Meta-analysis and Combining Information in Genetics

Rudy Guerra, David Allison and Darlene R. Goldstein

Contents

Introduction Microarrays

CHAPTER 1

Comparison of meta-analysis to combined analysis of a replicated microarray study

Darlene R. Goldstein[1], Mauro Delorenzi[2], Ruth Luthi-Carter[3], and Thierry Sengstag[2]
[1]École Polytechnique Fédérale de Lausanne (EPFL), Institut de mathématiques, CH-1015 Lausanne, Switzerland;
[2]Bioinformatics Core Facility, Institut Suisse de Recherche Expérimentale sur le Cancer (ISREC), and Swiss Institute of Bioinformatics, CH-1066 Epalinges, Switzerland;
[3]École Polytechnique Fédérale de Lausanne (EPFL), Laboratoire de neurogénomique fonctionnelle, CH-1015 Lausanne, Switzerland

1.1 Introduction

Microarray technologies measure mRNA abundance for thousands of genes in parallel. The high throughput nature of microarrays has contributed to their rise in importance for studying the molecular basis of fundamental biological processes and complex disease traits. Whereas only a few years ago microarray experiments were uncommon, they are now regularly used in a great variety of biological and medical studies.

Several different types of microarray platforms are available. Those currently in common use include high-density short oligonucleotide arrays, such as Affymetrix GeneChip ® arrays; long oligonucleotide arrays, such as those produced by Agilent; and cDNA arrays, fabricated in laboratories on site at many academic and commercial institutions.

The widespread use of microarrays has resulted in a large-scale, rapid expansion of data. Many research groups throughout the world are engaged in gene expression studies of the same or similar conditions – specific cancers, for example. Data from many microarray studies are deposited in publicly available databases such as Gene Expression Omnibus Edgar et al. (2002); Barrett et al. (2005). It is hoped that ready access to the data will facilitate the integration of information across different studies.

Each microarray study gives rise to its own list of ‘interesting’ genes. The lists from different studies, however, may not exhibit substantial concordance. Discordant results may produce scientific confusion or disagreement regarding the underlying biology, as well as lost time and misused resources. Consequently, the ability to synthesize information across studies is essential.

Meta-analysis consists of statistical methods for combining results of independent studies addressing related questions. One aim of combining results is to obtain increased power – studies with small sample sizes are less likely to find effects even when they exist. Putting results together increases the effective sample size, thereby allowing more precise effect estimation and increasing power. The uncovering of a significant effect from a combined analysis, where individual studies do not make positive findings at the same significance level, has been referred to in the microarray meta-analysis literature as ‘integration-driven discovery’ (IDD) Choi et al. (2003).

Given the limited size of most microarray studies to date, meta-analysis thus seems a natural approach to the problem of integrating conclusions from different microarray studies. Indeed, there is a recent and increasing literature for meta-analysis of microarray studies Rhodes et al. (2002, 2004a); Ghosh et al. (2003); Choi et al. (2003); Stevens and Doerge (2005). Meta-analysis is not without problems, however.

A major difficulty with synthesizing results is the occurrence of study heterogeneity. Studies which are apparently similar may in fact differ in many ways, some of which may be quite subtle Sutton et al. (2000). In general, studies carried out by different research groups may vary in:

- scientific research goals
- population of interest
- design
- quality of implementation
- subject inclusion and exclusion criteria
- baseline status of subjects (even with the same selection criteria)
- treatment dosage and timing
- management of study subjects
- outcome definition or measures
- statistical methods of analysis.

Additional issues more specific to the microarray context include:

- differences in the technology used for the study
- heterogeneity of measured expression from the same probe occurring multiple times on the array
- multiple (different) probes for the same gene
- variability in probes used by different platforms
- differences in quantification of gene expression, even when the same technology is used.

In this chapter, we examine properties of different methods for combining information from what is essentially a replicated experiment carried out with Affymetrix GeneChips. Our aim is to demonstrate that even in this almost ideal situation, several issues concerning appropriate data normalization and combination still arise. We first give some background on the study, then describe the statistical analyses and present results, and conclude with a discussion. Because the focus here is on meta-analysis, we treat only briefly the specifics of microarray data analysis. For further details on these aspects see e.g. Goldstein and Delorenzi (2004) for a review or `http://www.nslij-genetics.org/microarray/` for a bibliography of papers on microarray data analysis.

1.2 Study description

The data were obtained from two experiments on the R6/2 mouse. The R6/2 mouse line is transgenic for exon 1 of the human Huntington's disease (HD) gene, thus serving as an experimental model for the disease Mangiarini et al. (1996). These mice exhibit mRNA changes weeks in advance of neuronal death or gliosis phenotypes Luthi-Carter et al. (2000, 2002a).

Two separate studies were carried out to investigate the effects on gene expression of different drugs on HD and normal (or wild type (WT)) mice in order to identify genes differentially expressed between HD and WT mice. Each experiment was designed as a 2x2 factorial layout, where one factor is drug/placebo treatment and the other is HD/WT mouse.

We consider only the control groups for the two studies, which received the placebo (injected with normal saline 30-60 minutes prior to sacrifice). In Study I there were 8 control mice, while in Study II there were 6 control mice. In each study, half of the mice were HD and half WT.

The two experiments were carried out by the same laboratory a few months apart. In each experiment, the same protocols were used throughout with regard to mouse breeding, care and sacrifice, mRNA extraction, and hybridization to the microarray. Thus, these data are essentially those of a completely replicated study.

Affymetrix GeneChips contain several (usually 11 – 20) 25-mer oligonucleotides used to measure the abundance of a given target sequence, the perfect match (PM) probes, as well as an equal number of negative controls, the mismatch (MM) probes. The set of probes for a given target sequence is called a probe set. A single fluorescently labeled sample is hybridized to the array which is then scanned with a laser, yielding absolute measures of fluorescence intensity. The intensities are indicative of the amounts of mRNAs containing the target sequence in the sample, and thus provide a means of quantifying levels of gene expression.

The studies were carried out with the Affymetrix MOE 430A (Mouse Expression Array). These chips contain in total 22,690 probe sets, to which, with a slight abuse of terminology, we refer henceforth as 'genes'. The data are deposited in GEO with series accession number GSE1980, and should be publicly available by the end of 2005 at `http://www.ncbi.nlm.nih.gov/geo/`.

1.3 Statistical analyses

There are several components of the data analyses to be carried out. First, quality of the hybridizations should be assessed so that low quality chips are removed from further analysis. Before statistical analyses can take place, the primary data obtained from scanning and image analysis of arrays must first be quantified using a measure of gene expression. Once there is a measure of expression of each gene for each individual, we compute for each gene a statistic for assessing genes for differential expression between the HD and WT mice. Some determination of significance should also be made for these statistics, taking into account the multiplicity of hypotheses tested.

We compare analyses carried out under two scenarios: one, where the data are combined and analyzed as a single set; and two, where the two data sets are analyzed separately and their results are combined via meta-analysis. The steps are described in detail below. All analyses reported here were coded in the R statistical programming environment Ihaka and Gentleman (1996); R Development Core Team (200), using the following packages from R (2.0.1) and BioConductor (release 1.5) Gentleman et al. (2004): `affy` Irizarry et al. (2004), `affyPLM` Bolstad (2004), `car` Fox (2005), `limma` Smyth (2004), `qvalue` Dabney and with assistance from Gregory R. Warnes, and `rmeta` Lumley (2004).

1.3.1 Chip quality assessment

We assessed all chips for quality with the RMA-QC approach described in Collin (2004) and implemented in the BioConductor R package `affyPLM` Bolstad (2004). In this method, gene expression is modeled as the sum of chip and probe effects, with the model fit by robust regression (i.e. outliers are downweighted; see equation 1.1 below). Pseudoimages of the robust regression weights or residuals for each probe provide a graphical means to assess chip quality; numerical measures indicative of quality were also computed.

By these criteria, all 14 chips were of similar and suitably high quality that none required exclusion.

1.3.2 Quantifying gene expression

Several methods of quantifying gene expression from probe fluorescence intensities on Affymetrix GeneChips are in popular use, e.g. MAS5/GCOS Affymetrix (2001), the Li-Wong method, implemented in dChip Li and Wong (2001), and Robust Multichip Average (RMA) Irizarry et al. (2003a), among many others. For a comparison of methods see `http://affycomp.biostat.jhsph.edu/` Cope et al. (2004), where it is easily seen that no method is best under every circumstance. We have chosen to use RMA, due to its demonstrated favorable properties Irizarry et al. (2003a,b); Bolstad et al. (2003).

An important yet difficult aspect of gene expression quantification is normalization. (The term ‘normalization’ as used here is not related to the normal, or Gaussian, distribution.) The purpose of normalization is remove the effects of systematic variation other than that due to the effect of interest. Examples of such variation include differences in sample preparation, scanning intensities, and variability among chips. Ideally, any observed differences in gene expression remaining after normalization are due to differential expression rather than artifactual differences in measured expression.

RMA consists of three steps: a background adjustment, quantile normalization and probe set summarization. Background is estimated assuming that the observed signal is the convolution of an exponential signal with Gaussian background (noise). Quantile normalization forces equality of quantiles across samples. Such a normalization is appropriate assuming that the true distributions of intensities are the same in all samples (of course, the same probe may occur at different quantiles across samples). For each probe set on the chip, the $\log_2$ background-corrected normalized signal $\log_2 b(PM_{ij})$ is modeled as

$$\log_2 b(PM_{ij}) = \mu_i + \alpha_j + \epsilon_{ij}, \tag{1.1}$$

where μ_i is the summary measure of expression for the given probe set on chip i, α_j is a probe-specific effect, and ϵ_{ij} are independently and identically distributed mean 0 errors Irizarry et al. (2003b). For parameter identifiability, it is assumed that $\sum_j \alpha_j = 0$. The model is fit via median polish Tukey (1977); the estimated chip effect μ_i is the RMA value of the probe set for chip i. RMA values were computed with the `affy` package.

1.3.3 Identifying differential expression

A commonly addressed problem in microarray experiments is detection of genes differentially expressed under two or more conditions. A substantial number of statistical papers propose methods for this purpose, with new ones still being introduced (for an overview see Goldstein and Delorenzi (2004)). The high dimensionality of microarray data has also brought to the fore multiple hypothesis testing issues. The approach we adopt is described here.

Moderated t-statistic

Perhaps the most readily interpretable measure of differential expression is given by the fold change (ratio) in expression of a given gene between two types of samples

(HD and WT here). It is more convenient to consider fold change on the logarithmic scale, M = (average) $\log_2$(fold change).

The measure M has the shortcoming of not taking into account differing variability of different genes. The variability of M, though, is not the same across the range of signal intensities. In particular, genes with larger variance across arrays are likely to produce large values of M even when they are not truly differentially expressed between the two sample types.

An obvious way to deal with differing variability is by standardization. Here M is divided by its standard error, which is estimated based on expression measures of the corresponding gene. Thus, the difference in average expression between sample types is quantified with a t-statistic. However, a problem here is that the t-statistic performs very poorly at identifying true differential expression with the small sample sizes found in typical microarray studies.

Bayesian and empirical Bayes methods have been proposed as a compromise between single gene estimates of variability and no estimate of variability at all. These use data from all genes to improve estimation of differential expression for single genes Lönnstedt and Speed (2002); Smyth (2004). These methods have been shown to perform well, in terms of true and false positive and negative rates, at identifying differential expression. In addition, the methods have been extended to be applied to a large variety of experimental designs through a linear modeling approach Smyth (2004); Lönnstedt et al. (2001).

We follow the linear modeling approach here. For each gene g in a given study, the measured gene expression vector Y_g across samples is modeled as

$$Y_g = X\beta_g + \epsilon_g,$$

where X is the design matrix, β_g is a vector of coefficients, and ϵ_g is a vector of error terms. The design matrix X is the same for all genes within a study.

The moderated t-statistic for coefficient j and gene g is given by

$$\text{mod } t_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}},$$

where $\hat{\beta}_{gj}$ is the estimate of coefficient j for gene g, $\tilde{s}_g$ is the square root of the empirical Bayes shrinkage estimated variance, and v_{gj} is the scaling for the variance, reflecting sample size. That is, mod t is the ratio of M to its standard error, which has now been estimated taking into account expression levels not only of gene g but of all genes. It is similar to the ordinary t-statistic, but with a moderated standard error estimate and correspondingly an increased number of degrees of freedom. For a detailed explanation, refer to Smyth (2004). We base inference about effects on mod t.

Multiple hypothesis testing

The biological question of differential expression can be restated as a problem in multiple hypothesis testing: the simultaneous test for each gene of the null hypothesis of identical mean expression in the two sample types.

A multiplicity problem arises when attempting to assess the statistical significance of the results on tests carried out on several thousands of genes simultaneously. With whole genome coverage arrays consisting of probes for many thousands of genes, most genes will not be differentially expressed between the conditions under investigation. Thus, even a nominal p-value of say 0.01 cannot be characterized as 'significant', since such small p-values will occur by chance when such a large number of tests are made.

Classical approaches to correction for multiple testing focus on control of the family-wise error rate (FWER), or probability of at least one false positive result in all tested hypotheses. The resulting procedures tend to be depressingly conservative though. Recent developments in controlling the false discovery rate (FDR), or expected proportion of false positive findings among the rejected hypotheses, appear to provide a promising way to come up with meaningful significance measures among thousands of genes Benjamini and Hochberg (1995a); Reiner et al. (2003). In general, procedures controlling the FDR are typically less conservative than those controlling the FWER. FDR control thus seems well suited for microarray studies.

The (nominal, unadjusted) p-value of a test reflects significance only for a single gene considered in isolation. The q-value of a test measures the proportion of false positives (FDR) incurred among rejected nulls when that test is called significant. It has been described as the expected proportion of false positives among all test results as or more extreme than the one obtained Storey (2002b); Storey and Tibshirani (2003).

We make use of q-values to take into account the large number of individual hypotheses tested. The mod t p-values from a set of single gene tests are transformed to q-values with the `qvalue` package Dabney and with assistance from Gregory R. Warnes. We call a test result 'significant' by fixing a q-value (or FDR) threshold, usually at 0.05. In many microarray studies a higher threshold may be more relevant. For example, a FDR of 0.25 still suggests that three of four significant findings are real. This may be all that is attainable with study sizes available in practice.

1.3.4 Combined data analysis

In the combined data analysis, we consider all 14 chips as a single data set from the same experiment. This is not a completely artificial treatment, as most large experiments take place over a period of time and include hybridizing groups of chips at different times. RMA measures are obtained by quantifying expression, including normalization, on all chips together.

The linear modeling approach is used to identify genes differentially expressed between HD and WT mice. We consider a series of models, each of which includes an effect on gene expression of HD over WT. There might also be additional variability due to study, so we also allow for a study, or ‘batch’, effect as well as entertain the possibility of an HD by study interaction.

The design matrices are set up using treatment contrasts so that the effects of interest are included as coefficients in the models. Thus, for each gene g, the three combined data linear models are given by (the subscript g is suppressed; I represents an indicator random variable):

Model A: $y = \beta_0 + \beta_{HD} I_{\{HD=1\}} + \epsilon$

Model B: $y = \beta_0 + \beta_{HD} I_{\{HD=1\}} + \beta_{batch} I_{\{batch=1\}} + \epsilon$

Model C: $y = \beta_0 + \beta_{HD} I_{\{HD=1\}} + \beta_{batch} I_{\{batch=I\}} + \beta_{HD \times batch} I_{\{HD \times batch=1\}} + \epsilon$.

The coefficients are estimated by ordinary least squares. The `limma` package is used to compute for each gene under each model the statistic mod t and corresponding p-values as a prelude to obtaining q-values.

1.3.5 Meta-analysis

In the meta-analyses, each experiment is first analyzed as a separate study. After heterogeneity analysis, results from the two studies are combined under three meta-analytic techniques: fixed effects meta-analysis, random effects meta-analysis, and Fisher p-value combination. Computations were done with the R package `rmeta` Lumley (2004).

In the separate study analyses, gene expression is again quantified with RMA, but values are computed using only chips from the same study (8 chips for Study I or 6 chips for Study II). Linear modeling is carried out as above, but because each study is analyzed individually the model includes only the HD effect (Model A). Fitting the model produces effect estimates (coefficients) for each gene, while the empirical Bayes procedure produces shrinkage estimates of variance, moderated t-statistics and p-values, which in turn yield q-values and gene rankings based on evidence in favor of differential expression between HD and WT mice.

Heterogeneity analysis

Prior to combining effect sizes from different studies, it is important to verify that they are homogeneous – that is, that they all seem to be estimating the same underlying population parameter. Existing graphical methods for assessing inter-study heterogeneity, such as forest plots of individual study confidence intervals, seem of limited usefulness in the microarray setting, as one such plot would be required for each individual gene. We thus depend on numerical assessments to screen genes for heterogeneous treatment effects across studies.

The standard test of homogeneity Cochran (1954) tests, for each gene g, the null hypothesis of homogeneity of treatment effects β_i in k studies (the subscript g is suppressed)

$$H_0\colon\ \beta_1 = \beta_2 = \cdots = \beta_k$$

against the general alternative that at least one β_i is different. The test statistic Q is given by

$$Q = \sum_{i=1}^{k} w_i(\hat{\beta}_i - \bar{\beta}_.)^2, \tag{1.2}$$

where $\hat{\beta}_i$ estimates the treatment effect (the HD coefficient in the linear model for a given gene) in study i, w_i is the weight given to study i (most commonly taken as the reciprocal of the variance of the outcome estimate), and $\bar{\beta}_.$ is the weighted average treatment effect

$$\bar{\beta}_. = \frac{\sum_i w_i \hat{\beta}_i}{\sum_i w_i}. \tag{1.3}$$

Under the null hypothesis, the distribution of Q is approximately χ^2_{k-1}.

In the event that the null hypothesis is not rejected, any differences between studies are assumed to be due to chance variation, and it is considered appropriate to combine estimates via a fixed effects model. A major limitation of this approach, though, is the low power of the test to detect even substantial heterogeneity due to small sample sizes or a small number of studies. One way to avoid the risk of combining heterogeneous results is to relax the significance criterion from 0.05 to 0.10, say.

If instead the test shows that significant heterogeneity exists between study results, then combination via a random effects model is typically favored. Where possible, heterogeneity should be scrutinized rather than ignored, with an aim toward explaining important study differences Bailey (1987). Because our studies were carried out by the same laboratory using identical protocols, tracking down reasons that some genes show heterogeneity across studies while others do not seems an unsolvable problem.

It should also be kept in mind that there is one homogeneity test per gene, so the usual caveats regarding multiple hypothesis testing apply.

Fixed effects meta-analysis

Fixed effect (FE) meta-analysis assumes no heterogeneity between results of the different studies and therefore that a fixed effects model can be used to estimate the assumed common underlying treatment effect. In FE meta-analysis, each individual study estimate receives weight inversely proportional to its variance. The weighted estimates are pooled as above to yield the estimate of the treatment effect given by equation 1.3, where the weights w_i are inversely proportional to the variances. These weights are used as they minimize the variance of the combined estimate $\bar{\beta}_.$ Cooper and Hedges (1994). The variance of the weighted estimator is just $1/\sum_{i=1}^{k} w_i$. Under the assumption of normality of $\bar{\beta}_{.}$, a p-value for each single gene test of HD effect (i.e. differential expression between HD and WT for the given gene) is readily obtained; corresponding q-values are obtained from the set of p-values across all genes.

Random effects meta-analysis

If the study results do exhibit heterogeneity, then there is assumed to be no single underlying value of HD effect but rather a distribution of values. In the presence of heterogeneity, differences among study results are considered to arise from inter-study variation of true effect size as well as chance variation. Use of a FE model understates the true degree of variability of $\bar{\beta}_{.}$, resulting in p-values which are artificially low. A more conservative approach is to use a model which accounts for the additional source of variability due to study.

Random effects (RE) meta-analysis assumes that individual studies may be estimating different treatment effects. The aim is to estimate characteristics of the distribution of effects, particularly the mean population effect size and between study variance of effect sizes. As in the FE case we use weighted estimates, but the weights are adjusted to take into account the additional variability between studies:

$$w_i^* = \frac{1}{(1/w_i) + \hat{\tau}^2},$$

where $\hat{\tau}$ estimates inter-study variability (see Cooper and Hedges (1994) for a derivation). The estimated mean treatment effect is given by equation 1.3, but with w_i^* in place of the w_i. Similarly, the variance of the weighted estimator is now given by $1/\sum_{i=1}^{k} w_i^*$. When the inter-study variance is estimated as 0, the RE model reduces to the FE model.

As for the FE model, single gene p-values from the RE model are obtained assuming normality of the effect distribution; q-values are then computed from the p-values across all genes.

Meta-analysis by Fisher p-value combination

In FE and RE meta-analysis, combined estimates of effect size provide the basis for analysis. Other methods of meta-analysis, dating back to at least the 1930s, are based on combining the p-values from independent studies. Although it is usually preferable to base inference on effect sizes, there are situations for which combining p-values may be considered justified – for example, when only p-values are reported without a corresponding estimate of effect size, or when study characteristics (design, treatment levels) are sufficiently different that combining effect estimates seems unacceptable Hasselblad (1995).

Several methods exist for combining p-values. One popular method is due to Fisher

Fisher (1932). Under the null hypothesis of no treatment effect, the individual study p-values p_i are independent uniformly distributed $U(0,1)$ random variables. Upon rescaling, the Fisher summary test statistic is given by

$$S = -2\sum_{i=1}^{k} \log(p_i). \tag{1.4}$$

To assess the significance of the Fisher statistic S we need to determine its p-value. The theoretical null distribution of S should be χ^2_{2k} (here $2k$ = 4).

We compute p-values for the Fisher combined p-value statistic S in two ways: first, with the χ^2_4 approximation and second, by a resampling procedure proposed in Rhodes et al. (2002). In the resampling procedure, rather than choosing a p-value at random from $U(0,1)$ a p-value is instead chosen at random from each of the sets of p-values from the two studies. These are then combined as in equation 1.4 into a randomized summary statistic S^R. We obtain an empirical distribution of S^R by repeating the resampling procedure 100,000 times. The p-value for the Fisher S statistic is estimated as the proportion of the resampling-generated statistics S_i^R greater than or equal to the original observed value S. This method yields a more conservative estimate for the p-value of S because the distribution of actual study p-values is not uniform.

1.4 Results

Here we present detailed results of the analyses outlined above for the combined data set and for meta-analyses of separate experimental study outcomes.

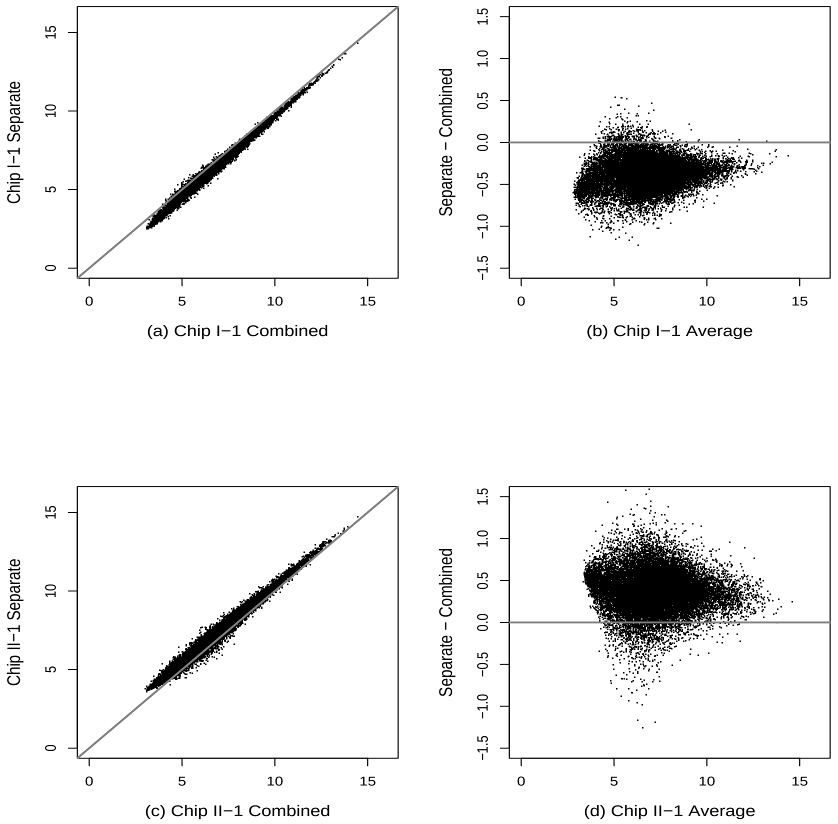
1.4.1 Combined data

Combined vs. separate gene expression quantification

Because the first step of analysis requires a measure of gene expression, we compared quantification of expression with the combined data (RMA values based on all 14 chips) to individual study quantification (we separately compute RMA values based on the 8 chips from Study I and RMA values from the 6 chips from Study II). Figure 1.1 contains plots to explore this comparison.

Figures 1.1(a) and (c) show separate versus combined RMA values for one chip from each study (chip 1 from Study I and chip 1 from Study II, or Chips I-1 and II-1). These chips are representative of all chips in the respective studies – all plots were quite similar within study – so our remarks on the plots apply to all chips. Each gene on the chip is represented by a point in the plot, with the diagonal line representing equal expression by each method.

As it is difficult to detect differences from the line of equality in these scatter plots,



(a) Chip I−1 Combined

(b) Chip I−1 Average

(c) Chip II−1 Combined

(d) Chip II−1 Average

Figure 1.1: Comparison of RMA values when studies combined and separate. (a) RMA values for chip I-1 computed within Study I vs. values computed from all chips combined; (b) Difference (Separate – Combined) vs. Average RMA values for chip I-1; (c) RMA values for chip II-1 computed within Study I vs. values computed from all chips combined; (d) Difference vs. Average RMA values for chip II-1. Diagonal and horizontal lines indicate equal values under both methods.

we have also plotted the corresponding rotated and rescaled version, the Difference-Average plot Tukey (1977) (a specific version of this plot is also called an MA plot in the microarray literature; figures 1.1(b) and (d)). In this representation, the difference between RMA values computed separately and combined is plotted against the average of the two values for the chip. If both RMA values were identical, all points would fall on the horizontal line at 0. Differences are more readily detected in this version of the plot.

It is easily seen that Study I chips tends to have higher RMA values when all chips are combined, while Study II chips have lower RMA values in the combined data set. The tendency persists throughout the range of ($\log_2$) signal intensities.

Many investigators have assumed that normalization of a set of chips together would remove artifacts of this nature. In fact, this does not appear to be the case at all. The persistence of the study batch artifact can be seen, for instance, using cluster analysis Everitt et al. (2001); Kaufman and Rousseeuw (1990). When we cluster samples (chips) based on gene expression, the ones from the same study cluster together. The clustering details (algorithm, dissimilarity measure, number of genes) do not seem to affect the cluster results to any great degree.

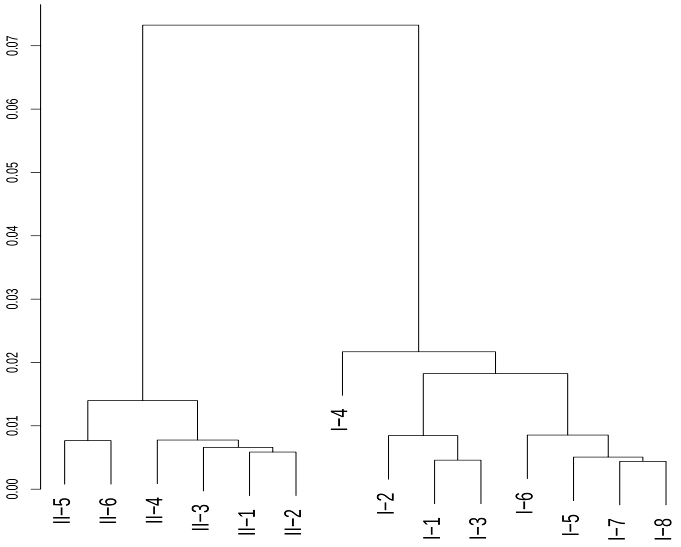


Figure 1.2: Cluster dendrogram of combined data RMA values. Samples are clustered using all genes, Ward's method and 1 – correlation dissimilarity.

Figure 1.2 shows an example of a dendrogram obtained clustering samples using all genes with Ward's method of clustering and 1 – correlation dissimilarity. The major cluster split occurs between Study I chips and Study II chips. There is a minor, and less clean, split on HD status: Study I samples 1 – 4 and Study II samples 1 – 3 are from the HD mice, the rest are WT mice. Thus, we see that even in what we might expect to be quite homogeneous studies, the most striking difference is in fact purely artifactual and is exactly of the sort that normalization is meant to remove. It is not quite clear why the effects persist, they do not appear to vary systematically with intensity (data not shown). However, quantile normalization of the entire set together does not remove the study batch effect. The batch effect must be removed in another way before reliable inference relating to differential expression can take place.

HD-study batch interaction

We next turn attention to linear modeling of gene expression in terms of the effects of interest. Here, gene expression is obtained by computing RMA values for the combined set of 14 chips. Although the primary focus is on the HD effect, we must also consider the ramifications of other potential terms for the model. We have just seen the need to include study batch in the model. We now consider Model C to assess the need to include the HD by batch interaction term.

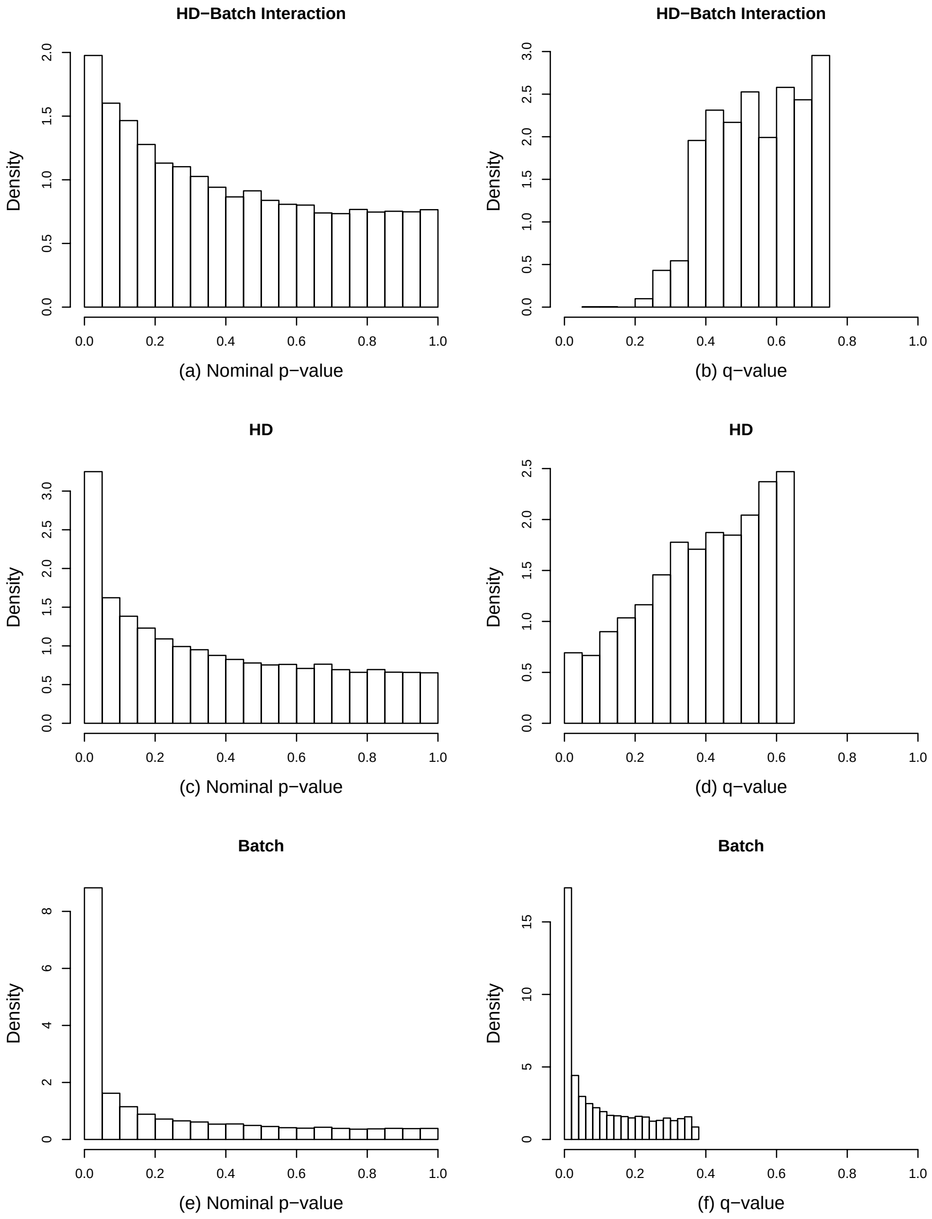
Histograms of p-values and q-values for the estimated interaction effects are shown in figures 1.3(a) and (b). There are 2242 genes with unadjusted p-values less than 0.05, but only 3 q-values less than 0.10 and thus indication of interaction between HD status and batch for only a few genes. In the face of this mild evidence, we discard Model C and ignore the possibility of interaction in the rest of the analyses.

Detection of differentially expressed genes

On the other hand, we have seen that there is strong evidence of batch effects for many genes. Using Model B to estimate HD and batch effects, we find evidence of significant HD effects for several genes along with a staggering number of genes with strong batch effects (figure 1.3(b) – (d)). While there are 785 genes with HD q-values < 0.05, nearly one half of the genes (10571 out of 22690) have batch effects with q-values < 0.05. It is not necessary to believe in the exactness of the p- and q-value estimation to conclude that there are many genes with strong batch effects.

We consider Model A, which contains only the HD term, in order to compare genes identified as differentially expressed between HD and WT mice with and without batch effects. Not surprisingly, the significance of the HD effect is always higher (q-value lower) for Model B, where a study batch effect is included in the model. This is because we have controlled for an important source of variability here by introducing an effective stratification factor (batch). There is within stratum homogeneity but heterogeneity between strata, resulting in increased power to detect HD differences.

Figure 1.4 displays the q-values for individual gene HD effects both with and without batch effects in the model. This plot shows the importance of the batch effect in



(a) Nominal p−value

(b) q−value

(c) Nominal p−value

(d) q−value

(e) Nominal p−value

(f) q−value

Figure 1.3: Histograms of p-values and q-values for Model C interaction term (a, b), and Model B HD (c, d) and study batch (e, f) effects.

uncovering differential expression due to HD status. Use of Model B produces an additional 681 significant genes at the same FDR of 0.05 (points in black to the left of the vertical line and above the horizontal line at 0.05).

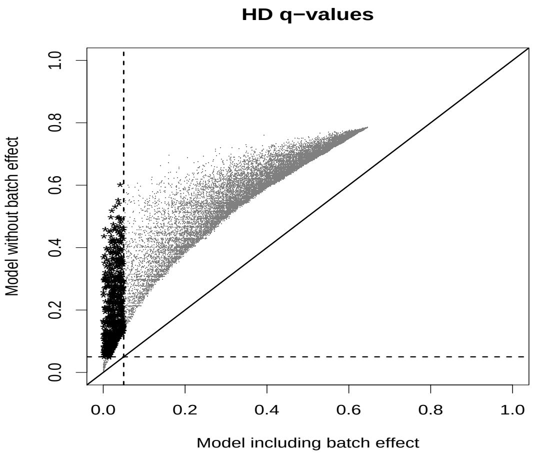


Figure 1.4: q-values for HD effects without (Model A) vs. with (Model B) batch effect. Solid diagonal line indicates equal values; dashed vertical and horizontal lines indicate a FDR of 0.05. Highlighted points are genes with significant HD effects for Model B but not for Model A.

A list of 'interesting' genes, those identified by the analysis as differentially expressed between HD and WT mice, can be produced upon selection of a significance threshold for the HD effect q-value, such as a FDR of 0.05. For comparison with results of meta-analyses of Studies I and II (below), we retain not only the gene list but all of the mod t p-values and corresponding q-values obtained using Model B.

1.4.2 Meta-analysis of Study I and Study II

Another strategy for dealing with study batch effects is to quantify gene expression separately for each study and then combine the studies by meta-analytic techniques. This is how the problem would necessarily be handled in the case of unrelated studies carried out by different research groups. Here, we are able to examine how meta-analysis would compare with a combined data analysis.

Heterogeneity analysis

We investigate heterogeneity of gene-specific HD estimates from Study I and Study II by computing the statistic Q (equation 1.2) as well as estimating the inter-study standard deviation (SD) τ for each gene. Characteristics of these are plotted in figure 1.5.

There is evidence of HD effect heterogeneity for some genes. Several genes have small nominal, unadjusted p-values: 3273 for $p < 0.05$; 5230 for $p < 0.10$ (figure 1.5(a)). If we choose a more stringent criterion of significance, say a q-value of 0.10, there are still 802 genes with significant heterogeneity. This is substantially larger than the number of genes for which an interaction effect was detected, but also very much smaller than the number with a significant batch effect.

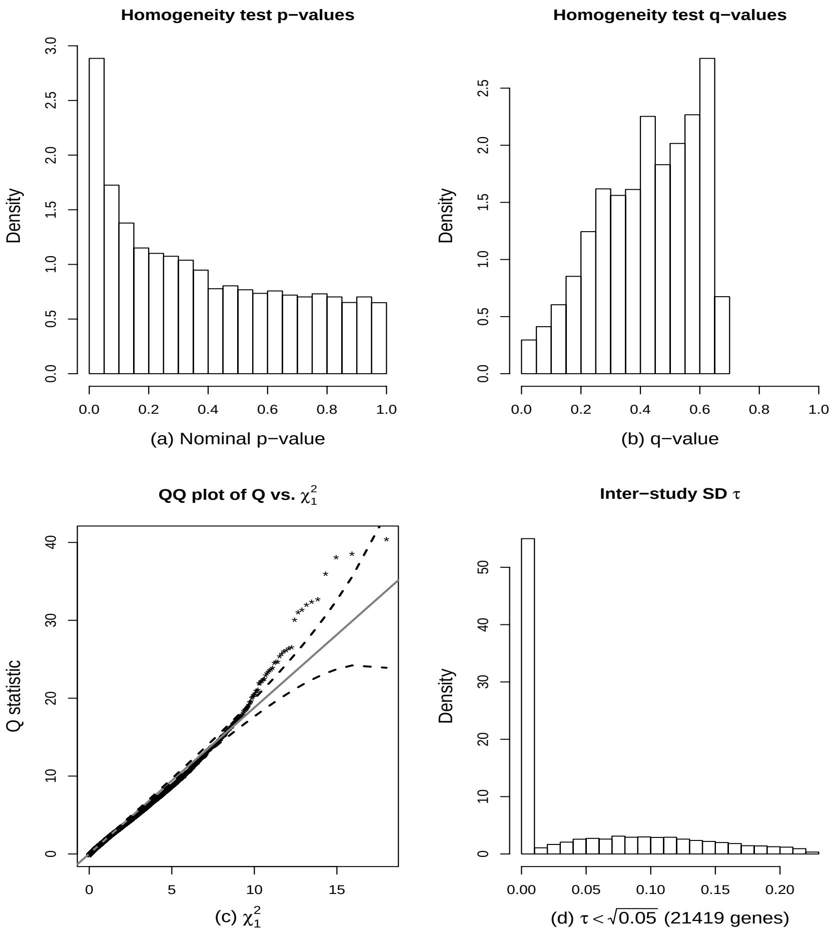
The quantile-quantile plot (figure 1.5(c)) shows some deviation from the assumed χ_1^2 null distribution. This could be due to inadequacy of the χ_1^2 approximation, but as it is unlikely that all the nulls are in fact true we instead interpret this as indicative of the presence of genes for which the alternative holds, i.e. there is some true heterogeneity. Since the χ^2 test has low power to detect heterogeneity for the small study number and sample sizes that we have, there is likely to be a greater degree of heterogeneity, and for more genes, than suggested here.

The distribution of estimated inter-study SD $\hat{\tau}$ is highly skewed. Over 70% (16428) of genes have $\hat{\tau} < 0.01$ (figure 1.5(d)). The value of $\hat{\tau}$ corresponding to a FDR of 0.10 is about 0.066. This gives some idea of what a ‘large’ value of τ is in this context. An example of intensities for a gene displaying heterogeneity is provided in Table 1.1, which gives the individual study RMA values of each chip for a gene with $\hat{\tau} \approx 0.1$.

Table 1.1: Individual Chip RMA Values for a Gene with $\hat{\tau} \approx 0.1$

	Chip number								Summary	
	1	2	3	4	5	6	7	8	Mean	SD
Study I	8.67	8.88	8.91	8.64	9.08	9.02	9.02	9.27	8.94	0.21
Study II	9.86	9.75	9.96	9.83	9.57	9.73			9.78	0.13

One purpose of testing homogeneity is for deciding between the FE and RE model for combining effect estimates. There will not be a large difference between FE and RE meta-analysis for genes with small τ. A general recommendation which has been made is to carry out both then compare similarity of results. If the results are similar then there is unlikely to be important heterogeneity and the FE model would typically be reported. If results are different, it is usually considered preferable to use RE meta-analysis to estimate the mean and SD of the effect size distribution. In the case of extreme, unexplained heterogeneity, it is probably more suitable to avoid combining the study results at all.



(a) Nominal p−value

(b) q−value

(c) χ_1^2

(d) $\tau < \sqrt{0.05}$ (21419 genes)

Figure 1.5: Plots for heterogeneity analysis: histograms of (a) p-values and (b) q-values for the homogeneity statistic Q; (c) quantile-quantile plot of Q compared to the theoretical χ_1^2 distribution with 95% confidence region (dashed lines); (d) histogram of gene-specific estimated inter-study SD for the 21419 genes for which $\hat{\tau} < \sqrt{.05}$.

Fixed effects and random effects meta-analysis

Based on the results of the heterogeneity analyses, we would choose to adopt the RE model for combining HD effect size estimates. Nevertheless, we investigate both approaches here in order to compare them.

Figure 1.6(a) compares the combined HD (mean) effect estimated by the RE model versus the FE model combined estimate. With the exception of a few genes, these combined estimates tend to be remarkably similar.

Due to the additional variability included by the RE model, however, there is a great deal of difference between the standardized estimates (figure 1.6(b)). Here, we can see that the FE standardized estimates are stochastically larger than the RE ones: about half of the genes have identical results for FE and RE, but for only 949 genes (or 4%) is the RE standardized estimate larger than that of FE. This phenomenon is also clearly reflected by the distributions of the corresponding q-values (figures 1.6(c) and (d)) – at any FDR, many more genes are called differentially expressed between HD and WT by the FE model.

Figure 1.7 shows how the methods compare for compare for different degrees of heterogeneity. In figures 1.7(a), (b) and (c), q-values are transformed by $-\log_{10}$ so that larger values are more significant. We see that significant effects in the RE meta-analysis tend to be for more genes with more homogeneous effects across studies (figures 1.7(a), (b)); that is, genes for which the inter-study variability does not overwhelm the size of the estimated effect. The q-values from FE and RE are compared directly in figure 1.7(c), where it is seen that the FE combined HD effect estimate is more significant than that of RE for virtually all genes. Finally, the location (centered at 0) and flatness of the loess curvefigure 1.7(d) show that heterogeneity is not more frequently found for larger estimated mean HD effect sizes.

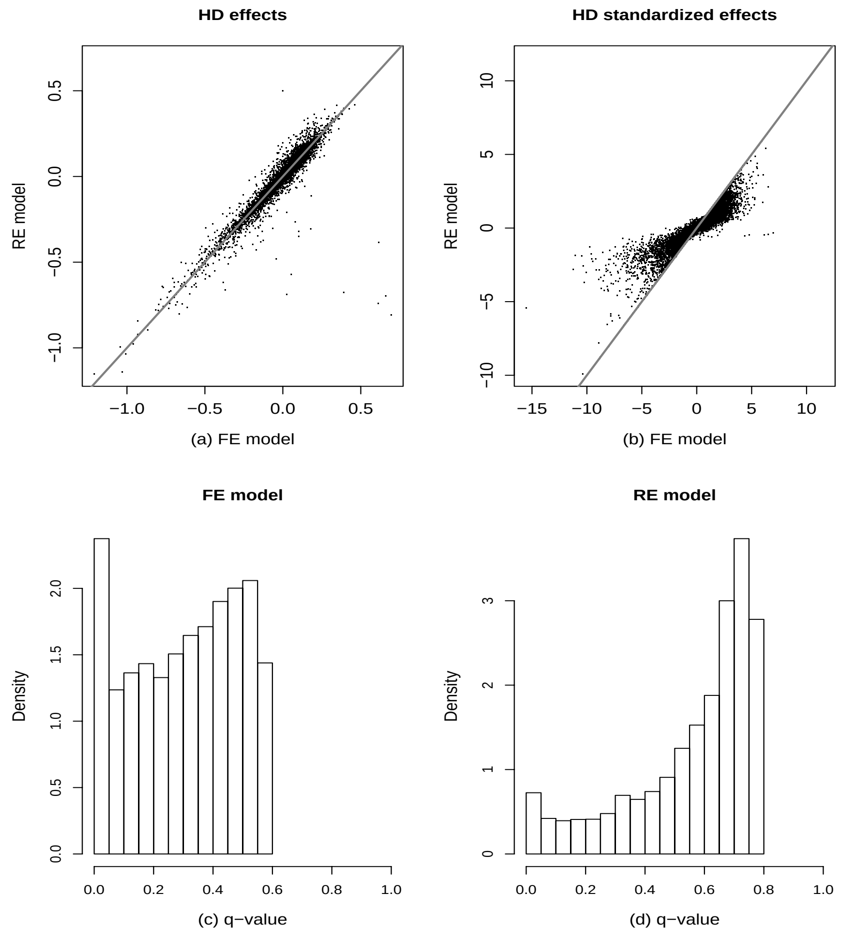


Figure 1.6: Comparison of HD effects for FE and RE meta-analysis. (a) HD effects estimated by RE vs. FE; (b) HD standardized effects estimated by RE vs. FE; q-values for HD (standardized) effects estimated by FE (c) and RE (d). Diagonal lines indicate equal values under both methods.

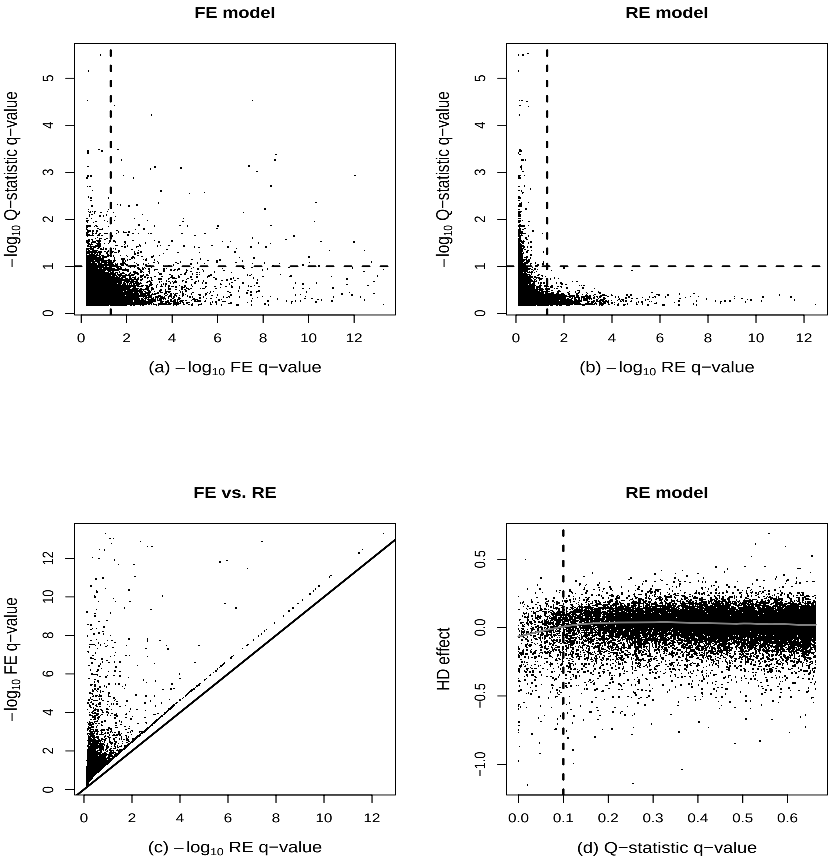


Figure 1.7: Characteristics of FE and RE inference for varying heterogeneity. $-\log_{10}$ q-value of homogeneity Q-statistic vs. $-\log_{10}$ q-value of combined HD effect estimate from FE (a) and RE (b), vertical line indicates HD effect FDR of .05, horizontal line indicates Q-statistic FDR of .10; (c) $-\log_{10}$ q-values for FE vs. RE, diagonal line indicates equal values; (d) RE model estimated mean HD effect size vs. q-value of Q-statistic. vertical line indicates Q-statistic FDR of .1, horizontal smooth line is a loess curve.

Fisher p-value meta-analysis

Figure 1.8 displays results obtained by combining for each gene the HD mod t p-values from Study I and Study II. The distribution of q-values obtained from p-values derived from the χ_4^2 distribution is compressed downward toward significance (a). Resampling p-values are exceedingly conservative compared to χ_4^2-derived p-values (b). Compared to RE model p-values the χ_4^2 p-values are liberal (c), while the resampling p-values are again conservative, although somewhat less than in comparison to the χ_4^2 p-values (d).

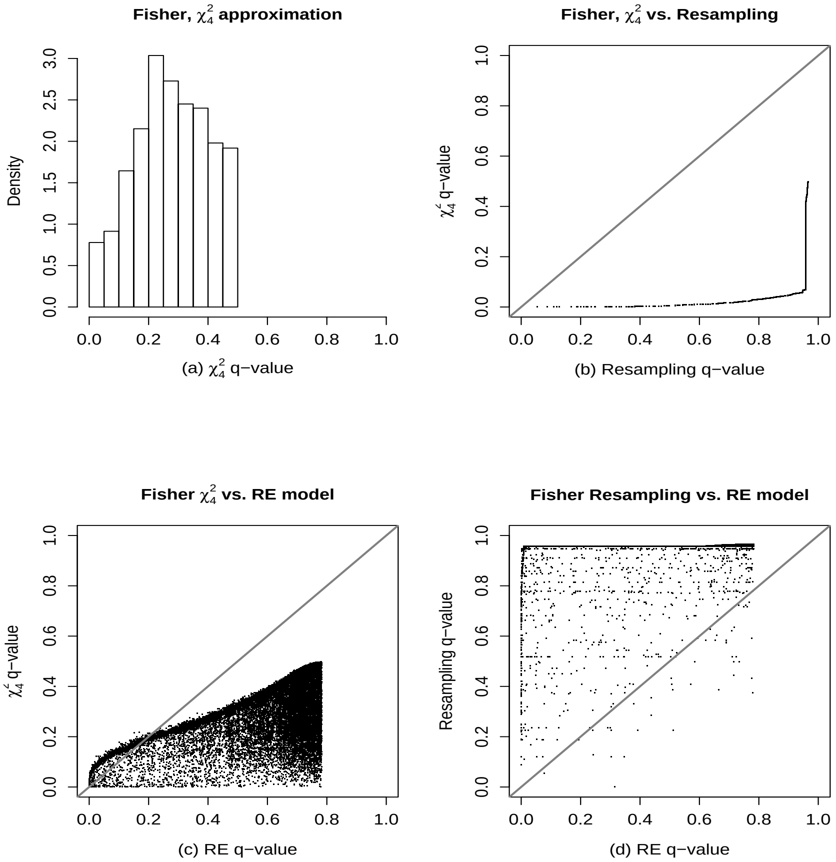


Figure 1.8: Comparison of Fisher S statistic q-values. (a) Histogram of χ_4^2 q-values; scatter plot of q-values obtained by χ_4^2 vs. resampling (b), χ_4^2 vs. RE (c), resampling vs. RE (d). Diagonal lines indicate equal values under both methods.

The lack of agreement indicates the degree to which inference depends on the specific method chosen. Even where the same statistic is used, there is a real problem determining its p-value – the χ^2_4 assumption results in very different p-values from the resampling-based ones. Caution must therefore be exercised in choosing a method and interpreting results.

Comparison of results stratified by heterogeneity status

The findings presented thus far consider the entire set of genes in aggregate. However, the set of all genes can be viewed as a mixture of two types: genes for which the HD effects are homogeneous and genes for which the effects are heterogeneous across studies. It is therefore worth looking at characteristics of the analyses when genes are stratified by heterogeneity status.

Defining heterogeneity status requires a criterion for significance. In the microarray context, one must consider its impact on the subsequent identification of differential expression. To be more conservative in calling a gene differentially expressed, a fairly liberal heterogeneity criterion would seem in order. Taking into consideration the outcome of the heterogeneity analysis above, we decided on a FDR cut-off of 0.10 for Q. For this threshold, the number of genes for which studies are heterogeneous is 802; there are thus 21888 homogeneous ones.

Table 1.2 gives the proportions of genes with significant HD effects for the four meta-analysis methods, as well as the combined data, for all genes together and also stratified by heterogeneity status (Hom. or Het.) at varying FDR for the HD effect. The methods are: C = combined data, FE = fixed effects model, RE = random effects model, FX = Fisher p-value combination method, χ^2 p-values, and FR = Fisher p-value combination method, resampling method. The Fisher resampling method proportions are extremely low, so the numbers of genes are also reported. For RE, proportions given as zero are actual zeros.

Table 1.2: Significance Proportions for Meta-analysis Methods

	Sig. at FDR = .10			Sig. at FDR = .05			Sig. at FDR = .01		
Method	All	Hom.	Het.	All	Hom.	Het.	All	Hom.	Het.
C	0.07	0.06	0.19	0.03	0.03	0.12	0.01	0.01	0.05
FE	0.18	0.17	0.38	0.12	0.11	0.30	0.06	0.05	0.21
RE	0.06	0.06	0.01	0.04	0.04	0.00	0.02	0.02	0.00
FX	0.08	0.06	0.70	0.04	0.03	0.29	0.01	0.01	0.10
FR	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FR Number	(5)	(3)	(2)	(3)	(2)	(1)	(3)	(2)	(1)

FE finds the most significant effects, followed by Fisher χ^2 (FX). These two methods as well as the combined data method also find drastically higher rates of significant effects for studies which are heterogeneous, pointing to the need for caution when combining information. In contrast, RE finds lower rates of significant effects under heterogeneity.

Pairwise agreement of meta-analysis results

Lastly, we look at agreement for pairs of methods stratified by heterogeneity status, varying the FDR for calling a gene differentially expressed between HD and WT (figure 1.9). The simple agreement rate is just the proportion of genes for which both methods agree on whether or not the HD effect is significant. The correspondence between plotting symbol and comparison pair is given in table 1.3.

Table 1.3: Correspondence of Plotting Symbol and Pair

Symbol	0	1	2	3	4	5	6	7	8	9
Pair	FE FR	FE RE	FE C	FE FX	FX RE	FX FR	FX C	C FR	C RE	RE FR

The pairs appear to fall roughly into three groups when homogeneity and heterogeneity rates are considered jointly. Pairs 0 – 3 form one group. This group consists of all pairwise comparisons with FE. These pairs have the lowest agreement under homogeneity. Pairs 4, 5 and 6 have high homogeneity agreement and lower, but increasing with decreasing FDR, heterogeneity agreement. FX appears in each of these pairs. Finally, pairs 7, 8 and 9 have highest agreement under both homogeneity and heterogeneity. These are all pairs are formed from C, RE, FR.

1.5 Discussion

Pooling raw data from different studies for analysis is not always possible; even when it is possible it might not be recommended (e.g. to avoid Simpson's paradox Simpson (1951)). However, in the simple setup we have described here, where a single lab has carried out the same experiment twice, one would think that combining the raw data should be a fundamentally sound approach. In particular, carrying out the normalization step on the aggregated data would seem not only desirable but also necessary.

We have illustrated, however, that even in such an uncomplicated scenario, without issues of different platforms or experimental designs and protocols, integrating the available information might not be completely straightforward. We have seen persistent batch effects that must be taken into account. We would recommend that new methods developed for more complex situations also be tested in simpler cases so that the properties of the methods may be better understood.

