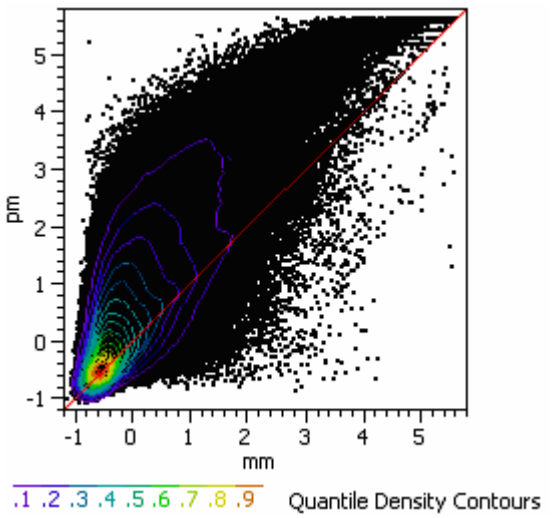


Figure 4.1. Scatter plot of $\log_2$ (PM) *vs.* $\log_2$ (MM) probes from data of six arrays

Table 4.1. Correlation coefficients between pairs of arrays of the yeast experiment. Arrays 1-3 are perfect match probes hybridized with control samples and 4-6 are perfect match probes hybridized with treated samples.

Array 1	2	3	4	5	6
1					
0.981	1				
0.952	**0.953**	1			
0.904	0.906	0.886	1		
0.894	0.889	0.883	**0.973**	1	
0.897	0.898	0.887	**0.973**	**0.980**	1

The range of the intensity is approximately the same for each array. That means we do not have significant global patterns with respect to either treatments or arrays. To remove systematic variation at the array level, a reasonable and simple normalization method is to center the log intensities so that the mean of each array is zero. This is based on the observation that the proportion of differentially expressed genes is very small and will not affect the mean value of each array.

Gene-specific models

ANOVA analysis is often used to decompose the effect into a linear combination of different factors (Kerr and Churchill 2000). To extend inference to random populations, Wolfinger *et al*. (Wolfinger et al 2001) propose a mixed ANOVA approach to model array variation as a random effect. Chu *et al*. further extended it to probe level data (Chu et al 2002), and used only PM probes to detect significant effects. The mixed model from Chu *et al* is the first model compared in this report and it is called the PM-only model throughout the discussion. The model is constructed on a gene-by-gene basis to account for the heterogeneity of variation across genes and also to avoid handling a huge data set in one complex model.

I. PM-only model

The model formulation is

$$\log_2(\mathrm{PM}_{ijk}) = T_i + P_k + TP_{ik} + A_{j(i)} + \varepsilon_{ijk}$$

The symbols *PM, T, P, TP, A* and ε represent perfect match probe intensity, treatment effect, probe effect, treatment by probe interaction effect, chip random effect and

stochastic error term. Both of the random terms are assumed to follow normal distributions as follows:

$\varepsilon_{ijk} \sim N(0,\sigma^2)$ is independent of $A_{j(i)} \sim N(0,\sigma_a^2)$,

II. Bivariate model

The correlation between the PM probes and MM probes is 0.819 as shown in Figure 4.1. To borrow the information from MM probes, I include both PM and MM intensity as paired repeated measures under the framework of the mixed model. The model is formulated as follows

$$\log_2(\mathrm{Y_{ijkm}}) = T_i + M_m + TM_{im} + P_k + TP_{ik} + MP_{mk} + A_{j(i)} + \varepsilon_{ijkm}$$

The symbol Y_{ijkm} represents the intensity for MM probes if *m* equals to 1 and it represents the PM intensity if *m* equals to 2. The symbols *T, M, P, TM, TP, MP, A* and ε represent the treatment effect, mean effect for PM and MM probes, probe effect, treatment by probe type interaction effect, probe type by probe interaction effect, chip random effect and stochastic error terms. I assume a bivariate normal distribution for the error terms of the PM and MM probes as follows:

$$\begin{pmatrix} \varepsilon_{ijk1} \\ \varepsilon_{ijk2} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{MM}{}^2 & \rho\sigma_{MM}\sigma_{PM} \\ \rho\sigma_{MM}\sigma_{PM} & \sigma_{PM}{}^2 \end{pmatrix} \right)$$ is independent of $A_{j(i)} \sim N(0,\sigma_a^2)$,

PM and MM are assumed to follow a bivariate normal distribution with correlation coefficient $\frac{\sigma_a^2 + \rho\sigma_{PM}\sigma_{MM}}{\sqrt{\sigma_a^2+\sigma_{PM}^2}\sqrt{\sigma_a^2+\sigma_{MM}^2}}$. Under this probabilistic description, σ_a^2 is constrained to be non-negative. However, it is permissible to allow the estimate of σ_a^2 to

be negative as long as the marginal variance-covariance matrix of the data is positive definite. The implication of this will be explored in the next section.

