Statistical Analysis of Gene Expression Profile: Transcription Network Inference and Sample Classification

Nan Bing

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Genetics, Bioinformatics and Computational Biology

Ina Hoeschele, Chair
Saghai Maroof
Pedro Mendes
Naren Ramakrishnan
Keying Ye

April 5, 2004
Blacksburg, Virginia

Keywords: Microarray, Gene Network, Genetical Genomics, Structural Equation Model, Classification, Mixture Model

Statistical Analysis of Gene Expression Profile: Transcription Network Inference and Sample Classification

Nan Bing

(ABSTRACT)

The copious information generated from transcriptomes gives us an opportunity to learn biological processes as integrated systems; however, due to numerous sources of variation, high dimensions of data structure, various levels of data quality, and different formats of the inputs, dissecting and interpreting such data presents daunting challenges to scientists. The goal of this research is to provide improved and new statistical tools for analyzing transcriptomes data to identify gene expression patterns for classifying samples, to discover regulatory gene networks using natural genetic perturbations, to develop statistical methods for model fitting and comparison of biochemical networks, and eventually to advance our capability to understand the principles of biological processes at the system level.

Acknowledgements

I would like to thank my advisor, Dr. Ina Hoeschele, for accepting me and introducing me to the field of Statistical Genomics and Bioinformatics. I have benefited a lot from her permission to let me freely choose courses in my first two years of graduate study, her rigorous training on statistics, her willingness to let me freely develop ideas, and finally, her criticism and guidance in research and writings. I greatly appreciate her effort and time, and cherish the four years I have spent in her group.

The diverse background and kind help from my committee members made this dissertation possible. I have always been able to obtain assistance from Dr. Ye, when there are statistical puzzles in my research, and he always give me quick and clear answers. Dr. Mendes introduced me to the field of biochemical network simulations and I am impressed by his sharp mind and good sense of science. Dr. Maroof sparked my first interest in Microarray data analysis, when I took his class on Genomics, and he has always been supportive and provided suggestion on my study of "Genetical Genomics". Dr. Ramakrishnan is one of the smartest people I know. I learned a lot from his class on Artificial Intelligence and the learning is reflected in the thesis.

I appreciate many helpful suggestions from my colleagues Drs. Guiming Gao, Xiao Yang, Hua Li, Peter Sorensen and especially Alberto de la Fuente. Alberto and I have very good discussions from time to time and he has been helping me on biochemical network simulations.

Finally and foremost, thanks to my parents and especially my wife, Di Chen, for their love and support.

Contents

List of Figures

List of Tables

Chapter 1

Literature Review

1.1 From Genetics to Functional Genomics

It was exactly at the turn of the last century, that Mendel's principles (Mendel, 1865) for how traits are inherited were rediscovered by the scientific community, which marked the beginning of modern genetics. Within the following 100 years, great progresses have been made in biology. Definition of chromosomes and genes as the carrier of inheritance (Baltzer and Boveri, 1964; McKusick, 1960), double helix three dimensional structure determination of DNA molecules (Watson and Crick, 1953), technology breakthroughs such as recombinant DNA (Jackson et al., 1972), DNA molecular sequencing (Sanger et al., 1977), and mapping genes by DNA polymorphisms (Botstein et al., 1980) have dramatically advanced our knowledge of biology and fundamentally changed the way people perceive life and disease. Now, at the turn of new century, we are facing another opportunity of revolution in biology and medicine from the development of Genomics. The successful completion of Human Genome Project gives us an atlas of the human genes (IHGSC, 2001; Venter et al., 2001). Genome sequencing of more than a hundred of other important organisms has been completed with hundreds of others undergoing. With a genome blueprint in hand, we will gain more information about organization of their structures, linkage between genome location and disease, distribution of regulatory sequence in genome, molecular evolution among species. However, sequence by itself doesn't tell researchers what genes do. Knowing sequence is merely the first chapter of genome biology. It is more

important to understand how genes work to comprise diverse functioning cells and organisms, as well as how genes interact and respond to stimulus from environment.

Technology such as complementary DNA (cDNA) microarray (Schena et al., 1995) or oligonucleotide DNA chips (Lockhart et al., 1996) are already allowing scientists to monitor genome-wide transcription profiles under different physical, environmental and genetic circumstances. Both oligonucleotide and cDNA microarray technologies are based on the hybridization property between nucleic acids. A description of cDNA microarray schema will demonstrate the principles for both. In a cDNA array experiment, templates for genes of interest are obtained and amplified by PCR (Polymerase Chain Reaction). Following purification and quality control, cDNAs are printed to specific locations on coated glass microscope slides using a computer–controlled, high–speed robot. To compare the relative abundance of each of these gene sequences in two DNA or RNA samples, the two samples are fluorescent labeled with either Cyanine3 (green) or Cyanine5 (red) dUTP using a single round of reverse transcription. The fluorescent samples are pooled and hybridize to the DNA spots on the array. The same DNA molecule from different samples will compete for the binding of their partners on the slide. After hybridization, laser excitation of the incorporated fluorescence (red or green) is measured using a scanning confocal laser microscope. The relative ratio or the direct measurement can be used to determine abundance of the sequence for each specific gene in the two mRNA or DNA samples. Then it is possible to compare gene expression between two cell states (Duggan et al., 1999).

Since its introduction in mid 90's, microarray technology has been more and more applied to biology and medicine to address a broad range of questions: general application on gene differential expression analysis under various environmental or biochemical conditions (Kim et al., 2001; Ooi et al., 2001; Schena et al., 1996); experimental annotation of human genome (Shoemaker et al., 2001); functional discovery of genes (McDonald and Rosbash, 2001); prediction of drug target (Hughes

et al., 2000); molecular expression profiles of tumors (Alon et al., 1999); identification genetic markers associated with tumors (Welsh et al., 2001); tumor classification, prediction and subclass discover (Alizadeh et al., 2000; Bittner et al., 2000; Golub et al., 1999; Khan et al., 2001); prediction clinical outcomes of diseases by expression profiles (Pomeroy et al., 2002; Shipp et al., 2002; Takahashi et al., 2001; van 't Veer et al., 2002); analysis of development pathways (Davidson et al., 2002; White et al., 1999); systematic determination of the genetic network architecture maintaining metabolic networks, cell growth and cycles (DeRisi et al., 1997; Ideker et al., 2001; Jorgensen et al., 2002; Lee et al., 2002; Tavazoie et al., 1999); and candidate gene confirmation in QTL mapping (Wayne and McIntyre, 2002).

The successful applications of microarray have provided researchers with a new way to think about biology. It has enabled investigators to progress from studying the expression of one gene in several days to hundreds or thousands of gene expressions in a single day. In contrast to the traditional one gene per experiment approach, genome-wide surveys of gene expression are leading biologists to a more systemic view of biology. It is predicted that microarray, combined with currently developing Proteomics and techniques quantifying small molecules in cells such as metabolites, will greatly enhance our ability to understand the fundamental mechanisms of life, significantly benefit agriculture in selecting genetically improved animals and plants of economic importance, and eventually revolutionize medical diagnosis, treatment of diseases, and the design of effective drugs.

While the data from microarrays contain plenty of valuable information, they are usually buried with substantial error and variability, and are highly dimensional with many latent variables. Analyzing and interpreting the data in the large scale presents a daunting challenge to scientists. Without appropriate study design and tools for data inference and information extraction, the promised benefits for the future of molecular genetics and medicine will not fully and efficiently materialize. To make sense of the

complex and large amount of data generated from microarray, reliable statistical and computational methods must be applied. Usually some or all of the following steps must be included to analyze microarray data: (1) Identification sources of variation and statistical design of experiments so that the efficiency and reliability of the obtained data can be improved. (2) Identification of genes, which are differentially transcribed across the different mRNA samples included in the experiment. (3) Exploratory analysis of transcription profiles such as clustering to reveal the relationship either among genes, among samples, or between genes and samples. It can be used for gene function discovery or sample grouping. (4) Discriminant or classification analysis to separate samples and predict unknown samples to known classes. (5) Inferences about gene interactions under different circumstances and reconstruction regulatory network underlying biochemical pathways.

1.2 Experimental Design, Data Normalization and Detection of Differential Expression

1.2.1 Why Experimental Design and Principles

Microarray experiment is a new and promising technology. On the other hand, it is expensive, time consuming and contains many source of variation in its data. A good experimental design can improve the efficiency of the experiment to reach the same or better conclusion of interest without increase the cost and time, most important is that a good design will help to identify and remove systemic variation which usually impairs our ability to answer the question of biological interest.

Usually, design of experiments depends on the experimental purposes and the availability of the experimental resources. It is hard to give a universal recommendation for all situations, but general principles such as replication and balance are usually required. Microarray data is rather variable such that single observation of gene

expression may reflect a deviation from the true value due to different source of variability. Without knowing the scale of the variability for genes expression among and across classes and the contribution of variability from technical resources, any conclusion will be spurious and misleading. Systematic errors are biases; they result in a constant tendency to over- or underestimate true values and may depend on other factors (e.g. spatial locations, signal intensities). Replication and balance of theses systematic factors will reduce systemic variability by average values among replicates. Such replication will enable us to reduce systematic error and estimate biological error against which differences among treatments of interest are judged.

Replications can be implemented in different levels. Place different aliquots of the same sample to different location of the same array, to different arrays and by different dye labeling will unable us to detect the variability from spot in the same slide, from different slides and from different dyes. Such kind of replication is called subsampling or repeated measure since it uses DNA from the same biological individual. This replication will give the information about the systematic variability, which needs to be reduced to obtain the real estimates of interest. Another important replication is to replicate for different individuals of the same biological class such as in tumor study, tissue samples from different individuals with the same kind of tumor and samples from different normal persons. Biological replication will allow us to detect variability of biological interests, which should be the major concerns of microarray experiments.

Balance is usually required to average out suspected systematic variability factors. Then an average of the factors of interest will balance the random error from factors not interested such that effects of interest are not confounded with other sources of variation.

1.2.2 What Kind of Designs, Pros and Cons

Reference sample designs are widely used in miaroarray studies, in which a common

reference sample on every array is used, so that different samples on different arrays can be compared with each other by relative ratios to the common reference samples. Realization that comparisons are compared indirectly and half of the experimental information on reference samples is of little or no interest, a balanced dye swap design using two arrays is proposed, which compare the samples directly without a common reference sample (Kerr and Churchill, 2001b; Kerr and Churchill, 2001c). Dye-swap design provides technical replication and avoids confounding of effects. If there is only one factor of interest, e.g. treatment vs. control, dye-swap design can compare treatment vs. control directly without introduce a reference and the comparison will be more precise than reference design. If the factors of interest are more than two, a loop design can be implemented where each sample is labeled once with Cye 3, the other with Cye 5 and comparisons among factors are connected via a loop. The loop design has the advantage to collect as twice amount of data as reference design with the same number of arrays, and is more efficient than reference design when there is more than two but less than ten factors of interest (Kerr and Churchill, 2001b). However, in loop design, some factors can't be compared directly but through other samples in the loop. So it is possible that comparison of some factors will not be as precise as in reference design when the link for the two samples is large in terms of connection distance in the loop, since in reference design the link of two factors will never be more than two steps. Reference design has other practical advantages: simple to implement; one bad quality array will not affect other array's inference; and the possibility to expand experiments across time, even across several different labs. So both statistical and practical issues need to be considered when an experiment is designed.

1.2.3 Data Transformation and Normalization

Data transformation and normalization are needed after the raw fluoresces information is obtained. This is necessary to reduce the systematic variability that occurs in every microarray experiment, so that meaningful and precise comparison of gene expression

across slides can be used for further computational and biological inference. Several reasons promoting normalization of data are different initial concentrations of RNA, differences of labeling and detection efficiency for different dyes and different slide contribution to the expression value.

Different normalization procedures have been proposed with different purposes. Background correction is usually the first step so that different background from different slides will not be confounded in expression level. Log transformations are usually used, which can naturally transform the proportional differences between intensities into additive ones and transform the errors form proportional to intensity into additive errors (Rocke and Durbin, 2001). Global total intensity normalization shifts the center of the distribution of log ratios to zero: $\log_2 R/G \rightarrow \log_2 R/G - c$, where c is the mean or median of the intensity log ratio. Global total intensity normalization adjusts for the difference of initial RNA quantities from different samples for each array based on the assumption that approximately the same number of labeled molecules from each sample should hybridize to the arrays and the total hybridization intensities summed over all elements in the arrays should be the same for each sample (Quackenbush, 2001). Apart from global normalization methods, several reports have indicated that systemic errors associated with intensity value and spatial arrangements. The intensity dependent error has been illustrated by RI (Ratio by Intensity) plot (also referred as MA plot) by several studies (Dudoit, 2002; Quackenbush, 2002; Yang et al., 2002). Under the assumption approximately even number of up and down regulated genes, the RI plot should show a random and even distribution around line 0 of log ratios. However, usually RI plot shows a stronger intensity for green dyes than red dyes and a curved line depending on intensity level. Locally weighted linear regression (LOWESS) analysis has been proposed to remove such intensity dependent effects in the log ratio values (Dudoit, 2002; Yang et al., 2002). LOWESS performs a local intensity dependent adjustment $\log_2 R/G \rightarrow \log_2 R/G - c(I)$, where $c(I)$ is the LOWESS fit value of the RI plot. LOWESS transformation can also be implemented to local area of slides or pin

groups, which will reduce spatial dependent error. There are other transformations with different assumptions and purposes: arsinh transformation can stabilize the variance of microarray data based on assumption of quadratic relationship between the variance and intensity values (Huber et al., 2002); shit-log transformation to minimize the curvature-causing background differences (Kerr et al., 2002); lin-log transformation, where linear for low intensity and log for high intensity transformation, is used to stabilize variance (Cui et al., 2003).

Transformations are usually based on some assumptions. We must be cautious when applying them. We may run the risk to over-adjust the data to what we expected without preserve original structure of the data and introduce new errors larger than we remove. As indicated in (Quackenbush, 2002), data normalization can't compensate for poor quality data. Good experimental designs with sufficient replicates are necessary to collect highest quality data possible.

1.2.4 Statistical Detection of Differential Expression

Given suitable normalized data, determination of a list of differentially expressed gene across treatment or time series is usually the first biological concern. Several statistical methods have been proposed to deal with this problem. At the first stage of microarray experiments, significantly differential expression is assessed without replication (Chen, 1997) based on the assumptions of most of the genes are not differentially expressed and the distribution of the housekeeping genes can be used as the distribution of ratios under null hypothesis. As indicated in experimental design, single array detection without replication can't distinguish true difference from random error. Realization of this leads to a routine use of replications (Lee et al., 2000). With replicated data, classical statistical test could apply. Traditional two samples t-test is applied to each gene independently, which assume different expression error terms for each gene and significance level is determined by permutation test to account for multiple testing

(Dudoit, 2002). Analysis of variance (ANOVA) model based on careful design can integrate both data normalization and test of significant differential expression together (Kerr and Churchill, 2001c). The general framework of ANOVA model is extended to mixed model to include random terms in the model, which is more flexible and practical for assessing gene significance from microarray experiments (Wolfinger et al., 2001). Apart from classical statistical test, Bayesian approaches have also been developed (Baldi and Long, 2001; Newton et al., 2001; Theilhaber et al., 2001). A Bayesian model test of significance uses both prior knowledge and data. An empirical Bayes method using B-statistics modified from t-statistics is found to be more powerful than traditional t-test in detecting gene differential expression (Lönnstedt and Speed, 2002).

1.3 Profile Exploration – Clustering

Microarray experiments produce large amount of data difficult to comprehend, and it is useful to have methods of summarizing and extracting information from them. Various clustering and a variety of dimensional reduction methods such as principal component analysis have been proposed to summary and overview the large data set from microarray experiments. These methods are representatives of statistical exploratory and unsupervised pattern recognition techniques, which are usually used as the first step in data analysis and serve as guidance for further investigation.

Clustering is a multivariate statistical method that investigates the relationships within a set of objects in order to establish a small number of clusters for which the observations within each group are similar, but the clusters are dissimilar to each other. Similarity can be defined by different measurements such as Euclidean distance, Pearson correlation, rank correlation, and information measures. Hierarchical agglomerative clustering has been widely used in miaroarray analysis (Alon et al., 1999) since its first introduction to microarray community (Eisen et al., 1998). It returns a hierarchy of nested clusters, where each cluster typically consists of the union of two or more

smaller clusters. The structures are easily viewed and understood, and provide information about the relations of the clusters. K-means clustering (Tavazoie et al., 1999) and self organization maps (SOM) (Tamayo et al., 1999) are representatives of partition non-hierarchical clustering methods, where observations are grouped into predefined number of groups (K-means) or nodes in high dimensional space (SOM). K-means and SOM are more flexible than Hierarchical clustering, where they allow the items to be removed from one cluster to another during the iterative clustering process. SOM is more structured than K-means in that cluster centers are located on a grid. Various other techniques have also been applied in the broad sense of clustering analysis including mixture distribution modeling to find the best partition of the data by fitting a mixture of normal or other distributions via maximum likelihood or Bayesian posterior probability (Ghosh and Chinnaiyan, 2002; McLachlan et al., 2002; Medvedovic and Sivaganesan, 2002; Yeung et al., 2001); unsupervised neural network which applies the advantage of neural learning to clustering (Herrero et al., 2001); and dynamic clustering which takes into account the dynamic nature of gene expression time series during clustering (Ramoni et al., 2002). Different evaluation methods have also been proposed to evaluate the effectiveness and validity of the clusters formed (Kerr and Churchill, 2001a). However, the unsupervised and exploratory nature of the clustering methods makes it difficult to judge the validity of clusters and different methods or different choice of similarity measurements usually give different results and interpretations.

Principal components analysis (Yeung and Ruzzo, 2001), singular value decomposition (Alter et al., 2000; Holter et al., 2001; Wall et al., 2001), multidimensional scaling (Bittner et al., 2000) and correspondence analysis (Fellenberg et al., 2001) are a group of related multivariate statistical methods useful for dimensional reduction and visualization. These techniques rely on the assumptions that variability of highly dimensional gene expression data can be transformed into a set of uncorrelated variables (principal components) such that the first few components explain most of the

variation in the data. Original data will be projected to space based on the linear combination of the original variables so that the relations between the samples can be best visualized in the newly formed small (usually two or three) dimensions. However, it is usually difficult to determine the exact number of dimensions to retain and though the reduced dimensions maximize the variance of the whole gene expression, they may not reflect the relations between biological objects well.

1.4 Sample Classification

In contrast to exploratory methods where no class information is utilized, classification represents supervised methods where class information is used to construct classification and prediction models. The methods concern with separating distinct sets of objects and allocating new objects to previously defined groups. Though clustering can also be used for grouping objects, it is not as powerful as supervised methods, and usually group objects according to obvious criterion such as age but not the biological class label of interest. Various classification methods have been proposed to do sample classification and prediction from microarray data especially in tumor diagnosis. All of the methods require correctly labeled samples as learning data to construct classification models. Future samples' classes will be predicted from the model learned based on the training data.

Classification and discriminant analysis have been well studied both in statistics and machine learning field. Fisher linear discriminant analysis searches for a linear combination of original variables, which maximizes the ratio of variance for between-groups and within-groups. Based on assumptions that two multivariate normal distributions underlying distinct classes, maximum likelihood methods have been used to construct either linear discriminant analysis (LDA) rule with the assumption of same variance structures of two classes, or quadratic (QDA) rules for different variance-covariance matrix. In LDA and QDA, training data are used to construct the

multivariate normal distribution within each class. The new data will be assigned to class with the largest likelihood according the learned distribution functions. Logistic regression can also be used for class separation and prediction and is more robust to the violation of the assumptions than normal distribution based methods. The methods above are mostly parametric based on certain statistical assumptions. Other non-parametric statistics rules can also be applied. A mathematically very simple non-parametric classification procedure is the nearest neighbor method. In k-Nearest Neighbor (kNN), distance between new object and samples in training classes are calculated. The k nearest distance to new samples is recorded and the new sample is allocated to the group to which the majority of the k samples belong. Comparison of the above methods is conducted on three different microarray data sets. The result suggests simple methods such as kNN and LDA outperform complex ones (Dudoit et al., 2002). However, we believe more data and comparisons are needed to make a sound conclusion.

Other methods from machine learning community include artificial neural network (ANN) learning, classification trees, and support vector machines (SVM). Classification trees recursively partition samples by maximizing information gain from the partition process. It has been used on microarray study and found more accurate for discriminating among distinct colon cancer tissues than other statistical approaches used (Zhang et al., 2001). Classification trees have the potential to reveal gene interaction. However, with the large number of genes available, classification trees tend to be unstable and many different tree structures will have similar classification abilities. ANN is also trained for tumor sample classification (Khan et al., 2001). ANN has the advantage to model nonlinear relationship but is also prone to over fit the data. SVM can separate a given set of binary-labeled training data with a hyper-plane that is maximally distant from them. Combined with kernel functions, a nonlinear separation planes can be formed. Application of support vector machine shows that it performs comparably to other classification methods on many of microarray datasets (Furey et

al., 2000).

One specific problem in micorarray classification studies is the high dimension of the variable space. A tradition microarray experiments will usually contain measurements of thousands of genes with no larger than a hundred of observations. This is a 'large p small n' problem which will affect most of parametric statistical classification methods in that usually infinite number of models can fit the data perfectly without further constraints. Though k-NN and SVM can construct classification using all of the genes' expression values, it is believed that not all of the genes' expressions measured on a microarray experiment are relevant to the biological class or phenomenon of interest. And the existence of too many irrelevant variables can affect the classification precision. Furthermore, genes are interconnected through complex pathways and biological networks, expression of them tend to be highly correlated. The correlation of the variables gives us redundant information. Construction of classification models without accounting for collinearly in variables may give us unstable and wrong models (Myers, 1990). Realizing the existence of irrelevant and redundant information in miaroarray experiments, prior gene selection and dimensional reduction should be performed before constructing a classification model. Gene selection can be done in a similar process as we mentioned in detection differential expression, where a specific test criterion are used to select a predefined number of genes or by a significance level adjusted for multiple testing. One study (Dudoit et al., 2002) uses the ratio of between-group sum-squares and within-group sum-squares for each gene to rank their relevance to the biological classes and pre select up to 200 genes to test their effects on the performance of the classification models. A stepwise gene selection combined with Fisher's linear discriminant analysis indicates that only a subset of genes ranging from 3 to 10 can achieve high classification accuracy of a gene expression data set for colon cancer that consists of 22 normal and 40 tumor colon tissue samples (Wuju and Momiao, 2002). Principal component analysis (PCA) and Singular Value Decomposition (SVD) can also be applied as dimensional reduction methods in

classification model construction. West et al. (2001) develop a Bayesian regression framework by employing a singular value decomposition of the full matrix of expression measurements and pursuing regression on the resultant latent factor variables. PCA and SVD reduce the data to a few dimensions that explain as much of the variation in the data as possible. However, the reduced dimension may not correspond to the biological class of predication interest well. As an alternative, partial least square (PLS) is proposed so that the sample covariance between the response and the reduced dimension is maximized (Nguyen and Rocke, 2002). Gene selection and dimensional reduction are necessary for microarray classification. They can remove most of the irrelevant and redundant information and give us a precise model, which will be easy to understand. Many of the selection and reduction methods can be combined, and combination of the methods may give us better results. Above all, the interest of biology matters and should determine the choice of methods.

1.5 Genetic Network Simulation, Inference and Reconstruction

The massively parallel gene expression data acquired from cDNA microarray has shifted biologists' attention towards a more complex understanding of molecular biology. In addition to determining the roles of individual genes, genetic network analysis enables us to study cells as a complex network of biochemical factors, and helps us to understand how genes work together to comprise functional cell and life to cope with environments and disease. Understanding genetic networks will also provide us with novel drug targets, sensitive diagnostics for individualized therapy, and ways to manipulate its topology to cure disease. Several experimental strategies have been used for the purpose to learn the structure of genetic interactions or regulations: time course survey of gene expression during cell cycle (Spellman et al., 1998) or development (Davidson et al., 2002; Kalir et al., 2001; White et al., 1999); gene expression level change after treatment by biological or environmental factors (Causton et al., 2001;

Gasch et al., 2000; Gross et al., 2000; Lyons et al., 2000; Ogawa et al., 2000); genetic perturbation of specific gene and observe the change of other genes (Ideker et al., 2001); combining regulatory sequence on genome with clustered gene expression data (Tavazoie et al., 1999); genome wide location study to find interaction of transcription factor with regulatory sequences (Iyer et al., 2001; Ren et al., 2000; Simon et al., 2001) and study of gene interactions in QTL mapping (Garner et al., 2002). Though the experiments have provided us a lot of information, the key to a successful discovery of the regulatory network lies in properly matching experimental design with data mining and predictive modeling methods. In the following, I will briefly review the characteristics of genetic network and concentrate on different reverse engineering methods for genetic network reconstruction from experimental data.

1.5.1 Genetic Network and the Topological Features

Before we jump into various methods to infer the structure of gene network, we need to be clear about what genetic networks are, and what the characteristics of its organization are. Living cells respond to information from their environment, programmed regulate their growth and development on the basis of the interactions of a large number of molecules including DNA, RNA, proteins, and metabolites. These molecules work together to form an integrated complex cellular system. Genes encode proteins and proteins execute functions. Proteins and metabolites interact with each other and finally regulate and modify the expression of other genes or their own. Genes will not interact directly, rather through a chain of proteins and metabolites. Genetic network is just a projection or reflection of this complex cellular network on to gene space. Even on the genetic level, some activities other than gene expression such as differential splicing or modification will have effects on gene activity, but the regulation will not be observed by current experiment methods or need some specific detection experiments other than microarray analysis. Currently, genetic network are mostly referred to the interactive influence of gene expression by other genes expression due to

the type of data constraint from microarray experiments. With the development of proteome and metabolome techniques, we will finally reveal the cellular network at all the levels including protein and metabolites, and really understand the regulatory mechanisms in a living cell.

Recent study in the topology of metabolic and genetic network has revealed features of biological network both from large-scale system-levels and basic building blocks of "network motif". Uncovering the structure characteristics of network will help us in understanding the fundamental design principals of network and benefit our reconstruction of the network from experimental data.

Several global features for metabolic networks include: (1) 'Small world' -- the networks are both large and sparse, and any two nodes can be connected in very few steps. So the time required to spread information in a small world network is close to the theoretically possible minimum (Fell and Wagner, 2000; Watts and Strogatz, 1998); (2) 'Scale-free' -- the wiring architecture deviates significantly from the random, exponential networks. Scale-free networks are highly inhomogeneous, where the number of connections of the members fell off in a power-law relationship. A small number of members have a large number of connections, and a large number of members have few connections. Scale-free networks are robust to errors: random deletion of individual nodes does not significantly affect information flow in the network. However, they are vulnerable to the removal of a highly connected node (Jeong et al., 2000; Wolf et al., 2002); (3) 'Hierarchical Organization of Modularity' -- the network are organized into many small, highly connected topologic modules that combine in a hierarchical manner into larger, less cohesive units, with their number and degree of clustering following a power law. In the network, many highly integrated small modules group into a few larger modules, which in turn can be integrated into even larger modules. The organization of the networks is likely to combine a capacity for rapid flux reorganization with a dynamic integration with all other cellular function

(Ravasz et al., 2002). Though the large scale organization features are mostly found in metabolic network, it is believed other biological networks such as genetic network or even non-biological networks such as internet and social network shares the same feature. And 'scale-free' topology has been found to exist in genetic networks (Featherstone and Broadie, 2002).

In addition to the global topology, local features of genetic network also received much notice. 'Network motifs' is defined as patterns of interconnections that recur in many different parts of a genetic network at frequencies much higher than those found in randomized networks. Two transcription regulation 'network motifs' have been found to occur much greater than in the random networks: (1) Feed-forward loop: a transcription factor X regulates a second transcription factor Y and both jointly regulates a third gene Z. (2) Bi-fan: an overlapping interaction between two effect genes and two transcription factor. The feed-forward loop occurs where an external signal causes a rapid response of many systems. And Bi-fan allows a group of transcription factors to collaboratively regulate a group of functionally related effectors.

1.5.2 Boolean and Qualitative Network Models for Gene Network

Boolean Network model was first introduced (Kauffman, 1969) as a simple approximation to genetic networks. In this model, gene expression is quantified to two states: either ON or Off. The state of each gene at next time point is determined by Boolean function of its inputs at the current state and every gene synchronously updates its state. The network is deterministic since given identical start state and rules, the same state transition process and final state will be reached. One notable feature of this representation is that the network will have finite number of states: for n genes there will be at most 2^n states. Within 2^n transitions, a repeating state will be found. And deterministic of the updating rule will result in a repeating cycle of states.

Based on Boolean network model, an algorithm for inferring genetic network architectures from state transition tables which correspond to time series of gene expression patterns has been constructed (Liang et al., 1998). Basically, the algorithm uses information theory to establish how given input elements are connected in the network and then determine the functions that specify the output states from input. Mutual information was used to quantify the existence of a constant mapping for a set of input vectors to output. Mutual information measures mutual dependency or the amount of information one object model contains about another. As a result, mutual information can be used to measure mutual agreement between object models. The algorithm proceeds by trying different combination of input elements, which maximizes the mutual information. Another algorithm for learning Boolean network (Akutsu et al., 1999) used exhaustive search for hypothesis in the space of conjunctions, disjunctions and exclusive ORs. The algorithm suggested that if the number of input to each node is bounded by a constant, only O(log n) state transition pairs (from 2n pairs) are necessary and sufficient to identify the Boolean network of n nodes correctly with high probability. By employing combinatorial optimization techniques, Ideker et al.(Ideker et al., 2000) can provide Boolean network models that are consistent with expression data.

Boolean network has the advantages of simplicity and modeling nonlinear relations. However, the idealized assumption about binary representation of gene expression makes it unreliable to most of the conditions. The criterion of "On" and "Off" will affect the final connectivity of network a lot and is hard to define. The approximate representation by binary function of continuous variable will result in loss of information and erroneous inference. The number of regulatory inputs to determine the functions is unknown, which makes the search for functions an intractable job. Furthermore, synchronous may not meet the real situation of nature and the transition between states is rather probabilistic than deterministic. Though probabilistic Boolean network (Shmulevich et al., 2002) which incorporate uncertainty regarding to state transition, and qualitative network (Akutsu et al., 2000) which a number of discrete

value more than two represents gene expression, have been proposed, the fundamental problem about the number of discrete value and how to discretize still exists.

1.5.3 Linear Models for Gene Network

Various linear models were proposed for gene network modeling in which gene expression variables take on real continuous values. The model represents regulatory relationships between genes as linear coefficients or weights, with the 'net' regulation influence on a gene's expression being the mathematical summation of the regulatory inputs. Though several groups of researchers coin different names for their methods, the basic principals are the same. The proposed method names include: additive model (D'Haeseleer et al., 1999), differential equations (Chen et al., 1999), and weight matrix (Weaver et al., 1999). Basically, the model can be written in a linear multiple regression equation, where the output value of a gene is a weighted summation of its input gene expressions:

$$x_i = \sum_j w_{ij} x_j + b_i \quad \text{or} \quad \frac{\partial x_i}{\partial t} = \sum_j w_{ij} x_j(t) + b_i \tag{1.1}$$

where x_i is the expression level of gene i at time t, b_i is a bias term indicate the basal gene expression level without influence form other inputs. w_{ij} indicates the influence of gene j on the regulation of gene i. Weaver et al. (Weaver et al., 1999) derive the model from environmental input of simulated data and find it could predict correct components of the model even in noisy expression data. Chen et al. (Chen et al., 1999) solved the linear model using minimum weight solutions and Fourier transformation to model periodic data in cell cycle. D'Haeseleer uses least square to fit the model and showed it helpful to infer biologically relevant regulatory relationships from real data (D'Haeseleer et al., 2000).