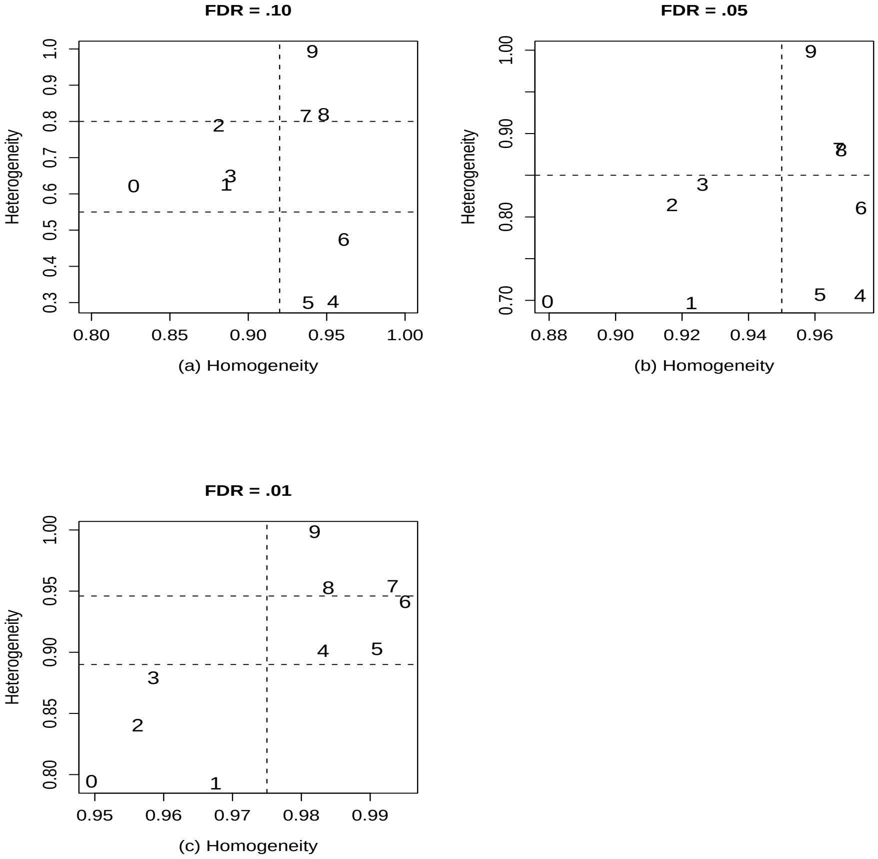


Figure 1.9: Simple agreement rates of pairs of methods for varying FDR. Agreement rate under heterogeneity vs. agreement rate under homogeneity for (a) FDR = 0.10, (b) FDR =0.05 and (c) FDR = 0.01. Dashed vertical and horizontal lines separate groups of pairs.

The importance of considering variability across different labs has been noted in the literature Irizarry et al. (2005); our work here suggests that within lab variability may also need to be considered.

Our results also have substantial implications for large single studies, where patients are recruited over time and arrays are not all hybridized at the same time. Avoidance of problems before they arise calls for careful study design in advance. In addition, comprehensive exploratory data analyses are required once data are collected, to identify and adjust for sources of variability which could obscure the underlying biology. The presence of strong batch effects may be indicative of aspects of laboratory practice in need of improvement. Analysis of data for batch effects can reveal such problems and could therefore help in their rectification.

In this work we can compare results from different methods of analysis, but we are unable to rigorously assess method performance or robustness because it is not feasible at present to determine the truth of the findings. To date we can identify as true positives only a subset of genes that are likely to be differentially expressed in R6/2 mice Luthi-Carter et al. (2000, 2002a), and we also do not yet know which identified genes are not (false positives), or which genes missed are in fact differentially expressed (false negatives). It is advisable to build some truth into the experiment where feasible, for example by using spike-in controls (specific RNAs added to the sample in known quantities). Nevertheless, we hope that this survey of single methods and method agreement can provide some guidance to investigators in selecting appropriate procedures.

In a similar investigation, Stevens and Doerge (2005) use a model to generate a known truth and simulate data from their model to examine properties of meta-analysis of microarray studies. Although this approach may provide some useful broad guidelines, further empirical evidence is required to gain more refined insight into the sources and magnitudes of variability and their effects on properties of meta-analysis of microarray studies.

In most studies, researchers have the resources for further investigation into only a few of the findings. A typical validation study consists of following up on a few genes, often in the range of 5 – 20. The research community would benefit from larger scale follow-up studies to enable the properties of different methodologies for synthesis to be judged more critically.

Although a number of intriguing methods have been introduced for meta-analysis of microarray data, the literature in this field is not yet fully developed. Clearly there is a need for further empirical and theoretical research in this challenging area. Large scale validation studies would provide a welcome opportunity to advance both biological and methodological knowledge.

1.6 Acknowledgements

This work was supported by funds from the Swiss National Science Foundation to the National Centers for Competence in Research (NCCR) in Plant Survival (DRG) and Molecular Oncology (MD, TS), and also by the US National Institutes of Health (RL-C). The authors also acknowledge R. J. Ferrante and J. K. Kubilus for the samples used to generate the HD data set for this analysis.

CHAPTER 2

Alternative Probeset Definitions for Combining Microarray Data Across Studies Using Different Versions of Affymetrix Oligonucleotide Arrays

Jeffrey S. Morris, Chunlei Wu, Kevin R. Coombes, Keith A. Baggerly,
Jing Wang, & Li Zhang
University of Texas MD Anderson Cancer Center
Houston, TX, USA

2.1 Introduction

Many published microarray studies have small to moderate sample sizes, and thus have low statistical power to detect significant relationships between gene expression levels and outcomes of interest. By pooling data across multiple studies, however, we can gain power, enabling us to detect new relationships. This type of pooling is complicated by the fact that gene expression measurements from different microarray platforms are not directly comparable.

In this chapter, we discuss two methods for combining information across different versions of Affymetrix oligonucleotide arrays. Each involves a new approach for combining probes on the array into probesets. The first approach involves identifying ”matching probes” present on both chips, and then assembling them into new probesets based on Unigene clusters. We demonstrate that this method yields comparable expression level quantifications across chips without sacrificing much precision or significantly altering the relative ordering of the samples. We applied this method to combine information across two lung cancer studies performed using the HuGeneFL and U95Av2 chips, revealing some genes related to patient survival. It appears that the gain in statistical power from the pooling was key to identifying many of these genes, since most were not found by equivalent analyses performed separately on the two data sets. We have found that this approach is not feasible for combining information across the U95Av2 and U133A chips, which share fewer probes in common. Our second method defines probesets as sets of probes matching the same full-length mRNA transcripts in current genomic databases. We found this method yielded comparable expression levels across U95Av2 and U133A chip types, and had better correlation across chip types than Affymetrix's matching probeset definitions.

2.2 Combining Microarray Data across Studies and Platforms

In recent years, microarrays have been used extensively in biomedical research. This is evident from the fact that there are over 9000 articles published since 2000 that involve microarrays, with over 3000 published in 2004 alone (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed). Generally, these studies involve the identification of individual genes or sets of genes whose expression profiles are related to clinical or biological factors of interest, including tissue type, disease status, disease subtype, patient prognosis, and biological pathway, to list a few. While microarrays measure the expression levels for thousands of genes, because of cost limitations, most studies are performed using only a small number of samples. As a result, individual studies often have limited power for detecting relevant biological relationships.

More recently, there has been a movement within the scientific community to make data from microarray studies publically available. This movement has been propelled by the establishment of standards for minimal information to provide when posting data (MIAME, (Brazma et al., 2001)) and the requirement of many major journals to make such data publically available. There are currently a number of public repositories in which microarray data are posted, including ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and Gene Expression Omnibus (GEO; http://www.ncbi.nlm.nih.gov/geo/). This explosion of publically-available data makes it possible to consider meta-analyses that combine information across multiple studies, which allow one to assess the reliability of results reported in the individual studies and also to uncover new biological insights not discovered in any individual study. If done properly, this pooling of information across studies can provide increased power to detect small consistent relationships that may have gone undetected in the individual analyses, and can provide results that are more likely to prove reproducible.

There is a small but growing number of studies in existing literature that attempt to combine information across multiple data sets. Generally, there are three approaches that are used: 1. Identify an intersection of genes that are significant across multiple studies, 2. Validate results from a single individual study using data from other studies, or 3. Perform a single analysis after combining data across multiple studies. We now briefly discuss the merits and drawbacks of each approach.

The idea behind the first approach is that if a gene is truly differentially expressed, then this differential expression should be manifest across multiple data sets. However, this Venn diagram-based approach often reveals a shockingly small number of genes that are found to be differentially expressed in multiple data sets. In a study comparing normal and CLL B-cells, Wang et al. (2004) found that only 9 genes were found to be differentially expressed in all three studies conducted on three different microarray platforms, out of 1172 that were differentially expressed in at least one study. Similarly, in a study involving pancreatic cells, Tan et al. (2003) found only 4 genes differentially expressed across 3 different platforms, among the 185 deemed differentially expressed on at least one platform. While perhaps identifying the most reliably differentially expressed genes, this approach actually results in reduced sensitivity for detecting biological relationships, since each (perhaps underpowered) study must find the gene significant before it is declared so. Other less conservative approaches focused on identifying genes that are consistent across studies include methods discussed in Rhodes et al. (2002) and Rhodes et al. (2004a), which involve combining p-values across studies, and the integrative correlation method of Parmigiani et al. (2004), which involves computing gene-gene pairwise correlations on the expression levels and/or tests statistics for each individual study, then computing a “correlation of correlations” across studies. This approach results in a list of reproducible genes whose absolute or relative expression levels are correlated across studies and platforms. It does not, however, provide additional power for detecting biological relationships.

A number of studies take the second approach, identifying biological relationships using the data from a single study, then using data from other studies for validation of these relationships (Beer et al., 2002; Sorlie et al., 2003; Stec et al., 2005; Wright et al., 2003). Since the studies may differ with respect to their patient populations, microarray platforms, and sample handling and processing, results surviving this stringent form of validation are likely to be real. However, like the first approach, this use of multiple data sets does not yield any additional power for detecting biological relationships since only a single data set is used in the discovery process.

In the third approach, the data is actually combined across studies and a single analysis is performed on the pooled data set. This is our primary interest in this chapter. The clear advantage of this approach is the possibility of increased power for detecting biological relationships, since the pooled data set is significantly larger than any of the individual data sets. The difficulty is that there are important differences between the studies that must be taken into account before it is possible to successfully pool the data. The studies may differ with respect to their patient populations, sample handling, or sample preparations. These differences can be manifest in both the clinical outcomes and the microarray data, and may affect the genes in a differential manner. It has been shown that it is possible to obtain comparable microarray data from different laboratories on a common platform if rigorous experimental protocols are established and followed across the different sites (Dobbin et al., 2005). However, posted data from different studies were likely generated using different protocols, so these factors come into play in the meta-analysis context. These problems are further exacerbated if the studies are conducted on different microarray platforms, which have technical differences that make their gene expression levels fundamentally incomparable (Kuo et al., 2002; Tan et al., 2003; Mah et al., 2004; Marshall, 2004; Mecham et al., 2004a).

Some of this heterogeneity can be handled by modeling study effects for each gene using fixed or random effects in the context of mixed models or Bayesian hierarchical models, standard approaches used in meta-analysis (Normand, 1999; Ghosh, 2004; Wang et al., 2004). These approaches appropriately account for the study-to-study variability when performing inference in the meta-analysis, and provide a simple first-order correction for each gene that aligns the mean expression levels for the different studies. Other approaches involve first-order corrections, but use methods that are more sophisticated mathematically. One is based on the singular value decomposition (Alter et al., 2000; Nielsen et al., 2002), and normalizes the raw expression levels within studies using the first eigenvectors for the genes and arrays. This approach assumes that these eigenvectors represent the study-to-study variability, which is assumed to dominate all other factors. Another approach (Benito et al., 2004) normalizes using a new method called "distance weighted discrimination" (DWD), which performs supervised discrimination to identify linear combinations of genes associated with the study effect, which is subsequently removed. However, these approaches, when applied to the raw expression levels, do not appear to be sufficient to make data comparable across different platforms. For one, they only adjust the mean of the distributions for the two studies, but do not adjust for higher order distributional properties like the variances or quantiles. In a study comparing data from spotted cDNA glass arrays and Affymetrix oligonucleotide arrays, Kuo et al. (2002) concluded that "data from spotted cDNA microarrays could not be directly combined with data from synthesized oligonucleotide arrays," and further, that it is unlikely that the data could be normalized using a common standardizing index.

For this reason, many studies do not attempt to combine the raw expression profiles across platforms, but instead only combine unitless summary measures derived from the raw data. The assumption is that, while the raw expression levels for the different studies may not be comparable, these unitless statistics should be, since they are at least on a common scale. For example, Wang et al. (2004) and Choi et al. (2003) first compute the standardardized log fold changes between two experimental conditions, then combine these across studies using hierarchical models. Similarly, Ghosh et al. (2003) and Tan et al. (2003) first compute t-statistics comparing two experimental conditions, then combine these t-statistics across studies. Shen et al. (2004) combine the posterior probabilities of being over-expressed, under-expressed, or similarly expressed between two experimental conditions across data sets. These approaches are promising and all result in increased power to detect biological relationships in the data, and can in principle be used across different platforms. However, we believe it would be inherently better to work with the raw expression levels, if we could get them to be comparable. In that case, we would not be limited to dichotomous comparisons, but could relate gene expression levels with any type of outcome (*e.g.* survival or time to progression). Also, these summary measures make implicit assumptions about the comparability of the reference populations in the different studies that, if not true, may adversely affect inference. For example, using t-statistics assumes that the mean and standard deviation of the true gene expression levels should be the same across studies, and are only different because of technical reasons. By using the raw expression levels, one could avoid making such assumptions.

Some studies have explicitly used sequence information to try to obtain comparable expression levels across platforms (Morris et al., 2005; Mecham et al., 2004a; Mah et al., 2004; Wu et al., 2005; Ji et al., 2005). This idea is natural, since much of the systematic variability between expression level measurements between (and even within) platforms is attributable to sequence-related factors, such as cross-hybridization, alternative splicing, inaccurate annotation of gene sequences, and RNA degradation. Cross-hybridization occurs when a gene hybridizes to "near matches" on the array, which can attenuate estimates of gene expression. Certain sequences are more likely to cross-hybridize (Zhang et al., 2003), so may result in less reliable measurements of gene expression. Also, single genes may be transcribed into multiple different mRNA variants. These alternatively spliced variants may cause some sequences corresponding to different exons from the same gene to be discordant. Additionally, not all probes on microarrays map to annotated sequences in public databases. These probes tend to be less reliable (Mecham et al., 2004b), which may explain some of the lack of concordance across platforms. In a study involving matched samples run on Affymetrix and nylon cDNA arrays, Ji et al. (2005) showed that the correlation of expression levels these platforms was greater for sequences with matches in the RefSeq database. Finally, RNA degradation can affect probes differentially, since sequences closer to the endpoints of the gene may be more susceptible to this degradation than sequences near the middle. These factors are relevant when comparing completely different technologies, *e.g.* spotted glass cDNA arrays and Affymetrix oligonucleotide arrays, as well as when comparing different versions of the same technologies, *e.g.* different versions of Affymetrix arrays or glass cDNA arrays constructed using different clones. We believe that methods that explicitly take into account these known biological and technological factors ultimately will result in the most successful methods for combining information across platforms.

2.3 Overview of Affymetrix Oligonucleotide Arrays

Generally speaking, there are two major types of microarrays, cDNA arrays and oligonucleotide arrays. One key difference between these technologies is that on cDNA arrays, genes are represented by a single cDNA clone spotted on the array, while on oligonucleotide arrays (Lockhart et al., 1996), genes are represented by "probes," or short sequences of nucleotides from the target gene sequence. Affymetrix, Inc. (Santa Clara, CA) is the largest producer of oligonucleotide arrays, which they call GeneChips. Affymetrix GeneChips contain multiple probes for each gene. For the remainder of this chapter, we focus our attention on Affymetrix oligonucelotide arrays, which in practice are the most commonly used arrays today.

The Affymetrix probes each consist of a sequence of 25 bases from the target gene, which generally contains a total of several hundred or thousand base pairs. Since not all sequences bind equally well, there is natural variability between the expression level measurements for different probes taken from the same gene. In order to average over some of this variability, each gene is represented by a number of probes, which together form a "probeset." These probes are scattered across the array. For each probe, there is also a corresponding “mismatch” probe, which contains the identical sequence except with the 13^{th} base replaced by its Watson-Crick complement. The mismatch probes are intended for normalization, although they have not been shown to be clearly useful for that purpose (**?**).

The probes are constructed based on sequence information contained in GenBank (http://www.psc.edu/general/software/packages/genbank/genbank.html), a public archive of DNA sequence information, Unigene (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unigene), which partititions these sequences into non-redundant clusters presumably corresponding to genes, and RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/), which is constructed by the NCBI to represent the state of the art in terms of the sequences of known genes. As this information has evolved over time, Affymetrix has produced different versions of its GeneChip. The most commonly used chip types used in human studies include the HuGeneFL, the U95Av2, and the U133A.

The HuGeneFL was introduced in November 1998, and its sequence clusters are based upon Unigene build 18. It contains information on roughly 5600 genes, and each gene is represented by roughly 20 probe pairs. The probes corresponding to the same probeset are placed together in the same region of the array. The U95Av2 was introduced in April 2000, and is based upon Unigene build 95. It contains information on roughly 10,000 genes, each of which is represented by 16 probe pairs. The probes are randomly distributed across the array. The U133A was first introduced in January 2002, and is based upon Unigene build 133. It contains information on 14,500 genes, and contains 11 probes per gene. The probes are arranged on the array in such a way as to optimize the probe synthesis efficiency.

Frequently, researchers wish to combine information across experiments conducted using different versions of Affymetrix GeneChips. As new studies are conducted using more recent versions of the chips, researchers want to still use information from previous studies performed using older generations. Also, some researchers may want to perform meta-analyses on data collected from multiple studies performed at different institutions. It is not easy to merge information across chip types, since there are some genes represented on newer chips that were not on previous ones, and even the common genes are represented by different sets of probes on the different chips, so their expression levels are not generally comparable.

In the remainder of this chapter, we describe in detail two methods we have developed (Morris et al., 2005; Wu et al., 2005) to combine information across studies using different Affymetrix chip types. These methods use sequence information to define new probesets that yield comparable expression levels across different chip types. Our hope is that the raw expression level values using these redefined probesets are sufficiently comparable that they can be combined across versions. For each method, we describe the method and use an example data set to demonstrate the concordance of expression levels across different array types.

2.4 Partial Probesets

The incompatibility of expression levels across chip types is largely due to the fact that different sets of probes are used to represent the same genes on different chips. We expect, however, that individual probes present on multiple chips should yield comparable expression levels across chips. Thus, one approach for obtaining comparable expression levels across studies using two different chip types is to only use "matching probes" that are present on both chip types.

For example, suppose we have microarray data from two studies, one performed on the HuGeneFL chip and the other on the U95Av2. The HuGeneFL contains a total of roughly 130,000 probes partitioned into 6,633 probesets, each containing 20 probe pairs, while the U95Av2 contains a total of roughly 200,000 probes partitioned into 12,625 probesets, each containing 16 probe pairs. There are a total of 34,428 "matching probes" that are present on both chip types.

After identifying these matching probes, we then recombined these into new probesets based on the most current build of Unigene. We refer to these new probesets as "partial probesets." Note that because they are explicitly based on Unigene clusters, these probesets will not precisely correspond to Affymetrix-determined probesets. Frequently, multiple Affymetrix probesets map to the same Unigene cluster. We then eliminated any probesets containing just one or two probes, since we expected the gene expression measurements based on so few probes to be less reliable. When performed based on Unigene build 160, this left us with 4,101 partial probesets. In general, we expect these probesets to be smaller than the Affymetrix-defined probesets, since they only use the matching probes. Figure 2.1 contains a plot of the number of probes within each of these partial probesets. Most of the probesets (84%) contained 10 or fewer probes, and the median probeset size was seven. There were several probesets containing more than 20 probes.

2.5 Example: CAMDA 2003 Lung Cancer Data

Two independent studies were performed at Harvard University (Bhattacharjee et al., 2001) and University of Michigan (Beer et al., 2002), both focusing on the same question of relating gene expression data to survival in lung cancer patients. These data were part of the 2003 critial assesssment of microarray data analysis (CAMDA) competition (http:/www.camda.duke.edu/camda2003). These studies both used Affymetrix GeneChips, but the Michigan study used the HuGeneFL while the Harvard study used the U95Av2. Our goal in analyzing these data was to combine information across both data sets to identify prognostic genes, whose expression levels provided prognostic information on patient survival over and above what is already provided by known clinical factors. We used partial probesets to quantify the gene expression levels, and demonstrated that this resulted in comparable expression levels across the two chip types, without any loss of precision from using only a subset of the probes. We identified a number of prognostic genes in our pooled analysis that were not discovered in the analyses performed on the individual studies, highlighting the benefit

Figure 2.1: Histogram of number of probes in each "partial probeset."

of pooling data across studies. We first summarize these data sets, then describe our analyses to validate the partial probeset method and obtain prognostic genes. More details of this analysis can be found in Morris et al. (2005).

2.5.1 Overview of Data Sets

The Harvard study analyzed 186 lung tumor samples using U95Av2 Affymetrix GeneChips. From these, 125 were adenocarcinomas for which clinical information on the corresponding patients was available, including gender, age, stage of disease, and survival time. Applying hierarchical clustering to these data, Bhattacharjee et al. (2001) identified four distinct subtypes of adenocarcinoma with different molecular profiles, and further demonstrated that these subtypes had different survival prognoses.

The Michigan study analyzed 86 lung adenocarcinoma samples using HuGeneFL Affymetrix GeneChips. All of these samples also had corresponding clinical information, including gender, age, stage of disease, and survival time. Using univariate Cox regressions, they identified a number of genes whose expression levels were associated with patient survival. They subsequently constructed a "risk index" using the top 50 genes, and demonstrated that this risk index helped predict patient survival both in their own data and in independently obtained data from another experiment (Bhattacharjee et al., 2001).

In our own analysis, we first performed various quality control checks, after which we removed 10 arrays from the Michigan study and one from the Harvard study that demonstrated poor quality. This left us with a total of 200 arrays, 124 from the Harvard study and 76 from the Michigan study. Using the partial probeset definitions described above, we quantified the gene expression levels for each partial probeset using the Positional Dependent Nearest Neighbor (PDNN) model (Zhang et al., 2003). Other quantification methods could have been used, but we chose this one because we believe its use of probe sequence information to predict patterns of specific and nonspecific hybridization intensities can lead to more reliable and accurate quantifications.

We also performed other preprocessing steps. We removed the half of the probesets with the lowest mean expression levels across all samples, then normalized the log expression values by using a linear transformation to force each chip to have a common mean and standard deviation across genes. We next removed the probesets with the smallest variability across chips (standard deviation < 0.20), since we considered them unlikely to be discriminatory and more likely to be spuriously flagged as prognostic. Finally, we removed the probesets with poor relative agreement (Spearman correlation< 0.90) between the partial probeset and full probeset quantifications (see next section). After this preprocessing, 1036 probesets remained and were considered in our subsequent analyses.

2.5.2 Validation of Partial Probesets

Before analyzing the microarray data to identify prognostic genes, we assessed whether our method for combining information across different Affymetrix chip types performed acceptably. First, we checked whether the expression levels appeared to be comparable across chip types. Specifically, we computed the median and median absolute deviation (MAD) log expression level for each partial probeset across the Michigan samples run on the HuGeneFL chip and also for the Harvard samples run on the U95Av2 chip. Since the patient populations in the two studies appeared to reasonably similar, we expected to see high concordance in these quantities between the two chips if the expression levels were comparable. We did not, however, expect perfect concordance, since different patients were used in the two studies. Figure 2.2 contains a plot of these quantities, and demonstrates good concordance between the center and spread in the distribution of gene expression values on the two chips. The concordance between these values was 0.961 for the median and 0.820 for the MAD, so it appears that using the partial probeset method yielded reasonably comparable expression levels across the two chips.

Recall that partial probesets use only the matching probes, while completely ignoring expression level information for the non-matching probes. This means that partial probesets are generally smaller than the Affymetrix-defined probesets. The median size of our partial probesets was seven, while the Affymetrix-defined probesets for the HuGeneFL and U95Av2 chips have 20 and 16 probes, respectively. Since additional probes can increase the precision in measuring the expression level of the corresponding gene, one might expect a loss of precision when using the partial probesets to quantify expression levels. To investigate this possibility, we quantified the expression levels for the full probesets of the Harvard samples using the PDNN

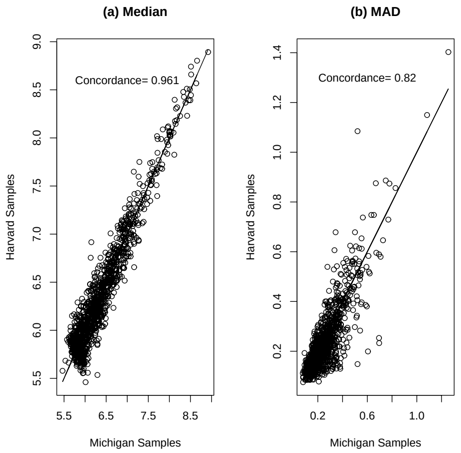


Figure 2.2: Median (a) and median absolute deviation (b) expression levels for each partial probeset based on the Harvard samples run on the U95Av2 chips vs. the Michigan samples run on the HuGeneFL chip. The high concordance in these measures suggests we obtain reasonably comparable expression levels by using the matched probes.

model. The full probesets consisted of all probes on the array mapping to the Unigene cluster, i.e., not just the matching ones. We plotted the standard deviation for each gene using the full probeset versus the standard deviation for the partial probeset, given in Figure 2.3. If the partial probeset quantifications were considerably less precise, we would expect measurement error to cause the standard deviation to be larger for the partial probesets. There was no evidence of significant precision loss in this plot, as there is strong agreement between the standard deviations for each gene using the two methods (concordance=0.942). This may seem surprising at first, but upon further thought is reasonable, since we expect that the probes Affymetrix retained in formulating the new chips may in some sense be the “best” ones.

We computed Spearman correlations between the partial and full probeset quantifications for each probeset to confirm that our method preserved the relative ordering of the samples, i.e., the ranks. For example, we expected that a sample with the largest expression level for a given gene using the full set of probes will also demonstrate the largest expression level for that gene when using only the matched probes. The median Spearman correlation across all probesets was 0.95, suggesting that our method did a good job of preserving the relative ordering of the samples. Interestingly, but not surprisingly, most of the lower Spearman correlations occur for probesets with less heterogeneous expression levels across samples and/or probesets containing smaller numbers of probes. It appears that our partial probeset method worked quite well.

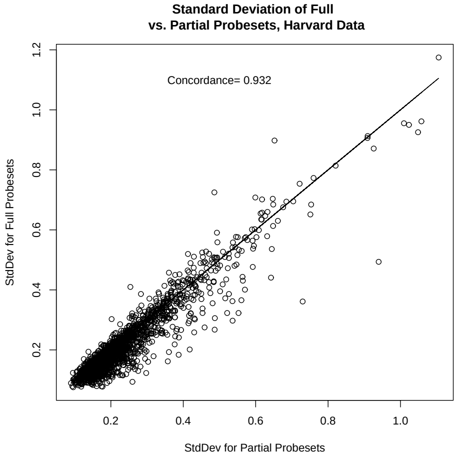


Figure 2.3: Standard deviation across Harvard samples for each gene based on full and partial probesets. A “full probeset” contains all probes on the U95Av2 chip mapping to a unique Unigene ID, while the corresponding “partial probeset.”

2.5.3 Pooling Across Studies to Identify Prognostic Genes

We pooled the data across these two studies to identify prognostic genes offering predictive information on patient survival. We were not primarily interested in finding genes that were simply surrogates for known clinical prognostic factors like stage, since these factors are easily available without collecting microarray data. Rather, we were interested in finding genes that explained the variability in patient survival that remained after modeling the clinical predictors. Thus, we fit multivariable survival models, including clinical covariates in all survival models we used to identify prognostic genes.

We screened the 1036 genes to find potentially prognostic ones by fitting a series of multivariable Cox models containing age, stage (dichotomized as low, stages I-II, and high, stages III-IV), institution, and the log-expression of one of the genes as predictors. The institution effect was included in the model to account for differences in survival that were evident between the two studies, even after accounting for known clinical covariates. We obtained the exact p-values for each gene’s coefficient using a permutation approach. In this approach, we first generated 100,000 datasets by randomly permuting the gene expression values across samples while keeping the clinical covariates fixed. We subsequently obtained the permutation p-value for each gene by counting the proportion of fitted Cox coefficients that were more extreme than the coefficient for the true dataset. A small p-value for a given gene indicated potential for that gene to provide prognostic information on survival beyond the clinical covariates. We also obtained p-values using asymptotic likelihood ratio tests (LRT) and the bootstrap to assess robustness of our results.

If there were no prognostic genes, statistical theory suggests that a histogram of these p-values should follow a uniform distribution. An overabundance of small p-values would indicate the presence of prognostic genes. We fit a Beta-Uniform mixture model to this histogram of p-values using a method called the Beta-Uniform Mixture method (BUM, Pounds and Morris, 2003), which partitions the histogram into two components, a Beta component containing the prognostic genes and Uniform component containing the non-significant ones. We used this model to identify a p-value cutoff that controlled the false discovery rate (FDR, (Benjamini and Hochberg, 1995a) to be no more than 0.20. This means that of the genes flagged as prognostic, we expect at most 1 in 5 were false positives.

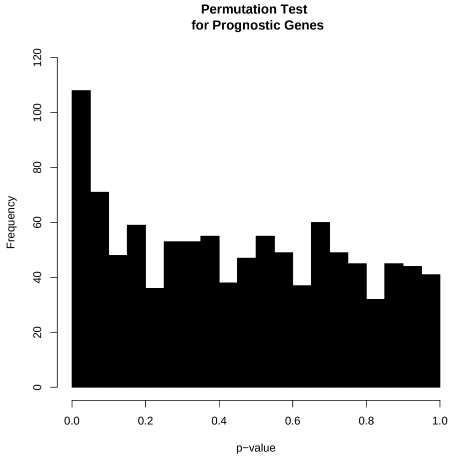


Figure 2.4: Histogram of p-values from permutation test on gene coefficient in Cox model containing clinical covariates and each one of the 1036 candidate genes. The corresponding histogram for the LRT is nearly identical.

Figure 2.4 contains the histogram of permutation test p-values. The overabundance of very small p-values indicates the presence of some genes providing information on patient prognosis beyond what is offered by the modeled clinical factors. Table 2.1 contains a set of 26 genes that are flagged by the BUM method using FDR< 0.20, which are those genes with p-values less than 0.0025. Many of these genes appear to be biologically interesting and worthy of future consideration. We were able to link 10 of our 26 prognostic genes to lung cancer based on the existing literature. Four others could be linked to cancer in general or other lung disease in the literature. These genes are discussed in more detail in Morris et al. (2005).

None of the genes we identified appeared in the list of top 100 genes from the Michigan analysis (Beer et al., 2002), and we only found one (CPE) that was mentioned in the Harvard paper (Bhattacharjee et al., 2001). CPE was one of the genes defining a neuroendocrine cluster that they identified and associated with poor prognosis. We

Table 2.1: Set of genes flagged as prognostic by applying BUM on the permutation p-values with $FDR < 0.20$. Also included are the LRT and bootstrap p-values and estimates of the Cox model coefficient. A '*' indicates the p-value was below the BUM significance threshold. The identity of the genes is also given. A negative coefficient indicates that larger expression levels of that gene correspond to a better survival outcome.

Gene Identity	Coef	Prognostic p-values		
		Permut.	LRT	Bootstrap
FCGRT	-2.07	< 0.00001*	0.00014*	0.0006*
ENO2	1.46	0.00001*	0.00002*	< 0.0001*
NFRKB	-2.81	0.00001*	0.00435	0.00404*
RRM1	1.81	0.00002*	0.00008*	< 0.0001*
TBCE	-2.35	0.00004*	0.00069*	0.0006*
Phosph. mutase 1	1.92	0.00008*	0.00020*	0.0004*
ATIC	1.81	0.00009*	0.00153*	0.0004*
CHKL	-1.43	0.00010*	0.02305	0.0260
DDX3	-2.37	0.00017*	0.00012*	0.0002*
OST	-1.64	0.00020*	0.00010*	0.0010*
CPE	0.72	0.00031*	0.00053*	0.0010*
ADRBK1	-2.20	0.00044*	0.00678	0.0030*
BCL9	-1.64	0.00067*	0.03602	0.0460
BZW1	1.33	0.00068*	0.00279*	0.0006*
TPS1	-0.64	0.00106*	0.00217*	< 0.0001*
CLU	-0.52	0.00109*	0.00239*	0.0024*
OGDH	-2.19	0.00118*	0.00405	0.0020*
STK25	2.29	0.00122*	0.00152*	0.0080
KCC2	-1.70	0.00143*	0.00988	0.0220
SEPW1	-1.29	0.00145*	0.01026	0.0160
FSCN1	0.66	0.00150*	0.00241*	0.0103
MRPL19	1.12	0.00211*	0.03213	0.0340
ALDH9	-1.18	0.00223*	0.00378*	0.0020*
PFN2	0.63	0.00248*	0.00351*	0.0020*
BTG2	-0.75	0.00232*	0.00580	0.0140

repeated our analysis separately for the Harvard and Michigan data sets, *i.e.* without pooling, and only eight and one of the 26 genes, respectively, were flagged as having p-values less than 0.0025, while 17 are not flagged, including the top gene in our list (FCGRT). Thus, it appears that our pooled analysis revealed new biological insights contained in these data that were not identified when analyzing them separately.

2.6 Full-Length Transcript Based Probesets

The analyses presented in the previous section suggest that by using partial probesets, we were able to obtain comparable expression levels across studies conducted at different institutions using different chip types (HuGeneFL and U95Av2), allowing us to perform a pooled analysis that revealed new biological insights into lung cancer. Unfortunately, this approach is not feasible when combining information across the U95Av2 and U133A chips, since these chips share fewer probes in common than the HuGeneFl and U95Av2. There are 34,428 probes (14%) on the U95Av2 that are also present on the HuGeneFl, while there are only 11,582 probes (6%) that are also present on the U133A. If we form partial probesets and eliminate those with less than 3 probes, we are left with only 628 probesets. Thus, we have explored less stringent alternative approaches to use for combining information across these chip types.

One of the primary reasons probes yield discordant measurements is that they may be responding to different transcripts alternatively spliced from the same gene. When the transcripts are differentially regulated, the corresponding probes can yield conflicting signals. The current design of arrays ignores the effects of alternative splicing. Thus, if we differentiate the probes that match sets of alternatively spliced transcripts, we may be able to resolve the discordant measurements. Based on this idea, we developed a new method to regroup the probes into probesets. In our new definition of a probeset, all probes in the probeset must match the same set of full-length gene sequences. We refer to such a probeset as a “Full-Length Transcript Based Probeset” (FLTBP, (Wu et al., 2005). Assuming complete inclusion of alternatively spliced transcripts, we can in principle ensure concordant behavior of the probes within these probesets.

We now describe how we obtained these transcript-based probesets. First, we constructed a comprehensive library of full-length mRNA transcript sequences in the human genome by combining records in RefSeq (http://www.ncbi.nlm.nih.gov/RefSeq/) and HinvDB (http:// hinvdb.ddbj.nig.ac.jp/index.jsp) databases. As of January 2005, RefSeq (build 111504, human section) contained 28,712 full-length transcript sequences representing 23,809 genes. H-InvDB (version 1.7) contained 41,118 sequences representing 21,037 genes. All of the sequences in this database were validated by full-length cDNA clones. We estimate that collectively the two databases represent approximately 29,000 genes with 50,000 non-redundant transcripts.

We used this library as the basis for defining our probesets. For each probe sequence used on the U133A and U95Av2 arrays, we identified all matching full-length transcripts using the Blast program (http://www.ncbi.nlm.nih.gov/blast/). We aggregated the IDs of those transcripts with exact matches to construct a matched target list. We found that 15% of the probes on the U95Av2 and 13% of the probes on the U133A had no exact match in our library, and 38% of the probes on the U133A and 33% of the probes on the U95Av2 matched more than two targets in our library, demonstrating that it was very common for one probe to match multiple targets.

By grouping the probes within the same matched target lists, we formed 23,972 and 14,148 probesets on the U133A and U95Av2, respectively. We call these probesets "Full-Length Transcript Based Probesets" (FLTBPs). Because multiple probes in a probeset are essential to reduce noise and bias, we discarded all small probesets containing less than 3 probes, leaving us with 18,011 and 11,228 FLTBPs on the U133A and U95Av2, respectively. Collectively, these FLTBPs contained 82% of the probes on the arrays.

These new probesets were very different from the original ones. Only 9,893 of the original probesets on U133A and 5,257 original probesets on U95Av2 were the same after regrouping. Figure 2.5 shows a histogram of the number of probes in each FLTBP. The probesets outside of the major peaks reflect division and fusion of the original probesets. Detailed information of our probesets are stored on our web site (http://odin.mdacc.tmc.edu/∼zhangli/FLTBP). This website also contains chip design files (CDF) using FLTBPs following the format designed by Affymetrix (http://www. affymetrix.com/index.affx). These CDF files can be used to run MAS5, RMA and dChip algorithms in Bioconductor (http://www.bioconductor.org/).

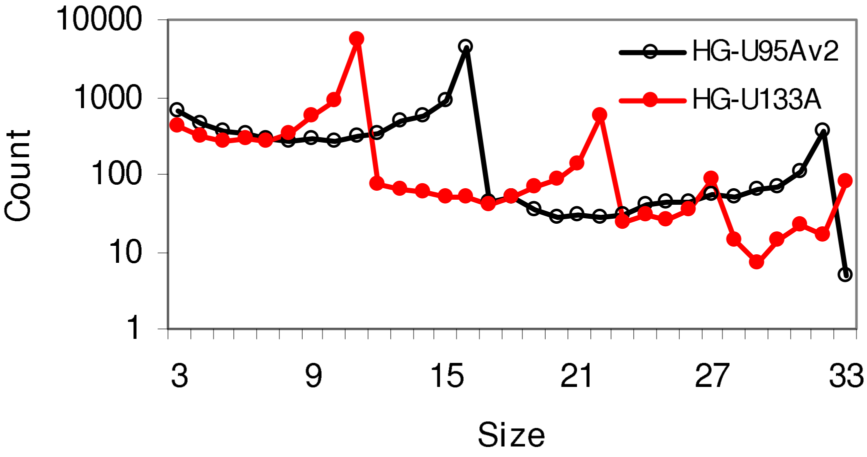


Figure 2.5: Histogram of number of probes per FLTBP.

By matching the matched target lists of FLTBPs on the two arrays, we found 9,642 pairs of FLTBPs that can be mapped between the U133A and U95Av2. Affymetrix has their own method for mapping probesets between different chip types (http://www. affymetrix.com/Auth/support/downloads/comparisons/best_match.zip), which yields

9,480 pairs of probesets between the U95Av2 and U133A chips. There are numerous differences between these Affy-defined mappings and our FLTBPs. Only 52% of the probe sets on the U133A and 48% of the probesets on the U95Av2 are mapped the same way as our FLTBPs.

2.7 Example: Lung Cell Line Data

To compare our mapping method with that of Affymetrix, we used a data set consisting of 28 paired measurements obtained by hybridizing identical samples on both the U133A and U95Av2 arrays. Because of this paired design, we expect very little biological variability between paired measurements on the two arrays, so any differences observed should be attributable to technical sources. We now describe this dataset and use it to demonstrate that the FLTBPs result in quantifications that are more comparable across chip types than Affymetrix- based probesets.

2.7.1 Overview of Data Set

Thirty RNA samples from variant lung cancer or normal lung cell lines and one human reference sample were hybridized on both U133A and U95Av2 arrays. Our quality control procedures revealed that three array images had obvious defects, so were discarded. This left us with 28 pairs of samples that we used in this study.

We preprocessed and quantified the gene expressions with PDNN (Zhang et al. 2003) using the PerfectMatch software (ver2.2) (http://odin.mdacc.tmc.edu/∼zhangli/ PerfectMatch). For comparison, we also preprocessed and quantified the data using other competing methods, RMA (Irizarry et al., 2003a), MAS5 (http://www. affymetrix.com/ products/software/specific/mas.affx) and dChip (Li and Wong, 2001), using BioConductor (v1.5, http://www.bioconductor.org/), following the default settings in the `affy` package (Irizarry et al., 2004).

2.7.2 Validation of Transcript-Based Probesets

In order to assess comparability across chip types, for each gene, we computed the correlations between the paired U95Av2 and U133A measurements across samples. To enhance the contrast between two different mapping methods, in our comparisons we focused on the probesets that differed between the two methods. Approximately 1/3 of the probesets were mapped differently, which resulted in 3,309 and 3,527 paired probesets for FLTBP method and Affymetrix method, respectively.

Figure 2.6 contains a histogram of these correlations across probesets for the two mapping methods and four quantification methods. These histograms summarize the observed distribution of the paired correlations across probesets. Figure 2.6A clearly demonstrates that, when using the PDNN quantification method, the FLTBP mapping tends to yield better correlations than the Affymetrix mapping ($p < 0.00001$,

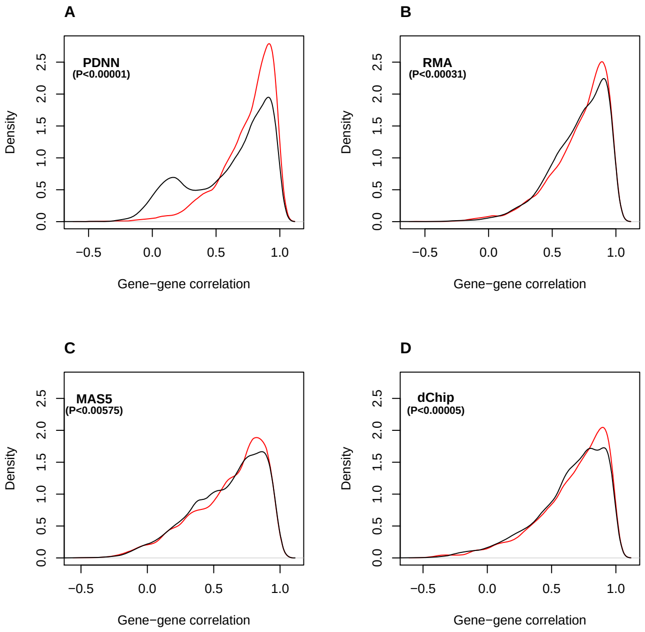


Figure 2.6: Distribution of gene-to-gene correlation between probesets on two U95Av2 and U133A arrays, combining information over all samples, using both Affymetrix-defined probesets and FLTBPs. The correlations were computed using four different quantification methods, (A) PDNN, (B) RMA, (C) MAS5.0, and (D) dChip.

Kolmogorov-Smirnov [KS] test). Notice the two peaks evident in the distribution of correlations for the Affymetrix mapping. The minor peak contains a large group of probesets with poor correlation across chip types. With other quantification methods, there is also evidence that the FLTBP method tends to result in better correlation across chip types than the Affymetrix method, although this evidence is not as strong (Figures 2.6B-D, $p = 0.00031$, 0.00575, and 0.00005 respectively). This improvement from using the FLTBPs is likely due to the fact that the FLTBP adjusts for some of the heterogeneity that is due to alternative splicing.

Note also that, when compared with Figure 2.6A, the distributions in Figure 2.6B-D are shifted more towards low correlations. This suggests that, for these data, the PDNN quantification tended to yield generally higher correlations than the RMA, MAS5, or dChip quantifications. This is even more evident in the sample-by-sample correlations between the chip types computed across genes, as shown in Figure 2.7. This increased correlation observed from the PDNN method may reflect the man-

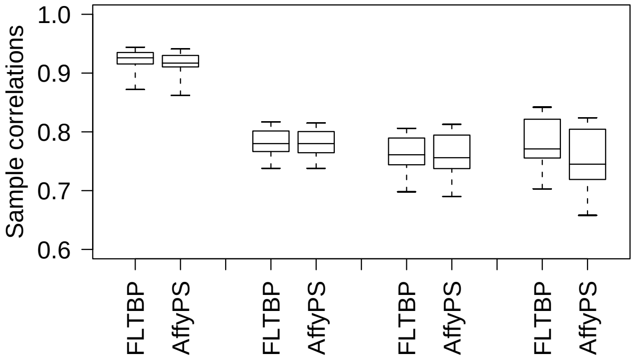


Figure 2.7: Distribution of sample-to-sample correlation between probesets on two U95Av2 and U133A arrays, combining information over all genes, using both Affymetrix-defined probesets and FLTBPs. The correlations were computed using four different quantification methods, PDNN, RMA, MAS5.0, and dChip, respectively.

ner in which the PDNN model estimates and adjusts for the effects of non-specific binding.

¿From Figure 6A, we see that even when using the FLTBPs, not all genes displayed high correlations across chip types. Many of these low correlations were observed for genes that appeared to have low biological variability in these data. Low variability would make the noise component of the measurements dominate, resulting in low correlations. There are, however, some probesets with low correlations that do not have small variances. It is possible that some of the sequences corresponding to these probesets were strongly affected by RNA degradation, or the currently available collection of transcripts may not include certain alternatively spliced variants that were differentially expressed across the sample tests, causing the correlations to become attenuated. Further work needs to be done to further reduce the effects of cross-hybridization and RNA degradation, which will hopefully lead to even more comparable expression levels across platforms.

2.8 Summary

In this chapter, we have illustrated the benefit of pooling data across multiple microarray studies. We performed a pooled analysis over two lung cancer microarray studies, and identified new prognostic genes that were not detected by separate analyses performed on the individual data sets. We also described two new probeset definitions that result in more comparable expression levels across different versions of Affymetrix oligonucleotide chips. The first method is based on partial probesets, which only use probes present on both chip types and combine them together based on Unigene cluster information. This approach works very well, but has limited applicability, since it is only feasible to apply across chip types that share many probes in common. The second method does not restrict us solely to matching probes, but works by recombining probes based on the set of full-length mRNA transcripts to which they map. In this way, the probesets map to the same set of alternatively spliced transcripts. Combined with the PDNN quantification method which accounts for non-specific binding, this approach appears to result in more comparable expression levels across chip types than Affymetrix's matched probesets. The benefit of this approach is that it does not restrict attention to matched probes, so can be widely applied to combine data across any chip types. It may even be possible to use this principle to match up oligonucleotide array data with cDNA data, although this remains to be seen.

CHAPTER 3

Significance testing for small microarray experiments

Charles Kooperberg, Aaron Aragaki, Charles C. Carey, and Suzannah Rutherford
Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109

3.1 Introduction

When there are many degrees of freedom it is sometimes less critical which significance test is carried out, as most analysis will give approximately the same result. However, when there are few degrees of freedom the choice of which significance test is being used can have a strong effect on the results of an analysis. Unfortunately, this is often the case for microarray experiments, as research laboratories often perform such experiments with only a few (say less than five) repeats, Reasons for the small number of repeats include availability of specimens and economics. Kooperberg et al. (2005) compared several approaches to significance testing for experiments with a small number of oligonucleotide (one-color) arrays. In this paper we summarize the results from that analysis, include a couple of additional methods, and describe a similar comparison for methods of carrying out significance testing for two-color (red-green) arrays.

The limited number of repeats, together with the large variability that even the best microarray platforms have, make small sample comparisons unattractive. A standard T-test for an experiment with six two-color arrays has, depending on whether other variables are controlled for, at most five degrees of freedom. The resulting two-sided test, with $\alpha = 0.05$ and a Bonferoni correction for 10000 genes requires a T-statistic of 20.6 or more for significance. The lack of degrees of freedom is really what drives the extremely large significance threshold for T-statistics: the same α and Bonferoni correction for 20 arrays requires a T-statistic of 6.3 or more while a normal distribution only requires a Z-statistic of 4.6 or more, on the other hand reducing the number of genes of interest on the original array from 10000 to 500 only reduces the required T-statistic to 11.3.

Nonparametric (Wilcoxon) or permutation tests are no easy way out. For example, for an experiment with k two-color (spotted) arrays, a P-value for a permutation test can be no smaller than 2^{-k}; if we want a two-sided test with $\alpha = 0.05$ and a Bonferoni correction for 10000 genes, we need k to be at least 19. Reducing the number of genes to 500 reduces the minimum k to 15. Similarly, for a one-color (oligonucleotide) array the P-value for a permutation tests with k cases and k controls a P-value cannot be smaller than $\binom{2k}{k}$; so for a two-sided test with $\alpha = 0.05$ and a Bonferoni correction for 10000 genes, we need at least $2k = 22$ arrays. Reducing the number of genes to 500 reduces the minimum number of arrays to 18.

As permutation tests are not going to help us, we need to obtain a better estimate for the residual variance to overcome the lack of repeats. There are two obvious choices: we can combine different genes in the same experiment or we can combine different experiments, if similar experiments were carried out. When genes are combined we can either choose to combine those genes for which the general expression level is similar as do, for example, Huang and Pan (2002) and Jain et al. (2003) or we can choose to combine all genes. An alternative approach to obtain more power with small experiments is to add a stabilizing constant to the estimate of the variance for each gene or to use some (Bayesian) model for the expression levels. SAM (Tusher et al., 2001) is a methodology that adds a constant to the estimate the variance. The approaches by Baldi and Long (2001), Lönnstedt and Speed (2002), Smyth (2004) and Cui et al. (2005) are four related (empirical) Bayesian approaches. Wright and Simon (2003) discuss a closely related frequentist approach.

In this paper we do not control for multiple comparisons. In practice, when one carries out tests for many thousands of genes simultaneously, a multiple comparisons correction or a correction of the false discovery (FDR) rate is essential. See Dudoit et al. (2003) for an extensive overview of multiple comparisons corrections. While several of these proposals use permutation arguments to correct for multiple comparisons, permutation typically either requires a substantial number of replicates (that are not available in small experiments), or they require implicit assumptions about similarities in the variational properties of different genes. In either scenario, we believe that only well calibrated marginal P-values are going to yield good multiple comparison corrected P-values.

P-values have the advantage that there are well established measures such as Type I error and power that can be used to judge the performance of a test. The FDR (Benjamini and Hochberg, 1995a) does not have such a simple measure, to check whether estimates of the FDR are accurate on a single experiment In addition, just like for multiple comparison procedures, there are procedures to approximate the FDR from P-values.

3.2 Methods

Most of the methods that we compare in this paper can be used either for one-color (oligonucleotide) arrays or for two-color (spotted) arrays. We assume that the arrays have been properly normalized; see Section 3.6 for how we normalized our arrays.

3.2.1 Notation

Two-color spotted arrays For each gene and each two-color array we have an expression ratio x^m_{ijl} summarizing the (log-)expression ratio between experimental conditions $k = 1$ and $k = 2$ (that may be different between experiments) for gene $i = 1, \ldots, n$ in experiment $j = 1, \ldots J$ on replicate array $l = 1, \ldots, L_j$. For each gene on each array we also have an estimate of the overall expression x^a_{ijl}, typically this will be the (geometric) average of the normalized expression for both channels of the array. Unless there is confusion we will write x_{ijl} instead of x^m_{ijl} for the log-expression ratios.

Let μ_{ij} be the "true" (log-)expression ratio of gene i in experiment j for condition 1 relative to condition 2. Set $\widehat{\mu}_{ij} = \sum_l x_{ijl}/L_j$, $s^2_{ij} = \sum_l (x_{ijl} - \widehat{\mu}_{ij})^2$, and $x^a_{ij} = \sum_l x^a_{ijl}/L_j$.

One-color oligonucleotide arrays Similarly, for each gene and each one-color array we have a (log-)expression x_{ijkl}, for experimental conditions $k = 1$ and $k = 2$, for gene $i = 1, \ldots, n$ in experiment $j = 1, \ldots J$ on replicate array $l = 1, \ldots, L_{jk}$.

Let μ_{ijk} be the "true" mean (log-)expression level of gene i in experiment j under condition k. Set $\widehat{\mu}_{ijk} = \sum_l x_{ijkl}/L_{jk}$ and $s^2_{ijk} = \sum_l (x_{ijkl} - \widehat{\mu}_{ijk})^2$.

3.2.2 Significance Tests

All significance tests that we consider in this paper can be written in the form

$$\frac{\widehat{\mu}_{ij}}{\widetilde{\sigma}_{ij}/\sqrt{L_j}},$$

for two-color arrays and

$$\frac{\widehat{\mu}_{ij1} - \widehat{\mu}_{ij2}}{\widetilde{\sigma}_{ij}\sqrt{\frac{1}{L_{j1}} + \frac{1}{L_{j2}}}},$$

for one-color arrays. Here $\widetilde{\sigma}_{ij}$ is an estimate of the variance of x_{ijl}. The methods that we discuss differ primarily in how the estimate $\widetilde{\sigma}_{ij}$ is obtained. The traditional test statistics estimate $\widetilde{\sigma}_{ij}$ uses only the data on gene i and experiment j. The approaches that inflate the variance and those that combine genes also use data on genes i^*, $i^* \neq i$; implicitly to estimate hyper-parameters for the empirical Bayes approach that inflates the variance, or explicitly to smooth the estimates for $\widetilde{\sigma}_{ij}$. Finally the approaches that combine experiments use data on experiments j^*, $j^* \neq j$. Most of the methods below have a proper reference distribution, but alternatively significance levels can be obtained using permutations (see Section 3.2.3); in fact, some of the authors recommend permutations as the method to obtain P-values.

Below we describe the test-statistics we are including in our comparison. We provide details for the two-color arrays, modifications for one-color arrays are indicated.

T-statistic. The traditional T-statistic is

$$t_{ij} = \frac{\widehat{\mu}_{ij}}{\widehat{\sigma}_{ij}/\sqrt{L_j}},$$

where $\widehat{\sigma}^2_{ij} = s^2_{ij}/(L_j - 1)$, provided $L_j > 1$. The reference distribution is the T-distribution with $L_j - 1$ degrees of freedom, and the main assumption is that for each gene i and experiment j the x_{ijkl} are independent having a normal distribution with variance σ_{ij}, although the T-test is generally considered to be robust against departures from normality.

The two-sample T-statistic is the equivalent test for one-color arrays. This statistic assumes that the variance for both experimental conditions is the same. An alternative is the Welch (1938) two-sample T-statistic that does not make that assumption. In Kooperberg et al. (2005) it was shown that this approach has almost no power for small sample sizes, and should probably be avoided for small microarray experiments.

Methods combining genes: smoothing the variance

There have been several proposals in the literature to combine the estimates of the variance for several genes to obtain better estimates, so that the resulting test has more degrees of freedom. Typically the assumption that is made is that genes with the same expression level have approximately the same variance. Under this assumption estimates for the variance can be obtained by smoothing the variance as a function of the expression level. For one-color arrays there are methods which smooth the variances jointly and methods which smooth variances separately for both experimental conditions.

LPE Jain et al. (2003) describe a method they call “Local Pooled Error test” (LPE). As described in this paper, LPE only is applicable to one-color arrays. In their approach, let $\widehat{\sigma}_{ijk}$ be the the sample variance of the x_{ijkl}, for $l = 1, \ldots, L_{jk}$. LPE regularizes these estimates for each j and k separately by smoothing the $\widehat{\sigma}_{ijk}$ versus $\widehat{\mu}_{ijk}$. The assumption being made here is that genes with the same expression level for the same experiment and the same condition have (approximately) the same variance. As the smoothing spline that is used effectively involves averaging a large number of genes, the authors use a normal reference distribution. In our study we have used the implementation by the authors, available in the R-package (Ihaka and Gentleman, 1996) LPE, which is available from CRAN/Bioconductor [*] Since the method averages the variance separately for two conditions, it is currently only available for one-color arrays, where both experimental conditions are measured separately.

Loess Huang and Pan (2002) make several related proposals. The main difference between their approach and the approach by Jain et al. (2003) is that they first compute $\widehat{\sigma}_{ij}$ and smooth these estimates against $\widehat{\mu}_{ij} = \widehat{\mu}_{ij1} + \widehat{\mu}_{ij2}$ for one-color experiments and against x^a_{ij} for two-color experiments. Their simulation results

[*] CRAN: The Comprehensive R Archive Network; see `http://www.r-project.org`.

show that, not unexpectedly, for the null-model a normal reference distribution is appropriate. We reimplemented their approach using a `loess` smoother.