This model is similar to that of Wu *et al.* (Wu *et al*. 2004), except that ρ can be estimated directly using restricted maximum likelihood. Wu *el al*. used information from other experiments to score the optical noise and non-specific binding strength. Such extra experiments need to be performed under parallel conditions and are often unavailable.

Hypothesis Tests

The main interest here is the difference of expression levels between treatment and control groups for each gene. The hypothesis tested in the PM-only model is simply that the difference of the two treatment effects is zero:

$$H_0 : T_1 - T_2 = 0$$

For the bivariate model, the treatment effects are tested using parameters corresponding to the PM data:

$$H_0 : T_1 - T_2 + TM_{12} - TM_{22} = 0$$

The two hypotheses are comparable in the sense that they have exactly the same point estimates from the data. See the Appendix A for SAS code describing the two models and the tested hypothesis.

Results

The two models were first compared based on the significance level from the above described hypothesis tests. The scatter plot of negative log p-values is shown in Figure

4.2. All but a few hundred genes have a higher significance level in the Bivariate model than in PM-only model.

An intriguing pattern for the scatter plot in Figure 4.2 is the distinctly separated groups. They turn out to be associated with genes that have zero estimates of σ_a (the random Array effect) in the PM-only or Bivariate model. This is caused by the non-negativity constraint on this parameter applied during model fitting. About 7% of the genes have a zero estimates of σ_a in PM-only model and 6% of the genes in Bivariate model. Those genes have signals that are highly consistent across the six arrays.

Zero variance component estimates increase the degrees of freedom used for hypothesis testing to a different extent for the two models. Some details are provided in Appendix B. To investigate this phenomenon more thoroughly, I dropped the constraint of non-negativity for the estimate of the variance component σ_a. It is referred to as the NOBOUND case (corresponding to the name of the SAS Proc Mixed option used to specify it). When the array variation is small, there is a good chance of getting estimates less than zero. See Appendix C for details. Besides the concern of degrees of freedom, the type I error is better controlled without the non-negativity constraint (Murray 1998). The estimates are forced to be positive or zero with the constraint, which in turn inflates the mean of variance component estimates, and decreases test size. The scatter plot of negative log p-values in the NOBOUND case is plotted in Figure 4.3. There are no distinct groups caused by discontinuous degrees of freedom estimates.

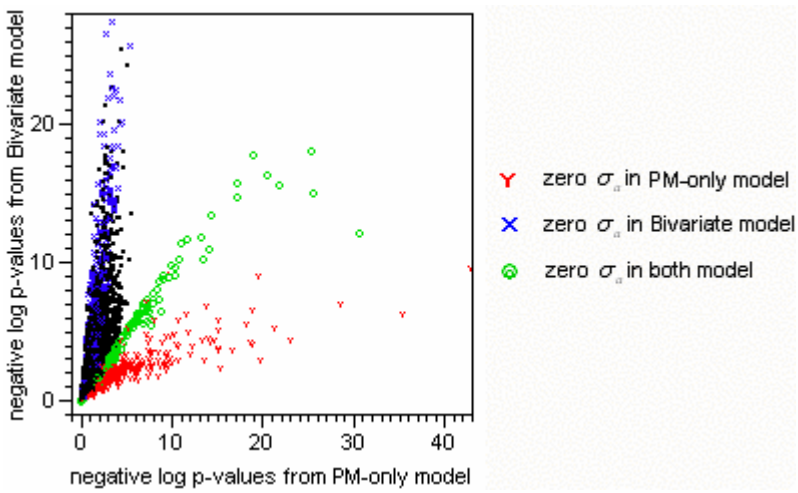


Figure 4.2. Scatter plot of negative log p-values from the two models for yeast data. The random components are estimated under the non-negativity constraint.

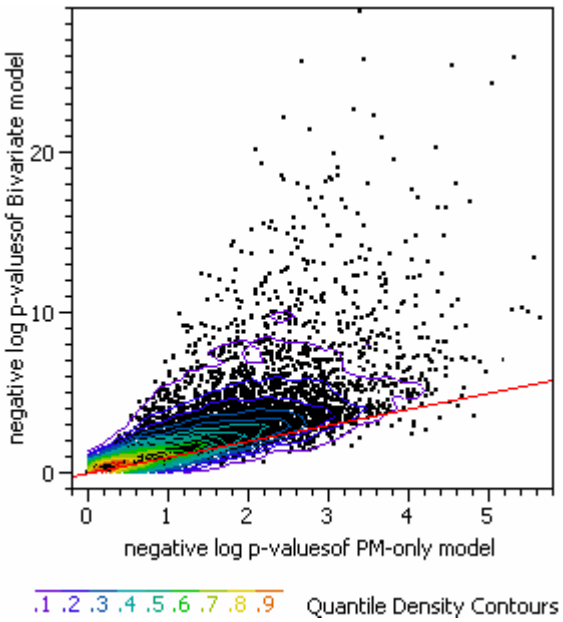


Figure 4.3. Scatter plot of negative log p-values from the two models for yeast data with NOBOUND option. The contours model a smooth surface that describes how dense the data points are at each point in that surface. The red line indicates equal p-values in both models.