Linear model is a good approximation to nonlinear biological system especially when no information about what functions between input and output should be. However, because of the multicollinearity problem within the data, the inferred network will not be estimated precisely or even in a wrong way. And usually the data contains more variables than experimental observations, which leads to nonunique solutions. Further, even when correct coefficients were inferred, it is hard to tell whether it is due to direct influence or through influence of other intermediate genes.

Recently, variant methods based on linear model have been proposed. A combined singular value decomposition and robust regression method (Yeung et al., 2002) was introduced to overcome the nonunique solution problem due to relative small number of experiments. They first use SVD to construct a family of candidate networks, all being consistent with the experimental data, and then use robust regression to identify the sparsest network in this family as the most likely solution. Though it is relatively correct that the genetics network is sparse, from the study of global and local topology of genetic network and our intuition, we doubt that genetic network should be the sparsest in the candidate solutions. So the network inferred from this method is hard to say to be a realistic one. Another method uses theory of Metabolic Control Analysis to infer the genetic network on the basis of perturbation data from appropriately designed microarray experiments, which can identify the direct regulatory relationship between genes (de la Fuente et al., 2002). This method infers the local relations of genes based on an inverse of the matrix representing global regulation relationships. Genetics network structure and its quantitative regulatory strength will be inferred form the local regulation. However, microarray experiments are full of error and variance, and the inverse of a matrix is subject to the influence of an even small amount of variance.

1.5.4 Nonlinear Models for Genetic Networks

It is believed that the regulatory relationships between genes are intrinsically nonlinear.

Several complex functions have been used with the purpose to mimic the nonlinear properties of the regulation including nonlinear differential equations (Koza et al., 2001; Wahde and Hertz, 2000), neural network (Wahde and Hertz, 2000), and genetic algorithms (Wahde and Hertz, 2001). However, the nonlinear models usually have a relatively complex parameter space and need more data to get the estimates for parameters, and the interpretation of the parameter is vague. Nonlinear network such as neural network has the tendency to over fit the data, which leads to difficulties in extension to other dataset.

1.5.5 Probability Models — Bayesian Network

Both theoretical consideration and experimental results suggest that in reality genetic networks are stochastic (Elowitz et al., 2002; McAdams and Arkin, 1997; Ozbudak et al., 2002; Swain et al., 2002), which means a given gene-expression state can generate more than one successive gene-expression state. And the experimental data are full of noise. Then probabilistic representation of genetic network is needed. A Bayesian network (BN) is a graphical model that encodes probabilistic relationships among variables of interest. Its probabilistic nature is most suitable to describe the expression data with variance and uncertainty; its well-defined statistical foundations and successful applications in other fields imply it a suitable tool to find causal relations within genetic network. Bayesian network has been used in modeling genetic network in several researches: Friedman et al. show that BN can describe interactions between genes, and apply it to uncover biological features from *S. cerevisiae* cell-cycle measurements (Friedman et al., 2000); Hartemink et al. use BN to correctly differentiate between alternative hypotheses of the galactose regulatory network in *S. cerevisia* (Hartemink et al., 2001); Pe'er et al. extend BN to handle perturbations, and identify significant sub-networks of interesting genes from *S. cerevisia* mutant data (Pe'er et al., 2001); Yoo et al. also deal with perturbation data and develop an algorithm to find hidden causal effect (Yoo et al., 2002). However, most of the applications use data

discretization which results in a loss of information; furthermore due to difficulties in search of the large possible network model space, most of the methods are constructing pairwise relationship, features of network or sub-networks rather than learning the whole structure of the network.

1.5.6 "Genetical Genomics"

A term of "Genetical Genomics" was coined in (Jansen and Nap, 2001), which involves expression profiling and marker-based fingerprinting of each individual in a segregating population, and exploits the statistical tools used in genetic (QTL) mapping. In QTL mapping, quantitative trait is usually referred to trait or measurement of ecological or medical interest such as the yield of crop, the milk production of cows, or the disease susceptibility of human. In Genetical Genomics, gene expression is viewed as a quantitative trait that will change continuously according to the status of other genes and environmental stimuli. The variation of its expression reflects the regulation from other genes within a complex genetic network. Gene expression data can then be treated as quantitative trait and mapped onto genetic map in a segregating population. By appropriate analysis, gene expression can be resolved into the contributing loci on the genetic map. The genes at the contributing loci will be identified as the candidates that have the influence on the transcription of the gene studied. Correspondingly, a list of genes from different QTL as causal effects for the expression of the gene studied can be obtained. Combining the list together will enable us to reveal the relationship between genes and it is possible to construct a regulatory network among them.

Genetical Genomics is likely to become instrumental for unravelling metabolic, regulatory and developmental pathways. Although clear in concept, combinatory method to construct such network and the search for optimal structure are needed to further develop this concept. Intensive simulations and real data analysis are needed to establish the type of population, the kind of map and to test the effects of its

implementation.

1.6 Statistical Methods in QTL mapping

There are several issues to be solved before we can do genetic mapping: What is the type of the trait, whether it is qualitative or quantitative? What is the mapping population, whether experimental inbreeding or outbreed pedigree? What kind of molecular marker available? Is there a dense or sparse genetic map? And finally which method should we use to do the genetic mapping.

1.6.1 Qualitative vs. Quantitative Trait

Genetic mapping of trait-causing genes dates back to the work of Sturtevant in 1913 (Sturtevant, 1913). Then it became the main tool for geneticist in discovering the genetic cause of a trait by comparison the inheritance pattern of a trait with the inheritance pattern of chromosomal regions. The traits can usually be classified into qualitative discrete such as flower color and quantitative continuous such as individual's height. Discrete trait follows exactly Medelian rules and discovering the association of genetic variants with the trait is relative easy compared to quantitative trait where usually more than one gene and their interaction with environment will be responsible for the trait. Specific regions in the genome underlying quantitative or complex traits are commonly called quantitative trait loci (hereafter QTL). The methods that reveal the association of QTL to quantitative traits are called QTL mapping.

1.6.2 Inbreed Lines

Both QTL mapping in plants and animals and complex disease mapping in human population have been studied extensively. The principles behind them are the same other than the population structures. Complex disease gene mapping depends on linkage analysis using pedigree information or other association methods. QTL mapping in

plants is relative easy since the ease of cross and maintenance of a certain population. All of the mapping population are crossed from two highly inbreed parent lines, differing both in trait values and in the marker variants they carry. The resulting F1 lines will be heterogeneous at all markers and QTLs that differ in two parent lines. From the F1 population, crosses are made to get F2 and F3 progeny, recombinant inbreed line (RILs), doubled-haploid lines (DHLs), backcross (BC) progeny, or near isogenic lines (NILs). The final mapping population will have variation in both the trait of interest and the underlying quantitative trait loci and marker genotypes. Then QTL mapping will be executed on this population to find the correlation between the trait in question and marker genes that has been marked on the map previously. I will use RILs to illustrate the cross process and its advantage in mapping QTLs.

A recombinant inbred line comes from the cross of two wildly separated lines for the phenotype, (also different in the genotype in most loci). After the first cross, F1 lines are self crossed in plant or sib-crossed in animal continuously (Burr and Burr, 1991). This process continues usually more than 6 generations. For each cross the genetic heterogeneity will reduce by half. After seven or more crosses, we will get many homogeneous lines with different combination of the markers (QTLs). Each RIL represents different multilocus genotypes. The advantage is that most lines are homogeneous in genetic marker and have no within-line genetic variance, whereas the genetic variance between lines is considerable. Furthermore, these more or less permanent populations permit many geneticists to contribute to the mapping effort and to profit from each other's work.

1.6.3 Genetic Marker and Genetic Map

It was in the early 1980s that people first found that naturally occurred DNA sequence variation can be used to dissect quantitative and complex traits (Botstein et al., 1980). Since then, genetic markers have sparked an explosion of genetic maps in humans and

economically important plants and animals. Genetic markers are specific aspects of DNA with an identifiable physical location on a chromosome and whose inheritance can be followed. Since the first discovery of RFLP (restriction fragment length polymorphism), the genetic marker family has been evolved to include microsatellite, RADP, VNTR, AFLP, SSR and SNP et al. Genetic maps are the ordering of the genetic markers along a chromosome and the relative distance between them. Though markers are mostly non-functional or selectively neutral sites, its tight linkage to functional genes or QTLs provides structure for QTL mapping. Because DNA segments that lie near each other on a chromosome tend to be inherited together, markers are often used as indirect ways of tracking the inheritance pattern of a gene of interest. Nowadays, more and more markers are available especially SNPs. It is possible that every gene will be mapped exactly on the map. Our work will assume a tight or ultra-tight map available where each gene will link to at least one genetic marker.

1.6.4 Methods for QTL Mapping

The task of mapping QTL is to detect and estimate the association between the variation at the phenotypical level (trait data) and the variation at genetic level (marker data) in terms of number, positions, effects and interaction of QTL. I will review methods for QTL mapping especially in experimental animals or plants, since this proposal will mostly apply to inbreed plants.

Single Marker Analysis

A simple method for QTL mapping is the single marker test such as simple regression or analysis of variance (ANOVA) at the marker loci which assess the segregation of a phenotype with respect to a marker genotype (Soller and Genizi, 1978). This method has been utilized with various experimental designs such as backcross and intercross designs. The presence of a QTL is determined by F-statistic (or corresponding t-statistics), and this significance test is equivalent to that using LOD score that is

logarithm of the likelihood favoring linkage. The test can indicate which markers are associated with the quantitative trait of interest, and, therefore, point to the existence of potential QTL. However, this test can't separate out the QTL location and QTL effects. It can only detect QTL effect but can't locate its position. The relative distance, or the recombination frequency between markers and the QTL cannot be revealed.

Interval Mapping

Simple marker analysis is extended to interval mapping where both position and effect of a QTL can be inferred (Lander and Botstein, 1989). A previous defined genetic map is used as the framework for the location of QTL. The intervals that are defined by ordered pairs of markers are searched incrementally. The probability of the QTL genotypes is calculated from their flanking markers genotype and their distance. The likelihood of the QTL in the region are calculated and tested against the null hypothesis that there is no QTL in this region. The most likely significant regions (exceed some kind of criterion based on either permutation or some empirical such as LOD > 3) are assigned as QTLs. The advantages of interval mapping are: both position and effect of QTL can be inferred, and the QTL can be searched between the two markers interval. However, some problems exist with this method. First, QTL effect in the nearby region may lead to false positive in the tested region. Second, multiple QTL in the same region can't be separated. And other marker information outside the flanked marker is not used.

Composite Interval Mapping

An extension of interval mapping called composite interval mapping was proposed (Jansen, 1993; Zeng, 1993; Zeng, 1994), which involves regression both on QTL within an interval and on marker loci outside that interval as cofactors to control the genetic variation of other possibly linked or unlinked QTL. This method can be viewed as a combination of the interval mapping with multiple regression and the inference is made

by maximum likelihood from an existing genetic map. The model allocate out the effect of other QTLs by including other marker information when dissect the region of interest. One disadvantage of CIM is that it can't detect epistasis the interactions of QTLs. And the test result is affected by the uneven distribution of the markers in the genome.

Multiple QTL models

Like composite interval mapping, multiple QTL models or marker-QTL-marker (MQM) (Jansen, 1993; Jansen and Stam, 1994) combines the utility of interval mapping with the strength of multiple regression. MQM fit multiple QTL and their epistasis together. This involves the test of the number of QTLs, the position of QTLs, the effect of QTLs and the epistatis of QTLs simultaneously. The purpose of MQM is to include the most important cofactors into the regression model. This is a kind of model selection and model fitting problem. Different criterion can be applied such as R-square, likelihood, BIC, AIC, or Bayesian score for the model selection. Extensive simulation shows the power of MQM mapping over interval mapping through the control of Type I and Type II error (Jansen, 1994).

Bayesian QTL mapping

During mapping QTL, there are much uncertainty and complexity involved: the number of QTL is unknown; polygenic background and gene interactions may all act on the trait under consideration; genotypes of QTL can't be observed directly but need to be inferred from observed marker information. Bayesian analysis has the ability to incorporate the unknown and known quantity together to form the joint posterior probability, and then the unknown quantity including QTL effect and position can be inferred based on samples of all unknowns obtained from the joint posterior distribution via Markov chain Monte Carlo (MCMC) algorithm. Details on current implementation of the Bayesian QTL analysis can be found in several contributions (Heath, 1997;

Hoeschele and Vanraden, 1993; Satagopan et al., 1996; Sillanpaa and Arjas, 1998; Sillanpaa and Arjas, 1999; Thomas et al., 1997; Uimari and Hoeschele, 1997).

Reference

Akutsu, T., Miyano, S., and Kuhara, S. (1999). Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. Pac Symp Biocomput, 17-28.

Akutsu, T., Miyano, S., and Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics *16*, 727-734.

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X.*, et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503-511.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A *96*, 6745-6750.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci U S A *97*, 10101-10106.

Baldi, P., and Long, A. D. (2001). A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. Bioinformatics *17*, 509-519.

Baltzer, F., and Boveri, T. (1964). Science *144*, 809-815.

Bittner, M., Meltzer, P., Chen, Y., Jiang, Y., Seftor, E., Hendrix, M., Radmacher, M., Simon, R., Yakhini, Z., Ben-Dor, A.*, et al.* (2000). Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature *406*, 536-540.

Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet *32*, 314-331.

Burr, B., and Burr, F. A. (1991). Recombinant inbreds for molecular mapping in maize: theoretical and practical considerations. Trends Genet *7*, 55-60.

Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. (2001). Remodeling of yeast

genome expression in response to environmental changes. Mol Biol Cell *12*, 323-337.

Chen, T., He, H. L., and Church, G. M. (1999). Modeling gene expression with differential equations. Pac Symp Biocomput, 29-40.

Chen, Y., E.R. Dougherty, and M.L. Bittner. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics *2*, 364-374.

Cui, X., Kerr, M. K., and G.A., C. (2003). Data Transformation for cDNA Microarray Data. submitted.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C.-H., Minokawa, T., Amore, G., Hinman, V., Arenas-Mena, C.*, et al.* (2002). A Genomic Regulatory Network for Development. Science *295*, 1669-1678.

de la Fuente, A., Brazhnik, P., and Mendes, P. (2002). Linking the genes: inferring quantitative gene networks from microarray data. Trends Genet *18*, 395-398.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science *278*, 680-686.

D'Haeseleer, P., Liang, S., and Somogyi, R. (2000). Genetic network inference: from co-expression clustering to reverse engineering. Bioinformatics *16*, 707-726.

D'Haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. Pac Symp Biocomput, 41-52.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association *97*, 77--87.

Dudoit, S., Yang, Y. H, Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica *12*, 111-140.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. Nat Genet *21*, 10-14.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*,

14863-14868.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. Science *297*, 1183-1186.

Featherstone, D., and Broadie, K. (2002). Wrestling with pleiotropy: genomic and topological analysis of the yeast gene expression network. Bioessays *24*, 267-274.

Fell, D. A., and Wagner, A. (2000). The small world of metabolism. Nat Biotechnol *18*, 1121-1122.

Fellenberg, K., Hauser, N. C., Brors, B., Neutzner, A., Hoheisel, J. D., and Vingron, M. (2001). Correspondence analysis applied to microarray data. Proc Natl Acad Sci U S A *98*, 10781-10786.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J Comput Biol *7*, 601-620.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics *16*, 906-914.

Garner, C. P., Tatu, T., Best, S., Creary, L., and Thein, S. L. (2002). Evidence of genetic interaction between the beta-globin complex and chromosome 8q in the expression of fetal hemoglobin. Am J Hum Genet *70*, 793-799.

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., and Brown, P. O. (2000). Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell *11*, 4241-4257.

Ghosh, D., and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. Bioinformatics *18*, 275-286.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A*., et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Gross, C., Kelleher, M., Iyer, V. R., Brown, P. O., and Winge, D. R. (2000). Identification of the copper regulon in Saccharomyces cerevisiae by DNA microarrays. J Biol Chem *275*, 32310-32316.

Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Using

graphical models and genomic expression data to statistically validate models of genetic regulatory networks. Pac Symp Biocomput, 422-433.

Heath, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet *61*, 748-760.

Herrero, J., Valencia, A., and Dopazo, J. (2001). A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics *17*, 126-136.

Hoeschele, I., and Vanraden, P. (1993). Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor Appl Genet *85*, 946-952.

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., and Banavar, J. R. (2001). Dynamic modeling of gene expression data. Proc Natl Acad Sci U S A *98*, 1693-1698.

Huber, W., Von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002). Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics *18*, S96-S104.

Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D*., et al.* (2000). Functional discovery via a compendium of expression profiles. Cell *102*, 109-126.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science *292*, 929-934.

Ideker, T. E., Thorsson, V., and Karp, R. M. (2000). Discovery of regulatory interactions through perturbation: inference and experimental design. Pac Symp Biocomput, 305-316.

IHGSC (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M., and Brown, P. O. (2001). Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature *409*, 533-538.

Jackson, D. A., Symons, R. H., and Berg, P. (1972). Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA

molecules containing lambda phage genes and the galactose operon of Escherichia coli. Proc Natl Acad Sci U S A *69*, 2904-2909.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. Genetics *135*, 205-211.

Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. Genetics *138*, 871-881.

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

Jansen, R. C., and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. Genetics *136*, 1447-1455.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000). The large-scale organization of metabolic networks. Nature *407*, 651-654.

Jorgensen, P., Nishikawa, J. L., Breitkreutz, B.-J., and Tyers, M. (2002). Systematic Identification of Pathways That Couple Cell Growth and Division in Yeast. Science *297*, 395-400.

Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M. G., and Alon, U. (2001). Ordering Genes in a Flagella Pathway by Analysis of Expression Kinetics from Living Bacteria. Science *292*, 2080-2083.

Kauffman, S. A. (1969). Metabolic stability and epigenesis in randomly constructed genetic nets. J Theor Biol *22*, 437-467.

Kerr, M. K., Afshari, C. A., Bennett, L., Bushel, P., Martinez, J., Walker, N. J., and Churchill, G. A. (2002). Statistical analysis of a gene expression microarray experiment with replication. Statistica Sinica *12*, 203-217.

Kerr, M. K., and Churchill, G. A. (2001a). Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A *98*, 8961-8965.

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

Kerr, M. K., and Churchill, G. A. (2001c). Statistical design and the analysis of gene expression microarray data. Genet Res *77*, 123-128.

Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med *7*, 673-679.

Kim, S. K., Lund, J., Kiraly, M., Duke, K., Jiang, M., Stuart, J. M., Eizinger, A., Wylie, B. N., and Davidson, G. S. (2001). A gene expression map for Caenorhabditis elegans. Science *293*, 2087-2092.

Koza, J. R., Mydlowec, W., Lanza, G., Yu, J., and Keane, M. A. (2001). Reverse engineering of metabolic pathways from observed data using genetic programming. Pac Symp Biocomput, 434-445.

Lander, E. S., and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics *121*, 185-199.

Lee, M.-L. T., Kuo, F. C., Whitmore, G. A., and Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. PNAS *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. T., Thompson, C. M., Simon, I.*, et al.* (2002). Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Liang, S., Fuhrman, S., and Somogyi, R. (1998). Reveal, a general reverse engineering algorithm for inference of genetic network architectures. Pac Symp Biocomput, 18-29.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol *14*, 1675-1680.

Lönnstedt, I., and Speed, T. (2002). Replicated Microarray Data. Statistical Sinica *12*, 31-46.

Lyons, T. J., Gasch, A. P., Gaither, L. A., Botstein, D., Brown, P. O., and Eide, D. J. (2000). Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. Proc Natl Acad Sci U S A *97*, 7957-7962.

McAdams, H. H., and Arkin, A. (1997). Stochastic mechanisms in gene expression.

Proc Natl Acad Sci U S A *94*, 814-819.

McDonald, M. J., and Rosbash, M. (2001). Microarray analysis and organization of circadian gene expression in Drosophila. Cell *107*, 567-578.

McKusick, V. A. (1960). Walter S. Sutton and the physical basis of Mendelism. Bull Hist Med *34*.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. Bioinformatics *18*, 413-422.

Medvedovic, M., and Sivaganesan, S. (2002). Bayesian infinite mixture model based clustering of gene expression profiles. Bioinformatics *18*, 1194-1206.

Mendel, G. (1865). Versuche über Pflanzen-Hybriden. [Experiments in Plant Hybridisation.], Verhandlungen des naturforschenden Vereines in Brünn [Proceedings of the Natural History Society of Brünn]).

Myers, R. H. (1990). Classical and modern regression with applications, 2 edn, Boston, PWS-KENT).

Newton, M. A., Kendziorski, C. M., Richmond, C. S., Blattner, F. R., and Tsui, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol *8*, 37-52.

Nguyen, D. V., and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics *18*, 39-50.

Ogawa, N., DeRisi, J., and Brown, P. O. (2000). New components of a system for phosphate accumulation and polyphosphate metabolism in Saccharomyces cerevisiae revealed by genomic expression analysis. Mol Biol Cell *11*, 4309-4321.

Ooi, S. L., Shoemaker, D. D., and Boeke, J. D. (2001). A DNA microarray-based genetic screen for nonhomologous end-joining mutants in Saccharomyces cerevisiae. Science *294*, 2552-2556.

Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002). Regulation of noise in the expression of a single gene. Nat Genet *31*, 69-73.

Pe'er, D., Regev, A., Elidan, G., and Friedman, N. (2001). Inferring subnetworks from perturbed expression profiles. Bioinformatics *17*, S215-224.

Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C.*, et al.* (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature *415*, 436-442.

Quackenbush, J. (2001). Computational analysis of microarray data. Nat Rev Genet *2*, 418-427.

Quackenbush, J. (2002). Microarray data normalization and transformation. Nat Genet *32*, 496-501.

Ramoni, M. F., Sebastiani, P., and Kohane, I. S. (2002). Cluster analysis of gene expression dynamics. Proc Natl Acad Sci U S A *99*, 9121-9126.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabasi, A. L. (2002). Hierarchical organization of modularity in metabolic networks. Science *297*, 1551-1555.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E.*, et al.* (2000). Genome-wide location and function of DNA binding proteins. Science *290*, 2306-2309.

Rocke, D. M., and Durbin, B. (2001). A model for measurement error for gene expression arrays. J Comput Biol *8*, 557-569.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A *74*, 5463-5467.

Satagopan, J. M., Yandell, B. S., Newton, M. A., and Osborn, T. C. (1996). A bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics *144*, 805-816.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O., and Davis, R. W. (1996). Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci U S A *93*, 10614-10619.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S.*, et al.* (2002). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised

machine learning. Nat Med *8*, 68-74.

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002). Probabilistic Boolean Networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics *18*, 261-274.

Shoemaker, D. D., Schadt, E. E., Armour, C. D., He, Y. D., Garrett-Engele, P., McDonagh, P. D., Loerch, P. M., Leonardson, A., Lum, P. Y., Cavet, G.*, et al.* (2001). Experimental annotation of the human genome using microarray technology. Nature *409*, 922-927.

Sillanpaa, M. J., and Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics *148*, 1373-1388.

Sillanpaa, M. J., and Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics *151*, 1605-1619.

Simon, I., Barnett, J., Hannett, N., Harbison, C. T., Rinaldi, N. J., Volkert, T. L., Wyrick, J. J., Zeitlinger, J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001). Serial regulation of transcriptional regulators in the yeast cell cycle. Cell *106*, 697-708.

Soller, M., and Genizi, A. (1978). The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait loci in segregating populations. Biometrics *34*, 47-55.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell *9*, 3273-3297.

Sturtevant, A. H. (1913). The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. J of Exp Zool *14*, 43-59.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. Proc Natl Acad Sci U S A *99*, 12795-12800.

Takahashi, M., Rhodes, D. R., Furge, K. A., Kanayama, H., Kagawa, S., Haab, B. B., and Teh, B. T. (2001). Gene expression profiling of clear cell renal cell carcinoma: gene identification and prognostic classification. Proc Natl Acad Sci U S A *98*, 9754-9759.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A *96*, 2907-2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281-285.

Theilhaber, J., Bushnell, S., Jackson, A., and Fuchs, R. (2001). Bayesian estimation of fold-changes in the analysis of gene expression: the PFOLD algorithm. J Comput Biol *8*, 585-614.

Thomas, D. C., Richardson, S., Gauderman, J., and Pitkaniemi, J. (1997). A Bayesian approach to multipoint mapping in nuclear families. Genet Epidemiol *14*, 903-908.

Uimari, P., and Hoeschele, I. (1997). Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics *146*, 735-743.

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T.*, et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. Nature *415*, 530-536.

Venter, J. C., Adams, Mark D., Myers, Eugene W., Li, Peter W., Mural, Richard J., Sutton, Granger G., Smith, Hamilton O., Yandell, Mark, Evans, Cheryl A., Holt, Robert A., Gocayne, Jeannine D., Amanatides, Peter, Ballew, Richard M., Huson, Daniel H., Wortman, Jennifer Russo, Zhang, Qing, Kodira, Chinnappa D., Zheng, Xiangqun H., Chen, Lin, et. al. (2001). The Sequence of the Human Genome. Science *291*, 1304-1351.

Wahde, M., and Hertz, J. (2000). Coarse-grained reverse engineering of genetic regulatory networks. Biosystems *55*, 129-136.

Wahde, M., and Hertz, J. (2001). Modeling genetic regulatory dynamics in neural development. J Comput Biol *8*, 429-442.

Wall, M. E., Dyck, P. A., and Brettin, T. S. (2001). SVDMAN--singular value decomposition analysis of microarray data. Bioinformatics *17*, 566-568.

Watson, J., and Crick, F. (1953). A structure for deoxyribose nucleic acid. Nature *171*, 737-738.

Watts, D. J., and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature *393*, 440-442.

Wayne, M. L., and McIntyre, L. M. (2002). Combining mapping and arraying: An approach to candidate gene identification. Proc Natl Acad Sci U S A *1*, 1.

Weaver, D. C., Workman, C. T., and Stormo, G. D. (1999). Modeling regulatory networks with weight matrices. Pac Symp Biocomput, 112-123.

Welsh, J. B., Zarrinkar, P. P., Sapinoso, L. M., Kern, S. G., Behling, C. A., Monk, B. J., Lockhart, D. J., Burger, R. A., and Hampton, G. M. (2001). Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. Proc Natl Acad Sci U S A *98*, 1176-1181.

White, K. P., Rifkin, S. A., Hurban, P., and Hogness, D. S. (1999). Microarray analysis of Drosophila development during metamorphosis. Science *286*, 2179-2184.

Wolf, Y. I., Karev, G., and Koonin, E. V. (2002). Scale-free networks in biology: new insights into the fundamentals of evolution? Bioessays *24*, 105-109.

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 models. J Comput Biol *8*, 625-637.

Wuju, L., and Momiao, X. (2002). Tclass: tumor classification system based on gene expression profile. Bioinformatics *18*, 325-326.

Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucl Acids Res *30*, e15-.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics *17*, 977-987.

Yeung, K. Y., and Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. Bioinformatics *17*, 763-774.

Yeung, M. K., Tegner, J., and Collins, J. J. (2002). Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci U S A *99*, 6163-6168.

Yoo, C., Thorsson, V., and Cooper, G. F. (2002). Discovery of causal relationships in a gene-regulation pathway from a mixture of experimental and observational DNA microarray data. Pac Symp Biocomput, 498-509.

Zeng, Z. B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci U S A *90*, 10972-10976.

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

Zhang, H., Yu, C. Y., Singer, B., and Xiong, M. (2001). Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci U S A *98*, 6730-6735.

Chapter 2

A Mixture Model Approach to Classifying Uncertain Labeled Sample from Gene Expression Data

Abstract

Here, we discuss how current sample classification methods for microarray gene expression data can be extended to account for uncertainty in class labels. The method is motivated by a microarray study to predict the development competence of in vitro constructed embryos, where uncertain development competence prior probabilities associated with each embryo. A finite mixture model approach is proposed to perform this classification. Posterior probabilities of development competence learned from mixture model, which combined individual embryo's expression profile and the development competence prior probabilities, were used for classification. Simulations show that this method performs better than hierarchical clustering and linear discriminant analysis under uncertain class conditions. Applying this method to real expression data of in vitro constructed embryos suggested which embryos were viable and which were not. This study demonstrates that the mixture model approach is a powerful method for classifying uncertain labeled biological samples from gene expression data.

2.1 Introduction

Large-scale measurements of gene expressions generated by microarray experiments (Lockhart et al., 1996; Schena et al., 1995) have the potential to reveal the molecular distinctions of biological samples and classes (Alizadeh et al., 2000; Golub et al., 1999). Two kinds of methods have been proposed to group biological samples from gene expression data based on whether the sample class information is available or not. If the sample class labels are clearly known, many statistical and supervised machine-learning methods have been applied, such as linear and quadratic discriminant analysis, logistic regression, nearest neighbor, recursive partitioning trees (Dudoit et al., 2002), and support vector machines (Furey et al., 2000). If the sample class labels are unknown, many statistical exploratory and unsupervised machine-learning methods have been used, such as hierarchical clustering analysis (Eisen et al., 1998), self organization map (Tamayo et al., 1999; Tavazoie et al., 1999) and k-means clustering (Tavazoie et al., 1999). There is not much study on one specific situation when the sample class information is uncertain. This happens when some probabilities are associated with sample classes but they are not definite. Under such condition, supervised learning can't construct the classification model directly because no clear class information is available. Unsupervised learning is not suitable either, because these methods group samples solely based on the expression similarities, and the partial class information will be totally ignored. Here, we discuss a method for sample classification under such uncertain labeled condition. The method is motivated by and applied to a gene expression experiment for in vitro constructed embryo development competence classifications. The general principals and methods are applicable to other uncertain sample labeled conditions.

Live birth of cloned animals with correct genetic enhancements could greatly benefit human medicine by providing pharmaceutical proteins and human organ transplant products. The live birth of cloned animal requires successful development of the in vitro

constructed embryos. However, the molecular mechanisms underlying the embryo development process are unclear, and to date no mathematical models can reliably predict the development competence of the constructed embryo. The success of using expression profiles in cancer classification indicates that gene expression can reveal molecular distinctions of biological samples (Alizadeh et al., 2000; Beer et al., 2002; Golub et al., 1999; Ramaswamy et al., 2001; van de Vijver et al., 2002; West et al., 2001). It is possible that gene expression from in vitro constructed embryo can be used for embryo development competence classification. Based on this assumption, microarray experiments are conducted to measure gene expressions of in vitro constructed embryos.

Two groups of embryos are used in this study. One group of embryo is constructed from Nuclei Transfer (NT) method. Another group of embryo is constructed from In Vitro Fertilization (IVF) method. Only 2 percent of animals cloned from NT constructed embryos survive to birth (Colman, 2000). IVF embryos have a relative higher successful survival rate at 50 percent (Colman, 2000). Though the definite live birth outcome of each embryo is unknown, the different live birth rates can give partial development competence information. The purpose of this study is to develop a method that can classify NT and IVF embryos as development competent (DC) or incompetent (DI) using gene expression profiles. The cDNA microarray measured gene expression data are normalized via linear mixed model (LMM) first. A modified LR test based on finite mixture model (FMMA) was applied to select genes that are supposed to be differentially expressed across DC and DI. Principal component analysis (PCA) was used to reduce the data dimensions before the final FMMA for classification.