Methods combining genes: (empirical-)Bayesian model for σ

Rather than smoothing the variance explicitly as a function of the expression level, we can include information from other genes for the analysis of a particular gene by making assumptions about the distribution of the variance for all genes. The information about the other genes then allows us to estimate some (hyper-)parameters, that can be used to stabilize the variance estimate. There are a variety of such methods with different motivations: ad-hoc (e.g. SAM (Tusher et al., 2001) using an (empirical) Bayes argument (e.g. (Baldi and Long, 2001; Lönnstedt and Speed, 2002; Smyth, 2004), a James-Stein type estimator (Cui et al., 2005) or a frequentist approach (Wright and Simon, 2003).

The first three approaches that we discuss combine the sample variance $\widehat{\sigma}^2_{ij}$ with another estimate σ_{0ij} that has d_{ij} degrees of freedom, yielding a variance estimate of

$$\widetilde{\sigma}^2_{ij} = \frac{d_{ij}\sigma^2_{0ij} + (L_j - 1)\widehat{\sigma}^2_{ij}}{L_j + d_{ij} - 1}, \tag{3.1}$$

that can be used in a T-test with $L_j + d_{ij} - 1$ degrees of freedom. The three methods **Cyber-T, Limma, RVM** use this approach; they differ primarily in the methods to obtain σ_{0ij} and d_{ij}.

Cyber-T The Cyber-T approach of Baldi and Long (2001) is motivated as a fully Bayesian procedure. However as implemented in practice (see Section 5 of Baldi & Long 2001) the test is carried out using a T-test on (for two-color arrays) $L_j + \nu_0 - 1$ degrees of freedom, and an estimate of the variance (compare 3.1) of

$$\widetilde{\sigma}^2_{ij} = \frac{\nu_0\sigma^2_{0ij} + (L_j - 1)\widehat{\sigma}^2_{ij}}{L_j + \nu_0 - 1}, \tag{3.2}$$

where σ^2_{0ij} is an estimate of the "prior variance" that is obtained as a running average of the variance estimates of the genes in a "window" of size w of similar x^a_{ij}. Thus the Cyber-T approach uses the average of a smoothed variance (like **LPE** and **Loess**, only using another smoother) with the regular variance of the **T-statistic**. A non-Bayesian interpretation of Cyber-T is thus that it combines a smoothed estimated (as in **Loess** and **LPE**) with a traditional estimate from the **T-test**.

We used the defaults $\nu_0 = 10$ and the window width $w = 101$ from the R-software available on `http://visitor.ics.uci.edu/genex/cybert`. Note that the paper of Baldi and Long mentions another default of $\nu_0 = 10 - L_j$.

Limma Smyth (2004) generalizes the approach from Lönnstedt and Speed (2002) The main assumption in Smyth's model is a prior distribution on the variances σ^2_{ij}:

$$\frac{1}{\sigma^2_{ij}} \sim \frac{1}{d_{0j}s^2_{0j}}\chi^2_{d_{0j}}.$$

(We include the index j for the parameters of the prior, as they may be different for different experiments $j = 1, \ldots, J$.) The model also includes priors on the coefficients for each gene in a linear regression model, which in the two sample case reduces to the difference between the mean expression for the two groups. Using methods of moments estimators estimates d_{0j}, s_{0j}^2, and a few other parameters are obtained. An inflated variance

$$\widetilde{\sigma}_{ij}^2 = \frac{d_{0j} s_{0j}^2 + (L_j - 1)\widehat{\sigma}_{ij}^2}{L_j + d_{0j} - 1}, \tag{3.3}$$

(compare 3.2) is used for a "moderated T-test" with $d_{0j} + L_j - 1$ degrees of freedom. Thus, a main difference between the approach of Smyth (2004) and the approach of Baldi & Long (2001) is that Limma uses one single estimate for the prior variance (s_{0j}^2) for all genes and it estimates the prior degrees of freedom d_{0j} based on the data, while the latter uses a smooth estimate for the prior variance σ_{0ij}^2, but it uses a fixed number of prior degrees of freedom ν_0. The approach of Smyth (2004) is available from the Bioconductor package Limma. We used Limma with the default options.

RVM The Random Variance Model (RVM) of Wright and Simon (2003) inflate the variance similar to Baldi & Long (2001) and Smyth (2004), and obtain a model similar to (3.1). They assume an inverse Gamma model for σ^2, and estimate the two parameters from this model using the method of maximum likelihood. Implementation of their approach would require estimating of two parameters of an F-distribution. We do not include RVM this method in our comparisons, as we could not locate publicly available software.

Shrinking Cui and Churchill (2003) and Cui et al. (2005) develop a James-Stein shrinkage estimate $\widetilde{\sigma}_{ij}^2$. After appropriate transformations this estimator "shrinks" the **T-test** estimate $\widehat{\sigma}_{ij}^2$ towards the mean variance $\sum_i \sigma_{ij}^2/I$, where the exact amount of shrinkage differs from gene to gene, and depends on the variability for that gene. Easy to implement formulas are given in Cui et al. (2005). Note that the authors of this method recommend a permutation approach (see Section 3.2.3) to obtaining P-values. We still include this approach without permutations using a normal reference distribution, as well as using permutation P-values.

Methods combining experiments

Instead of combining different genes *within* one experiment, we can also combine expression levels of the same gene *between* experiments. This would potentially be useful if we have several smaller experiments, and it is thus reasonable to assume that for each gene the variance in each experiment is approximately the same.

Pooled-T We define the pooled T-test statistic, combining experiments, as

$$c_{ij} = \frac{\widehat{\mu}_{ij}}{\widehat{\sigma}_i \sqrt{\frac{1}{L_j}}},$$

where $\widehat{\sigma}_i^2 = \sum_j s_{ij}^2/L$ and $L = \sum_j (L_j - 1)$, provided $L > 0$. The reference distribution is the T-distribution with L degrees of freedom, and the main assumption is that the x_{ijl}^m are independent for each j and l, having a normal distribution with mean μ_{ij} and variance σ_i.

For most of the other methods that we discussed it is, in principle also possible to pool different experiments in obtaining a single variance estimates. As all these methods already regularize the estimates for σ in some way, pooling typically has no effect, and the corresponding method behaves similar to the "parent" method, as was confirmed for the **Loess** approach in Kooperberg et al. (2005) and for **Limma** in unpublished results.

Note that methods whose implementation allows for general design matrices (e.g. **Limma**) can yield pooled estimates by setting up an appropriate design matrix and testing appropriate contrasts.

3.2.3 Permutation P-values

Permutation of the arrays in an experiment can be an alternative to using a parametric reference distribution for a test statistic. Assume that we have a two-color experiment with L arrays, and that the test statistic for the ith gene is T_i. To compute the significance of T_i we also compute the test statistics for all genes for each of the $m = 1, \ldots, 2^L$ experiments that are obtained by "flipping" the signs of the x_{il}^m for some of the l. (We omit the index of experiment j.) Note that one of these permutations will be the original design. Let T_i^m be the test statistic for the ith gene for the mth permutation. We can use

$$\sum_{i^*=1}^{n} \sum_{m=1}^{2^L} I(T_i < T_{i*}^m)/n2^L$$

as an estimate of the P-value corresponding to T_i. If L is larger than, say, 8 we may want to sample permutations to save computing time; in this paper that is not an issue.

These estimates will be unbiased if (i) each T_i has the same distribution under the null-hypothesis, and (ii) no genes are differentially expressed. The first assumption is not as severe as it appears. When a parametric distribution is used the stronger assumption, that the distributions of each T_i under the null-hypothesis are the same as a particular parametric distribution, is made. The second assumption is much more severe, and it will lead to conservative P-values when in fact there are a substantial number of differentially expressed genes (Storey and Tibshirani, 2003).

For one-color (oligonucleotide) arrays we randomly rearrange the L_1 arrays with the first experimental condition and the L_2 arrays with the second experimental condition, and proceed in a similar manner.

Table 3.1: Organization of the two-color (spotted) data for our analysis. Experiments whose code start with a D are expected to have differences between both groups, while those starting with an S are repeats, the digit "2" refers to the two-color (spotted) array type. The arrays for experiments D2.3 and D2.4 and those for D2.5 and D2.6 are different; experiment S2.1 are arrays from a cell-line not used for the other experiments.

Exp.	sample one	sample two	L_j	different
S2.1	KC cell	KC cell	4	no
S2.2	SAM	SAM	2	no
S2.3	SAM	SAM	2	no
S2.4	SAM	SAM	4	no
D2.1	SAM	D-recomb 304	2	yes
D2.2	SAM	D-recomb 220	2	yes
D2.3	SAM	D-pure	2	yes
D2.4	SAM	D-pure	4	yes
D2.5	SAM	E-pure	4	yes
D2.6	SAM	E-pure	4	yes
D2.7	SAM	F-pure	6	yes

3.3 Data

For our analysis we use two sets of data. One comes from a one-color experiment, and is part of the data that was also used in Kooperberg et al. (2005) the other comes from a not yet published study on Drosophila.

The two-color experimental data that we use come from a series of spotted microarrays of *Drosophila melanogaster* that were grown in Suzannah Rutherford's lab at the Fred Hutchinson Cancer Research Center. The arrays are part of a larger set of experiments whose results have not yet been reported. The subset of arrays that we compare here include some experiments that are self-to-self hybridizations, and some experiments where both samples are genetically different, see Table 3.1. Thus, the experiments S2.1, S2.2, S2.3, and S2.4 are intended to establish that the tests have the right size Type I error, and the experiments D2.1, D2.2, D2.3, D2.4, D2.5, D2.6, and D2.7 are intended to establish the power of the tests.

For the SAM samples RNA from a large number of flies that were genetical identical, other than some being male and some being female, was combined and the RNA for the arrays was taken out of this large pool. For the D-recomb 304, D-recomb-220, D-pure, E-pure, and F-pure lines for each array samples from 15-30 flies that were genetical identical, other than some being male and some being female, was combined. In addition we included four unrelated Drosophila cell line arrays. We organized the experiments so that all experiments are "dye swapped": i.e. half of the

Table 3.2: Organization of the one-color (Affymetrix) data for our analysis. HD: Huntington's Disease mouse, WT: wildtype mouse. Experiments whose code start with a D are expected to have differences between both groups, while those starting with an S are repeats, the digit "1" refers to the one-color (Affymetrix) array type.

Exp.	Tissue	Mouse	Group 1	Group 2	L_{j1}	L_{j2}	different
S1.1	cerebellum	DRPLA 26Q	HD	HD	2	2	no
S1.2	cerebellum	DRPLA 26Q	WT	WT	2	2	no
S1.3	cerebellum	YAC	HD	HD	3	2	no
S1.4	cerebellum	YAC	WT	WT	3	2	no
D1.1	cerebellum	DRPLA 65Q	HD	WT	4	4	yes
D1.2	cerebellum	R6/2 12 weeks	HD	WT	2	2	yes
D1.3	cerebellum	N171	HD	WT	4	4	yes

arrays have sample one on the red channel, the other half have sample two on the red channel. There are 13,440 spots (genes) on each array.

One-color experimental data was obtained using Affymetrix Mu 11K-A microarrays generated for a series of experiments on Huntington's Disease mouse models. The results of these experiments were reported as a series of related papers (Chan et al., 2002; Luthi-Carter et al., 2002a,b). For this analysis we compare cerebellar gene expression in similarly aged mice carrying a wildtype or mutant form of the Huntington's gene. Every comparison reported in Chan et al. (2002), Luthi-Carter et al. (2002a) and Luthi-Carter et al. (2002b) showed some differentially expressed genes, although the amount of differentiation differed considerably between the experiments. For each of the experiments both groups had between 2 and 5 mice. Thus, all our repeats use different samples (sometimes referred to as "biological repeats") and are not repeat arrays using the same samples (sometimes refereed to as "technical repeats"), that could be expected to vary less. There are 6,595 probe sets (genes) on each array.

The experiments listed in Table 3.2 are the seven experiments comparing cerebellar tissue used in Kooperberg et al. (2005); the six experiments using striatum tissue used in that paper are not used here. As for the two-color experiments, some experiments are intended to establish that the tests have the right size and others are intended to establish the power of the tests.

3.4 Results

We analyze the experiments listed in Section 3.3 using the analysis methods described in Section 3.2.2. For the experiments where both groups are different (D2.x and D1.x) we prefer methods with the largest percentage of significant genes (the largest power), provided that the method does have the correct percentage of significant genes in the experiments where both groups are the same (S2.x and S1.x): at most $\alpha\%$ significant genes when tested at significance level α.

Typically we show results for $\alpha = 1\%$ and $\alpha = 0.01\%$. For the two-color arrays there are approximately 11,000 genes after removal of spots (genes) that were too close to background (see Section 3.6) . Assuming independence of genes a 95% confidence interval for the percentage of significance genes based upon the binomial distribution is between 0.8 and 1.2% at $\alpha = 1\%$ and between 0 and 0.03% at $\alpha = 0.01\%$. For the one-color arrays there are 6,595 genes, thus these confidence intervals are slightly larger (0.75 through 1.25% at $\alpha = 1\%$ and 0 and 0.045% at $\alpha = 0.01\%$). When we average four experiments and (incorrect) assume independence for both array types we expect between about 0.9 and 1.1% significant genes at $\alpha = 1\%$ and between 0 and 0.025% at $\alpha = 0.01\%$ for both array types.

3.4.1 Bandwidth selection for smoothers

Three methods (**Cyber-T**, **LPE**, and **Loess**) require the choice of a bandwidth or smoothing parameter. For **LPE** and **Loess** this determines over how many genes the variance is “averaged”. For **Cyber-T** the averaged variance is combined with the variance for the individual genes.

In Table 3.3 we summarize the results for the two-color experiment for the **Loess** approach. The parameter `span` for the `loess()` function in R is approximately linear in the bandwidth for a local linear smoother. From this table we note that the bandwidth has very little influence on the results. The explanation for this is that even for the smallest bandwidth the variances of several dozen genes are effectively averaged. Smaller values of `span` are not useful, as they will increasingly lead to numerical problems in regions where there is less data.

We note that for all four choices of `span` and for all S2.x experiments at $\alpha = 0.01\%$ and for two of the four of these experiments at $\alpha = 1\%$ the percentage of genes that are called significant is much too large. The same was concluded in Kooperberg et al. (2005) for the one-color arrays.

In the remainder of our comparisons we use a `span` of 0.1, which yielded the lowest average number of significant results for both $\alpha = 1\%$ and $\alpha = 0.01\%$ for the four S2.x experiments. As the influence of the bandwidth appears minimal, we will use **Cyber-T** and **LPE** with their default values.

3.4.2 Comparison of methods

In Tables 3.4 and 3.5 we show the results for seven of the methods described in Section 3.2.2 when applied to the two-color and one-color data described in Section 3.3, respectively. Results for the **LPE** method are not available for the two-color

Table 3.3: Performance of the **Loess** approach for various values of the bandwidth (`span`) parameter for the two-color experiments. We report the percentage of genes that are called differentially expressed at levels $\alpha = 1\%$ and $\alpha = 0.01\%$. Ideally the four S2.x experiments would have α differentially expressed genes, while the seven D2.x would have many such genes.

	$\alpha = 1\%$				$\alpha = 0.01\%$			
span	10	1	0.1	0.01	10	1	0.1	0.01
S2.1	1.1	1.1	0.7	0.7	0.340	0.306	0.198	0.159
S2.2	7.8	7.0	5.8	6.6	2.884	2.507	1.528	1.915
S2.3	2.2	2.1	2.0	2.0	0.984	0.922	0.982	0.942
S2.4	0.7	0.6	0.6	0.6	0.262	0.262	0.230	0.212
S2-ave	3.0	2.7	2.3	2.5	1.118	0.999	0.735	0.807
D2.1	25.8	25.9	26.8	27.1	11.941	11.994	12.698	12.827
D2.2	31.7	31.8	32.3	32.9	16.817	17.000	17.682	18.300
D2.3	53.5	53.6	53.8	53.8	38.170	38.354	38.368	38.457
D2.4	54.3	54.4	54.4	54.7	37.709	37.858	37.774	38.043
D2.5	43.3	43.5	43.5	44.2	28.006	28.190	28.225	28.574
D2.6	73.0	73.2	76.5	76.6	62.230	62.431	66.313	66.501
D2.7	62.1	62.3	64.3	64.3	47.863	48.003	50.124	50.471
D2-ave	49.1	49.2	50.2	50.5	34.677	34.833	35.883	36.168

data. Cui et al. 2005 recommends permutations to obtain P-values for the **Shrinking** approach, as in Tables 3.6 and 3.7 and Figure 3.3 and 3.4. In Tables 3.4 and 3.5 and Figure 3.1 and 3.2 we use a normal reference distribution; which distribution is used has a substantial impact on the results.

In Figure 3.1 we give a graphical display of how well these methods adhere to the significance levels, and in Figure 3.2 we display power. These figures are probability-probability plots on a logit-scale. That is, for a particular method and a particular experiment let p_i be the two-sided (sometimes called signed) P-values. That is, if p_i is close to 0 there is evidence of under-expression and if p_i is close to 1 there is evidence of over-expression of group one relative to group two. We now combine all p_i for a group of experiments and sort them. Assume that we have N P-values. We plot the sorted P-values (horizontal) against $(1, \ldots, n)/(N+1)$. When the experiments that we consider are self-versus-self comparisons we would like these plots to follow the identity line, as that implies that the significance levels are “unbiased”. Curves that flatten out are particularly worrisome, as they suggest significantly differentially expressed genes that are in fact false positives. Curves that are more vertical than the identity line suggest statistics that are too conservative: something that is not a concern when there is in fact no difference, but would likely hurt us when we use the same method to analyze data where some genes are differentially expressed. Second, for groups of experiments where there is a difference between both samples we want the most horizontal curves, among the methods that did not generate a substantial number of false positives for the repeat experiments.

From Figure 3.1 we see that the **Loess** and **LPE** approach identify substantially more differentially expressed genes than the nominal levels for the experiments where in fact the two samples being compared are repeats. The **Cyber-T** approach shows a mild number of increases, and none of the other approaches shows serious bias. For both groups of experiments a normal reference distribution for the **Shrinking** approach appears too conservative.

Table 3.4 elaborates on this. At a significance level of $\alpha = 1\%$ only the **Loess** method shows a substantial bias, and it does that for five out of eight data sets. For microarray experiments the more stringent level $\alpha = 0.01\%$ is very relevant, as multiple comparisons corrections often will imply selecting genes at low significance levels. We note that the **Loess** again shows substantial bias. The **LPE** approach also indicates ten times more significant genes than the nominal value; this bias is present for three of the four data sets. At this significance level the **Cyber-T** method shows a modest bias; in particular we notice that the bias is only substantial for one dataset (two-color experiment S2.2). The excess percentage of significant genes for the **Pooled-T** approach is minimal, and could just be due to chance.

From Figure 3.2 we note that for all methods far more genes are identified as differentially expressed by the two-color experiments than by the one-color experiments, as the curves for the two-color experiments are much more horizontal than those for the one color experiments. This is largely an effect of the actual data used, as the two-color Drosophila experiments involved substantially altered flies, while the differences between the mice involved in the one-color Huntington's disease experiments are much more subtle. We do note from this figure though that the ordering of the methods is largely unchanged, suggesting that since our conclusions remain the same for two dramatically different experiments (different technologies, different amounts of differential genes) they are likely fairly robust and may well generalize to many other situations.

For both the two-color and the one-color experiments the **Loess** approach is the most powerful. This is not a surprise, since the method does not maintain significance levels for the experiments where both samples are repeats. Similarly, we are not surprised that the **LPE** method is quite powerful for the one-color experiments. This method also did not maintain significance levels for the experiments where both samples are repeats. Among the remaining methods, we note that the **Pooled-T** approach performs best for the two-color experiments, followed by the **Cyber-T** and **Limma** approach, while for the one-color experiments the **Cyber-T** and **Limma** approach seem slightly more powerful than the **Pooled-T** approach.

Table 3.5 confirms all these conclusions. Interestingly for the D2.x (two-color) experiments we notice that for those experiments with two arrays (D2.1, D2.2, and D2.3) the **Pooled-T** approach is particularly more powerful. Maybe this is not sur-

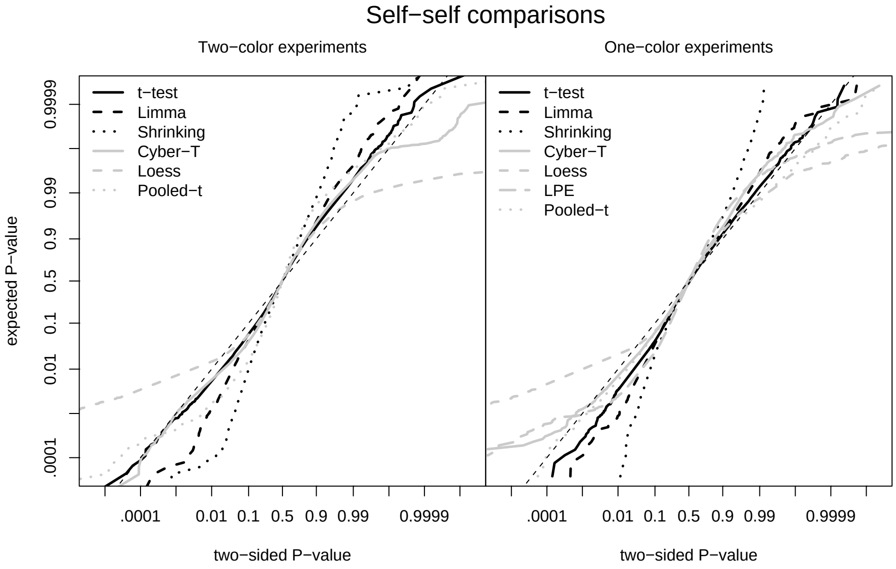


Figure 3.1: Performance of the various approaches to significance testing using an explicit reference distribution for small microarray experiments for the combined two-color and one-color self-versus-self experiments. For unbiased methods the curves should follow the identity line.

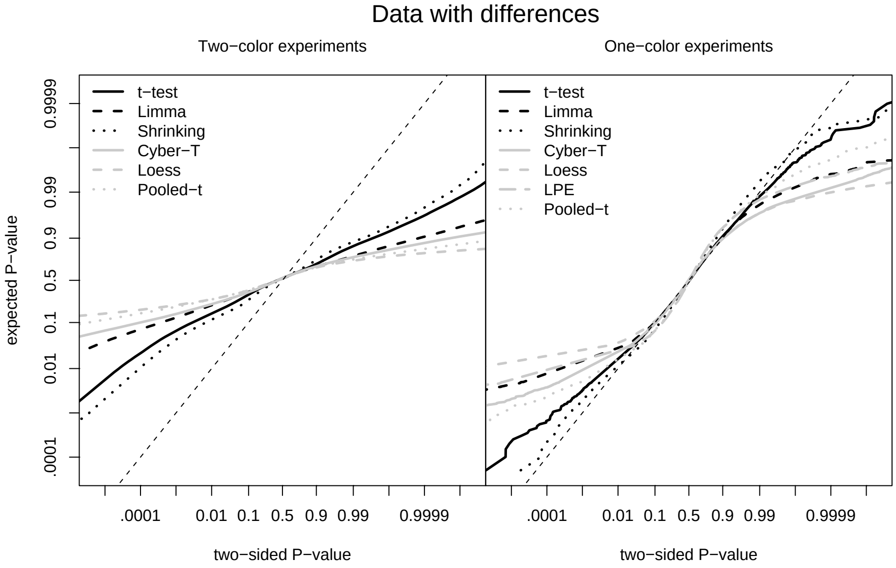


Figure 3.2: Performance of the various approaches to significance testing using an explicit reference distribution for small microarray experiments for the combined two-color and one-color experiments that involve different samples. More horizontal curves correspond to more powerful methods.

prising: the borrowing of degrees of freedom between experiments, as the **Pooled-T** approach is doing, is particularly useful when the number of degrees of freedom is small.

3.4.3 Permutation P-values

As detailed in Section 3.2.3, an alternative approach to obtaining P-values is a permutation approach in which the test statistics for all genes are combined. In Figure 3.3 we give a graphical display of how well each of the methods adhere to the significance levels when P-values are determined using such an approach, and in Figure 3.4 we display power for these situations. We do not show permutation results for the **Pooled-T** approach: since this procedure combines arrays from different experiments a permutation procedure is less standard, besides that the results using a T-distribution already give satisfactory results.

The displays in Figures 3.3 and 3.4 are organized similar to Figures 3.1 and 3.2. We notice that the permutation approach for computing P-values yields approximately unbiased results for all approaches as all curves in Figure 3.3 follow the diagonal. However, as expected, the permutation approach reduces power for any of the approaches using randomization. In Figure 3.4 we note that the procedures based on permutation are considerably less powerful than the procedures that do not use permutation (as shown in Figure 3.2). In particular, we notice that the curves in Figure 3.4 all stay within a “band” of the diagonal. This is in fact a consequence of using the permutation approach with a small number of repeats: irrespective of the actual number of differentially expressed genes, there is a maximum number of genes that can be differentially expressed at any particular significance level thanks to the experimental design. This is explained in detail below in the discussion of Table 3.7.

Tables 3.6 and 3.7 for the permutation based procedures are organized similar to Tables 3.4 and 3.5 for the procedures using a reference distribution. From these tables we draw the same conclusions as from Figures 3.3 and 3.4: while the permutation approach does control the significance level α appropriately, it limits the power. We note from these tables that no methods and no data sets are exceptions. The part of Table 3.7 for the two-color (D2.x) experiments with different samples clearly illustrate an artifact of the permutation approach. As we have seen before, the D2.x experiments have very many differentially expressed genes (see Table 3.5). But in Table 3.7 there seems to be a cap: at a significance level of $\alpha = 1\%$ for experiments D2.1, D2.2, and D2.3 all methods suggest at most 2% differentially expressed genes, for experiments D2.4, D2.5, and D2.6 all methods suggest at most 8% differentially expressed genes, and for experiments D2.7 all methods suggest at most 32% differentially expressed genes. Let’s focuss on experimant D2.4. This is an experiment with 4 arrays. There are thus at most $2^4 = 16$ permutations from “flipping” the arrays. Since each permutation arises twice (when all arrays are flipped relative to the first analysis), only 8 of these permutations are unique. Assume that for this experiment 40% of the genes are differentially expressed (as Table 3.5 suggest), and

Table 3.4: Percentage of differentially expressed genes using various approaches to significance testing using an explicit reference distribution for small microarray experiments for the individual two-color and one-color self-versus-self experiments at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. For unbiased methods the percentage of differentially expressed genes should be close to α.

$\alpha = 1\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE	Pooled-T
S2.1	0.2	0.1	0.0	0.1	0.7	NA	0.3
S2.2	1.1	0.1	0.0	2.3	5.8	NA	0.3
S2.3	0.6	0.2	0.0	0.3	2.0	NA	0.4
S2.4	0.2	0.1	0.0	0.0	0.6	NA	0.1
S2-ave	0.5	0.1	0.0	0.7	2.3	NA	0.3
S1.1	0.4	0.2	0.0	0.4	0.7	0.4	0.0
S1.2	0.6	0.3	0.0	1.4	2.7	1.1	0.2
S1.3	0.8	0.1	0.0	0.3	3.9	0.3	3.2
S1.4	0.3	0.0	0.0	0.1	2.6	0.1	1.3
S1-ave	0.5	0.2	0.0	0.6	2.5	0.5	1.2

$\alpha = 0.01\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE	Pooled-T
S2.1	0.000	0.000	0.000	0.000	0.198	NA	0.017
S2.2	0.009	0.000	0.000	0.277	1.528	NA	0.061
S2.3	0.018	0.000	0.000	0.000	0.982	NA	0.009
S2.4	0.000	0.000	0.000	0.000	0.230	NA	0.009
S2-ave	0.007	0.000	0.000	0.069	0.735	NA	0.024
S1.1	0.015	0.030	0.000	0.061	0.197	0.106	0.000
S1.2	0.000	0.000	0.000	0.045	0.697	0.243	0.000
S1.3	0.000	0.000	0.000	0.015	0.500	0.061	0.091
S1.4	0.000	0.000	0.000	0.000	0.728	0.000	0.000
S1-ave	0.004	0.008	0.000	0.030	0.531	0.102	0.023

Table 3.5: Percentage of differentially expressed genes using various approaches to significance testing using an explicits reference distribution for small microarray experiments for the individual two-color and one-color experiments that involve different samples at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. The larger the percentage of differentially expressed genes, the more powerful a method is.

$\alpha = 1\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE	Pooled-T
D2.1	1.9	12.1	0.0	15.8	26.8	NA	30.9
D2.2	2.3	16.0	0.0	21.9	32.3	NA	28.9
D2.3	4.0	34.8	0.0	43.6	53.8	NA	48.2
D2.4	31.0	44.8	22.6	45.5	54.4	NA	62.7
D2.5	20.9	31.6	13.1	35.1	43.5	NA	52.4
D2.6	53.6	66.5	46.3	66.9	76.5	NA	58.6
D2.7	51.8	57.6	46.9	55.9	64.3	NA	56.3
D2-ave	23.7	37.6	18.4	40.7	50.2	NA	48.3
D1.1	2.6	3.4	2.0	4.0	6.4	2.7	3.3
D1.2	1.2	5.3	0.1	5.6	6.7	5.0	1.5
D1.3	1.6	1.6	1.0	1.6	3.0	0.9	0.8
D1-ave	1.8	3.4	1.1	3.7	5.4	2.9	1.9

$\alpha = 0.01\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE	Pooled-T
D2.1	0.009	0.864	0.000	2.148	12.698	NA	10.835
D2.2	0.026	1.219	0.000	5.051	17.682	NA	11.928
D2.3	0.027	7.699	0.000	19.441	38.368	NA	26.722
D2.4	1.994	15.378	0.296	21.732	37.774	NA	44.632
D2.5	1.083	4.752	0.201	10.856	28.225	NA	31.806
D2.6	7.729	39.769	2.858	47.705	66.313	NA	40.295
D2.7	17.023	29.986	11.971	34.357	50.124	NA	38.347
D2-ave	3.984	14.238	2.189	20.184	35.883	NA	29.224
D1.1	0.121	0.349	0.030	1.046	2.593	0.788	0.516
D1.2	0.000	2.153	0.000	1.668	2.835	2.092	0.243
D1.3	0.106	0.243	0.061	0.379	1.410	0.288	0.182
D1-ave	0.076	0.915	0.030	1.031	2.280	1.056	0.313

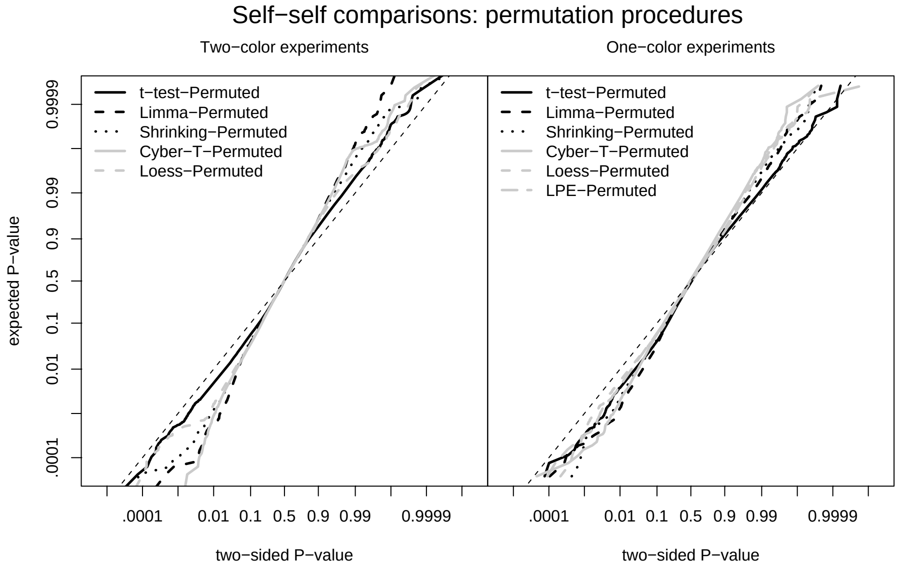


Figure 3.3: Performance of the various approaches to significance testing using a permutation approach rather than a reference distribution for small microarray experiments for the combined two-color and one-color self-versus-self experiments. For unbiased methods the curves should follow the identity line.

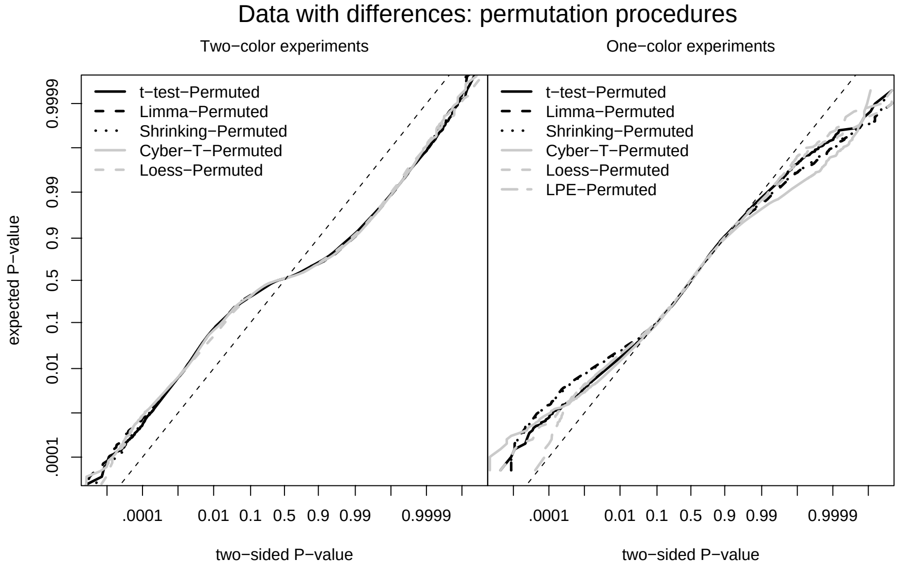


Figure 3.4: Performance of the various approaches to significance testing using a permutation approach for small microarray experiments for the combined two-color and one-color experiments that involve different samples. More horizontal curves correspond to more powerful methods.

Table 3.6: Percentage of differentially expressed genes using various approaches to significance testing using a permutation approach rather than a reference distribution for small microarray experiments for the individual two-color and one-color self-versus-self experiments at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. For unbiased methods the percentage of differentially expressed genes should be close to α.

$\alpha = 1\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE
	permuted	permuted	permuted	permuted	permuted	permuted
S2.1	0.1	0.0	0.0	0.0	0.0	NA
S2.2	1.0	0.0	0.2	0.4	0.6	NA
S2.3	0.6	0.1	0.1	0.0	0.4	NA
S2.4	0.2	0.1	0.1	0.0	0.2	NA
S2-ave	0.5	0.1	0.1	0.1	0.3	NA
S1.1	0.3	0.1	0.1	0.1	0.1	0.1
S1.2	0.6	0.4	0.4	0.3	0.4	0.4
S1.3	1.1	0.5	0.4	0.2	0.5	0.5
S1.4	0.3	0.1	0.1	0.1	0.4	0.2
S1-ave	0.6	0.2	0.2	0.1	0.4	0.3

$\alpha = 0.01\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE
	permuted	permuted	permuted	permuted	permuted	permuted
S2.1	0.000	0.000	0.000	0.000	0.000	NA
S2.2	0.000	0.000	0.000	0.000	0.000	NA
S2.3	0.017	0.000	0.000	0.000	0.000	NA
S2.4	0.000	0.000	0.008	0.000	0.000	NA
S2-ave	0.004	0.000	0.002	0.000	0.000	NA
S1.1	0.000	0.000	0.000	0.000	0.000	0.000
S1.2	0.000	0.000	0.000	0.000	0.000	0.000
S1.3	0.000	0.000	0.000	0.015	0.000	0.015
S1.4	0.000	0.000	0.000	0.000	0.000	0.000
S1-ave	0.000	0.000	0.000	0.004	0.000	0.004

Table 3.7: Percentage of differentially expressed genes using various approaches to significance testing using a permutation approach rather than a reference distribution for small microarray experiments for the individual two-color and one-color experiments that involve different samples at significance levels $\alpha = 1\%$ and $\alpha = 0.01\%$. The larger the percentage of differentially expressed genes, the more powerful a method is.

$\alpha = 1\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE
	permuted	permuted	permuted	permuted	permuted	permuted
D2.1	1.6	2.0	1.8	2.0	2.0	NA
D2.2	1.5	2.0	2.0	2.0	2.0	NA
D2.3	1.9	2.0	2.0	2.0	2.0	NA
D2.4	7.7	8.0	8.0	8.0	8.0	NA
D2.5	7.4	8.0	8.0	7.9	7.5	NA
D2.6	8.0	8.0	8.0	8.0	0.0	NA
D2.7	30.5	31.8	30.5	31.8	24.8	NA
D2-ave	8.4	8.8	8.6	8.8	7.8	
D1.1	2.8	3.8	3.8	3.6	2.8	2.8
D1.2	1.2	3.0	2.6	2.7	2.7	2.7
D1.3	1.9	1.8	1.8	1.4	1.3	1.0
D1-ave	2.0	2.9	2.7	2.6	2.3	2.1

$\alpha = 0.01\%$	T-test	Limma	Shrinking	Cyber-T	Loess	LPE
	permuted	permuted	permuted	permuted	permuted	permuted
D2.1	0.008	0.008	0.008	0.008	0.017	NA
D2.2	0.017	0.017	0.017	0.017	0.026	NA
D2.3	0.009	0.008	0.000	0.009	0.018	NA
D2.4	0.068	0.076	0.076	0.068	0.079	NA
D2.5	0.075	0.083	0.059	0.084	0.079	NA
D2.6	0.075	0.075	0.075	0.025	0.068	NA
D2.7	0.308	0.315	0.283	0.308	0.314	NA
D2-ave	0.080	0.083	0.074	0.074	0.086	NA
D1.1	0.121	0.258	0.212	0.243	0.106	0.030
D1.2	0.000	0.000	0.015	0.015	0.015	0.015
D1.3	0.136	0.243	0.258	0.212	0.121	0.045
D1-ave	0.086	0.167	0.162	0.157	0.081	0.030

these 40% of the genes have very large test-statistics. There are about 10,000 genes on these arrays, thus 4,000 test-statistics are large, say larger than A. Now assume that among the 7 other permutations none of the test-statistics are larger than A. Then out of $8 \times 10,000 = 80,000$ test-statistics 4,000 are larger than A. However, at the $\alpha = 1\%$ level at most $0.01 \times 80,000 = 800$ can be called significant at $\alpha = 1\%$. Which is 8%, rather than the 40% that are differentially expressed, of all the genes on the array. (In fact the percentage is slightly lower as a few rare permuted genes also have large statistics.) We could choose to ignore the "original" permutation in getting the percentiles of the permutation distribution, but this would violate the assumptions of exchangeability under the null-hypothesis of no differential expression. When the number of arrays increases, or when the number of differentially expressed genes is much smaller, this artifact clearly disappears.

3.5 Discussion

The choice of significance test in microarray experiments with low replication can dramatically influence the results. For both one-color and two-color arrays we set up our experiments so that we could both judge which approaches yield approximately unbiased P-values when the experimental conditions are identical, and which approaches are most powerful when both conditions differ. We focused on P-values, rather than for example the FDR, as we believe that a "good" P-value will yield a "good" multiple comparisons correction, and a multiple comparisons adjustment by itself can not save a procedure that yields badly calibrated P-values.

The two groups of experiments that we considered differed in another aspect besides technology: our one-color experiments had a modest number of differentially expressed genes, while our two-color experiments had many such genes. Given the difference between the two groups of experiments the similarity in results was striking.

Our main conclusions are:

- The **T-test** has almost no power when the sample size is small. When there are less than, say, six repeat arrays some of the alternative solutions are much more powerful. Kooperberg et al. (2005) concluded that the lack of power is even more extreme for the Welch statistic.
- Combining an estimate of the overall variance with an estimate of the individual variance, such as is done for **Limma** (Smyth, 2004) and **Cyber-T** (Baldi and Long, 2001) appear very effective. Apparently such a regularization reduces the noise in the variance estimates effectively. Because of the similarity of the results for these two approach, and the much worse results for the smoothing approaches, we hypothesize that for the **Cyber-T** approach the running average estimate of σ_{0ij} is effectively estimating an overall variance, rather than a local variance. In our experiments **Limma** performed slightly better than **Cyber-T**.
- An approach which borrows degrees of freedom from other experiments **Pooled-T**, first proposed in Kooperberg et al. (2005) performs equally well as the **Limma**

and **Cyber-T** approach. In fact, when the sample size is real small ($n = 2$) it seems to perform slightly better. Obviously for this approach the main question is "what to combine". In Kooperberg et al. (2005) a small simulation study was carried out suggesting that there can be a reasonable amount of experiment-to-experiment variation without seriously inflating the type-1 error. The fact that we can without much problem combine cell-line experiments with RNA harvested from fruit-flies (as was done for the two-color experiments in this paper) confirms that conclusion.

- Methods which solely use a smoothed estimate of the variance, such as the **LPE** approach (Jain et al., 2003) and the **Loess** approach (inspired by Huang and Pan (2002)) can give severely biased results by inflating the percentage of significant genes well beyond a pre-specified level α, when in fact there are no differences between the two samples. For the **Loess** approach this was evident at $\alpha = 1\%$ and $\alpha = 0.01\%$, for the **LPE** approach it was only evident at $\alpha = 0.01\%$. However, since for microarray experiments often multiple comparisons corrections are carried out very small significance levels are in fact used, we would want to avoid methods that solely use smoothing approaches. A reason for the bias because of smoothing the variance may be due to the fact that with the normalization methods developed in recent years (see Section 3.6) the relation between variance and expression level has been considerably reduced.
  In particular, in Figure 3.5 for an individual two-color array and one of the two-color experiments and in Figure 3.6 for one of the one-color experiments we show the relation between the difference between the two signals (left side of Figure 3.5) or the variance and the average signal (other panels). As can be seen, the relation between average signal and variance is minimal, and in fact the correlation between the variance from one experiment to the next experiment for the same gene is much larger than the correlations in these figures (data not shown). Thus, locally averaging the variances will sometimes yield variances that are too large and sometimes yield variances that are too small. When the variance is too small there is a substantial chance of incorrectly identifying a gene as differentially expressed.
- A permutation approach to obtaining P-values severely reduces the number of genes that are identified as differentially expressed for experiments with a lot of differential expression. This limits our conclusions about the **Shrinking** approach (Cui et al., 2005), as for this approach it is the only suggested method to obtain P-values.

All approaches that we studied are either available in R-packages available from CRAN or Bioconductor, or are easily implemented in R code.

3.6 Appendix: Normalization of arrays

Two-color arrays For the two-color arrays we first excluded all spots with a log base 2 expression of less than 5, and spots whose background level was higher than

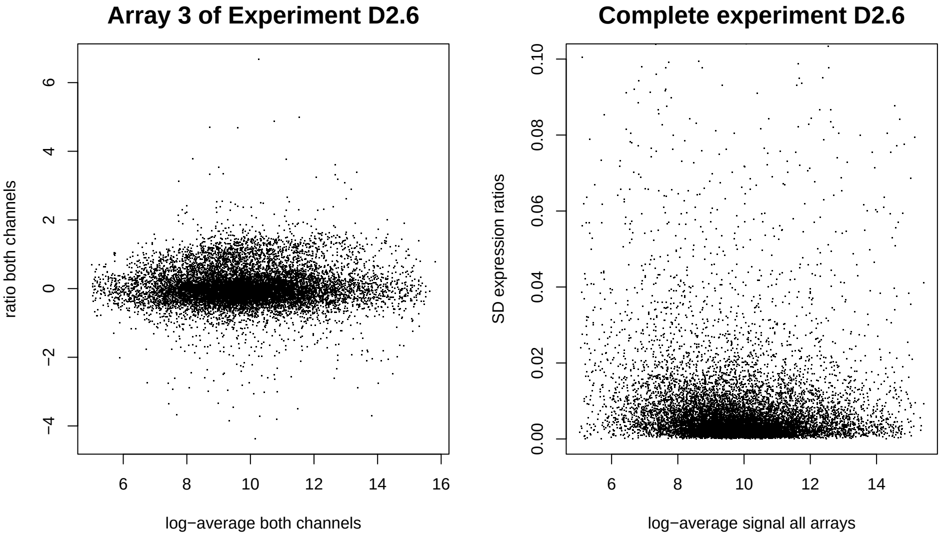


Figure 3.5: Relation between log expression ratio and average log expression for one normalized two-color array, and the standard deviation of log expression ratios with the average log expression ratio for all arrays from that experiment.

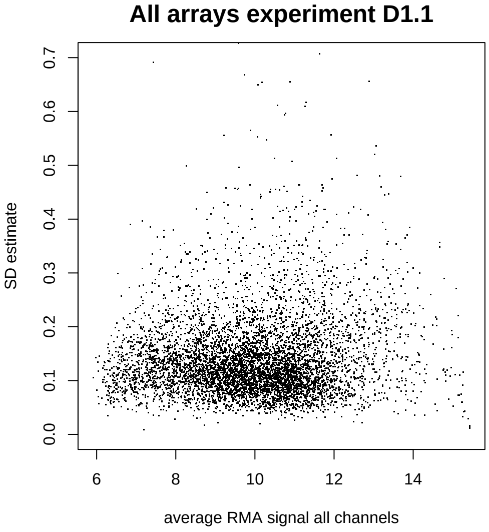


Figure 3.6: Relation between the residual standard deviation and the average RMA normalized expression for one of the one-color experiments.

the foreground level for either channel. This excludes about 11.5% of the spots, primarily spots that do not hybridize well. In particular of the 13,440 spots on our arrays, 1,296 were excluded on all 36 arrays: of the remaining spots only about 2% were excluded. We then subtracted the background and used a print-tip loess correction using the Limma function `normalizeWithinArrays()` with defaults. Any spot that had at least two estimates for a particular experiment was included in our analysis. We employed various graphical QC tools, and felt that all arrays were of good quality.

One-color arrays For all methods we analyzed gene expressions that were normalized by the RMA algorithm of Irizarry et al. (2003b). We also carried out the same analysis using the log of the MAS5 Average Difference summary and obtained essentially the same results. For RMA we normalized all arrays simultaneously; however when we analyzed each of the experiments separately, the results were again essentially the same. We employed various graphical QC tools, and felt that all arrays were of good quality.

Acknowledgments

Charles Kooperberg was supported in part by NIH grants CA 74841, CA 53996, and HL 74745. Aaron Aragaki was supported in part by NIH grant CA 74841. Charles C. Carey was supported in part by NIH interdiciplinary training grant 2T32CA80416-06. Suzannah Rutherford was supported in part by NIH grants GM 06673. The authors thank Andy Strand, Jim Olson, and the HDAG group for allowing us to use the one-color data.

Gene Mapping

CHAPTER 4

Utilizing Information from Multiple Genome Scan Studies in an Empirical Bayes Framework for Strengthening Inferences and Narrowing QTL Location Estimates

Kui Zhang[1], Howard Wiener[2], T. Mark Beasley[1], Christopher I. Amos[3], David B. Allison[1,4]
[1]Department of Biostatistics, Section of Statistical Genetics, The University of Alabama at Birmingham, Birmingham, AL, 35294
[2]Department of Epidemiology, The University of Alabama at Birmingham, Birmingham, AL, 35294
[3]Department of Epidemiology, The U.T. MD Anderson Cancer Center, Houston, TX, 77030
[4]Department of Nutrition Sciences and Clinical Nutrition Research Center, The University of Alabama at Birmingham, Birmingham, AL, 35294

4.1 Abstract

Compared with an individual genome scan study for quantitative trait locus (QTL) mapping, meta-analysis, which can formally utilize the data from multiple genome scan studies, generally can provide higher statistical power, more precise estimates of QTL location and effect, and tighter confidence intervals for QTL location and effect. Therefore, meta-analysis is very useful for prioritizing regions for subsequent follow-up studies. In some situations, investigators who already have initialized a genome scan study would like to evaluate the ”believability” of apparent linkage signals by examining the results of other genome scan studies of the same trait and informally update their beliefs about which linkage signals in their own scan most merit follow-up via a subjective-intuitive integration approach. In this situation, meta-analysis methods may not be suitable because they treat all genome scan studies ”equally”, which is not subjective to the initial genome scan study conducted by the investigator. In the contrast, empirical Bayes (EB) based methods that can formally

borrow information from other genome scan studies to update the estimates and adjust the confidence in finding in an objective fashion could be useful. In empirical Bayes based methods, the linkage statistics from other genome scan studies are used as prior information to update the linkage statistics obtained from the genome scan study conducted by the investigator. The updated linkage statistics can then be used to estimate the QTL location and effect. In this chapter, we summarize the empirical Bayes based methods for multiple genome scan studies using sib pairs. We also evaluate their performance in terms of their power to and their accuracy to estimate the QTL location and effect, using extensive simulations based on actual marker spacing and allele frequencies from available data. Results indicate that the empirical Bayes based methods are insensitive to between-study heterogeneity. The empirical Bayes based methods can yield higher statistical power, generate more precise estimates for the QTL location and effect, and provide narrower confidence intervals than results from an individual study.

4.2 Introduction

Genome scan studies for linkage analysis have been widely used to search candidate regions containing quantitative trait loci (QTLs). However, most genome scan studies for QTL mapping are analyzed without formal consideration of information provided by other genome scan studies of the same trait. The resulting candidate regions often contain a large number of functional genes due to the lower power of individual genome scan studies. Consequently, the subsequent fine mapping and positional cloning for these candidate regions may be problematic. When multiple genome scan studies of the same trait are available, we may increase the power to detect the linkage between markers and QTLs by using information provided from all these studies. Methods that can formally integrate data from multiple genome scan studies have been emerging as useful and powerful tools in the field of linkage analysis for QTL mapping.

Marked heterogeneity can exist in multiple genome scan studies and pose daunting challenges in such analysis. Different genome scan studies can use different genetic marker loci and marker maps, different statistical methods to test for linkage, and different sampling schemes. Furthermore, the QTL effect can vary across studies because of disparate environmental effects and population substructures. The combination of raw data from all studies with a well-designed pre-analysis procedure would be a preferred approach to overcome such difficulties. However, in many situations this is not feasible because only some statistics, rather than the raw data, are available. For these reasons, two closely related but distinct groups of methods have been developed to test and map QTLs by integrating same type of statistics obtained from multiple genome scan studies: Meta-analysis and Empirical Bayes (EB).

The first group of methods is meta-analyses, which can be viewed as a set of statistical procedures designed to summarize statistics across independent studies that address similar scientific questions. Several meta-analysis methods have been developed to detect linkage between genetic markers and QTLs (Allison and Heo, 1998;

Etzel and Guerra, 2002; Gu et al., 1998; Guerra, 2002; Guerra et al., 1999; Hedges and Olkin, 1985; Li and Rao, 1996; Rice, 1997; Wise et al., 1999). For example, Allison and Heo (1998) used Fisher's method (Fisher, 1925) to show strong evidence of linkage in OB regions by combining p-values from five published linkage studies on these regions. They also illustrated this method's applicability in the presence of marked heterogeneity across studies. This technique has also been used by other researchers (e.g., Guerra, 2002; Wise et al., 1999). However, it is difficult to use this technique to estimate the parameters of interested in, such as the location and effect of a QTL, because of the method's nonparametric nature. At the same time, several meta-analysis methods that can estimate the parameters of interest across studies, such as the location and effect of a QTL, by combining estimates of Haseman-Elston regression slopes and associated variances at marker loci (Haseman and Elston, 1972) have been developed too (Etzel and Guerra, 2002; Gu et al., 1998; Li and Rao, 1996). The weighted least-square estimator (WLSE) developed by Etzel and Guerra (2002) does not require the same marker map or the same QTL effect across all studies. For a more detailed review of these meta-analysis methods, please refer to (Some References; BOOK CHAPTER in the Same Book).

The second group of methods is based on the EB framework (Beasley et al., 2005; Bonney et al. 1992; Zhang et al., 2005). In EB based methods, the linkage statistics (e.g., Haseman-Elston regression slopes and their associated variances at marker loci) are obtained from each individual genome scan study and then the linkage statistics from an individual study of interest are updated by incorporating the linkage statistics from other studies. The updated linkage statistics can be used for detecting the linkage between a marker and the QTL and mapping the location of the QTL. It is worth emphasizing the key difference between the EB based methods and the meta-analysis methods. In the empirical Bayes analysis, an individual genome scan study of interest is identified as the primary study and the rest of studies are considered as the background studies. Theoretically, each individual study can be claimed as the primary study. However, the study of primary interest to an investigator would be the study conducted by the investigator; presumably the investigator would be able to obtain further genotypes from the individuals in the primary study for fine mapping, while this type of information would not necessarily be available from the background studies. In the meta-analysis methods, each individual study is equivalent to the other studies used and investigators are interested in the overall results.

The rest of this paper is organized as follows: the EB based method for the genome scan studies using sib pairs and the simulation design are described in Methods section; the assessment of the performance of the empirical Bayes based methods for detecting linkage between markers and the QTL and mapping the location of the QTL are presented in Results section; the conclusions, the implications, and the possible extensions of the empirical Bayes based method are given in Discussion section.

4.3 Methods

4.3.1 *Haseman-Elston Regression Analysis for a Single Genome Scan Study with m Markers using Sib Pairs*

The Haseman-Elston Regression has been widely used to detect the linkage between genetic markers and QTLs using sib pairs. Suppose that the trait values, the squared trait difference, and the estimated proportion of alleles shared identical-by-descent (IBD) at a marker locus for the ith sib pair are denoted as $y_i = (y_{i1}, y_{i2})$, $Y_i^D = (y_{i1} - y_{i2})^2$, and π_i, respectively. Then the Haseman-Elston method can be represented by a simple regression of Y_i^D on π_i :

$$Y^D = \beta_0 + \beta\pi + \varepsilon$$

The regression slope β has the expectation $E(\beta) = -2(1-2\theta)^2\sigma_g^2$, where θ is the recombination fraction between the marker locus and the QTL, and σ_g^2 is the phenotypic variance explained by the additive effect of this QTL. Thus, the regression slope β is 0 under the null hypothesis of no linkage, and is negative under the alternative hypothesis. Specifically, if there are m markers and the estimates of the slope and its associated variance at each marker are denoted by $\hat{\beta}_j$ and $\hat{S}_j^2$, $(j = 1, \ldots, m)$, then the t statistic $t_j = -\hat{\beta}_j/\sqrt{\hat{S}_j^2}$ asymptotically follows a standard normal distribution under the null hypothesis of no linkage. The null hypothesis is rejected with the 5% nominal level at the j th marker if t_j exceeds 1.645. It is also worth noting that the regression slope and its associated variance are only estimated at the marker loci with determined genotype in the original Haseman-Elston method. Due to the coarse marker map in linkage analysis, this method is more suitable for detecting linkage between markers and the QTLs rather than estimating the QTL location and effect.

The Haseman-Elston methods assumes the normality of trait values and is robust even this assumption is violated for a reasonably large sample size ($n > 100$ sib pairs) (Allison et al., 2000). However, the original Haseman-Elston regression tends to have lower power than the variance component method. Other modified Haseman-Elston regression methods were subsequently developed (Amos 1994; Drigalenko 1998; Elston et al., 2000; Feingold, 2002; Sham et al., 2001; Xu et al., 2000). For example, additional power can be acquired by regressing the mean corrected squared sums of trait values $Y_i^S = (y_{i1} - \bar{y} + y_{i2} - \bar{y})$ (Drigalenko 1998), the mean corrected cross product of trait values $Y_i^P = (y_{i1} - \bar{y})(y_{i2} - \bar{y}) = (Y_i^S - Y_i^D)^2$ (Drigalenko 1998; Elston et al., 2000), or a weighted combination of Y_i^D and Y_i^S (Xu et al., 2001) on π_i, where $\bar{y}$ is the mean trait value over all sib pairs.