The significance of the test statistics is determined by three components. They are estimates of treatment difference, standard deviation, and degrees of freedom. Figure 4.4 shows the comparison of the three components between the two models from the yeast data. Since our testing hypothesis is made to test the difference between two treatments in the PM data for both models, the estimates are identical. All gains in power in the bivariate case should be due to either smaller estimated standard errors or more estimated degrees of freedom. These two figures provide evidence for the improvement offered by the Bivariate model. To quantify the exact gains of power, Monte Carlo simulation was performed with different parameter settings.

4.4 SIMULATION

The parameters needed in the simulation were derived from the estimates of the yeast data. Figure 4.5 plots the distribution of standard deviation estimates from the PM-only and Bivariate models. The standard deviation estimates are mildly skewed, as would be expected for this kind of parameter. The estimated correlation coefficient between the errors for PM and MM in the Bivariate model has a bell-shaped distribution between -1 and 1 with mean and median close to 0.13. Note that this number is a lot smaller than the value of 0.8 from Figure 4.1. The Bivariate model captures the correlation of PM and MM not only through ρ but also through the array random effect, which models a common correlation amongst all the PM and MM measurements on the same chip.

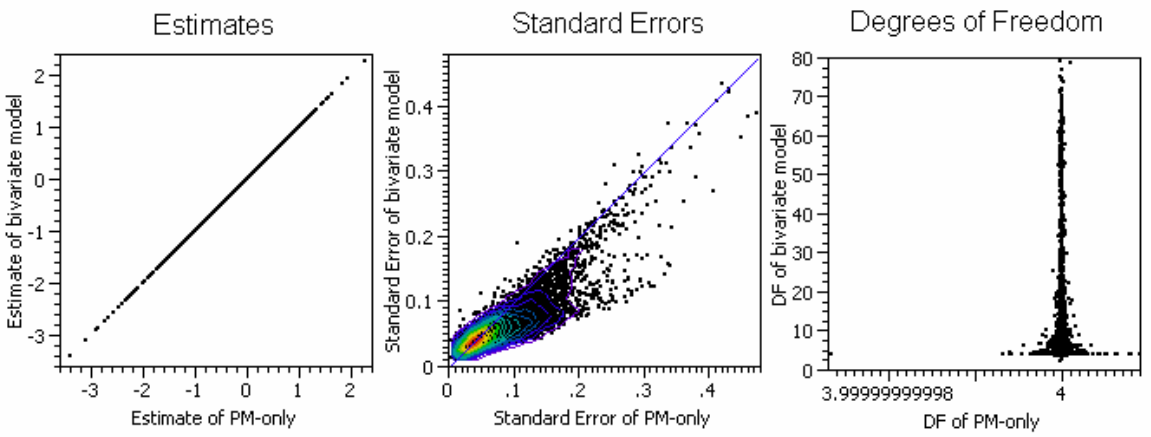


Figure 4.4 Comparison of (a) estimates of treatment effect difference, (b) standard deviation of the estimates, and (c) estimates of degrees of freedom from Kenward-Roger's method for the two models

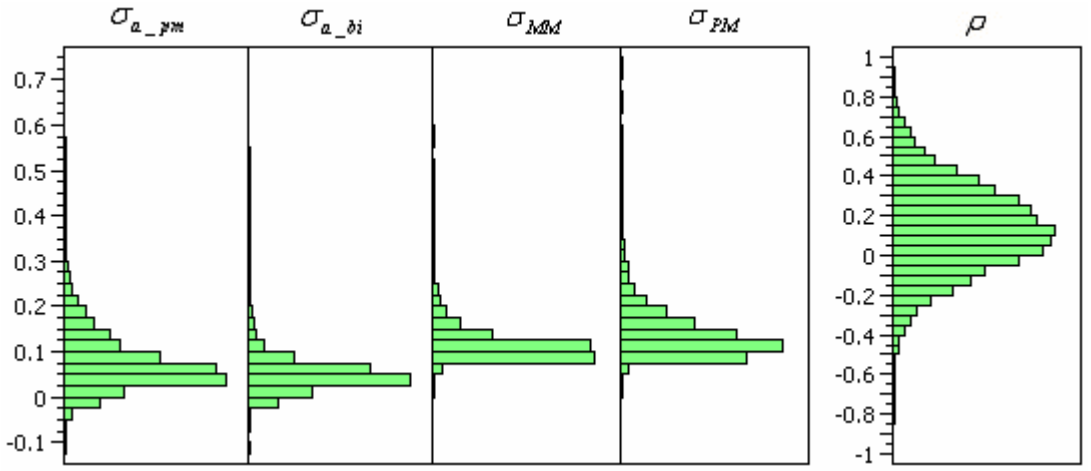


Figure 4.5. Histograms of covariance parameter estimates from the yeast data. $\sigma_{MM}, \sigma_{PM}, \rho$ are from the residual matrix of bivariate model. σ_{a_bi} and σ_{a_pm} are the array random components for bivariate model and PM-only model respectively. The parameters are derived under the NOBOUND case, thus permitting negative estimates of the array standard deviations. The standard deviation estimates are computed from the variance component estimates using a signed square root. If the variance is smaller than zero, the standard deviation is -1 times the square root of the absolute value of the variance. Otherwise, it is the simple square root.