2.2 Data and Methods

2.2.1 Microarray Experiment Design and Linear Mixed

Model for Data Processing

The cDNA microarray experiment is designed as an interwoven loop design and analyzed as a linear mixed model (LMM). Ten embryos were collected from each group (NT or IVF) separately. An NT is paired with an IVF embryo directly in one microarray slide. There are a total of 100 microarray slides with 2640 genes spotted twice on each of these 100 arrays. The following linear mixed model was used for the analysis:

$$y_{ijklgm}=\mu+A_i+D_j+A_i\times D_j+G_g+A_i\times G_g+T_k+T_k\times G_g+E_l(T_k)+E_l(T_k)\times G_g+e_{ijklgm}$$

where y_{ijklgm} was the log2 transformed measured intensity from array *i*, dye *j*, type *k*, embryo *l*, gene *g*, and replication *m*. The terms in the linear model are either the main effects of array (*A*), dye (*D*), gene (*G*), treatment (*T*, NT or IVF) or their interactions. Embryo effects (*E*) are nested within the main effects of type. Here the *A, A×D, A×G, E(V), E(V)×G* and *e* are assumed random effects. Other terms are assumed fixed effects. The observed data is fitted into the proposed linear mixed model, and the variance components are estimated via restricted maximum likelihood (Harville, 1977). Fixed effects are estimated and random effects are predicted from the model. This mixed linear model could detect the experimental source of variance such as the array, the dye effects and their interactions, which do not represent the biological difference between embryos. The individual gene main effect represents a specific gene effect across all samples, which will not contribute to the expression difference between embryos. Construction of classification model should be based on normalized data that experimental sources of variation are removed or minimized from such model rather than use the raw data.

2.2.2 Finite Mixture Analysis

Direct test of the gene expression difference across treatment via the linear mixed model as described above is not the most powerful method for identifying those genes,

which are differentially expressed between developmentally competent (DC) and incompetent (DI) embryos. This is so because LMM compares IVF and NT embryos rather than DC and DI embryos, where the IVF and NT classes are both mixtures of DC and DI embryos with different mixing proportions (the proportion of DC embryos is .50 among IVF embryos while it is .02 among NT embryos). A finite mixture model (McLachlan and Peel, 2000; Titterington *et al*., 1985) assumes that an observation comes from a distribution that is a mixture of several distributions, *e.g.*, two normal distributions with different means and possibly also different variances, *i.e*. the observation comes from the first normal distribution $N(\mu_1,\sigma^2_1)$ with mixing probability π and from the other normal distribution $N(\mu_2,\sigma^2_2)$ with probability $1 - \pi$. In this study, the expression value for a particular gene has this mixture distribution with probability $\pi_1 = .50$ ($\pi_2 = .02$) for an embryo of type IVF (NT). The entire expression vector of an embryo (containing the expression values for all genes) follows a mixture of two multivariate normal distributions with the same mixing probabilities as above.

Several recent studies have used mixture models in cluster analysis of gene expression data (Ghosh and Chinnaiyan, 2002; McLachlan *et al*., 2002; Yeung *et al*., 2001), implemented in an unsupervised manner to estimate the number of classes, the parameters of the class densities, and the mixing probabilities. Here we assume two classes (DC, DI) with known but different mixing probabilities for IVF and NT embryos. We identify genes, which discriminate between DC and DI, and we perform embryo classification as DC or DI using Principal Components Analysis (PCA) based dimension reduction, based on finite mixture model analysis (FMMA) as described below.

Data for FMMA

Normalized expression data as input for FMMA were obtained from the LMM as the estimates of the linear predictors $X_{gkl} = T_k + T_k \times G_g + E_l(T_k) + E_l(T_k) \times G_g$. The input data

are then represented by the 20×2640 matrix *X*, where the rows refer to the embryos and the columns to the genes.

Mixture Distribution Inference

Let ϕ_1 (ϕ_2) denote the probability density of the distribution of the expression vectors of DC (DI) embryos. Then, the expression vector (x) of embryo *k* of type IVF (*c = 1*) or NT (*c = 2*) has a mixture distribution with probability density

$$f(\mathbf{x}_{ck}) = \pi_c \phi_1(\mathbf{x}_{ck}; \theta_1) + (1 - \pi_c)\phi_2(\mathbf{x}_{ck}; \theta_2) \qquad (2.1)$$

Assuming that observations on different embryos are independent, the log likelihood function for parameter vector θ containing the parameters of ϕ_1 (θ_1) and ϕ_2 (θ_2) is

$$\log L(\theta) = \sum_{c=1}^{2} \sum_{k=1}^{n_c} \log f(\mathbf{x}_{ck}; \theta) = \sum_{c=1}^{2} \sum_{k=1}^{n_c} \log \left[\pi_c \phi_1(\mathbf{x}_{ck}; \theta_1) + (1 - \pi_c)\phi_2(\mathbf{x}_{ck}; \theta_2) \right] \quad (2.2)$$

Maximum Likelihood (ML) estimates of the parameters in θ were obtained by using an Expectation-Maximization (EM) (Dempster et al., 1977) algorithm (see Appendix). The two component distributions were assumed to be multivariate normal with different mean vectors and common covariance matrix.

Once the parameters have been estimated, an embryo can be classified as DC or DI based on its posterior probability of being DC (see Appendix), which combines the embryo's expression data with the prior probability of DC for the embryo's type (c):

$$P_c(DC \mid \mathbf{x}_{ck}, \pi_c) = \frac{\pi_c \phi_1(\mathbf{x}_{ck})}{\pi_c \phi_1(\mathbf{x}_{ck}) + (1 - \pi_c)\phi_2(\mathbf{x}_{ck})} \qquad (2.3)$$

The posterior probability allows modification of the prior probability through the expression data. For example, an NT (c = 2) embryo has a very low prior probability of

being DC (π_2 = .02), however, if $\phi_1(x_{ck})$ is much larger than $\phi_2(x_{ck})$, then this embryo's posterior probability of DC may exceed 0.5.

Relevant Gene Selection by Modified Likelihood Ratio Test and Permutation

For identification of those genes, which are differentially expressed between the unobserved classes of DC and DI, a likelihood ratio (LR) test was used. The LR test is constructed according two hypothesis: the null hypothesis assumed a univariate normal distribution for the expression data of a gene not differentially expressed, and the alternative hypothesis of a mixture of two univariate normal distributions (with different means) for the expression data of a gene differentially expressed between DC and DI. To avoid large likelihood ratio values resulting from an extremely small variance of expression for some genes (a problem previously observed for t-statistics, see Tusher et al. (2001)), a small positive constant $S(0)$ was added to every estimated variance when calculating the LR, analogous to the modification of the two sample t-statistic by Tusher et al. (2001). The value of $S(0)$ was obtained by minimizing the coefficient of variation of the modified LR as suggested by (Tusher et al., 2001). The modified LR for gene i was then $LR_i = -2ln[L(\mu_i, \sigma^2_i + S(0)) / L(\mu_{i1}, \mu_{i2}, \sigma^2_i + S(0))]$ for the hypotheses H_0: $\mu_{i1} = \mu_{i2} = \mu_i$ and H_1: $\mu_{i1} \neq \mu_{i2}$, where μ_{i1}, μ_{i2}, and σ^2_i are evaluated at their Maximum Likelihood estimates under the respective hypothesis.

To determine the modified LR test threshold, permutations of the original data are performed. Labels of each embryo as NT or IVF are randomly permutated. For each permuted data, modified LR statistic is calculated for each gene as described above. After 1000 repeat, modified LR statistics for all genes in all permutations are ranked together. The 99%, 99.5%, and 99.9% largest test statistics are recorded as the threshold value for type I error rate of 0.01, 0.005 and 0.001 respectively. Then genes, whose test statistics are larger than the threshold, are selected as statistically significant genes

differentiate expressed across DC and DI.

Dimension Reduction by Principal Component Analysis

After gene selection, the number of retained genes may still exceed the number of observations (unless an extremely conservative multiple testing adjustment, Bonferroni, is applied). Therefore, we used principal component analysis (PCA) as a dimensional reduction method to project the expression matrix of the selected genes into a few principal component spaces. Principal component (PC) variables are linear combinations of the original variables according to k eigenvectors, where k is smaller than the number of observations (samples). If the original gene expression values have a mixture distribution with multivariate normal component distributions, then it is easy to show that the new PC variables also have a mixture distribution with multivariate normal components (see Appendix). The PC variables then replace the expression data in (2.1), (2.2) and (2.3), so embryo classification is based on the PC rather than the original variables.

2.3 Results

2.3.1 Finite Mixture Model Analysis of Simulated Data

Prior to application of the FMMA to the real data, we tested this method with simulated data. For a total of 10 genes, 8 were not differentially expressed, and their values were simulated from $N(0,\sigma^2_i)$ with $1/\sigma^2_i \sim Gamma(2,1)$ for genes $i=1,...,8$. For genes 1 and 2, expression values were simulated from $N(\mu_{ic},\sigma^2_i)$ with $i=1,2,\ c=1(2)$ for DC(DI), $\mu_{i=1,c=1} \sim N(3,1)$, $\mu_{i=1,c=2} \sim N(0,1)$, $\mu_{i=2,c=1} \sim N(1,1)$, $\mu_{i=2,c=2} \sim N(0,1)$, and $1/\sigma^2_i \sim Gamma(2,1)$. A total of 200 samples (embryos) were generated with samples 1 to 100 being IVF and samples 101 to 200 NT, and with samples 1 to 50 and 199 to 200 being DC. A total of 100 data sets each consisting of 200 samples were simulated. On these data sets, FMMA was compared with the following methods: Linear Discrimination

Analysis (LDA) using the DC/DI labels known from the simulation as the gold standard; LDA trained on the IVF/NT labels, assuming that we don't know the DC/DI labels but use IVF/NT as the substitute label instead; cluster analysis (Eisen et al., 1998) ignoring the IVF/NT labels with link functions complete, single, average and centroid. Number of misclassifications was obtained as the number of false negatives (FN; DC sample classified as DI) plus the number of false positives (FP; DI sample classified as DC).

Average numbers of misclassification and standard deviations over the 100 simulated data sets are in Table 2.1. The gold standard had an average of 5.73 misclassifications in 200 samples. This number is 45.96, if we use IVF/NT labels as substitution to DC/DI labels in LDA learning, but compare the classification results to the true DC/DI labels. Cluster analysis performed slightly to considerably worse, depending on link function. When the difference in mean expression is not very pronounced relative to the standard deviation within DC/DI classes, then cluster analysis tends to assign a large number of samples to one group with some outliers in another group, hence there are large numbers of false positives and few false negatives. For FMMA, average FP decreases with increasing posterior probability threshold, while FN and total number of misclassifications increase. FMMA classification based on the first two genes performed somewhat better than that based on all ten genes. The threshold can be used to balance the FP and FN errors. If overall misclassification is of concern, a posterior probability threshold of 0.5 may be a good choice. If control of FP is important, a higher threshold is preferred. At a threshold of 0.6, FP achieved the level of the gold standard.

2.3.2 Finite Mixture Model Analysis of Real Data

A scatter plot of the LR test statistics versus within-gene residual expression variances (Fig. 2.1 A) revealed that higher LR values tended to be associated with lower variances. A scatter plot of the modified LR test statistics versus within-gene residual

expression variances (Fig. 2.1 B) showed a somewhat more even distribution of the statistics across the range of variances. The constant S(0) in the modified LR statistic was determined to equal 0.0456 in this study. Permutation thresholds corresponding to type I error rates of 0.01, 0.005 and 0.001 identified 360, 218 and 50 genes, respectively, with significant LR test.

The first five eigenvalues and the proportion of variance explained by each eigenvector of the retained 360, 218 and 50 genes expression were shown in Table 2.2. The first eigenvalue is much larger (containing more than 99% of the variance) than the others in three of the retained gene sets, and first eigenvector explains most of the variation in the expression. This finding indicates substantial correlation among the expression patterns of the selected genes. The first two principal components of the NT and IVF embryos are shown in Fig. 2.2 (based on the set of 360 genes). All NT embryos are clustered together, while the IVF embryos are more spread out, and two IVF embryos are very close to the NT group. When considering only the first, dominant eigenvector, essentially all of the NT embryos and half of the IVF embryos fall into the higher value region of the first principal component, indicating that they are probably DI embryos; while the other half of IVF embryos is located in the lower value region of the first principal component, suggesting that they may be DC embryos.

Classification of embryos was performed based on the first principal component, since the first eigenvalue is much larger than the others. Posterior probabilities of developmental competency for each of the 20 embryos in this study are listed in Table 2.3. Six IVF embryos are classified as DC based on the 0.50 threshold. These are the same for all three sets of selected genes (360, 218 and 50), although there is some (minor) variation in the values of the posterior probability of individual embryos among sets.

2.4 Discussion

Using gene expression data, several statistical methods were combined to achieve the goal of classifying uncertain labeled samples. LMM was used to remove systemic errors. FMMA was proposed for gene selection, and sample classification based on PCA projected variables. The LMM could identify genes, which are differentially expressed between NT and IVF embryos. Since only 2% of the NT embryos are expected to be DC, as opposed to 50% of the IVF embryos, the LMM analysis should be able to identify genes that are differentially expressed between DC and DI, but at less than maximally possible power. FMMA is expected to have higher accuracy both in gene selection and in classification, which is confirmed by our simulation.

Mixture model analysis of microarray expression data has been applied in several previous studies mostly for clustering of samples, which considered the case of an unknown number of component distributions in the mixture and unknown mixing probabilities (Ghosh and Chinnaiyan, 2002; Liu *et al*., 2003; McLachlan *et al*., 2002; Yeung *et al*., 2001). Here, we know that there are two underlying classes (DC, DI), which we wish to discriminate between, and we assume that the mixing weights are known for different groups of individuals (NT, IVF). Our FMMA implementation can be improved in various ways, for example by allowing for unequal variances or covariance matrices of the component distributions (due to the small number of observations in this study, we assumed equality), by replacing the multivariate normal component distributions with multivariate t-distributions (McLachlan *et al*., 2002), and by considering other methods of dimension reduction, such as factor analysis (McLachlan *et al*., 2002) and correspondence analysis (Liu *et al*., 2003).

Many classification or predication studies, which assume that the class is known, are in fact the situations that the class is uncertain. For example, in classifying solid tissue samples, the gene expression measured are indeed from a mixture of the various cellular types contained in a set of samples directly from the measurements taken on the whole sample. This situation has been noticed by others (Venet et al., 2001), and we believe

FMMA could provide a powerful classification in this condition. The finite mixture model analysis may have other useful applications. For example, different probabilities of breast cancer survival are known for different statuses of lymph nodes (van de Vijver *et al*. 2002), which could be incorporated in a FMMA for survival prediction.

Appendices 2

2A. EM algorithms for mixture of multivariate normal distributions

The form of multivariate normal distribution:

$$\phi_k(x_i;\mu_k,\Sigma) = (2\pi)^{-d/2} \mid \Sigma \mid^{-1/2} e^{-\frac{1}{2}(x_i-\mu_k)'\Sigma^{-1}(x_i-\mu_k)} \qquad \text{(2A.1)}$$

Here k= 0 or 1, corresponding to development incompetent and competent. x_i is a vector of p measured gene expression from microarray experiments. μ_1 is the mean vector of the p genes of developmental competent embryos, μ_0 is the mean vector of the p genes of developmental incompetent embryos, and Σ is the common variance covariance matrix of the genes expression for both classes.

The EM algorithms can be divided into Expectation and Maximization two steps. Generally, after the initial setting of the parameters, the algorithm proceeds by cycling over the two steps: E-step, based on the current parameter estimates, the posterior probability that ith observation is from class k is,

$$\tau_k(x_i;\Theta) = \frac{\pi_k \phi_k(x_i;\mu_k,\Sigma_k)}{\sum_{h=1}^{g} \pi_h \phi_h(x_i;\mu_h,\Sigma_k)} \qquad \text{(2A.2)}$$

where π_h is the prior probability that the ith observation belongs to class h.

M-step new parameter estimates are obtained by utilizing the expected value of the hidden variable from the E-step:

$$\mu'_k = \frac{\sum_{i=1}^{n}[\tau_k(x_i;\Theta)x_i]}{\sum_{l=1}^{n}\tau_k(x_l;\Theta)} \tag{2A.3}$$

$$\Sigma'_k = \frac{\sum_{i=1}^{n}[\tau_k(x_i;\Theta)(x_i-\mu'_k)(x_i-\mu'_k)^T]}{\sum_{l=1}^{n}\tau_k(x_l;\Theta)} \tag{2A.4}$$

The E- and M- steps are alternated repeatedly until convergence of the parameters. The EM algorithm has reliable convergence in that regardless of the starting value, the likelihood never decreases after each EM iteration.

2B. Derivation of posterior probability of development competent for specific embryo

$$\begin{aligned} P(T_i=1\mid x_i,Y_i) &= \frac{P(T_i=1,x_i\mid Y_i)}{P(T_i=1,x_i\mid Y_i)+P(T_i=0,x_i\mid Y_i)} \\ &= \frac{P(x_i\mid T_i=1,Y_i)P(T_i=1\mid Y_i)}{P(x_i\mid T_i=1,Y_i)P(T_i=1\mid Y_i)+P(x_i\mid T_i=0,Y_i)P(T_i=0\mid Y_i)} \\ &= \frac{P(x_i\mid T_i=1)P(T_i=1\mid Y_i)}{P(x_i\mid T_i=1)P(T_i=1\mid Y_i)+P(x_i\mid T_i=0)P(T_i=0\mid Y_i)} \\ &= \frac{\phi_{i1}P(T_i=1\mid Y_i)}{\phi_{i1}P(T_i=1\mid Y_i)+\phi_{i0}P(T_i=0\mid Y_i)} \end{aligned} \tag{2B.1}$$

Where $P(T_i=1|Y_i=0)=0.02$, $P(T_i=0|Y_i=0)=0.98$, $P(T_i=1|Y_i=1)=0.5$, $P(T_i=0|Y_i=1)=0.5$, ϕ_{i1} and ϕ_{i0} are likelihood values for the ith embryo defined above with parameters substituted by MLEs learned from the EM algorithms. In the above we assume $P(x_i|T_i,Y_i)=P(x_i|T_i)$, which means that given we know the developmental competence status T_i, gene expression x_i is independent to its origin Y_i no matter it is NT or IVF

embryo. In this representation, we will classify the embryos as development competent or not according to their posterior probabilities.

2C. Proof of linear combination of mixture normal distribution still a mixture normal distribution

Assume that $\underline{x}$ is a vector, which follows a mixture of multivariate normal distribution. Its pdf can be written as a finite mixture distribution:

$f_X(\underline{x}) = \sum_{i=1}^{k} \pi_i f_i(\underline{x})$ where $\sum_{i=1}^{k} \pi_i = 1$ and $f_i(\underline{x})$ follows multivariate normal distribution.

Now transformed variable $\underline{z}$ is a vector which is transformed from $\underline{x}$ by an orthogonal matrix B containing eigenvectors. So each individual variable in $\underline{z}$ is a linear combination of the original variable with one eigenvector.

So: $\underline{z} = B\underline{x}$, and $\underline{x} = B^{-1}\underline{z}$

The pdf of $\underline{z}$ can be obtained according Jacobi transformation:

$$f_Z(\underline{z}) = \frac{f_X(B^{-1}\underline{z})}{|B|} = \frac{\sum_{i=1}^{k} \pi_i f_i(B^{-1}\underline{z})}{|B|} = \sum_{i=1}^{k} \pi_i \frac{f_i(B^{-1}\underline{z})}{|B|} \qquad (2C.1)$$

for each $\frac{f_i(B^{-1}\underline{z})}{|B|}$,

$$\begin{aligned}\frac{f_i(B^{-1}\underline{z})}{|B|} &= \frac{1}{|B|(2\pi)^{n/2}\,|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(B^{-1}\underline{z}-\underline{\mu}_i)'\Sigma_i^{-1}(B^{-1}\underline{z}-\underline{\mu}_i)} \\ &= \frac{1}{|B'B|^{1/2}\,(2\pi)^{n/2}\,|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\underline{z}-B\underline{\mu}_i)'(B^{-1})'\Sigma_i^{-1}(B^{-1})(\underline{z}-B\underline{\mu}_i)} \\ &= \frac{1}{(2\pi)^{n/2}\,|B\Sigma_i B'|^{1/2}} e^{-\frac{1}{2}(\underline{z}-B\underline{\mu}_i)'(B\Sigma_i B')^{-1}(\underline{z}-B\underline{\mu}_i)}\end{aligned} \qquad \text{(2C.2)}$$

it follows a normal distribution with mean $B\underline{\mu}_i$ and variance $B\Sigma_i B'$. So the variable $\underline{z}$ follows a mixture of normal distribution.

Reference

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X.*, et al.* (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503-511.

Beer, D. G., Kardia, S. L., Huang, C. C., Giordano, T. J., Levin, A. M., Misek, D. E., Lin, L., Chen, G., Gharib, T. G., Thomas, D. G.*, et al.* (2002). Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med *8*, 816-824.

Colman, A. (2000). Somatic Cell Nuclear Transfer in Mammals: Progress and Applications. *1*, 185 - 200.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society *39*, 1-38.

Dudoit, S., Fridlyand, J., and Speed, T. P. (2002). Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. Journal of the American Statistical Association *97*, 77--87.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.

Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. (2000). Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics *16*, 906-914.

Ghosh, D., and Chinnaiyan, A. M. (2002). Mixture modelling of gene expression data from microarray experiments. Bioinformatics *18*, 275-286.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A.*, et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science *286*, 531-537.

Harville, D. A. (1977). Maximum likelihood approaches to variace component esti-

mation and to related problems. J Amer Statist Assoc *72*, 320-340.

Liu, J. S., Zhang, J. L., Palumbo, M. J., and Lawrence, C. E. (2003). Bayesian clustering with variable and transformation selections. Bayesian Statistics *7*, 249-275.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol *14*, 1675-1680.

McLachlan, G. J., Bean, R. W., and Peel, D. (2002). A mixture model-based approach to the clustering of microarray expression data. Bioinformatics *18*, 413-422.

McLachlan, G. J., and Peel, D. (2000). Finite mixture models (New York, Wiley).

Ramaswamy, S., Tamayo, P., Rifkin, R., Mukherjee, S., Yeang, C. H., Angelo, M., Ladd, C., Reich, M., Latulippe, E., Mesirov, J. P.*, et al.* (2001). Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A *98*, 15149-15154.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E. S., and Golub, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A *96*, 2907-2912.

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J., and Church, G. M. (1999). Systematic determination of genetic network architecture. Nat Genet *22*, 281-285.

Titterington, D. M., Smith, A. F. M., and Makov, U. E. (1985). Statistical analysis of finite mixture distributions (New York, John Wiley& Sons).

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., Voskuil, D. W., Schreiber, G. J., Peterse, J. L., Roberts, C., Marton, M. J.*, et al.* (2002). A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med *347*,

1999-2009.

Venet, D., Pecasse, F., Maenhaut, C., and Bersini, H. (2001). Separation of samples into their constituents using gene expression data. Bioinformatics *17*, 279S-287.

West, M., Blanchette, C., Dressman, H., Huang, E., Ishida, S., Spang, R., Zuzan, H., Olson, J. A., Jr., Marks, J. R., and Nevins, J. R. (2001). Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A *98*, 11462-11467.

Yeung, K. Y., Fraley, C., Murua, A., Raftery, A. E., and Ruzzo, W. L. (2001). Model-based clustering and data transformations for gene expression data. Bioinformatics *17*, 977-987.

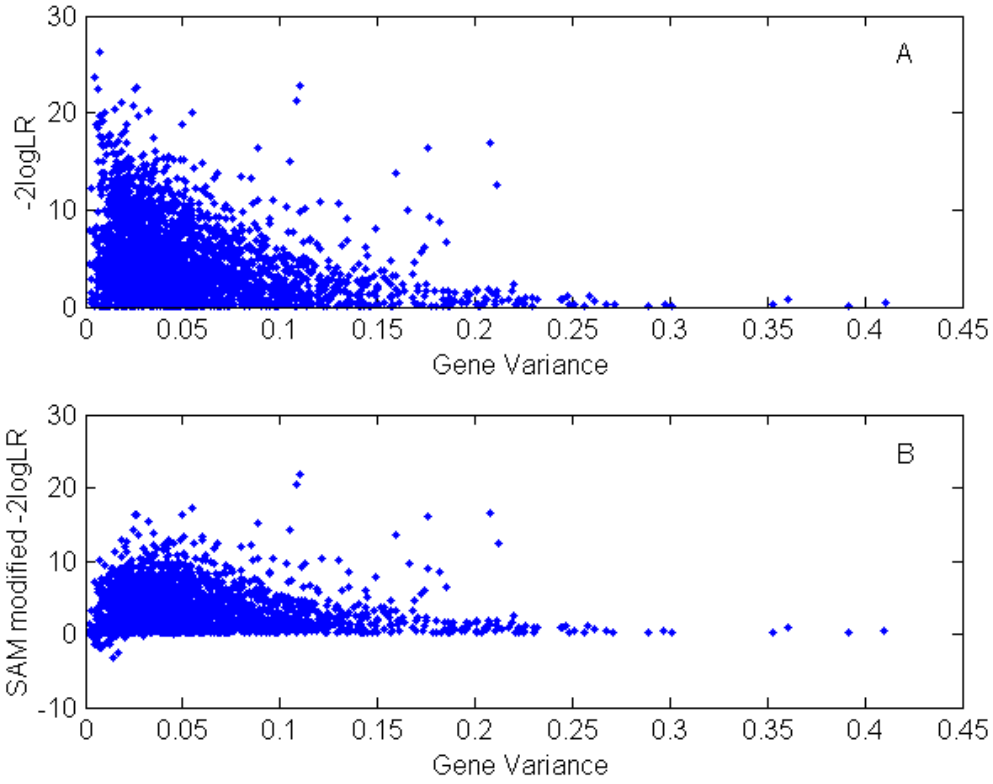

Fig 2. 1 Scatter plot of LR vs. gene residual variance

Scatter plots of the log likelihood ratio (LR) statistics versus estimated within-gene residual expression variance (*A*) and of the modified (see main text) log LR statistics versus estimated within-gene residual expression variance (B).

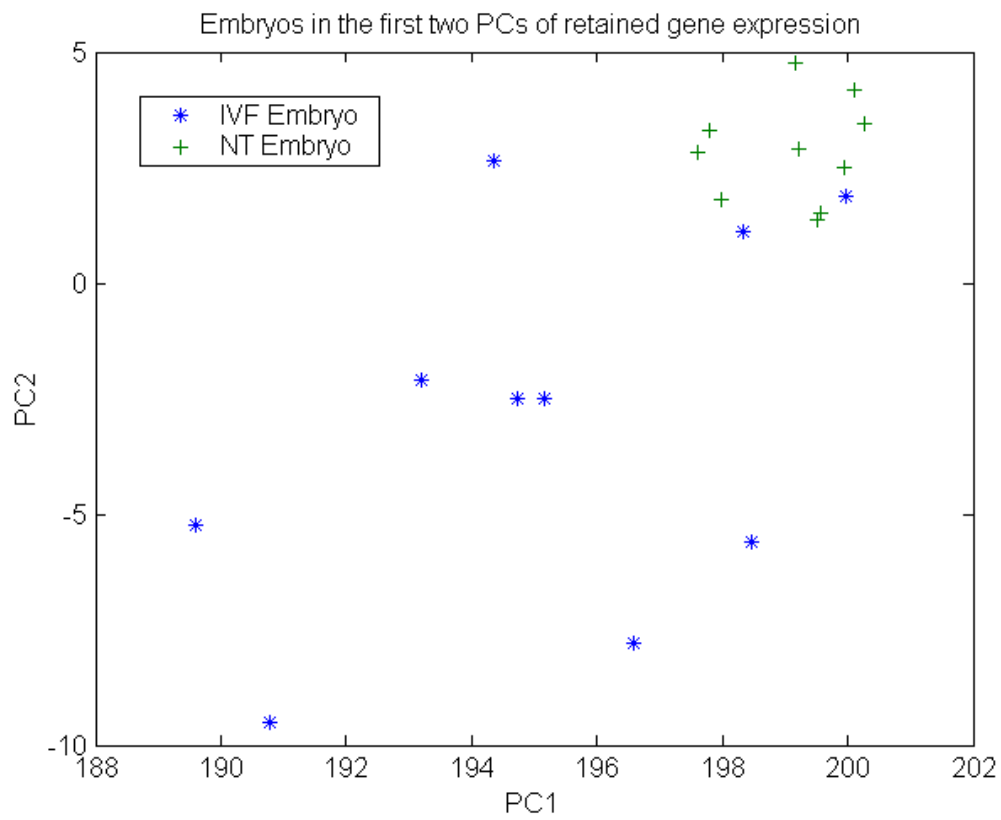

Fig 2. 2 Scatter plot of 20 embryos in first two principal components

Scatter plot of 20 embryos at the first two principal components space of retained 360 genes expression. IVF embryos have much spreader expression than NT embryos.

Table 2. 1 Comparison of different classification method via simulation

Average and standard deviation of misclassification count over 100 simulated data sets for different prediction methods. False positive count (FP): number of DI embryos classified as DC. False negative count (FN): number of DC embryos classified as DI.

Prediction Methods			False Positives		False Negatives		Total Misclassifications	
			Mean	Std	Mean	Std	Mean	Std
LDA using DC/DI labels			3.67	*7.50*	2.06	*3.85*	5.73	*10.94*
LDA using IVF/NT labels			43.17	*6.09*	2.79	*2.06*	45.96	*6.61*
Cluster Analysis	Complete linkage		52.12	*53.48*	9.64	*12.56*	61.76	*59.52*
	Single linkage		125.41	*52.95*	0.97	5.08	126.38	*52.17*
	Average linkage		92.76	*70.57*	2.79	*8.10*	95.55	*70.01*
	Centroid linkage		126.76	*51.05*	1.63	*6.82*	128.39	*49.26*
Mixture Model Based Method	Threshold 0.5	10 genes	6.00	*7.57*	6.80	*8.43*	12.80	*15.85*
		2 genes	5.78	*7.56*	6.55	*8.28*	12.33	*15.69*
	Threshold 0.6	10 genes	3.20	*4.15*	10.76	*14.52*	13.96	*17.17*
		2 genes	3.04	*4.16*	10.39	*14.28*	13.43	*16.91*
	Threshold 0.7	10 genes	1.74	*2.50*	13.44	*17.37*	15.18	*18.59*
		2 genes	1.71	*2.54*	12.98	*17.05*	14.69	*18.33*

Table 2. 2 First five eigenvalues of the three sets of retained gene expression

Eigenvalue Number	Eigenvalues of 360 Retained Gene	Proportion of Total Variance	Eigenvalues of 218 Retained Gene	Proportion of Total Variance	Eigenvalues of 50 Retained Gene	Proportion of Total Variance
1	762290	99.91%	460030	99.91%	104550	99.89%
2	279.558	0.04%	167.936	0.04%	61.507	0.06%
3	107.845	0.01%	68.999	0.01%	16.196	0.02%
4	74.195	0.01%	48.097	0.01%	12.328	0.01%
5	53.714	0.01%	31.330	0.01%	11.350	0.01%

Table 2. 3 Posterior probability of DC for embryos

Posterior probability of developmental competency for individual NT and IVF embryos, for three sets of selected genes corresponding to significance levels of 0.01, 0.005, and 0.001. Embryos exceeding the threshold of .6 are marked with a "*" and classified as DC.