4.3.2 *The Interval Mapping (IM) Method to Detect Linkage between Markers and QTLs and Estimate the QTL Location and Effect Based on m Markers from a Single Genome Scan Study Using Sib Pair*

Fulker at al. (1995) developed an interval mapping (IM) method to detect linkage between markers and QTLs and to estimate the QTL location and effect. They first used the estimated proportion of alleles shared IBD at all marker loci on a single chromosome and the genetic distance between these markers to estimate the proportion of IBD sharing at virtually any location on the chromosome and then perform the Haseman-Elston regression at this location (Fulker et al., 1995). Suppose that the estimates of regression slope and its associated variance at each analysis point q along the chromosome are denoted by $\hat{\beta}_q$ and $\hat{S}_q^2$, respectively. At any analysis point q, the null hypothesis of no linkage is rejected at the nominal 5% level if the value of test statistic $\hat{t}_q = -\hat{\beta}_q/\sqrt{\hat{S}_q^2}$ is greater than 1.645. The analysis point, $\hat{q}$, that gives the maximum value of the test statistic $\hat{t}_q = -\hat{\beta}_q/\sqrt{\hat{S}_q^2}$, is taken as the estimate for the QTL location. The point estimate of QTL effect, σ_g^2, is given by $\hat{\sigma}_g^2 = -\hat{\beta}_{\hat{q}}/2$.

4.3.3 *Empirical Bayes Model (Bayesian Hierarchical Normal Model)*

In Bayesian analysis, the choice of reasonable prior distribution for parameters is sometimes not obvious. However, if data from several independent studies are available, the prior information can be extracted from the data. Such approaches are called empirical Bayes methods (Carlin and Louis, 2000a). These methods can be viewed as approximations to a complete hierarchical Bayesian analysis; hybrid approaches between classical frequentist methods and fully Bayesian methods. Both parametric and non-parametric approaches exist (Carlin and Louis 2000b), but even the parametric varieties do not depend on strong distributional assumptions (Efron and Morris, 1973).

The empirical Bayes approach as proposed by Efron and Morris (1973; 1975) can be described by a widely used two-level hierarchical normal model. Suppose β is the parameter of interest and there are k populations available to estimate β_i in each population, where β_i can be different among k populations. At the first level, the maximum likelihood estimators $\hat{\beta}_i (i = 1, \ldots, k)$ for β_i can be obtained and we assume that $\hat{\beta}_i|\beta_i$ asymptotically follows a normal distribution, $N(\beta_i, S_i^2)$. At the second level, β_i is specified by a normal model with an r -dimensional predictor x_i, a common regression coefficient μ , and an unknown variance $A \geq 0$; i.e., $\beta_i|\mu \sim N(x_i'\mu, A)$. Using the Bayesian rule, it is easy to compute the marginal distribution of $\hat{\beta}_i$ (given μ and A) and conditional distribution of β_i (given $\hat{\beta}_i$, μ, and A):

$$\hat{\beta}_i|\mu, A \sim N(x_i'\mu, S_i^2 + A), i = 1, \ldots, k \tag{4.1}$$

and

$$\beta_i|\hat{\beta}_i, \mu, A \sim N((1-B_i)\hat{\beta}_i + B_i x_i'\mu, S_i^2(1-B_i)), i = 1, \ldots, k \tag{4.2}$$

where $B_i = S_i^2/(S_i^2 + A)$ is an unknown shrinkage factor. Generally, S_i^2 is unknown and is replaced by $\hat{S}_i^2$, the estimates of associated variance of β_i. A and μ can be estimated by the maximum likelihood methods or by more advanced techniques developed by Tang and Morris (2003). Then we can use $\tilde{\beta}_i = (1-B_i)\hat{\beta}_i + B_i x_i'\mu$ and $\tilde{S}_i^2(1-B_i)$ as the final estimator for β_i and its associated variance, respectively.

4.3.4 Application of the Empirical Bayes Method to Each Marker Based on k Studies with m markers and Identical Marker Map

Empirical Bayes methods have been used in many contexts, including genetic research (Bonney et al., 1992; Li and Rao, 1996; Lockwood et al., 2001; Witte, 1997). We tailored the general empirical Bayes procedure for linkage analysis. Assume that data for the detection of linkage to the same QTL are available from k genome scans using sib pairs. Within each of the k studies, a set of m markers with the identical map are used. For each marker locus from each study the regression coefficient, β_{ij}, is the parameter of interest and describes the effect of the putative QTL on the phenotype. The expectation of β_{ij} equals to $-2(1-2\theta_{ij})^2\sigma_{gi}^2$ at marker locus $j (j = 1, \ldots, m)$ in study $i (i = 1, \ldots, k)$, where θ_{ij} is the recombination fraction between the QTL and the marker j in study i and σ_{gi}^2 is the total genetic variance of the QTL in study i. From the Haseman-Elston regression analysis, we can obtain the estimator $\hat{\beta}_{ij}$ for β_{ij} and its estimated sampling variance$\hat{S}_{ij}^2 (i = 1, \ldots, k; j = 1, \ldots, m)$. For k available studies, all k studies are first used to estimate parameters μ_j and A_j ($j = 1, \ldots, m$), then the empirical Bayes estimators $\tilde{\beta}_{ij}$ and $\tilde{S}_{ij}^2$ for β_{ij} and associated variance can be easily obtained using formulas (4.1) and (4.2) for each of k studies.

4.3.5 The IM-EB Method to Detect Linkage between Markers and QTLs and Estimate the QTL Location and Effect Based on m Markers and k Genome Scan Studies Using Sib Pairs

In this section, we give the detailed desription for the IM-EB method to detect linkage between markers and QTLs and estimate the QTL location and effect from multiple genome scan studies using sib pairs. We assume that data of genome scans using sib pairs with the same trait are available and consider the first study as the primary study. Within each of the studies, a set of markers are used within the same chromosomal region and are denoted as $M_{ij} (i = 1, \ldots, k; j = 1, \ldots, m)$.

For the IM-EB method, the estimates of the regression slope and its associated variance, $\hat{\beta}_{iq}$ and $\hat{S}_{iq}^2$ ($i = 1, \ldots, k$) at each analysis point q on the chromosome, are obtained using the IM method (Fulker et al., 1995). Then, the empirical Bayes estimates, $\tilde{\beta}_{iq}$ and $\tilde{S}_{iq}^2$ ($i = 1, \ldots, k$), are obtained from each of the k studies by using GRIMM (Tang and Morris, 2003). GRIMM is independently applied to each analysis point along the chromosome. The test statistic for the primary study is then calculated on the basis of $\hat{t}_{1q} = -\hat{\beta}_{1q}/\sqrt{\hat{S}^2_{1q}}$and $\tilde{t}_{1q} = -\tilde{\beta}_{1q}/\sqrt{\tilde{S}^2_{1q}}$ at the analysis point q. The analysis point $\hat{q}$ having a maximum value $\hat{t}_{1\hat{q}}$ over the entire chromosome is considered as the IM estimate of QTL location and consequently, the IM estimate of σ^2_{1g} is given by $\hat{\sigma}^2_{1g} = -\hat{\beta}_{i\hat{q}}/2$. The same procedure can be applied to $\tilde{t}_{1q}$ to obtain $\tilde{q}$ and $\tilde{\sigma}^2_{1g}$, the IM-EB estimates of QTL location and effect, respectively. At each analysis point q , the null hypothesis of no linkage is rejected with the 5% nominal level by the IM estimator if the value of test statistic $\hat{t}_{1q} = -\hat{\beta}_{1q}/\sqrt{\hat{S}^2_{1q}}$ is greater than 1.645. Similarly, the null hypothesis is rejected at the 5% nominal level if the value of test statistic $\tilde{t}_{1q} = -\tilde{\beta}_{1q}/\sqrt{\tilde{S}^2_{1q}}$ is greater than 1.645.

4.3.6 Simulation Designs

To investigate the performance of the empirical Bayes method to incorporate data from multiple genome scan studies using sib pairs, we conducted the following simulations. We assumed that there is only one QTL with no background polygenic variation and no shared sib environment effect, or equivalently that such effects are subsumed into the residual variance. There were two alleles at the QTL with the high-risk allele having a frequency of 0.05. We chose 15 microsatellite markers on chromosome 11 that were used for a recent genome scan of Alzheimer's disease (Blacker et al., 2003) because it provides known parameters, including the location and the allele frequencies at each marker locus, for simulations. The trait value of each individual was generated according to the genetic model, $y = \mu + g + \varepsilon$, where μ is the overall trait mean across the population, g is additive effect of the high-risk allele, and ε is the normally distributed random error. We set $E(\varepsilon) = 0$, $cov(g, \varepsilon) = 0$, and $\mu = 70$ and set the total variance of g and ε, $\sigma^2_g + \sigma^2_\varepsilon$ as 1 for all studies.

For each simulation, 5, 10, or 15 studies were generated corresponding to a single study of interest with 4, 9, or 14 background studies, respectively. We generated 500 unrelated sib pairs in each study. We used the same marker map for all studies. For the primary study, the QTL was positioned 65cM from the p-terminus of the chromosome. The heritability of QTL was set either to 0 (without QTL effect) or 15% (with non-zero QTL effect). For background studies, the location and the heritability of QTL could be same as, or different from that of the primary study. The location of QTL in background studies was set either at 35cM or at 65 cM from the p-terminus of the chromosome. The marker locations along with the QTL location are shown in Figure 4.1. The heritability of QTL in each background study varied between 0 and 25% in increments of 5%, which represented variation from the primary study. The number of background studies with non-zero QTL effect varied but all background studies having a non-zero QTL effect were given the same value of heritability. This simulation strategy can accommodate different degrees of heterogeneity among the primary study and the background studies. It can also include a variety of combinations of weak to strong linkage signals among the primary study and background studies. For example, we can set the QTL heritability in the primary study to 15%, the heritability of half of the background studies to 0, and the heritability of the other half of background studies to 25% to represent the situation that the primary study has the moderate linkage signal while some background studies show small to no QTL effect and some of background studies have stronger QTL effect. Other situations can be easily accommodated by varying the number of background studies with non-zero QTL effect and their heritability.

Once the genotypic and phenotypic data were generated, the estimates of the Haseman-Elston regression slopes and their associated variances at each marker or analysis point in each study were determined by regressing the weighted combination of Y^D and Y^S on π (Xu et al., 2001).

4.4 Results

To assess the performance of the empirical Bayes based method, IM-EB, in terms of its power to detect linkage between markers and QTLs and its accuracy to estimate the QTL location and effect, we adapted different simulation strategies and recorded and used different summary statistics.

4.4.1 The Type I Error Rate and Power of the IM-EB method to Detect Linkage between Markers and QTLs

We first investigate the type I error rate of the IM-EB method. It is important to understand that a null model in this context refers only to the study of interest, whether or not the background studies contain a linked QTL. We generated 1,000 data sets with 5, 10, and 15 studies. In all studies, the QTL was positioned 65cM from the p-terminus of the chromosome. In the primary study, the heritability of QTL was set to 0. In the background studies, the heritability was set either to 0 or some value between 5% and 25% in increments of 5%. The number of background studies with non-zero QTL effect varied. Under any particular condition, all background studies with non-zero QTL effect had the same heritability. In the primary study, the null hypothesis was rejected when the IM-EB statistics at 65cM from the p-terminus of the chromosome exceeded 1.645. Figure 4.2 shows the type I error rate of the IM-EB method, which is the proportion of simulations in which the null hypothesis was rejected.

For 5 studies, the number of background studies with non-zero QTL effect was set to 1, 2, 3, or 4. It can be seen when three or fewer background studies have non-zero QTL effect, the type I error rate stays below to the nominal 5% error rate. When all 4 background studies have a heritability of 10%, the type I error rate can be greater than the nominal 5% error rate. The highest type I error rate is 8.5% for all 4 background studies having a heritability of 25%. For 10 studies, the number of background studies with non-zero QTL effect was set to 2, 4, 6, 8, or 9. When there are fewer than 4 background studies having non-zero heritability as high as 20%, the type I error rate is below to the nominal 5% error rate. The type I error rate is inflated when 8 or more background studies have a heritability greater than 10%. The highest type I error is 16%. For 15 studies, the number of background studies with non-zero QTL effect was set to 3, 6, 9, 12, or 14. When fewer than 6 background studies have a heritability as high as 25%, the type I error rates do not t exceed the nominal 5% rate. Again, the type I error rate is inflated when 9 or more background studies have a heritability greater than 10%. The type I error rate is 20% when all 14 background studies have a heritability of 25%. In summary, the type I error rate of the IM-EB stays below to the nominal 5% rate when most of the background studies have a heritability less than 10%. At the same time, we did find some inflated type I error rates of the IM-EB method when most of background studies have a higher heritability. This is expected because the empirical Bayes based method borrows the information from the other studies. If there are a large number of studies with the large QTL effect, the empirical Bayes based method will detect a QTL even the results from the primary study shows small to no effect. However, from an EB perspective, it is debatable whether this situation is truly a ”null’ situation.

We then investigate the power of the IM-EB method. We generated 1,000 data sets with 5, 10, and 15 studies. In the primary study, the heritability of QTL was set to 15%. In the background studies, the heritability was set either 0 or some value between 5% and 25% in increments of 5%. The number of background studies with non-zero QTL effect varied and all background studies with non-zero QTL effect had the same heritability.

Figure 4.3 shows the power of the IM-EB method, which is the proportion of simulations in which the null hypothesis was rejected. In these simulations, the QTL was positioned 65cM from the p-terminus of the chromosome. In the previous subsection, we used 1.645 as the 95% cutoff value to reject the null hypothesis of no linkage between the marker and the QTL. This value is only valid for one single study. When the empirical Bayes based method was used, this cutoff value tends to be conservative. We followed the method proposed by Beasley et al. (2005) to determine the cutoff value. We simulated 1,000 data sets with 5, 10, and 15 studies. All studies had no QTL effect. For the IM-EB method, the 95% cutoff values were 1.464, 1.406, and 1.224 for 5, 10, and 15 studies. These simulated cutoff values were used as critical values to reject the null hypothesis at the nominal 5% level.

It can be seen from Figure 4.3, the power of the IM-EB estimator can be substantially increased when a majority of background studies have the same or higher QTL effect. When all 4, 9, and 14 background studies have a heritability of 15%, the power of the IM-EB estimator increases from 0.191 (the power of the IM estimator for an individual study) to 0.266, 0.343, and 0.466, respectively. When all 4, 9, and 14 background studies have a heritability of 25%, the power of the IM-EB estimator increases to 0.322, 0.525, and 0.688, respectively. The power of the IM-EB estimator also increases even when some of the background studies disagree with the primary study. For example, when about half of the background studies have no QTL effect and half of the background studies have the same heritability of 15%, the power of the IM-EB estimator is 0.224, 0.221, and 0.345 for 4, 9 and 14 background studies, respectively. As would be expected, the increase in power is slightly less than the situation when all of the studies agreed.

To see how the existence of other QTLs along the same chromosome affects the power of the IM-EB method, we simulated data sets in which all background studies had a heritability of 15% but half of them had the QTL positioned 35cM from the p-terminus of the chromosome. The power of the IM-EB estimator at each marker locus is shown in Figure 4.4. We find that the IM-EB estimator increases the power to detect linkage near the QTL of interest at a very small cost of inflated type I error rates at 35cM.

4.4.2 The Accuracy of the IM-EB estimates for QTL Location and QTL Effect

To investigate the accuracy of the IM-EB estimates for QTL location, we recorded their mean value (MEAN), their standard error (STD), and the square root of the mean squared difference between the estimates and the true value (MSE). The simulations adapted here were same with those described in the previous subsection. Specifically, we generated 1,000 data sets with 5, 10, and 15 studies. In the primary study, the QTL was positioned 65cM from the p-terminus of the chromosome and the heritability of the QTL was set to 15%. In the background studies, the heritability was set either to 0 or some value between 5% and 25% in increments of 5%. The number of background studies with non-zero QTL effect varied. Under any particular condition, all background studies with non-zero QTL effect had the same heritability.

The mean and MSE of the IM and IM-EB point estimates for the QTL location and effect under several different simulation strategies are presented in Table 4.1 and 4.2. Several general conclusions emerged from these two tables. First, as expected, the empirical Bayes based method (the IM-EB method) using multiple studies estimate the QTL location and effect more precisely and supply a smaller MSE than does the IM method using an individual study in most situations we simulated. This improvement becomes more notable with more independent studies having larger QTL heritability included in the analysis. For 5 studies, the MSE of the estimates for the QTL location is reduced 5% (from 41.4 to 39.0) when all 4 background studies have a heritability of 25%. For 10 studies, the MSE of the estimates for the QTL location is reduced 18% (from 42.4 to 34.9) when all 9 background studies have a heritability of 25%. Second, the heterogeneity among background studies and the disagreement between the primary study and background studies only slightly affect the accuracy of the IM-EB estimates. In addition, we did not observe a large bias for the estimates of either the QTL location or effect in the presence of other QTLs and different QTL effects in the background studies, as observed in Zhang et al. (2005).

4.5 Discussion

With availability of multiple genome scan studies detecting the linkage between the same QTL and the marker, there is a need to develop novel methods that can borrow or combine information from all available studies. Historically, there are two kinds of methods: the meta-analysis methods and the empirical Bayes based methods. In this paper, we summarized the empirical Bayes based methods (Beasley at el., 2005; Zhang et al., 2005) and assessed their performance using extensive simulations. We found that the empirical Bayes based methods have more power to detect the QTL and provide more precise estimates of QTL location and effect than do methods using an individual study.

To assess the effect of the heterogeneity among studies, we assumed the background studies could have no QTL effect, have a non-zero QTL effect different from that of the primary study, or have the QTLs different from that of the primary study. Although the influence of these factors varies, the empirical Bayes methods were generally robust under all simulated situations. That is, they had more power to detect the QTL and yielded more precise estimates for QTL location and effect, with type I error increased only under extreme situations. In simulations, we assumed that all studies had identical marker maps. This is not required by the empirical Bayes based methods. In addition, varied marker maps across studies had the slight impact on the empirical Bayes based methods and could be helpful in a few situations (Zhang et al., 2005).

We did not compare the empirical Bayes based methods with meta-analysis methods in this paper. Zhang et al. (2005) compared several empirical Bayes based methods with a weighted least-square methods developed by Etzel and Guerra (2002). Their results showed that no method was superior to any other under all simulation situations. Although it is great of interest to conduct such comparison, it is important to point out that the empirical Bayes based methods introduced here are not meta-analysis methods. In the meta-analysis methods, results from several studies of the same relationship are combined to obtain an overall inference or estimate of that relationship. In such an analysis, the results of the studies are combined with equal regard weighted by their relative precisions. In the empirical Bayes based methods, there is one study of primary interest, whereas the rest of studies are regarded as background studies. The results obtained from background studies are incorporated as prior information to improve the inference or estimate for the primary study.

In summary, we conclude that the empirical Bayes based methods can account for between-study heterogeneity. They can have more power to detect linkage between markers and QTL and provide more precise estimates for the QTL location and effect.

4.6 Acknowledgement

This paper is supported by grant R01ES09912 from National Institutes of Health.

4.7 Figures and Tables

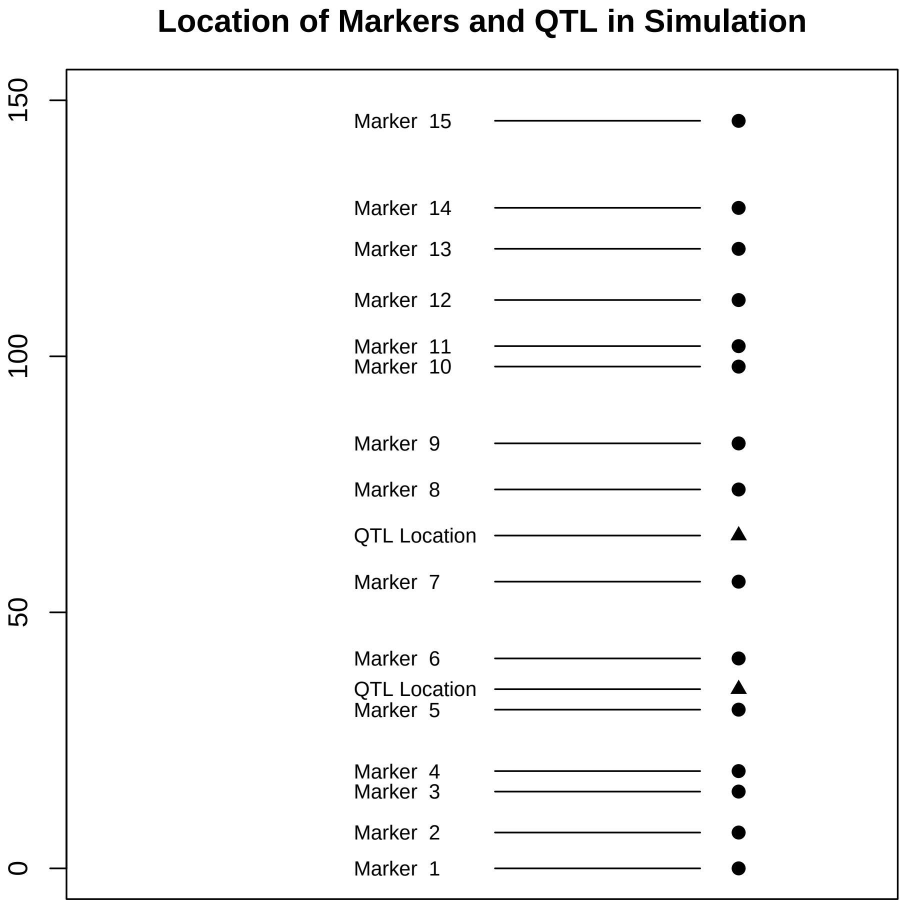


Figure 4.1: The actual map for 15 micro-satellite markers from the National Institute of Mental Health Alzheimer's Diseases Genetics Initiative and the locations of two hypothetical QTLs used in simulations. The minimum distances between the marker and two QTLs, 65cM and 35cM from the p-terminus of the chromosome, are 9cM and 4cM, respectively.

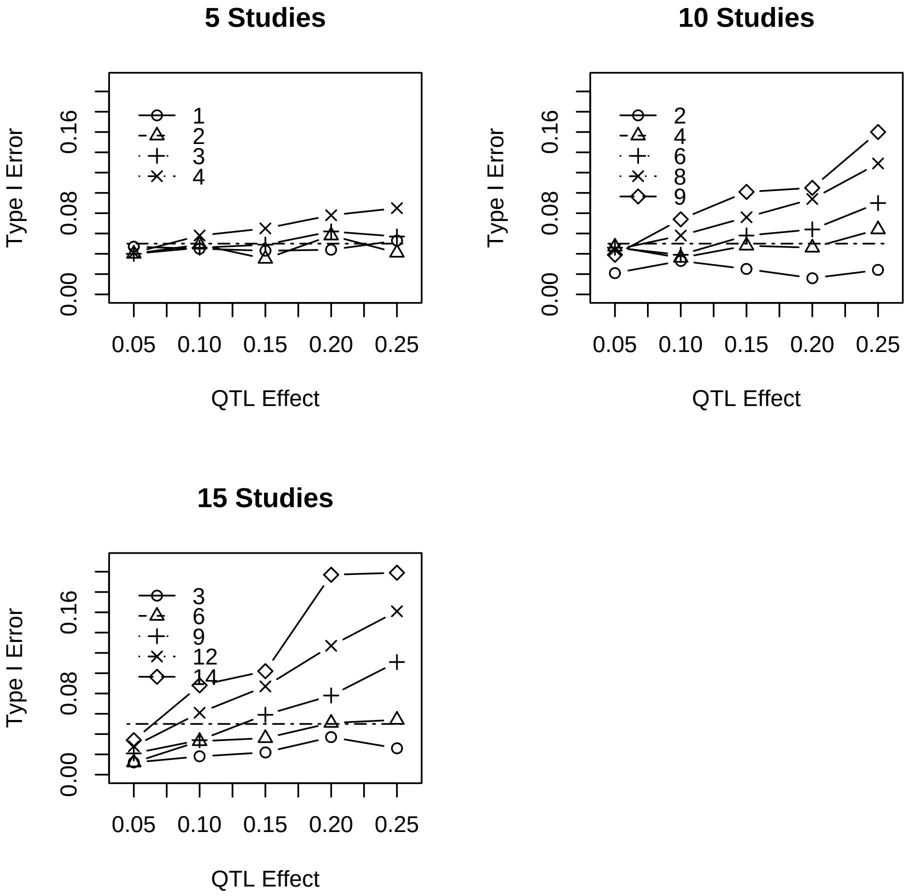


Figure 4.2: The type I error rates of the IM-EB estimator at the 65cM from the p-terminus of the chromosome with 5, 10, and 15 studies. The QTL in all studies were positioned 65cM from the p-terminus of the chromosome. In the primary study, the heritability of QTL was set to 15% and the number of background study having non-zero QTL effect varied.

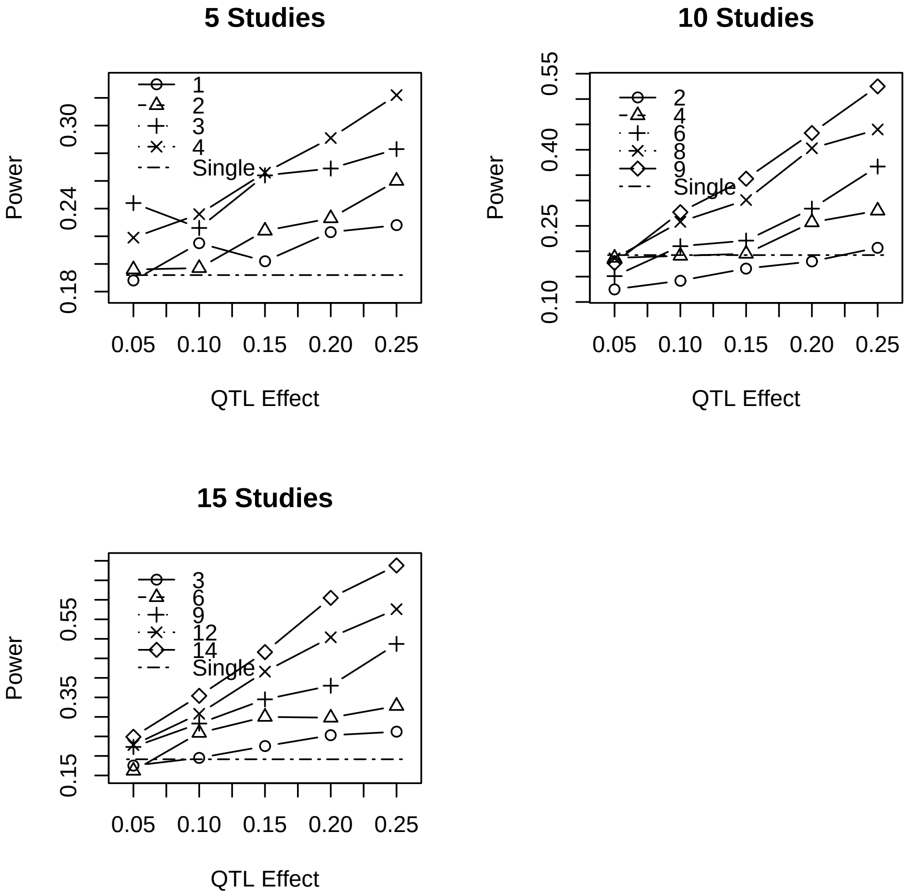


Figure 4.3: The power of the IM estimator and IM-EB estimator with 5, 10, and 15 studies at 65cM from the p-terminus of the chromosome. The QTL in all studies was positioned 65cM from the p-terminus of the chromosome. In the primary study, the heritability of QTL was set to 15% and the number of background study having non-zero QTL effect varied.

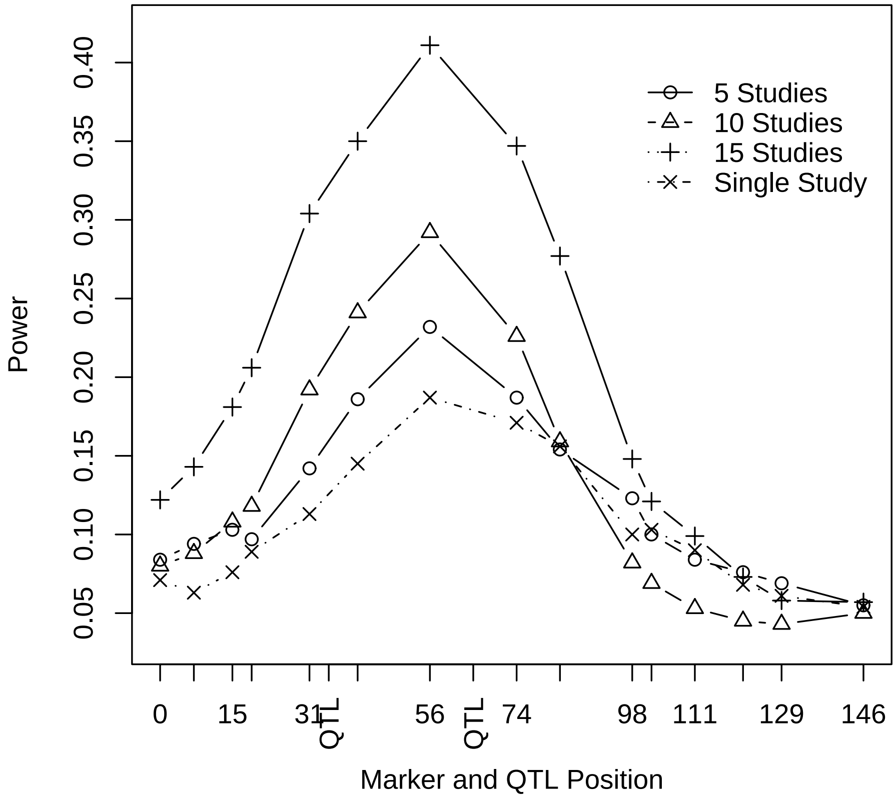


Figure 4.4: The power of the IM estimator and IM-EB estimator with 5, 10, and 15 studies at the marker loci. In all studies, including the primary study and background studies, the heritability was set to 15%. In the primary study and half of the background studies, the QTL was positioned 65cM from the p-terminus of the chromosome. In another half of the background studies, the QTL was positioned 35cM from the p-terminus of the chromosome.

Table 4.1: The mean and MSE (in parentheses) for the point estimates of QTL location.

Number of QTLs in Background Studies	Number of Studies	Number of Background Studies with non-zero QTL effect	Method	The Heritability in Background Studies		
				5%	15%	25%
One QTL at 65cM	5	2	IM	71.8(43.4)	69.2(43.6)	69.6(43.9)
	5	2	IM-EB	72.2(44.2)	70.6(42.1)	68.7(42.5)
	5	4	IM	65.4(41.7)	66.8(41.9)	69.5(41.4)
	5	4	IM-EB	64.9(40.6)	67.8(40.7)	69.3(39.0)
	10	4	IM	71.1(41.8)	68.2(42.2)	69.6(42.4)
	10	4	IM-EB	72.0(41.4)	68.1(40.2)	70.0(40.1)
	10	9	IM	70.8(43.5)	71.0(41.8)	69.9(42.2)
	10	9	IM-EB	68.9(41.9)	69.5(37.5)	67.2(34.9)
	15	6	IM	72.4(42.6)	70.4(42.6)	70.8(41.9)
	15	6	IM-EB	71.9(43.3)	68.4(40.4)	70.8(38.3)
	15	14	IM	69.8(41.2)	68.3(42.7)	71.0(43.2)
	15	14	IM-EB	69.9(39.3)	69.4(38.5)	68.7(33.2)
One QTL at 65cM One QTL 65cM at 35cM	5	2	IM	70.8(42.7)	70.8(42.6)	66.0(42.2)
	5	2	IM-EB	69.4(42.3)	70.2(42.4)	64.8(41.9)
	5	4	IM	67.8(41.5)	70.0(42.9)	70.641.9)
	5	4	IM-EB	66.3(41.1)	68.6(41.2)	65.8(40.2)
	10	4	IM	72.0(42.2)	69.1(42.5)	72.0(43.0)
	10	4	IM-EB	71.6(41.7)	68.4(42.9)	67.7(41.2)
	10	9	IM	70.2(42.6)	68.7(42.5)	68.2(42.4)
	10	9	IM-EB	65.8(42.0)	65.2(38.3)	62.7(36.8)
	15	6	IM	69.9(41.3)	69.7(41.9)	69.8(42.5)
	15	6	IM-EB	70.6(41.8)	66.0(40.7)	64.8(39.5)
	15	14	IM	69.8(43.1)	70.7(42.7)	69.7(41.4)
	15	14	IM-EB	67.4(41.5)	63.6(37.9)	60.3(34.2)

Table 4.2: The mean and MSE (in parentheses) for the point estimates of QTL effect.

Number of QTLs in Background Studies	Number of Studies	Number of Background Studies with non-zero QTL effect	Method	The Heritability in Background Studies		
				5%	15%	25%
One QTL at 65cM	5	2	IM	0.27(0.18)	0.27(0.18)	0.26(0.18)
	5	2	IM-EB	0.22(0.14)	0.23(0.14)	0.23(0.14)
	5	4	IM	0.27(0.18)	0.27(0.17)	0.26(0.18)
	5	4	IM-EB	0.22(0.14)	0.22(0.14)	0.23(0.14)
	10	4	IM	0.27(0.19)	0.27(0.18)	0.26(0.18)
	10	4	IM-EB	0.17(0.10)	0.17(0.10)	0.18(0.10)
	10	9	IM	0.27(0.18)	0.27(0.18)	0.27(0.18)
	10	9	IM-EB	0.17(0.10)	0.19(0.10)	0.19(0.10)
	15	6	IM	0.27(0.18)	0.27(0.18)	0.27(0.18)
	15	6	IM-EB	0.14(0.08)	0.15(0.09)	0.16(0.09)
	15	14	IM	0.27(0.18)	0.27(0.18)	0.27(0.18)
	15	14	IM-EB	0.15(0.08)	0.16(0.09)	0.17(0.08)
One QTL at 65cM One QTL 65cM at 35cM	5	2	IM	0.26(0.17)	0.26(0.17)	0.27(0.18)
	5	2	IM-EB	0.21(0.13)	0.22(0.13)	0.22(0.14)
	5	4	IM	0.27(0.18)	0.27(0.18)	0.26(0.17)
	5	4	IM-EB	0.22(0.14)	0.22(0.14)	0.23(0.14)
	10	4	IM	0.27(0.18)	0.27(0.17)	0.27(0.18)
	10	4	IM-EB	0.16(0.09)	0.17(0.09)	0.17(0.10)
	10	9	IM	0.27(0.18)	0.27(0.18)	0.27(0.18)
	10	9	IM-EB	0.17(0.09)	0.18(0.10)	0.19(0.10)
	15	6	IM	0.27(0.18)	0.27(0.18)	0.27(0.18)
	15	6	IM-EB	0.14(0.08)	0.15(0.08)	0.15(0.08)
	15	14	IM	0.27(0.17)	0.27(0.18)	0.28(0.19)
	15	14	IM-EB	0.14(0.07)	0.16(0.08)	0.17(0.08)

CHAPTER 5

Meta-analysis methods for genome-wide linkage studies

Cathryn M. Lewis
Department of Medical and Molecular Genetics
Guy's, King's and St. Thomas' School of Medicine
King's College London, UK

5.1 Introduction

Genome-wide linkage studies have been extensively used to identify chromosomal regions which may harbour susceptibility genes for complex diseases. The early enthusiasm for such studies has been replaced by the realisation that most complex disease genes have only a minor effect on risk, and consequently many linkage studies have low power to detect such genes (Risch and Merikangas, 1996). This was well illustrated by a compilation of 101 genome-wide linkage studies in 31 diseases, which found that few studies achieved significant evidence for linkage, and there was little replication within each disease (Altmuller et al., 2001). Replication of linkage is an important concept in genome-wide linkage studies: two studies obtaining high (if not significant) LOD scores in the same approximate region lends further weight to these results. This *ad hoc* method of comparing results across studies is formalised in meta-analysis, which provides statistical evidence for the co-localisation of linkage evidence across studies. Meta-analysis can also provide a solution to the lack of power in individual studies: combining weak evidence of linkage from several studies may show an overall significant effect.

Several methods for meta-analysis of linkage studies have been proposed. The gold standard is a complete analysis of genotype data from all contributing studies (often termed 'mega-analysis'). However, many study groups are reluctant to share raw genotype data, particularly if they are restricted by industrial partnerships. There are also technical problems of pooling different marker maps, and difficulties in finding an analysis method that is suitable for all studies. Pooling genotypes in short candidate regions has worked well in many collaborative studies (Demenais et al., 2003; Levinson et al., 2002).

5.2 Statistical methods for meta-analysis of linkage studies

The meta-analysis methods used in epidemiological studies are difficult to apply directly to genetic linkage studies. Methods that pool effect sizes (*e.g.* odds ratios) across studies are inappropriate as linkage studies frequently report results as a test statistic or p-value. In addition, we wish to assess linkage evidence across a region, not at a single location. Novel meta-analysis methods have therefore been developed to take account of the unique design and analysis strategies used in genetic studies.

For a meta-analysis of p-values at a single point, Fisher's method for pooling p-values can be used, provided LOD score values of zero are treated correctly (Province, 2001). However, unless testing for linkage at a strong candidate gene, specifying a single location for the analysis may not be optimal. Simulation studies show that maximum LOD scores have poor localisation, and can arise up to 30cM from a susceptibility gene (Cordell, 2001). Assessing evidence across a region therefore improves the power to detect linkage in a meta-analysis; this strategy is implemented in the Multiple Scan Probability (MSP) method (Badner and Gershon, 2002b). This method extends Fisher's p-value method, using the minimum p-values attained in a region, with a correction to the p-value for the total region length included in the analysis (see below for further details). The meta-analysis of identity-by-descent (IBD) sharing in affected sib pairs has been proposed for both discrete and quantitative traits (Gu et al., 2001) (***see also chapters in this book). Performing meta-analysis on this parameter of effect size is methodologically appealing. However, the IBD sharing statistic is rarely reported in publications, and some methods rely on identical markers being genotyped in each study, which severely restricts their application.

5.3 Genome Search Meta-Analysis method

The Genome Search Meta-Analysis (GSMA) method (Wise et al., 1999) was developed to circumvent some common problems of performing meta-analysis on genome-wide linkage studies. The GSMA is a non-parametric method, with few restrictions or assumptions, so that any genome-wide linkage search can be included, regardless of study design or statistical analysis method.

*** RG: Add intro comment on *types* of studies leading to the lod scores or p-values for the GSMA. In general, can one have *any* test stat?

***phone: find part where she mentions this and move/add to here

*** RG: Add a comment regarding association studies: (a) does GSMA work for these? (b) can/should assoc. studies be included in a MA with linkage studies? (discuss)

***CL: mention that this applies to linkage and it's not obvious how would extend to assoc (since testing assoc with particular alleles, which might not be common across studies)

In the GSMA, the entire genome is divided into bins of approximately equal width (measured in cM). We conventionally use 120 bins of 30cM length, so that for chromosome 1, the region between 0 and 30cM is assigned to bin 1.1, between 30-60cM to bin 1.2, *etc.*.

***RG: (a) include sex chromosomes? ***DG fix: (b) redo labeling so that it's clearer (c) what to do when the chromosome doesn't partition into 30 cM regions? (d) no overlap across chromosomes?

Let the number of bins be n , and the number of studies be m. For each study, the maximum LOD score (or minimum p-value) within each bin is identified, and the bins are ranked, with the most significant result achieving a rank of n, the next highest result a rank of $n-1$, *etc.*. Across studies, the ranks for each bin are summed; the summed rank forms the test statistic for this bin. A high summed rank implies that the bin has high LOD scores within individual studies, and may contain a susceptibility locus. Under the null hypothesis of no linkage, the summed rank for a bin will be the sum of m ranks, randomly chosen from $1, 2, \ldots, n$ with replacement. Significance levels for each bin can be determined from the distribution function of summed ranks (Wise et al., 1999) or by simulation.

***CL: Is there a preference? On what parameters does the sampling distribution depend?

Under no linkage, the probability of attaining a summed rank r in a specific bin, from m studies and n bins is:

***DG: check formula

$$P(\sum_{i=1}^{m} X_i = r) \quad = \quad \begin{cases} 0 & \text{for} \quad r < m \\ \frac{1}{n^m} \sum_{k=0}^{d} (-1)^k \binom{m}{k} \binom{r-kn-1}{m-1} & \text{for} \quad m \leq r \leq mn \\ 0 & \text{for} \quad R > m, \end{cases}$$

where X_i = rank of study i and d = integer part of $(r-m)/n$ (Wise et al., 1999). Hence the probability of obtaining a summed rank of r or greater (*i.e.* the p-value) in a bin can be calculated. This bin-wise p-value, p_{SR}, can also be obtained by simulation, permuting the bin-location of the assigned ranks.

***DG: 'bin-location of the assigned ranks' - not quite right wording permute ranks across bin location labels

For each study, the ranks within a study are randomly re-assigned to bins, and then across studies the summed rank calculated for each bin. For d permutation replicates, dn summed rank values are obtained, and the p-value for an observed summed rank r_{obs} associated with a given bin is calculated from r_{sim}, the number of simulated bins with summed rank greater than or equal to the observed summed rank . The p-value is then $p_{SR} = (r_{sim}+1)/(dn+1)$, where n is the number of simulated bins (North et al., 2003). Calculating critical values by simulation is particularly appropriate when the assigned ranks depart from the integer values $1, 2, \ldots, n$ assumed in the distribution function above, as happens through tied ranks or missing values (see Table 5.1).

Table 5.1: Common sources of incomplete data in the GSMA, and possible solutions

Missing data problem	Possible solutions
Many bins with a maximum LOD score of zero	Use tied ranks, so 20 bins with a maximum LOD score of zero would be assigned ranks 10.5.
Bins with no genotyped markers or no linkage data	Assign the median rank (*i.e.* $(n+1)/2$ for n bins), or assign a rank which is the weighted average of flanking bins (since multipoint LOD scores are correlated in adjacent bins).
Results are only reported from regions with the strongest evidence for linkage	Contact study authors for full information, and carry out the study collaboratively. Alternatively, if the observed results fall into b bins, assign these ranks $n, n-1, n-2, \ldots, n-(b+1)$, and assign all remaining bins the average remaining rank. For many missing bins, or bins missing in several studies, this method is not advisable, as the distribution function no longer provides a good fit.
Different chromosomes have been included (*e.g.* some studies have not tested the X chromosome)	Analyse all relevant subsets of studies to obtain maximum information, and for each bin/region, report results from the analysis with most complete data. If chromosome X is missing for r studies (out of m), analyse the remaining $m-r$ studies for the whole genome, and report these results from this analysis for chromosome X. Autosomes can then be analysed will all studies.
Two-stage genome wide study, with some regions genotyped on additional families	Use only the first stage analyses: the distribution of the maximum LOD score per bin depends on the number of families included, and a consistent study design should be used across the genome.
High-density genotyping in previously identified candidate regions	Obtain original LOD scores from markers used in the genome search. The maximum evidence for linkage within a bin increases with denser genotyping, thus inflating the evidence for linkage in more densely-genotyped bins.

The GSMA was developed to encompass diverse study designs and analysis methods. The linkage evidence may be extracted from any analysis method: for example, multipoint LOD scores calculated at each 1 cM, LOD scores calculated at each marker genotyped with the bin, or parametric LOD scores calculated at a series of recombination fractions for each marker. For parametric LOD scores, linkage is often tested using a series of models with different modes of inheritance or different penetrance/frequency parameters. The evidence for linkage can be assessed across all models analysed, provided the underlying distribution of LOD scores is approximately equal in each model; this can be determined from the distribution of LOD scores across the genome. Thus, the maximum evidence for linkage within a bin would be the highest LOD score calculated, regardless of the model under which it was obtained.

The bin-wise summed rank p-value p_{SR} assesses the information in multiple binsand should therefore be corrected for multiple testing. With 120 bins, under no linkage, 6 bins would be expected to attain $p_{SR} < 0.05$, and 1.2 bins to attain $p_{SR} < 0.01$. Following Lander and Kruglyak (Lander and Kruglyak, 1995), we define genome-wide evidence for linkage as that expected to occur by chance once in 20 GSMA studies, and suggestive evidence for linkage as that expected to occur once in a single GSMA study (Levinson et al., 2003). Using a Bonferroni correction on 120 bins gives p = 0.00042 (= 0.05/120) for genome-wide significance within a study, and p = 0.0083 (= 1/120) for suggestive evidence of linkage.

***RG: Doesn't seem right; genomewide: 1 in 20 studies, suggestive: 1 in a single study

***DG: tighten up or ask CL

For a genome-wide assessment of linkage, an ordered rank (OR) p-value (p_{OR}) may be used (Levinson et al., 2003).

***RG: Give some interpretation of ordered p-values? ***DG: tighten up or ask CL

This uses simulations of the complete GSMA to compare the summed rank of the observed k^{th} highest bin with the simulated distribution of summed ranks of the k^{th} highest bin, *i.e.* compares the 'place' of the bins in the full listing of results. Therefore, in a simulation of 5000 complete GSMAs, the bin with the highest observed summed rank is compared to all 5000 bins with highest summed rank, and the ordered rank p-value p_{OR} calculated. Similarly, the summed rank of the bin in the k^{th} place is compared to summed ranks of all bins lying in k^{th} place. This test can identify evidence for many bins with increased evidence for linkage, although the evidence for linkage within each bin may be modest. In the study of 20 genome wide searches for schizophrenia, 12 bins in the weighted analysis had significant summed rank and significant ordered ranks ($p_{SR} < 0.05$, $p_{OR} < 0.05$). Our simulations based on these studies showed that this combination of significant results was not consistent with occurring by chance (not observed in 1000 GSMA simulations of an unlinked study). The combination of a significant p_{SR} and p$_{OR}$ is therefore highly predictive of a linkage within a bin, however empiric criteria for linkage for an arbitrary number of studies have not yet been developed (Levinson et al., 2003).

***RG: Is there a recommendation for multiple testing correction of ordered p-values? ***DG: goes along with above

In assessing linkage we recommend the following hierarchy for interpreting results:

***DG: how is this a hierarchy (confusing within/between studies)

1. A genome-wide significant summed rank p-value ($p_{SR} <$ 0.05/#bins),
2. Nominal evidence for linkage in both statistics ($p_{SR} < 0.05$, p$_{OR} < 0.05$)
3. Nominal evidence for linkage in the summed rank ($p_{SR} < 0.05$).

No evidence for linkage should be declared where bins do not have a significant summed rank p-value. Within bins with a significant summed rank, a significant ordered rank p-value can be considered to enhance the evidence for linkage. Clearly, if the k^{th} bin has nominal evidence for linkage under both statistics, then any bin with higher summed rank must also be considered significant. By plotting the observed summed ranks by size, with the distribution of ordered ranks, a ‘scree slope’ may be seen where the summed ranks decrease rapidly and the ordered ranks become non-significant (see Figure 2, in the inflammatory bowel disease GSMA (van Heel et al., 2004)).

***DG: see if can get permission to include this plot

***DG: ask CL if can get example plots of ind. studies and meta-an p-value of summed ranks, a figure of the method (like fig 1 in vanHeel, not like fig 1 here, which addresses the robustness of the method)

In regions where the $p_{SR} > 0.05$ but $p_{OR} < 0.05$, one interpretation is that the power to identify linkage in these bins is low, and a larger meta-analysis might increase significance of p_{SR}, whilst retaining the significance of the ordered rank statistic.

5.4 Collaborative or published information?

Two main approaches are used to carry out a GSMA analysis. Firstly, the GSMA may be based on published information, for example extracting linkage statistics (NPL/MLS scores, p-values, *etc.*) from graphs and tables. In some cases, investigators may have posted detailed genome-wide results or original genotype data on a website. In papers, genome-wide studies are frequently displayed as line graphs of linkage statistics along each chromosome. This may be used in the GSMA by dividing each chromosome into the required number of equal length bins, and reading off the maximum statistic attained in each bin. Inaccuracies in the method arise from different marker maps used in each study, or different chromosome lengths (so that bins will not be exactly compatible across studies). If marker names are given, bins may be designated more accurately by mapping the bin boundary markers relative to the genotyped markers. In some studies, tables of linkage statistics attained at each marker genotyped are given. These markers may be placed into relevant bins, and the maximum linkage statistic for each bin identified. Common problems arising from the use of published data are listed in Table 5.1, with possible solutions.

A more satisfactory method of performing a meta-analysis study is to form a collaboration of relevant research groups, and use computer files of LOD scores (*e.g.* output files generated from Genehunter, Allegro, *etc.*). This gives full information on the location and magnitude of linkage statistic, and should improve the accuracy of the resulting study. However, if some researchers do not wish to participate, the organisers must then choose between an incomplete meta-analysis of high quality data and a complete meta-analysis of lower quality data. In practice, meta-analyses of genetic studies have been widely supported by researchers (*e.g.* schizophrenia (Lewis et al., 2003), bipolar disorder (Segurado et al., 2003), and inflammatory bowel disease (van Heel et al., 2004)).

In any meta-analysis, the investigators rely on the high quality of results generated by the original studies. Any errors due to genotyping problems, inaccurate phenotype definition, incorrect pedigree reconstruction, or poor analysis methods will be carried through to the meta-analysis, and will reduce power to detect evidence for linkage. Errors seem likely to be random in each study, and should therefore not introduce a bias to the meta-analysis results.

5.5 Summed ranks or average ranks?

The GSMA was originally formulated using summed ranks, where the highest rank n is assigned to the bin with the strongest evidence for linkage. This follows the statistical convention that high test statistics (*i.e.* summed rank) show more evidence against the null hypothesis. An alternative, more intuitive, approach is to assign rank 1 to the ‘best’, most significant bin, and then use the average rank as a test statistic so that low average ranks give stronger evidence for linkage (Levinson et al., 2003). Statistically these approaches are equivalent, and a summed rank of R from n bins and m studies can be converted to an average rank as $(n + 1) - R/m$.

5.6 Bin width

The GSMA is heavily dependent on the chosen bin width. Our original description of the GSMA listed 120 bins, defined by specific boundary markers (see table at `http://www.kcl.ac.uk/depsta/memoge/gsma/` for full marker-bin information). The exact bin width depends on both chromosome length (to give equal width bins on each chromosome) and marker location. Other studies have chosen different bin widths (see Table 5.2). Although narrow bins may intuitively provide more information (see Figure 5.1), localisation through linkage information is broad. Adjacent bins may show evidence for linkage (see, for example, rheumatoid arthritis (Fisher et al., 2003), inflammatory bowel disease (van Heel et al., 2004) GSMA studies) and simulation studies have shown that the strongest information for linkage may arise in the bin flanking the true location (Levinson et al., 2003). In a study of age-related macular degeneration (Fisher et al., 2005), the original 120 bins (of 30cM length) were then bisected, and ranks (for 240 bins) re-assigned to determine whether more bins would improve localisation information or identify novel loci. The results were disappointing, with similar evidence for linkage spreading across several 15cM-width bins, and no novel regions were identified. The relative advantages of narrow or wider bins are listed in Table 5.3.

5.7 Weighted analysis

The original formulation of the GSMA assumed that all studies contributed equally.

However, a study of 500 affected sibling pairs (ASPs) has higher power to detect a true locus than a study of 100 ASPs. This aspect can be reflected in the meta-analysis by weighting the studies by sample size. The function sqrt(#genotyped affected individuals) has been used in many studies (see Table 5.2) and increased the power to detect linkage by approximately 7% compared to unweighted analyses in a simulation study based broadly on studies in the schizophrenia GSMA (Levinson et al., 2003). The optimal weighting function is unclear, particularly when some studies have used extended pedigrees and others have used ASPs. The power to detect linkage will depend on the locus effects (mutation frequency, penetrance), and for some loci, extended pedigrees may have higher power to detect linkage while affected sib pairs may be the optimal sampling unit for other genes. Defining a single weighting parameter is therefore somewhat unsatisfactory.

The chosen weighting function can be standardised by its average value for all studies, so that the mean weight is 1. Using a narrow range of weights (*e.g.* 0.9 – 1.1) will give an analysis that is very close to the unweighted analysis. However, using one study with a very high weight (*e.g.* four studies with weights 3.0, 0.4, 0.3, 0.3) will give results close to those obtained in this single study. Both these situations should be avoided, and alternative weighting functions may need to be tested.

5.8 GSMA software

Software to perform GSMA on genome-wide linkage studies is available from `http://www.kcl.ac.uk/depsta/memoge/gsma/` (Pardi et al., 2005). This program is written in C++ and available on Windows, Mac, and Unix/Linux platforms. The data input is a table of maximum linkage statistics for each bin, for each study. The program allows for an arbitrary number of bins and studies. Missing values are permitted, and bins replaced with the median linkage statistic for that study. For studies reporting p-values, the entry values should be $1 - p$-value to ensure correct ranking of results. The program calculates the summed rank, then determines the summed rank and ordered rank p-values (p_{SR}, p_{OR}) by simulation. The user may determine the number of simulations, and the program is rapid, completing 10,000

Table 5.2: Summary of published GSMA studies (*geno*: genotyped individuals; *aff*: affecteds; *arp*: affected relative pairs; *asp*: affected sib pairs; Significance – Nom: nominal; Sugg: suggestive; Gen: genome-wide)

						# bins with SR	$p_{SR} < 0.05$,
Disease	Publication	# studies	# families	# bins	Weights	Nom./Sugg./Gen.	$p_{OR} < 0.05$
Multiple sclerosis	Wise, 1999	4	257	120	–	8/2/1	–
Type 2 diabetes	*Demanais, 2003	4	1127	120	–	6/1/0	–
Schizophrenia	*Lewis, 2003	20	1208	120	$\sqrt(\#aff)$	12/4/1	12
Bipolar disorder[a]	*Segurado, 2003	18	370	120	$\sqrt(\#aff)$	9/2/0	2
Coeliac disease	*Babron, 2003	4	442[b]	115	#ped	5/5/2	–
Rheumatoid arthritis	Fisher, 2003	4	570	120	#asp	10/3/1	–
Coronary heart disease	Chiodini, 2003	4	807	124	$\sqrt(\#asp)$	4/3/1	–
Inflammatory bowel disease	Williams, 2003	5	709	117	–	8/4/1	–
Crohn's disease	Williams, 2003	5	472	117	–	9/4/0	–
Inflammatory bowel disease	*van Heel, 2004	10	1253	105	$\sqrt(\#arp)$	8/5/1	6
Crohn's disease	*van Heel, 2004	10	711	105	$\sqrt(\#arp)$	10/5/0	8
Ulcerative colitis	*van Heel, 2004	7	314	195	$\sqrt(\#arp)$	5/1/0	0
Hypertension/blood pressure	*Koivukoski, 2004	9	1992	120	$\sqrt(\#aff)$	9/3/1	2
Psoriasis	†Sagoo, 2004	6	493	110	–	5/2/2	–
Cleft Lip/Palate	†Marazita, 2004	13	574	120	$\sqrt(\#geno)$	12/3/1	12[c]
Body mass index	*Johnson, 2005	5	505	121	$\sqrt(\#geno)$	–/1/0	–
Age-related macular degeneration	*Fisher, 2005	6	908	120	$\sqrt(\#aff)$	15/2/1	11

* = collaborative study; † = partially collaborative; [a] very narrow phenotype definition; [b] based on fine-scale mapping; [c] maximum number, including candidate region follow-up

Table 5.3: Comparison of properties affecting choice of bin width

Property	Narrower bins (*e.g.* 120 x 30cM bins)	Wider bins (*e.g.* 60 x 60cM bins)
Bin width	Little variability	Unequal bin widths for different length chromosomes
Correlation in ranks in adjacent bins	Highly correlated, particularly for multipoint linkage analysis. May violate distributional assumptions for test statistic.	Low correlation
Localisation	Reasonable, although adjacent bins may be significant	Poor
Power to detect linkage	High, except where maximum LOD scores occur in different bins	Lower, except where wider bins substantially increases the study rank in linked regions
Consistency of bin definition across studies	Poor, especially based on published information	More overlap between bins in adjacent studies, even when poorly defined

simulations in under 3 seconds on a desktop PC. Weighted and unweighted analysis is performed, using user-defined weights. Three results files are output: (a) results for the most significant bins only, (b) a full genome listing of bin, summed rank, p_{SR}, p_{OR} (weighted and unweighted analyses), and (c) ranks assigned to each study, for data checking.