I picked several values to represent each factor and simulate the data from the Bivariate model. The values are $\sigma_{MM} = 0.1, 0.15$, $\sigma_{PM} = 0.1, 0.15$, $\rho = -0.8, -0.4, 0, 0.4, 0.8$, $\sigma_a = 0.01, 0.03, 0.05$, 0.1. For each combination of parameters, 1000 sets of data were generated from a bivariate normal distribution. The treatment difference $T_1 - T_2$ was specified from 0 to 0.2. The treatment effect $T_1 - T_2$ was tested at significance level 0.05 and the power curves are shown in Figure 4.6. The standard deviation of detection rate is bounded by $\sqrt{\frac{0.5 \times 0.5}{1000}} \approx 0.016$. The parameters represented in Figure 4.6 are close to median values of Figure 4.5.

It is clear that the bivariate model shows more power than the PM-only model when the random components are small and when the treatment difference is less than 0.1 on the log2 scale. It creates the most margin when the random components are $\sigma_{MM} = 0.1$, $\sigma_{PM} = 0.1$, and $\sigma_a = 0.01$. I will focus on this combination in the discussion of different correlation coefficients.

There is a trend for the power to increase with respect to correlation coefficients of PM and MM errors (Figure 4.6). The power increase is the smallest when ρ is close to 1. It seems that the statistics formed from the bivariate model possess more efficiency when the PM and MM errors are negatively correlated, a state perhaps induced by competitive binding. This trend is conceptually close to that for antithetic variates in Monte Carlo simulation (Fishman 1972).