Embryo		Posterior Probability (360 selected genes)	Posterior Probability (218 selected genes)	Posterior Probability (50 selected genes)
IVF	1	0.0022	0.0059	0.0113
	2	0.9991*	0.9986*	0.9974*
	3	0.0000	0.0001	0.0000
	4	0.0016	0.0025	0.0002
	5	0.1667	0.3618	0.0538
	6	1.0000*	1.0000*	1.0000*
	7	0.9559*	0.9066*	0.8275*
	8	1.0000*	1.0000*	1.0000*
	9	0.8810*	0.8010*	0.6435*
	10	0.9828*	0.9685*	0.9021*
NT	1	0.0000	0.0000	0.0000
	2	0.0000	0.0000	0.0000
	3	0.0000	0.0000	0.0000
	4	0.0001	0.0003	0.0001
	5	0.0000	0.0000	0.0001
	6	0.0002	0.0002	0.0001
	7	0.0003	0.0002	0.0001
	8	0.0000	0.0000	0.0000
	9	0.0000	0.0000	0.0001
	10	0.0000	0.0000	0.0000

Chapter 3

Transcription Network Inference Using Natural Multigenetic Perturbations

Abstract

Joint analysis of gene expression and genotype data in a segregating population could recover genetic effect on gene expression and has the potential to reveal naturally occurred gene networks. This strategy treats gene expression as a quantitative trait (QT) and maps quantitative trait loci (QTL) for each expression; however, identification of candidate regulatory genes is usually difficult due to the fact that the QTL region is relatively large. Here, we propose to utilize the expression correlation information to narrow the candidate genes list in the expression QTL region. This method is applied to a Yeast data set. The number of candidate genes in each QTL confidence interval is greatly reduced by the correlation test. About one third of candidate genes fall in the marker sub-intervals QTL fine mapped from a three marker sliding regression analysis. The genetic networks are constructed by combining directional links from candidate genes within the QTL region to the analyzed gene, whose expression is affected by this QTL. Several biologically meaningful regulations are detected. Highly interconnected groups of genes, as disclosed in the final reconstructed gene networks, tend to be involved in common biological processes. Our extension to Genetical Genomics provides an efficient approach to constructing gene regulatory networks from natural multigenetic perturbations, with the potential to generate valid and new hypotheses for future studies.

3.1 Introduction

The identification of genes and genetic regulatory networks underlying complex traits is a fundamental aim of genetics. Quantitative trait locus (QTL) mapping is a method which identifies genomic regions associated with a phenotype of interest (Korstanje and Paigen 2002). Large-scale gene expression data acquired from microarray experiments (Schena, Shalon et al. 1995; Lockhart, Dong et al. 1996) provide information about regulatory relationships between genes. Most approaches inferring gene networks from microarray experiments are either based on external environment perturbations (Causton, Ren et al. 2001) or on single gene perturbations (commonly achieved through knockouts) in the otherwise same, homogeneous genetic background (Ideker, Thorsson et al. 2001). Recently, a strategy to infer genetic networks using natural and multifactorial genetic perturbations was proposed, and was coined "Genetical Genomics" (Jansen and Nap 2001). This method combines QTL mapping and microarray technologies via joint analysis of genotype and expression profiling in a segregating population. Each gene expression profile is treated as a quantitative trait (QT), which is affected by multiple genetic variants. QTL analysis of gene expression profiles helps to identify the genomic regions, which are likely to contain at least one causal gene with regulatory effect on analyzed gene expression. If the causal gene, underlying a QTL affecting the expression profile of a given gene, could be identified, then the link from the causal gene to the expressed gene would represent a regulatory relationship. Because a large number of genes are expression profiled in microarray experiments, many such links should be detected. By combining the links, genetic networks would be constructed. Although this strategy is clear in concept, how to construct a set of candidate genetic networks, which reflect the underlying true network structure as closely as possible given the limited information, is not obvious. This task is difficult because QTL regions are generally large (several cM) and hence contain many putative causal genes. Identification of causal genes usually requires application of a functional validation procedure, which is expensive, time consuming, and feasible

for at most tens of genes. Computational methods are needed to reduce the number of candidate genes in each QTL region to a smaller number, which either allows us to propose a finite set of candidate genetic network structures immediately, or to perform a smaller and feasible number of validation studies prior to network inference.

Here we propose a novel method to reduce the number of candidate genes to one or few genes in each QTL region by using correlations between the expression values of the gene undergoing QTL analysis and the candidate genes found in a given QTL region. The assumption is that genes belonging to the same pathway or network tend to have strong correlations between their expression values. This assumption has been used extensively in cluster analysis (Eisen, Spellman et al. 1998) and construction of co-expression gene networks (Stuart, Segal et al. 2003). Correlation analysis of all expression-profiled genes in a micorarray experiment, without QTL analysis, may produce many spurious associations; however, after QTL analysis has determined one or several QTL regions for each expression profile, then significant expression correlation of candidate genes in a QTL region with the gene, whose expression profile has undergone QTL analysis, should tend to indicate real functional relationships. We applied our approach to data from a Yeast study (Brem, Yvert et al. 2002), which is the first experiment on genetic dissection of genome-wide expression profiling. The authors performed QTL analysis in the form of nonparametric marker analysis of expression profiles in 40 haplotypes from a cross between a laboratory and a wild strain of Yeast. A total of 570 gene expression profiles were affected by at least a significant QTL. Thirty-two percent of these expression profiles were QTL mapped to the Yeast genomic region, where the genes themselves were located, indicating cis-regulation. Moreover, eight trans-acting loci were found to affect the expression of groups of genes with 7 to 94 members representing genes of related function (Brem, Yvert et al. 2002).

This experiment (Brem, Yvert et al. 2002) was very successful in identifying QTL regions for gene expression profiles; however, the identification of causal genes within

the QTL regions is difficult, but necessary for gene network reconstruction. An extended study of the Yeast experiment identified two genes responsible for the trans-acting loci via positional cloning and functional analysis (Yvert, Brem et al. 2003). Here, we reanalyzed the data from the Yeast study (Brem, Yvert et al. 2002). We allowed each expression profile to be affected by multiple QTL residing on different chromosomes, and we computed a confidence interval for each QTL by bootstrapping. A set of genes physically located within each QTL confidence region were found using the sequenced Yeast genome map. For each confidence region, a single gene or a subset of genes were identified as likely causal gene(s) based on expression correlations. Directional links were established from these likely causal genes to the gene whose expression profile had undergone QTL analysis. These links were then connected for gene network inference. A three-marker sliding regression method is used for QTL fine mapping. The candidate gene locations are compared to the fine mapped marker sub-intervals. We believe that we have provided an initial, efficient method for inferring gene regulatory networks in experiments using natural genetic perturbations.

3.2 Methods

3.2.1 QTL Analysis of Gene Expression Profiles

We used the gene expression and genotype data from the 40 *Saccharomyces cerevisiae* haplotypes (Brem et al., 2002). The data set contains 6215 gene expression values and genotypes at 3312 markers for each haplotype. In the previous analysis of this data set (Brem et al., 2002), a significant QTL was identified for 570 gene expression profiles using nonparametric, single marker analysis based on the Wilcoxon-Mann-Whitney test and a significance threshold of $p<5\times10^{-5}$. Only the most significant marker across the entire Yeast genome was determined for each of the 570 gene expression profiles. It is certainly possible that there are multiple significant QTL for some of the gene expression profiles. We used the same nonparametric analysis and the same p-value

threshold to detect QTL, but we retained the most significant QTL per chromosome. Hence multiple QTL residing on different chromosomes can be detected for a gene expression profile.

3.2.2 QTL Confidence Intervals

A confidence interval (CI) was computed for each retained, significant QTL via a bootstrap resampling method (Visscher et al., 1996). Bootstrap samples were created via sampling, with replacement, the set of expression values and the set of marker genotypes of any of the 40 haplotypes. Marker analysis was performed on each of 1,000 bootstrap data sets. For each chromosome with a significant QTL affecting a given expression profile in the original data set, the QTL position with the highest test statistic was retained for each of the 1,000 bootstrap samples. These 1,000 QTL positions were ordered across the chromosome. The 95% confidence interval of the QTL position was then determined by taking the bottom and top 2.5% of the ordered QTL positions.

3.2.3 Identification of Candidate Genes via Expression Correlation Test

Genes physically located in the confidence intervals of the QTL were identified from the Yeast physical map (Goffeau et al., 1996). One or several of the genes located in the CI of a QTL affecting the expression profile of a particular gene, say gene A, should have regulatory effects on the expression level of gene A. Spearman correlation coefficients were computed between these genes and gene A for each CI. Each correlation coefficient was transformed using Fisher's Z transformation and its significance was evaluated via the Z-test (Appendix). The *p*-value threshold was Bonferroni adjusted as $0.01/n$, where n is the number of genes in each confidence interval. Only those genes whose correlation was significant with the Bonferroni adjustment were retained. If more than one gene was retained within the same CI, then

the differences between the most significant correlation with the other retained correlations were tested via the two sample Z-test (Appendix). If a difference was significant at a *p*-value threshold of 0.05, then the less correlated gene was removed from the candidate gene list of the CI.

3.2.4 Fine Mapping of QTLs and Comparison with Candidate Gene Locations

To narrow the QTL regions to small intervals and differentiate multiple possible QTLs in one chromosome, a (sliding) three-marker regression analysis is proposed for fine mapping of the QTL. Such method is described previously (Thaller and Hoeschele, 2000) and applied here to the detected 95% confidence intervals of the QTL from single marker bootstrap analysis. Three consecutive markers, whose genotype constituents are required to be different in at least 2 of the 40 haplotypes, are selected from the beginning to the end of the previously determined confidence intervals. Gene expression affected by this QTL is regressed on such sliding consecutive three markers. The partial regression coefficient for the intermediate marker *i* has a non-zero expected value if and only if there is at least one QTL located between markers *i*-1 and *i*, or between *i*+1 and *i* (Zeng, 1993). Hence, if there is a single QTL in the confidence interval, only the two markers flaking the sub-interval containing the QTL have non-zero expected partial correlation. In this study, since significant single marker analysis has indicate that there is at least one QTL is the confidence interval, we select the sub-interval flanked by markers *i* and *i*+1 with the largest *t*-values as the fine mapped QTL region. We also assign multiple QTL sub-intervals in one QTL confidence interval, where both the flanking markers and markers within (if there is any) are significant (at the 0.05 level of partial regression) and at least one marker next to the flanking markers to be not significant.

The genomic locations of the retained candidate genes are compared with the fine

mapped QTL marker sub-intervals. The distances are calculated as the nearest location differences between candidate genes with the subintervals. If there is any overlap between the gene location and the sub-intervals, the distance is recorded as zero.

3.2.5 Construction of the Network

For any QTL confidence interval, where the set of candidate genes was reduced to one or a few genes as described above, a directional link was drawn from each of these retained candidate genes to the expressed gene (the gene whose expression profile was affected by the QTL). The genetic network was constructed by combining all links. The network structure was displayed using the network drawing software Cytoscape (Shannon et al., 2003), where a node represents a gene and a directional arrow represents a putative regulatory relationship. Transcription factors identified in the inferred network were obtained from the studies of Lee et al. (Lee et al., 2002) and Guelzim et al. (Guelzim et al., 2002). The common functions of the genes contained in an independent network or in highly connected sub-networks were analyzed through Yeast Gene Ontology (Ashburner et al., 2000).

3.3 Results

In addition to the previously identified detected 570 QTL (Brem et al., 2002), an additional eleven QTL were detected at the $p<5\times10^{-5}$ level using single marker analysis based on the Wilcoxon-Mann-Whitney test, which we applied via identifying the most significant QTL at each chromosome separately rather than genome-wide. The genomic positions of expression-profiled genes were plotted against the genome locations of DNA markers significantly affecting expression profiles (Fig. 3.1). A diagonal point represents a DNA marker with a genome location very close to the location of the gene, whose expression profile is significantly affected by this marker. An off-diagonal point indicates an expression-profiled gene whose map position is different from the location

of a marker significantly affecting the expression. Hence, a diagonal point represents a putative *cis*-regulation, while an off-diagonal point is a putative *trans*-regulation. The dense distribution of diagonal points across the Yeast genome indicates a large proportion of *cis*-regulations. As indicated in previous study (Brem et al., 2002), groups of expression profiled genes shared common QTL regions. Within a group, the expression-profiled genes did not exhibit any significant co-location on the Yeast genome, which means that these genes with common QTL do not merely represent coordinated regulated genes in the same chromosome region.

The length of the QTL confidence intervals, obtained by bootstrapping, varied with a minimum distance of 66 bp, a maximum distance of 1,319,588 bp and a median distance of 93,476 bp. The number of genes within an interval ranged from zero to 717 genes, with a median number of 49. The number of genes in a confidence interval was highly correlated (correlation coefficient of 0.98) with the length of the interval.

The number of candidate genes in each interval was reduced by requiring a significant Spearman correlation between the expression profiles of the candidate gene and the expression profiles of gene affected by the QTL, as described in Methods. In nearly 60% of the QTL regions, a single gene was retained as the candidate gene (Fig. 3.2). In one extreme case, nineteen significant genes were retained in the QTL region; however, the frequencies of larger numbers retained decreased quickly with the increase in number. We note that in 13% of the QTL regions, no significantly correlated candidate gene was identified.

Overall, 1027 pairs of significant correlation were revealed between the candidate genes in QTL confidence intervals and the genes whose expression are affected by such QTL. Though these correlations are valid hypothesis for gene regulations, we believe not all of them are correct ones due to the fact that the QTL confidence intervals are very large in some cases. To obtain more confident interactions, a (sliding) three-marker regression analysis is applied for fine mapping of QTL based on original obtained

confidence intervals. From the analysis, 848 marker sub-intervals were discovered from the previous defined 581 QTL confidence intervals. The length of the 848 sub-interval are much smaller than the previous defined QTL confidence intervals, with a minimum distance of 1 bp, a maximum distance of 84,219 bp and a median distance of 6,286 bp. Of the 1027 pairs of significant correlation from the original QTL confidence intervals, we compared the genomic location of the causal candidate genes with the corresponding fine mapped QTL sub-intervals. Among the comparisons, 305 (out of 1027) causal genes overlapped with the corresponding marker sub-intervals. Another 58 genes located within 1kb of the inferred sub-intervals. These causal genes and their corresponding regulations would represent more confident hypothesis for future studies. The interactions (causal and expressed genes), their corresponding fine mapped QTL sub-intervals, and distance between causal genes and QTL sub-intervals are listed in supplementary materials.

Gene regulatory networks were constructed by combining the directional links from retained candidate genes in QTL region to expression-profiled genes affected by the corresponding QTL. Several network motif structures similar to previously defined structures (Lee et al., 2002; Milo et al., 2002; Shen-Orr et al., 2002) were found (Fig. 3.3): (a) The feedback loop motif represents the case where either gene was physically located at the QTL region of the other gene. The two gene products (YGL051W, YAR033W, Fig. 3.3) in the feedback loop motif were also previously shown to interact based on a Yeast protein interaction experiment (Uetz et al., 2000). (b) The feedforward loop motif represents the case where one gene regulates another gene, and these two genes jointly regulate a third gene. This case shows that the effect of regulation between genes can be executed directly or indirectly through other genes. (c) The single input motif represents the case where the expression profiles of multiple or many genes are influenced by the same QTL region and the same gene was retained as the candidate regulatory gene in this interval. Hence this motif depicts a set of functionally related genes coordinately regulated by a simple input. All the genes in the single input motif of

Fig. 3.3 are involved in protein metabolism. (d) The multiple input motif represents the case where a set of genes are regulated jointly by the combination effect of another set of genes. In the multiple input motif of Fig. 3.3, multiple expression-profiled genes (MFA1/YDR461W, STE3/YKL178C, AGA2/YGL032C, MFALPHA2/YGL089C, STE2/YFL026W) are involved in response to pheromone functions. The expression profiles of all these genes are affected by the same significant QTL region, with the same candidate regulatory genes retained in the region: YCR039C and YCR040W. YCR039C and YCR040W encode MATALPHA2 and MATALPHA1. One serves as a transcription co-repressor and the other as a transcription co-activator. MATALPHA2 and MATALPHA1 act together and are both involved in the regulation of alpha-specific genes in response to pheromone functions.

The entire constructed networks included 721 genes and 1027 interactions (Fig. 3.4). These structures ranged from simple self-regulation, pairwise regulation, and interactions among a few genes, to highly connected networks. In the previous study (Brem et al., 2002), groups of genes were found to link to eight transacting loci. From the biological function descriptions, six genes were proposed as the possible *trans*-acting regulators for six groups of expression profiled genes (Brem et al., 2002). In our inferred set of networks, most of the profiled genes in the eight groups were included. Five out of the six putative regulators (*CST14, LEU2, MAT, URA1, SIR3*) were identified to regulate the corresponding groups of genes.

Our largest network was constructed by linking several densely connected regions with a few connections. The biological functions and processes of the genes in the highly interconnected sub-networks were obtained from the Gene Ontology database (Ashburner et al., 2000). Overall, one or several biological processes were statistically significantly over-represented in our independent network structures or in the highly interconnected sub-networks (Fig. 3.4 and Table 3.1). Genes involved in protein synthesis were found to be highly interconnected among themselves, as well as

connected to other groups of genes involved in sterol metabolism, oxidative phosphorylation, cytokinesis and response to stress. These findings show coordinated regulation of different biological processes. Other processes over-represented in the network included amino acid metabolism, development and structural constituents of ribosome.

Because the segregating population in this Genetical Genomics study was a cross between two strains of Yeast, the bioprocesses represented in the networks should be those pathways whose gene constituents carry different genetic variants resulting in phenotypic differences between the two strains. The role of transcription factors was also investigated in the inferred network structures. A list of 166 known transcription factors were obtained by combining two studies (Guelzim et al., 2002; Lee et al., 2002). Twenty-three of these transcription factors were found in the inferred networks (green color nodes in Fig. 3.4). They did not appear in the center of the entire interconnected network displayed in Fig. 3.4. Instead, most of these transcription factors had a putative regulatory link to only one other gene or exhibited *cis*-regulation. This result is in agreement with the finding that transcription factors showed no enrichment in *trans* variations (Yvert et al., 2003).

3.4 Discussion

In this investigation, we reanalyzed a Yeast cross segregating population with gene expression and DNA marker data recorded for all individuals and the entire Yeast genome. The goal of this implementation and extension of a Genetical Genomics analysis was to obtain results on the network structures to guide us in future investigations of a methodology for gene network reconstruction based on data from segregating populations. The following steps are critical in a Genetical Genomics analysis: (1) Determination of the required sample size for a segregating population such that an acceptable false discovery rate is achieved while sufficient power is

maintained for the identification of expression QTL and causal links in the network. (2) An optimal implementation of the QTL analysis, including (i) careful experimental design so that linear mixed models can be used to account for systematic sources of variation of microarray data (*e.g.*, array, spot) in the QTL mapping software, (ii) the computation of QTL confidence intervals of minimal length providing the desired coverage probabilities, (iii) the allowance of multiple QTL on the same chromosome to prevent the detection of ghost QTL regions (Zeng, 1993) and not containing the actual causal genes or regulators, (iv) the consideration of multiple trait QTL analysis for a group of expression profiles, which vary jointly (possibly determined by prior cluster analysis), and (v) the use of a multiple testing approach which considers not only the entire genome but also the many correlated traits (expression profiles). (3) An evaluation of statistical methods for further refinement of the constructed network structures (to eliminate some direct or indirect links not supported by the data).

Here we used bootstrapping to construct QTL confidence intervals. This bootstrap procedure tends to be conservative and to produce relatively large confidence regions, especially when the QTL effect is small (Dupuis and Siegmund, 1999). Selective bootstrap resampling was advocated to reduce the length of the confidence interval (Lebreton et al., 1998). Because it is not clear what selection criterion should be applied, we compared two selection strategies, which retained only those bootstrap samples whose highest significant QTL on a given chromosome and for a given expression profile achieved a p-value equal to or smaller than a threshold of 0.001 and 0.00005, respectively. The average confidence intervals for these two selective bootstrapping methods were smaller than the average confidence interval from the original bootstrap analysis; however, the lists of candidate regulatory genes retained in the intervals after the correlation test were very similar across all three bootstrapping methods (results not shown). Unnecessarily large QTL confidence intervals can also result from the presence of multiple QTL on the same chromosome affecting the same expression profile, hence QTL mapping procedures capable of resolving multiple linked

QTL are required.

Spearman rank correlation coefficients between the expression profiles of the candidate genes in a QTL region and the gene whose expression profile was affected by the QTL were used to determine a final, short list of candidate regulatory genes for each QTL interval. Spearman correlations are suitable for quantifying the strength of monotonic relationships and are more robust in the presence of nonlinear regulatory relationships between genes, when compared with the Pearson correlation coefficient. In nearly 60% of the QTL regions, a single correlated candidate gene was retained; however, in 13% of the QTL regions, no significantly correlated candidate gene was identified. In part, this finding may have resulted from the fact that some regulatory mechanisms do not exhibit expression correlations. For example, a protein coding polymorphism may affect the binding activity of a transcription factor to its downstream genes. This polymorphism may not change the transcription factor's transcript level, but would affect the expression of downstream genes. In this case, a gene expression profile would be found to be affected by a QTL representing the genomic location of the transcription factor, but the expression correlation of the transcription factor with the profiled gene may be low, and hence no candidate gene could be identified in the QTL region. To determine the candidate gene underlying such QTL regions, further functional analyses are needed. Other reasons for not identifying any candidate regulatory genes in 13% of all QTL regions could be lack of power (resulting from the limited sample size of 40 in the segregating Yeast population), and biased QTL intervals resulting from multiple QTL on the same chromosome and the use of single marker analysis. For 30% of all the QTL regions, where more than one candidate gene was retained based on significant expression correlation, additional functional information and analyses are needed to determine whether one or several genes are responsible for the regulatory effect of the QTL region. In the inferred mating development network structure (Fig. 3.4), where most of the genes are involved in response to pheromone functions, both regulators MATALPHA1 and MATALPHA2 have been shown to act together to regulate alpha-

specific genes in response to pheromone functions.

In this study, we used fine mapping of QTL via sliding three-marker regression. Such method can assign QTL to very narrow marker sub-intervals, which represent a further step after the original linkage analysis. However, since the underlying QTL regulation mechanisms are unknown, such as non-normal distributed expression data and possible epistasis among several QTLs, such narrowed sub-interval could be biased or wrong. In this study, we did not depend on such sub-intervals for causal gene identification. Rather, we utilize such information as a complementary validation approach for the inferred causal gene after significant correlation test. And the result shows that almost one third of the inferred causal genes falls in such small fine mapped sub-intervals. The interactions involving such causal genes would be more likely to be true, since they are supported by both the sub-intervals and the correlation test.

Gene network inference by joint analysis of gene expression and DNA marker genotype data in a segregating population, representing a natural, multifactorial perturbation experiment, has many advantages over single factor, extreme perturbation (e.g., gene knockout) experiments (Jansen, 2003). Another study on genetic dissection of gene expression in mice indicates that gene expression can be used to identify distinct disease subtypes, and that these subtypes are under the control of different loci (Schadt et al., 2003). For Genetical Genomics to be successful, researchers must be provided with optimal statistical and computational tools for designing and analyzing Genetical Genomics experiments, which can be developed quickly by utilizing the rich sets of tools previously developed in the QTL mapping and microarray analysis communities. Different implementations, algorithms and methods of analysis for Genetical Genomics studies should be compared objectively with artificial data (Quackenbush, 2001), which have been generated not from simple stochastic distributions, but rather were based on an existing regulatory network of genes. A system to simulate data based on gene networks has been developed recently (Mendes et al., 2003). Further developments

should also consider the use of metabolome and proteome data.

Appendices 3

3A. Rank based Spearman correlation

$$D = \sum_{i=1}^{n} [R(x_i) - R(y_i)]^2$$

$$r_s = 1 - \frac{6D}{n(n-1)(n+1)} \qquad (3A.1)$$

where n is the number of observations; $R(x_i)$ is the rank of x_i in the group of x and $R(y_i)$ is the rank of y_i in the group of y; r_s is the calculated Spearman correlation.

3B. Fisher Z transformation and significant test of correlation

Correlations are transformed into variables, which are approximately normally distributed by using Fisher's Z transformation, $Z = \frac{1}{2}\ln(\frac{1+r_s}{1-r_s})$.

The variance of Z is independent of r_s and is given by, $s_Z^2 = 1/(n-3)$.

Then $z = \frac{Z}{s_Z} = \frac{1}{2}\sqrt{n-3}\ln(\frac{1+r_s}{1-r_s}) \sim N(0,1)$, the significance p-value of the correlation can be obtained from the normal distribution of the transformed z value.

3C. Test of the difference between two correlations

The absolute values of the two correlations are first transformed via Fisher's Z

transformation, then $z = \frac{Z_1 - Z_2}{s_p} \sim N(0,1)$ and $s_p^2 = s_1^2 + s_2^2 = 2/(n-3)$

Reference

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T.*, et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet *25*, 25-29.

Brem, R. B., Yvert, G., Clinton, R., and Kruglyak, L. (2002). Genetic Dissection of Transcriptional Regulation in Budding Yeast. Science *296*, 752-755.

Causton, H. C., Ren, B., Koh, S. S., Harbison, C. T., Kanin, E., Jennings, E. G., Lee, T. I., True, H. L., Lander, E. S., and Young, R. A. (2001). Remodeling of yeast genome expression in response to environmental changes. Mol Biol Cell *12*, 323-337.

Dupuis, J., and Siegmund, D. (1999). Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics *151*, 373-386.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci U S A *95*, 14863-14868.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M.*, et al.* (1996). Life with 6000 genes. Science *274*, 546, 563-547.

Guelzim, N., Bottani, S., Bourgine, P., and Kepes, F. (2002). Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet *31*, 60-63.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science *292*, 929-934.

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

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

Korstanje, R., and Paigen, B. (2002). From QTL to gene: the harvest begins. Nat Genet

31, 235-236.

Lebreton, C. M., Visscher, P. M., Dupuis, J., and Siegmund, D. (1998). Empirical nonparametric bootstrap strategies in quantitative trait loci mapping: conditioning on the genetic model

Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics *148*, 525-535.

Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I.*, et al.* (2002). Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science *298*, 799-804.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol *14*, 1675-1680.

Mendes, P., Sha, W., and Ye, K. (2003). Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics *19*, II122-II129.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. Science *298*, 824-827.

Quackenbush, J. (2001). Computational analysis of microarray data. Nat Rev Genet *2*, 418-427.

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

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science *270*, 467-470.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the

transcriptional regulation network of Escherichia coli. Nat Genet *31*, 64-68.

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003). A gene-coexpression network for global discovery of conserved genetic modules. Science *302*, 249-255.

Thaller, G., and Hoeschele, I. (2000). Fine-mapping of quantitative trait loci in half-sib families using current recombinations. Genet Res *76*, 87-104.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P.*, et al.* (2000). A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature *403*, 623-627.

Visscher, P. M., Thompson, R., and Haley, C. S. (1996). Confidence intervals in QTL mapping by bootstrapping. Genetics *143*, 1013-1020.

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

Zeng, Z. B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc Natl Acad Sci U S A *90*, 10972-10976.

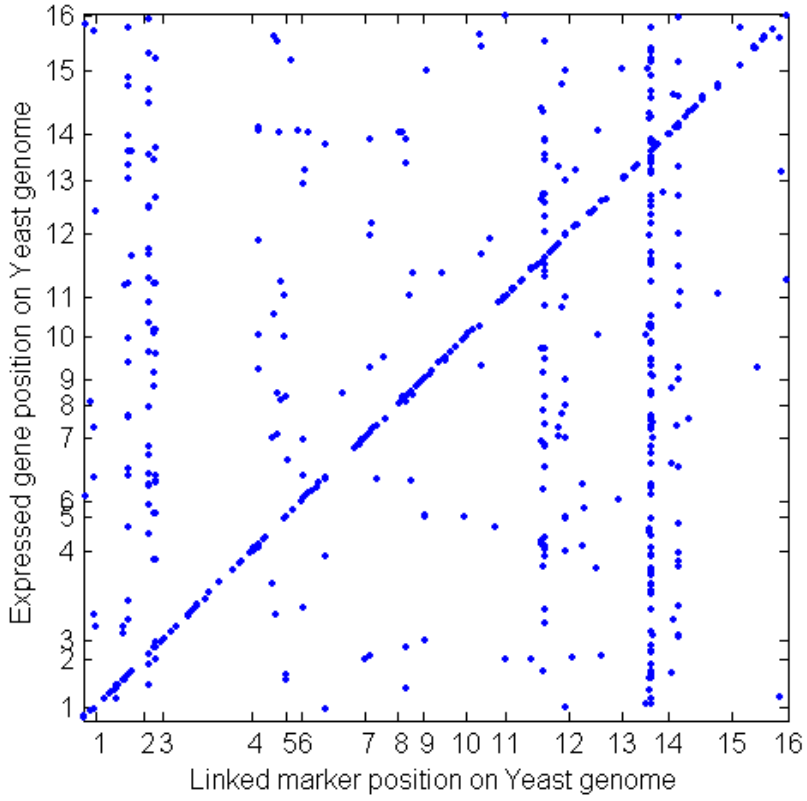

Fig 3. 1 Linked expressed gene genome location vs. linked marker genome location

Plot of the genome location of an expression profiled gene (Y axis) versus the genome location of a DNA marker significantly affecting the expression profile (X axis). The X and Y axes represent the entire Yeast genome consisting of 16 chromosomes of unequal length. A diagonal point represents a DNA marker with genome position very close to the location of the gene, whose expression profile is significantly affected by this marker. An off-diagonal point indicates an expression profiled gene whose map position is different from the location of a marker significantly affecting the expression.

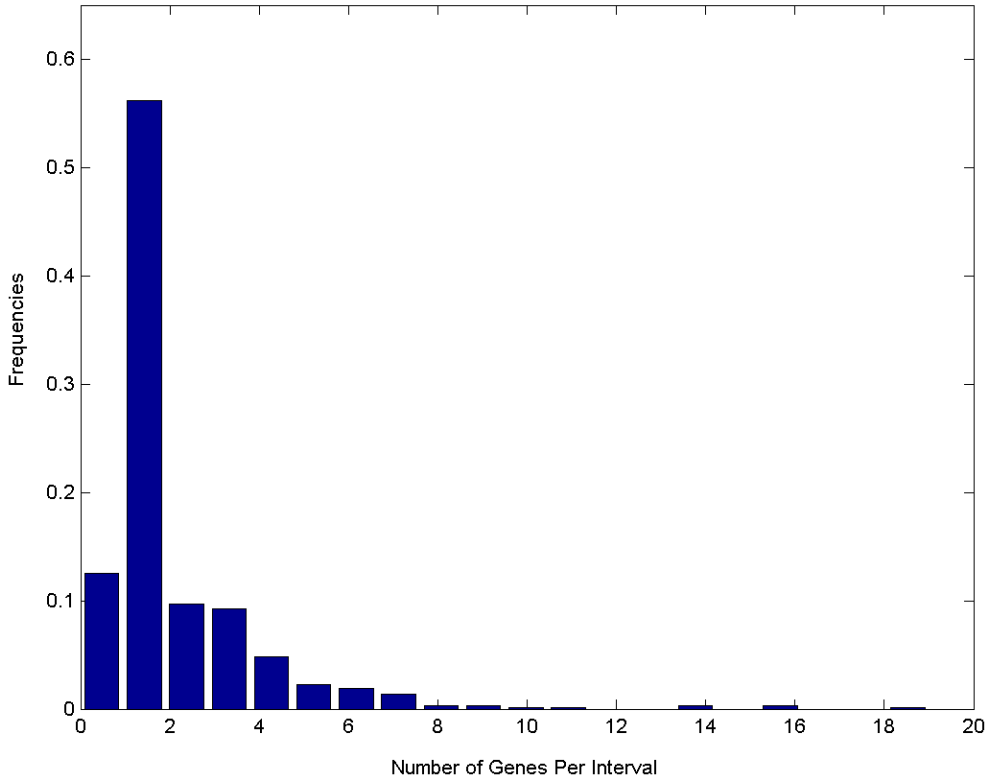

Fig 3. 2 Number of genes retained in each QTL region

The number of genes retained in each QTL confidence interval after expression correlation test ranged from 0 to 19. In 57% of the QTL regions, a single gene was retained. In another 13% of the regions, no gene was retained due to lack of significant expression correlation tests. In the remaining regions, more than one gene were retained.