5.9 Power to detect linkage using the GSMA

***phone: clarify power here, is there some parameter? Define what power means here. No effect size anywhere in the power discussion. What is the genetic effect you are trying to detect?

An extensive simulation study of the GSMA was carried out by Levinson et al. (2003) based on genome scans contributed to the meta-analyses of schizophrenia (Lewis et al., 2003) and bipolar disorder (Segurado et al., 2003). For the simulation, a number of sib pairs with broadly equivalent information to the pedigrees from the original studies were used, with 1625 ASPs for schizophenia, 1017 ASPs for bipolar disorder (narrow phenotype definition), and 501 ASPs for bipolar disorder (very narrow phenotype definition). These three studies therefore give a wide range of study sizes covering those seen in many GSMA studies (Table 5.2).

The schizophrenia study had high power to detect linkage with a locus conferring a sibling relative risk (λ_s) of 1.3 at a significance level of $p < 0.01$.

***RG: ‘detect linkage’ - bin containing the disease gene?

For a significance level of 0.05, a power of at least 70% was attained in the following situations:

- 1625 ASPs (schizophrenia), for a locus with λ_s = 1.15,
- 1017 ASPs (bipolar disorder, narrow phenotype) for a locus with λ_s = 1.3,
- 501 ASPs (bipolar disorder, very narrow phenotype) for a locus with λ_s = 1.4.

Full details of other assumptions required in the simulation, including the number of genotyped parents, marker density, and number of loci simulated are given in the original paper (Levinson et al., 2003).

***RG: (below): ‘power’ seems ill-defined, or at least something is unclear.

The power of a study to detect linkage depends on the number of studies m and the number of bins n, in addition to the genetic effect size in each study. The average rank threshold for declaring genome-wide, suggestive or nominal linkage changes with the number of studies $(m = 4, 7, 10, 15, 20)$ and the number of bins $(n = 60, 120)$, as shown in Figure 5.1. Note that the thresholds for genome-wide (p_{GW}) and suggestive (p_{SUG}) linkage depend on the number of bins used: p$_{GW}$ = 0.00042 and p_{SUG} = 0.0083 for 120 bins, and p_{GW} = 0.00056 and p_{SUG} = 0.017 for 60 bins; nominal evidence for linkage was fixed at p = 0.05 throughout.

***RG: where do the thresholds come from? Fig 1? What reported ranks?

With 120 bins, an average rank threshold for nominal linkage is 32 for 4 studies, but over 48 for 20 studies – so the average rank is not even within the top third of reported ranks.

***RG: meaning between 1 and 40?

An average rank of 32 gives nominal evidence for linkage with 4 studies, but provides genome-wide evidence for linkage with 20 studies. For a given study size, relative to 120 bins an analysis with 60 bins requires smaller average ranks for linkage (Figure 5.1). Thus, the evidence must be stronger by pooling smaller correlated bins into wider ones. Provided the maximum LOD scores for a locus localise to a narrow region, using narrow bins provides the most evidence for linkage: with 10 studies, an average rank of $\approx$ 20 gives genome-wide evidence for linkage if this is obtained using 120 bins, but only nominal significance with 60 bins.

***RG: The setting does not take account of the assumption that the locus is narrowly defined.

Reducing the number of bins could, however, increase the power to detect linkage if the LOD scores’ peaks are too widely spread to be contained in a single bin (for example if the locus lies close to a bin boundary), so that the average ranks decrease using fewer bins.

***RG: Does the figure correspond to a simulation? (Details of simulation given by Levinson et al).

One critical issue is the loss of information arising when the GSMA divides the genome into discrete bins. ***Two simulation studies have compared the power of the GSMA to the power of ‘mega-analysis’, based on pooling the raw genotype data

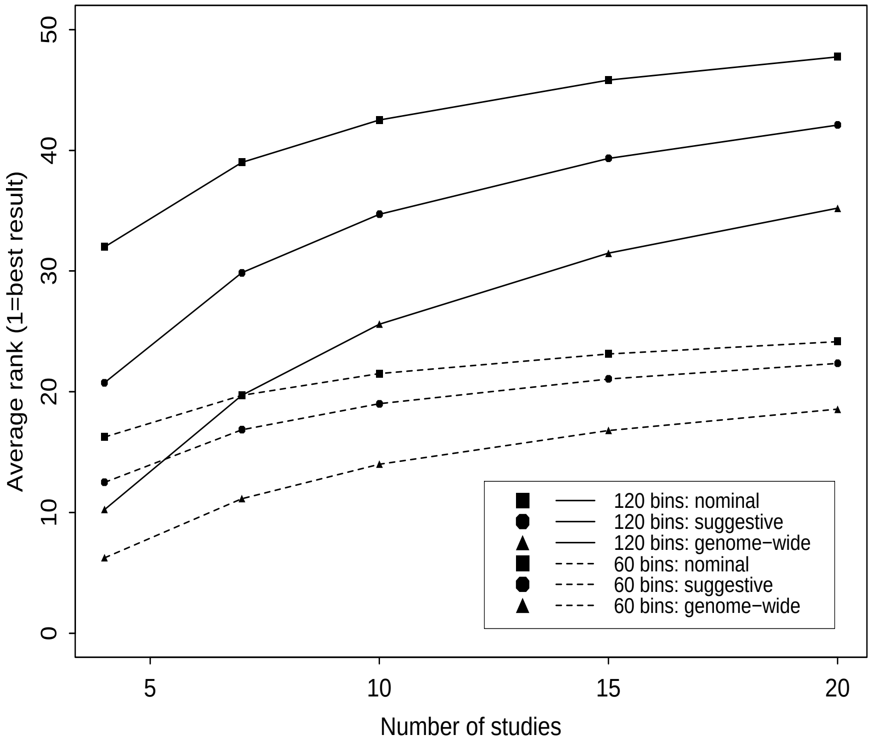


Figure 5.1: Critical values of the average rank required for genome-wide, suggestive, and nominal evidence for linkage, by number of bins.

from each study. Demple and Loesgren (Dempfle and Loesgen, 2004) showed that the power of the GSMA was less than the mega-analysis approaches tested, but they applied the Lander and Kruglyak criteria for genome-wide significance, which is much more stringent than using a Bonferroni multiple testing correction (0.05/#bins). Using this appropriate, less stringent, correction, Levinson et al. (2003) showed that the power of the GSMA to detect linkage was actually higher than for the analysis of pooled genotypes.

*** RG/DG: !!! This result seems surprising and possibly counter-intuitive and requires additional comment.

*** RG: Also see Guerra and Goldstein papers

5.10 Extensions of the GSMA

Many different diseases have been studied using the GSMA, but little further methodological development has been carried out. Some authors have proposed minor enhancements to the method. For example in their study of celiac disease, Babron et al. (2003) used a summed rank function that was a weighted average of the ranks of a bin and two flanking bins. This extends the potential area in which evidence for linkage can be shown, since high linkage statistics in a flanking bin will be included. However, it will also increase the correlation between summed ranks in adjacent bins. An alternative approach to the problem of maximum LOD scores being attained in adjacent bins in different studies is ‘pooled bins’ used in the rheumatoid arthritis study (Fisher et al., 2003). Here, adjacent bins are pooled, and the original analysis of n bins is reanalysed as two analyses of $n/2$ bins each, where bins 1+2, 3+4, $\ldots$ are pooled in the first analysis, and 2+3, 4+5 $\ldots$ are pooled in the second analysis. This analysis would be valuable where a true locus lies close to a bin boundary, and the bin-location of maximum linkage evidence is inconsistent across studies. However, as Figure 5.1 shows, reducing the total number of bins reduces the power to detect linkage.

***RG: Has argued both ways: increasing power with increasing number of bins, increasing power with decreasing number of bins.

In their study of cleft lip/palate, Marazita et al. (2004) use a series of overlapping bins from 0-30cM, then 10-40cM, 20-50cM, *etc.* and assess the maximum evidence for linkage across each possible bin. This should give better localisation information, and may determine whether two linkage peaks exist in one region. However, there are unresolved problems of multiple testing.

Recently, Zintzaras and Ioannidis (2005b) provided a major extension to the GSMA in developing methods to test for heterogeneity of linkage evidence within a bin. Heterogeneity testing is a standard component of meta-analysis in epidemiological studies, where researchers test for evidence of different effect sizes across studies, but has not previously been implemented in the GSMA. They apply these methods directly to the rank statistics of each study, introducing three highly correlated heterogeneity statistics. The significance of each statistic is assessed by simulation, randomly reassigning the ranks to bins within each study, and recalculating each heterogeneity statistic. The proportion of simulated bins with Q-statistics above the observed value (for high heterogeneity), or below the observed value (for low heterogeneity) is then tabulated for a p-value. Zintzaras and Ioannidis (2005b) applied the methods to published ranks in GSMA studies of rheumatoid arthritis (Fisher et al., 2003) and schizophrenia (Lewis et al., 2003). They identify several bins in each study that show evidence for high heterogeneity (different evidence for linkage across studies) or low heterogeneity (consistent linkage evidence). The authors acknowledge that the distribution of the heterogeneity statistics may depend on the summed rank statistic attained within the bin. They therefore test for heterogeneity under two scenarios: where the observed heterogeneity statistic is compared to all simulated bins, and where the observed heterogeneity statistic is only compared to simulated bins with similar summed rank values (±2).

5.11 Limitations of the GSMA

Three classic sources of error in meta-analysis studies are listed below and discussed with their relevance to the GSMA.

5.11.1 File drawer problem

This error arises when unpublished studies are not included in the meta-analysis, as their existence is unknown to the investigators. For linkage studies of candidate regions, a publication bias exists as negative studies are less likely to be published, which will bias the results of the meta-analysis. For genome-wide studies this is not a major concern: these studies are large, expensive to perform, and publishable, regardless of the significance of LOD scores obtained. No single hypothesis is being tested, so publication bias is not relevant.

5.11.2 Garbage in, garbage out

Any meta-analysis is reliant on the quality of both the data and the results from the individual studies. We assume that each study has a high quality of phenotype and genotype data, and that standard quality control checks have been performed (*e.g.* testing for non-paternity, genotyping errors). The most challenging problem in the GSMA is ensuring a consistent bin definition, particularly where studies have used marker maps that differ in order or distance.

5.11.3 Apples and Oranges

Pooling data from many different studies is statistically appealing, but it is only of value if a common effect is occurring across the studies. There are several sources of heterogeneity that can limit the value of a meta-analysis of genetic linkage studies. Potential sources of heterogeneity are population, family sampling units (extended pedigrees or affected sibling pairs), and clinical characteristics (diagnostic criteria, age of diagnosis, severity of disease). Heterogeneity for evidence of linkage can be tested using the methods of Zintzaras and Ioannidis (2005b). A subset analysis can also be performed to analyse a more homogeneous set of studies. We have little understanding of how the distribution of genetic variants contributing to complex disease may be affected by these features, although the common disease, common variant (CDCV) hypothesis for complex diseases implies that a variant would be present across a wide range of study designs. Some GSMA studies have detected linkage to several genetic regions (schizophrenia, inflammatory bowel disease), suggesting that at least some common disease genes can be detected across diverse studies.

5.12 Disease studies using the GSMA

The GSMA has been applied in 14 studies of complex diseases, summarised in Table 5.2 (Demenais et al., 2003; Wise et al., 1999; van Heel et al., 2004; Lewis et al., 2003; Segurado et al., 2003; Fisher et al., 2003, 2005; Babron et al., 2003; Marazita et al., 2004; Chiodini and Lewis, 2003; Williams et al., 2002; Koivukoski et al., 2004; Sagoo et al., 2004; Johnson et al., 2005). Most studies have analysed qualitative diseases, but quantitative traits (hypertension, body mass index) have also been studied. The average number of linkage studies included per meta-analysis was 7.9 (range 4-20), and the average number of families was 736 (range 257-1992). (These figures omit the overlapping studies of inflammatory bowel disease, Crohn's disease and ulcerative colitis). Of 14 studies, 8 were full collaborations, while others relied at least partially on published information. All studies found at least one suggestive result (approximately $p < 0.01$), and in 12 studies, at least one result of genome-wide significance was found.

***CL: This p-value adjusted for multiple testing?

In the auto-immune diseases, genome-wide significance was found in the HLA region on chromosome 6 (multiple sclerosis (Wise et al., 1999), rheumatoid arthritis (Fisher et al., 2003), psoriasis (Sagoo et al., 2004), inflammatory bowel disease (van Heel et al., 2004)), confirming findings of the original linkage studies. In other studies, a region of genome-wide significance was observed on chromosome 2 for schizophrenia (Lewis et al., 2003), which had not previously been highlighted as a strong candidate region for schizophrenia (O'Donovan et al., 2003). Similarly, regions of genome-wide significance were detected on chromosome 4 for psoriasis (Sagoo et al., 2004), on chromosome 3 for coronary heart disease (Chiodini and Lewis, 2003), on chromosome 2 for cleft lip/palate (Marazita et al., 2004), on chromosome 3 for hypertension (Koivukoski et al., 2004) and on chromosome 10 for age-related macular degeneration (Fisher et al., 2005). No susceptibility genes have yet been localised in these regions for these diseases, but they provide strong candidate regions for follow-up linkage or association studies. Genome-wide significance is an extremely stringent criteria (occurring only once in 20 GSMAs by chance), and this is illustrated by the results for Crohn's disease in the region of CARD15 on chromosome 16. This region attained a p-value of 0.003 (weighted analysis) (van Heel et al., 2004), despite the presence of this confirmed susceptibility gene. Across the diseases, there was no correlation between the number of bins with nominal or suggestive significance and the number of studies included. Only five studies had used the Ordered Ranks test to assess clustering of linkage results, but the easy availability of this method in the GSMA software package (Pardi et al., 2005) should make this analysis more widely used.

These results show that the GSMA can play an important role in synthesizing data across genome-wide linkage studies and directing follow-up studies. The number of significant regions arising from GSMA studies has raised enthusiasm for the potential utility of linkage studies, these studies suggest that susceptibility genes for complex diseases are detectable using linkage studies, provided the sample sizes are large enough.

5.13 The Multiple Scan Probability method (MSP)

Badner and Gershon (2002b) developed a novel method of meta-analysis of linkage data, based on the maximum evidence for linkage obtained within a genetic region. This method is ‘region-wide’ rather than genome-wide, as the region for analysis can be specific by investigators, and is usually triggered by one low p-value within a study (*e.g.* $p < 0.01$). For each study, the strongest evidence for linkage within 30cM of the triggering-locus is noted, and the p-values combined, accounting for the length of the region of the final analysis and the genotyping density of original studies (see Badner and Gershon (2002b) for full details). A replication analysis excluding the original linkage finding is also recommended.

This method has been applied to autism (Badner and Gershon, 2002b), schizophrenia and bipolar disorder (Badner and Gershon, 2002a). In schizophrenia, significant evidence for linkage was detected on chromosome 8p, 13q and 22q. These regions on chromosome 8p and 22q were also detected in the GSMA study of schizophrenia (Lewis et al., 2003), but the 13q region was absent. Linkage to 13q and 22q were also found in bipolar disorder, neither of which was detected in the GSMA study (Segurado et al., 2003), however for both schizophrenia and bipolar disorder, the studies included in the GSMA and the MSP differed substantially.

The major contrast between the GSMA and the MSP methods is in the test statistic. The MSP uses a p-value, and therefore retains the magnitude of the significance of the original study. In contrast, the GSMA is a non-parametric rank method, and the maximum contribution from any study is the maximum number of bins (i.e. rank 120 in a study of 120 bins). The MSP should therefore have higher power to detect regions which have strong evidence for linkage in some studies, but with genetic heterogeneity present. Interestingly, the analysis of heterogeneity in the schizophrenia GSMA showed significant genetic heterogeneity on chromosome 13q, which may contribute to the different GSMA and MSP meta-analysis results in this region (Zintzaras and Ioannidis, 2005b). The MSP would have lower power to detect regions where linkage evidence is moderate in all studies, as this would not trigger the investigation of a region.

5.14 Conclusions

Millions of dollars have been spent on linkage studies of complex genetic disorders, but the results have been overwhelmingly disappointing. In hindsight, many of these studies are under-powered to detect linkage to genes that confer only a modest increase in risk for a complex disease. However, the utility of linkage studies has been demonstrated by the localisation of a few genes (*e.g.* CARD15 in inflammatory bowel disease, NRG1 in schizophrenia, CAPN10 in type 2 diabetes) following fine-mapping of regions detected in linkage analysis. Linkage studies still have an important role in localising disease genes: genotyping of many large cohorts is in progress, and linkage studies are still widely published. Meta-analysis of linkage studies is therefore a timely approach. It provides a rapid and cost-effective method to ensure that maximum information is extracted from the many linkage studies already performed. The regions highlighted in meta-analysis of linkage can be used to prioritise future gene localisation studies, whether these are based on fine-scale linkage, on association studies of candidate genes, or on follow-up of whole genome association studies.

5.15 Acknowledgements

I thank Douglas Levinson, Sheila Fisher and Lesley Wise for invaluable discussions on the GSMA and meta-analysis in genetics, Fabio Pardi for developing the GSMA software, and all researchers of genome-wide linkage studies for generously contributing their data to meta-analysis studies.

CHAPTER 6

Heterogeneity in Meta-Analysis of Quantitative Trait Linkage Studies

Hans C. van Houwelingen and Jérémie J. P. Lebrec
Leiden University Medical Center, The Netherlands

6.1 Introduction

In complex diseases where many genes might be involved in the genetic causation of the disease, individual loci influencing a quantitative trait are most likely to explain only a small proportion of its total variance. Consequently, there is a huge problem of lack of statistical power. Most linkage studies published to date only consist of a few hundred pedigrees with a limited number of individuals and, therefore, have little power to detect linkage of any but the ”largest” quantitative trait loci (QTL). In order to enhance power, it is now common practice to retrospectively pool evidence for linkage from several different studies. However, in pooling data from different studies, one should be aware of the possible heterogeneity between studies. The aim of this chapter is to present statistical models for describing this heterogeneity and approaches to analyze heterogeneous data

We distinguish two types of heterogeneity: locus and size heterogeneity. The populations used in each of the studies often have different genetic backgrounds and a locus affecting the trait of interest in one population might have no effect in another one; we will refer to this type of heterogeneity as *locus heterogeneity*. In other instances, the same locus may influence the trait in all populations, but there are many reasons to believe that the size of the effect will vary. For instance, the frequency of the causal allele may be much smaller in some populations or it may interact with other loci, or with environments and risk factors. We will refer to this type of heterogeneity as *size heterogeneity*.

Besides those biological sources of heterogeneity, some common logistic sources of variation often arise: typically, genotyping will have been carried out on different marker maps (and even when identical markers are used, their allele frequencies may vary across populations) and families may have been sampled according to different schemes. More simply, the phenotypes measured may vary in their method of collection from study to study.

When the raw data are available, one obvious way to gather evidence from several studies is to pool the data into a meta-file and proceed with an overall analysis. In the case of linkage studies with different marker maps, the data manipulations involved are very tedious. Moreover, the data sets become unnecessarily large because of the artificially created missing data on markers that are used in other studies. Furthermore, running standard methods of analysis on such large data files usually requires uncommon computing capacities. Therefore, we advocate the meta-analytic approach that collect all relevant summary information for each study and uses that as starting point for further analysis. Of course another simple reason for favoring meta-analysis is that researchers usually simply cannot access the raw data for each study and have to be content with individual test statistics along with (at best) parameter estimates.

We refer the reader to Dempfle and Loesguen (2003) and Rao and Province (2001) for recent overviews of meta-analytic methods for linkage studies. Most methods are in the spirit of the classical meta-analysis. An interesting, widely applicable, alternative are the rank-based methods such as the GSMA (Wise *et al.*, 1999). They might be sub-optimal compared to approaches based on the pooling of estimates of a common linkage parameter, but much more robust because of the built-in genomic control. Note that associated methods that assess heterogeneity have recently been developed (Zintzaras and Ioannidis, 2005). The idea of pooling different estimates of a common linkage effect across studies is not new although it has only been described for sib pair designs to date. Gu *et al.*(1998) use the excess identical-by-descent (IBD) sharing as a common effect, but their approach appears to be limited to studies with the same marker maps. Li and Rao (1996) and Etzel and Guerra (2002) both use the slope in a classical Haseman-Elston regression as a common effect, the former suffering the same restriction as Gu *et al.*(1998) regarding location of markers. Interestingly in the latter, the authors explicitly adjust for the (study-specific) marker to locus distance and allow for heterogeneity across studies by means of a random effect. Unfortunately, they do not seem to efficiently take into account the within-study dependence structure between markers.

Classical methods of meta-analysis originally introduced in the field of clinical trials (DerSimonian and Laird, 1986) can be adapted to linkage studies. The sufficient statistics used to perform such approaches are some measure of effect on a common grid of putative locations and its associated standard error. In the case of quantitative traits, a natural estimate of common linkage effect is the proportion of total variance explained by a putative location. We first describe the meta-analytic tools, assuming that QTL effect estimates and standard errors are available for all studies on a *common grid* of locations. In Section 7.2 the traditional meta-analytic approach in the context of linkage is reviewed, including how to test and allow for *size heterogeneity*, while in Section 6.2.4 we introduce a simple finite mixture model to account for potential *locus heterogeneity*. A complication that arises in both approaches for heterogeneous data is that variance components are nonnegative by definition. We will discuss the consequences of that for estimation and testing. In Section 6.3, we quickly review the methods which should be used for the analysis of individual studies in order to yield the relevant statistics required for meta-analysis as advocated in Sections 7.2. All methods are illustrated by means of four data sets used for a genome-wide scans for lipid levels in Section 6.4.

6.2 The classical meta-analytic method

Introductions to classical meta-analysis can be found in two Tutorials in Biostatistics in Statistics in Medicine, namely Normand (1999) and van Houwelingen *et al.*(2002). In this section, we recall briefly how meta-analysis is classically carried out and introduce some refinement that is specific to the variance component model used in linkage studies. We assume that at a given *common putative position*, each study (indexed by $i = 1, \ldots, K$) provides a consistent estimate $\hat{\gamma}_i$ of the true QTL effect γ_i of that locus and an associated standard error s_i. The link with the traditional lodscore is given by $\mathrm{LOD_i} = (\hat{\gamma}_\mathrm{i}{}^2/\mathrm{s_i^2})/(2 \times \ln(10))$. Details of the definition of the variance component and its estimation are given in Section 6.3.

6.2.1 *Analysis under homogeneity*

The simplest approach to meta-analysis assumes that the effects γ_i's are all equal to a common value γ so that $\hat{\gamma}_i \sim N(\gamma, s_i^2)$. This is known as the *homogeneity assumption* In this situation the corresponding maximum likelihood estimator of γ is given by the weighted average

$$\hat{\gamma}_{\mathrm{hom}} = \frac{\sum_i \hat{\gamma}_i / s_i^2}{\sum_i 1/s_i^2} \text{ with standard error } SE_{\mathrm{hom}} = 1/\sqrt{\sum_i 1/s_i^2}\,. \qquad (6.1)$$

The null hypothesis of no effect, that is $\gamma = 0$ versus the alternative $\gamma > 0$, can be tested by means of the one-sided statistic

$$\left(z^+_{\mathrm{hom}}\right)^2 = \begin{cases} \left(\hat{\gamma}_{\mathrm{hom}}/SE_{\mathrm{hom}}\right)^2, & \text{if} \hat{\gamma}_{\mathrm{hom}} > 0 \\ 0 & \text{if} \hat{\gamma}_{\mathrm{hom}} \le 0 \end{cases}$$

which follows the mixture distribution $\frac{1}{2}\chi^2_0 + \frac{1}{2}\chi^2_1$ under the null hypothesis, where χ^2_0 denotes the degenerate density with all mass in 0. The corresponding $\mathrm{LOD}_{\mathrm{hom}}$ score can be calculated as $\left(z^+_{\mathrm{hom}}\right)^2 / (2 \times ln(10))$. Observe that we do not truncate the estimated $\hat{\gamma}_i$ at zero, if negative, because that would complicate the pooling considerably. However, truncation is no problem in the final stage.

6.2.2 *Test for heterogeneity*

Even when the same locus is affecting a trait in different populations, it seems difficult to believe, for reasons given in Section 7.1, that the QTL effects are all equal. In the setting introduced earlier, this situation of *size heterogeneity* can be tested:

$$\begin{aligned}\mathrm{H}_0 &: \quad \gamma_1 = \gamma_2 = \cdots = \gamma_K \equiv \gamma_{\text{hom}} \\ \mathrm{H}_1 &: \quad \text{at least one } \gamma_i \text{ is different },\end{aligned}$$

the hypothesis of homogeneity H_0 can be tested using the following statistic

$$X^2 = \sum_{i=1}^{K} \frac{(\hat{\gamma}_i - \hat{\gamma}_{\text{hom}})^2}{s_i^2}$$

whose approximate null distribution is χ^2_{K-1}. In practice, any test for heterogeneity is likely to have little power because individual studies tend to have low precision. Nonetheless, the test can formally suggest heterogeneity in some instances, as will be seen in Section 6.4. Note that the X^2 statistic has an appealing interpretation (at least for researchers with experience in parametric linkage). Indeed, it can be re-written as

$$\begin{aligned}X^2 &= \sum_{i=1}^{K} \frac{\hat{\gamma}_i^2}{s_i^2} - \frac{\hat{\gamma}_{\text{hom}}^2}{(\sum_i 1/s_i^2)^{-1}} \\ &= 2 \times \ln 10 \times (\sum_{i=1,\ldots,K} \mathrm{LOD}_i - \mathrm{LOD}_{\text{hom}}) \,.\end{aligned}$$

In other words, the individual LODs add up only when the effect is perfectly homogeneous.

6.2.3 Modeling size heterogeneity

The classical way to allow for heterogeneity between studies is to introduce an additional layer in the earlier homogeneous model by assuming that the true study specific effects γ_i's themselves arise from some distribution. The usual model is a normal distribution with common mean γ and a between study variance σ^2. This is referred to as a normal mixture model (or random effect model) and results in marginal distributions for the observations given by $\hat{\gamma}_i \sim N(\gamma, s_i^2 + \sigma^2)$. If the between study variance σ^2 were known, the estimate of γ would be

$$\hat{\gamma}_{\text{het}}(\sigma^2) = \frac{\sum_i w_i \hat{\gamma}_i}{\sum_i w_i} \text{ with } w_i = \frac{1}{\sigma^2 + s_i^2} \text{ and with standard error } SE_{\text{het}} = 1/\sqrt{\sum_i w_i}\,,$$

So, one way to carry out estimation is by maximization of the profile log-likelihood $pl(\sigma^2) = l(\hat{\gamma}_{\text{het}}(\sigma^2), \sigma^2)$.

In the context of linkage where the actual effects γ_i's are standardized variance components themselves, this model only makes sense if the probability $\Phi(-\gamma/\sigma)$ of negative γ's is negligibly small. In practice that is achieved if the coefficient of variation $\sigma/\gamma < 1/2$. For the same reasons, the null hypothesis of no locus effect requires that all γ_i's should be equal to 0 with probability 1. Hence, the null hypothesis specifies both $\gamma = 0$ and $\sigma^2 = 0$, which is different from the usual situation in meta-analyzes of clinical trials. The test for linkage is then given by the corresponding log-likelihood difference

$$2 \times \left[pl(\hat{\sigma}^2) - l\left(\gamma = 0, \sigma^2 = 0\right)\right]$$

so that evidence for heterogeneity potentially contributes to the rejection of the null hypothesis of no linkage. The use of the usual mixture $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ for the null distribution of this non-standard likelihood is anti-conservative, the correct asymptotic distribution is given by a mixture $(\frac{1}{2}-p)\chi_0^2+\frac{1}{2}\chi_1^2+p\chi_2^2$ (Self and Liang, 1987). However, asymptotic results are unlikely to be useful since we typically have very few observations (i.e. studies) to pool together. In practice, we use the anti-conservative limits dictated by the $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ mixture as a screening tool and resort to parametric bootstrapping for refinement of the level of significance once interesting positions have been identified.

6.2.4 A two-point mixture model for locus heterogeneity

In some cases, the previous model will not be adequate to model differences between studies because heterogeneity is qualitative rather than quantitative, in other words the locus influences the trait in some studies/populations and not at all in others. There is an indication of such qualitative heterogeneity when the normal mixture model yields a large coefficient of variation σ/γ allowing negative γ 's under the normal mixture . In analogy to what is done routinely at the family level in parametric linkage (e.g. Ott (1999), see also Holliday *et al.*(2005) for a recent application) and can be done in the variance components setting (Ekstrom and Dalgaard, 2003), one can fit a two-point mixture model at the study level as follows: $\hat{\gamma}_i|\gamma_i \sim N(\gamma_i, s_i^2)$ with

$$\gamma_i = \begin{cases} \gamma, & \text{with probability } \alpha; \\ 0, & \text{with probability } 1-\alpha \end{cases}$$

so that, marginally,

$$\hat{\gamma_i} \sim \alpha N(\gamma, s_i^2) + (1-\alpha)N(0, s_i^2)\ .$$

The basic idea is that only a proportion α of the studies show linkage to the putative locus and γ is the QTL effect among those studies only. (Hence, γ is not longer the mean value of the γ_i's as in the normal mixture model. Care is needed when comparing the models) . For estimation purposes, this mixture of normal distributions naturally lends itself to the EM algorithm (Dempster *et al.*, 1977). Denoting by $\phi(x;\mu,\sigma^2)$ the normal density function with mean μ and variance σ^2, the E (estimation) step at stage $k+1$ of the iterative procedure consists in calculating the posterior probabilities $\tau_i^{(k+1)}$'s that the $\hat{\gamma}_i$'s have arisen from a normal distribution with mean $\gamma^{(k)}$ given the prior mixing proportion $\alpha_{(k)}$ i.e.

$$\tau_i^{(k+1)} = \frac{\alpha^{(k)}\phi(\hat{\gamma}_i, \gamma^{(k)}, s_i^2)}{\alpha^{(k)}\phi(\hat{\gamma}_i, \gamma^{(k)}, s_i^2) + (1-\alpha^{(k)})\phi(\hat{\gamma}_i, 0, s_i^2)}\ ,$$

whereas the M (maximization) step gives the updated parameters $\alpha^{(k+1)}$ and $\gamma^{(k+1)}$ as

$$\begin{aligned} \alpha^{(k+1)} &= \sum_{i=1}^{K} \tau_i^{(k+1)}/K \\ \gamma^{(k+1)} &= \frac{\sum_{i=1}^{K} \hat{\gamma}_i \tau_i^{(k+1)}/s_i^2}{\sum_{i=1}^{K} \tau_i^{(k+1)}/s_i^2} . \end{aligned}$$

The model parameters α and γ are constrained in $[0,1]$ and $[0,+\infty[$ respectively and although the EM estimation procedure described above ensures that $\alpha \in [0,1]$, the estimate of γ will sometimes be negative in which case we set $\hat{\gamma} = 0$ and $\hat{\alpha} = 0$ too. Under usual regularity conditions, the corresponding likelihood ratio test would be asymptotically distributed as a $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ under the null hypothesis. However, here the situation is further complicated by the fact that the model parameters are not identifiable under the null hypothesis (indeed if $\gamma = 0$, any choice of α will give the same likelihood). One way to tackle this problem is to slightly modify the likelihood as done by Chen *et al.*(2001) and derive corresponding simple asymptotics, but for the same reason alluded to in Section 7.2, we prefer to resort to parametric bootstrapping techniques in order to assess significance of the likelihood ratio test.

The model for size heterogeneity and locus heterogeneity could be combined into a model where either $\gamma = 0$ with probability $1 - \alpha$ or γ follows a normal distribution with probability α .

6.3 Extracting the relevant information from the individual studies

As we described in Section 7.2 the basic ingredients of a classical meta-analysis are study specific QTL effects' estimates $\hat{\gamma}_i$'s in the $i = 1, \ldots, K$ studies available and their associated standard errors s_i's on a *common* fine *grid* of genome locations. In this section, we explain how to obtain these estimates in practice and how to adjust for varying information across studies.

6.3.1 General approach

For random samples of the trait values, the variance components method (Almasy and Blangero, 1998; Amos, 1994) is the standard way of testing for linkage to a quantitative trait. Unfortunately, the emphasis of most computer programs implementing the variance components method has been placed on testing rather than estimating and they rarely provide both QTL effect estimates and associated standard errors. In the context of linkage, two exceptions that we know of are the `MENDEL` (Lange, 2001) and `Mx` softwares (Neale *et al.*, 1999). However, in principle, this is not so much of a problem because asymptotic standard errors s can be obtained provided the QTL effect estimate $\hat{\gamma}$ is present (and differs from 0) in addition to its statistical significance, using the approximate relation $(\hat{\gamma}/s)^2 \simeq \chi^2$ with $\chi^2 = \text{LOD} \times 2 \times \ln(10)$.

At positions where the QTL estimate is 0, one could interpolate values of s at neighboring positions where $\hat{\gamma} \neq 0$. One problem with the variance components method, as far as pooling of estimates is concerned, is that $\hat{\gamma}$ is constrained to remain nonnegative and pooling of several imprecise estimates $\hat{\gamma}_i$'s could result in a positively biased estimate of the true QTL effect γ. Whenever possible, we would personally favor adequate regression or score test approaches (Lebrec *et al.*, 2004) to linkage whose slope is equal to $\hat{\gamma}$ and is allowed to be negative. As shown by Putter *et al.*(2002), such approaches are equivalent to the variance components method.

Often, data are selected based on phenotype values (selected sample such as affected sibpairs, extremely discordant pairs, etc ...), the variance components method is no longer valid and appropriate methods that take into account the sampling scheme need to be employed. These so-called inverse regression methods first introduced by Sham and Purcell (2001) have been implemented in `MERLIN-regress` (Sham *et al.*, 2002). A typical output from the software will provide a signed estimate of the QTL effect $\hat{\gamma}$ and associated standard error s at an arbitrary grid of positions. The software can also be used in case of random samples as an alternative for the variance components modules. Because of its very convenient output we advocate the use of `MERLIN-regress` when analyzing linkage data whenever suitable. One outstanding problem with `MERLIN-regress` is the use of an imputed covariance for IBD sharing which can lead to bias in estimation especially in genome areas where markers information is very low. In practice, one clear indication that the imputed covariance is not a good approximation is when the software either gives out QTL estimates larger than 1 with huge associated LOD scores or no estimates at all (NA). In practice, marker maps and densities vary widely and one often ends up with areas of the genome with scarce information. In this case, we advocate the use of a more reliable IBD covariance matrix which we calculate by Monte Carlo simulations. In Section 6.3.2, we provide more details on how we do this in practice.

6.3.2 *Special case: sib pair designs*

In order to show how we adjust for differing marker maps (or different allele frequencies on the same map), we now outline the inverse regression approach in the simplest and most widespread case of sib pair studies. The trait values $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2)'$ are assumed to have been standardized and to follow the usual additive variance components model i.e. the vector $\boldsymbol{x}$ is assumed to follow a bivariate normal distribution with mean 0 and covariance matrix Σ

$$\Sigma = \begin{bmatrix} 1 & \rho + \gamma(\pi - \frac{1}{2}) \\ \rho + \gamma(\pi - \frac{1}{2}) & 1 \end{bmatrix} .$$

Here π is the proportion of alleles shared IBD measured exactly at the QTL position and γ therefore represents the proportion of total variance explained by the QTL, ρ is the marginal sib-sib correlation for the trait of interest. An extension of a relation shown in Putter *et al.*(2003) under complete information gives an approximate regression (valid for small values of γ) between excess IBD sharing and a function of the phenotype trait values which is the basis of the inverse regression approach:

$$\mathrm{E}(\hat{\pi} - \frac{1}{2}|\boldsymbol{x}, \gamma) \simeq \gamma \, \mathrm{var}_0(\hat{\pi}) \, C(\boldsymbol{x}, \rho)$$

where

$$\hat{\pi} = \frac{1}{2} \times \mathrm{P}_0(\pi = \frac{1}{2}|M) + 1 \times \mathrm{P}_0(\pi = 1|M)$$

is the usual estimate of IBD sharing given marker data M available while

$$C(\boldsymbol{x}, \rho) = \left[(1+\rho^2)x_1x_2 - \rho(x_1^2 + x_2^2) + \rho(1-\rho^2)\right] / (1-\rho^2)^2$$

and is sometimes referred to as the optimal Haseman-Elston function. For a sample of $j = 1, \ldots, N$ sib pairs, the method of least squares provides an approximately consistent estimate of γ given by

$$\hat{\gamma} \quad = \quad \frac{\sum_{j=1}^{N}(\hat{\pi}_j - \frac{1}{2})C(\boldsymbol{x}_j, \rho)}{\mathrm{var}_0(\hat{\pi}) \times \sum_{j=1}^{N} C^2(\boldsymbol{x}_j, \rho)} \,, \qquad (6.2)$$

$$\text{with standard error } s \quad = \quad \left(\mathrm{var}_0(\hat{\pi}) \times \sum_{j=1}^{N} C^2(\boldsymbol{x}_j, \rho)\right)^{-1/2} \,. \qquad (6.3)$$

Here $\mathrm{var}_0(\hat{\pi})$ represents the variance of $\hat{\pi}$ under the null hypothesis and would equal $\frac{1}{8}$ under complete information and although an exact calculation is extremely tedious it can be closely approximated by simple Monte Carlo simulations. For example, one can use the options `--simulate` and `--save` in `Merlin` (Abecasis *et al.*, 2002) to generate a large number of pedigrees with a given structure (sib pairs here), markers' characteristics (i.e. allele frequencies and inter-marker distances) and possibly missing pattern for genotypes, the true $\mathrm{var}_0(\hat{\pi})$ can then be accurately approximated by the sample variance of $\hat{\pi}$. We show in Figure 6.1 how widely this measure of marker information may vary within and between studies. It is therefore crucial to appropriately account for this variation when estimating γ, failure to do so may introduce bias in the QTL estimates. If no such information is available, it is possible in principle to calibrate scan by comparing mean or median QTL variance components over the whole genome between studies, but in small studies such methods might be prone to error.

6.3.3 Retrieving information on the common grid from an individual study

For the meta-analysis we need to define a common grid of locations and obtain QTL estimates on that grid for each study. However, it can happen that in the individual studies, the only data at hand are QTL estimates ($\hat{\gamma}$'s) and their standard errors (s's) on an original grid of locations which is not the common one we wish to use. Typically this original grid would be a set of say $t = 1, \ldots, M$ markers' positions. We show how to obtain QTL estimates and associated standard errors on this new common grid of locations, if the characteristics of the original map are available and from the IBD distribution for that map under the null hypothesis.

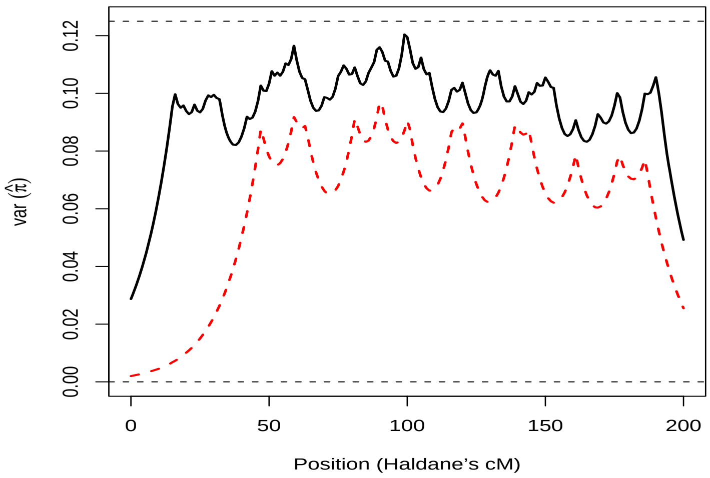


Figure 6.1: Marker information ($\text{var}_0(\hat{\pi})$) in the Australian (continuous line) and Dutch (broken line) data sets Vs. position (Haldane's cM) - Chromosome 6

For the sake of simplicity, we stick to sib-pair designs as in the previous section. Given the $M \times 1$ vector of original QTL effect estimates $\hat{\gamma} = (\hat{\gamma}_t)_{t=1,\ldots,M}$ and associated standard errors $(s_t)_{t=1,\ldots,M}$, the best linear approximation of the QTL effect $\hat{\gamma}_q$ at an arbitrary position denoted q is given by a weighted least squares estimate

$$\begin{aligned} \hat{\gamma}_q &= \frac{\omega_q' V^{-1} \hat{\gamma}}{\omega_q' V^{-1} \omega_q} , \\ \text{with standard error } s_q &= \left(\omega_q' V^{-1} \omega_q\right)^{-1/2} . \end{aligned}$$

Here $'$ denotes the transpose of a vector. The matrix V is proportional to the variance-covariance matrix of the vector $\hat{\gamma}$ under the null hypothesis of no linkage and is given by

$$V_{kl} = \begin{cases} \text{var}_0(\hat{\pi}_k)^{-1} & \text{if } k = l \\ \text{Cov}_0(\hat{\pi}_k, \hat{\pi}_l) \ \left(\text{var}_0(\hat{\pi}_k)\ \text{var}_0(\hat{\pi}_l)\right)^{-1} & \text{if } k \neq l \end{cases} ,$$

Furthermore, ω_q is the $M \times 1$ vector whose k^{th} element is given by

$$\omega_{q,k} = \frac{\text{Cov}_0(\hat{\pi}_k, \hat{\pi}_q)}{\text{var}_0(\hat{\pi}_k)} .$$

All the var_0 and Cov_0 terms can in principle be calculated by Monte Carlo simulations provided the map characteristics and pedigree structure are known.

In the idealized case of a saturated map which would supply perfect IBD knowledge at any location on a chromosome, all var_0 terms are equal to $\frac{1}{8}$ and $\text{Cov}_0(\hat{\pi}_{t_1}, \hat{\pi}_{t_2}) =$

$\frac{1}{8}(1 - 2\theta_{t_1,t_2})^2$, where θ_{t_1,t_2} is the recombination fraction between loci at t_1 and t_2 (Risch, 1990). Taking the off-diagonal terms in V to be equal to 0 (i.e. assuming that markers are not linked), one obtains the estimate of QTL effect advocated by Etzel and Guerra (2002) (in the special case that between-study variance $\sigma^2 = 0$). In the context of meta-analysis, it is important to properly account for differences in marker information between studies, unless the marker maps are close to saturated in all studies. Remarkably, the elements needed to calculate $\hat{\gamma}_q$ and s_q at any arbitrary location are just the corresponding estimates at M marker locations and map characteristics, none of the subject-specific data (traits values, individual IBD estimates $\hat{\pi}_i$) are needed.

6.4 Example

We applied the methods previously described to four data sets on lipid levels originating from Australia (aus), The Netherlands (nlj and nlo) and Sweden (swe). The full results are reported in Heijmans *et al.*(2005) and we have selected only one endpoint (LDL cholesterol levels) for illustration purposes. The data available for linkage analysis consisted almost entirely of sib pairs (371, 83, 110 and 36 pairs in the aus, nlj, nlo and swe data sets, respectively) with the exception of the Australian data set which also had 1 family with three siblings and 3 families with four siblings. Genotyping has been carried out using a common marker map for the nlj, nlo and swe data sets but with a different denser map for the aus data set. We actually had access to the raw data sets and could therefore easily obtain QTL estimates and standard errors on a common grid of positions.

Prior to linkage analysis (using `MERLIN-regress`), raw phenotypic data were adjusted for sex and age, within country. The analysis of the actual data revealed little differences between the three methods described in Section 7.2, this is partly due to the small sample sizes in the individual data sets which does not allow to clearly establish heterogeneity between studies. We present graphically the original results for two interesting chromosomes: chromosome 2 (Figure 6.2) and chromosome 13 (Figure 6.3). Note that the QTL variance estimates and LOD scores of the Pooled analyses have been multiplied by 0.95 and 1.05 for the random effect model (labelled ’het’) and the two-point mixture model (labelled ’2-p mixt’) respectively, this was necessary to make all curves visible.

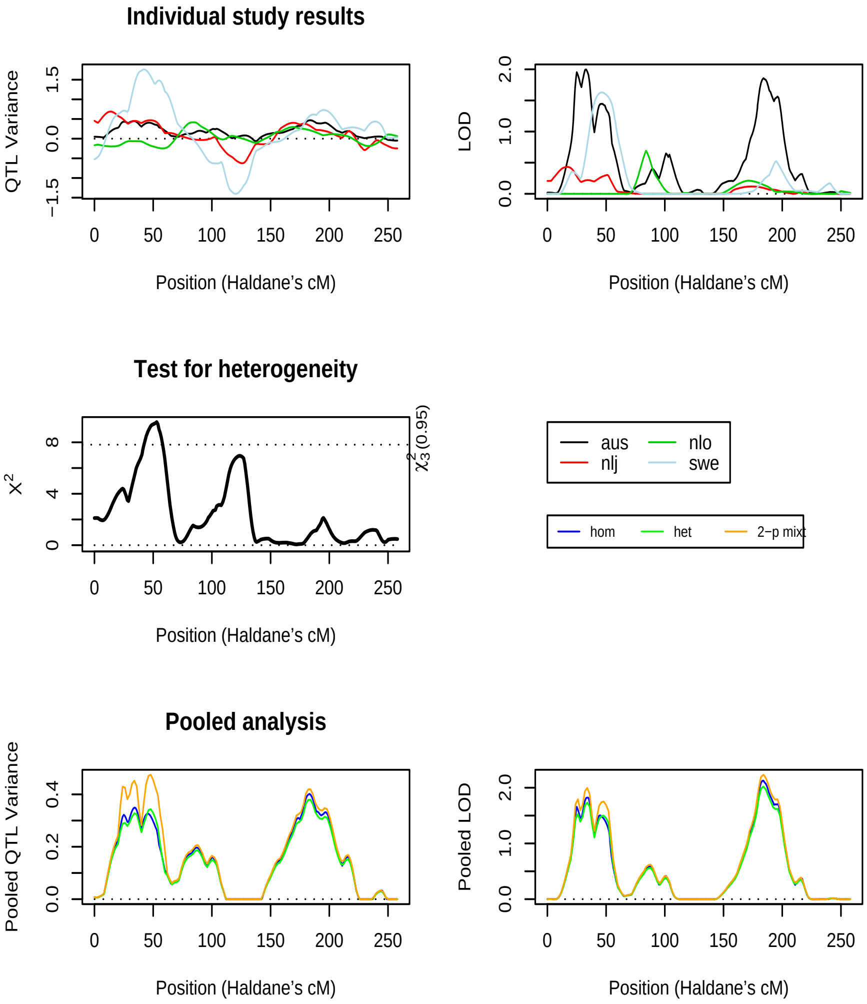


Figure 6.2: Original data - Chromosome 2 - LDL cholesterol level

In chromosome 2, the three pooled estimates of QTL variance coincide everywhere apart from the 20-60cM region where the two-point mixture model gives a higher estimate with corresponding estimate of proportion of study linked $\hat{\alpha} = 0.75$ (i.e. the 'nlo' data set is not linked) at 32cM where the maximum LOD score is attained. The corresponding pooled LOD score is roughly the same as the maximum LOD score obtained in the 'aus' data set and therefore there seems to be no gain in pooling the three linked data sets in this case. On chromosome 13, the pooling results in a very slight increase in LOD score in the region around 20cM compared to the maximum

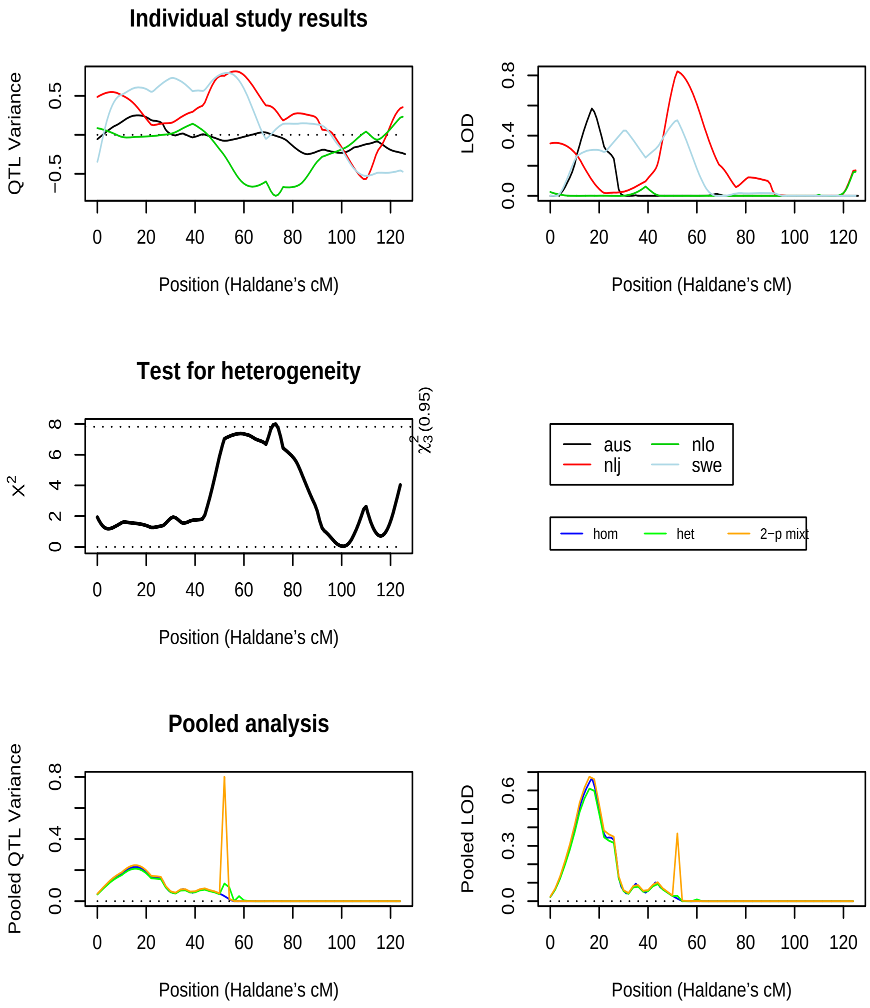


Figure 6.3: Original data - Chromosome 13 - LDL cholesterol level

of the individual LOD scores and the three methods give the same score. Note the sudden rise and fall in the estimate of QTL variance $\hat{\gamma}$ for the two-point mixture at 52cM which corresponds to a decrease in $\hat{\alpha}$ from 1.0 to 0.36. The fitting algorithm of the two-point mixture actually gave negative values for $\hat{\gamma}$ right of 54cM so the estimates were truncated to 0. Given those unconvincing real-life examples, one can legitimately asks the next two questions:

1. In practice, is there any gain in pooling data sets at all? I.e. can we obtain higher LOD scores than the maximum of the individual LOD scores?
2. Does allowance for heterogeneity help in enhancing statistical significance? I.e. are the LOD scores for the random effect model and/or the two-point mixture model ever higher than the LOD score of the homogeneity model?

The answer to question 1. is ’Yes’ even when individual studies are small provided the QTL effects are more or less the same in all studies i.e. the assumption of homogeneity is verified. The answer to question 2. is also ’Yes’ but only when the sample size in the individual studies are large enough as we show by means of a simulated example inspired from the original lipid levels data. We artificially increased the sample size of each of the four data sets by a factor 4 (i.e. the standard errors were divided by 2). The corresponding results are displayed graphically in Figure 6.4 for chromosome 2 and in Figure 6.5 for chromosome 13. In the 20-70cM region of chromosome 2, studies ’aus’ and ’swe’ both show clear linkage signals, QTL estimates vary quite widely across studies which is now unambiguously shown by the heterogeneity test. We are probably in presence of both quantitative and qualitative heterogeneity here since study ’nlo’ shows no QTL effect at all. As a result, the significant signals observed in the ’aus’ and ’swe’ studies (maximum LOD score $\simeq 8$) weaken in the homogeneous model (maximum LOD score $\simeq 7$) while both the heterogeneity model and the two-point mixture enhance it further (maximum LOD score $\simeq 10$). Heterogeneity therefore contributes to the proof that a linkage effect is present. Similar outputs are displayed for chromosome 13 in Figure 6.5. In the 40-70cM region, heterogeneity of QTL effects is now clearly qualitative (both ’nlj’ and ’swe’ have similar QTL effects with corresponding suggestion for linkage) and the pooled homogeneous analysis is dominated by the large ’aus’ study with QTL variance estimates close to 0 which entirely obliterates the individual linkage signals of ’nlj’ and ’swe’. The two-point mixture works best here in pooling evidence from the two positive studies and enhancing the LOD score beyond 4 in a much narrower region (maximum LOD score $\simeq 3.5$ in individual studies).

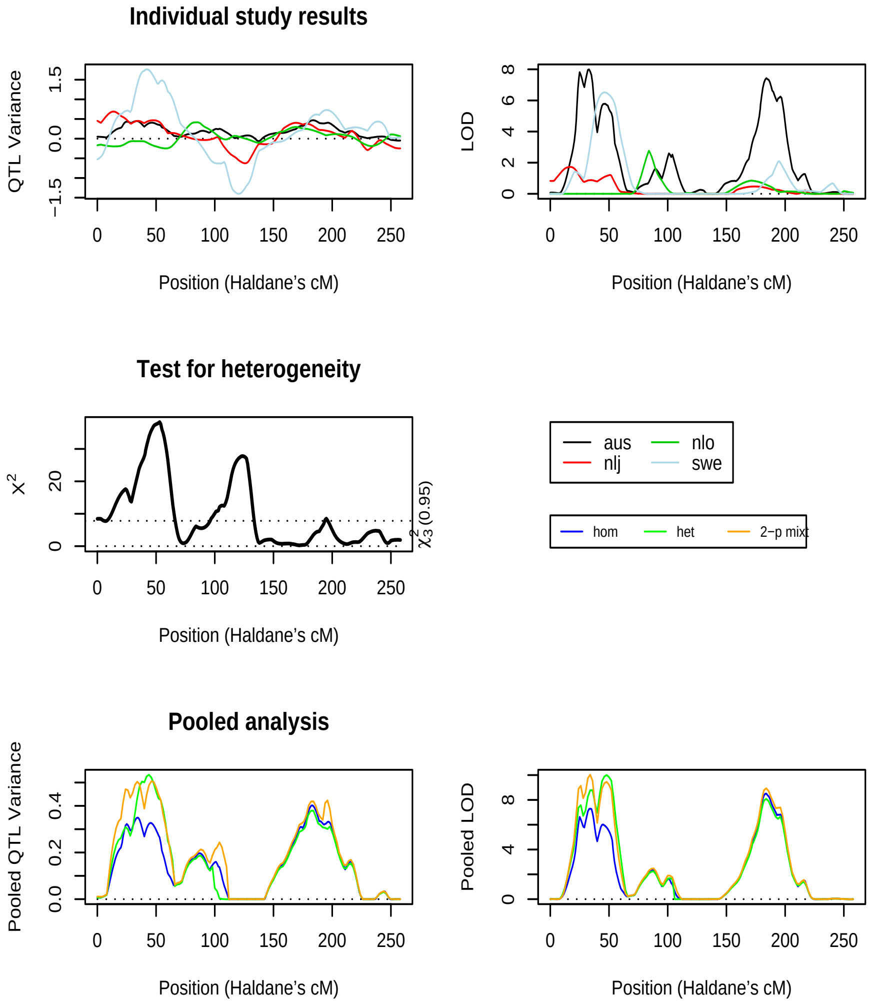


Figure 6.4: Artificial data - Chromosome 2 - LDL cholesterol level

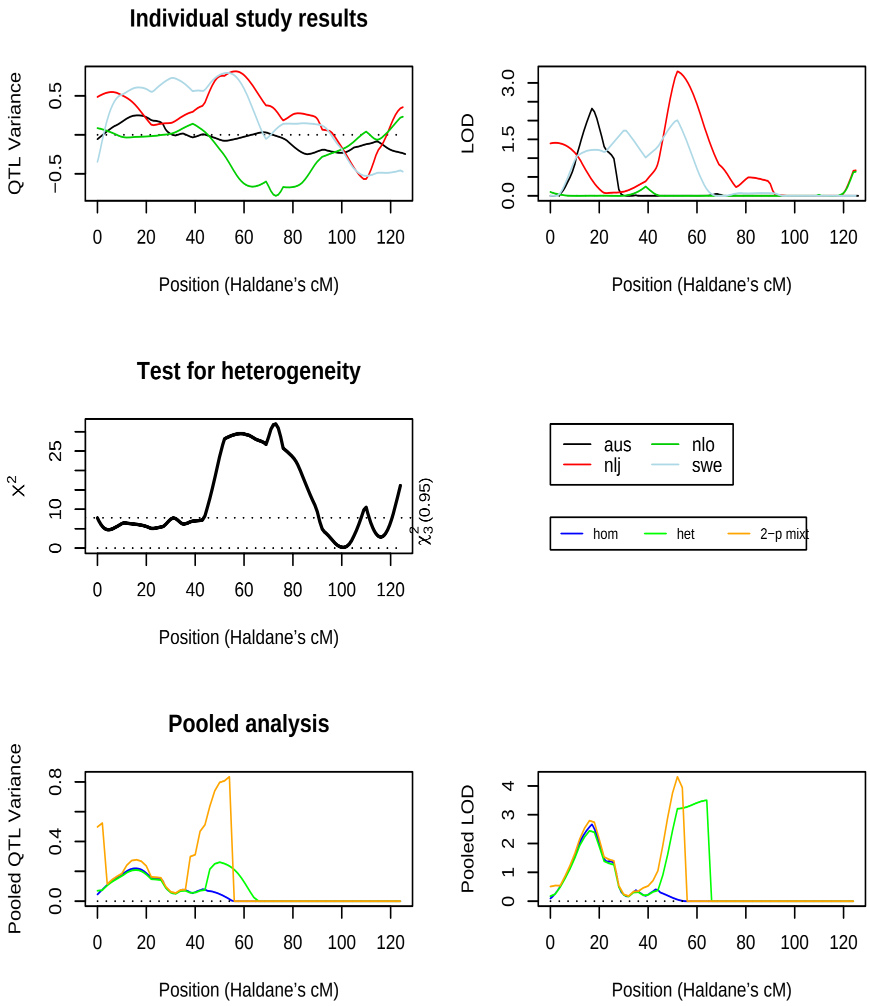


Figure 6.5: Artificial data - Chromosome 13 - LDL cholesterol level

6.5 Discussion

We have detailed how classical meta-analytic methods can be adapted to linkage provided consistent estimates of QTL effects along with standard errors are available for each study on a common grid of positions. The methods required to obtain such summary statistics are now well developed and their software implementation has been publicly available for a number of years. We realize, however, that most published studies to date will not have sufficient information in order to carry out the method advocated here. Indeed, it is still common practice nowadays in the literature, even for QTL mapping where the effect to be estimated is fairly uncontroversial, to publish statistics conveying statistical significance only (i.e. LOD scores) without any idea of the actual effect estimate. This heavily hinders powerful pooling of the many small linkage studies available in the community. Gu *et al.*(1998) presented guidelines on how to report linkage studies that would enable future meta-analysis using IBD sharing as a common linkage parameter. Since the analysis tools are available (e.g. `MERLIN-regress`), it should be expected by journals that researchers publish QTL effects and associated standard errors (at least as add-on information) on a grid of locations.