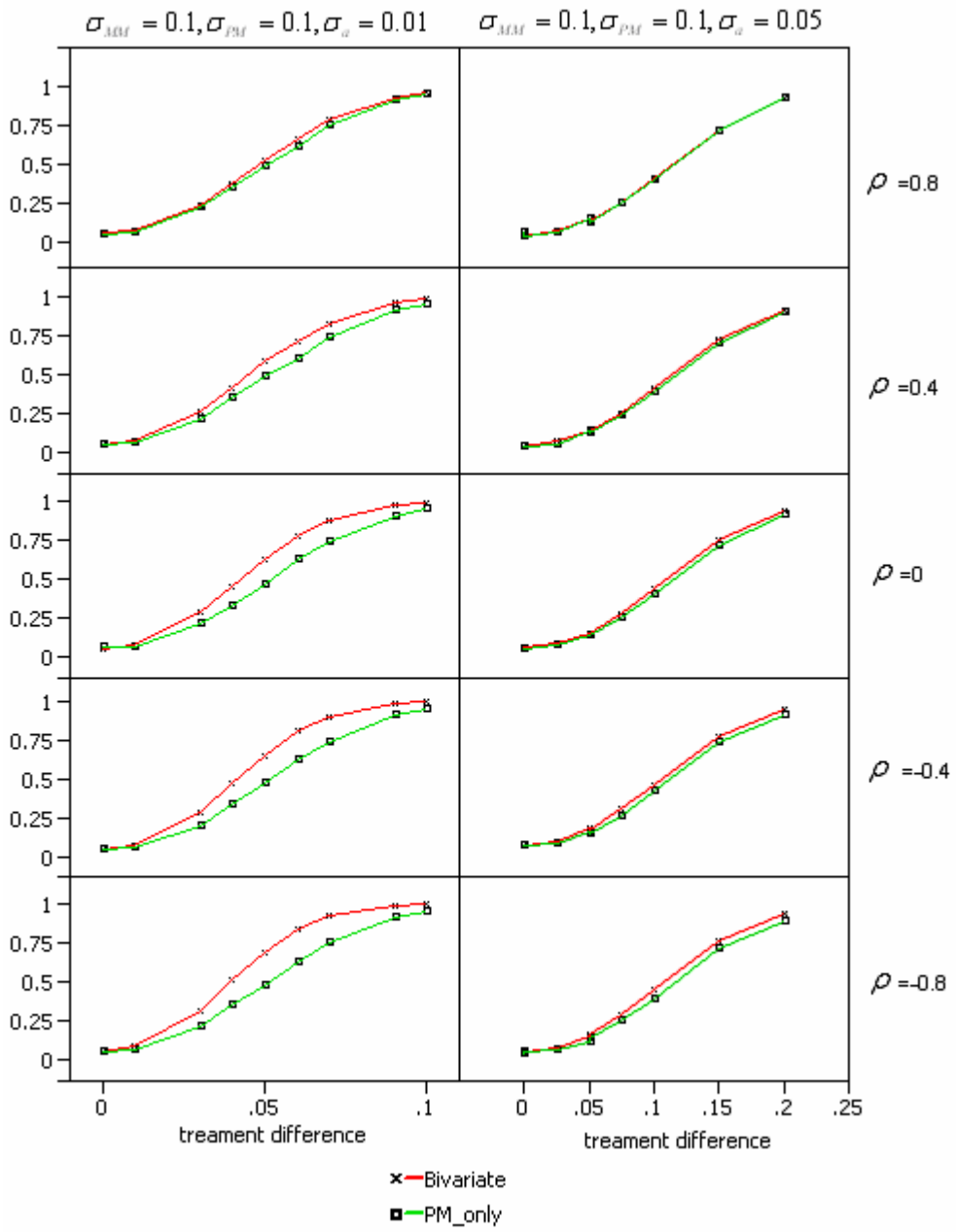


Figure 4.6 Power curves. The x-axis stands for the difference of treatment effects and the y-axis stands for the proportion of rejection in 1000 simulations. The type I errors at treatment difference zero meet the specified size 0.05.

To further explore this efficiency gain, I checked the estimates for the standard errors and degrees of freedom. The simulation results were considered under the scenario $\sigma_a = 0.01, \sigma_{MM} = 0.1,\ \sigma_{PM} = 0.1$, and $T_1 - T_2 = 0.05$, all combined with different values of ρ. 1000 sets from the Bivariate model were generated, and the statistics were derived with both models. The mean values of standard errors did not seem to differ across different ρ values although the variation grows larger for more positively correlated data (Figure 4.7). There was an interesting trend of degrees of freedom estimates with respect to ρ (Figure 4.8). The degrees of freedom had a much wider range when ρ took more negative values. This trend matches the increase of power observed in Figure 4.6.

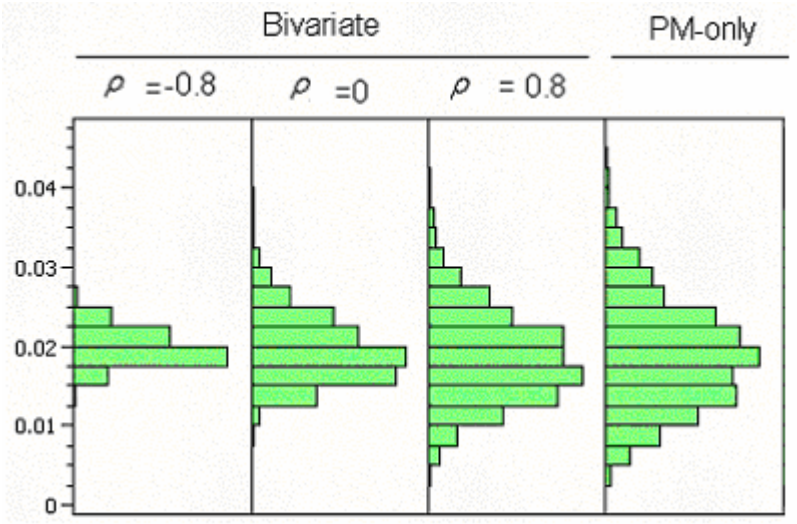


Figure 4.7 Histograms of standard errors from 1000 simulations under the condition of $\sigma_a = 0.01, \sigma_{MM} = 0.1$, $\sigma_{PM} = 0.1$ and $T_1 - T_2 = 0.05$

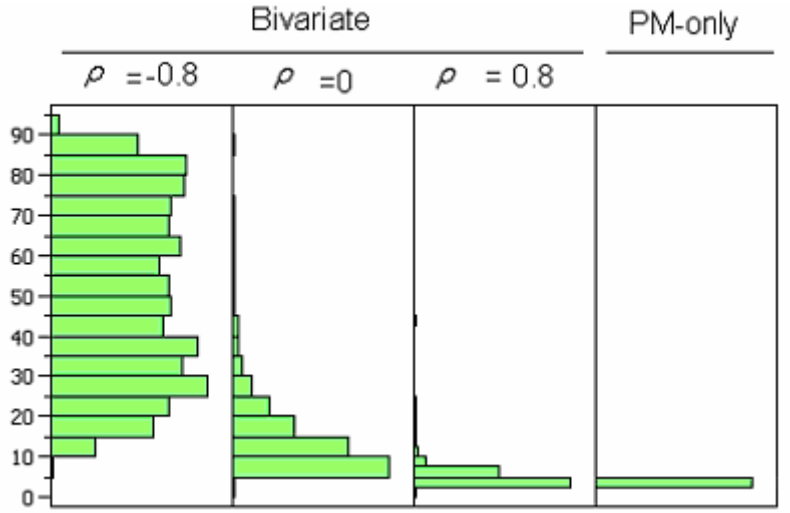


Figure 4.8 Histograms of estimates of degrees of freedom from 1000 simulations under the condition of $\sigma_a = 0.01, \sigma_{MM} = 0.1$, $\sigma_{PM} = 0.1$ and $T_1 - T_2 = 0.05$