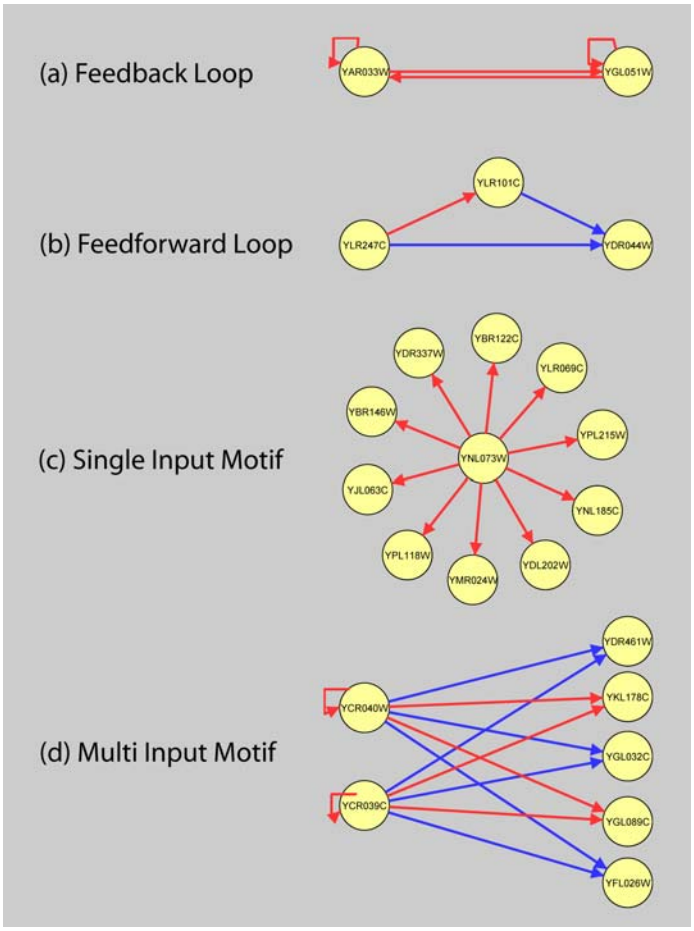

Fig 3. 3 Network motifs

Four different network motifs are presented. Directed links were drawn from retained candidate genes in a QTL region to the gene whose expression profile was affected by the QTL region. A link from one gene back to itself indicates *cis*-regulation, while a link from one gene to another gene represents putative *trans*-regulation. Red links depict positive correlations and blue links negative correlations.

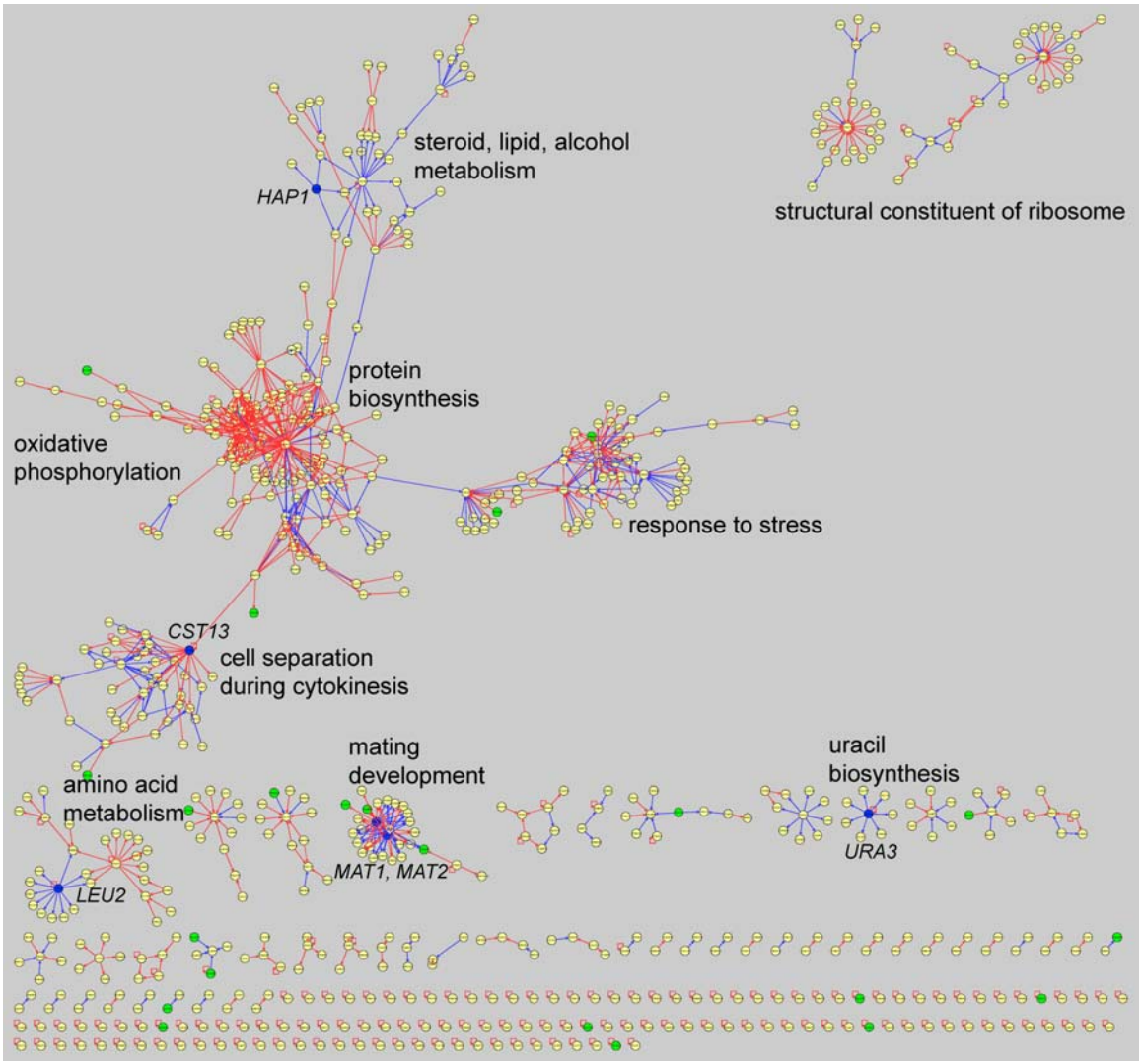

Fig 3. 4 Entire network topology.

The nodes represent genes and directed arrows indicate significant expression correlation between a candidate gene in a QTL region and the gene affected by the QTL. Transcription factors obtained from are plotted as green nodes. Five of the putative regulators (*CST13, LEU2, MAT, URA3, HAP1*) affecting the expression of groups of genes are plotted as blue nodes and are indicated. Biological processes over-represented in highly interconnected sub-networks are listed in the figure.

Table 3. 1 Over-represented biological processes in sub-networks

Group[a]	Biological Process[b]	Proportion in Group[c]		Proportion in Genome[d]		Enrichment *p* value[e]
1	Structural constituent of ribosome	14/53	26.4%	228/7270	3.1%	6.74E-10
2	Steroid metabolism	11/41	26.8%	42/7270	0.5%	6.44E-16
3	Oxidative phosphorylation	5/9	55.5%	34/7270	0.4%	2.77E-10
4	Protein biosynthesis	51/132	38.6%	766/7270	10.5%	2.64E-17
5	Response to stress	7/51	13.7%	362/7270	4.9%	1.28E-3
6	Cell separation during cytokinesis	3/45	6.6%	5/7270	0.07%	4.51E-6
7	Amino acid metabolism	12/30	40%	173/7270	2.3%	1.91E-12
8	Development	15/28	53.5%	343/7270	4.7%	2.65E-13
9	Uracil, Pyrimidine base biosynthesis	3/8	37.5%	13/7270	1.8%	3.78E-7

[a] The list of genes in each group and in corresponding biological processes can be found in the supplementary material.

[b] Biological processes were based on terms from Gene Ontology.

[c] The number of genes involved in the biological process within the inferred group divided by the total number of genes in the group.

[d] The total number of genes annotated in the biological process divided by the total number of annotated genes in the genome.

[e] The *p* value was computed as the probability of obtaining the observed or a larger number of genes in the group by chance under the hypergeometric distribution.

Chapter 4

Model Fit and Comparison of Biochemical Network Structures with Structural Equation Model

Abstract

Many algorithms have been proposed to infer genetic and biochemical networks from transcritomic, proteomic and metabolomic data. However, how to evaluate the fitness of these inferred networks against observed data and how to compare and differentiate between several possible hypothesized biochemical network structures especially in the presence of feedback loops have not been studied thoroughly. Here, we propose to use structural equation model (SEM) as a tool to evaluate the fitness of biochemical models against the data and to compare different biochemical network structures. The performance of SEM is examined against artificial biochemical data. Simulations indicate that the fitness test statistics of SEM follows the expected distribution under various disturbance magnitudes, different sample sizes, and is robust against the nonlinear relations among the simulated biochemical variables. Thus, the extent that the inferred network structures fit the observed data can be determined using the SEM χ^2 goodness-of-fit-test. Simulations from two sets of models also show that Bayesian information criterion (BIC) obtained from SEM is useful for comparison of different biochemical network structures.

4.1 Introduction

With the development of microarray technology, the amount of functional genomic data has been exponentially increasing. Additional biological data will be accumulating from proteomics and metabolomics studies. Several strategies have been proposed to infer biological network structures from such data. These include obtaining a gene network by perturbing one gene expression at a time, usually via knockouts (Ideker et al., 2001; Wagner, 2002); recovering naturally occurring gene networks via joint analysis of gene expression and genotype data in a segregating population – "Genetical Genomics" (Bing and Hoeschele, 2004; Jansen and Nap, 2001); inferring network structures from the statistical dependencies and independencies among measured expression levels using Boolean networks (Akutsu et al., 2000), linear models (D'Haeseleer et al., 1999), and Bayesian networks (Friedman et al., 2000). Although the methods described above could successfully discover biological network structures, how to statistically evaluate and justify inferred networks against measured gene expression data has not been well established. At the same time, biologists usually possess some knowledge of the underlying biochemical networks for the specific organisms and systems under study. How to use the burgeoning genomic data to confirm or improve our current knowledge about network structure and to differentiate different hypothesized candidate structures are unsolved questions in systems biology. Here, we propose to use structural equation models as a means to evaluate and compare biological network structures from the large amount of measured "omic" data.

Structural equation model (SEM) is a comprehensive statistical approach to testing hypotheses about relations among observed and latent variables. It is a powerful generalization of other statistical approaches, such as path analysis, factor analysis, and has been used extensively in psychology, sociology and econometrics as a tool for causal inference. SEM begins with the specification of linear models to be estimated. The parameters in the model are estimated from a set of observed data. Then these

estimates can be used to test whether the model is consistent with the observed data. This is achieved by evaluating the extent that the model implied covariance matrix is equivalent to the empirical covariance matrix. The most common index of fit is the χ^2 goodness-of-fit-test, which assumes that the data comes from a multivariate normal distribution (Bollen, 1989). Different models may be compared with each other using the Bayesian formalism of the posterior probability of the model given the observed data (Raftery, 1993).

The advantages of SEM in model biochemical networks is that it measures the variable in a continuous distribution rather than utilizing discretized data; it measures the over all fit of the structure to the data rather than merely fitting locally; it is also flexible enough to incorporate hidden variables into the structure if the theory supports this; and most importantly, it can model feedback in the structures, which is ubiquitous in biological systems but can't be modeled by other current computation methods such as Bayesian network. Although SEM has many advantages and has been successfully applied in many other fields, its effectiveness in modeling biochemical networks has not been shown or tested. SEM assumes multivariate normal distribution and linear relations among variables. How robustly SEM will perform in biochemical networks, where nonlinear relations and non-normal data distributions are possible, remains unknown. Here, the performance of SEM is examined using an artificial simulated biochemical network (Mendes, 1997). Two different networks are simulated with a combination of different perturbation magnitudes and sample sizes. The effectiveness of using the χ^2 goodness-of-fit-test has been evaluated from replicated simulations. Model comparisons by Bayesian Information Criterion (BIC) were conducted for two sets of simulated networks.

4.2 Theory and Methods

4.2.1 Model Fit in Structural Equation Models

A linear structural equation model (SEM) models each variable as a linear function of its direct causes and an error term. It can be represented as follows:

$$y = By + Fu + e; \quad Var(u) = U; \quad Var(e) = E \tag{4.1}$$

where y is a $p \times 1$ vector of dependent variables (observed), u is a $m \times 1$ vector of independent variables (some observed, some latent), e is a $p \times 1$ vector of unobserved residuals, and B and F are $p \times p$ and $p \times m$ matrices of unknown parameters (some may be fixed). We use θ, a $q \times 1$ vector, to represent the free parameters (those unfixed parameters in B and F, and the variance terms of E). The variance and covariance matrix of observed variable y from the implied structure is a function of θ:

$$\Sigma(\theta) = (I - B)^{-1}(FUF^T + E)(I - B)^{-T} \tag{4.2}$$

Let the $p \times p$ matrix, S, be the usual unbiased estimator of the population covariance matrix. Estimation of θ strives to minimize the discrepancies of the model implied covariance matrix $\Sigma(\theta)$ and observed covariance matrix S. By assuming multivariate normal distribution of y, a maximum likelihood discrepancy function can be derived (Bollen, 1989),

$$F_{ML} = \log|\Sigma(\theta)| + tr(S\Sigma^{-1}(\theta)) - \log|(S)| - p \tag{4.3}$$

The maximum likelihood estimate $\hat{\theta}$ for parameters can be estimated by minimizing this function. The fitness of the data to the model may be evaluated via the covariance hypothesis:

$$H_0 : \ S = \Sigma(\hat{\theta})$$

where $\Sigma(\hat{\theta})$ is the implied covariance matrix evaluated at the maximum likelihood estimate of θ. It can be shown that $(n\text{-}1)F_{ML}$ evaluated at $\hat{\theta}$ is approximately distributed as a chi-square variate with degrees of freedom $\frac{1}{2}p(p+1)-q$ (Bollen, 1989), where n is the number of observations in the data. For the chi-square test of the null hypothesis, a large p value indicates the failure to reject the null hypothesis that the model covariance matrix is equivalent to the sample covariance matrix. If a small p value is observed, this may indicate either a wrongly specified model, or a violation of the distribution assumptions.

4.2.2 Model Comparison in Structural Equation Models

The theory of p value is suitable to evaluate fitness of data with a specified model. However, there are many difficulties with such a strategy in comparing models, especially when both models have large p values from χ^2 goodness-of-fit-test. Comparison of different models can be achieved in the Bayesian formalism of the posterior probability of the model given the observed data.

$$p(M_k \mid D) = \frac{p(D \mid M_k)p(M_k)}{\sum_{k=1}^{K} p(D \mid M_k)p(M_k)} \tag{4.4}$$

$$p(D \mid M_k) = \int p(D \mid \theta_k,\ M_k)p(\theta_k \mid M_k)d\theta_k \tag{4.5}$$

Computation of the posterior probability involves multiple integrals and is not easy to achieve analytically. Bayesian Information Criterion (BIC) is usually used as the approximation to this posterior probability (Schwarz, 1978), which is defined as

following:

$$2\log p(M_k \mid D) \approx 2\log p(D \mid M_k, \hat{\theta}_k) - v_k \log n \qquad (4.6)$$

where $\hat{\theta}_k$ is the maximum likelihood estimate of the parameters, n is the number of samples, v_k is the number of parameters in the models. The first term is the log likelihood scoring with MLE, which usually increases when more parameters are present; the second term is a penalty function to restrict models containing a large number of parameters. Intuitively, BIC compares models by their fit to the data corrected/penalized to model complexity. The model that yields a smaller value of BIC is preferred. The use of BIC for model comparison and selection in SEM was suggested by Raftery (Raftery, 1993).

4.2.3 Biochemical Network Simulation

Evaluation of Model Fit of SEM in Biochemical Networks via Simulation

The asymptotic χ^2 goodness-of-fit-test is derived under the normal assumptions, linear relations and with comparatively large samples. How it will behave in nonlinear biochemical networks with relatively small sample sizes is not known. Here, we used artificial biochemical network data to evaluate the performance of a χ^2 goodness-of-fit-test for two simulated networks with different sample sizes, and different magnitudes of perturbations. The two biochemical network structures are illustrated in Fig. 4.1. One structure (model 1 in Fig. 4.1) represents a linear organization of six biochemical variables. Another structure (model 2 in Fig. 4.1) represents a fork configuration of six biochemical variables. Simulations of biochemical variables in networks are formulated in terms of ordinary differential equations (ODEs) and rely on phenomenological representations of reaction mechanisms. The rate of change for each biochemical

compound in the network is described by a synthesis term and a degradation term. A normally distributed random term is added in each equation, which represents unknown random perturbations to the biological system from unobserved perturbations such as uncontrolled environmental changes. Such random terms will propagate through the nonlinear system equations, and the distributions of the final simulated biochemical variable are not guaranteed to be normal.

The simulation equations for model 2 are illustrated below, and model 1 is simulated similarly but with different modifiers in the equation according to model 1's own structure:

$$\begin{aligned}
&\frac{dG1}{dt} = V_{G1} - k_{G1}G1 + \varepsilon_{G1}, \quad \frac{dG2}{dt} = \frac{V_{G2}}{1 + G1/K_{G1}} - k_{G2}G2 + \varepsilon_{G2} \\
&\frac{dG3}{dt} = V_{G3} - k_{G3}G3 + \varepsilon_{G3}, \quad \frac{dG4}{dt} = \frac{V_{G4}}{1 + G3/K_{G3}} - k_{G4}G4 + \varepsilon_{G4} \\
&\frac{dG5}{dt} = \frac{V_{G5}}{(1 + G2/K_{G2})(1 + G4/K_{G4})} - k_{G5}G5 + \varepsilon_{G5}, \\
&\frac{dG6}{dt} = \frac{V_{G6}}{1 + G5/K_{G5}} - k_{G6}G6 + \varepsilon_{G6}
\end{aligned} \tag{4.7}$$

All parameter values are set to unity. To use realistic perturbation terms, deterministic steady state values of the variables are simulated first without adding the random perturbations. Then different magnitudes of standard deviation of the normal distributed perturbations are decided in proportion (1%, 5%, 10%, 20% and 50% separately) to the pre-deterministic steady state values of each corresponding biochemical compound. The means of the perturbations are set to zero. Multiple steady state samples are obtained from such simulation. Different sample sizes, 50, 100, 200, 500 and 1000 samples are simulated from each model for all the perturbations. After the simulation runs, linear structure equation models are formed according to the simulated network structures. The simulated data are then fit to their corresponding structural equation models via a commercial statistical software package, SAS (SAS Institute, Cary, NC) using the

CALIS procedure. 500 replications of each combination of perturbation and sample size are run, and χ^2 values and *p*-values are recorded. The mean and variance of the 500 χ^2 values are obtained. The proportion of *p*-values <0.05 and <0.01 are recorded respectively.

Evaluation of Model Comparison of SEM in Biochemical Networks via Simulation

We simulated two sets of biochemical networks (Fig. 4.2). The networks within each set possess similar structures. Sets 1 (model 1.0, 1.1, 1.2 , and 1.3 in Fig. 4.2) are either the same as model 1 in Fig. 4.1 or with one link difference. Sets 2 (model 2.0, 2.1, 2.2 , and 2.3 in Fig. 4.2) are either the same as model 2 in Fig. 4.1 or with one link difference. Special motifs such as loop, local feedback and feed-forward are designed. Data are simulated as described above. The standard deviations of the perturbation terms are decided as the 10% of the predetermined steady state values. 500 samples are simulated from each structure. The simulated data from one model is fitted into all the structures via structure equation models. For example, the data simulated from network model 1.0 (Fig. 4.2) would fit into the SEM of not only model 1.0 but also model 1.1, 1.2, 1.3 plus the second sets of models (Fig. 4.2). The *p* values of χ^2 tests and BIC from each model are recorded for each fitted SEM.

4.3 Result

Test statistic results of models (Fig. 4.1) against the simulated data are summarized in Table 4.1. If the given test statistic were χ^2 distributed, the calculated sample mean and variance of the test statistics across the 500 replications in each condition should approximate the degrees of freedom (10 in these two cases) and twice of the degrees of freedom (20 in these two cases) respectively. From Table 4.1 and 4.2, we can see that χ^2 means of most samples at various conditions are close to 10. Positive bias of the χ^2

means tends to be more frequent than negative bias, although the absolute relative biases are usually smaller than 10%. On average, the sample variances at various conditions are a bit larger than 20 especially at larger perturbations such as 50%. It seemed that the moments of the sample test statistics are very close to the theoretical population parameters of the χ^2 distributions. As a test statistic, the distribution of tails is of more interest than the sample moments. The rejection frequencies of the model under each condition were recorded at the *p*-values of 0.05 and 0.01 respectively. In general, the statistics tend to over reject the model at these two *p* values. But the rejection percentages did not deviate too much from 0.05 and 0.01 at various conditions. These results are very close to the theoretical values, which indicate that the sample test statistics are very close to the χ^2 distributions. And this is true for sample sizes as small as 50 and perturbation standard deviations as large as 50% of the predetermined steady state value. At 50% perturbation, the simulation is already hard to converge. Further increase of the perturbation magnitude would hardly yield a steady state. Therefore, from extensive simulations across two different topology models possessing various perturbations and sample sizes, we find that the use of χ^2 statistics of SEM for testing the fitness of biochemical networks against a model is possible and informative.

The results of using SEM in comparisons of the two biochemical network sets are shown in Table 4.3. We can see that data simulated from model set 1 fit badly with model set 2, showing extremely low *p* values and large BIC (Table 4.3). The reverse is also true that data simulated from model set 2 fit poorly with model set 1 (Table 4.3). From this we conclude that if the model differs dramatically from the original model where the data come from, χ^2 fitness test of SEM can easily detect that the model is misspecified against the data. On the contrary if the correct model is used to evaluate fitness against the data, large *p* values and small BIC are observed (the shaded diagonal numbers in Table 4.3). So correct specified biochemical models fit the data well. At the same time, we can also compare how the biochemical networks with structures similar to the correct one fit the data. This is evaluated from both *p* values and BIC (Table 4.3).

For data simulated from model 1.0, all of the four models fit it well with *p* values larger than 0.01. However, comparisons of BIC would suggest model 1.0 the best-fitted model. For data simulated from model 1.1, only model 1.1 fit it well with *p* value larger than 0.01. For data from model 1.2, model 1.2 and 1.3 fit it nearly equally well at *p* values. It is also hard to differentiate these two models using a BIC comparison. For data from model 1.3, *p* value is informative enough to conclude that only model 1.3 fits well. Similar results can be found in the second set of simulated data (Table 4.3). For data simulated from model 2.0, all the four models fit well with *p* values larger than 0.01. BIC would correctly suggest that the model 2.0 is the most appropriate one. For data from model 2.1, only model 2.1 fits well with a *p* value larger than 0.01. For data from model 2.2, all four models fit well at *p* values. Comparison of BIC suggests that model 2.0, which is not the correct model in this case, is the most possible one among these four. Model 2.2 differs from model 2.0 at a local feedback from G5 to G2 (Fig. 4.2). It is hard for the global fitness criterion of SEM such as BIC to differentiate such local feedbacks. And model 2.2 is punished more than model 2.0 with one more parameter in the model. Thus BIC prefers a simple model of 2.0 rather than 2.2. Data from model 2.3, model 2.2 and 2.3 fit equally well with the same BIC.

These results show that if the data fit to a proposed model that differs significantly from true the model where the data are from, χ^2 fitness test of SEM can easily detect the model misspecification. If the proposed models are similar to the true model where the data are from, usually BIC can be used to find the correct model via model comparison. However, sometimes BIC cannot differentiate equally possible models. And it is also possible that we may obtain a close but a wrongly specified model if we solely depend on the smallest BIC for selecting the most possible model.

4.4 Discussion

This study represents an initial investigation of the use of SEM for examining both

biochemical network model fit and model comparisons using simulated biochemical network data. Current results from the simulations support the use of χ^2 fitness test for biochemical network evaluation and the use of BIC for biochemical network comparison. It seemed that the empirical distribution of the test statistics is close to the theoretic χ^2 distributions under various sample sizes, perturbation magnitudes and in the presence of nonlinear relations among the simulated biochemical variables. The results also indicate model comparison using BIC would usually select the correct model. However, it is also possible that a model similar to the true model possess smaller BIC than the true model. The true model could be among a set of models with relatively small BIC but not necessarily the smallest one. We suggest that BIC may be used as a criterion for model comparison and selection. However, selecting one "best" model is dangerous. Rather, it is recommended to select a set of models with similar small BIC values that fall within *Occam's window*, a generalization of the famous Occam's razor (the principle of parsimony), as people usually do in Bayesian model selection (Madigan and Raftery, 1994). Usually, the true model will be contained within this set. If the networks within the retained sets possess similar structures, a crude structure of the underlying real model can be constructed from their common links. Otherwise, more biological evidence may be needed for final determination of the structures.

Inferring and evaluating biochemical models from gargantuan quantities of genomic data is a daunting but important task in systems biology. Although many machine learning methods such as Bayesian networks, linear models and Boolean networks have been applied to real data, their efficiency and accuracy have not been validated. This is due to our lack of knowledge about the true biological systems. It is very hard to judge the results obtained from real data, thus the suitability of the methods remains controversial and uncertain. But the recent development of computational modeling of biochemical networks has provided a way to objectively evaluate these algorithms on simulated data (Kaern et al., 2003; Mendes et al., 2003). One requirement for such simulation is to mimic the reality as close as possible. In this study, we add a random

perturbation term in the deterministic differential equations to mimic random environmental perturbations to the systems. This random variable was sampled from a predetermined distribution and remained constant until the system reached steady state. We realized that this is not the optimal simulation mimicking the random environmental perturbations. The variable should vary within the entire time interval before the system reaches steady state. Such simulations require the use of stochastic differential equations and have been investigated recently (Oksendal, 1998). Further development of our study would be using such stochastic simulated data, and evaluating the efficiency of SEM for model fit and model comparison with more complex biochemical network structures and realistic parameter values.

Appendices 4

4A. Derivation of F_{ML}

In deriving F_{ML}, the set of N independent observations are of the multinormal random variables y. Assuming centered data, the probability density function of y is

$$f(y;\Sigma) = (2\pi)^{-p/2} \mid \Sigma \mid^{-1/2} \exp[(-\frac{1}{2}) y' \Sigma^{-1} y] \qquad (4A.1)$$

For a random sample of N independent observations of z, the joint density is,

$$f(y_1, y_2, ..., y_N; \Sigma) = f(y_1; \Sigma) f(y_2; \Sigma) ... f(y_N; \Sigma) \qquad (4A.2)$$

Once we observe a given sample, the likelihood function is

$$L(\theta) = (2\pi)^{-Np/2} \mid \Sigma(\theta) \mid^{-N/2} \exp[-\frac{1}{2} \sum_{i=1}^{N} y'_i \Sigma(\theta)^{-1} y_i] \qquad (4A.3)$$

Rewrite the last term in (4A.3) above as

$$\begin{aligned} -\frac{1}{2} \sum_{i=1}^{N} y'_i \Sigma(\theta)^{-1} y_i &= -\frac{1}{2} \sum_{i=1}^{N} tr[y'_i \Sigma(\theta)^{-1} y_i] \\ &= -(\frac{N}{2}) \sum_{i=1}^{N} tr[N^{-1} y_i y'_i \Sigma(\theta)^{-1}] \\ &= -(\frac{N}{2}) \sum_{i=1}^{N} tr[S^* \Sigma(\theta)^{-1}] \end{aligned} \qquad (4A.4)$$

where S^* is the sample ML estimator of the covariance matrix. This allows us to rewrite the log$L(\theta)$ as

$$\begin{aligned}\log L(\theta) &= \text{constant} - (\frac{N}{2})\log|\Sigma(\theta)| - (\frac{N}{2})tr(S^*\Sigma^{-1}(\theta)] \\ &= \text{constant} - (\frac{N}{2})\{\log|\Sigma(\theta)| + tr(S^*\Sigma^{-1}(\theta)]\}\end{aligned} \qquad (4A.5)$$

Compare (4A.5) to F_{ML}:

$$F_{ML} = \log|\Sigma(\theta)| + tr(S\Sigma^{-1}(\theta)) - \log(S) - p \qquad (4A.6)$$

The constant in log $L(\theta)$ and $-\log(S)-p$ in F_{ML} are constant terms given the data. They will not affect the choice of $\hat{\theta}$. The effect of $(-N/2)$ term present in (4A.5) but not in (4A.6) is to lead us to minimize (4A.6) at the same time maximize (4A.5) at $\hat{\theta}$. So the parameter that minimizes F_{ML} is the parameter that maximizes the likelihood.

4B. Likelihood ratio rationale for the asymptotic χ^2 distribution of $(N-1)F_{\text{ML}}$

The null hypothesis is that the population covariance matrix is the model implied covariance matrix, $H_0 : \Sigma = \Sigma(\theta)$

Under such hypothesis, we get the log of likelihood function as in (4A.5), here we use unbiased sample covariance matrix S instead of S^*, so the term N changed into N-1.

$$\log L_0(\theta) = \text{constant} - (\frac{N-1}{2})\{\log|\Sigma(\hat{\theta})| + tr(S\Sigma^{-1}(\hat{\theta})]\} \qquad (4B.1)$$

This is the log of the numerator for the likelihood ratio test.

The alternative hypothesis, H_1, is chosen for which the corresponding log of likelihood function $\log L_1$, is at maximum. The least restrictive H_1 possible is that Σ is any positive definite matrix. If $\hat{\Sigma}$ is set to S, the sample covariance matrix, the $\log L_1$ is at its

maximum value. The likelihood function for H1, $\log L_1$ is

$$\begin{aligned}\log L_1 &= \text{constant} - (\frac{N-1}{2})\{\log|S| + tr(S^{-1}S]\} \\ &= \text{constant} - (\frac{N-1}{2})\{\log|S| + p\}\end{aligned} \tag{4B.2}$$

So, the log likelihood ratio test is

$$\begin{aligned}-2\log\frac{L_0}{L_1} &= -2\log L_0 + 2\log L_1 \\ &= (N-1)[\log|\Sigma(\hat{\theta})| + tr(S\Sigma^{-1}(\hat{\theta}) - \log|S| - p] \\ &= (N-1)F_{ML}\end{aligned} \tag{4B.3}$$

So, $(N-1)F_{ML} \sim \chi^2$ with df, $\frac{1}{2}p(p+1) - q$.

Reference

Akutsu, T., Miyano, S., and Kuhara, S. (2000). Inferring qualitative relations in genetic networks and metabolic pathways. Bioinformatics *16*, 727-734.

Bing, N., and Hoeschele, I. (2004). Transcription network inference using natural, multigenetic perturbations. in preparation.

Bollen, K. A. (1989). Structural Equations with Latent Variables (New York, John Wiley).

D'Haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999). Linear modeling of mRNA expression levels during CNS development and injury. Pac Symp Biocomput, 41-52.

Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000). Using Bayesian networks to analyze expression data. J Comput Biol *7*, 601-620.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science *292*, 929-934.

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

Kaern, M., Blake, W. J., and Collins, J. J. (2003). The engineering of gene regulatory networks. Annu Rev Biomed Eng *5*, 179-206.

Madigan, D. M., and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's Window. Journal of the American Statistical Association *89*, 1335-1346.

Mendes, P. (1997). Biochemistry by numbers: simulation of biochemical pathways with Gepasi 3. Trends Biochem Sci *22*, 361-363.

Mendes, P., Sha, W., and Ye, K. (2003). Artificial gene networks for objective comparison of analysis algorithms. Bioinformatics *19*, II122-II129.

Oksendal, B. K. (1998). Stochastic Differential Equations, 5th edn (Berlin; New York, Springer).