Given the small individual study sizes one typically encounters, any test for heterogeneity of QTL effects across studies is bound to suffer from a lack of power. This was reflected in the test for heterogeneity of the real lipid levels data as well as in the estimate of the between study variance component σ^2 which very rarely differs from 0 (Heijmans *et al.*, 2005). Another way to test for heterogeneity in the random effect model setting is to test whether $\sigma^2 = 0$ and this is known to be asymptotically equivalent to the X^2 test that we have presented (Andersen *et al.*, 1999). Note that the classical random effects model is probably not the most appropriate in the case of linkage, indeed the fact that the QTL effect is a variance component precludes it from being negative (which is not impossible under the normal mixture model) and suggests that the random effects γ_i's could be more appropriately modelled as arising from a Γ distribution but estimation then becomes less straightforward.

The idea of applying the concept of finite mixture models to meta-analysis is also not new (Böhning *et al.*, 1998) although it is new for meta-analysis of linkage studies as far as we are aware. It is based on the simple idea that only studies with a positive effect should be pooled together to provide evidence for linkage. Instead of doing this by hand, we let the data decide which study exhibits positive linkage. Note that one can also formally test for locus heterogeneity by assessing whether α differs from 1. Ultimately, given a sufficiently large number of studies with decent precision, it would be possible to fit a model that adapts to both locus and size heterogeneity by combining the random effect and the two-point mixture models.

6.6 Acknowledgements

We wish to thank Bas Heijmans, Marian Beekman, Hein Putter, Nico Lakenberg, Henk Jan van der Wijk, John Whitfield, Danielle Posthuma, Nancy Pedersen, Nick Martin, Dorret Boomsma and Eline Slagboom for allowing us to use their lipid levels data as an example. Gathering these data was a tremendous effort of many individuals and in particular we would like to acknowledge the work of Kate van Duijn, Eka Suchiman and Dennis Kremer for expert technical assistance; Harry Beeby, Scott Gordon and Megan Campbell for assistance with cleaning Australian genotype data; Dr. Daniel O'Connor for his effort to complete the Australian genotype data set;

and the Mammalian Genotyping Service of the Center for Medical Genetics, Marshfield, WI, USA (director Dr. James Weber) for genotyping part of the Australian twin sample. This work was partially funded by the GENOMEUTWIN project which is supported by the European Union Contract No. QLG2-CT-2002-01254.

CHAPTER 7

Combining Information Across Genome-wide Scans

Carol J. Etzel and Tracy J. Costello
University of Texas M. D. Anderson Cancer Center, Houston, Texas

7.1 Introduction

With the formation of international consortia to investigate complex disorders and a variety of cancers, meta-analysis is quickly becoming a valuable tool to combine linkage results and narrow chromosomal regions of interest. The presumed etiology of a complex disease is a combination of effects from multiple genes and the environment. The possibility of identifying some of these genes, which most likely have small effects, from a single study using traditional linkage analysis methods, is small. Instead, pooling raw data across independent studies (*i.e.* a mega-analysis) or pooling linkage results across independent studies (*i.e.* a meta-analysis) may be the best means to identify these numerous genes with typically small effects. Among-study heterogeneity, which may include differing marker maps, marker informativity, sample sizes, phenotype definition, ascertainment schemes, and linkage tests, can be problematic for a meta-analysis. Methods proposed to handle such problems are discussed here.

The basis of meta-analytic methods in genetic linkage is derived from pooling methods that have been available in the field of statistics for over 75 years. Such distinguished statisticians as Fisher (1925), Tippett (1931), and Pearson (1933) provide the earliest references to meta-analysis. These methods were based on testing a consensus or omnibus null hypothesis (*i.e.*, all null hypotheses from the individual studies are true) by combining the p-values from each of the individual studies. These methods are nonparametric in the sense that they do not rely on any distributional assumptions regarding the data in the individual studies; however, it is assumed that each study tests a common (and combinable) null hypothesis. Folks (1984) provides an excellent and detailed review of these early meta-analytic methods.

Meta-analysis for genome-wide scans has roots in methods developed for individual marker meta-analysis. These methods involved either pooling p-values (using the method of Fisher (1925)) or pooling estimates of genetic effects or of proportion of alleles shared identical by descent (ibd) among relative pairs (Li and Rao, 1996; Gu et al., 1998). However, current technology has evolved to allow investigators to perform full genome scans and therefore, linkage testing is not done for a single marker anymore. In this chapter, we review recent applications and extensions of meta-analytic methods for combining information across independent genome scans. We also provide strategies to choose a method suited to the scientific goals.

7.2 Meta-Analytic Methods for Genome Scans

In this section, we review meta-analytic methods that have been proposed and applied to genome-wide scan studies. Our coverage of such methods may not be exhaustive as we have tried to focus on such methods where power and type I error have been evaluated or methods (due to their ease of application) that have been widely used.

7.2.1 Meta-analytic methods based on p-values and tests of significance

As mentioned in the Introduction, general applications of meta-analysis have been developed from methods based on combining p-values. The method proposed by Fisher (1925) has been widely used in genetic linkage and many extensions have been developed for meta-analyses involving genome-wide scans. Suppose that we wish to complete a meta-analysis on k studies. Each study k has m markers. Let M_{st} denote the t^{th} marker, $t = 1, \ldots, m$, from study s, for $s = 1, \ldots, k$. Further define p_{st} as the p-value that provides evidence for linkage at the marker M_{st}. We are not assuming that each study used the same sampling scheme or linkage test; however the studies must be testing the same null hypothesis of no linkage. Using Fisher's method, we can define

$$X_t^2 = -2 \sum_{s=1}^{k} \ln(p_{st}) \tag{7.1}$$

as the combined evidence for linkage at marker $M_{\cdot t}$ across all studies. We can further define the p-value associated with X_t^2 as

$$P_t = \mathrm{P}(\chi^2_{2k} > X_t^2), \tag{7.2}$$

where χ^2_{2k} is distributed as a chi-square variate with $2k$ degrees of freedom. The power and type I error of this method was evaluated by Guerra et al. (1999) where a per marker alpha level of 0.1% was used to account for genome-wide testing. They concluded that although Fisher's method is applicable for genome scans, the power to detect linkage using this method is not equivalent to that achieved by pooling raw data.

One of the caveats to using this method to carry out a genome-wide meta-analysis is that an investigator is not guaranteed that all of the studies included in a meta-analysis will have used the exact same marker map. Or if the investigator is relying on published data, it is not guaranteed that results of all linkage studies are published, or of those that have been published, that results for all markers involved in a particular study will be readily available. Instead only information on local minimum p-values may reach publication. Therefore, the straightforward application of Fisher's method may not be feasible. Alternatives to Fisher's method have been proposed (informally and formally) in order to apply this meta-analytic method across whole regions of the human genome instead of single loci. One such informal application was proposed by Allison and Heo (1998) to combine data from several studies that used different tests for linkage and different markers to detect linkage within the Human *OB* region. Their technique involved obtaining a single p-value within the *OB* region from each of five published studies that investigated linkage to body mass index using different testing procedures for different sets of markers. Fisher's method was then used to combine the p-values across the five studies. They concluded that meta-analysis is a vital statistical tool that highlights the importance of published literature in the absence of available raw data and increases the power to detect genes influencing complex traits. They note that their approach illustrates that one can conduct a meta-analysis over multiple linkage studies investigating a single phenotype despite what they describe as "worst case conditions." However, we argue that the situations that Allison and Heo describe are realistic of early linkage publications and worst case conditions are those in which no meta-analysis can be performed.

Badner and Gershon (2002b) formally considered a similar modification of Fisher's method so that meta-analysis can be performed for regions across the human genome instead of one marker at a time. In their paper, they defined equation (7.2) as the Multiple Scan Probability (MSP) with p^*_{st} substituting for p_{st}, where p^*_{st} is defined as the minimum observed p-value for study *s* over a specified linkage region *t* corrected for the size of the linkage region. Their correction factor was based on the Feingold et al. (1993) estimate of the probability of a p-value being observed in a specified region size, namely

$$p^*_{st} = Cp_{st} + 2\lambda GZ(p_{st})\phi(\Phi^{-1}(p_{st}))V[\Phi^{-1}(p_{st})\sqrt{4\lambda\Delta}] \qquad (7.3)$$

where p_{st} is the observed p-value from study s over region *t*, C is the number of chromosomes, λ is the rate of crossovers per Morgan (which varies based on the linkage method employed and family structure), G is the size of region *t* in Morgans, $\Phi^{-1}(\cdot)$ is the standard normal inverse function, $\phi(\cdot)$ is the normal density function, Δ is the average distance in Morgans between adjacent markers and the function *V* is a discreteness correction factor for Δ. Feingold et al. (1993) show that $V(x) \approx \exp(-0.583x)$, for $x < 2$. Under certain conditions, they also show that equation (7.3) is equivalent to the Lander and Kruglyak (1995) p-value correction factor. Badner and Gershon (2002b) show via simulation that the type I error rate for this modification is at least as low as for any single genome scan study and that power to detect linkage using this method is equivalent to that of pooling raw data. This method has been applied to studies involving autism (Badner and Gershon, 2002b) and bipolar disorder and schizophrenia (Badner and Gershon, 2002a).

Another caveat to applying Fisher's method to genome-wide scans is that many widely used linkage tests are one-sided (i.e., LOD scores have a lower bound of 0) whereas the distributional assumptions for Fisher's original method assume that the p-values were derived from two-sided tests. Province (2001) suggested an extension of Fisher's general method to adjust for the potential bias of combining linkage results from such one-sided tests. Citing the one-to-one correspondence between LOD scores and p-values (Ott, 1999)

$$p_{st} = 1 - \Phi[sign(LOD_{st})\sqrt{2\ln(10)|LOD|}], \tag{7.4}$$

where $\Phi(\cdot)$ is the standard normal distribution function, Province recommended that LOD scores equal to zero should be assigned a p-value equal to $\frac{1}{2\ln(2)} \approx 0.72$ instead of equal to 0.50 as given by equation (7.4) or equal to 1.0 as suggested by maximum-likelihood theory. By doing so, the resulting test statistic obtained from Fisher's method using p-values extracted from published or derived LOD scores would roughly follow the assumed chi-square distribution with the appropriate number of degrees of freedom (2 times the number of studies) under the null of no linkage. This extension of Fisher's method has been applied to genome scan studies involved in the National Heart, Lung and Blood Institute Family Blood Pressure Program looking for obesity- related genes (Wu et al., 2002), hypertension-related genes (Province et al., 2003) and diabetes (An et al., 2005).

The Fisher p-value method and its subsequent extensions do not necessarily account for among-study heterogeneity with one of the most obvious differences being sample size and hence admittedly are subject to potential biases from not accounting for such differences among studies. Although decision criteria could be developed such that only studies that are most homogeneous (with respect to sample size or pedigree selection) be included in a meta-analysis, this may exclude too many studies with viable linkage information and hence limit the sample size for the meta-analysis (see discussion below). Rice (1990) suggested a reparameterization of Fisher's method such that the evidence for linkage from each study can be weighted by the corresponding study's sample size. In doing so, he suggested that the p-value, p_{st}, be transformed into a standard normal variate, $z_{st} = \Phi^{-1}(p_{st})$ where $\Phi^{-1}(\cdot)$ is the standard normal inverse function. A weighted average of the z-values at marker *t* (or region *t* if applying this reparameterization to the Badner and Gershon extension) can be calculated

$$z_{\cdot t} = \frac{\sum_{s=1}^{k} N_s z_{st}}{\sum_{s=1}^{k} N_s}$$

where N_s is the sample size (number of pedigrees, number of sib-pairs, etc.) for study *s*. Under the omnibus null hypothesis of no linkage, $z_{\cdot t}/\sqrt{Var(z_{\cdot t})}$ follows a standard normal distribution where

$$Var(z_{\cdot t}) = \frac{\sum_{s=1}^{k} N_s^2}{(\sum_{s=1}^{k} N_s)^2}.$$

Other novel meta-analytic methods for genome scans that use p-values or other outcomes of significance tests involving linkage which are not extensions of Fisher's method have been proposed specifically for genome-scan meta-analysis. One such widely used method, the Genome Search Meta-analysis Method (GSMA), developed by Wise et al. (1999) is based on a nonparametric ranking of p-values or LOD scores within specified genetic regions (or bins). Suppose that we have split the chromosomes into *m* bins. For each genome-scan study *s* ($s = 1, \ldots, k$ =number of total studies) the most significant linkage result (whether it be p-value, LOD score or another linkage test statistic) within each bin *t* ($t = 1, \ldots, m$) is identified. The bins are then ranked within each study where the most significant bin receives the highest rank. The ranks for each bin are then summed across the studies, such that

$$V_t = \sum_{s=1}^{m} R(X_{st}) \tag{7.5}$$

where X_{st} is the most significant linkage result for bin *t* of study *s*, and $R(\cdot)$ is the ranking function. As with Fisher's method, there are no assumptions that each study used the same sampling scheme or linkage test, or that each genome scan used the same set of markers. Additionally, however, they showed through simulation that the GSMA is useful when studies use different ascertainment schemes, marker maps, or statistical methods to detect linkage. citetWise1999 derived the null distribution of V_t given in (7.5) and Koziol and Feng (2004) refined the derivation of the null distribution using probability generating functions and provided approximations to the GSMA null distribution.

Wise (2001) further proposed an extension of the GSMA method such that candidate region studies can be included in the meta-analysis with genome-wide studies. In this extension, a simulation procedure is developed to assign ranks to the candidate regions where the ranks reflect the expected ranks under the null hypothesis of no linkage for a genome-wide study. By assigning the ranks to the candidate regions in this manner, Wise concludes that the false positive rate is not inflated due to the higher marker density of candidate region studies.

Babron et al. (2003) updated the GMSA method by first replacing the rank V_t in equation (7.5) with the average rank of bin *t* and the ranks of its two flanking bins, defined as V_{-t} and V_{+t} in order to adjust for arbitrary bin construction. Second, they defined a weighting scheme for the ranks such that the rank of study *s* in bin *t*, namely X_{st} in (7.5), is weighted by the number of pedigrees in study *s* in order to account for differing information content across studies. Although Babron et al. (2003) suggested weights to account for differing information content, a formal test for heterogeneity among the studies for the GSMA method was not introduced until 2005. Zintzaras and Ioannidis (2005b) propose three weighted metrics to measure among-study heterogeneity for the GSMA method: 1. sum of the weighted squared mean rank deviations, 2. sum of the weighted absolute mean rank deviations and 3. weighted sum of the distinct absolute rank differences. Furthermore, Zintzaras and Ioannidis (2005a) have developed a software program HEGESMA to perform the GSMA meta-analysis (unweighted or weighted as specified by the user) as well as provide the user with heterogeneity results.

In their original paper, Wise (2001) suggested a bin width of 30 cM, but recently, Marazita et al. (2004) proposed repeating the GSMA with variable bin-length starting points in order to determine minimum regions of maximum significance (MRMS). The resulting bin-shifting method identifies narrower regions of positive findings compared to the original GSMA which then leads to narrower regions to be followed-up with fine-scale mapping.

Since its original publication, the GSMA has been the most widely used meta-analytic method for genome scans, specifically due to its ease of use and invariance to whether the studies are from one-sided or two-sided tests or if only the most significant results have been reported. A number of investigators have applied the GSMA method to a variety of complex diseases: multiple sclerosis and other autoimmune diseases (Wise et al., 1999; Fisher et al., 2003; Sagoo et al., 2004), inflammatory bowel disease (Williams et al., 2002; van Heel et al., 2004), asthma (Wise, 2001), celiac disease (Babron et al., 2003), schizophrenia and bipolar disorders (Levinson et al., 2003; Lewis et al., 2003; Segurado et al., 2003), obesity (Johnson et al., 2005), diabetes (Demenais et al., 2003), coronary heart disease (Chiodini and Lewis, 2003) and hypertension (Liu et al., 2004; Koivukoski et al., 2004) to name a few.

7.2.2 Meta-analytic methods based on effect sizes

A meta-analysis based on combining the results from significance tests can be limited or misleading, especially in cases where the concordance or discordance of significant linkage between two studies may not reflect the existence of true linkage, but rather may be based on the amount of heterogeneity between the studies. Although adjustments for heterogeneity have been proposed for these methods, combining effect sizes may be a better approach as many of these methods are based on random effects models that naturally allow the user to adjust for among-study heterogeneity.

Loesgen et al. (2001) developed a meta-analytic test that computes a weighted average estimate of score statistics

$$Z_{MA_t} = \frac{\sum_{s=1}^{k} w_{st} Z_{st}}{\sqrt{\sum_{s=1}^{k} w_{st}^2}} \tag{7.6}$$

where Z_{st} is the NPL score statistic and w_{st} is the assigned weight from study *s* at position *t*. They proposed several weighting schemes such as sample size, information content and an exponential function based on marker distance. Dempfle and Loesgen (2004) compared the power of the method proposed by Loesgen et al. (2001) to Fisher's method, the GSMA and other p-value based meta-analytic methods. They showed that meta-analysis performed using weighted effect sizes had more power to detect linkage than the p-value methods with nominal increases in false positive rates. Further, they found that their method based on effect sizes was more robust and consistent across simulation aspects compared to the p-value based methods.

Etzel and Guerra (2002) developed a meta-analysis technique to combine Haseman-Elston test statistics across studies that have distinct marker maps. For this method they suppose that $\hat{\beta}_{st}$, the Haseman-Elston slope estimate (Haseman and Elston, 1972), and S^2_{st}, the corresponding variance estimate of $\hat{\beta}_{st}$ for the marker *t* of study *s* are available for each of *k* studies. They further define $\{L_q, q = 1, \ldots, v\}$ as the set of analysis points such that L_1 and L_t are at each endpoint of a chromosome segment, respectively, and the distance between any two adjacent points L_i and L_{i+1} is constant and equal to L/t where L is the length of the chromosome segment. For each analysis point, they calculate the statistics $\hat{\beta}_{stq}$ and S^2_{stw} utilizing markers within D cM of L_q, where

$$\hat{\beta}_{stq} = \frac{\hat{\beta}_{st}}{[1 - 2\theta_{stq}]^2} \text{ and } S_{stq} = \frac{S^2_{st}}{[1 - 2\theta_{stq}]^4}.$$

The value θ_{stq} is the recombination fraction between marker *t* of study *s* and analysis point L_q as estimated using a general mapping function, for example, Kosambi. Next, they calculate the weighted least-squares estimate $\tilde{\beta}_q$ at L_q,

$$\tilde{\beta}_q = \frac{\sum_{s=1}^{k} \sum_{t=1}^{n_{sq}} w_{st} \hat{\beta}_{stq}}{\sum_{s=1}^{k} \sum_{t=1}^{n_{sq}} w_{st}} \text{ and } w_{st} = \frac{1}{\sigma^2_B + S^2_{stq}}$$

where *k* is the number of studies and n_{sq} is the number of markers within D cM of L_q for study *s* and σ^2_B is between-study variance. The estimator $\hat{\sigma}^2_{B_q}$ for σ^2_B at L_q is

$$\hat{\sigma}^2_{B_q} = \frac{1}{\sum_{s=1}^{k} n_{sq} - 1} \sum_{s=1}^{k} \sum_{t=1}^{n_{sq}} [\hat{\beta}_{stq} - \bar{\beta}_{\cdot\cdot q}]^2 - \frac{1}{\sum_{s=1}^{k} n_{sq}} \sum_{s=1}^{k} \sum_{t=1}^{n_{sq}} S^2_{stq},$$

where $\bar{\beta}_{\cdot\cdot q}$ is the average of the $\hat{\beta}_{stq}$ that are within D cM of L_q. The variance of $\tilde{\beta}_q$ is $1/\sum_{s=1}^{k} \sum_{t=1}^{n_{sq}} w_{st}$. The analysis point $L_{q'}$ such that $t_{q'} = \tilde{\beta}_{q'}/\sqrt{Var[\tilde{\beta}_{q'}]}$ is minimum and significant at a specified level is the point estimate of location of the QTL. Likewise, the estimate of genetic variance is given by $\hat{\sigma}^2_g = \frac{\tilde{\beta}_{q'}}{-2}$. Etzel and Guerra (2002) further describe a bootstrapping procedure to construct confidence intervals for location of the putative QTL and genetic variance. Through simulation, they show that the empirical power using this procedure remained high even when power at the individual study level was low. This procedure was used to assess linkage of immunoglobulin E (IgE), an asthma related quantitative trait, using the nine data sets provided by the Genetic Analysis Workshop 12 and found suggestive linkage for two regions on chromosome 4 and one region on chromosome 11.

The method proposed by Loesgen et al. (2001) assumes that all studies use the same marker map but different linkage tests and the method proposed by Etzel and Guerra allows for differing marker maps among the studies involved; however, the Etzel and Guerra method is limited by the fact all studies must use the same linkage test. Etzel et al. (2005) (***GAW14) proposed a meta-analytic procedure that combines the methods of Loesgen et al. (2001) and Etzel and Guerra (2002) and results in a more flexible procedure to combine effect sizes across linkage studies that perform different linkage tests on different marker maps. The resulting Meta-Analysis for Genome Studies (MAGS) method is based on a weighted average of effect sizes that are obtained through the reported linkage summary statistics. Suppose that we wish to complete a meta-analysis on *k* studies. Each study *k* has m_k number of markers.

It is not assumed that the studies have the same number of markers, $m_i \neq m_k, i \neq j$, nor it is assumed that the studies have the same marker maps. For a specified chromosome, let M_{st} denote the t^{th} marker from study s, for $s = 1, \ldots, k$ and $t = 1, \ldots, m_k$. Define $\{L_q, q = 1, \ldots, l\}$ as the set of analysis points such that the L_q are equally spaced across the chromosome. For each set of M_{st} on a chromosome, let Z_{st} be the associated score statistic. As noted by Dempfle and Loesgen (2004), Z_{st} can be the NPL score statistic as most standard multipoint linkage analysis software packages includes the calculation of such statistics. However, Z_{st} can also be derived from other linkage related statistics, such as an HLOD score or even a p-value with the correct transformation (see Appendix A). For each analysis point L_q, calculate the weighted normal variate:

$$Z_{MA_q} = \frac{\sum_{s=1}^{k} \sum_{t=1}^{m_k} I_{q\{M_{st}\}} w_{stq} Z_{st}}{\sqrt{\sum_{s=1}^{k} \sum_{t=1}^{m_k} I_{q\{M_{st}\}} w_{stq}^2}},$$

where w_{stq} is the weight given to marker M_{st}. The indicator function $I_{q\{M_{st}\}}$ is defined as 1 if marker is within a set distance D cM from analysis point L_q and 0 otherwise. The weight w_{stq} for marker M_{st} can be a function of study sample size, information content at that marker, and/or distance (recombination fraction, θ_{stq}) between marker M_{st} and analysis point L_q, say $w_{stq} = f(n_s)g(IC_{q\{M_{st}\}})h(\theta_{stq})$.

The p-value for each analysis location then be compared to a set level to determine areas with combined evidence for linkage. NOTE: If all studies use the same marker map, then the combined set of markers can replace the analysis points L_q and the expression for Z_{MA_t} simplifies to the statistic proposed by Dempfle and Loesgen (2004). Etzel et al. (2005) applied this procedure to the simulated data from the Genetic Analysis Workshop 14 and correctly identified the disease loci on chromosomes 1, 3 and 5; however, found low evidence of linkage to the disease modifier genes on chromosomes 2 and 10.

7.3 Choosing a method to best suit your analytic needs

Data can be obtained from published sources, open-source websites or through consortia group agreements. At times, the researcher may be limited in choosing a preferred meta-analytic method due to the type of data available for a meta-analysis: complete data on all studies through a consortium; data obtained by contacting corresponding authors from published articles; data from published reports; or some combination of these three. However, the researcher who is able to obtain the data of his/her choosing should then select the meta-analysis method based on the most robust methodology for identifying linkage within each individual study. Below, we propose some scenarios that reflect reasonable situations in which a meta-analysis would be performed and provide advice regarding the type of meta-analytic method to use.

7.3.1 Scenario 1: Raw data available on all studies

This scenario could arise when the researcher is a member of a data consortium whereby members of the consortium freely share all data from their individual studies. For a meta-analysis, this is the most ideal situation since the researcher is relatively free to reanalyze the data (separately from each study) using a preferred linkage method and then combine the resulting linkage outcome using any one of the above mentioned meta-analysis methods. In order fully account for between-study heterogeneity, the researcher should choose one of the meta-analysis methods that allows for such an adjustment (Dempfle and Loesgen (2004), Etzel et al. (2005) or Zintzaras and Ioannidis (2005b)). Even if the marker maps are different among the studies in the consortium, the researcher could develop a simple scheme to align the marker maps in order to perform the meta-analysis. The researcher even has the option to not perform a meta-analysis, but to complete a mega-analysis instead, such that the raw data from each of the studies are combined into one common database. Some notable examples of this approach were applied to multiple sclerosis (Cooperative”, 2001; GAMES and Cooperative”, 2003), celiac disease (Babron et al., 2003), asthma (Iyengar et al., 2001), diabetes (Demenais et al., 2003) and obesity related phenotypes (Heo et al., 2002). A master marker map can be established by using a marker location database. If there are any missing values, one could consider imputation as in Heo et al. (2002). The combined data is then analyzed using a standard linkage method. It has been shown (Guerra et al., 1999), that a mega-analysis may have more power to detect linkage than a meta-analysis; however, one should consider the different types of heterogeneity that may be inherent in each of the different studies. This heterogeneity may adversely confound or overshadow the results from a mega-analysis and may arise from differing study designs (linkage results on extended pedigrees may not combine well with linkage results from sib-pairs, discordant pairs or parent-offspring triads), varying ethnic/racial groups across study populations (different genes acting in different populations) and varying sample sizes.

7.3.2 Scenario 2: All studies use similar linkage tests and similar marker maps

This scenario could also arise when the researcher is a member of a data consortium whereby the members individually analyze their own data using a common linkage method and freely share linkage results instead of raw data. Likewise, this scenario could occur when the researcher personally contacts corresponding authors from published studies and requested complete linkage analysis results from their data. If these data are obtained from corresponding authors, or extracted from the literature, the researcher should collect the most detailed information possible: i.e., score statistics instead of p-values, marker information content, recruitment criteria and sample schemes. For this scenario, we once again recommend that the researcher choose a meta-analysis method that is flexible enough to account for between-study heterogeneity: (Dempfle and Loesgen (2004) or Etzel et al. (2005) if score statistics are available or Zintzaras and Ioannidis (2005b) if only p-values are provided.

7.3.3 Scenario 3: All studies used similar linkage tests but with different marker maps

This scenario is similar to scenario 2 except for the commonality of the marker maps between the studies and likewise, this scenario could occur for the same reasons as scenario 2. The added complexity of differing marker maps will not hinder a meta-analysis over the individual studies, as long as the researcher uses a method that is flexible in this respect. Once again, we advise that the researcher request as detailed linkage information as possible and apply a meta-analysis based on the effect size method proposed by Etzel et al. (2005) if score statistics are available or the GSMA modification proposed by Zintzaras and Ioannidis (2005b) if only p-values are provided.

7.3.4 Scenario 4: p-values or LOD scores from different linkage tests and different marker maps from published data are available from all studies

In this scenario, it is assumed that the researcher is basing the meta-analysis on summary linkage results (p-values or LOD scores) that are available from published articles with no follow-up information obtained from the corresponding authors. Although the availability of data in this scenario may seem limited and can vary greatly depending on the disease of interest, manuscript type and journal of publication, many meta-analyses are based on such data (Allison and Heo (1998) for instance). For this case, the GSMA method (Wise et al., 1999) would be the best method to employ as long as the available data allow. If possible, the researcher could also employ any of the modifications to the GSMA method if s/he has ample auxiliary information to do so. In cases where application of the GSMA method is not possible (such as the scenario posed by Allison and Heo (1998)), then application of Fisher's method is still viable.

7.4 Discussion

Herein, we review current meta-analytic techniques for the combination of linkage data across studies in order to arrive at a consensus for linkage to a complex disease. We also propose several scenarios to help guide the researcher in their choice of which meta-analytic technique to employ. However, we caution that meta-analysis is more than just a method one can use to combine data together. Although the choice of method is important, the researcher must also keep in mind that the application of a method is just a small part of a complete meta-analysis. Just as study design and participant recruitment is important at the beginning of any linkage study, a researcher who is about to embark on a meta-analysis should also develop a study design and participant study plan which includes a literature review plan, as well as study inclusion/exclusion criteria. The researcher must also gather as much information on original studies as possible, which may include contacting corresponding authors. If raw data are provided, the researcher needs to decide how to treat missing data. The researcher may have ample data to complete a meta-analysis; however, roadblocks to complete the meta-analysis may exist. Most of these roadblocks include differences among the studies with respect to: marker maps or denseness of maps, family structure, environmental factors, population substructure, distinct genetic etiology/different pathways within the disease of interest, marker informativity, sample sizes, ascertainment schemes, phenotype definitions and/or linkage tests. Additional challenges include publication bias and time-lag bias. Although we presented meta-analytic methods that can handle some of these problems, no one single meta-analysis method exists that can handle all such problems. Therefore, a researcher must be willing to accept the limitations of his/her own meta-analysis.

Two topics that we have not discussed in detail within this chapter involve determining an appropriate significance level for a meta-analysis performed on genome scans and the effect of publication bias (only positive linkage results published). The topic of genome-wide significance levels for individual studies remains in controversy and to fully detail the debate with respect to a meta-analysis would be a lengthy chapter in itself. Instead, we leave it to the researcher to consider an appropriate significance level, but advise the researcher to look to Morton (1955), Lander and Kruglyak (1995), Feingold et al. (1993), Sawcer et al. (1990), Rao (1998), Rao and Gu (2001), and Levinson et al. (2003) to gain more insights into the determination of an appropriate significance level.

Publication bias in a meta-analysis may become a factor when the results of the study impact the probability that it will be published in the literature. In this event, if the published literature was biased in favor of statistically significant results, you would find a relative lack of studies reporting negative evidence for linkage and you could incorrectly conclude a region to be more significantly involved in the disease in question than it really is. Iyengar and Greenhouse (1988) present two procedures to handle this potential bias by estimating what they term the ’fail safe sample size.’ They first describe the procedure presented by Rosenthal (1979) which determines the minimum number of unpublished studies with null results required to reverse the conclusion of the meta-analysis over the published studies and note that Rosenthal (1984) provides some ad hoc guidelines for interpretation. Iyengar and Greenhouse (1988) extend the approach described by Rosenthal (1979) and present a second procedure based on selection models that uses a maximum likelihood approach to model the reporting process by weighting the results in the meta-analysis. They note that by using the MLE approach, you can examine how changing your assumptions about the selection model change the parameter estimates and inference of the meta-analysis.

7.5 Acknowledgements

This work was partially funded by National Cancer Institute K07 CA093592 and R03 CA110936 (CJE) and R25 CA57730 (TJC).

[This research was supported by a cancer prevention fellowship funded by the National Cancer Institute grant R25 CA 577730 and K07 CA 093592-02 and R03 CA110936]

7.6 Appendix A

Example transformation of a linkage summary to a score statistic

1. Transform an HLOD to Chi-square variate: $X_{st} = 4.6 * HLOD_{st}$

2. Obtain p-value for each chi-square variate (Faraway, 1993): $p_{st} = 0.5 * [1 - \mathrm{P}^2(\chi_1^2 < X_{st})]$

3. Transform the resulting p-value to a normal variate by the inverse of the normal distribution: $Z_{st} = \Phi^{-1}(p_{st})$

APPENDIX A

Combining Different Data Types

CHAPTER 8

A Misclassification Model for Inferring Transcriptional Regulatory Networks

Ning Sun, Hongyu Zhao
Yale University

8.1 Introduction

Understanding gene regulations through the underlying transcriptional regulatory networks (referred as TRNs in the following) is a central topic in biology. A TRN can be thought of as consisting of a set of proteins (transcription factors), genes, small modules, and their mutual regulatory interactions. The potentially large number of components, the high connectivity among various components, and the transient stimulation in the network result in great complexity of TRNs. With the rapid advances of molecular technologies and enormous amounts of data being collected, intensive efforts have been made to dissect TRNs using data generated from the state-of-the-art technologies, including gene expression data and other data types (e.g. Chu *et al.*, 1998; Ren *et al.*, 2000; Davison *et al.*, 2002; Lee *et al.*, 2002; Bar-Joseph *et al.*, 2003; Zhang and Gerstein, 2003). The computational methods include gene clustering (e.g. Eisen *et al.*, 1998; Roberts *et al.*, 2000), Boolean network modeling (e.g. Liang *et al.*, 1998; Akutsu *et al.*, 1999, 2000; Shmulevich *et al.*, 2002), Bayesian network modeling (e.g. Friedman *et al.*, 2000, Hartemink *et al.*, 2001, 2002), differential equation systems (e.g. Gardner *et al.*, 2003; Tegnr *et al.*, 2003), information integration methods (e.g. Gao *et al.*, 2004), and other approaches. For recent reviews, see de Jong *et al.*(2002) and Sun and Zhao (2004). As discussed in our review (Sun and Zhao, 2004), although a large number of studies are devoted to infer TRNs from gene expression data alone, such data only provide very limited amount of information. On the other hand, other data types, such as protein-DNA interaction data (which measure the binding targets of each transcription factor, denoted by TF in our following discussion, through direct biological experiments), may be much more informative and should be combined together for network inference.

In this article, we describe a Bayesian framework for TRN inference based on the combined analysis of gene expression data and protein-DNA interaction data. The statistical properties of our approach are investigated through extensive simulations, and our method is then applied to study TRNs in the yeast cell cycle.

8.2 METHODS

In this article, we model a TRN as a bipartite graph: a one-layer network where a set of genes are regulated by a set of TFs. The TFs bind to the regulatory regions of their target genes to regulate (activate or inhibit) the transcription initiation of these genes. Transcription initiation is a principal mode of regulating the expression levels of many, if not most, genes (Carey and Smale, 1999). Because the number of the genes largely exceeds the number of TFs in any organism (e.g. there are 374 TF entries in the updated Transfac database (http://www.gene-regulation.com/pub/databases.html) and more than 6000 genes in yeast), there is combinatorial control of the TFs on genes. That is, for a given gene, its expression level is controlled by the joint actions of its regulators. Two well-known facts on the joint actions of TFs include cooperativity, which in the context of protein-DNA interaction refers to two or more TFs engaging in protein-protein interaction stabilize each other's binding to DNA sequences, and transcriptional synergy, which refers to the interacting effects among the Polymerase II general transcriptional machinery and the multiple TFs on controlling transcription levels. In our previous work (Zhao *et al.*, 2003), we assumed that the expression level of a specific gene is controlled through the additive effects of its regulators. Liao *et al.* (2003) applied Hill's equation for the cooperative TF bindings on the regulatory regions of their target genes and the first order kinetics for the rate of gene transcription. Under a quasi-steady state assumption, they proved that the relative gene expression level has a linear relationship with the relative activities of the TFs that bind on the gene's regulatory region. In order to obtain a unique solution of the regulation matrix, they required the full column-rank of the regulation and its reduced matrices. In this article, we extend our previous work (Zhao *et al.*, 2003) to fully incorporate gene expression data and protein-DNA binding data to infer TRNs. Before the discussion of our model, we first give a brief overview of the protein-DNA binding data used in our method.

As the primary goal of TRN inference is to identify the regulation targets of each TF, the most direct biological approach for this goal is to experimentally identify the targets of various TFs. Many different biological methodologies are available to serve this purpose. The large-scale chromatin immunoprecipitation microarray data (ChIP-chip data) provide the in vivo measurements on TFs and DNA binding in yeast (Ren *et al.*, 2000; Lee *et al.*, 2002). In our study, the protein-DNA binding data thus collected are viewed as one measurement of the TRN with certain level of measurement errors due to biological and experimental variations, e.g. physical binding is not equivalent to regulation. We use the ChIP-chip data collected by Lee *et al.* (2002) as the data source for protein-DNA binding. These data represent a continuous measurement of the binding strength between each TF and its potential targets, and a p-value is derived based on replicated experiments to assess the statistical significance of binding. In our following work, the inferred binding p-values between a TF and its potential target genes are transformed into binary observations using a significance level cut off of 0.05. That is, for all TF-gene pairs whose p-value is below 0.05, we denote the observation as 1, representing evidence for binding, and for those pairs whose p-value is larger than 0.05, we denote the observation as 0, representing not sufficient evidence for binding. The reason that we utilize protein-DNA binding data is because we believe that the information from such data serves as a close measurement for the true underlying TRN.

In our previous work (Zhao *et al.*, 2003), we treated protein-DNA binding data as representing the true underlying network, and used a simple linear model to describe the relationship between the transcript amounts of the genes considered and their regulators' activities. In our current work, we extend this linear model to incorporate potential errors associated with protein-DNA binding data to integrate three components that are biologically important in transcription regulation, namely, the TRN as characterized by the covariate (or design) matrix in the linear model, protein regulation activities as defined by the predictors in the model, and gene expression levels as defined by the response variables. We propose a misclassification model to simultaneously extract information from protein-DNA binding data and gene expression data to reconstruct TRNs.

8.2.1 Model Specification

Our model relating gene expression levels, TRNs, and TF activities can be described through three sub-models:

- A linear regression model relating gene expression levels with the true underlying TRNs and regulators' activities;
- A misclassification model relating the true underlying networks and the observed protein-DNA binding data;
- Prior distribution on the TRNs.

The information on the measurement error can be built in a flexible way into a graphical model (Richardson and Gilks, 1993; Richardson, 1999). The hierarchical structure of our graphical model is summarized in Figure 8.1 and we describe each component in detail in the following.

The first sub-model: the linear regression model

Let N denote the number of genes and M denote the number of TFs related to the regulation of these genes. We consider a total number T of gene expression experiments, where these experiments may represent a time-course study, e.g. yeast cell cycle studies, or different knock out experiments. We focus on time-course experiments in our following discussion. In this case, we use t represents a specific time point. The observed gene expression levels at time t, $\mathbf{Y}_t$, are the vector of N expression levels normalized over all time points for each gene i and serve as the response

Figure 8.1: The hierarchical structure of the misclassification model discussed in this paper. The unknown parameters are in the ovals, and the known parameters are in the rectangles.

in the linear model (8.1) with the following form:

$$Y_t = X\beta_t + \epsilon_t \tag{8.1}$$

$$\epsilon_{it} \sim N(0, \sigma_t^2) \tag{8.2}$$

where **X** represents the true TRN, β represents the time dependent regulator activities of the *M* TFs, and ϵ_t represents the errors that are associated with gene expression measurements. In matrix **X**, each row corresponds to a gene and each column corresponds to a TF. Therefore, the (i,j)th entry in this matrix represents the regulation pattern of the *j*th TF to the *i*th gene. The value of this entry is 1 if the *j*th TF affects the transcription level of the *i*th gene, and the value is 0 otherwise. Therefore, if our primary interest is to infer the TRN, the overall objective is to infer the values in this matrix, either 0 or 1.

This model states that (1) the expression level of a gene is largely controlled by the additive regulation activities of its regulators, (2) the same regulator has the same relative effect on all its targets, (3) the TRN is identical across all time points, and (4) the errors associated with gene expression measurements have the same distribution across all the genes. We note that these assumptions are simplistic and may only provide a first order approximation to reality. This model has nevertheless (implicitly or explicitly) been used in the analysis of TRNs by many research groups and found good success. The limitations and modifications of these assumptions are further discussed in the Summary section.

Because protein-DNA binding data are often obtained from a mixture of biological samples across all the time points, e.g. the asynchronized cells, they measure an averaged protein-DNA binding over the whole cell cycle. Although we may use the time-course gene expression data to investigate the fluctuation of the network over time, the information at one time point may not be sufficient for statistical inference (see results in the simulation study in the following). Therefore, we make the assumption that the network is time independent and combine the information across time points. Consequently, the variation of the response variable, gene expression, across time points is accredited to the change in activities of the TFs, β_t. With the given activities of the predictors, the TRN of gene i ($\mathbf{X}_i$) is independent of the network of any other gene $\mathbf{X}_{i'}$, where $i' = 1, 2, ..., (i-1), (i+1), ..., N$.

The second sub-model: the misclassification model

In our model set-up, both the true and observed covariates are binary, where 0 corresponds to no regulation and 1 corresponds to regulation. We assume the following model (8.3-8.6):

$$P(W_{ij} = 1 | X_{ij} = 1) = 1 - p \tag{8.3}$$

$$P(W_{ij} = 0 | X_{ij} = 1) = p \tag{8.4}$$

$$P(W_{ij} = 0 | X_{ij} = 0) = 1 - q \tag{8.5}$$

$$P(W_{ij} = 0 | X_{ij} = 1) = q \tag{8.6}$$

where the values of p and q are the false-negative and false-positive rates of the protein-DNA data. In practice, these values may be directly estimated from some control experiments, thus we treat these parameters as known or prior information in the misclassification model and specify their values. In the case these values may not be precisely known, we also study the robustness of their misspecifications on statistical inference. Note that the false-positive and false-negative rates may be gene-TF specific, therefore, our assumption here represents a first-order approximation to reality that may need further extension in future studies. The binary binding matrix **W** serves as the measurement for the true TRN **X**.

The third sub-model: the exposure model

For this submodel, we need to specify the prior distribution of the regulatory matrix **X**. The prior distribution of **X** (π_X) describes the probability of X_{ij} being 1, where X_{ij} represents the regulation between TF j and gene i. We assume that the X_{ij} are independent and have an identical distribution π_X. For a given true network **X**, the value of π_X can be calculated from the data. When X is unknown and W serves as the surrogate of **X**, π_X is a model parameter to be specified.

8.2.2 MCMC algorithm for statistical inference

In our model set-up, a large number of unknown parameters $\{$**X**$, \beta_t, \sigma_t^2\}$ need to be inferred based on the observations $\mathbf{Y}_t$, t=1, , T, and **W**. We propose to use the Gibbs sampler for statistical inference. The Gibbs sampler is alternated between two steps: (1) sample $\{\beta_t, \sigma_t^2\}$ conditional on **X**; and (2) sample **X** conditional on $\{\beta_t, \sigma_t^2\}$. These two steps are described in detail in the following.

Given current estimate of **X**, the model reduces to a standard linear regression model. The parameters $\{\beta_t, \sigma_t^2\}$ are sampled through (8.7 and 8.8)

$$\sigma_t^2 \sim Inv - \chi^2(df, s_t^2) \tag{8.7}$$

$$\beta_t \sim N(\widehat{\beta}_t, \mathbf{V}_\beta \sigma_t^2) \tag{8.8}$$

where $df = N - M$, $\mathbf{V}_\beta = (\widehat{\mathbf{X}}^T \widehat{\mathbf{X}})^{-1}$, $\widehat{\beta}_t = \mathbf{V}_\beta \widehat{\mathbf{X}}^T Y_t$, and s_t is the sample standard deviation. The matrix is the current estimate for the TRN.

Given current estimates of $\{\beta_t, \sigma_t^2\}$, we individually update the TRN for each gene. If there are M TFs, there are a total of $K = 2^M$ possible combined patterns among the TFs to jointly regulate a specific gene. The likelihood L_{ik} for each pattern k can be evaluated as

$$L_{ik} = L_{ik}^X + L_{ik}^Y \tag{8.9}$$

where

$$L_{ik}^X = n_1 \log \pi_X + n_{11} \log(1-p) + n_{10} \log p + n_0 \log(1-\pi_X) + n_{01} \log q + n_{00} \log(1-q) \tag{8.10}$$

$$L_{ik}^Y = -\sum_{t=1}^{T} \frac{(Y_{it} - \widehat{Y_{ikt}})^2}{2\sigma_t^2} \tag{8.11}$$

In the above expression, L_{ik}^X and L_{ik}^Y represent the likelihood contributions from the protein-DNA binding data and the expression data, respectively. In the expression for L_{ik}^X, n_{so} represents the number of TF-gene pairs whose true regulation is s and the observed binding is o, where the values of s and o are 0 or 1. For example, n_{11} corresponds to the number of pairs whose true regulation and observed binding are both 1, $n_1 = n_{10} + n_{11}$, and $n_0 = n_{00} + n_{01}$. The expression for L_{ik}^Y represents the likelihood component derived from gene expression data across all time points.

After evaluating the log-likelihood for all the patterns, we sample one pattern based on the following multinomial distribution:

$$L_{ik}^{Y} \sim multinomial(1, \frac{\exp(L_{ik})}{\sum_{k=1}^{K} \exp(L_{ik})}) \qquad (8.12)$$

Therefore, in the updating of the TRN, our algorithm does an exhaustive search over all possible network patterns for each gene, and sample a specific network based the relative likelihood of all possible networks. We repeat this for each of the N genes to obtain the updated $\widehat{\mathbf{X}}$ for the next iteration.

Based on the sampled parameter values, we can derive the posterior distributions for all the unknown parameters. For example, we can obtain the inferred TRN describing the binding between the jth TF and the ith gene through the marginal posterior distribution, i.e. the proportion of samples that the value of X_{ij} is 1. These posterior probabilities can then be used to infer the presence or absence of regulation through specifying a cut-off value, e.g. 0.5, such that all the entries below this cut-off are inferred not to have regulation effect, whereas all the entries having values above this cutoff are inferred to have regulation.

8.2.3 Data analysis and simulation set-up

As our simulation model is based on the real data to be analyzed, we describe the data sources first. According to the literature, we select eight important cell cycle TFs, namely Fkh1, Fkh2, Ndd1, Mcm1, Ace2, Swi5, Mbp1, and Swi4, and based on protein-DNA interaction data reported in Lee *et al.* (2002), we obtain a binary binding matrix for these regulators and all yeast genes. The binary observation is obtained by applying a 0.05 p-value cut-off to the p-values reported by Lee and colleagues. We then remove those genes with no *in vivo* binding evidence with any of the eight TFs from the binding matrix, and further focus only on yeast cell cycle genes defined by Spellman *et al.* (1998). These steps result in a total of 295 genes to be analyzed, and the observed protein-DNA binding matrix has a dimension of 295 (genes) by 8 (TFs). For gene expression data, we use the α arrest cell cycle data with 18 time points collected by Spellman *et al.* (1998).

Now we describe our set-up used to conduct simulation experiments to evaluate the performance of our proposed procedure. In our simulation model, we need to specify (1) the true TRN, (2) true protein regulation activities, (3) false-positive and false-negative rates in the observed binding matrix, and (4) measurement errors associated with microarray data. We consider all 295 genes used in the real data analysis, and select five TFs (Fkh2, Mcm1, Ace2, Mbp1, and Swi4, which are reported to control the gene expression at the four cell cycle stages) out of the total eight in our simulations to simplify the analysis and summary. For the specification of the "true" TRN in our simulations, we use the observed binding data to represent the true TRN. As for TF activity specifications, we estimate the activities of the chosen five TFs from the linear regression model using the above "true" TRN and the expression levels of all 295 genes at each time point. The activity levels of the five TFs over 18 time points are shown in Figure 8.2. As for false-positive and false-negative rates, we vary their levels from 0.1 to 0.9 to examine their effects on statistical inference. Finally, we assess the effect of the measurement variation associated with microarray data on statistical inference. For the majority of simulations, we assume that the microarray data are collected from 18 time points as in Spellman *et al.* (2002). In one case, we vary the number of time points available to investigate the effect of the number of time points on statistical inference.

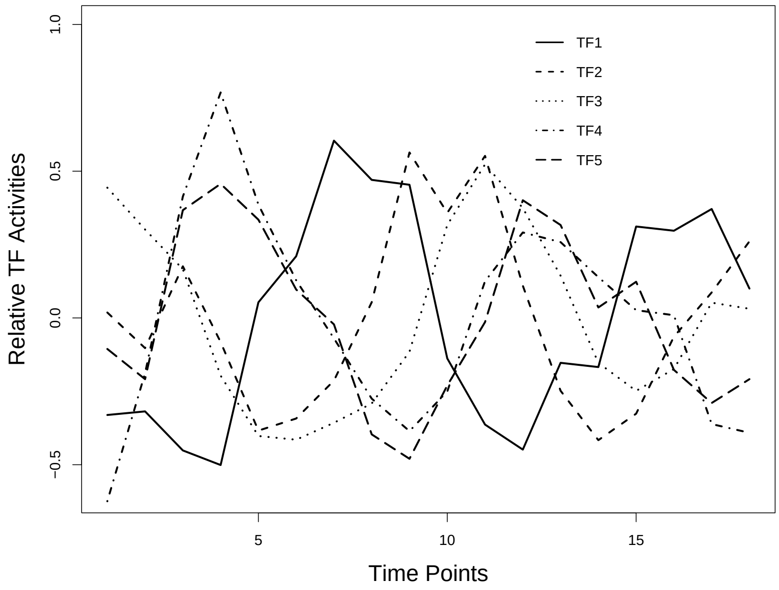


Figure 8.2: The activities of five transcription factors vary over 18 time points. Two of the five transcription factors share similar variation, which may lead to identifiable problem of the model. However, our results show that the slight difference between the TF activities prevents the problem.

8.3 Simulation Results

8.3.1 Convergence diagnosis of the MCMC procedure

Based on our simulation runs, we generally find good mixing of the proposed MCMC procedure. Both the traces of the parameter values and the autocorrelation of the parameter curves indicate that a burn-in run of 1,000 iterations out of 10,000 iterations is stable enough to obtain reliable posterior distributions. The posterior distributions of the five TF activities (β_t) and measure variations from microarrays σ_t^2 at a time point from a randomly chosen simulated data set are shown in Figure 8.3. We also investigate the effect of the initial network (covariate matrix) on MCMC results. When the measurement errors in gene expression data are low, the MCMC procedure has good convergence regardless of the initial network. In general, the observed protein-DNA binding data provide a good starting point for statistical inference.

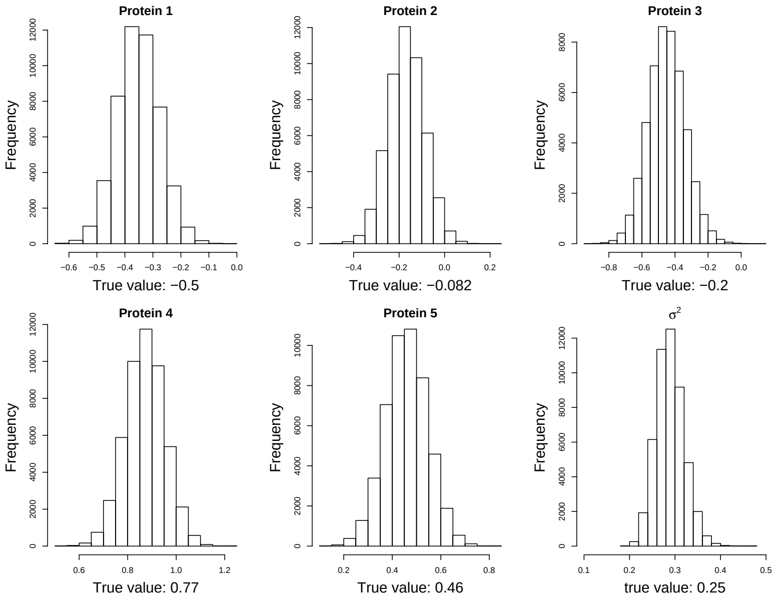


Figure 8.3: The posterior distributions for the model parameters β_t and σ_t^2 at t = 4. The standard deviations of these posterior distributions are 0.075, 0.078, 0.092, 0.077, 0.091, and 0.027, respectively.

In our model specification, there are two types of errors: the errors associated with the measured gene expression levels (responses, denoted by σ) and those associated with the observed protein-DNA binding data (denoted by p and q). In order to systematically investigate the effect of both types of errors, we consider seven pairs of p and q as (0.1,0.1), (0.2,0.2), (0.2,0.4), (0.4,0.2), (0.3,0.3), (0.4,0.4), and (0.5,0.5). For each pair of p and q values, we simulate the observed protein-DNA binding data as well as gene expression data under 22 different σ values, ranging from 0.001 to 1.5. For each specification of the $22 \times 7 = 154$ sets of parameter values, we simulate data sets consisting of protein-DNA interaction data and gene expression data. Each data set is analyzed through our proposed MCMC approach with a burn-in of 1,000 iterations and a further run of 5,000 iterations. The posterior distribution for each unknown parameter is summarized and compared to the true underlying network. We use a cut-off of 0.5 to infer the presence or absence of interactions between TFs and genes. The inferred network is then compared to the true network to calculate the proportion of false-positive and false-negative inferences for each TF-gene pair. The overall false-positive and false-negative rates are then estimated through the average of all TF-gene pairs across all the simulated data sets. The results are summarized in Figure 8.4. In Figure 8.4(a), we plot the false-positive rates for the inferred network. As can be seen from this figure, the false-positive rates for the inferred network increase as σ, p, and q increase. The false-negative rates for the inferred networks show a similar pattern. The major feature is that the information from gene expression data may significantly improve the estimation on **X**. When s is small and p and q are not too high, there is a very good chance that the true network can be recovered from the joint analysis of gene expression data and protein-DNA binding data. For example, with a 30% false-positive and 30% false-negative rates, when σ is less than 0.2, the whole network may be fully recovered. Even when σ is large, the false-positive rates in the inferred network using both binding data and gene expression data still outperform the false-positive rates in the observed protein-DNA expression data, i.e. gene expression data are not considered in the inference. The results for the false-negative rates as shown in Figure 8.4(b) show similar patterns.