4.5 ANOVA

To summarize these simulation findings, I put all factors together in an ANOVA model to see which one contributed most to the power improvement. The response variable is the power of detecting the treatment difference at $T_1 - T_2 = 0.05$. It was calculated as the number of simulations in which the difference was detected correctly, divided by the total number of simulations. The main effects included the standard deviation of the MM probes (σ_{MM}), the standard deviation of the PM probes (σ_{PM}), the standard deviation of the random effect (σ_a), the correlation between PM and MM errors (ρ), an indicator variable for either of the two models (M), and an indicator variable for the NOBOUND case (NB). Since I did not simulate all combinations of these parameters, the data were fit with a reduced model consisting of only main effects and all pairwise interactions. The Type III effect tests are provided in Table 4.2. The rank order of significant main effects was as follows: 1. σ_a, 2. σ_{PM}, 3. model, and 4. correlation coefficients ρ. Our choice of settings for these parameters make their F-statistics roughly comparable. To sum up, the values of chip random effect and the stochastic error term seem to have the most crucial effect on the detection power. Their relative values also contribute significantly.

Table 4.2. Type III table for the ANOVA analysis of the factors contributing to the explanatory power of the model.

Effect	Num DF	F values	Pr > F
σ_{MM}	1	1.57	0.212
σ_{PM}	**2**	**681.11**	**<.0001**
$\sigma_{MM} * \sigma_{PM}$	1	1.08	0.2993
σ_a	**3**	**1916.67**	**<.0001**
$\sigma_{MM} * \sigma_a$	2	5.33	0.0055
$\sigma_{PM} * \sigma_a$	**3**	**390.9**	**<.0001**
ρ	**4**	**11.48**	**<.0001**
$\sigma_{MM} * \rho$	4	0.9	0.463
$\sigma_{PM} * \rho$	8	0.87	0.5427
$\sigma_a * \rho$	**12**	**3.74**	**<.0001**
nb	1	0.84	0.3592
σ_{MM} *nb	1	1.3	0.2561
σ_{PM} *nb	1	1.87	0.1732
σ_a *nb	3	0.85	0.4674
ρ *nb	4	0.24	0.9171
model	**2**	**139.25**	**<.0001**
σ_{MM} *model	2	2.41	0.0921
σ_{PM} *model	3	1.05	0.373
σ_a *model	**4**	**28.87**	**<.0001**
ρ *model	**8**	**8.25**	**<.0001**
nb*model	2	0.93	0.397

4.6 DISCUSSION

Owing to the high cost of microarrays, many research projects can only afford a small number of replications, which adversely affects statistical power for detecting differential expression. It is thus desirable to explore statistical procedures that increase inferential power from datasets with small numbers of replicates. I propose a bivariate model under the mixed ANOVA framework and demonstrate the power increase relative to the PM-only model, which utilizes only half of the data from each array. The bivariate model not only integrates more data points but also dynamically estimates the covariance structure between PM and MM probes.

I have made a relatively thorough investigation concerning the standard errors and degrees of freedom estimates for mixed model test statistics and find an antithetic trend of power improvement with respect to the correlation coefficients between PM and MM probes. This is an interesting behavior of the bivariate model and is worth more theoretical discussion.

Overall, at the cost of additional theoretical and computational complexity, the bivariate model improves power over the PM-only model by utilizing extra information from all MM probes. It works well especially when both the treatment difference and chip-to-chip variability are small. For an experiment with a greater number of replicates, the improvement will likely become insignificant since then we will have better estimates from the PM-only model and the margin created by the degrees of freedom difference will decrease. The computational time needed for the bivariate model is approximately 10-20 times more than the PM-only model and so the bivariate model is only recommended when the treatment effects are expected to be small and when the

experiment has a small sample size. However, under these circumstances it must be recognized that biological or technical experimental sources of variation are more likely to contribute to apparent differential expression. Caution is urged in interpreting higher statistical significance as true biological difference.

In general, the mixed model approach provides flexibility for modeling different sources of variation and correlation. In the bivariate model discussed here, the variance components can be refined to separate the variation from PM and MM. It can also be constrained so that $\sigma_{MM} = \sigma_{PM} = \sigma$. Reducing the number of parameters can result in even better estimates of other parameters and improved power, as long as the reduced model still fits the data well.

4.7 REFERENCES

Chu T-M, Weir B, Wolfinger R. A. (2002) A systematic statistical linear modeling approach to oligonucleotide array experiments. Mathematical Biosciences 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(4):500-6

Fishman, G.S. (1972) Variance Reduction in Simulation Studies. Journal of Statistical Computation and Simulation, 1:2, pp. 173-182.

Hubbell E, Liu WM, Mei R. (2002) Robust estimators for expression analysis. Bioinformatics 18:1585–1592.

Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP. (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Biostatistics. Vol. 4, Number 2: 249-264

Kenward MG, Roger JH. (1997) Small sample inference for fixed effects from restricted maximum likelihood. Biometrics. Sep;53(3):983-97.