Raftery, A. E. (1993). Bayesian model selection in structural equation models. In Testing structural equation models, K. A. Bollen, and J. S. Long, eds. (Newbury Park, CA, SAGE), pp. 163-180.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics *6*, 461-464.

Wagner, A. (2002). Estimating Coarse Gene Network Structure from Large-Scale Gene Perturbation Data. Genome Res *12*, 309-315.

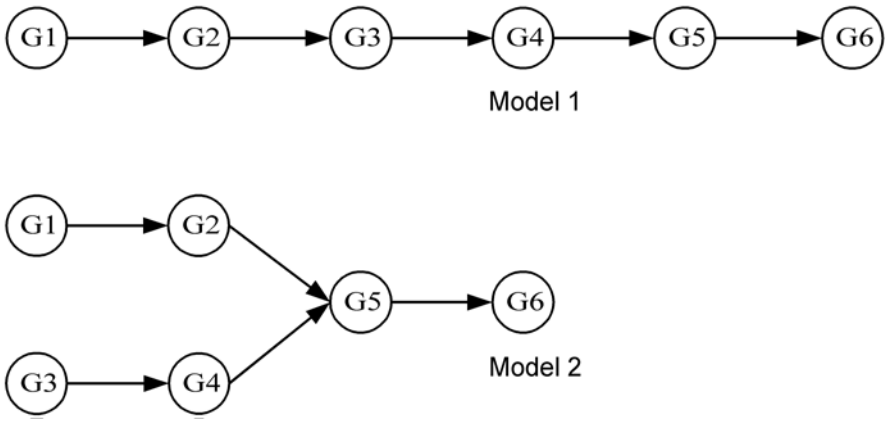

Fig 4. 1 Graphs of two artificial biochemical networks

Model 1 represents a linear structure between the 6 biochemical variables. Model 2 represents a fork structure between the 6 biochemical variables.

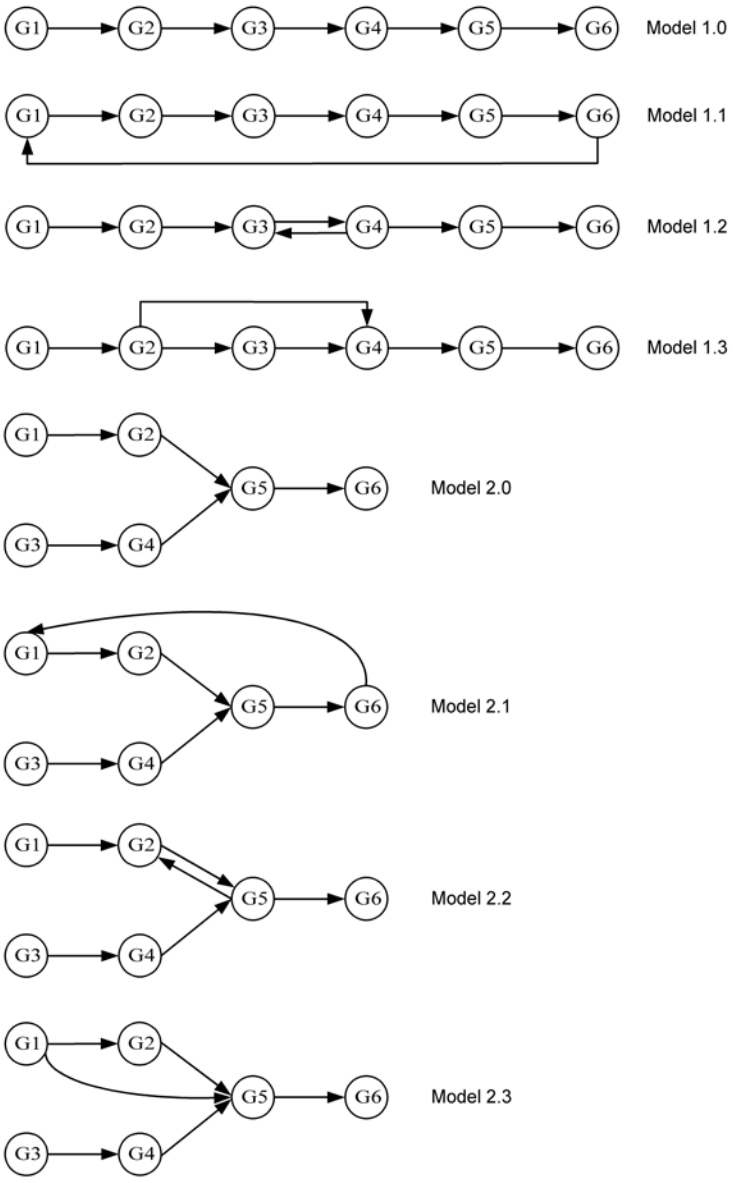

Fig 4. 2 Graphs of two sets of artificial biochemical networks

Models are similar within sets but different between sets. Model 1.1 differs from model 1.0 with a feedback from G6 to G1. Model 1.2 differs from model 1.0 with a local feedback from G4 to G3. Model 1.3 differs from model 1.0 with a feedforward structure from G2 to G4. Similar feedback, feedforward loops exist in model set two.

Table 4. 1 Summary of simulation results for model 1

Error Perturbation Percentage	Sample Size	Degree Freedom	Sample Chi Square Mean	Relative Bias	Sample Chi Square Variance	Percentage rejected at p<0.05	Percentage rejected at p<0.01
1%	50	10	9.8363	-1.64%	19.0549	0.044	0.004
	100	10	10.2064	2.06%	19.87	0.052	0.004
	200	10	10.0067	0.07%	21.3426	0.052	0.012
	500	10	9.8363	-1.64%	19.0549	0.044	0.004
	1000	10	10.0127	0.13%	18.8062	0.048	0.012
5%	50	10	10.5995	6.00%	22.1731	0.072	0.016
	100	10	10.311	3.11%	18.0806	0.044	0.006
	200	10	10.1731	1.73%	21.4464	0.050	0.012
	500	10	10.4298	4.30%	22.3654	0.068	0.014
	1000	10	10.3867	3.87%	20.0479	0.052	0.014
10%	50	10	10.8962	8.96%	25.4907	0.088	0.024
	100	10	10.0734	0.73%	20.3204	0.062	0.008
	200	10	10.0163	0.16%	19.9779	0.054	0.008
	500	10	10.0998	1.00%	19.2654	0.052	0.010
	1000	10	9.7856	-2.14%	17.7774	0.040	0.006
20%	50	10	10.6504	6.50%	25.805	0.076	0.024
	100	10	10.2945	2.94%	20.7171	0.062	0.006
	200	10	10.8128	8.13%	24.5714	0.072	0.022
	500	10	10.2861	2.86%	21.2791	0.060	0.014
	1000	10	10.1168	1.17%	22.3571	0.064	0.012
50%	50	10	10.7476	7.48%	24.2181	0.072	0.018
	100	10	10.4958	4.96%	25.7891	0.072	0.026
	200	10	9.9957	-0.04%	22.3188	0.062	0.012
	500	10	10.1537	1.54%	23.8762	0.082	0.014
	1000	10	10.0668	0.67%	21.5396	0.059	0.009

Table 4. 2 Summary of simulation results for model 2

Error Perturbation Percentage	Sample Size	Degree Freedom	Sample Chi Square Mean	Relative Bias	Sample Chi Square Variance	Percentage rejected at p<0.05	Percentage rejected at p<0.01
1%	50	10	10.5547	5.55%	20.7542	0.072	0.006
	100	10	10.7605	7.61%	23.74	0.072	0.030
	200	10	10.6732	6.73%	22.8069	0.072	0.014
	500	10	10.2269	2.27%	21.2474	0.052	0.012
	1000	10	9.9464	-0.54%	19.0538	0.044	0.004
5%	50	10	10.6023	6.02%	21.2028	0.056	0.012
	100	10	10.5165	5.17%	22.1543	0.064	0.008
	200	10	10.415	4.15%	21.4463	0.062	0.008
	500	10	10.0748	0.75%	21.648	0.070	0.012
	1000	10	9.6933	-3.07%	20.310	0.048	0.006
10%	50	10	10.9433	9.43%	24.5222	0.094	0.016
	100	10	10.482	4.82%	21.4191	0.074	0.010
	200	10	10.1897	1.90%	22.1338	0.056	0.010
	500	10	9.8587	-1.41%	18.7185	0.044	0.004
	1000	10	10.1168	1.17%	21.2754	0.052	0.016
20%	50	10	11.1294	11.29%	21.3139	0.058	0.018
	100	10	10.6083	6.08%	20.4179	0.062	0.014
	200	10	10.5287	5.29%	22.456	0.056	0.012
	500	10	10.702	7.02%	24.2506	0.084	0.022
	1000	10	10.7492	7.49%	24.8526	0.066	0.020
50%	50	10	10.5024	5.02%	22.0927	0.062	0.018
	100	10	10.4136	4.14%	21.7362	0.070	0.010
	200	10	10.3442	3.44%	24.12	0.062	0.020
	500	10	10.4036	4.04%	24.1311	0.074	0.016
	1000	10	10.8373	8.37%	26.019	0.078	0.022

Table 4. 3 Model comparison

Model	Criterion	Data1.0	Data1.1	Data1.2	Data1.3	Data2.0	Data2.1	Data2.2	Data2.3
1.0	Pr > Chi-Square	0.8992	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	BIC	-57.2690	-26.4851	62.1707	59.9705	148.6414	200.8225	149.5986	312.7565
1.1	Pr > Chi-Square	0.9271	0.1671	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
	BIC	-52.1825	-43.0294	39.6782	23.8531	150.8236	200.7700	150.1991	278.0073
1.2	Pr > Chi-Square	0.8617	0.0017	0.1145	<.0001	<.0001	<.0001	<.0001	<.0001
	BIC	-51.2575	-29.4807	-39.6977	42.7120	111.1740	131.8793	104.8157	36.9778
1.3	Pr > Chi-Square	0.8644	0.0098	0.1077	0.1572	<.0001	<.0001	<.0001	<.0001
	BIC	-51.2898	-34.2085	-39.4895	-42.8102	153.5258	197.0652	155.1600	317.8868
2.0	Pr > Chi-Square	<.0001	<.0001	<.0001	<.0001	0.1017	0.0060	0.5371	<.0001
	BIC	1753.8900	182.4910	303.3847	274.4925	-46.2191	-37.4510	-53.1989	120.6378
2.1	Pr > Chi-Square	<.0001	<.0001	<.0001	<.0001	0.1491	0.5662	0.4461	<.0001
	BIC	1740.7581	176.3315	302.1483	189.5447	-42.6217	-48.2464	-47.0269	116.4318
2.2	Pr > Chi-Square	<.0001	<.0001	<.0001	<.0001	0.0689	0.0038	0.6887	0.3126
	BIC	1760.0434	188.0780	307.9702	280.7040	-40.0251	-31.5915	-49.4283	-45.4459
2.3	Pr > Chi-Square	<.0001	<.0001	<.0001	<.0001	0.0689	0.0038	0.6887	0.3126
	BIC	1760.0434	188.0727	307.9702	280.7040	-40.0251	-31.5915	-49.4283	-45.4459

Supplement Data

The data contains expressed gene, the corresponding causal gene via significant correlation test, their spearman correlation, causal gene genomic locations, nearest corresponding fine mapped QTL sub-intervals, and their location distances.

ArrayGene	CausalGene	SpCor	CGChrom	CGBegin	CGend	QTLSIBegin	QTLSIend	DisDiff
YAR027W	YAR027W	1	1	183764	184471	184243	187276	0
YAR028W	YAR028W	1	1	184886	185590	184243	187276	0
YAR033W	YAR033W	1	1	188101	188805	187402	188546	0
YAR033W	YGL051W	0.888743	7	403688	404392	402833	427476	0
YBL038W	YNL081C	0.896811	14	476185	476616	449639	502316	0
YBL044W	YBL044W	1	2	135960	136328	133749	151686	0
YBL090W	YNL081C	0.908443	14	476185	476616	449639	502316	0
YBL113C	YLR446W	0.542214	12	1025209	1026510	1023790	1042072	0
YBR017C	YNL093W	-0.702627	14	449865	450527	418293	486861	0
YBR017C	YNL088W	0.58424	14	457701	461987	418293	486861	0
YBR017C	YNL086W	-0.710882	14	466331	466639	418293	486861	0
YBR017C	YNL085W	0.792495	14	467128	469620	418293	486861	0
YBR017C	YNL081C	-0.638462	14	476185	476616	418293	486861	0
YBR017C	YNL078W	0.679362	14	479765	480988	418293	486861	0
YBR037C	YIL124W	0.56454	9	126204	127097	98949	133663	0
YBR066C	YBR066C	1	2	369997	370659	352257	380938	0
YBR120C	YNL081C	0.914259	14	476185	476616	449639	502316	0
YBR122C	YNL073W	0.576923	14	488383	490113	486861	502316	0
YBR127C	YNL093W	-0.612383	14	449865	450527	418293	486861	0
YBR127C	YNL086W	-0.647842	14	466331	466639	418293	486861	0
YBR127C	YNL085W	0.703377	14	467128	469620	418293	486861	0
YBR129C	YNL137C	0.663227	14	368592	370052	368178	393903	0
YBR132C	YBR132C	1	2	499609	501399	499012	521415	0
YBR146W	YNL073W	0.736773	14	488383	490113	449639	502316	0
YBR166C	YOL164W	0.63227	15	6175	8115	1152	16691	0
YBR170C	YBR170C	1	2	576301	578043	573407	582419	0
YBR197C	YBR197C	1	2	615160	615813	610333	624500	0
YBR262C	YNL086W	0.758161	14	466331	466639	449639	502316	0
YBR262C	YNL081C	0.854034	14	476185	476616	449639	502316	0
YCL018W	YCL018W	1	3	91323	92417	92247	92391	0
YCL065W	YCR039C	0.864916	3	199538	200170	175802	201167	0

YCL065W	YCR040W	0.939963	3	200434	200961	175802	201167	0
YCL065W	YCR041W	0.928143	3	200903	201235	175802	201167	0
YCL066W	YCR039C	0.852158	3	199538	200170	175802	201167	0
YCL066W	YCR040W	0.931895	3	200434	200961	175802	201167	0
YCL066W	YCR041W	0.906942	3	200903	201235	175802	201167	0
YCL067C	YCR039C	0.964353	3	199538	200170	177850	201167	0
YCL069W	YKR102W	0.608068	11	645985	649494	643655	650334	0
YCL073C	YLR172C	-0.633583	12	501262	502164	500589	508029	0
YCR003W	YNL081C	0.866229	14	476185	476616	449639	502316	0
YCR024C	YNL093W	0.639024	14	449865	450527	449639	502316	0
YCR024C	YNL081C	0.63546	14	476185	476616	449639	502316	0
YCR024C	YNL073W	0.680488	14	488383	490113	449639	502316	0
YCR040W	YCR040W	1	3	200434	200961	175802	201167	0
YCR041W	YCR041W	1	3	200903	201235	201166	201167	0
YCR096C	YCR039C	0.932833	3	199538	200170	177850	201167	0
YCR097W	YCR038C	0.650094	3	197613	199541	175802	209932	0
YCR097W	YCR040W	-0.750469	3	200434	200961	175802	209932	0
YCR097W	YCR041W	-0.77955	3	200903	201235	175802	209932	0
YCR097W-A	YCR041W	-0.75272	3	200903	201235	201166	209932	0
YDL055C	YBR132C	0.746154	2	499609	501399	489202	499889	0
YDL168W	YDL168W	1	4	159605	160765	154436	167440	0
YDL227C	YDL227C	1	4	46272	48032	33214	54228	0
YDL231C	YDL231C	1	4	38868	42245	33214	54228	0
YDR005C	YAR047C	-0.599719	1	201459	201779	193251	201472	0
YDR038C	YDR038C	1	4	527416	530691	527508	527875	0
YDR041W	YNL137C	0.829268	14	368592	370052	368178	393903	0
YDR119W	YBR137W	-0.637899	2	513001	513540	499012	521415	0
YDR175C	YNL081C	0.875797	14	476185	476616	449639	502316	0
YDR177W	YDR177W	1	4	816871	817518	802851	828741	0
YDR197W	YNL137C	0.774296	14	368592	370052	368178	393903	0
YDR237W	YNL081C	0.856848	14	476185	476616	449639	502316	0
YDR261C	YNL086W	-0.645028	14	466331	466639	449639	502316	0
YDR261C	YNL081C	-0.649156	14	476185	476616	449639	502316	0
YDR322W	YNL081C	0.749156	14	476185	476616	449639	502316	0
YDR337W	YNL073W	0.657786	14	488383	490113	486861	502316	0
YDR347W	YNL081C	0.725328	14	476185	476616	449639	502316	0
YDR367W	YDR367W	1	4	1212836	1213602	1213416	1220683	0
YDR375C	YNL081C	0.825891	14	476185	476616	449639	502316	0
YDR423C	YDR423C	1	4	1318034	1319263	1310380	1321550	0
YDR441C	YDR441C	1	4	1344505	1345050	1344550	1344670	0
YDR451C	YDR451C	1	4	1361108	1362169	1344550	1364751	0

YDR453C	YOL084W	0.805253	15	162355	165330	154309	174364	0
YDR461W	YCR039C	-0.682176	3	199538	200170	175802	201167	0
YDR461W	YCR040W	-0.729456	3	200434	200961	175802	201167	0
YDR461W	YCR041W	-0.685366	3	200903	201235	175802	201167	0
YDR492W	YGL052W	-0.628518	7	403438	403743	402833	415585	0
YDR492W	YGL051W	-0.603377	7	403688	404392	402833	415585	0
YDR492W	YGL049C	-0.635835	7	406861	409605	402833	415585	0
YDR493W	YNL086W	0.675797	14	466331	466639	449639	502316	0
YDR493W	YNL081C	0.815385	14	476185	476616	449639	502316	0
YDR494W	YNL071W	-0.60788	14	491520	492968	486861	502496	0
YDR511W	YNL086W	0.626829	14	466331	466639	449639	486861	0
YDR511W	YNL081C	0.757786	14	476185	476616	449639	486861	0
YDR539W	YDR539W	1	4	1512078	1513589	1510883	1525328	0
YEL021W	YEL021W	1	5	116167	116970	116812	117705	0
YEL050C	YNL081C	0.718386	14	476185	476616	449639	502316	0
YEL052W	YEL021W	-0.673265	5	116167	116970	109310	117705	0
YEL057C	YEL057C	1	5	45020	45721	44611	48845	0
YEL073C	YEL073C	1	5	7230	7553	6334	15817	0
YER047C	YER047C	1	5	243809	246502	243215	244117	0
YER058W	YNL071W	-0.579174	14	491520	492968	486861	502316	0
YER187W	YER188W	-0.519418	5	568035	568754	568668	568734	0
YER190W	YLR446W	0.583114	12	1025209	1026510	1023795	1042072	0
YFL025C	YCR041W	-0.766792	3	200903	201235	201166	201167	0
YFL026W	YCR039C	-0.802627	3	199538	200170	175802	201166	0
YFL026W	YCR040W	-0.793058	3	200434	200961	175802	201166	0
YFL026W	YCR041W	-0.767542	3	200903	201235	175802	201166	0
YFL027C	YCR041W	-0.68349	3	200903	201235	201166	209932	0
YFL052W	YJL212C	0.64878	10	33850	36249	24745	34098	0
YFL055W	YJL212C	0.606191	10	33850	36249	24745	34098	0
YFR013W	YMR206W	-0.634709	13	675895	676836	667328	686206	0
YGL028C	YBR132C	0.642964	2	499609	501399	499012	521415	0
YGL028C	YBR148W	-0.740901	2	537833	539662	530481	555787	0
YGL032C	YCR041W	-0.729456	3	200903	201235	201166	209932	0
YGL049C	YGL049C	1	7	406861	409605	402833	427476	0
YGL051W	YGL051W	1	7	403688	404392	403924	427476	0
YGL051W	YAR033W	0.888743	1	188101	188805	187544	188546	0
YGL052W	YGL051W	0.652158	7	403688	404392	403925	427476	0
YGL053W	YGL053W	1	7	402590	403303	402833	427476	0
YGL064C	YHR038W	0.582364	8	184057	184749	176994	188671	0
YGL089C	YCR041W	0.899812	3	200903	201235	201166	209932	0
YGL090W	YCR041W	0.932833	3	200903	201235	201166	201167	0

YGL107C	YNL081C	0.745403	14	476185	476616	449639	502316	0
YGL125W	YCL018W	-0.628518	3	91323	92417	79091	92391	0
YGL129C	YNL081C	0.72758	14	476185	476616	449639	502316	0
YGL129C	YNL073W	0.673546	14	488383	490113	449639	502316	0
YGL143C	YNL081C	0.815197	14	476185	476616	449639	502316	0
YGL143C	YNL073W	0.634897	14	488383	490113	449639	502316	0
YGL169W	YGL169W	1	7	186063	187343	187185	192140	0
YGL195W	YNL093W	-0.645779	14	449865	450527	449639	502316	0
YGL195W	YNL086W	-0.731895	14	466331	466639	449639	502316	0
YGL195W	YNL085W	0.838086	14	467128	469620	449639	502316	0
YGL195W	YNL078W	0.669043	14	479765	480988	449639	502316	0
YGL201C	YGL201C	1	7	117856	120909	110813	117900	0
YGR076C	YNL137C	0.793621	14	368592	370052	368178	393903	0
YGR219W	YNL081C	0.814822	14	476185	476616	449639	502316	0
YGR219W	YNL071W	-0.605816	14	491520	492968	449639	502316	0
YGR220C	YNL137C	0.82833	14	368592	370052	368178	393903	0
YGR221C	YCL026C-A	-0.714822	3	74704	75285	75021	76127	0
YGR234W	YLR260W	0.565103	12	665844	667907	663370	668249	0
YGR296W	YLR446W	0.521764	12	1025209	1026510	1023790	1042072	0
YHL003C	YHL003C	1	8	100642	101877	80014	111690	0
YHL009C	YHL009C	1	8	84063	85055	80014	111690	0
YHL010C	YHL010C	1	8	81959	83716	80014	98519	0
YHL022C	YHL022C	1	8	62958	64154	56246	63386	0
YHL047C	YHL044W	0.815572	8	13563	14270	13926	17636	0
YHR015W	YHR015W	1	8	134545	136524	128732	137233	0
YHR028C	YHR028C	1	8	164969	167425	161987	188851	0
YHR031C	YAR031W	-0.484615	1	186830	187726	187544	188546	0
YHR036W	YLR389C	0.655535	12	899693	902659	899898	909226	0
YHR038W	YNL137C	0.814071	14	368592	370052	368178	393903	0
YHR058C	YHR058C	1	8	218998	219885	213539	221933	0
YHR059W	YNL081C	0.795497	14	476185	476616	449639	502316	0
YHR091C	YNL081C	0.598874	14	476185	476616	449639	502316	0
YHR143W	YBR132C	0.623827	2	499609	501399	499889	521415	0
YHR143W	YBR148W	-0.703189	2	537833	539662	530481	555787	0
YHR143W	YBR150C	-0.605066	2	541166	544450	530481	555787	0
YHR147C	YNL137C	0.820263	14	368592	370052	368178	393903	0
YIL010W	YIL010W	1	9	334879	335526	323975	341216	0
YIL014W	YIL014W	1	9	326101	327993	325242	341216	0
YIL015W	YCR039C	-0.761351	3	199538	200170	175802	209932	0
YIL015W	YCR040W	-0.837523	3	200434	200961	175802	209932	0
YIL015W	YCR041W	-0.782176	3	200903	201235	175802	209932	0

YIL061C	YIL061C	1	9	244654	245556	242417	246551	0
YIL070C	YNL081C	0.80319	14	476185	476616	449639	502316	0
YIL089W	YIL089W	1	9	195596	196213	190794	196145	0
YIL093C	YNL081C	0.810882	14	476185	476616	449639	502316	0
YIL093C	YNL071W	-0.641276	14	491520	492968	449639	502316	0
YIL098C	YNL086W	0.618574	14	466331	466639	449639	502316	0
YIL098C	YNL081C	0.762852	14	476185	476616	449639	502316	0
YIL136W	YIL136W	1	9	93619	94800	88155	98949	0
YIL166C	YIL166C	1	9	30938	32566	27128	31216	0
YIR031C	YIR031C	1	9	413012	414676	410028	415311	0
YIR038C	YOR185C	0.704878	15	681444	682106	679086	690475	0
YJL035C	YJL035C	1	10	379945	380697	372838	380085	0
YJL046W	YNL086W	0.571857	14	466331	466639	449639	502316	0
YJL046W	YNL081C	0.783302	14	476185	476616	449639	502316	0
YJL046W	YNL076W	0.582364	14	483553	485307	449639	502316	0
YJL046W	YNL071W	-0.639775	14	491520	492968	449639	502316	0
YJL051W	YJL051W	1	10	339483	341951	341703	353027	0
YJL057C	YJL057C	1	10	327814	329817	325500	336317	0
YJL063C	YNL081C	0.827392	14	476185	476616	449639	502316	0
YJL063C	YNL073W	0.649906	14	488383	490113	449639	502316	0
YJL096W	YNL137C	0.843152	14	368592	370052	368178	393903	0
YJL116C	YOL084W	0.669418	15	162355	165330	154309	174364	0
YJL130C	YEL021W	-0.604221	5	116167	116970	109310	117705	0
YJL170C	YCR041W	-0.757974	3	200903	201235	201166	201167	0
YJR004C	YCR041W	0.893058	3	200903	201235	201166	209932	0
YJR015W	YJR015W	1	10	462414	463946	461201	464261	0
YJR101W	YNL137C	0.901126	14	368592	370052	368178	393903	0
YJR147W	YBR137W	-0.622889	2	513001	513540	513408	514035	0
YJR147W	YBR149W	-0.66379	2	539944	540978	530481	555778	0
YJR162C	YER188W	0.618387	5	568035	568754	568425	568734	0
YKL053C-A	YNL081C	0.889493	14	476185	476616	449639	502316	0
YKL127W	YNL081C	-0.600563	14	476185	476616	418293	486861	0
YKL138C	YNL137C	0.870169	14	368592	370052	368178	393903	0
YKL148C	YKL109W	0.560037	11	231869	233533	229052	238310	0
YKL167C	YNL137C	0.654034	14	368592	370052	368178	393903	0
YKL169C	YNL137C	0.830582	14	368592	370052	368178	393903	0
YKL170W	YNL081C	0.911257	14	476185	476616	449639	502316	0
YKL177W	YCR039C	0.67167	3	199538	200170	175802	209932	0
YKL177W	YCR040W	0.802627	3	200434	200961	175802	209932	0
YKL177W	YCR041W	0.814822	3	200903	201235	175802	209932	0
YKL178C	YCR039C	0.779737	3	199538	200170	175802	209932	0

YKL178C	YCR040W	0.870919	3	200434	200961	175802	209932	0
YKL178C	YCR041W	0.872608	3	200903	201235	175802	209932	0
YKL187C	YKL187C	1	11	89285	91537	85837	97761	0
YKL209C	YCR039C	-0.724578	3	199538	200170	175799	201166	0
YKL209C	YCR040W	-0.78743	3	200434	200961	175799	201166	0
YKL209C	YCR041W	-0.787617	3	200903	201235	175799	201166	0
YKL216W	YEL021W	-0.573265	5	116167	116970	109310	117705	0
YKL225W	YER188W	0.665478	5	568035	568754	568425	568734	0
YKR049C	YOL094C	-0.629081	15	141583	142554	141621	180180	0
YKR049C	YOL084W	0.776923	15	162355	165330	141621	180180	0
YKR062W	YKR062W	1	11	559302	560288	527129	566015	0
YKR071C	YCL018W	-0.515572	3	91323	92417	79091	91324	0
YKR087C	YKR087C	1	11	602831	603775	596152	612781	0
YKR102W	YKR102W	1	11	645985	649494	643655	649174	0
YKR103W	YKR102W	0.780488	11	645985	649494	643655	650334	0
YKR104W	YKR104W	1	11	656465	657385	657299	666052	0
YKR106W	YKR102W	0.636398	11	645985	649494	643655	650800	0
YLL007C	YLL007C	1	12	134301	136298	131338	145786	0
YLL010C	YLL010C	1	12	129329	130612	126934	145780	0
YLL013C	YLL013C	1	12	122074	124713	112275	126934	0
YLL057C	YLL057C	1	12	25756	26994	19609	26184	0
YLR040C	YCR039C	0.786867	3	199538	200170	175802	209932	0
YLR040C	YCR040W	0.876173	3	200434	200961	175802	209932	0
YLR040C	YCR041W	0.884991	3	200903	201235	175802	209932	0
YLR041W	YCR039C	0.689681	3	199538	200170	175802	201166	0
YLR041W	YCR040W	0.808255	3	200434	200961	175802	201166	0
YLR041W	YCR041W	0.798124	3	200903	201235	175802	201166	0
YLR050C	YLR050C	1	12	245588	246073	238376	248307	0
YLR073C	YLR073C	1	12	281020	281622	274240	282397	0
YLR155C	YLR172C	-0.585178	12	501262	502164	501510	508029	0
YLR156W	YLR172C	-0.657411	12	501262	502164	500589	508029	0
YLR157C	YLR169W	0.582552	12	500337	500690	500589	508029	0
YLR157C	YLR172C	-0.703189	12	501262	502164	500589	508029	0
YLR158C	YLR169W	0.569043	12	500337	500690	500493	508029	0
YLR158C	YLR172C	-0.62045	12	501262	502164	500493	508029	0
YLR159W	YLR172C	-0.669794	12	501262	502164	500589	508029	0
YLR160C	YLR172C	-0.686867	12	501262	502164	500589	508029	0
YLR161W	YLR172C	-0.633959	12	501262	502164	489760	508029	0
YLR204W	YNL081C	0.910694	14	476185	476616	449639	502316	0
YLR239C	YNL086W	0.613696	14	466331	466639	449639	486861	0
YLR239C	YNL081C	0.697186	14	476185	476616	449639	486861	0

YLR285W	YBR147W	-0.646154	2	536532	537422	530481	537314	0
YLR312W-A	YNL071W	-0.634709	14	491520	492968	486861	502316	0
YLR353W	YLR353W	1	12	834351	836162	829705	849446	0
YLR389C	YLR389C	1	12	899693	902659	899898	909226	0
YLR420W	YEL021W	-0.590338	5	116167	116970	109310	117705	0
YLR452C	YHL003C	0.614259	8	100642	101877	80014	111679	0
YML046W	YHL003C	0.593996	8	100642	101877	98519	111680	0
YML050W	YNL137C	0.766604	14	368592	370052	368178	371953	0
YML066C	YML066C	1	13	140424	141533	124876	144766	0
YMR024W	YNL073W	0.696998	14	488383	490113	449639	502316	0
YMR064W	YAR047C	-0.527298	1	201459	201779	199760	201472	0
YMR096W	YCL018W	-0.64803	3	91323	92417	79091	92391	0
YMR108W	YCL018W	-0.502251	3	91323	92417	79091	92157	0
YMR157C	YNL137C	0.798311	14	368592	370052	368178	393903	0
YMR158W	YNL081C	0.888368	14	476185	476616	449639	502316	0
YMR171C	YMR171C	1	13	603867	605519	597711	649250	0
YMR220W	YNR043W	0.627204	14	701892	703082	695114	705553	0
YMR292W	YGL020C	0.706191	7	457166	457873	457215	460945	0
YNL005C	YNL081C	0.826454	14	476185	476616	449639	502316	0
YNL040W	YNL019C	-0.516698	14	598372	599226	591234	602405	0
YNL066W	YBR132C	0.679174	2	499609	501399	499012	521415	0
YNL066W	YBR138C	-0.578987	2	513719	515293	499012	521415	0
YNL066W	YBR148W	-0.661351	2	537833	539662	530481	555787	0
YNL137C	YNL073W	0.697749	14	488383	490113	486861	502316	0
YNL145W	YCR039C	-0.625516	3	199538	200170	175802	201166	0
YNL145W	YCR040W	-0.689306	3	200434	200961	175802	201166	0
YNL145W	YCR041W	-0.653096	3	200903	201235	175802	201166	0
YNL177C	YNL081C	0.837148	14	476185	476616	449639	502316	0
YNL185C	YNL081C	0.675234	14	476185	476616	449639	502316	0
YNL185C	YNL073W	0.642026	14	488383	490113	449639	502316	0
YNL202W	YNL202W	1	14	259566	260453	254156	263964	0
YNL208W	YBR147W	0.587805	2	536532	537422	530481	555787	0
YNL208W	YBR149W	0.741463	2	539944	540978	530481	555787	0
YNL208W	YBR155W	-0.694747	2	549728	550885	530481	555787	0
YNL254C	YGL256W	0.803377	7	14910	16307	15891	21177	0
YNL284C	YNL081C	0.806191	14	476185	476616	449639	502316	0
YNL327W	YBR132C	0.658349	2	499609	501399	499012	521415	0
YNL327W	YBR148W	-0.622514	2	537833	539662	530481	555778	0
YNR040W	YNL073W	0.667917	14	488383	490113	486861	502316	0
YNR044W	YHL003C	0.64409	8	100642	101877	80014	111690	0
YNR067C	YBR148W	-0.697936	2	537833	539662	530481	555778	0