8.3.2 Misspecification of the model parameters p, q, and π_X

In the results summarized above, we assume that the true values of p and q are precisely known to us. However, their exact values may not be accurately inferred. Therefore, we conduct simulation experiments to examine the performance of the proposed procedure when the values of p and q are misspecified. In this set of simulations, we simulate data from three sets of p and q values: (0.1,0.1), (0.3,0.3), and (0.2,0.4). For each simulated data set under a given set of parameter values, we perform statistical analysis under different sets of specifications for p and q, including (0.9,0.9), (0.8,0.8), (0.7,0.7), (0.6,0.6), (0.5,0.5), (0.4,0.4), (0.3,0.3), (0.2,0.2), (0.1,0.1), (0.05, 0.05), (0.01,0.01), and (0.05, 0.4). Throughout these simulations, we assume $\sigma = 0.2$. The performance of our procedure in terms of false-positive and false-negative rates is summarized in Figures 8.5(a) to 8.5(c). These results suggest that the statistical inference is robust to the misspecification of the parameters p and

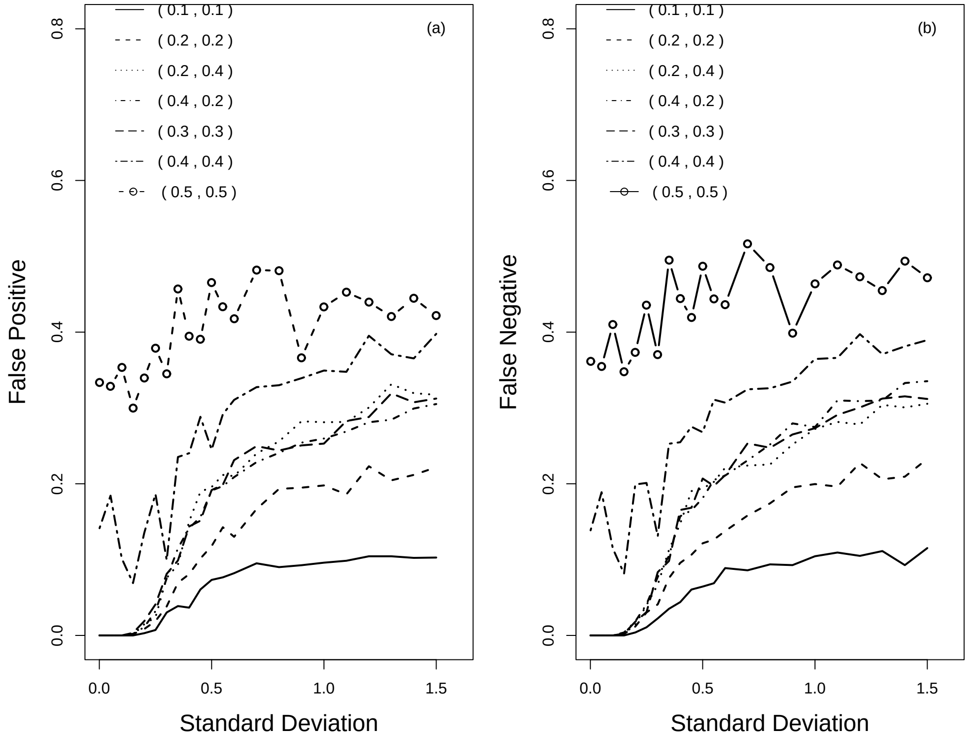


Figure 8.4: The false positive and false negative rates of the inferred network. The X-axis is the standard deviation in the gene expression data, while the Y-axis is either the false positive rate or false negative of the posterior network with respect to the true regulatory network in the cell cycle. Different lines correspond to different levels of quality of the protein-DNA binding data.

q when the specified values are not too distinct from the true parameter values. We observe similar patterns for other values of σ.

As another parameter that needs to be specified in our approach is the prior probability, π_X, that there is an interaction between a TF and a gene, we further investigate the performance of our approach when π_X is misspecified. The true value of π_X is about 0.46 ($683/(295 \times 5)$), where there are 683 regulation pairs in the protein-DNA binding data) in the given true network **X**, but we consider 0.1, 0.2, 0.3, 0.4, 0.46, 0.5, 0.6, 0.7, 0.8, and 0.9 in the specification of π_X in our analysis. The results are summarized in Figure 8.5(d). Compared to the results for p and q, the statistical inference is more sensitive to the value of π_X. However, when the specified parameter value is reasonably close to the true value, our approach yields generally robust estimates.

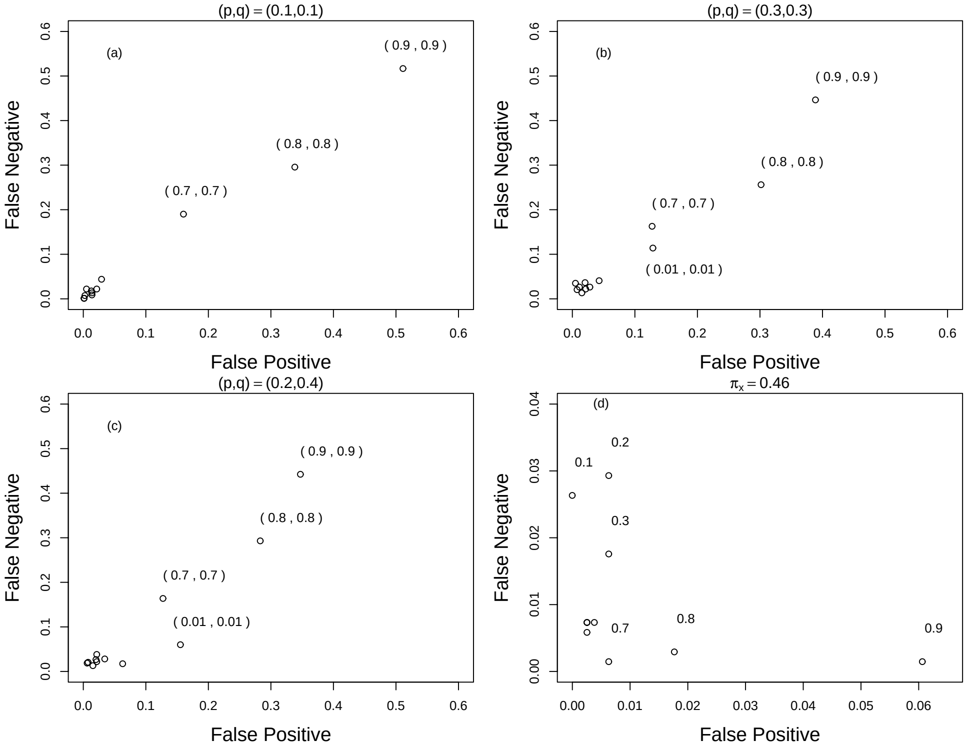


Figure 8.5: The effects of the misspecification of the model parameters p, q, and π_X on the inferred network. The standard deviation of the simulated gene expression data is 0.2. The real values of parameters (p,q) or π_X are indicated in the title of each plot. In the first three plots, the true value of π_X is 0.46, but (p,q) are specified as (0.9,0.9), (0.8,0.8), (0.7,0.7), (0.6,0.6), (0.5,0.5), (0.4,0.4), (0.3,0.3), (0.2,0.2), (0.1,0.1), (0.05, 0.05), (0.01,0.01), and (0.05, 0.4). For the last plot, the values of (p, q) are (0.1,0.1), but π_X is specified at various levels: 0.1, 0.2, 0.3, 0.4, 0.46, 0.5, 0.6, 0.7, 0.8, and 0.9.

Overall, our simulation studies suggest that misspecifications of model parameters p, q, and π_X within a reasonable range will not substantially affect the statistical inference of the true network.

8.3.3 Effect of the number of experiments used in the inference

In the above simulations, we simulate data from 18 time points and use all of them in the inference of the underlying network. In this subsection, we consider the effect of the number of time points on the inference. For this set of simulations, we simulate the protein-DNA binding data by fixing the values of p and q at 0.1, select the value of σ at 0.001, 0.2, and 0.5, and vary the number of time points used in the analysis from 1 to 18. When there is little error associated with gene expression data, i.e. σ = 0.001, the data at one time point can carry enough information to fully recover the true network. With increasing σ values, the number of time points affects the results on the inferred network (Figure 8.6). When σ is 0.2, our previous results show that there is a significant improvement of the inferred network from the binding data. As more time points are included in the analysis, we observe a more accurate inference of the underlying network. When σ is 0.5, the improvement of the inferred network from the binding data is still obvious but limited by too much noise in gene expression data.

8.4 Application to Yeast Cell Cycle Data

In this section, we apply our method to jointly analyze gene expression data from 295 genes over 18 time points (Spellman *et al.* 2002) and protein-DNA binding data of Fkh1, Fkh2, Ndd1, Mcm1, Swi5, Ace2, Mbp1, and Swi4 (Lee *et al.* 2002). We consider eight sets of model parameters for $\{p, q, \pi_X\}$: $\{$0.1, 0.1, 0.5$\}$, $\{$0.2, 0.2, 0.5$\}$, $\{$0.2, 0.1, 0.5$\}$, $\{$0.1, 0.2, 0.5$\}$, $\{$0.2, 0.2, 0.4$\}$, $\{$0.2, 0.2, 0.6$\}$, $\{$0.1, 0.1, 0.4$\}$, and $\{$0.1, 0.1, 0.6$\}$. For each set of parameter specifications, we run MCMC with a burn-in of 1,000 runs and an additional 5,000 runs to obtain the posterior distributions for the parameters of interest. The overall inference is based on the average posterior probabilities over the eight model parameter settings, which yield similar results among different settings.

The posterior distributions of the protein activities for the eight TFs and the σ at every time point are summarized in Table 8.1. The average value of σ across 18 time points is about 0.55. Based on our simulation studies, at this level of expression errors, the incorporation of gene expression data should improve the inference of TRNs.

8.5 Summary

In this article, we have developed a misclassification model to integrate gene expression data and protein-DNA binding data to infer TRNs. Compared to other models, our model (1) integrates gene expression data and protein-DNA binding data

Table 8.1: The estimates of the regulation activities of the transcription factors and σ based on our model.

Time Point	Fkh1	Fkh2	Ndd1	Mcm1	Ace2	Swi5	Mbp1	Swi4	σ
1	0.09	-0.81	-0.55	0.54	1.84	-0.29	-0.79	-0.27	0.88
	±0.13	±0.12	±0.13	±0.13	±0.14	±0.13	±0.12	±0.12	
2	-0.36	-1.00	0.24	0.28	1.18	-0.46	-0.18	-0.01	0.75
	±0.11	±0.11	±0.11	±0.11	±0.13	±0.12	±0.10	±0.11	
3	-0.53	-0.63	0.14	0.09	0.98	-0.35	1.43	0.06	0.66
	±0.10	±0.10	±0.10	±0.10	±0.14	±0.11	±0.09	±0.10	
4	-0.34	-0.31	-0.25	-0.29	0.17	-0.42	1.86	0.27	0.58
	±0.08	±0.09	±0.09	±0.08	±0.13	±0.10	±0.07	±0.08	
5	0.73	0.12	-0.62	-0.63	0.26	-0.67	0.79	0.13	0.54
	±0.07	±0.08	±0.08	±0.07	±0.09	±0.08	±0.07	±0.08	
6	0.72	0.20	-0.42	-0.49	-0.17	-0.49	0.28	-0.04	0.6
	±0.08	±0.08	±0.09	±0.08	±0.10	±0.09	±0.08	±0.08	
7	1.31	0.16	0.41	-0.61	-0.07	-0.55	-0.28	-0.28	0.53
	±0.08	±0.09	±0.08	±0.08	±0.10	±0.09	±0.08	±0.08	
8	0.44	0.18	0.61	0.01	-0.47	-0.31	-0.43	-0.57	0.44
	±0.06	±0.06	±0.06	±0.06	±0.08	±0.07	±0.06	±0.06	
9	0.17	0.09	1.03	0.58	-0.46	-0.00	-0.57	-0.74	0.5
	±0.07	±0.07	±0.07	±0.07	±0.09	±0.08	±0.07	±0.07	
10	-0.27	-0.48	0.81	0.47	-0.54	1.11	-0.39	-0.42	0.57
	±0.07	±0.08	±0.07	±0.07	±0.10	±0.08	±0.07	±0.07	
11	-0.90	0.02	-0.01	0.79	-0.32	1.23	0.13	0.08	0.75
	±0.10	±0.11	±0.11	±0.10	±0.13	±0.12	±0.10	±0.11	
12	-1.07	0.22	-0.29	0.14	-0.45	0.93	0.56	0.65	0.44
	±0.07	±0.06	±0.07	±0.06	±0.08	±0.07	±0.07	±0.06	
13	-0.20	0.44	-0.82	-0.28	-0.15	0.35	0.16	0.63	0.45
	±0.07	±0.07	±0.07	±0.06	±0.08	±0.07	±0.06	±0.06	
14	-0.35	0.42	-0.68	-0.37	-0.31	-0.08	-0.31	0.52	0.45
	±0.06	±0.07	±0.07	±0.07	±0.08	±0.07	±0.06	±0.06	
15	0.44	0.68	-0.61	-0.51	-0.08	-0.32	-0.44	0.38	0.44
	±0.06	±0.07	±0.07	±0.06	±0.08	±0.07	±0.06	±0.07	
16	0.09	0.59	-0.10	-0.16	-0.58	-0.04	-0.45	0.13	0.6
	±0.08	±0.08	±0.08	±0.08	±0.10	±0.09	±0.07	±0.08	
17	0.26	0.26	0.46	-0.02	-0.27	-0.08	-0.71	-0.26	0.62
	±0.08	±0.09	±0.09	±0.08	±0.10	±0.09	±0.07	±0.08	
18	-0.20	-0.15	0.66	0.48	-0.57	0.44	-0.63	-0.26	0.57
	±0.08	±0.09	±0.09	±0.08	±0.10	±0.10	±0.07	±0.08	

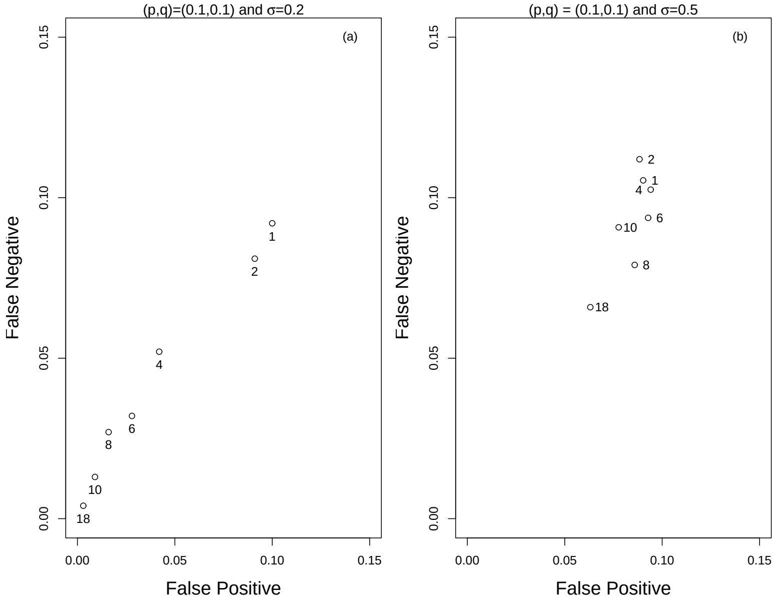


Figure 8.6: The effect of sample size on the inferred network. The number besides each symbol indicates the number of the time points used in the simulated gene expression data. The value of π_X is 0.46, and the values of other parameters are indicated in the title of each plot.

through a consistent framework, (2) considers the misclassification associated with protein-DNA binding data explicitly, and (3) consists of a flexible model structure. The systematic simulation results indicate that this model performs well in the reconstruction of the underlying networks when the misclassification associated with gene expression data and (more importantly) protein-DNA binding data are within reasonable ranges. For example, in the case of less than 30% to 40% false-positive and false-negative rates in the observed binding data, our method may significantly reduce both types of error rates in the inferred network when the standard deviation in gene expression measurements is around 0.5 or less. In all the cases, the inclusion of gene expression data leads to improved inference of the underlying network compared to that solely based on the binding data even when the measurement error in gene expression data is very high.

In this article, we have considered five TFs in simulation studies and eight TFs in the application to the yeast cell cycle data. Because there are 133 TFs in yeast protein-

DNA binding data, the inclusion of all TFs in the same model will create both statistical and computation challenges. In the context of yeast cell cycle data, protein-DNA binding data suggest that close to 20 TFs may be involved in the regulation of cell cycle genes (data not shown). The results of the application of our method to a more complete TF set and biological interpretations of the results will be reported in a separate article. From this study, we have found that (1) protein-DNA binding data can serve as a good starting point in the proposed MCMC procedure, and (2) the larger the number of gene expression data sets used, the more accurate we expect our procedure performs, especially when the gene expression data have low to moderate measurement errors. Therefore, in general, when the number of TFs increases, we hope to collect more samples on relevant gene expressions. More samples can be achieved by increasing the number of experimental conditions or the number of replicates per experimental condition or both. The advantage of increasing the number of experimental conditions is to introduce more variations of TF activity profiles so as to better infer the underlying network. However, more parameters are needed to specify the model for the additional conditions. We also need to be cautious on how to pool the experiments to infer the TRN. In this work, we have assumed a time independent TRN throughout the yeast cell cycle. This assumption may be true in this context and it allows us to pool information from across all time points. However, the TRN may differ under different conditions, and the transient behavior of the TRN needs to be taken into account when using all the microarray data. The advantage of increasing the number of replicates per condition is to reduce errors associated with measured gene expression levels at each point without introducing more model parameters. In this study, the replicates were not included in the model set-up, however, the flexible structure of our model allows an easy incorporation of such information into the model.

In our simulation studies, we have investigated the sensitivity of our method when some of the model parameters are misspecified, including the prior distribution on the network connections and our belief (measured by p and q) on the quality of protein-DNA binding data. We found that the method is not sensitive to the misspecifications of these model parameters unless the specified model parameters are drastically different from the true model parameters. In the analysis of yeast cell cycle data, we considered eight sets of model parameters and observed general agreements among results from different parameter specifications. In practice, we may take a full Bayesian approach to inferring the network through averaging inferred networks under certain prior distributions for the model parameters.

As discussed above, although we have treated the observed protein-DNA binding data as a 0-1 variable, the observed data are, in fact, continuous. In this case, our model can be modified within the measurement model framework so that the measured and true covariate values are continuous. To specify the prior distribution for the covariate values, we may use normal mixtures or more sophisticated models for the binding intensity. However, the interpretation of the model parameters will be somewhat different if the intensity levels are used because the parameter β_t cannot be simply interpreted as TF activities.

In our model set-up, we assume that all the TFs act additively to affect the transcription levels of their target genes and this linear relationship between TF activities and the normalized expression levels is a key assumption for this model. Because of the complexity in transcription regulation, such as synergistic effects among TFs, a linear model can serve as an approximation at best. Nevertheless, linear models have been used in this context by various authors (Bussemaker *et al.*, 2001; Liu *et al.*, 2002; Wang *et al.*, 2002; Liao *et al.*, 2003; Gao *et al.*, 2004). The potential departure from linearity may result from synergistic regulation effects of TFs bound to the upstream region of the same gene, and we are in the process of developing statistical approaches for analyzing nonlinear models.

To conclude, we note that our model can be extended in different ways to be more comprehensive and better represent the underlying biological mechanisms. For example, the linear form of the model can be extended to incorporate nonlinear interactions among different TFs as discussed above; the replicates per experiment can be considered into the model to improve the data quality; more prior information or more sophisticated statistical models can be used to construct the prior distribution of the network (π_X). In addition, our general framework has the potential to integrate more data types into the model, such as sequence data and mRNA decay data to further infer the transcriptional regulatory networks.

8.6 Acknowledgment

This work was supported in part by NSF grant DMS-0241160.

CHAPTER 9

An Overview of Statistical Approaches for Expression Trait Loci Mapping

Christina Kendziorski and Meng Chen
***

9.1 Introduction

Karl Sax was a pioneer in the field of quantitative trait loci (QTL) mapping. In his ground breaking 1923 paper, Sax identified a quantitative trait locus (QTL) for seed weight by associating the trait with seed color (a “marker” for which genotype information could be inferred). The next 60 years saw only a handful of similar studies, largely due to limitations imposed by the difficulty in arranging crosses with a reasonably large number of genetic markers. This changed in the 1980s following the discovery that abundant, highly polymorphic variation could be used to derive molecular markers densely spaced throughout the genome (Botstein *et al.* 1980). This advance, combined with statistical methods for QTL mapping (Lander and Botstein 1989), led to hundreds of QTL mapping studies.

A recent advance of comparable significance has been made in the area of phenotyping. With high throughput technologies now widely available, investigators today can easily measure thousands of traits for QTL mapping. Gene expression abundances measured via microarrays are particularly amenable to QTL mapping, and most scientists agree that the mapping of gene expression has the potential to impact a broad range of biological endeavors (Cox 2004; Broman 2005).

The optimism is based largely on the first expression trait loci (ETL) studies which have demonstrated utility in identifying candidate genes (Schadt *et al.* 2003; Bystrykh *et al.* 2005 Hubner *et al.* 2005), in inferring not only correlative but also causal relationships between modulator and modulated genes (Brem *et al.* 2002; Schadt *et al.* 2003; Yvert *et al.* 2003), in elucidating subclasses of clinical phenotypes (Schadt *et al.* 2003; Bystrykh *et al.* 2005; Chesler *et al.* 2005; Hubner *et al.* 2005), and perhaps most importantly, in identifying “hot spot” regions, genomic regions where multiple transcripts map (Schadt *et al.* 2003; Brem *et al.* 2002; Morley *et al.* 2004; Bystrykh *et al.* 2005; Chesler *et al.* 2005; Hubner *et al.* 2005). Hot spot regions are attractive for follow up studies as they putatively contain master regulators that affect transcripts of common function. The identification of master regulators could give critical information on mechanisms of regulation that remain poorly characterized and ultimately lead to targets of gene therapies (Cox 2004; Schadt *et al.* 2003). As a result of these successes, a number of efforts are now underway to localize the genetic basis of gene expression.

It is clear that the experimental set up in an ETL mapping study is structurally similar to a traditional QTL mapping study, but with thousands of phenotypes; and, as a result, most published studies to date have used methods developed for the QTL mapping problem in the ETL mapping setting. Lan *et al.* (2003) reduced the expression measurements to a few summary scores using a principal components analysis and then used single-trait QTL mapping methods to map the summary phenotypes. Doing so proved useful; however, transcript specific information could not be recovered. Others have used a "transcript-based" approach. In a transcript-based approach, each transcript is treated separately as a one-dimensional phenotype for QTL mapping. Single trait QTL analysis is then carried out thousands of times (once for each transcript). Notably, although adjustments are made for multiple tests across the genome, no adjustments are made for multiple tests across transcripts. This leads to a potentially serious multiple testing problem and an inflated false discovery rate (FDR).

An alternative approach recognizes the similarities between ETL mapping and the problem of identifying differentially expressed (DE) transcripts in a standard microarray experiment. By grouping animals with similar marker genotypes, the ETL mapping problem at a particular marker reduces to identifying DE transcripts across the genotype groups. Any method developed for identifying DE transcripts could be applied. Similar to the transcript-based approach, this "marker-based" approach is also subject to inflated FDR as here multiplicities across markers are not accounted for. For some labs, an inflated FDR is tolerable as many genes can be tested quickly for certain properties and discarded if found to be false positives. However, for many labs, validation tests are prohibitively expensive and statistical methods that control error rates across both markers and transcripts are needed. Kendziorski *et al.* (2004) proposed such an approach, the mixture over markers (MOM) model.

In this chapter, we will review transcript-based approaches, marker-based approaches, and the MOM model approach to ETL mapping. The advantages and disadvantages of these approaches are discussed in Sections 9.2 and 9.4. Utility is evaluated using simulated data and data from two case studies (Section 9.3).

9.2 ETL Mapping Data and Methods

9.2.1 Data

The general data collected in an ETL mapping experiment consists minimally of a genetic map, marker genotypes, and microarray data (phenotypes) collected on a set of individuals. A genetic marker is a region of the genome of known, or estimated, location. These locations make up the genetic map. At each marker, genotypes are obtained. ETL mapping studies take place in both human and experimental populations. We focus here on the latter. For these populations, the number of possible marker genotypes is relatively small.

Studies with experimental populations most often involve arranging a cross between two inbred strains differing substantially in some trait of interest to produce F1 offspring. Segregating progeny are then typically derived from a B1 backcross (F1 x Parent) or an F_2 intercross (F1 x F1). Repeated intercrossing ($F_n x F_n$) can also be done to generate so-called recombinant inbred (RI) lines. For simplicity of notation, we focus on a backcross population. This is not required and is relaxed in the simulation and case studies sections. Consider two inbred parental populations P_1 and P_2, genotyped as AA and aa, respectively, at M markers. The offspring of the first generation (F_1) have genotype Aa at each marker (allele A from parent P_1 and a from parent P_2). In a backcross, the F_1 offspring are crossed back to a parental line, say P_1 resulting in a population with genotypes AA or Aa at a given marker. We denote AA by 0 and Aa by 1.

For each member of the backcross population, phenotypes are collected via microarrays. For the k^{th} animal, let $y_{t,k}$ denote the expression level for transcript t and $g_{m,k}$ denote the genotype at marker m; $t = 1, 2, \ldots, T$ and $k = 1, 2, \ldots, n$. To avoid confusion when referring to genes on a genetic map and gene expression levels measured on a microarray (where the physical location of the gene is often not known), when referring to the former, we use the term gene; when referring to the latter, we use transcript or trait.

Most questions addressed in an ETL mapping study rely on the ability to identify a list of significant linkages between transcripts and markers. To be precise, a transcript t is linked to marker m if $\mu_{t,0} \neq \mu_{t,1}$, where $\mu_{t,0(1)}$ denotes the latent mean level of expression of transcript t for the population of animals with genotype $0(1)$ at marker m. Suppose observations $y_{t,k}$ have density $f_{obs}(y_{t,k}|\mu_{t,g_{m,k}}, \theta)$ where θ denotes any remaining unknown parameters. Assuming independence across animals, under the null hypothesis of no linkage, the data is governed by $\prod_{k=1}^{n} f_{obs}(y_{t,k}|\mu_{t,0} = \mu_{t,1}, \theta)$; and under the alternative, $\prod_{k=1}^{n} \left[f_{obs}(y_{t,k}|\mu_{t,0}, \theta)\right]^{1-g_{m,k}} \left[f_{obs}(y_{t,k}|\mu_{t,1}, \theta)\right]^{g_{m,k}}$. As discussed below, a main difference between the transcript-based (TB) and marker-based (MB) approaches arises from different assumptions regarding the latent means.

9.2.2 Transcript Based Approach

A TB approach refers generally to the repeated application of any single phenotype QTL mapping method to each mRNA transcript, with locations identified as important if the test statistic of interest exceeds some critical value. The LOD score

$$\log_{10}\left(\frac{\prod_{k=1}^{n} f_{obs}(y_{t,k}|\hat{\mu}_{t,0}, \hat{\mu}_{t,1}, \hat{\theta})}{\prod_{k=1}^{n} f_{obs}(y_{t,k}|\hat{\mu}, \hat{\theta})}\right)$$

is often used as the statistic measuring evidence in favor of linkage, where $(\,\hat{\cdot}\,)$ denotes the maximum likelihood estimate of the associated parameter(s) and μ denotes the mean common across genotype groups (Lander and Botstein 1989). Critical values that adjust for multiplicities across genome locations can be obtained theoretically (Dupuis and Siegmund 1999) or via permutations (Churchill and Doerge 1994).

The specific TB approach considered here assumes a Gaussian density for f_{obs} with critical values determined by the formulas given in Dupuis and Siegmund (1999). We consider the output from this approach at markers and refer to this as a TB marker regression (TB-MR) approach. The restriction to consider output only at markers is done to facilitate comparisons with MB methods, discussed below. For TB-MR, the genome wide type I error rate per transcript is controlled at 5% (Dupuis and Siegmund 1999).

9.2.3 Marker Based Approaches

To identify transcripts significantly linked to genomic locations, instead of testing each transcript for significant linkage across markers, one could test at each marker for significant linkage across transcripts. This amounts to identifying DE transcripts at each marker, with groups determined by marker genotypes. The MB approach refers generally to the repeated application, at each marker, of any method for identifying DE transcripts. In this setting, a number of approaches could be used. We here consider four.

The first is an empirical Bayes approach, *EBarrays*, with the log-Normal Normal model (LNN) described in detail in Kendziorski *et al.* (2003; 2004). This approach calculates the posterior probability of differential expression for every transcript. Thresholds can be chosen to control the expected posterior FDR across transcripts. For example, by specifying the threshold to be the smallest posterior probability such that the average posterior probability of all transcripts exceeding the threshold is larger than $1 - \alpha$, the posterior expected FDR is controlled at $\alpha \cdot 100\%$ (Newton *et al*. 2004). This marker-based empirical Bayes approach will be referred to as MB-EB. As in TB-MR, the LNN model assumes a Gaussian density for f_{obs}.

The second marker-based approach consists of obtaining p-values from a Student t-test followed by p-value adjustment; and the last two approaches consider moderated t-statistics followed by p-value adjustment. The details of the moderated statistic construction are given in Smyth et al. (2004) and Tusher et al. (2003), respectively. Adjustment for these last three methods is done using q-values to control the overall false discovery rate (FDR). In particular, to control the FDR at α, transcripts with q-values $<= \alpha$ are considered significant (Storey and Tibshirani 2003). MB-Q, MB-LIMMA, and MB-SAM will denote the three marker-based approaches, respectively.

9.2.4 Other Approaches

Although the TB and MB approaches are in many ways fundamentally different, they share an important flaw. Separate tests are conducted for each transcript-marker pair, and each measures evidence that the transcript maps to that marker relative to evidence that it maps nowhere. Since a transcript can map to any of many marker locations, the evidence that a transcript maps to a particular marker should not be judged relative only to the possibility that it maps nowhere, but rather relative to the possibility that it maps nowhere *or* to some other marker. This idea motivates the mixture over markers (MOM) model (Kendziorski *et al.* 2004). Briefly, MOM assumes a transcript t maps nowhere with probability p_0 or to marker m with probability p_m where $p_0 + \sum_{m=1}^{M} p_m = 1$ and M denotes the total number of markers. The marginal distribution of the data $\mathbf{y}_t$ is then given by

$$p_0 f_0(\mathbf{y}_t) + \sum_{m=1}^{M} p_m f_m(\mathbf{y}_t) \tag{9.1}$$

where f_m describes the distribution of data if transcript t maps to marker m (f_0 describes the data for non-mapping transcripts). The component densities are predictive distributions that can be derived under different parametric assumptions. For comparison, we take Gaussian observation components for the log measurements with Normal priors on the latent expression levels.

9.3 Evaluation of ETL Mapping Methods

The methods discussed above were evaluated using simulated data and data from two case studies. The simulations are in no way designed to capture the many complexities of ETL mapping data. Nevertheless, they do provide some insight into operating characteristics of each of the approaches. The first case study concerns an experiment in yeast and the second a study of diabetes in mouse.

9.3.1 Simulation

Recall that for a backcross population, a subject has one of two genotypes (AA or Aa) at each marker locus. For an F_2, three genotypes are possible (AA, Aa, or aa) and, as a result, a given transcript may be equivalently expressed (EE) or may be in any one of 4 DE patterns ($AA|Aa, aa$; $AA, Aa|aa$; $AA, aa|Aa$; $AA|Aa|aa$). Here | denotes inequality among the latent genotype group means. We performed a simulation of an F_2 population in which pattern membership was determined by a multinomial where the expected proportion of transcripts in each DE pattern was specified at 3%, 3%, 1% and 3%, respectively (1% is used for the pattern that is least biologically plausible).

Care was taken to protect against biasing the results in favor of any of the methods considered. The details are given in Kendziorski *et al.* (2004). In short, a major difference among methods lies in the estimation of transcript variance ${\sigma_t}^2$. To set the variance for a simulated transcript t, we used the posterior mean of ${\sigma_t}^2$, given by $\frac{\sum_{k=1}^{n}(y_{t,k}-\bar{y}_{t,\cdot})^2+\nu_0{\sigma_0}^2}{\nu_0+n-2}$ (derived assuming the transcript specific variance is distributed as scaled inverse chi-square: ${\sigma_t}^2 \sim \text{Inv}\chi^2\left(\nu_0, {\sigma_0}^2\right)$). As $\nu_0 \to 0$, the posterior mean approaches $\frac{(n-1)s^2}{n-2} \approx s^2$, the transcript specific sample variance, which is the naive estimate of ${\sigma_t}^2$ for an EE transcript under TB-MR assumptions. Data simulated with small ν_0 is therefore consistent with assumptions made in TB-MR. As $\nu_0 \to \infty$, the posterior mean approaches a constant value ${\sigma_0}^2$, which is assumed in MB-EB (note that this assumption implies a constant coefficient of variation on the raw gene expression scale). By varying ν_0, operating characteristics could be evaluated without biasing the results in favor of one method. Data simulated by this empirical method had marginal distributions that were virtually indistinguishable from the observed data.

We consider a single ETL simulation with 100 animals and 2 chromosomes. Marker genotype data was obtained from chromosomes 2 and 3 of the F_2 data described in the next section. Chromosome 2 (3) contained 17 (6) markers with an average intermarker distance of 7.6 (17.7) cM. An ETL at marker 5 on chromosome 2 was simulated; no ETL was simulated on chromosome 3. Seven sets of simulations were obtained for ν_0 between 5^{-5} and 5^5 (ν_0 for the actual F_2 data was estimated near 5). For each value of ν_0, 20 simulated data sets were generated. At each fixed ν_0, the profile marginal MLE was obtained for ${\sigma_0}^2$.

FDR gives the proportion of transcripts identified incorrectly as mapping to chromosome 2; i.e. they were EE or they were DE but mapped outside the region flanking the true ETL. Table 1 reports the operating characteristics. FDR is well above the target level of 0.05 for most methods and most values of ν_0. MOM is the only approaches capable of FDR control in this simple simulation setting. Power measures the ability to identify the DE transcripts exactly at marker 5 or either of the flanking markers which are 16.5 and 5.8 cM away, respectively. There is little variation in power across ν_0. MB-Q is the most powerful method, followed by TB-MR, MB-EB, and MOM. The difference in power between MOM and the others is statistically significant, but perhaps not *practically* significant as power is still near 80%.

As shown in Table 1, the results from MB-Q, MB-LIMMA and MB-SAM were very similar, most likely because the relatively large sample size (100 animals) yields statistics in MB-LIMMA and MB-SAM that have been “moderated” only slightly. A similar result was reported in Smyth *et al.* (2004), where an experiment with 16 animals was considered. For this reason, only results for MB-Q will be discussed hereinafter.

9.3.2 Case Studies

To further compare these approaches, we consider ETL mapping data from the yeast experiment described in Brem *et al.* (2002). It is structured as a backcross between a standard laboratory strain (BY) and a wild isolate from a California vineyard (RM). There are 6215 transcripts and 3312 markers. With only 40 segregants in the cross, recombinants are limited. We removed pairs of markers with fewer than 10 recombinants in between leaving 88 markers.

Brem *et al.* (2002) identified 8 regions enriched for linkage across the genome. Many transcripts in these hot spot regions have been at least partly validated using independent experiments. As noted in the Introduction, these regions are of much interest as they may contain a master regulator responsible for the control of transcripts sharing common biological function. A statistical test for enrichment of common function can done via *GOHyperG* in Bioconductor (Bioconductor Core Team 2004)). *GOHyperG* uses data from Gene Ontology (GO), where transcripts are categorized at varying levels of biological detail (the three broadest levels are molecular function, cellular component, and biological process - there are many subcategories within each). For a given set of mapping transcripts and a given function, a hypergeometric calculation is performed to test for enrichment of that function across the transcripts.

Interpretation of resulting p-values is not straightforward due to the many dependent hypotheses tested. Furthermore, the hypergeometric calculation tends to result in small p-values when GO nodes with few transcripts are considered. For these reasons, it has been suggested that one only consider interesting small p-values obtained from a relatively large set of transcripts (> 10) (Gentleman, 2005). Applying this criterion to the results from Brem *et al.* (2002) gives 5 regions, shown in Table 2.

Table 3 shows information similar to Table 2, for the top 5 regions (5 regions with the largest number of mapping transcripts) identified by MOM, TB-MR, and MB-Q. We see that TB-MR identifies 3 of the 5 regions identified by Brem *et al.* (2002) on chromosomes 3, 12, and 14. The location identified by Brem *et al.* (2002) on chromosome 2 is missed by TB-MR; and the location identified by TB-MR on chromosome 9 is not found using any other method and shows little evidence for enrichment of common function. This is likely a false positive. Similar results are obtained from MB-Q, with 3 of the 5 regions identified, and one potentially spurious identification on chromosome 8.

The MOM model performs better: 4 of the 5 regions identified by Brem *et al.* (2002) (on chromosomes 2, 3, 12, and 14) are also identified by MOM. The one region identified by Brem *et al.* (2002) but not MOM is a second location on chromosome 3. There are not enough markers considered (using the selected 88) to distinguish between these two regions using MOM. In addition to improved hot spot localization, MOM is generally more sensitive than the other methods. We suspect that the increased number of identifications made by MOM are not false discoveries as the additional transcripts maintain evidence for enrichment of the common function.

It is insightful to check the results from these approaches when control of particular error rates is not used for hot spot identification. For example, instead of defining hot spots in terms of the number of mapping transcripts (which depends on particular thresholds to generate binary calls), one could consider average evidence (across transcripts) of mapping at each location (average LOD, average posterior probability, or the average of 1 - q-value). Given hot spots identified in this way, one can simply rank transcripts at each hot spot by LOD score, posterior probability, or 1-q-value and then consider the top N transcripts for some N. In terms of regions identified and tests for enrichment of common function, we found results similar to those shown in Table 3 for N of 50 and 100.

The ETL mapping approaches were also evaluated using data from a study of diabetes in mouse. For details on the experiment, see Kendziorski *et al.* (2004). Briefly, it is well known that the *ob* mutation in the C57BL/6J mouse background (B6-*ob/ob*) causes obesity, but only mild and transient diabetes (Coleman and Hummel, 1973), while the same mutation in the BTBR genetic background (BTBR-*ob/ob*) causes severe type 2 diabetes (Stoehr *et al.* 2000). To gain insight into the genetic basis of these differences, a (B6 x BTBR)F_2-cross was generated yielding 110 animals. Selective phenotyping (Jin *et al.* 2004) was employed to identify 60 F_2 *ob/ob* mice. For each of the 60 mice, liver tissue was isolated and 45,265 mRNA abundance traits were collected at 10 weeks of age using Affymetrix Gene Chips (MOE430A,B). The probe level data was processed using Robust Multi-array Average (RMA) to give a single, normalized, background corrected summary score of expression for each transcript (Irizarry *et al.* 2003). Low abundance transcripts, defined as transcripts with average expression level below the tenth percentile, were removed leaving 40,738 traits. Genotypes for 145 markers were also obtained (over 90% of the animals provided genotype data at any given marker).

Each method was applied to identify ETL. Hot spot regions are shown in the left panel of Figure 2. The first marker, D2Mit241, is adjacent to D2Mit9, which has recently been identified as an obesity modifier locus (Stoehr *et al.* 2004). Two additional regions identified by 4 of the 5 methods (on chromosomes 4 and 10) are not yet known to be involved in diabetes although we note that the region identified on chromosome 4 has been implicated in other analyses done in the Attie lab. The two regions identified by MOM alone on chromosomes 5 and 8 have been identified by other groups in earlier studies: D5Mit1 is a location known to affect triglyceride levels (Colinayo *et al.* 2003) and D8Mit249 is the marker on our map closest to the “fat” gene which is known to affect both diabetes and obesity (Naggert *et al.* 1995). This provides some evidence for the MOM approach, but much more biological validation is required.

It is interesting to note that the agreement between FDR controlled and rank based inferences observed for the yeast study was not observed here. Figure 2 (right panel) gives results from the diabetes case study using the binary scores. As shown, there is much less agreement across methods when the binary scores are used. We expect there are conditions under which averaging evidence across transcripts is more advantageous than reducing to a binary score (and vice versa). This is currently an area under investigation.

9.4 Discussion

The field of QTL mapping was reignited in the 1980’s by advances that allowed for the relatively easy identification of genetic markers and their genotypes. Today, with major developments in high throughput technologies, a similar advance has taken place that allows for measurement of thousands of phenotypes. The number and nature of these phenotypes are what distinguish QTL from ETL mapping. In fact, ETL mapping is exactly traditional QTL mapping, but with thousands of expression traits considered as phenotypes. The simplicity with which this difference can be stated perhaps obscures the resulting challenges posed for the statistical analysis of ETL data.

When faced with just about any statistical problem, it is often best to first consider methods that are currently available. This was done for ETL mapping. The earliest ETL papers applied traditional QTL mapping methods to each transcript in isolation. Doing so does not account for multiple tests across transcripts; and we found this to have a real impact on increased FDR even in very simplified simulation settings. For some labs, an inflated FDR is tolerable as many genes can be tested quickly for certain properties and discarded if found to be false positives. However, for many labs, such tests are prohibitively expensive and more appropriate statistical methods are needed.

The most recent ETL studies have made attempts at adjusting for multiplicities across both markers and transcripts using a two stage approach (Chesler *et al.* 2005; Hubner *et al.* 2005). The first stage obtains a single p-value for each transcript that is adjusted for multiple tests across markers; stage two controls the FDR across transcripts by calculating q-values from these p-values. With this approach, mapping transcripts are identified, along with the single most likely location to which these transcripts map. Preliminary simulation results (not shown) show very low power if attempts are made to control the FDR at 5%. This is consistent with the results reported in Chesler *et al.* (2005), where an FDR cutoff of 25% is used so that 101 transcripts can be identified (out of $12,422$ total transcripts).

Our general conclusion is that a clever application of statistical methods developed in the context of QTL mapping and/or multiple testing is not sufficient to address the complexities of the ETL mapping problem. As a result, we continue to investigate MOM. The MOM approach was designed explicitly to address the ETL mapping question. Operating characteristics evaluated via simulations as well as results from case studies are encouraging. Another nice feature of the MOM framework is that it can be extended to account for interval and multiple ETL mapping. This work is underway.

In summary, much more work is required before the analysis of ETL data becomes routine. In practice, we suggest an investigator apply a number of tools and focus initially on genomic locations at which most methods agree (such as the 4 regions shown in the left panel of Figure 2), keeping in mind that assumptions across different methods are often very similar and therefore by no means are the results of different methods independent confirmations. Statisticians can contribute to the ETL mapping effort by method development, evaluation, and validation; and by carefully considering those genomic regions that *do not* agree across methods. Such regions can provide valuable insights so that specific conditions under which different methods work best can be identified. Advances in each area and communication between the two are required to maximize the amount of information that can be derived from ETL mapping studies.

9.5 Figures and Tables

Figure 2: Evidence of linkage for each approach (LOD for TB-MR, posterior probability for MB-EB and MOM, and 1 - q-value for MB-Q). TB-MR, MB-EB, MOM, and MB-Q are colored by blue, red, purple and green, respectively. The left panel averages evidence of mapping over transcripts; the right panel gives normalized totals of mapping transcripts based on thresholding to control FDR. The 5 markers with the strongest evidence of mapping transcripts are indicated by triangles for each method.

2*OC	2*Method	ν_0						
		5^{-5}	5^{-3}	5^{-1}	5^{0}	5^{1}	5^{3}	5^{5}
4*FDR	TB-MR	0.286	0.286	0.293	0.285	0.286	0.28	0.301
	MB-EB	0.282	0.281	0.285	0.279	0.269	0.117	0.034
	MB-Q	0.24	0.246	0.246	0.24	0.245	0.23	0.226
	MB-LIMMA	0.238	0.236	0.232	0.237	0.235	0.237	0.229
	MB-SAM	0.233	0.238	0.235	0.232	0.238	0.236	0.221
	MOM	0.038	0.041	0.046	0.037	0.036	0.005	0.002
4*Power	TB-MR	0.884	0.886	0.887	0.886	0.889	0.919	0.868
	MB-EB	0.820	0.817	0.815	0.823	0.833	0.895	0.837
	MB-Q	0.911	0.912	0.913	0.912	0.917	0.949	0.918
	MB-LIMMA	0.900	0.910	0.909	0.900	0.914	0.935	0.899
	MB-SAM	0.897	0.908	0.906	0.898	0.913	0.933	0.899
	MOM	0.848	0.851	0.853	0.850	0.856	0.860	0.811

Table 1: Average operating characteristics (OCs) for TB-MR, MB-EB, MB-Q, MB-LIMMA, MB-SAM, and MOM. Averages are calculated over 20 data sets; standard errors were less than 0.005. OC definitions and details of the simulation are given in the text (see Section 9.3.1).

Chromosome (BP)	Number of Mapping Transcripts	Common Function	p-value
2(550)	18	Cell Separation	$\sim 10^{-7}$
3(90)	21	Leucine Biosynthesis	$\sim 10^{-7}$
3(190)	28	Mating	$\sim 10^{-10}$
12(670)	28	Fatty Acid Metabolism	$\sim 10^{-7}$
14(490)	94	Mitochondrial Induction	$\sim 10^{-6}$

Table 2: Results reproduced from Brem *et al.* (2002). Chromosomal locations, number of transcripts mapping to each region, biological function common to these transcripts, and p-values from GoHyperG are shown. BP gives the number of bases ($/1000$) from the 5’ end of the chromosome.

Method	Chromosome	Number of	Common Function	p-value
	(BP)	Mapping Transcripts		
TB-MR	3(75)	29	Leucine Biosynthesis	$\sim 10^{-6}$
TB-MR	12(607)	21	Fatty Acid Metabolism	$\sim 10^{-7}$
TB-MR	14(502)	644	Mitochondrial Induction	$\sim 10^{-6}$
TB-MR	15(1)	27	Glucan Metabolism	> 0.2
TB-MR	9(99)	19	Iron Transport	0.03
MOM	2(602)	56	Cell Separation	$\sim 10^{-5}$
MOM	3(75)	56	Leucine Biosynthesis	$\sim 10^{-6}$
MOM	12(872)	55	Fatty Acid Metabolism	$\sim 10^{-8}$
MOM	14(502)	94	Mitochondrial Induction	$\sim 10^{-6}$
MOM	15(1)	288	Glucan Metabolism	$\sim 10^{-3}$
MB-Q	3(75)	31	Leucine Biosynthesis	$\sim 10^{-5}$
MB-Q	12(607)	36	Fatty Acid Metabolism	$\sim 10^{-7}$
MB-Q	14(502)	78	Mitochondrial Induction	$\sim 10^{-5}$
MB-Q	15(1)	29	Glucan Metabolism	10^{-1}
MB-Q	8(80)	81	Response to Pheromone	0.001

Table 3: Top 5 regions identified by TB-MR, TB-Q, and MOM. For each method and region, chromosomal locations, number of transcripts mapping to each region, biological function common to these transcripts, and p-values from GoHyperG are shown. BP gives the number of bases ($/1000$) from the 5' end of the chromosome. Note that the region identified by all methods on chromosome 15 is one of the 8 originally identified by Brem *et al.* (2002). It was excluded when constructing the list of 5 due to a relatively large p-value (0.02). It is difficult to judge whether or not this region is a false positive. Considering all methods point to this region, perhaps it is not.

CHAPTER 10

Combining genomic data in human studies

Debashis Ghosh, Daniel Rhodes and Arul Chinnaiyan
University of Michigan

10.1 Introduction

With the development of technology that has allowed for the high-throughput miniaturization of standard biochemical assays, it has become possible to globally monitor the biochemical activity of populations of cells. This has led to the emergence of cDNA microarrays in medical and scientific research and has allowed for large-scale transcriptional characterization. It should also be noted that the microarray technology would have limited ability without the existence of large-scale genome sequencing projects, such as the Human Genome Project (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001). Having such sequence data available allows for the characterization of the probes on the microarray. In this chapter, we will be using the term “genomic data” to generically refer to any genetic data that is generated using large scale technologies.

While transcript mRNA microarrays have received much attention in the literature, there has been work on other types of microarrays. Examples include chromatin-immunoprecipitation (ChIP) microarrays, which measure transcription factor-DNA binding expression (Lee et al., 2002) and methylation microarrays (Yan et al., 2001), which assess DNA methylation on a global scale. In addition, there has also been much attention on high-throughput assays that measure protein-protein interactions, such as yeast two-hybrid systems (**?**). Because of all the large-scale data that is being generated, there is much interest in attempting to integrate the data to provide a more complete understanding of the biological mechanisms that are at play. This type of analysis has been given the name “systems biology” in the bioinformatics literature (Ideker et al., 2001).

For the statistician, this area brings many interesting and challenging problems. While the term “meta-analysis” is familiar among most statisticians (Normand, 1999), the term here takes a very different meaning. The situation statisticians are familiar with involves attempting to combine information from relatively homogeneous data structures from multiple similar experiments. However, in much of the genomic area, the issue is one of trying to combine relatively inhomogeneous data structures from multiple experiments that may or may not be similar.

Another complication is that data availability depends on the type of organism studied. In this chapter, we focus on data from human studies. Thus, protein-protein interaction data from two-hybrid experiments are not currently available for humans. We will talk about approaches for combining genomic data in human studies, primarily focusing on methods developed in the cancer setting. Some familiarity with microarray technologies is assumed; the reader is referred to the first and second volumes of *The Chipping Forecast*, a supplement to the journal *Nature Genetics* that has been made publicly available online (Chipping Forecast, 1999, 2002). Our goal here is to seek to outline the major issues involved in such analyses and describe some solutions that have been proposed. It is not our intent to provide an up-to-the date listing of all methodologies that have been used, as the literature is constantly changing. Given the dynamic nature of the field, an important component will be benchmarking of methods to see which should be used in practice.

10.2 Genomic data integration in cancer

10.2.1 Goals

Our group has focused primarily on the analysis of genomic data in cancer studies. There are two broad goals of this research. One is the discovery of new biomarkers that might be used potentially as screening tests or to better predict patient prognosis. Examples of potential promising biomarkers found using gene expression technology include enhancer of zeste homolog 2 (EZH2) in prostate cancer (Varambally et al., 2002). In this study, the transcript mRNA expression EZH2 gene transcript was found to be highly expressed in metastatic prostate cancer. A key point to make at this stage, which we will address later, is that mRNA expression does not necessarily perfectly correlate with protein expression. In terms of diseases, the action is happening at the protein level. In protein validation studies done by Varambally et al. (2002)), the EZH2 protein was also found to be highly expressed in metastatic prostate cancer. Another example of a potential biomarker found using genomic data technologies is prostasin in ovarian cancer (Mok et al., 2001). In that study, the authors reported a sensitivity of 92% and a specificity of 94% for discriminating ovarian cancer cases from controls using validation by ELISA of serum. Thus, prostasin might serve as a potential biomarker for early detection of ovarian cancer.

The second is to better understand the biology of the disease. In the past, cancer was thought of as a heterogeneous collection of diseases. However, a more integrative view of the disease is currently being put forward by many researchers; this view was summarized eloquently in a review article by Hanahan and Weinberg (2000). According to their paradigm, there are six principles that underlie tumorigenesis (the initiation and development of a tumor); equivalently, for a cancer to develop, it must acquire six “hallmark capabilities”:

- Self-sufficiency in growth signals;
- Insensitivity to anti-growth signals;
- Evading apoptosis (cell death);
- Limitless replicative potential;
- Sustained angiogenesis;
- Tissue invasion and metastasis.

With the current availability of large-scale genomic data, we can address the Hanahan and Weinberg model in two ways. First, we can analyze the data to see the relative contributions of the six “hallmark capabilities.” Second, we can use genomic data to further refine and identify the pathways that comprise each of the individual hallmark capabilities described above.

10.3 Combining data from related technologies: cDNA microarrays

The statistical problem closest in spirit to classical meta-analysis involves trying to combine multiple datasets in which the same type of cellular activity was assessed. As an example here, we consider multiple microarray studies in which the same comparison was considered, namely cancer versus normal.

There are several issues that must be considered when attempting such an analysis. First, one must consider the problem of study-specific artifacts, such as sampling bias, variations in experimental protocols and differences in laser scanners. However, there are two bigger issues in the analysis of such data. The first is that of matching genes from two studies. This is where the availability of large-scale genomic data figures in hugely. Each spot on a microarray corresponds to a DNA sequence. What one can do is to match up each spot to a putative gene in the Unigene Database, which is a collection of clusters of orthologous genes. The Unigene link can then be used to identify common genes across multiple datasets. Such a task can be done for Affymetrix chips from their website (http://www.netaffx.com/) or for two-color cDNA microarrays using the SOURCE tool at Stanford (Diehn et al., 2003).

While such a mapping is useful, there still might be errors that remain. A more challenging issue involves the fact that the numbers from different microarray platforms represent different things. That is, an expression value of 20 from a cDNA two-color microarray is much different from an expression value of 20 measured on an Affymetrix array. Another technique that has proven to be useful as a filtering device to enhance comparability across arrays of different platforms is known as the integrative correlation coefficient or correlation of correlation coefficients (Lee et al., 2002; Parmigiani et al., 2004). The idea underlying this method is that while raw expression values vary from study to study, the intergene correlations do not vary as much.

Thus, one would consider combining genes that have similar intergene correlations across the studies.

In terms of meta-analysis methods put forward, many have been based on the fact that the standardized effect size is combinable across studies. This is the approach advocated by Parmigiani et al. (2004) after filtering based on the integrative correlation coefficient. In Rhodes et al. (2002), the t-statistic was transformed into a p-value, a transformation of which was combined across multiple studies. By contrast, in Ghosh et al. (2003), the t-statistic was combined directly. An approach that was more Bayesian in nature was taken by Wang et al. (2004), in which expression values from one study were used to develop a prior distribution for the standardized effect size; data from the remaining studies were used to generate posterior distributions. A fully hierarchical approach was taken by Choi et al. (2003), who then used Markov Chain Monte Carlo methods to sample from the posterior distributions. It should be noted that all of these methods make the assumption that a standardized effect size can be estimated directly for each individual study.

Another approach more in line with classification or supervised learning analyses is to build a classifier or find a gene expression signature on one dataset and to see how well it predicts in an independent microarray dataset. Such approaches were taken by Beer et al. (2002), **?** and Jiang et al. (2004). An alternative method using hierarchical clustering, which is an unsupervised learning procedure, was taken by Sorlie et al. (2003). They found a gene expression signature that defined molecular subtypes in breast cancer; they found through interrogation of other datasets that the subtypes were present there as well. Given the increasing availability of publicly available large-scale gene expression datasets, it is increasingly important that results found by one investigator on a particular dataset be validated using other datasets as well.

A large-scale comprehensive meta-analysis was performed by Rhodes et al. (2004a). They performed a meta-analysis of 40 independent datasets ($>$3,700 array experiments) across ***??? tissue sites. They found a universal profile of 67 genes that could differentiate cancer versus noncancer tissue for a variety of cancers. In addition, they determined 36 cancer-specific signatures for determining a tissue-specific cancer. The signatures also demonstrated good discrimination performance on three independent datasets.

A more sophisticated method for meta-analysis was put forward by Shen et al. (2004), based on an idea of Parmigiani et al. (2002). Namely, the idea is that for a given gene from a given sample in a given study, it is either over-, under- or non-differentially expressed with respect to a baseline cohort of genes. Each of the three states defines a latent category, which induces a mixture model for gene expression values. The latent states of over-, under- or non-differentially expressed are inferred using a Markov Chain Monte Carlo sampling algorithm. The estimated probabilities of the latent states are then transformed to define a “probability of expression,” which is then used as input for a meta-analysis.

Much of the meta-analysis methods have studied differential expression across multiple studies. A notable exception is the study by Lee et al. (2004), in which intergene correlations across multiple studies was considered. The authors sought pairs of genes that were consistently coexpressed across several datasets. As will be described in the next section, such coexpression is the first step needed in building gene regulatory networks.