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

Lipshutz, R. J., Fodor, S., Gingeras, T. & Lockhart, D. (1999) High density synthetic oligonucleotide arrays. Nat. Genet., supplement 21, 20–24.

Murray DM, Hannan PJ, Wolfinger RD, et al. (1998) Analysis of data from group-randomized trials with repeat observations on the same groups. Statistics in Medicine. 17 (14): 1581-1600

Naef, F., Lim,D.A., Patil,N. and Magnasco,M. (2002) DNA hybridization to mismatched templates: a chip study. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., **65**, 040902.

Searle, S. R., Casella, G., and McCulloch, C.E. (1992) Variance Components, New York: John Wiley & Sons, Inc.

Teng, C.-H. (2001) A Repeated Measure Approach for Analyzing Affymetrix Feature Data, JSM Proceeding.

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 Comput Biol. 8(6):625-37.

Wu, Zhijin, Irizarry, RA, Gentleman, R, Martinez Murillo, F, Spencer, F. A. (2004) Model Based Background Adjustment for Oligonucleotide Expression Arrays. To appear in JASA.

Zhang L, Aldape KA, Miles MF. (2003) A model of molecular interactions on oligonucleotide microarrays. Nat. Biotech. 21(7):818-821

4.8 APPENDIX

A. SAS code

```
*PM-only model;

proc mixed data=pharmacia;
   by unit;
   where flag=pm;
   class probe treatment chipid;
   model logi = probe|treatment /ddfm=kr;
   random intercept / subject=chipid(treatment);
   estimate 'trt' treatment 1 -1;
run;


* Bivariate model;

proc mixed data=pharmacia;
   by unit;
   class probe flag treatment chipid;
   model logi = probe|flag|treatment@2 / ddfm=kr;
   random intercept / subject=chipid(treatment);
   repeated flag / type=un subject=probe(chipid*treatment) r rcorr;
   estimate "trt:conditioned on PM" treatment 1 -1 flag*treatment 0 0 1 -1;
run;
```

B. Degrees of freedom estimates with and without NOBOUND

When the array variation is greater than zero, the degrees of freedom estimated by Kenward-Roger's method (Kenward and Roger 1997) approximately equal (number of subjects - number of parameters estimated for main effects], which is 6-2=4 in the yeast data and is comparable in both the PM-only and bivariate models. When the estimate of array variation is zero, the test statistic is the same as that in fitting a fixed-effect ANOVA and the degrees of freedom is [n - total number of parameters estimated], where

n is the total number of observations. The number *n* is about 100 for one gene in PM-only model of the yeast data, depending on the number of probes representing the gene. This number is doubled in the bivariate model. Increasing the degrees of freedom from 4 to 100 greatly improves the power of a t-test.

C. Calculating the probability of getting negative estimates

A simple model considered here has two treatments and three replicates in the PM-only model. The probe set under consideration is assumed to have 16 probes, which is a general case.

$$y = \mu + T_i + P_j + TP_{ij} + A_{ik} + \varepsilon_{ijk}$$
$$A_{ik} \sim N(0, \sigma_a^2), \varepsilon_{ijk} \sim N(0, \sigma_\varepsilon^2)$$

T: treatment, i=1,2
P: probe, j=1~16
A: array effect, k=1~3

Table 4.3. Model expected mean squares

Source	Type III Expected Mean Square	D.F.
Treatment	Var(Error)+16Var(ChipID(Treatment))+ Q(treatment, Probe*Treatment) $=\sigma_\varepsilon^2 + 16\sigma_a^2 + Q(T, P*T)$	1
Probe	Var(Error)+Q(Probe, Probe*Treatment) $=\sigma_\varepsilon^2 + Q(P, P*T)$	15
Probe*Treatment	Var(Error)+Q(Probe*Treatment) $=\sigma_\varepsilon^2 + Q(P*T)$	15
ChipID(Treatment)	Var(Error)+16Var(ChipID(Treatment)) $=\sigma_\varepsilon^2 + 16\sigma_a^2$	4
Error	Var(Error) $=\sigma_\varepsilon^2$	60

Let MSC represent the mean sum squares of variation from array effect and MSE represent the mean sum squares of error. We have

$$\hat{\sigma}_a^2 = (MSC - MSE)/16$$

Under the normality assumption: $\begin{bmatrix} A \\ \varepsilon \end{bmatrix} \sim N\left(0, \begin{bmatrix} \sigma_a^2 I_6 & 0 \\ 0 & \sigma_\varepsilon^2 I_{96} \end{bmatrix}\right)$

It can be shown (Searle 1992) that $\frac{SSE}{E(MSE)} \sim \chi^2_{60} \Rightarrow \frac{60MSE}{E(MSE)} \sim \chi^2_{60}$