YOL014W	YOL014W	1	15	299693	300067	290670	301077	0
YOL023W	YNL086W	0.564916	14	466331	466639	449639	486861	0
YOL023W	YNL081C	0.759099	14	476185	476616	449639	486861	0
YOL023W	YNL076W	0.566041	14	483553	485307	449639	486861	0
YOL089C	YOL089C	1	15	150397	153489	132423	154309	0
YOL091W	YOL091W	1	15	145333	147162	141621	154309	0
YOL114C	YEL021W	-0.612477	5	116167	116970	109310	117705	0
YOL126C	YEL021W	-0.744371	5	116167	116970	116812	117705	0
YOL143C	YEL021W	-0.67833	5	116167	116970	116812	117705	0
YOL159C	YIL166C	0.587992	9	30938	32566	27026	38608	0
YOL162W	YOL162W	1	15	10118	10765	1152	17842	0
YOR019W	YOR019W	1	15	368126	370318	338018	370304	0
YOR150W	YNL081C	0.810131	14	476185	476616	449639	502316	0
YOR150W	YNL070W	0.666604	14	493363	493545	449639	502316	0
YOR225W	YCL018W	-0.546529	3	91323	92417	81843	91496	0
YOR263C	YBR147W	-0.609193	2	536532	537422	530481	537314	0
YOR354C	YNL081C	0.848968	14	476185	476616	449639	502316	0
YPL016W	YPL016W	1	16	521009	524953	523102	523450	0
YPL097W	YNL086W	0.585553	14	466331	466639	449639	502316	0
YPL097W	YNL081C	0.802439	14	476185	476616	449639	502316	0
YPL184C	YNL071W	0.759475	14	491520	492968	486861	525061	0
YPL187W	YCR040W	0.846341	3	200434	200961	175802	209932	0
YPL187W	YCR041W	0.850469	3	200903	201235	175802	209932	0
YPR085C	YPR085C	1	16	708493	709839	704388	711614	0
YPR099C	YNL137C	0.82364	14	368592	370052	368178	393903	0
YPR106W	YBR148W	-0.812383	2	537833	539662	530481	555787	0
YPR128C	YAL054C	-0.533583	1	42881	45022	42591	55329	0
YPR199C	YKR104W	-0.701876	11	656465	657385	657329	666052	0
YNL242W	YNL242W	1	14	191323	196101	191240	191243	80
YLR179C	YLR179C	1	12	514110	514715	514835	516242	120
YGL064C	YHR028C	-0.658349	8	164969	167425	167566	175255	141
YHR033W	YHR033W	1	8	175539	176810	176994	188671	184
YCR097W-A	YCR040W	-0.742026	3	200434	200961	201166	209932	205
YFL025C	YCR040W	-0.811257	3	200434	200961	201166	201167	205
YFL027C	YCR040W	-0.685366	3	200434	200961	201166	209932	205
YGL032C	YCR040W	-0.712946	3	200434	200961	201166	209932	205
YGL089C	YCR040W	0.86848	3	200434	200961	201166	209932	205
YGL090W	YCR040W	0.918387	3	200434	200961	201166	201167	205
YJL170C	YCR040W	-0.782176	3	200434	200961	201166	201167	205
YJR004C	YCR040W	0.902814	3	200434	200961	201166	209932	205
YAL056W	YAL056W	1	1	39260	41803	38491	38887	373

YCL017C	YCL017C	1	3	92776	94269	79091	92391	385
YER073W	YCL017C	0.690807	3	92776	94269	79091	92391	385
YGL009C	YCL017C	0.715947	3	92776	94269	92241	92391	385
YGL125W	YCL017C	0.690056	3	92776	94269	79091	92391	385
YHR208W	YCL017C	0.74409	3	92776	94269	92241	92391	385
YJR016C	YCL017C	0.710131	3	92776	94269	79091	92391	385
YKL120W	YCL017C	0.666792	3	92776	94269	92247	92391	385
YMR096W	YCL017C	0.653283	3	92776	94269	79091	92391	385
YNL104C	YCL017C	0.759662	3	92776	94269	79091	92391	385
YOR108W	YCL017C	0.803002	3	92776	94269	79091	92391	385
YBR166C	YOL149W	0.630769	15	44936	45631	44542	44548	388
YIL082W	YIL082W	1	9	205632	206504	197948	205191	441
YIL082W-A	YIL082W-A	1	9	205632	210129	197948	205191	441
YBR148W	YBR148W	1	2	537833	539662	517365	537314	519
YDR119W	YBR148W	-0.665103	2	537833	539662	530481	537314	519
YOR263C	YBR148W	-0.618199	2	537833	539662	530481	537314	519
YLR343W	YLR343W	1	12	816094	817761	808707	815552	542
YDR390C	YLR369W	0.635647	12	859551	861524	855389	858996	555
YNL329C	YNL329C	1	14	19541	22633	23220	28788	587
YBR037C	YIL134W	0.570732	9	97395	98330	98949	133663	619
YDR384C	YMR146C	0.584803	13	557480	558523	556726	556835	645
YGR279C	YGR279C	1	7	1048802	1049962	1007545	1048152	650
YLR303W	YCL018W	-0.627955	3	91323	92417	81843	90610	713
YER124C	YBR158W	0.80319	2	556505	558154	548401	555787	718
YGL028C	YBR158W	0.833396	2	556505	558154	530481	555787	718
YGR041W	YBR158W	0.8394	2	556505	558154	548401	555787	718
YHR143W	YBR158W	0.784803	2	556505	558154	530481	555787	718
YJL078C	YBR158W	0.818386	2	556505	558154	548401	555787	718
YNL066W	YBR158W	0.776548	2	556505	558154	530481	555787	718
YOR264W	YBR158W	0.704878	2	556505	558154	548401	555787	718
YPR106W	YBR158W	0.830582	2	556505	558154	530481	555787	718
YJR147W	YBR158W	0.587617	2	556505	558154	530481	555778	727
YNL327W	YBR158W	0.718011	2	556505	558154	530481	555778	727
YNR067C	YBR158W	0.85666	2	556505	558154	530481	555778	727
YIL120W	YIL120W	1	9	134414	136105	98949	133663	751
YFR030W	YCL018W	-0.594747	3	91323	92417	81832	90412	911
YGL010W	YCL018W	-0.665478	3	91323	92417	81832	90412	911
YPL135W	YCL018W	-0.676173	3	91323	92417	81832	90412	911
YPR167C	YCL018W	-0.590244	3	91323	92417	81832	90412	911
YCR039C	YCR039C	1	3	199538	200170	201166	201167	996
YCR097W-A	YCR039C	-0.581051	3	199538	200170	201166	209932	996

YFL025C	YCR039C	-0.832645	3	199538	200170	201166	201167	996
YGL032C	YCR039C	-0.648405	3	199538	200170	201166	209932	996
YGL089C	YCR039C	0.774296	3	199538	200170	201166	209932	996
YJL170C	YCR039C	-0.703002	3	199538	200170	201166	201167	996
YGR205W	YGR205W	1	7	909215	910087	911217	916675	1130
YGL178W	YGL178W	1	7	167356	170575	171769	181140	1194
YML055W	YML055W	1	13	164790	165326	158910	163328	1462
YDL178W	YNL076W	0.638649	14	483553	485307	486861	502496	1554
YCR038C	YCR038C	1	3	197613	199541	201166	209932	1625
YMR173W-A	YCR038C	0.593246	3	197613	199541	201166	201167	1625
YJL172W	YJL172W	1	10	97730	99460	101187	101193	1727
YLL065W	YER187W	-0.659193	5	566225	566650	568425	568734	1775
YMR009W	YLR247C	0.728518	12	628684	633354	635380	635396	2026
YLR089C	YCL017C	0.65272	3	92776	94269	81843	90610	2166
YGL010W	YCL017C	0.622139	3	92776	94269	81832	90412	2364
YGL114W	YCL017C	0.687805	3	92776	94269	81832	90412	2364
YLR042C	YBR149W	0.621576	2	539944	540978	517365	537314	2630
YOR138C	YOR138C	1	15	584309	586324	589145	608832	2821
YLL050C	YOL026C	0.649719	15	274012	274353	277470	282525	3117
YLR079W	YPR106W	0.642777	16	740058	741389	744524	744530	3135
YJR010C-A	YJR010C-A	1	10	457769	458053	461201	464261	3148
YNL316C	YNL316C	1	14	42070	43176	27632	38852	3218
YHR117W	YBR145W	-0.627767	2	533719	534774	521415	530481	3238
YHR036W	YLR397C	0.646717	12	912549	914891	899898	909226	3323
YDR039C	YDR039C	1	4	531301	534576	527509	527875	3426
YDL182W	YBR104W	0.661163	2	449625	450614	454091	465933	3477
YNL317W	YNL317W	1	14	40618	42015	27632	37071	3547
YBR115C	YBR115C	1	2	469705	473883	454091	465933	3772
YDL182W	YBR115C	0.77955	2	469705	473883	454091	465933	3772
YKR049C	YOL097C	-0.726642	15	136526	137824	141621	180180	3797
YBR068C	YCL026C-A	0.574672	3	74704	75285	79091	90412	3806
YGR146C	YCL026C-A	0.55122	3	74704	75285	79091	90412	3806
YMR096W	YCL026C-A	0.641651	3	74704	75285	79091	92391	3806
YPL245W	YPL245W	1	16	85586	86950	90762	128687	3812
YAL041W	YAL054C	-0.539962	1	42881	45022	37068	38887	3994
YLR178C	YOL083W	0.615009	15	165713	166951	170945	180180	3994
YGR266W	YLR278C	0.573921	12	699999	704024	708041	710924	4017
YJL116C	YOL090W	-0.683302	15	147381	150275	154309	174364	4034
YBR129C	YNL122C	0.720638	14	398020	398367	368178	393903	4117
YDR041W	YNL122C	0.884803	14	398020	398367	368178	393903	4117
YDR197W	YNL122C	0.77167	14	398020	398367	368178	393903	4117

YGR076C	YNL122C	0.842214	14	398020	398367	368178	393903	4117
YGR220C	YNL122C	0.830019	14	398020	398367	368178	393903	4117
YHR038W	YNL122C	0.900375	14	398020	398367	368178	393903	4117
YHR147C	YNL122C	0.852158	14	398020	398367	368178	393903	4117
YJL096W	YNL122C	0.910319	14	398020	398367	368178	393903	4117
YJR101W	YNL122C	0.896435	14	398020	398367	368178	393903	4117
YKL138C	YNL122C	0.885366	14	398020	398367	368178	393903	4117
YKL167C	YNL122C	0.746154	14	398020	398367	368178	393903	4117
YKL169C	YNL122C	0.886492	14	398020	398367	368178	393903	4117
YMR157C	YNL122C	0.669418	14	398020	398367	368178	393903	4117
YNL122C	YNL122C	1	14	398020	398367	371953	393903	4117
YPR099C	YNL122C	0.912383	14	398020	398367	368178	393903	4117
YPR100W	YNL122C	0.930394	14	398020	398367	368178	393903	4117
YGL064C	YHR043C	-0.645591	8	192796	193536	176994	188671	4125
YHR043C	YHR043C	1	8	192796	193536	176994	188671	4125
YBR037C	YIL136W	0.657974	9	93619	94800	98949	133663	4149
YKL217W	YNL117W	0.689306	14	406355	408019	412275	418269	4256
YHL020C	YHL020C	1	8	66238	67452	71742	72233	4290
YNL245C	YNL245C	1	14	186345	186884	191240	191243	4356
YOR081C	YOR081C	1	15	476940	479189	469541	472577	4363
YBR127C	YNL071W	0.723452	14	491520	492968	418293	486861	4659
YKL127W	YNL071W	0.70394	14	491520	492968	418293	486861	4659
YNL071W	YNL071W	1	14	491520	492968	418293	486861	4659
YOL023W	YNL071W	-0.730206	14	491520	492968	449639	486861	4659
YDL182W	YBR103W	0.616886	2	447667	449274	454091	465933	4817
YLL028W	YOL108C	-0.628143	15	111430	111885	106152	106272	5158
YDL204W	YOL084W	0.853846	15	162355	165330	170945	180180	5615
YDL223C	YOL084W	0.871482	15	162355	165330	170945	174364	5615
YDR070C	YOL084W	0.852908	15	162355	165330	170945	174364	5615
YDR533C	YOL084W	0.820638	15	162355	165330	170945	180180	5615
YER150W	YOL084W	0.871857	15	162355	165330	170945	180180	5615
YGR043C	YOL084W	0.869606	15	162355	165330	170945	174364	5615
YLR178C	YOL084W	0.792683	15	162355	165330	170945	180180	5615
YML128C	YOL084W	0.854972	15	162355	165330	170945	180180	5615
YMR196W	YOL084W	0.846341	15	162355	165330	170945	174364	5615
YNL234W	YOL084W	0.673921	15	162355	165330	170945	175594	5615
YOL084W	YOL084W	1	15	162355	165330	170945	180180	5615
YOR173W	YOL084W	0.762289	15	162355	165330	170945	180180	5615
YPL223C	YOL084W	0.837711	15	162355	165330	170945	174364	5615
YPR184W	YOL084W	0.85591	15	162355	165330	170945	174364	5615
YOR342C	YBR214W	-0.611445	2	651372	652955	658746	670435	5791

YPL184C	YNL078W	0.640338	14	479765	480988	486861	525061	5873
YGL205W	YAL060W	0.607692	1	35156	36304	42255	55329	5951
YNL208W	YBR186W	-0.605816	2	600510	602317	608310	609055	5993
YPL002C	YPL002C	1	16	553622	554323	542307	547621	6001
YHR117W	YBR147W	-0.683865	2	536532	537422	521415	530481	6051
YKR089C	YKR089C	1	11	605268	608000	596152	599170	6098
YGL081W	YIL084C	0.638837	9	202273	203256	190794	196145	6128
YER188W	YER188W	1	5	568035	568754	549142	561876	6159
YCR046C	YIL136W	0.624765	9	93619	94800	101011	133693	6211
YJR060W	YJR060W	1	10	548452	549507	535311	541818	6634
YBL091C	YNL118C	0.563977	14	402649	405561	412269	412758	6708
YGL096W	YGL096W	1	7	325332	326162	318608	318611	6721
YBL005W	YBL005W	1	2	217432	220362	227290	227302	6928
YBR185C	YNL005C	0.696998	14	621310	622425	610996	614342	6968
YHR116W	YNL005C	0.824203	14	621310	622425	610996	614342	6968
YJL180C	YNL005C	0.8803	14	621310	622425	610996	614342	6968
YOL042W	YNL005C	0.621951	14	621310	622425	610996	614342	6968
YPL118W	YNL005C	0.858912	14	621310	622425	610996	614342	6968
YDR492W	YGL057C	-0.659287	7	394970	395833	402833	415585	7000
YGL057C	YGL057C	1	7	394970	395833	402833	415585	7000
YEL016C	YEL016C	1	5	124737	126218	109310	117705	7032
YDR005C	YDR005C	1	4	456832	458099	465157	478080	7058
YDR040C	YDR040C	1	4	535186	538461	527508	527875	7311
YHR117W	YBR148W	-0.673734	2	537833	539662	521415	530481	7352
YGL028C	YBR162C	0.673734	2	563160	564527	530481	555787	7373
YHR143W	YBR162C	0.645966	2	563160	564527	530481	555787	7373
YOR285W	YOR285W	1	15	849632	850051	838599	842027	7605
YBR026C	YBR026C	1	2	292836	293978	301671	310928	7693
YPR006C	YPR006C	1	16	567264	568991	542307	559544	7720
YBR256C	YCL018W	-0.667542	3	91323	92417	100213	105042	7796
YMR292W	YGR106C	0.637523	7	698990	699787	707605	707700	7818
YDL086W	YLR244C	-0.610507	12	625168	626331	634225	634227	7894
YBR044C	YBR044C	1	2	324298	326019	334016	334022	7997
YJR162C	YER185W	0.669043	5	559449	560360	568425	568734	8065
YLL065W	YER185W	0.641276	5	559449	560360	568425	568734	8065
YEL039C	YLR256W	-0.558537	12	646415	650923	659357	663370	8434
YMR009W	YLR256W	-0.571107	12	646415	650923	659357	668249	8434
YHR046C	YHR046C	1	8	197389	198276	176994	188671	8718
YOR264W	YBR148W	-0.754784	2	537833	539662	548401	555787	8739
YMR292W	YGR105W	0.787805	7	698600	698833	707605	707700	8772
YOR009W	YNL080C	0.681426	14	476929	478029	486861	502496	8832

YIR038C	YOR178C	0.784615	15	667860	670241	679086	690475	8845
YHR024C	YNL004W	0.695872	14	623329	624618	610996	614342	8987
YGR052W	YOL084W	0.876735	15	162355	165330	174364	180180	9034
YHR044C	YOL084W	0.622514	15	162355	165330	174364	180180	9034
YKR046C	YLR244C	-0.714822	12	625168	626331	635380	635396	9049
YMR009W	YLR244C	-0.604503	12	625168	626331	635380	635396	9049
YJL171C	YJL171C	1	10	99697	100887	110038	116255	9151
YIL034C	YOL022C	-0.689118	15	280272	281498	290670	338018	9172
YCR046C	YIL116W	-0.627955	9	142925	144082	101011	133693	9232
YBR037C	YIL116W	-0.571107	9	142925	144082	98949	133663	9262
YDR384C	YMR139W	-0.609006	13	546124	547236	556726	556835	9490
YDR128W	YDR128W	1	4	709542	712988	723161	744450	10173
YMR193W	YNL081C	0.825141	14	476185	476616	486861	502496	10245
YOR009W	YNL081C	-0.727767	14	476185	476616	486861	502496	10245
YPL184C	YNL081C	-0.666979	14	476185	476616	486861	525061	10245
YBR017C	YNL117W	-0.639024	14	406355	408019	418293	486861	10274
YIL034C	YOL023W	-0.670732	15	278056	280086	290670	338018	10584
YLR285W	YBR189W	0.625704	2	604465	605465	592863	593761	10704
YHR190W	YLR247C	0.566417	12	628684	633354	644136	662627	10782
YHR160C	YLR414C	0.65197	12	953348	954139	964989	965230	10850
YOR264W	YBR147W	-0.687617	2	536532	537422	548401	555787	10979
YLR153C	YLR265C	-0.566229	12	674427	675455	642137	663370	11057
YLR205C	YLR265C	-0.640619	12	674427	675455	659357	663370	11057
YML126C	YLR265C	-0.562289	12	674427	675455	642137	663370	11057
YNR043W	YLR265C	-0.664728	12	674427	675455	659357	663370	11057
YOR157C	YOR157C	1	15	630966	631751	618868	619862	11104
YDR261C	YNL100W	-0.685929	14	437610	438314	449639	502316	11325
YNL177C	YNL100W	0.685366	14	437610	438314	449639	502316	11325
YCR052W	YCR052W	1	3	214986	216437	228125	240265	11688
YDR453C	YOL094C	-0.693433	15	141583	142554	154309	174364	11755
YJR016C	YCL009C	0.706567	3	104614	105543	79091	92391	12223
YFL061W	YJR145C	-0.531144	10	701721	702762	715254	730438	12492
YDR390C	YLR375W	-0.741651	12	871696	872727	855389	858996	12700
YDL131W	YBR115C	0.735272	2	469705	473883	486640	499012	12757
YOR049C	YOR049C	1	15	422668	423732	407684	409779	12889
YOR283W	YOR283W	1	15	847450	848142	861183	862479	13041
YJR048W	YLR244C	-0.595685	12	625168	626331	611810	611997	13171
YMR292W	YGR170W	-0.697373	7	837144	840560	853749	858473	13189
YGR235C	YLR244C	-0.577674	12	625168	626331	611854	611967	13201
YLR295C	YKL087C	0.69137	11	276830	277504	258243	263612	13218
YPL078C	YKL087C	0.601689	11	276830	277504	258243	263612	13218

YEL039C	YLR244C	-0.632833	12	625168	626331	611810	611854	13314
YGL160W	YLR244C	-0.668668	12	625168	626331	607076	611854	13314
YLR205C	YLR244C	-0.668949	12	625168	626331	611810	611854	13314
YLR244C	YLR244C	1	12	625168	626331	611810	611854	13314
YMR134W	YLR244C	-0.656285	12	625168	626331	611810	611854	13314
YNR019W	YLR244C	-0.572233	12	625168	626331	607076	611810	13358
YHR091C	YNL058C	-0.612758	14	515759	516709	449639	502316	13443
YNL208W	YBR182C	0.605816	2	593463	594821	608310	609055	13489
YIL101C	YIL101C	1	9	175304	177247	190794	196145	13547
YER023W	YER023W	1	5	201075	201935	183958	187458	13617
YER150W	YOL073C	0.723077	15	193831	194799	170945	180180	13651
YLR178C	YOL073C	0.723077	15	193831	194799	170945	180180	13651
YDL055C	YBR138C	-0.621201	2	513719	515293	489202	499889	13830
YHR091C	YNL057W	-0.599812	14	516399	516731	449639	502316	14083
YGL114W	YCL009C	0.690619	3	104614	105543	81832	90412	14202
YER182W	YNL122C	0.809568	14	398020	398367	412758	418293	14391
YDR390C	YLR356W	-0.645216	12	840320	840913	855389	858996	14476
YBL091C	YNL123W	0.590619	14	394682	397675	412269	412758	14594
YKL152C	YNL123W	0.662477	14	394682	397675	412269	418293	14594
YKL217W	YNL123W	-0.638086	14	394682	397675	412275	418269	14600
YOR263C	YBR170C	-0.608068	2	576301	578043	592863	593761	14820
YBR157C	YFL054C	0.635272	6	20847	22787	5870	5877	14970
YGR117C	YFL054C	-0.653846	6	20847	22787	5852	5877	14970
YOL104C	YOL104C	1	15	116395	117453	132423	136327	14970
YBR298C	YGR289C	0.687617	7	1073967	1075817	1057534	1058947	15020
YEL039C	YLR253W	0.557598	12	642627	644336	659357	663370	15021
YPL078C	YKL085W	0.649343	11	278764	279768	258243	263612	15152
YDR005C	YAR069C	-0.587148	1	224001	224294	208493	208497	15504
YBR183W	YLR244C	-0.762101	12	625168	626331	642137	644136	15806
YNL007C	YGL049C	0.732458	7	406861	409605	425445	427476	15840
YDL120W	YAL053W	-0.6606	1	45899	48250	23813	29981	15918
YDR453C	YOL097C	-0.633021	15	136526	137824	154309	174364	16485
YEL033W	YLR332W	0.632645	12	790676	791806	808623	808707	16817
YMR134W	YLR247C	0.77242	12	628684	633354	611810	611854	16830
YNL111C	YLR247C	0.57242	12	628684	633354	607076	611854	16830
YPL184C	YNL085W	0.692683	14	467128	469620	486861	525061	17241
YOR342C	YBL004W	0.6803	2	227598	235079	252538	252640	17459
YMR292W	YGL030W	0.652533	7	439094	439641	457215	460945	17574
YHL050C	YLR446W	0.529268	12	1025209	1026510	1044161	1050447	17651
YJL225C	YLR446W	0.541651	12	1025209	1026510	1044161	1050447	17651
YOR238W	YOR238W	1	15	783668	784588	802724	808276	18136

YDR384C	YMR160W	-0.674484	13	575065	577515	556726	556835	18230
YFL010C	YFL010C	1	6	115102	115737	134102	134252	18365
YMR036C	YMR036C	1	13	341855	343519	362310	362346	18791
YJR162C	YER179W	0.608255	5	548416	549512	568425	568734	18913
YBR158W	YBR158W	1	2	556505	558154	530481	537314	19191
YDR119W	YBR158W	0.748405	2	556505	558154	530481	537314	19191
YLR042C	YBR158W	-0.615385	2	556505	558154	517365	537314	19191
YLR285W	YBR158W	0.669231	2	556505	558154	530481	537314	19191
YOR263C	YBR158W	0.633583	2	556505	558154	530481	537314	19191
YCR071C	YNL100W	0.671107	14	437610	438314	418269	418293	19317
YER182W	YNL100W	0.822326	14	437610	438314	412758	418293	19317
YJR147W	YBR128C	-0.651595	2	493038	494072	513408	514035	19336
YDR453C	YOL073C	0.595122	15	193831	194799	154309	174364	19467
YCR071C	YNL122C	0.700563	14	398020	398367	418269	418293	19902
YER087W	YNL122C	0.645591	14	398020	398367	418293	449639	19926
YKL217W	YNL125C	0.658161	14	390143	392164	412275	418269	20111
YPL184C	YNL086W	-0.639024	14	466331	466639	486861	525061	20222
YGL028C	YBR170C	-0.650469	2	576301	578043	530481	555787	20514
YHR143W	YBR170C	-0.648968	2	576301	578043	530481	555787	20514
YJL078C	YBR170C	-0.691557	2	576301	578043	548401	555787	20514
YNL327W	YBR170C	-0.657223	2	576301	578043	530481	555778	20523
YIR038C	YOR173W	0.819137	15	657132	658325	679086	690475	20761
YIL134W	YIL134W	1	9	97395	98330	74600	76380	21015
YLL050C	YOL012C	0.687242	15	303579	303983	277470	282525	21054
YDR384C	YMR131C	0.655722	13	533162	534697	556726	556835	22029
YGL110C	YMR219W	0.63621	13	707132	712108	667328	685092	22040
YDL124W	YDL124W	1	4	240258	241196	211612	217399	22859
YLR035C	YBR109C	-0.62758	2	457876	458319	481439	486640	23120
YBR064W	YBR064W	1	2	367723	368151	334020	343931	23792
YNL254C	YML099C	0.705816	13	74398	77040	49894	49903	24495
YGL081W	YIL074C	-0.635272	9	221078	222487	190794	196145	24933
YMR098C	YNL081C	0.826266	14	476185	476616	502316	502496	25700
YNR020C	YNL081C	0.843902	14	476185	476616	502316	502496	25700
YJL078C	YBR173C	-0.632083	2	581683	582129	548401	555787	25896
YNL035C	YNL058C	0.605816	14	515759	516709	542648	547072	25939
YJL167W	YLR247C	0.620263	12	628684	633354	659357	663370	26003
YLR101C	YLR247C	0.622889	12	628684	633354	659357	663370	26003
YNL156C	YLR247C	0.711632	12	628684	633354	659357	663370	26003
YHR117W	YBR158W	0.697186	2	556505	558154	521415	530481	26024
YML050W	YNL122C	0.715197	14	398020	398367	368178	371953	26067
YPR013C	YER081W	0.577298	5	322682	324091	350744	420589	26653

YDL086W	YLR231C	0.703189	12	605758	607119	634225	634227	27106
YHR117W	YBR159W	0.69925	2	558641	559684	521415	530481	28160
YKR046C	YLR231C	0.700375	12	605758	607119	635380	635396	28261
YDL204W	YOL094C	-0.70075	15	141583	142554	170945	180180	28391
YDR070C	YOL094C	-0.713133	15	141583	142554	170945	174364	28391
YDR533C	YOL094C	-0.674109	15	141583	142554	170945	180180	28391
YLR178C	YOL094C	-0.760038	15	141583	142554	170945	180180	28391
YMR196W	YOL094C	-0.676548	15	141583	142554	170945	174364	28391
YOR173W	YOL094C	-0.676173	15	141583	142554	170945	180180	28391
YPR184W	YOL094C	-0.705816	15	141583	142554	170945	174364	28391
YBR166C	YOL131W	-0.659287	15	73030	73356	44542	44548	28482
YMR096W	YCR005C	0.579737	3	120940	122322	79091	92391	28549
YDR054C	YGL241W	0.648218	7	45445	48459	16571	16619	28826
YDR262W	YDR262W	1	4	993125	993943	963733	963769	29356
YGR049W	YLR265C	-0.589493	12	674427	675455	704828	705220	29373
YGL129C	YNL052W	0.669794	14	531721	532182	449639	502316	29405
YOR354C	YNL052W	0.678612	14	531721	532182	449639	502316	29405
YMR068W	YMR068W	1	13	406303	407583	437167	437329	29584
YCL026C-A	YCL026C-A	1	3	74704	75285	105042	175799	29757
YDR033W	YDR033W	1	4	508141	509103	465157	478080	30061
YFR013W	YMR188C	0.623077	13	636290	637003	667328	686206	30325
YIL098C	YNL051W	0.621576	14	532655	533866	449639	502316	30339
YJR048W	YLR253W	0.654784	12	642627	644336	611810	611997	30630
YKR049C	YOL109W	-0.71257	15	110296	110637	141621	180180	30984
YGL262W	YGL241W	0.543058	7	45445	48459	10161	12939	32506
YMR208W	YLR265C	-0.629831	12	674427	675455	708041	710924	32586
YIL034C	YOR086C	0.699062	15	483220	486780	519764	519776	32984
YJL048C	YLR244C	-0.799062	12	625168	626331	659357	663370	33026
YLR100W	YLR244C	-0.567542	12	625168	626331	659357	663370	33026
YMR015C	YLR244C	-0.549906	12	625168	626331	659357	663370	33026
YNL156C	YLR244C	-0.757036	12	625168	626331	659357	663370	33026
YDR070C	YOL097C	-0.729268	15	136526	137824	170945	174364	33121
YLR178C	YOL097C	-0.763227	15	136526	137824	170945	180180	33121
YMR196W	YOL097C	-0.705629	15	136526	137824	170945	174364	33121
YER011W	YLR239C	-0.611069	12	616332	617318	570433	582499	33833
YGR235C	YLR256W	-0.639775	12	646415	650923	611854	611967	34448
YDL048C	YOR028C	0.679925	15	383532	384419	301077	348934	34598
YJR047C	YLR256W	-0.573358	12	646415	650923	607076	611810	34605
YNR013C	YNL073W	-0.599625	14	488383	490113	525068	542648	34955
YIL111W	YIL111W	1	9	155219	155762	190794	191491	35032
YGR086C	YOL164W	-0.613321	15	6175	8115	43172	43217	35057