Similarly $\frac{4MSC}{E(MSC)} \sim \chi^2_4$

When $\sigma_\varepsilon = 0.1, \sigma_a = 0.01$, according to Searle (1992), we can calculate the probability of getting negative estimates as

$$\begin{aligned}
\Pr(\hat{\sigma}_a^2 < 0) &= \Pr\{MSC / MSE < 1\} \\
&= \Pr\{\frac{MSC / E(MSC)}{MSE / E(MSE)} < \frac{E(MSE)}{E(MSC)}\} \\
&= \Pr\{F_{60}^4 < \frac{0.1^2}{0.1^2 + 16 \times 0.01^2}\} \\
&= \Pr\{F_{60}^4 < 0.862\} \\
&= 0.508
\end{aligned}$$

Chapter 5. Conclusions and Future direction

My thesis has focused on two major issues in gene expression profile analysis. One was to measure the variation across and among species and make biological interpretation. The long term goal is to understand how evolution occurs in organisms at the gene expression level. The second part of my work was to develop better statistical estimates that can account for different sources of variation for significant gene detection.

In chapter 2, the experiment carried out by Enard et al was re-analyzed. In this study, I used linear mixed model to decompose the expression variation into species, tissues and probe effects. With a gene-by-gene approach, there were more genes differentially expressed between human and chimpanzee within the liver tissues than that within the brain tissues, which provides a different perspective to the observations described in the original paper by Enard *et al* (2002) who suggested that there is accelerated evolution in the human brain.

A second observation was the tendency of higher expression levels for species in the order of human>chimpanzee>orangutan. Although this may simply explained by sequence divergence increasing with genetic distance of non-human species to human, whose sequences the array was designed for, there are significantly different degrees of asymmetry between brain tissues and liver tissues. This suggests that part of the bias might be contributed by real increases in expression in addition to the effect due to sequence divergence.

One of the important features specific to short oligonucleotide arrays is the probe patterns. The relative hybridization strength of probes within a probe set maintains the

same profiles across samples most of the time. Based on this assumption, I used a correlation-based heuristic rule to filter probes that are inconsistent between chimpanzee and human probe profiles. It hypothetically removes some of the probes that have sequence divergence between the two species. It improved the asymmetry somewhat although still maintained the order mentioned above.

Although this dataset contains a limited number of biological replicates for each species, the results indicated significant divergence across species. The hypothesis testing approach proposed by Rifkin *et al.* was applied to compare the divergence observed with the divergence expected under mutation-drift equilibrium. Under conservative conditions of effective population size and the time since divergence of human and chimpanzee lineage, only a few genes were suggested to be experiencing diversifying selection.

Given the issues that probe variation raises for interpretation of short oligonucleotide based gene expression profiles, the remainder of my research focused on development of statistical methods for teasing apart probe effects. These were pursued using two published experiments where perfect match and mismatch probes were both considered. Genetic variation among individuals in a population is expected to be small.

In chapter 3, a spike-in dataset was used to evaluate the linear mixed model. The Latin Square design with known genes and known concentration allows us to compare the estimates with true answers. There were three observations in this project. First, by comparing the correlation across arrays, the log 2 scale was shown to have the best consistency across the arrays. Second, when the least squares mean were used to represent the expression pattern across experiments, all 14 patterns were correctly

identified but the estimates and the true concentration did not follow a strict linear relation; instead, they fell into an S-shape. The linearity held better only for a certain range of concentrations. The last observation was the significant cross-hybridization to genes other than the 14 spiked ones. Some examples identified by Gibb's sampling method were given and they demonstrated how complicated the issue could be since a motif with length as short as 15 bases can possibly cause strong binding.

Following the observation in Chapter 3, a bivariate model was proposed in Chapter 4 to combine the mismatch probes as repeated measures to the perfect match probes under the framework of linear mixed models. Power increase was demonstrated for small samples through both simulation and application to a real yeast dataset. Although there is no significant improvement to PM-only models when the sample size is large, it is still a good alternative when there are only limited resources to do the experiment.

Based on the conclusion and observations above, there are some issues needed to be resolved in the future. One is cross-species hybridization, which was partly taken care of in this study by using a heuristic rule. For highly variable species such as Drosophila, intraspecific sequence polymorphism is greater than cross-species divergence among primates, so may have a large impact on inference of differential expression. A more detailed understanding about sequence similarity and hybridization strength will help to evaluate expression variation within and across species.

The second problem that needs to be studied is to understand the mechanism behind the expression divergence. It is the first time of the history that scientists are able to look at the behavior of tens of thousands of genes at the same time. By treating

expression as quantitative traits, the high-dimension data provide us with a different angle to look at evolution at the molecular level.

Throughout the thesis, the normalization methods were based on a strong assumption that most of the genes are not differentially expressed across experiments. This assumption holds for most whole genome experiments, but new applications for a small set of target gene detection could be problematic. The current available normalization methods are more or less based on the same assumption. A more general approach should be developed to fit all circumstances.

How to handle the cross-hybridization signals has been discussed in the microarray community but no conclusions have been reached so far. Motif finding based on sequence similarity was demonstrated to be an informative strategy in the thesis but it is not possible to have a comprehensive survey with spike-in data. It is sensible to consider the chemical and physical properties of the oligonucleotide sequences when dealing with this problem.