YHR108W	YHR108W	1	8	328305	330062	224267	293131	35174
YPL184C	YNL093W	-0.578612	14	449865	450527	486861	525061	36334
YIL034C	YOR085W	0.648405	15	482034	483086	519764	519776	36678
YKL210W	YKL210W	1	11	39164	42238	79182	85837	36944
YDL131W	YBR103W	0.709944	2	447667	449274	486640	499012	37366
YDL055C	YBR148W	-0.579925	2	537833	539662	489202	499889	37944
YOR062C	YOR062C	1	15	442726	443532	481586	488329	38054
YHR117W	YBR121C	0.74409	2	481321	483324	521415	530481	38091
YOR277C	YLR264W	0.729644	12	673131	673334	634225	634227	38904
YLR042C	YBR170C	0.634709	2	576301	578043	517365	537314	38987
YLR265C	YLR265C	1	12	674427	675455	635380	635396	39031
YHR160C	YLR401C	-0.621388	12	922617	924446	964989	965230	40543
YPL083C	YPL083C	1	16	396697	398100	439789	440563	41689
YJL072C	YJL072C	1	10	304917	305558	227894	262593	42324
YHR160C	YLR400W	-0.572608	12	922062	922535	964989	965230	42454
YDR453C	YOL109W	-0.782739	15	110296	110637	154309	174364	43672
YJL116C	YOL109W	-0.784428	15	110296	110637	154309	174364	43672
YBR129C	YNL100W	0.673358	14	437610	438314	368178	393903	43707
YDR197W	YNL100W	0.67955	14	437610	438314	368178	393903	43707
YGR220C	YNL100W	0.677486	14	437610	438314	368178	393903	43707
YHR147C	YNL100W	0.767917	14	437610	438314	368178	393903	43707
YKL167C	YNL100W	0.8	14	437610	438314	368178	393903	43707
YKL169C	YNL100W	0.758537	14	437610	438314	368178	393903	43707
YNL273W	YPL063W	0.63227	16	429934	431364	475487	479892	44123
YLL028W	YOL089C	-0.646717	15	150397	153489	106152	106272	44125
YGL020C	YGL020C	1	7	457166	457873	502131	557236	44258
YPL118W	YNL073W	0.737148	14	488383	490113	534449	547071	44336
YLR042C	YBR173C	0.620263	2	581683	582129	517365	537314	44369
YDR044W	YLR247C	-0.730582	12	628684	633354	677957	693790	44603
YDR197W	YNL099C	0.672045	14	438564	439280	368178	393903	44661
YFR013W	YMR179W	0.721013	13	619857	622133	667328	686206	45195
YGL110C	YMR179W	0.679925	13	619857	622133	667328	685092	45195
YNL078W	YBR158W	0.706754	2	556505	558154	603790	609055	45636
YMR152W	YMR122C	-0.549531	13	510700	511074	556841	562907	45767
YBR103W	YBR103W	1	2	447667	449274	391856	401568	46099
YER011W	YLR247C	-0.643715	12	628684	633354	570433	582499	46185
YIL034C	YOR078W	-0.643152	15	472726	473370	519764	519776	46394
YNL327W	YBR187W	0.630394	2	602591	603433	530481	555778	46813
YLR313C	YLR313C	1	12	760750	762342	713638	713644	47106
YKR049C	YOL120C	-0.631895	15	93394	94401	141621	180180	47220
YGL110C	YMR230W	0.642777	13	732413	733140	667328	685092	47321

YLL028W	YOL088C	-0.631144	15	153911	154744	106152	106272	47639
YKR049C	YOL047C	0.781801	15	241612	242745	180407	193911	47701
YKL152C	YNL086W	-0.609193	14	466331	466639	412269	418293	48038
YCR071C	YNL137C	0.6606	14	368592	370052	418269	418293	48217
YIL034C	YOL054W	-0.633021	15	228612	229832	180210	180222	48390
YDR089W	YDR089W	1	4	622105	624714	569429	573400	48705
YKL152C	YNL085W	0.638086	14	467128	469620	412269	418293	48835
YGL193C	YGL193C	1	7	141920	142231	85112	92848	49072
YEL033W	YMR009W	0.659099	13	284101	284640	234845	234851	49250
YMR292W	YGR010W	-0.670356	7	511546	512733	457215	460945	50601
YLR206W	YLR206W	1	12	554578	556419	607076	611997	50657
YBL038W	YNL122C	0.863039	14	398020	398367	449639	502316	51272
YDR347W	YNL122C	0.799625	14	398020	398367	449639	502316	51272
YGR165W	YNL122C	0.89531	14	398020	398367	449639	502316	51272
YHR059W	YNL122C	0.849343	14	398020	398367	449639	502316	51272
YKL170W	YNL122C	0.881051	14	398020	398367	449639	502316	51272
YNL185C	YNL122C	0.792495	14	398020	398367	449639	502316	51272
YGL195W	YNL123W	0.712946	14	394682	397675	449639	502316	51964
YJL048C	YLR231C	0.706379	12	605758	607119	659357	663370	52238
YPL028W	YLR231C	0.569981	12	605758	607119	659357	663370	52238
YPL091W	YPL091W	1	16	375497	376948	429188	436455	52240
YOR342C	YBR034C	0.743527	2	304889	305935	252538	252640	52249
YFL060C	YJR128W	-0.583396	10	662615	662974	715254	730438	52280
YOR277C	YLR222C	0.829831	12	579318	581771	634225	634227	52454
YDL202W	YNL073W	0.743902	14	488383	490113	542648	547072	52535
YLR069C	YNL073W	0.709381	14	488383	490113	542648	547072	52535
YHR116W	YNR020C	0.884428	14	667407	668219	610996	614342	53065
YJL180C	YNR020C	0.856473	14	667407	668219	610996	614342	53065
YOL042W	YNR020C	0.6	14	667407	668219	610996	614342	53065
YDR533C	YOL104C	-0.688931	15	116395	117453	170945	180180	53492
YOR173W	YOL104C	-0.63227	15	116395	117453	170945	180180	53492
YGR235C	YLR260W	0.604878	12	665844	667907	611854	611967	53877
YMR292W	YGL054C	0.669418	7	400872	401288	457215	460945	55927
YKL127W	YNL045W	0.617448	14	542959	544974	418293	486861	56098
YPL088W	YPL088W	1	16	381960	382988	439789	440563	56801
YKR049C	YOL041C	-0.671857	15	251265	252644	180407	193911	57354
YCR071C	YNL081C	0.670544	14	476185	476616	418269	418293	57892
YER182W	YNL081C	0.882927	14	476185	476616	412758	418293	57892
YPL215W	YNL081C	0.643152	14	476185	476616	418269	418293	57892
YDL244W	YJL176C	-0.627017	10	92050	94527	26519	34098	57952
YGL110C	YMR172C-A	0.651407	13	607827	608210	667328	685092	59118

YDR390C	YLR314C	0.782739	12	762575	764137	823996	824230	59859
YLR178C	YOL048C	0.684053	15	240202	240522	170945	180180	60022
YML128C	YOL048C	0.677861	15	240202	240522	170945	180180	60022
YGL110C	YMR260C	0.629081	13	789377	789838	849966	850514	60128
YDL204W	YOL109W	-0.691745	15	110296	110637	170945	180180	60308
YDL223C	YOL109W	-0.741839	15	110296	110637	170945	174364	60308
YDR070C	YOL109W	-0.690807	15	110296	110637	170945	174364	60308
YDR533C	YOL109W	-0.733208	15	110296	110637	170945	180180	60308
YER150W	YOL109W	-0.746341	15	110296	110637	170945	180180	60308
YLR178C	YOL109W	-0.687242	15	110296	110637	170945	180180	60308
YOR173W	YOL109W	-0.637711	15	110296	110637	170945	180180	60308
YPL223C	YOL109W	-0.745966	15	110296	110637	170945	174364	60308
YPR184W	YOL109W	-0.76773	15	110296	110637	170945	174364	60308
YKR049C	YOL127W	-0.606004	15	80347	81189	141621	180180	60432
YDR502C	YLR297W	-0.594747	12	724044	724433	659357	663370	60674
YIR038C	YOR152C	0.704315	15	617518	618288	679086	690475	60798
YOR277C	YLR276C	0.801126	12	695046	696830	634225	634227	60819
YNL254C	YML078W	-0.653471	13	111002	111550	49894	49903	61099
YOL113W	YOL113W	1	15	104325	106292	43172	43217	61108
YJL154C	YJL154C	1	10	130799	133633	40238	69578	61221
YDL204W	YOL047C	0.814259	15	241612	242745	170945	180180	61432
YER150W	YOL047C	0.795497	15	241612	242745	170945	180180	61432
YGR052W	YOL047C	0.726079	15	241612	242745	174364	180180	61432
YLR178C	YOL047C	0.742401	15	241612	242745	170945	180180	61432
YML128C	YOL047C	0.770732	15	241612	242745	170945	180180	61432
YOR173W	YOL047C	0.714071	15	241612	242745	170945	180180	61432
YKL152C	YNL078W	0.611257	14	479765	480988	412269	418293	61472
YBL043W	YBR045C	0.57167	2	328330	330051	391850	392138	61799
YFL052W	YJL173C	-0.564165	10	96158	96526	24745	34098	62060
YDR384C	YMR112C	-0.593246	13	494099	494494	556726	556835	62232
YLL028W	YOL082W	-0.606004	15	168726	169973	106152	106272	62454
YBR129C	YNL177C	0.712758	14	303683	304612	368178	393903	63566
YDR041W	YNL177C	0.887617	14	303683	304612	368178	393903	63566
YDR197W	YNL177C	0.79925	14	303683	304612	368178	393903	63566
YGR220C	YNL177C	0.817824	14	303683	304612	368178	393903	63566
YHR038W	YNL177C	0.850844	14	303683	304612	368178	393903	63566
YHR147C	YNL177C	0.81651	14	303683	304612	368178	393903	63566
YJL096W	YNL177C	0.913884	14	303683	304612	368178	393903	63566
YJR101W	YNL177C	0.881801	14	303683	304612	368178	393903	63566
YKL167C	YNL177C	0.770919	14	303683	304612	368178	393903	63566
YKL169C	YNL177C	0.866792	14	303683	304612	368178	393903	63566

YMR157C	YNL177C	0.681051	14	303683	304612	368178	393903	63566
YPR100W	YNL177C	0.91576	14	303683	304612	368178	393903	63566
YER114C	YNL123W	0.749343	14	394682	397675	330365	330368	64314
YNL234W	YOL047C	0.638274	15	241612	242745	170945	175594	66018
YLR069C	YNL081C	0.714822	14	476185	476616	542648	547072	66032
YIR038C	YOL109W	-0.713696	15	110296	110637	43172	43217	67079
YOL109W	YOL109W	1	15	110296	110637	43172	43217	67079
YGL114W	YCR023C	0.616698	3	158530	160365	81832	90412	68118
YPL215W	YNL073W	0.763415	14	488383	490113	418269	418293	70090
YCL033C	YMR152W	0.629081	13	563095	564192	492201	492207	70888
YPL195W	YFL030W	-0.642402	6	76829	77986	5852	5854	70975
YLR178C	YOL041C	-0.715385	15	251265	252644	170945	180180	71085
YLR295C	YKL137W	0.666979	11	185984	186295	258243	263612	71948
YHR160C	YLR384C	-0.655159	12	888851	892900	964989	965230	72089
YDR390C	YLR309C	0.770169	12	749034	751769	823996	824230	72227
YIR038C	YOR227W	0.693058	15	762825	766565	679086	690475	72350
YKL167C	YNL086W	0.668668	14	466331	466639	368178	393903	72428
YLR295C	YKL138C	0.705816	11	185287	185682	258243	263612	72561
YIL151C	YIL151C	1	9	57338	60694	133693	133699	72999
YMR292W	YGR027C	0.79137	7	534133	534459	457215	460945	73188
YIL034C	YOR142W	0.633208	15	593057	594046	519764	519776	73281
YJL183W	YJL183W	1	10	84066	85334	159479	185313	74145
YLR069C	YNL005C	0.867167	14	621310	622425	542648	547072	74238
YNL036W	YCR007C	0.585741	3	126005	126724	201167	210748	74443
YGR220C	YNL184C	0.781051	14	292554	292880	368178	393903	75298
YHR147C	YNL184C	0.754972	14	292554	292880	368178	393903	75298
YDR197W	YNL185C	0.72439	14	292190	292666	368178	393903	75512
YGR220C	YNL185C	0.741088	14	292190	292666	368178	393903	75512
YHR147C	YNL185C	0.850844	14	292190	292666	368178	393903	75512
YJL096W	YNL185C	0.817073	14	292190	292666	368178	393903	75512
YJR101W	YNL185C	0.819887	14	292190	292666	368178	393903	75512
YKL169C	YNL185C	0.760413	14	292190	292666	368178	393903	75512
YMR157C	YNL185C	0.767917	14	292190	292666	368178	393903	75512
YER114C	YNL117W	-0.598874	14	406355	408019	330365	330368	75987
YCL033C	YMR155W	0.63621	13	568550	570193	492201	492207	76343
YIR038C	YOR228C	0.730769	15	766869	767777	679086	690475	76394
YOR277C	YLR287C	0.803002	12	710991	712058	634225	634227	76764
YLR295C	YKL141W	0.65272	11	179668	180264	258243	263612	77979
YOR277C	YLR287C-A	0.817073	12	712537	713158	634225	634227	78310
YNR013C	YNL005C	-0.645403	14	621310	622425	525068	542648	78662
YBR185C	YNL052W	0.662664	14	531721	532182	610996	614342	78814

YOL042W	YNL052W	0.599437	14	531721	532182	610996	614342	78814
YGR251W	YGR251W	1	7	995642	996232	1075580	1081936	79348
YDR347W	YNL137C	0.850657	14	368592	370052	449639	502316	79587
YGR165W	YNL137C	0.926079	14	368592	370052	449639	502316	79587
YHR059W	YNL137C	0.813321	14	368592	370052	449639	502316	79587
YKL170W	YNL137C	0.808818	14	368592	370052	449639	502316	79587
YNL185C	YNL137C	0.823827	14	368592	370052	449639	502316	79587
YGL114W	YCR028C	0.627204	3	170878	172416	81832	90412	80466
YIL034C	YOR143C	-0.711632	15	601383	602342	519764	519776	81607
YBR129C	YNL081C	0.758724	14	476185	476616	368178	393903	82282
YDR041W	YNL081C	0.898124	14	476185	476616	368178	393903	82282
YDR197W	YNL081C	0.789306	14	476185	476616	368178	393903	82282
YGR076C	YNL081C	0.879925	14	476185	476616	368178	393903	82282
YGR220C	YNL081C	0.837336	14	476185	476616	368178	393903	82282
YHR038W	YNL081C	0.825328	14	476185	476616	368178	393903	82282
YHR147C	YNL081C	0.827205	14	476185	476616	368178	393903	82282
YJL096W	YNL081C	0.87167	14	476185	476616	368178	393903	82282
YJR101W	YNL081C	0.790431	14	476185	476616	368178	393903	82282
YKL138C	YNL081C	0.863039	14	476185	476616	368178	393903	82282
YKL167C	YNL081C	0.834897	14	476185	476616	368178	393903	82282
YKL169C	YNL081C	0.892308	14	476185	476616	368178	393903	82282
YMR157C	YNL081C	0.670544	14	476185	476616	368178	393903	82282
YNL081C	YNL081C	1	14	476185	476616	368178	393903	82282
YPR099C	YNL081C	0.880113	14	476185	476616	368178	393903	82282
YPR100W	YNL081C	0.913133	14	476185	476616	368178	393903	82282
YGR223C	YGR223C	1	7	940869	942215	833786	858473	82396
YIL034C	YOR144C	-0.651407	15	602717	605092	519764	519776	82941
YLR295C	YKL148C	0.637336	11	169208	171130	258243	263612	87113
YPR020W	YKL148C	0.643715	11	169208	171130	258243	263612	87113
YOR277C	YLR196W	0.741088	12	543970	545700	634225	634227	88525
YLR178C	YOL029C	-0.624765	15	269815	270420	170945	180180	89635
YMR048W	YMR048W	1	13	366980	367933	261719	277071	89909
YGR117C	YFL021W	-0.671107	6	95964	97496	5852	5877	90087
YPL195W	YFL021W	-0.599625	6	95964	97496	5852	5854	90110
YNL254C	YML066C	-0.657786	13	140424	141533	49894	49903	90521
YDR406W	YOL094C	-0.614259	15	141583	142554	44548	47951	93632
YOR185C	YOR185C	1	15	681444	682106	581307	587316	94128
YDR197W	YNL073W	0.645028	14	488383	490113	368178	393903	94480
YMR292W	YGR034W	0.749156	7	555930	556673	457215	460945	94985
YNL018C	YNL018C	1	14	599932	601770	502316	502496	97436
YIR038C	YOL094C	-0.694559	15	141583	142554	43172	43217	98366

YMR292W	YGR038W	-0.737242	7	560683	561351	457215	460945	99738
YDL204W	YOL022C	-0.702627	15	280272	281498	170945	180180	100092
YLR178C	YOL022C	-0.695497	15	280272	281498	170945	180180	100092
YJR113C	YNL185C	0.759475	14	292190	292666	191240	191243	100947
YJR113C	YNL184C	0.635835	14	292554	292880	191240	191243	101311
YNR020C	YNL122C	0.886304	14	398020	398367	502316	502496	103949
YKL043W	YER038C	-0.594934	5	226857	228251	332264	350744	104013
YML050W	YNL081C	0.792871	14	476185	476616	368178	371953	104232
YEL024W	YLR231C	0.615572	12	605758	607119	712135	712139	105016
YIL034C	YOR155C	0.651782	15	626628	627980	519764	519776	106852
YNL185C	YIL157C	0.754972	9	46949	47542	154733	155027	107191
YER114C	YNL100W	-0.602439	14	437610	438314	330365	330368	107242
YLR460C	YNL100W	-0.623827	14	437610	438314	314157	330365	107245
YOR342C	YBR155W	0.648405	2	549728	550885	658746	670435	107861
YPL195W	YFL010C	-0.586116	6	115102	115737	5852	5854	109248
YOR277C	YLR185W	0.842402	12	522665	523290	634225	634227	110935
YOR342C	YBR061C	0.635835	2	364746	365678	252538	252640	112106
YGL110C	YMR042W	-0.621576	13	352602	353135	238291	239892	112710
YAR002W	YAR002W	1	1	152258	153877	37068	38887	113371
YCR071C	YNL052W	0.601876	14	531721	532182	418269	418293	113428
YIL034C	YOR159C	-0.630394	15	633282	633566	519764	519776	113506
YCR071C	YNL177C	0.707692	14	303683	304612	418269	418293	113657
YDR406W	YOL084W	0.751032	15	162355	165330	44548	47951	114404
YOR277C	YLR181C	0.797186	12	516680	517672	634225	634227	116553
YDR261C	YNL006W	-0.652908	14	620064	620975	449639	502316	117748
YLR253W	YNL006W	0.689869	14	620064	620975	486861	502316	117748
YNL233W	YLR314C	0.617636	12	762575	764137	642137	644136	118439
YDL178W	YNL005C	0.701689	14	621310	622425	486861	502496	118814
YMR098C	YNL005C	0.842964	14	621310	622425	502316	502496	118814
YDR261C	YNL005C	-0.610694	14	621310	622425	449639	502316	118994
YDR322W	YNL005C	0.870357	14	621310	622425	449639	502316	118994
YGL107C	YNL005C	0.867917	14	621310	622425	449639	502316	118994
YGL129C	YNL005C	0.844841	14	621310	622425	449639	502316	118994
YLR253W	YNL005C	0.700938	14	621310	622425	486861	502316	118994
YOR354C	YNL005C	0.836585	14	621310	622425	449639	502316	118994
YOR342C	YBR148W	-0.748218	2	537833	539662	658746	670435	119084
YGR086C	YOL084W	0.609193	15	162355	165330	43172	43217	119138
YIR038C	YOL084W	0.751407	15	162355	165330	43172	43217	119138
YER114C	YNL093W	-0.621764	14	449865	450527	330365	330368	119497
YDL178W	YNL004W	0.678799	14	623329	624618	486861	502496	120833
YBR185C	YNL073W	0.630206	14	488383	490113	610996	614342	120883

YDR390C	YLR277C	0.641088	12	697156	699495	823996	824230	124501
YCR071C	YNL185C	0.701501	14	292190	292666	418269	418293	125603
YOR342C	YBR143C	0.612383	2	530826	532139	658746	670435	126607
YLR295C	YKL169C	0.731707	11	130686	131069	258243	263612	127174
YPR020W	YKL169C	0.717448	11	130686	131069	258243	263612	127174
YLR295C	YKL170W	0.742777	11	130635	131051	258243	263612	127192
YPR020W	YKL170W	0.723077	11	130635	131051	258243	263612	127192
YMR202W	YLR331C	0.675797	12	790669	791046	659357	663370	127299
YDR390C	YLR275W	0.658724	12	694378	694800	823996	824230	129196
YNL339C	YLR385C	-0.693902	12	892992	893390	1023790	1042072	130400
YIL147C	YAR015W	0.67242	1	169370	170290	36900	38491	130879
YNR020C	YNL137C	0.798124	14	368592	370052	502316	502496	132264
YNL233W	YLR321C	0.637523	12	776584	777864	642137	644136	132448
YKR049C	YOR001W	-0.632833	15	326832	329033	180407	193911	132921
YAR033W	YGR038W	-0.902158	7	560683	561351	402833	427476	133207
YBR185C	YNL081C	0.748218	14	476185	476616	610996	614342	134380
YHR116W	YNL081C	0.837899	14	476185	476616	610996	614342	134380
YJL180C	YNL081C	0.836585	14	476185	476616	610996	614342	134380
YOR277C	YLR167W	0.820826	12	498949	499407	634225	634227	134818
YDL048C	YOL084W	0.687617	15	162355	165330	301077	348934	135747
YER114C	YNL086W	-0.707317	14	466331	466639	330365	330368	135963
YBR127C	YNL004W	-0.618199	14	623329	624618	418293	486861	136468
YER114C	YNL085W	0.707129	14	467128	469620	330365	330368	136760
YKR049C	YOR003W	0.602439	15	331455	332891	180407	193911	137544
YBR129C	YNL051W	0.663602	14	532655	533866	368178	393903	138752
YIL034C	YOR176W	0.757599	15	662401	663582	519764	519776	142625
YDR347W	YNL177C	0.772045	14	303683	304612	449639	502316	145027
YGR165W	YNL177C	0.901689	14	303683	304612	449639	502316	145027
YHR059W	YNL177C	0.863602	14	303683	304612	449639	502316	145027
YKL170W	YNL177C	0.868856	14	303683	304612	449639	502316	145027
YNL252C	YNL177C	0.958349	14	303683	304612	449639	502316	145027
YMR292W	YGL102C	0.674296	7	311506	311934	457215	460945	145281
YER114C	YNL081C	-0.657598	14	476185	476616	330365	330368	145817
YGR287C	YGR038W	-1.15	7	560683	561351	707605	707700	146254
YOR277C	YLR325C	0.871857	12	781143	781379	634225	634227	146916
YGL110C	YMR056C	-0.658349	13	387314	388243	238291	239892	147422
YER114C	YNL078W	0.602627	14	479765	480988	330365	330368	149397
YHR084W	YNL078W	0.603565	14	479765	480988	314157	330365	149400
YGL110C	YMR124W	0.667542	13	514455	517286	667328	685092	150042
YIR038C	YOL073C	0.771857	15	193831	194799	43172	43217	150614
YHL043W	YER084W	0.673171	5	327061	327447	480027	480051	152580

YGL110C	YMR061W	0.61576	13	392754	394787	238291	239892	152862
YDR347W	YNL184C	0.763602	14	292554	292880	449639	502316	156759
YNL184C	YNL184C	1	14	292554	292880	449639	502316	156759
YDR347W	YNL185C	0.839587	14	292190	292666	449639	502316	156973
YHR059W	YNL185C	0.731332	14	292190	292666	449639	502316	156973
YDR366C	YDR366C	1	4	1212426	1212824	1022764	1053084	159342
YOR277C	YLR333C	0.87955	12	795573	795899	634225	634227	161346
YIL034C	YOR185C	0.644841	15	681444	682106	519764	519776	161668
YDL178W	YNR020C	0.624203	14	667407	668219	486861	502496	164911
YMR098C	YNR020C	0.810319	14	667407	668219	502316	502496	164911
YGL107C	YNR020C	0.818949	14	667407	668219	449639	502316	165091
YGL129C	YNR020C	0.783302	14	667407	668219	449639	502316	165091
YOR354C	YNR020C	0.812195	14	667407	668219	449639	502316	165091
YOR277C	YLR336C	0.726454	12	799697	802396	634225	634227	165470
YPL195W	YFR016C	0.660225	6	177033	180734	5852	5854	171179
YLL050C	YOR068C	-0.657036	15	453870	454214	277470	282525	171345
YBR185C	YNL100W	0.771295	14	437610	438314	610996	614342	172682
YJL207C	YNL262W	0.621388	14	148211	154879	330368	340077	175489
YJR113C	YNL137C	0.696998	14	368592	370052	191240	191243	177349
YEL033W	YLR245C	-0.733959	12	626502	626930	808623	808707	181693
YDL048C	YOL109W	-0.687617	15	110296	110637	301077	348934	190440
YPL195W	YFR022W	-0.60394	6	196820	199021	5852	5854	190966
YIR038C	YOR086C	0.662664	15	483220	486780	679086	690475	192306
YOR277C	YLR150W	0.790994	12	440468	441289	634225	634227	192936
YOR277C	YLR149C	-0.733584	12	437632	439824	634225	634227	194401
YEL033W	YLR237W	-0.654784	12	612367	614163	808623	808707	194460
YNR020C	YNL177C	0.894372	14	303683	304612	502316	502496	197704
YLR425W	YKL010C	0.627767	11	421062	425513	209627	222724	198338
YIR038C	YOL047C	0.681614	15	241612	242745	43172	43217	198395
YDR078C	YDR078C	1	4	602192	602863	802851	828741	199988
YMR292W	YGL135W	0.691745	7	254644	255297	457215	460945	201918
YDL204W	YOR028C	0.703752	15	383532	384419	170945	180180	203352
YGR052W	YOR028C	0.757786	15	383532	384419	174364	180180	203352
YML128C	YOR028C	0.736585	15	383532	384419	170945	180180	203352
YOR173W	YOR028C	0.632458	15	383532	384419	170945	180180	203352
YHR036W	YLR143W	0.628143	12	427330	429387	634225	634227	204838
YJR113C	YNL122C	0.722326	14	398020	398367	191240	191243	206777
YPL195W	YFR031C	0.69212	6	216581	220093	5852	5854	210727
YBR185C	YNL122C	0.754597	14	398020	398367	610996	614342	212629
YHR116W	YNL122C	0.897561	14	398020	398367	610996	614342	212629
YER114C	YNL043C	-0.615197	14	545585	545905	330365	330368	215217

YKR049C	YOR046C	-0.637148	15	414459	415907	180407	193911	220548
YOR277C	YLR367W	0.797186	12	856441	857316	634225	634227	222214
YKR049C	YOR048C	-0.627204	15	418630	421650	180407	193911	224719
YIR038C	YOL029C	-0.660976	15	269815	270420	43172	43217	226598
YNL004W	YNL004W	1	14	623329	624618	368178	393903	229426
YKR049C	YOR051C	-0.618574	15	424847	426085	180407	193911	230936
YBR185C	YNL137C	0.751782	14	368592	370052	610996	614342	240944
YHR024C	YNL137C	0.666792	14	368592	370052	610996	614342	240944
YHR116W	YNL137C	0.829081	14	368592	370052	610996	614342	240944
YCR071C	YNL252C	0.681238	14	171440	172285	418269	418293	245984
YJR113C	YNL100W	0.663978	14	437610	438314	191240	191243	246367
YIL034C	YOR228C	0.700188	15	766869	767777	519764	519776	247093
YKR049C	YOR063W	-0.618011	15	444687	445850	180407	193911	250776
YIL034C	YOR230W	0.691932	15	770800	772113	519764	519776	251024
YIR038C	YOL014W	-0.622702	15	299693	300067	43172	43217	256476
YLL067C	YLR325C	-0.606567	12	781143	781379	1043086	1044161	261707
YOR277C	YLR388W	0.802627	12	898651	898821	634225	634227	264424
YIL034C	YOR242C	-0.722139	15	788742	789857	519764	519776	268966
YLL050C	YOR122C	0.705816	15	552298	552887	277470	282525	269773
YFR013W	YMR061W	0.661726	13	392754	394787	667328	686206	272541
YGR247W	YGR247W	1	7	984969	985688	711992	712004	272965
YJR113C	YNL086W	0.676173	14	466331	466639	191240	191243	275088
YOR200W	YNL086W	0.658724	14	466331	466639	191183	191243	275088
YPL008W	YPL008W	1	16	539380	541965	819251	825431	277286
YJR113C	YNL081C	0.679362	14	476185	476616	191240	191243	284942
YOR277C	YLR400W	0.784615	12	922062	922535	634225	634227	287835
YOR200W	YNL078W	-0.730394	14	479765	480988	191183	191243	288522
YPL273W	YMR062C	0.677486	13	395053	396378	686206	697899	289828
YOR277C	YLR406C	0.8606	12	931062	931752	634225	634227	296835
YOR277C	YLR409C	0.774859	12	934410	937229	634225	634227	300183
YDR390C	YLR183C	0.641276	12	520545	522014	823996	824230	301982
YOR173W	YOR086C	0.659099	15	483220	486780	170945	180180	303040
YDR390C	YLR182W	0.691557	12	517942	520353	823996	824230	303643
YHR116W	YNL177C	0.903189	14	303683	304612	610996	614342	306384
YMR292W	YGL189C	0.643715	7	148233	148592	457215	460945	308623
YHR109W	YOR150W	0.64015	15	611999	612490	937036	945781	324546
YLR165C	YLR165C	1	12	494496	495260	823996	824230	328736
YMR292W	YGL200C	0.670544	7	122697	123308	457215	460945	333907
YDR044W	YLR102C	-0.739962	12	342971	343768	677957	693790	334189
YDR044W	YLR101C	-0.641839	12	342567	342962	677957	693790	334995
YDR044W	YLR100W	-0.654409	12	341811	342854	677957	693790	335103

YKR021W	YLR111W	-0.633959	12	370392	370724	708041	710924	337317
YLL050C	YOR171C	-0.750657	15	652010	653884	277470	282525	369485
YMR292W	YGL226C-A	0.820263	7	72747	73156	457215	460945	384059
YNL254C	YMR083W	-0.623827	13	434787	435914	49894	49903	384884
YML079W	YML079W	1	13	110247	110852	503713	503715	392861
YOL159C	YGR038W	-1.27739	7	560683	561351	131386	135426	425257
YMR292W	YGL251C	-0.634522	7	27921	31636	457215	460945	425579
YGR052W	YOR173W	0.750657	15	657132	658325	174364	180180	476952
YDR390C	YLR103C	0.781426	12	343990	345942	823996	824230	478054
YLL050C	YOR235W	-0.655722	15	779870	780184	277470	282525	497345
YGL251C	YMR168C	0.649531	13	597331	599157	64970	69122	528209
YER185W	YER185W	1	5	559449	560360	6334	6335	553114

Vita

Nan Bing was born the first son of Sukun Zhang and Qingchen Bing on January 3, 1973 in Shenyang, Liaoning Province, China, where he grew up and received education until he came to the States. He attended Shenyang No. 2 High School, where he graduated 1st in class and the top 2% of the whole grade. He then attended China Medical University for a special seven-year program combining Bachelor and Master study. He received Bachelor Degree of Medicine (M.D. equivalent) and Master Degree of Medicine in July 1998. Then he entered University of Cincinnati, graduate program of Cell and Molecular Biology. After spending two years in Cincinnati, he decided to enter the field of Bioinformatics. In August 2000, he transferred to Virgin Tech (VPI&SU) to pursue the Ph.D. in Genetics, Bioinformatics and Computational Biology.