Because of the fact that information on thousands of genes are typically considered, there is an inherent multiple testing problem. A popular method for calibrating results in this setting has been through use of the false discovery rate (Benjamini and Hochberg, 1995a). The false discovery rate, or FDR, is roughly defined as the expected proportion of falsely rejected null hypotheses among the set of rejected null hypotheses. A smaller FDR indicates that there are more "real" discoveries found by the investigator. This can be visualized by considering the cross-classification of n single-gene hypotheses by whether they are rejected based on the data and their true status (i.e. null hypothesis is true or alternative hypothesis is true). Such a table is given here:

Table 10.1: Outcomes of n tests of hypotheses

	Accept	Reject	Total
True Null	U	V	n_0
True Alternative	T	S	n_1
	W	Q	n

The definition of false discovery rate (FDR) as put forward by Benjamini and Hochberg (1995a) is

$$FDR \equiv E\left[\frac{V}{Q} \mid Q > 0\right] P(Q > 0).$$

The conditioning on the event $[Q > 0]$ is needed because the fraction V/Q is not well-defined when $Q = 0$. Methods for controlling the false discovery rate have been proposed by several authors (Benjamini and Hochberg, 1995a; Benjamini and Liu, 1999; Benjamini and Yekutieli, 2001; Sarkar, 2002). In addition, methods for directly estimating the false discovery rate (Storey, 2002b) are also available.

A more recent innovation put forward by Storey and Tibshirani (2003) has been estimation of a quantity known as the q-value, which represents the minimum positive FDR rate at which significance is attained. It represents an analog of the p-value that takes multiple testing into account. It is quite commonplace for investigators to rank genes based on a q-value threshold.

Another technique that is done is to adjust p-values for multiple testing; a variety of methods for doing so is found in **?**. The p-value corresponds to the minimum significance level at which significance is attained. For multiple testing as described in Table 10.1, the an analog of the significance level is the familywide error rate (FWER), defined as $P(V \geq 1)$. Further discussion for FWER-controlling procedures can be found in Ge et al. (2003) and in a collection of papers by van der Laan and colleagues (Dudoit et al., 2004; van der Laan et al., 2004b,a). One unintended result of the development of high-throughput genomic data technologies has been the development of new statistical methodologies for addressing the multiple testing problem.

10.3.1 Functional and Pathway Analyses

Once these meta-analyses are performed and a calibrated list of genes are generated, the gene lists can be entered into databases representing functional processes. A simple visualization exercise, done in Rhodes et al. (2002), is to find metabolic pathways in which multiple genes exist. One example of such a database is the Kyoto Encyclopedia of Genes and Genomes (KEGG). Based on a list of genes that were consistently dysregulated across multiple studies comparing prostate cancer to non-prostate cancer, pathways such as the purine biosynthesis were found to have multiple genes. This leads to the hypothesis that the purine biosynthesis pathway is dysregulated in prostate cancer. While the study is only generating a hypothesis and not confirming it, such a computational prediction can help to inform investigators as to the next series of experiments to perform. Also, a visual display such as that given by KEGG does not allow for any formal statistical assessment of significance.

More formal statistical analyses for enrichment of functional terms can be done using the hypergeometric distribution. This requires a database of functional annotation terms such as Gene Ontology (GO) (Ashburner et al., 2000a). The idea behind this procedure is to see if the frequency of certain Gene Ontology terms in a list of genes is similar to or significantly larger than that in an external database. If it is determined that there is statistically significant enrichment of functional annotation terms in a list, then again this generates the hypotheses that certain pathways are dysregulated in the disease process. This can be easily seen with the following 2 × 2 table: There are l genes in the list and N genes total, i.e. on the chip. The null hypothesis is that there is no association between the rows and columns of the table; no association means that there is no functional enrichment of GO term X in the list of genes. This is tested for by calculating a p-value based on the hypergeometric distribution, which conditions on the row and column totals. An exact test is known as Fisher's exact test.

Table 10.2: Fisher's test example

	Gene List	Non-Gene List	Total
GO term X	a	b	G
Non GO term X	c	d	N-G
	l	N-l	N

There are now many publicly available tools for performing such a test (Draghici et al., 2003; Al-Sharour et al., 2004; Beissbarth and Speed, 2004). Note that the methods discussed in the last two paragraphs are post-hoc types of procedures in that the pathway analysis is done conditional on selecting a list of genes. An alternative is to directly model the information contained in the Gene Ontology databases with gene expression data. However, this raises the problem of what constitutes a proper metric by which the heterogeneous information from the two diverse databases can be related; this currently remains an open question. We later discuss the use of graphical models later in this chapter as well.

A resource initiated by our group is a database known as ONCOMINE (Rhodes et al., 2004b), located at the URL http://www.oncomine.org/. The database represents an effort to systematically curate, analyze and make available all public cancer microarray data via a web-based database and data-mining platform. Within the database, one can perform over 100 types of differential expression analyses based on disease/non-diseased, stage of disease, subtype, etc., reported with study-specific q-values. These analyses are based on standard differential expression analysis with correction for multiple testing using the q-value. In addition, one can query individual genes for known available genetic and proteomic information that is stored at other databases (e.g., GenBank, Swiss-Prot, etc.). There are links with pathway databases for visualization and assessing functional enrichment of the gene lists that are found. One can also search for individual genes of interest to see their expression patterns across multiple cancer studies.

10.4 Combining Data from Different Technologies

In the traditional statistical view of meta-analysis, one thinks of attempting to combine information from multiple similar experiments. However, the challenge of bioinformatics is that high-throughput functional genomics data are being generated on a variety of platforms and stored in different databases. The challenge then becomes how to integrate diverse data. This leads to a new definition of “meta-analysis.”

10.4.1 Bayesian networks

One tool that has been utilized quite heavily for this type of problem has been graphical models Lauritzen (1996); Jensen (2001). These are also referred to as Bayesian networks and belief networks as well. The idea of graphical models is to estimate dependencies between random variables through calculation of measures of covariation between them. As a simple example, let us consider three random variables, A, B and C. If we assume that the joint distribution of (A, B, C) is multivariate normal, then assuming the random variables have mean zero, the distribution is summarized by the pairwise correlation coefficients between them. Thus, if we can estimate the correlations, then we have “learnt” about the system characterized by A, B and C. There was a lot of interest in attempting to construct regulatory networks by fitting graphical models to gene expression data only. However, given the amount of experimental variability in such data, this turned out not to be a major direction, so the focus has been on building networks with multiple sources of data.

One major goal of Bayesian networks has been to predict protein-protein interactions. While much of the genomic data is measured at non-protein levels, actual cellular activity and disease occurs at a protein level. Thus, it is of interest to figure out how well functional genomic correlations predict protein-protein interactions. This was first studied in yeast by Jansen et al. (2003). However, they had the advantage of having high-throughput protein-protein interaction data available from yeast two-hybrid experiments. Such experiments currently do not exist for humans.

In a recent application (Rhodes et al., 2005b), we used Bayesian networks to predict protein-protein human interactions using functional genomic data. We used several different types of information in order to develop the graphical model:

1. interactions between orthologs of human proteins;
2. intergene correlations from gene expression profiles;
3. shared functional annotations from Gene Ontology;
4. shared enrichment domains.

The idea was to develop a graphical model using known positive and negative protein-protein interactions in order to develop a scale of evidence for predicting a protein-protein interaction. To define the positives, we used the Human Protein Reference Database (HPRD) (Peri et al., 2003), a bioinformatics resource that contains known protein-protein interactions manually curated from the literature by expert biologists. We queried 11,678 distinct literature-referenced protein-protein interactions among 5,505 proteins. For the negatives, we identified all protein pairs in which one protein was assigned to the plasma membrane cellular component and the other to the nuclear cellular component based on Gene Ontology. Based on fitting model, we predicted approximately 10,000 interactions with a false positive rate of 20% and about 40,000 interactions with a false positive rate of 50%. Several of the predicted protein-protein interactions were verified by subsequent experimentation, while other predictions mimicked what was found in the reported experimental literature. This model has been integrated into ONCOMINE and is available at the URL http://www.himapp.org/.

While there have been some successes with the graphical models approach, this area definitely remains in its infancy. One limitation of the graphical model is that it only uses pairwise covariation information. Furthermore, the graphical models used by Jansen et al. (2003) and Rhodes et al. (2005b) involve a binning procedure that seems somewhat *ad hoc*. One interesting alternative has been proposed by Balasubramanian et al. (2004), who propose using a graph-theoretic approach to combining functional genomics data from diverse platforms and test for significance of the nodal connections using permutation testing. Interestingly, the appear to be similarities with the use of graph-theoretic ideas in this area with those in social network literature (Wasserman and Faust, 1994). This suggests that there may exist techniques from that field that may be of use here.

Another point of the Bayesian networks is that they are bidirectional and do not attempt to impose any directionality. However, we know that activity in biological systems consists of a series of ordered steps. Thus, there might be some advantage to incorporating directionality into the system. Let us take the transcription process as an example. First, the must be binding of DNA to the upstream promoter regions in the genome so that transcription is "turned on." Thus, one could imagine a model for expression as a function of upstream promoter sequence for this scenario. Models like this have been proposed for lower-level eukaryotes (Bussemaker and Siggia, 2001; Conlon and Liu, 2003) and are referred to as "dictionary models." They take a view that the expression value is a function of a score computed using the sequence data, which is a conditional model. It remains to be seen whether such models could work for human genomic data.

10.4.2 Toward an understanding of regulatory mechanisms

In the previous sections, we have described methods for combining information in order to derive improved gene signatures and to make protein-protein interactions. Another goal of interest is to derive "regulatory" modules. It is likely that some gene expression patterns observed from microarray data represent a downstream readout of a small number of genetic aberrations (e.g., mutations, amplifications, deletions, translocations) that led to the activation or inactivation of a small number of transcription factors. In some cases, cancer-causing genetic aberrations may not be directly apparent from these downstream gene expression readouts. Recently, approaches to developing gene expression regulatory modules in human studies have been taken by Elkon et al. (2003), Segal et al. (2004) and Rhodes et al. (2005a).

The general approach requires a predefined list of genes. The list of genes can come from an external database, such as Gene Ontology (e.g. set of genes involved in a known process), or it may come from a differential expression analysis. Based on the gene list, the Segal et al. (2004) approach is to determine which arrays are commonly induced by multiple gene lists; the gene lists are then combined to form a "core" gene cluster. One then determines which arrays show significant differential expression based on the core gene cluster. One then determines if there is enrichment of clinical annotation in the set of arrays found at the previous step. Through this procedure, Segal et al. (2004) are able to find 456 regulatory modules from gene expression data consisting of measurements of 14,145 genes in 1917 samples across 22 tissue sites.

The approach taken by Elkon et al. (2003), while similar in spirit, involves a major difference. The difference is that sequence data are integrated with the gene expression profiling data. For the study by Elkon et al. (2003), approximately 13,000 putative promoter start sites were identified based on the NCBI Reference Sequence Database (`ftp://ftp.ncbi.nih.gov/genomes/H\_sapiens`). Next, a set of genes that were determined to be cell-cycle regulated from a human cell cycle gene expression profiling study ***() were used; of the 874 putative cell-cycle genes in that paper, promoter start sites were available for 568 of them. The authors searched for significantly enriched position weight matrices in the entire set of the 568 cell cycle-regulated promoters using the original 13K set as the background set and found enrichment of six binding sets. Thus, this provides a set of candidate transcription factors which may play a role in cell-cycle progression.

The study of Rhodes et al. (2005a) is similar to that of Elkon et al. (2003). They derive 265 gene lists from various differential expression analyses using a q-value cutoff of 0.10. Next, they identify putative transcription factor binding sites in the promoter sequences of human genes and come up with a database of 361 transcription factors. Next, enrichment of each transcription factor in each of the gene lists is done; again an adjustment for multiple testing based on false discovery rate calibration is performed. From this analysis, they defined 311 regulatory programs that displayed highly significant overlap ($P < 0.00033$) between a gene expression signature and a regulatory signature; these will serve as candidate regulatory modules that can be tested experimentally.

The crux of the analyses described in this section is that based on defined lists of genes, one calculates overlap measures of enrichment of a certain biological property (here binding sites) with the lists. It is fairly easy to see how other types of biological sequence information (e.g., protein structure information, etc.) might be used here as well. In addition, there are many ways of defining "interesting." It could be differential expression from a two-group comparison, or cell-cycle regulated (i.e., periodic expression) in a microarray time-course study. The overlap statistic is a very simple, and again, many other approaches are possible. This area will be a popular one for further study.

10.5 In vivo/in vitro genomic data integration

An area that is beginning to be considered more frequently in functional genomic studies in cancer is the integration of *in vitro*, i.e. experimental studies, with human gene expression studies, termed *in vivo* data. Integrating results from such experiments with in vivo cancer signatures holds the potential both to infer activity of specific oncogenic pathways in vivo and to identify relevant effectors of oncogenic pathways. For example, Huang et al. (2003) developed distinct in vitro oncogenic signatures for three transcription factors, Myc, Ras and E2F1-3. These signatures were able to predict Myc and Ras state in mammary tumors that developed in transgenic mice expressing either Myc or Ras, suggesting that specific oncogenic events are encoded in global gene-expression profiles.

To begin to understand the mechanisms by which oncogenes cause cancer, studies have used gene-expression profiling to identify downstream targets of oncogenic pathways in cell-culture systems. Conceptually, this involves manipulating a gene in an in vitro system and measuring a global profile using gene expression technology and then trying to relate the in vitro gene expression profile to an in vivo gene expression profile. Such an approach was taken by Lamb et al. (2003) to determine the direct transcriptional effects of oncogene Cyclin D1. In vitro experiments were performed in which the Cyclin D1 was both over and underexpressed, and global gene expression profiles were determined. Lists of differentially expressed genes were then generated. To correlate the lists with in vivo gene expression data, a two-step process was utilized in which genes were first ordered based on correlation with Cyclin D1. Then, a Kolmogorov-Smirnov statistic was used to determine if the lists clustered within the ordered list based on correlation. Since there was significant evidence of clustering, Lamb et al. (2003) found that the in vitro-defined targets of Cyclin D1 were correlated with Cyclin D1 levels in vivo. This suggests that the direct regulatory effects of Cyclin D1 may play an important role in tumorigenesis. The statistical problem brought up this type of analysis is determining clustering of a list of genes within an ordered list of genes. While a Kolmogorov-Smirnov statistic has the advantage of being a nonparametric statistic, the potential disadvantage to the use of such a method will be a loss of efficiency. Determining alternative methodologies for this type of problem will be important.

Another setting that leads to consideration of *in vitro* and *in vivo* genomic data is when the *in vitro* experiment is performed in a model organism system. For example, Sweet-Cordero et al. (2005) defined a signature by comparing lung tumors generated from a spontaneous KRAS mutation mouse model to normal mouse lung and correlating it with gene expression profiles in human lung cancer studies. The major issue in such an analysis is mapping mouse genes to orthologous human genes. Sweet-Cordero et al. (2005) found that the mouse signature shared significant similarity with human lung adenocarcinoma but not with other lung cancer types. Next, they looked for evidence of the KRAS signature in human tumors carrying activating KRAS mutations relative to wild-type tumors. Although no individual genes were significantly associated with KRAS mutation status in human tumors, the mouse KRAS signature was significantly enriched among genes rank-ordered by differential expression in human tumors with a KRAS mutation.

It is expected that experiments such as those described in the previous two paragraphs will become much more commonplace in the future. Thus, it will be critical to address issues and to develop methods for integrating in vivo and in vitro genomic data so that inferences regarding transcriptional regulatory pathways in cancer can be generated.

10.6 Software availability

Due to the recent innovations previously described, public use software for implementing these methods is still in their infancy. As mentioned earlier, our group has developed a database, ONCOMINE, located at the following URL:

http://www.oncomine.org/.

The database is geared towards biologists and does automated data analyses. Examples include differential expression analyses, analyses for functional enrichment of GO terms and Kolmogorov-Smirnov analyses in the spirit of Lamb et al. (2003). In addition, links to the protein-protein prediction project of Rhodes et al. (2005b) are available. The website for this is located at `http://www.himapp.org/`.

Many genomic data analysts primarily use software languages such as MATLAB and R (R Development Core Team, 200) for the analysis of genomic data. In particular, there has been a project towards the development of bioinformatics software packages in R, known as Bioconductor (Gentleman et al., 2004). The goals of the Bioconductor project are threefold: goals of the project include:

1. foster collaborative development and widespread use of innovative software;
2. reduce barriers to entry into interdisciplinary scientific research,
3. promote the achievement of remote reproducibility of research results.

One benefit of R is that it is a high-level interpretable language that allows for relatively fast development of methods. In addition, it has a nice ability for packaging related components.

Another language that is of great use in this type of bioinformatics research is Perl. Given that many of the databases are text databases, it is very important to be able to manipulate such databases relatively easily. Perl is a very useful language for such text manipulations.

10.7 Discussion

In this chapter, we have attempted to describe the current state of knowledge in the area of functional genomic analyses. Because of the different types of functional genomic datasets that are being generated, this has led to an extension of the statistical concept of meta-analysis. Now, analysts are faced with the prospect of combining different sources of information from different types of platforms.

One of the techniques described earlier, graphical models, is a tool from the area of machine learning. Machine learning algorithms tend to be black-box algorithms that are useful for predictive inference. While the application of machine learning algorithms to high-dimensional genomic datasets will lead to some predictions that will be borne out, it is also important to attempt to build in biological information as much as possible into the analyses. As an example, a central tenet of biology is that binding of DNA to the binding sites transcription factors leads to activation of gene expression. It would seem sensible that a model in which transcription factor information is the independent factor and gene expression is the dependent variable should be a better model for the system than a graphical model that assumes no directionality.

Finally, an important non-statistical issue that needs to be addressed is how to store information from these types of analyses such that they themselves can be combined. One can imagine that lists of genes from different analyses can be used to make inferences about various biological aspects in cancer studies. It then may be of interest to compare the lists themselves in another type of meta-analysis so that higher-order inferences about the biological network can be made. However, to do this will require work to develop database requirements and standardization, much as was done in the case of microarrays (Brazma et al., 2001).

10.8 Acknowledgments

The first author would like to acknowledge the support of grant GM72007 from the Joint NSF/NIGMS Biological Mathematics Program.

Proteomics

CHAPTER 11

Data integration for the study of protein interactions

Fengzhu Sun[1], Ting Chen[1], Minghua Deng[2], Hyunju Lee[1], Zhidong Tu[1]
[1]University of Southern California and [2] Beijing University

Abstract

With the development of genomic technologies, enormous amount of biological data have been and are continually being generated. They include genomic sequence data, gene expressions, protein-protein interactions, protein structures, protein localizations, protein functions, etc. For biological problems of interest, each data source contributes partially to the understanding of the problems. An important issue is how to integrate the different data sources to obtain a more complete understanding of the problems. In addition, most of the data sources from the high throughput experiments contain many false positive and false negative errors. Statistics plays an essential role in understanding the reliability of the observed biological data as well as to choose a more reliable data set from the observed ones. Statistics and machine learning techniques can help the integration of different data sources to understand the biological problems. We present two examples: to study the reliability of observed protein interaction data sets, and to predict protein functions combining different data sources.

11.1 Introduction

In recent years, an increasing number of genomes of model organisms have been sequenced. Using these genomic sequences, researchers have been able to make tremendous progress in the study of genomes, such as numerous successes in the identification of genes, the detection of protein-binding DNA motifs, and the determination of gene regulation. Beyond these successes is the far more challenging and rewarding task of understanding proteomes by means of, e.g., (1) discovering signal transduction pathways, (2) determining protein structures, (3) detecting protein-protein, protein-DNA, and protein-metabolite interactions, (4) detecting post-translational modifications of proteins, and ultimately (5) elucidating the functions of genes and their protein products.

Unlike a genome, which is a stable feature of an organism, a proteome varies with the state of the development, the tissue, and the environment. Among many features of a protein, the interaction with other proteins is one of the most important aspects of its function. Traditionally, protein interactions have been studied individually by biochemical techniques. However, the speed of discovering new interactions increased dramatically in the last couples of years; several high-throughput techniques have produced a total of about 80,000 interactions between yeast proteins, which constitute a rough view of the actual protein-protein interaction network. The successful methods include yeast two-hybrid assays Uetz et al. (2000b); **?**); **?**, protein complex purification-mass spectrometry **??**, microarray gene expression profiles **?**, genetic interactions **??**, and computationally predicted protein associations **???**. These protein interactions will be very useful to study gene regulatory networks, pathways, as well as functions of proteins. To understand the interaction network and its applications for protein function prediction, it is essential to design a joint approach using tools from mathematics, statistics, computer science, and molecular biology. In recent years, several groups have developed computational tools to analyze and compare the different interaction data sets.

Two issues are important in assessing the usefulness of an experimentally observed protein-protein interaction data set. One is the *reliability* which is defined as the fraction of real protein-protein interactions in the observed interactions and the other is the *coverage* which is defined as the fraction of real interactions in the observed data over all the real interactions. A database of high coverage is not very useful if its reliability is low. Results of comparative analysis of multiple data sets have shown significantly different coverage and reliability for each technique **???**. In this paper we review methods to study the following problems:

1. Estimate the reliability of a putative observed interaction data set.
2. Give a score that a pair of proteins interact by combining different data sources.

Assigning functions to novel proteins is one of the most important problems in the post-genomic era. Many researchers have undertaken the task of functionally analyzing one of the most well-studied species, the yeast genome, comprising approximately of around 6400 proteins, of which roughly one-third do not have known functions **?**, and the other two-thirds, most likely, have many other unknown functions. The annotation of the yeast genome will have a great impact on genomes of higher organisms such as the human: new genes can be annotated through their homologous yeast genes.

Several approaches have been applied to assign functions to genes, including analyzing gene expression patterns, phylogenetic profiles, protein fusions and protein-protein interactions. Gene expression analysis can cluster genes based on similar expression patterns. This makes it possible to assign a biological function to genes, depending on the knowledge of the functions of other genes in the cluster **?**. However, expression profiling gives an indirect measure of a gene product’s biological and cellular function, because many cellular processes and biochemical events are ultimately achieved by interactions of proteins. A more complete study of protein functions can be achieved by looking at not only the mRNA levels but also the protein interaction network. We will review the following methods for protein function prediction:

1. A Markovian random field (MRF) model for assigning functions to proteins using highly reliable protein-protein interaction data and other data sources including gene expression profiles, protein sequence similarities, and features of individual proteins, and correlations of protein functions.
2. The use of support vector machine (SVM) for protein function prediction combining different data sources.
3. A kernel-based MRF model for protein function prediction.

The paper is organized as follows. We first provide the data sources for the studies. Then we divide the paper into two major sections: estimating the reliability of observed putative protein interactions and predicting protein functions based on reliable protein interactions and other data sources. We then discuss the connections of the two topics and future research questions.

11.2 Data Sources

Protein interactions have traditionally been studied individually by genetic, biochemical, and biophysical techniques. However, these techniques are generally labor intensive and cannot keep up with the speed new proteins are discovered. Recently, several high-throughput methods for the detection of protein interactions have been developed. These include the yeast two-hybrid assays **??**Uetz et al. (2000b), mass spectrometry **?** and gene knockouts **?**. *In silico* (computational) methods for interaction prediction include the chromosomal proximity method **?**, the gene fusion method **??**, the phylogenetic method **?**, and the combined method **???**. Several databases have been developed to collect different sources of protein interaction data including the Munich Information Center for Protein Sequences (MIPS: http://mips.gsf.de/) **?**, Database of Interacting Proteins (DIP: http://dip.doe-mbi.ucla.edu/) **?**, Biomolecular Interaction Network Database (BIND: http://www.bind.ca/)**?**, and the General Repository for Interaction Datasets (GRID: http://biodata.mshri.on.ca/grid)**?**.

Gene Expressions are widely used to study the relationship between proteins. It is generally believed that a pair of interacting protein pair are more likely to be co-expressed than random protein pairs and thus gene expression data can be useful for evaluating the reliability of protein interaction data as well as the probability that two proteins interact. It is also generally believed that if two proteins are highly correlated, they are more likely to have similar functions. Therefore, gene expression data can also be useful for protein function prediction. For this study, we use the gene expression data from Spellman et al. (1998). Other gene expression data can also be used.

Protein localizations. Proteins belong to different localizations in the cell and proteins within the same locations are more likely to interact. Therefore, protein localization data can be useful for predicting protein interactions. We use the protein localization data of Huh et al. (2003) in this study.

Domains. The amino acid sequence of a protein is extremely important for the proteins function. The sequence of a protein determines its secondary and tertiary structure and thus, determines its interaction partners and its biological functions. Protein domains are conserved regions of peptide sequences with relatively independent tertiary structures and represent important features for understanding protein function. We use Pfam domains as the source of domain information. The SwissPfam (ver7.5) (ftp://ftp.genetics.wustl.edu/pub/pfam/) defines the mapping between proteins SWISS-PROT/TrEMBL accession numbers and Pfam domains.

Gene Ontology (GO) (http://www.geneontology.org/) describes gene products (proteins or RNA) based on three principles: Cellular component, Molecular function, and Biological process. GO has a directed acyclic graph (DAG) structure. The high level categories are more general and contain many more genes than low level categories. For protein function prediction, we base on the known gene annotation given in GO.

All the databases listed above are publicly available.

11.3 Assessing the reliability of protein interaction data

Many protein interaction data sets generated from various laboratories using different techniques are available. It is difficult to compare different interaction data because different conditions and experimental techniques may not detect the same type of interactions. Another difficulty comes from the fact that the true interaction data is unknown. Two issues need to be considered in comparing different interaction data sets. One is the *reliability* of the observed interaction data set defined as the overlap between the true interactions and the observed interactions over all the observed interactions. The other is the *coverage* defined as the overlap between the true interactions and the observed interactions over the true interactions. Without knowing the true interaction data, it is difficult to study the coverage of a certain observed interaction data set. On the other hand, it is possible to study the reliability of an observed interaction data set using gene expressions and localizations.

Mrowka et al. (2001) first observed that the distribution of correlation coefficients of gene expressions for true interacting protein pairs is stochastically larger than that for random protein pairs. The distribution of gene expression correlation coefficients for observed interacting protein pairs from high-throughput yeast-two-hybrid assays is between that for random protein pairs and that for true interaction pairs. The observations indicate that the set of observed protein interactions from high-throughput experiment is a mixture of random protein pairs and true interaction pairs. Several problems are of interests:

1. How do we choose the true interaction set (the gold standard)?
2. How do we estimate the fraction of true interactions among a set of observed interactions?
3. Is it possible to give a reliability score for an individual observed interaction?

11.3.1 Estimating the reliability of putative protein interactions based on gene expressions

There is no consensus choosing the gold standard set of true protein interactions. Mrowka et al. (2001) used MIPS physical interactions (excluding those from high-throughput experiments) as the gold standard. They used a bootstrap method to count how many random pairs need to be added to the reference data such that it has the same statistical behavior of gene expression correlation coefficients as that of the observed protein-protein interaction data, and then estimate the reliability using the sampling data. On the other hand, Deane et al. (2002) used INT, a subset of DIP interactions which are derived from small-scale experiments, as the gold standard for real interactions. They formalized the above idea assuming that the distribution of the square of Euclidian distance between expression profiles of putative interacting pairs is a mixture of that for the real interacting pairs and that of random pairs. They then used a least square approach to estimate the reliability of the putative protein interaction data. Deng et al. (2003) further extended the idea in Deane et al. (2002) and used a maximum likelihood estimation (MLE) approach to estimate the reliability of a putative interaction data set. Similar to Mrowka et al. (2001), they used MIPS physical interactions as a reference set for true interactions. The same approach can be applied to estimate the fraction of protein pairs that belong to the same complex in an observed complex data set. The method can be briefly described as follows.

Let α be the reliability of a given set of putative protein interactions. Let $O_e(\cdot)$, $T_e(\cdot)$ and $R_e(\cdot)$ be the distribution of the correlation coefficients for gene pairs based on gene expressions for the given set of putative protein interactions, the true protein interaction set, and the random protein pairs, respectively. Then

$$O_e(\cdot) = \alpha T_e(\cdot) + (1-\alpha) R_e(\cdot). \tag{11.1}$$

$T_e(\cdot)$ and $R_e(\cdot)$ can be approximated based on the correlation coefficients for pairs of proteins within the golden standard set of protein interactions and the correlation coefficients of all the protein pairs, respectively.

Deng et al. (2003) split the values of correlation coefficients into $K = 20$ bins. Let n_k be the number of observed interaction pairs in the k-th bin. Let p_k and q_k be the fractions of real interactions and random pairs in the k-th bin, respectively. Then the likelihood function can be defined as:

$$L(\alpha) = \prod_{k=1}^{K} (\alpha p_k + (1-\alpha) q_k)^{n_k}. \tag{11.2}$$

$L(\alpha)$ is a convex function and a classical gradient algorithm can be used to estimate the parameter α, $\widehat{\alpha}$, by maximizing $L(\alpha)$.

The following equation was used to calculate the variance of $\widehat{\alpha}$,

$$\mathtt{Var}(\widehat{\alpha}) = \left(\sum_{k=1}^{K} n_k \frac{(p_k - q_k)^2}{(\widehat{\alpha} p_k + (1-\widehat{\alpha}) q_k)^2} \right)^{-1}. \tag{11.3}$$

11.3.2 Estimating the reliability of putative protein interactions based on gene expressions and protein localizations

Huh et al. (2003) generated a large-scale protein localization map of yeast and showed that protein interactions are strongly enriched among co-localized proteins and proteins between specific cellular locations. Therefore both gene expressions and localizations can be used for reliability estimation **?**. Again we model the putative interaction data set as a mixture of true interactions and random pairs. Let $\theta_{ll'}$ and $\delta_{ll'}$ be the probability that a true interacting pairs and random protein pair belong to locations (l, l'), respectively. Let $n_{kll'}$ be the number of observed protein pairs within the putative interaction data set with correlation coefficient in the k-th bin and with localizations (l, l'). Combining gene expression data and protein localization data results in the following likelihood function

$$L(\alpha) = \prod_{k=1}^{K} \prod_{l,l'=1}^{L_0} (\alpha p_k \theta_{ll'} + (1-\alpha) q_k \delta_{ll'})^{n_{kll'}}, \tag{11.4}$$

where L_0 is the number of locations being considered. α can again be estimated by maximizing $L(\alpha)$.

The following equation was used to calculate the variance of $\widehat{\alpha}$,

$$\mathtt{Var}(\widehat{\alpha}) = \left(\sum_{k=1}^{K} \sum_{l,l'=1}^{L_0} n_{kll'} \frac{(p_k \theta_{ll'} - q_k \delta_{ll'})^2}{(\widehat{\alpha} p_k \theta_{ll'} + (1-\widehat{\alpha}) q_k \delta_{ll'})^2} \right)^{-1}. \tag{11.5}$$

11.3.3 Applications to protein interactions from high throughput experiments

We applied the above methods to protein interaction data sets from several high throughput experiments. Two groups of interaction data sets were studied. The first group includes pairwise physical interactions including the MIPS, DIP, Uetz's Uetz et al. (2000b) and Ito's **??** interaction data sets. The Ito$i$IST indicates the set of protein pairs that are observed to interact i times. The MIPS physical interactions are used as a true interaction data set. The estimated reliability together with their standard deviations of the estimates using gene expressions and protein localizations alone or combined are given in Table 1.

The second group includes the protein complexes such as the MIPS complex data, the

	Localization		Gene Expression		Both	
Data	Reliability	Standard Err.	Reliability	Standard Err.	Reliability	Standard Err.
			Physical Interactions			
DIP	0.587	0.0082	0.815	0.0244	0.619	0.0076
Uetz	0.685	0.0273	0.529	0.0843	0.699	0.0257
Ito1IST	0.268	0.0140	0.167	0.0383	0.293	0.0133
Ito2IST	0.411	0.0259	0.558	0.0831	0.470	0.0253
Ito3IST	0.532	0.0345	0.753	0.1144	0.611	0.0321
Ito4IST	0.552	0.0397	0.895	0.1436	0.640	0.0366
Ito5IST	0.547	0.0429	0.964	0.1567	0.640	0.0394
Ito6IST	0.556	0.0491	0.676	0.1768	0.641	0.0451
Ito7IST	0.608	0.0544	0.791	0.1942	0.682	0.0492
Ito8IST	0.614	0.0572	0.878	0.2054	0.684	0.0514
			Complexes			
TAP	0.4544	0.0063	0.585	0.0081	0.516	0.0056
HMS-PCI	0.1975	0.0042	0.248	0.0053	0.205	0.0037

Table 11.1: Reliability of the protein physical interaction data (Uetz's, DIP, and Ito's with different IST hits), and the protein complex data (the TAP and the HMS-PCI) using the protein localization data, the gene expression data and both data sets.

TAP complex data, and the HMS-PCI complex data. Any pair of proteins within the same complex are considered interacting. We treat the MIPS complex data as a true protein complex data set. Table 1 gives the estimated reliability and the corresponding standard deviation for the various protein complex data. The standard deviation of the estimate using gene expression alone is very large with the estimated reliability showing irregular patterns. For example, the estimated reliability for Ito4IST (0.895) is much higher than the estimated reliability of Ito6IST (0.676) contradicting with our intuition. The standard deviation of the estimated reliability using localization alone is much smaller and the estimated reliability for Ito$i$IST increases as i increases consistent with our intuition. Finally the standard deviation of the estimate based on the combined data is smaller than that using gene expressions or protein localizations alone.

11.3.4 Estimating the probability of interaction for individual protein pairs

The above approach can only estimate the fraction of true interactions in a putative interaction data set. However, it does not give a reliability score for a particular observed interaction. Saito et al. **?** proposed the criterion "interaction generality" to assess the reliability of a particular interaction protein pair based on the idea that a protein cannot interact with too many interacting partners. If a protein interact with a large number of proteins, it is most likely a "stick" protein and the observed interactions associated with this protein does not have real functional associations. Recently, Troyanskaya et al. (2003) and Jansen et al. (2003) developed Bayesian approaches to give a reliability measure for a particular putative interaction based on the observations that interacting protein pairs are more likely to have similar functions, to have similar gene expression patterns, and to be in the same location. Troyanskaya et al. (2003) gave a reliability score for two proteins to be functionally related and Jansen et al. (2003) gave a reliability score for two proteins to be in the same complex. Methods have also been developed to evaluate the contributions of individual features as well as combined features for predicting protein interaction **??**. It is found that only relatively small number of features, for example, protein function, is adequate for predicting protein interactions. **?** proposed a Markov random field (MRF) model for predicting protein interactions. They assumed a MRF model for the interaction network based on the theory of random graphs **?**. Conditional on the true interaction network, they assumed probability models for the observed data. Machine learning approaches were used to estimate the parameters as well as to predict the posterior probability of interactions for protein pairs conditional on the observations from different data sources. More details can be found in **?**.

11.4 Protein function prediction using protein interaction data

It has been observed that interacting proteins are more likely to have similar functions **?**. Therefore, protein interaction networks can be useful for protein function prediction. For a given protein, all the proteins interacting with the given protein form its neighbors. Fellenberg et al. **?** and Schwikowski et al. **?** developed a neighbor counting method for protein function prediction. For an unknown protein, they counted the number of known proteins of its neighbors for each function of interest and assigned the unknown protein with the function category having the highest frequency. One problem with this approach is that it does not consider the frequency of the proteins having certain functions of interest. Hishigaki et al. **?** developed a χ^2-statistic based approach for protein function prediction. For an unknown protein and a function of interest, a χ^2-statistic is calculated by comparing the observed frequency with the expected frequency of neighbors having the function of interest. The unknown protein is assigned the function with the highest χ^2 statistic. Both the counting method and the χ^2 method do not consider unknown protein neighbors. Several novel methods have been developed for protein function prediction based on interaction networks and other data sources. In this section we review these approaches.

Suppose a genome has N proteins $P_1, \cdots, P_N$. Let $P_1, \cdots, P_n$ be the unknown proteins and $P_{n+1}, \cdots, P_{n+m}$ be the known proteins, $N = n + m$. A protein may have several different functions. To simplify the problem, we study each functional category separately. For a function of interest, let $X_i = 1$ if the i-th protein has the function and 0 otherwise. The problem is to assign values to $X = (X_1, \cdots, X_n)$ conditional on the protein interaction networks, other pairwise relationships, features of individual proteins, and the functions of the known proteins.

11.4.1 A Markov Random Field (MRF) model for protein function prediction

Based on the idea of guilty-by-association, Deng et al. **?** first developed a MRF model for protein function prediction. The basic idea is to assign a prior probability for $X = (X_1, \cdots, X_{n+m})$, the configuration of function labelling based on the protein interaction network. Under this model, they calculated the posterior probability distribution for $(X_1, \cdots, X_n)$ conditional on the network and $(X_{n+1}, \cdots, X_{n+m})$. The key is how to assign the prior probability distribution. Different priors give different accuracy for protein function prediction.

A MRF model based on one network

In **?**, they assigned the prior as follows. Let π be the probability of a protein having the function of interest. Without considering the interaction network, the probability of a configuration of X is proportional to

$$\prod_{i=1}^{N} \pi^{x_i}(1-\pi)^{1-x_i} = \left(\frac{\pi}{1-\pi}\right)^{N_1} (1-\pi)^N, \tag{11.6}$$

where $N_1 = \sum_{i=1}^{N} x_i$.

Deng et al. **?** then considered one interaction network. Let S denote all the interacting protein pairs. The probability of the functional labelling conditional on the network is proportional to

$$\exp(\beta N_{01} + \gamma N_{11} + \kappa N_{00}), \tag{11.7}$$

where $N_{ll'}$ is the number of (l, l')-interacting pairs in S, and

$$\begin{aligned} N_{11} &= \sum_{(i,j)\in S} x_i x_j \\ &= \#\{(1 \leftrightarrow 1) \text{ pairs in S}\}, \\ N_{10} &= \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\ &= \#\{(1 \leftrightarrow 0) \text{ pairs in S}\}, \text{and} \\ N_{00} &= \sum_{(i,j)\in S} (1-x_i)(1-x_j) \\ &= \#\{(0 \leftrightarrow 0) \text{ pairs in S}\}. \end{aligned} \tag{11.8}$$

Therefore, the total probability of the functional labelling is proportional to $\exp(-U(x))$, where

$$\begin{aligned} U(x) &= -\alpha N_1 - \beta N_{10} - \gamma N_{11} - \kappa N_{00} \\ &= -\alpha \sum_{i=1}^{N} x_i - \beta \sum_{(i,j)\in S} x_i x_j \\ &\quad - \gamma \sum_{(i,j)\in S} (1-x_i)x_j + (1-x_j)x_i \\ &\quad - \kappa \sum_{(i,j)\in S} (1-x_i)(1-x_j), \end{aligned} \tag{11.9}$$

and $\alpha = \log(\frac{\pi}{1-\pi})$.

$U(x)$ is referred as the *potential function* in the field of MRF and defines a global Gibbs distribution of the entire network,

$$\Pr(X \mid \theta) = \frac{1}{Z(\theta)} \exp(-U(x)), \tag{11.10}$$

where $\theta = (\alpha, \beta, \gamma, \kappa)$ are parameters and $Z(\theta)$ is a normalized constant calculated by summing over all the configurations:

$$Z(\theta) = \sum_{x} \exp(-U(x)).$$

$Z(\theta)$ is called the partition function.

Several other approaches for protein function prediction based on one interaction network have been developed. In particular, Vazquez et al. (2003) considered multiple function categories and proposed to maximize the number of interactions within the same function categories. For one function of interest, it is equivalent to maximize

$$N_{00} + N_{11}$$

where N_{00} and N_{11} are defined as above. The **?** model differs from the **?** model in two significant ways. (1) Vazquez et al. (2003) used only the interaction network and did not consider the fraction of proteins having the function of interest in the known proteins. (2) Vazquez et al. (2003) gave an equal weight to intra-function class interactions. Letovsky and Kasif **?** proposed a model to assign functions to proteins based on a probabilistic analysis of graph neighborhoods in a protein-protein interaction network, which is fundamentally a MRF model, and the belief propagation algorithm was used to assign function probabilities for proteins in the network.

A Markov Random Field (MRF) model for multiple networks

Deng et al. **?** further extended the above model to multiple networks and to include features of individual proteins. Assume that L sources of protein pairwise relationships that may be useful for protein function prediction are available. A network can be built based on each pairwise relationship denoted as $\text{Net}_1,\ \text{Net}_2, \cdots, \text{Net}_L$, respectively. The entire network we consider is the union of all the networks denoted as S.

Similar to equation (11.7), our belief for the functional labelling of all the proteins based on network Net$_l$ is proportional to

$$P\{\text{ labelling }|\text{Net}_l\} \propto \exp(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}), \tag{11.11}$$

where $(N_{10}^{(l)}, N_{11}^{(l)}, N_{00}^{(l)})$ are defined similarly as equation (11.8).

Multiplying over all the networks, our belief for the functional labelling of all the proteins is proportional to

$$\begin{aligned} P\{\text{ labelling }|\text{networks }\} &\propto \prod_{l=1}^{L} \exp(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}) \\ &= \exp \sum_{l=1}^{L} \left(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}\right). \end{aligned} \tag{11.12}$$

Our total belief for the functional labelling of all the proteins is proportional to the multiplication of equations (11.6) and (11.12).

Then an MRF over all the functional labelling is defined by

$$P\{\text{labelling, networks}\} = \exp(-U(x))/Z(\theta), \tag{11.13}$$

where

$$U(x) = -\sum_{i=1}^{n+m} x_i \alpha - \sum_{l=1}^{L} \left(\beta_l N_{10}^{(l)} + \gamma_l N_{11}^{(l)} + \kappa_l N_{00}^{(l)}\right), \tag{11.14}$$

θ indicates the vector of parameters, and $Z(\theta)$ is the summation of $\exp(-U(x))$ over all the functional labelling. Under the above model, all the parameters $(\kappa_1, \kappa_2, \cdots, \kappa_L)$ are redundant and are set to 1. In the terminology of MRF, $U(x)$ is called the potential function.

Incorporating features of individual proteins

In addition to protein pairwise relationships, features of individual proteins can be very important for protein function prediction. A feature refers to an observation about a protein. It can be the presence or absence of a motif signal, the protein's conservation and localization, the protein's isoelectric point, its absolute mRNA expression level, or mutant phenotypes from experiments about the sensitivity or resistance of disruption mutants under various growth conditions. Several investigators have developed protein function prediction methods based on features of individual proteins **????????**. Deng et al. **?** integrated features into the MRF models for protein function prediction.

Suppose we have M features of interest, $F_1, F_2, \cdots, F_M$. The m-th feature can take values $0, 1, 2, \cdots k_m - 1$ where k_m is the number of categories for the m-th feature. Let the feature vector corresponding to protein P_i be $f_i = (f_{i1}, f_{i2}, \cdots, f_{iM})$, where f_{im} is the index for the m-th feature of the i-th protein. For the m-th feature, let

$p_{1m}(k)$ ($p_{0m}(k)$) be the conditional probability that a protein has feature index k given that a protein has (does not have) the function of interest. For simplicity, we assume that all the features contribute independently to the functions of proteins.

For a given feature vector $f = (f_1, f_2, \cdots, f_M)$, define

$$P_1(f) = \prod_{m=1}^{M} p_{1m}(f_m),$$
$$P_0(f) = \prod_{m=1}^{M} p_{0m}(f_m).$$

The probability of the features of all the proteins given the functional labelling is

$$P\{\text{features} \mid \text{labelling}\} = \prod_{i:X_i=1} P_1(f_i) \times \prod_{i:X_i=0} P_0(f_i). \qquad (11.15)$$

Multiplying equations (11.13) and (11.15), we have the following probability model

$$\begin{aligned} &P\{\text{labelling, networks, domain features}\} = \\ &P\{\text{labelling, networks}\} \times P\{\text{domain features} \mid \text{labelling}\}. \end{aligned} \qquad (11.16)$$

Deng et al. **?** described methods to estimate the posterior distribution of the functions of the unknown proteins given the features of all the proteins, the different sources of protein pairwise relationship, and the annotations of the known proteins.

Computational Issues

Given the above models, the problem is to estimate the posterior probability distribution given the annotation of the known proteins, the features of all the proteins, and the network. The parameters are also unknown. Using equation (11.16), it can be shown that

$$\begin{aligned} &\log \frac{Pr(X_i = 1 \mid F, X_{[-i]}, \theta)}{1 - Pr(X_i = 1 \mid F, X_{[-i]}, \theta)} \\ =&\alpha_i + \sum_{l=1}^{L} (\beta_l - 1) M_0^{(i)}(l) + (\gamma_l - \beta_l) M_1^{(i)}(l), \end{aligned} \qquad (11.17)$$

where F is the feature information for all the proteins, $X_{[-i]} = (X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_{n+m})$, $\alpha_i = \log \frac{\pi P_1(f_i)}{(1-\pi)P_0(f_i)}$, $M_0^{(i)}(l)$ and $M_1^{(i)}(l)$ are the numbers of neighbors of protein P_i labelled with 0 and 1 according to the l-th network, respectively. The parameters can be estimated based on the network consisting of the known proteins by an S-plus routine **?** using equation (11.17).

Once all the parameters have been defined, Gibbs sampler **?** can be used to estimate the posterior probability distribution of $(X_1, \cdots, \cdots, X_n)$. The algorithm can be described as follows:

1. Randomly set the value of missing data $X_i = \lambda_i, i = 1, \cdots, n$ with probability π.
2. For each protein P_i, update the value of X_i using equation (11.17).
3. Repeat step 2 T times until all the posterior probabilities Pr$(X_i \mid D,\ X_{[-i]}, \theta)$ are stabilized.

11.4.2 Kernel-based methods for protein function prediction

In the MRF formulation, we only consider immediate neighbors for proteins. The protein interaction network can be used to define similarity between any pair of proteins using the diffusion kernel **?**. In the following we first briefly describe kernel based methods of Lanckriet et al. **???** to combine different data sources for protein function prediction. Then we describe our effort to combine the idea of kernel based method with the MRF model.

Support vector machine (SVM) and semidefinite programming (SDP)

In a series of recent papers, Lanckriet et al. **???** developed kernel-based methods for protein function prediction using SVM. Suppose that there are L data sources such as protein interactions, gene expressions, domains, localizations, etc. For the l-th data source, a kernel matrix K_l (semi-positive definite) is defined. For continuous data such as gene expressions, the Gaussian diffusion kernel can be used. For protein interactions, diffusion kernel on graphs can be used **?**. Several other kernel matrixes have been developed for different sources of data structures in **???**. To integrate different data sources, Lanckriet and colleagues considered the linear combinations of the kernel matrixes

$$K = \sum_{l=1}^{L} \mu_l K_l$$

where $\mu_l \geq 0, l = 1, 2, \cdots$ are parameters to be determined.

They used SVM with 1-norm soft margin to build a classifier. The problem can then be solved by solving the following constraint maximization problem:

$$\begin{aligned} &\max_{\alpha,t} 2\alpha^T e - ct \\ \text{subject to } &t \geq \frac{1}{r_i}\alpha^T \text{diag}(y) K_l \text{diag}(y)\alpha, \quad l = 1, 2, \cdots, L \\ &\alpha^T y = 0, \\ &C \geq \alpha \geq 0, \end{aligned} \tag{11.18}$$

where $r_i = \text{trace}(K_i)$, $c = \mu^T r$ and y is the annotation of the known proteins. This problem is a quadratically constraint quadratic program (QCQP) problem (Boyd and Vandenberghe 2001) and can be solved using standard software such as SeDuMi (Sturm 1999). The computational time is $O(n^3)$, where n is the number of proteins in the training set.

Combining kernel with the MRF model for protein function prediction

Lanckriet et al. (2004a) showed that SVM described above outperformed the MRF approach in almost all the function categories considered. One of the main reasons probably is due to the inclusion of multiple level neighbors in the kernel based methods. Note that $K_l(i,j)$ defines a similarity between protein P_i and protein P_j based on the l-th data source. Similar to equation (11.11), the probability of the labelling based on the l-th network N_l can be modelled as

$$\exp(\beta_l D_{10}(l) + \gamma_l D_{11}(l) + \kappa_l D_{00}(l)) \tag{11.19}$$

where β_l, γ_l, and κ_l are constants, and

$$\begin{aligned} D_{11}(l) &= \sum_{i<j} K_l(i,j) I\{x_i = 1, x_j = 1\}, \\ D_{10}(l) &= \sum_{i<j} K_l(i,j) I\{(x_i = 1, x_j = 0) \text{ or } (x_i = 0, x_j = 1)\}, \\ D_{00}(l) &= \sum_{i<j} K_l(i,j) I\{x_i = 0, x_j = 0\}. \end{aligned} \tag{11.20}$$

The summations are over all the protein pairs. Multiplying equation (11.6) and equation (11.19) for $l = 1, 2, \cdots, L$, we obtain the the total probability proportional to

$$\exp\left(\alpha N_1 + \sum_{l=1}^{L} (\beta_l D_{10}(l) + \gamma_l D_{11}(l) + \kappa_l D_{00}(l))\right) \tag{11.21}$$

¿From equation (11.21), it can be shown that

$$\begin{aligned} &\log \frac{\mathrm{P}(X_i = 1 \,|X_{[-i]}, \theta)}{1 - \mathrm{P}(X_i = 1 \,|X_{[-i]}, \theta)} \\ =&\alpha + (\beta_l - \kappa_l) K_0^{(i)}(l) + (\gamma_l - \beta_l) K_1^{(i)}(l). \end{aligned} \tag{11.22}$$

where

$$\begin{aligned} K_0^{(i)}(l) &= \sum_{j \neq i} K_l(i,j) I\{x_j = 0\}, \\ K_1^{(i)}(l) &= \sum_{j \neq i} K_l(i,j) I\{x_j = 1\}. \end{aligned}$$

Note that if we let $K_l(i,j) = 1$ when protein i interacts with protein j and $K_l(i,j) = 0$ otherwise in the l-th network, this new model is the same as the MRF model of Deng et al. **?**. We can similarly develop a MCMC approach to approximate the probability that an unknown protein having the function of interest. We refer the above approach as kernel-based MRF (KMRF)

11.4.3 Applications to real data

All the methods described above have been applied to predict protein functions. The MRF model has been used for protein function prediction first based on the MIPS function classification **??** and later were extended to functions defined in GO **?**. The SVM approach has been used to predict protein functions based on MIPS **?**, to predict ribosomal proteins and memberane proteins **?**. A summary paper for protein function prediction based on SVM is given in **?**. The new KMRF method has been applied for protein function prediction based on GO **?** and for prediction of protein essentiality **?**. The KMRF approach can be easily extended to incorporate correlated functions. For most functions that have been considered so far, the SVM approach outperformed the MRF approach. The KMRF approach has similar performance as the SVM approach. For example, for predicting protein essentiality, the receiver operating characteristic (ROC) scores for the MRF, SVM, and the KMRF approaches are 0.804. 0.812, and 0.831, respectively, based on the core interaction data set. Integrating protein function based on cellular processes, conservation, and localizations into the model increased the ROC score of the KMRF model to 0.869.

11.5 Discussion

Enormous amount of biological data have been generated and stored in public and private databases. These data sources are extremely important for biological studies. However, the data are generally noisy and contain many false positive and false negative errors. There are no systematic statistical tools to choose the most reliable data from the noisy data. The various data sources can most likely contribute to our understanding of the biological problems of interest. The data sources are usually correlated and their contributions to our understanding of the biological problems are not independent. An important issue is how to integrate the usually noisy and correlated data sources to understand the biological problems.

In this chapter, we review our recent efforts in integrating different data sources for biological studies. First we describe likelihood based methods for estimating the reliability of putative interaction data sets. We show that the localization data give more accurate estimation of the reliability than using the gene expression data. Integrating the localization and gene expression data can give even more accurate estimates of the reliability of the different data sets. Other statistical methods for estimating the probability of two proteins being interact integrating different data sources have also been developed.

Second we describe methods for protein function prediction based on interaction networks, genetic interactions, other pairwise relationships, as well as features of individual proteins. These approaches include MRF, SVM, and KMRF. As far as we know, the combination of kernels with MRF is novel in protein function prediction. The simplicity of KMRF and its high accuracy in protein function prediction warrant further studies of the this approach in other fields. As in most model based approaches, the KMRF model can be understand the contributory factors for the protein function of interest.

In protein function prediction, we implicitly assume that the networks under consideration, such as the protein interaction network and genetic interaction network, are highly reliable. Therefore we used the core interaction data in DIP in all our studies on protein function prediction. We tried to use all the interactions (not reliable) in DIP for protein function prediction and, as expected, the prediction accuracy is lowered. A problem is how best to use all the interactions for protein function prediction. The effect of incompleteness of the interaction data on protein function prediction is also unknown.

In summary, we show the power of integrating multiple data sources for biological studies. Significant questions remain as to how to integrate noisy and incomplete data in biological studies. It is also important to develop methods to evaluate the dependence among the different data sources and to integrate the correlated data sources for biological studies.

Acknowledgments

The research was partly supported by NIH/NSF joint mathematical biology initiative DMS-0241102.

CHAPTER 12

Gene Trees, Species Trees, and Species Networks

Luay Nakhleh Derek Ruths
Department of Computer Science
Rice University
Houston, TX 77005, USA
{nakhleh,druths}@cs.rice.edu

Hideki Innan
University of Texas Health Science Center at Houston
Houston, TX 77030, USA
Hideki.Innan@uth.tmc.edu

12.1 Introduction

The availability of whole-genome sequence data has provided a rich resource of deep insights into many biological, medical and pharmaceutical problems and applications, and is promising even more. Yet, along with these insights and promises, genomic data have given rise to many challenging problems, mainly due to the quantity and heterogeneity of such data. One of these major challenges is the phylogenetic analysis of multiple gene datasets that whole genomes provide.

Phylogeny, i.e., the evolutionary history of a set of organisms, has become an indispensable tool in the post-genomic era. Emerging techniques for handling essential biological tasks (e.g., gene finding, comparative genomics, and haplotype inference) are usually guided by an underlying phylogeny. The performance of these techniques, therefore, depends heavily on the quality of the phylogeny. Almost all phylogenetic methods, however, assume that evolution is a process of strict divergence that can be modeled by a phylogenetic tree. While the tree model gives a satisfactory first-order approximation for many families of organisms, other families exhibit evolutionary events that cannot be represented by a tree. In particular, the evolutionary history of bacterial genomes is characterized by the occurrence of processes such as horizontal gene transfer (HGT)—transfer of genetic material across the boundaries of of distantly related species—and inter-specific recombination—exchange of genetic material. Further, hybrid speciation occurs among various groups of plants, fish, and frogs. In the presence of such evolutionary processes, the evolutionary relationship of a set of organisms is modeled by a *phylogenetic network*.

Accurate reconstruction of these processes bears significant impact on many domains. The Tree of Life—the phylogeny of all organisms on Earth—is one of the grand challenges in evolutionary biology. The prokaryotic branch of this tree is believed to have a large number of horizontal gene transfer events, in addition to recombination events. Efforts to reconstruct a phylogeny for the prokaryotic branch may prove futile without developing phylogenetic network models and reconstruction methods.

A significant aspect of these complex evolutionary mechanisms is their contribution to microbial genome diversification. Like all forms of life, bacteria undergoes evolution. However, unlike many other organisms, bacterial evolution is not one of strict divergence. Recombination usually occurs within populations; in bacteria, however, recombination occurs among different strains. Further, HGT is ubiquitous in the prokaryotic branch of the Tree of Life. **?** has recently written of the various health risks that recombination and HGT pose, including: (1) antibiotic resistance genes spreading to pathogenic bacteria, (2) disease-associated genes spreading and recombining to create new viruses and bacteria that cause diseases, and (3) transgenic DNA inserting into human cell, triggering cancer. Hence, detecting and reconstructing these processes in bacteria play a major role in developing effective antibiotics, and bears a great impact on human health.

Biologists have long acknowledged the presence of these processes, their significance, and their effects. The computational research community has responded in recent years and proposed a plethora of methods for reconstructing complex evolutionary histories. The general theme of most existing methods can be summarized by: construct gene trees and reconcile them (this is known as the *separate analysis* approach). Gene tree reconciliation presents two major issues, namely identifying the (biological) source of incongruence, and (computationally) reconciling the trees. Many processes may lead to *incongruent* gene trees:
(1) *Stochastic factors*, such as wrong assumptions, insufficient data, incomplete sampling, and differential rates of sequence evolution across lineages. These factors do not violate the tree model of organismal evolutionary relationships; rather, the incongruence they cause must be eliminated in the early stages of phylogenetic analyses.
(2) *Intra-species factors*, such as gene loss and duplication. Although these events may lead to incongruent gene trees, they do not violate the tree model of organismal evolutionary relationships.
(3) *Inter-species factors*, such as horizontal gene transfer (whose rate is very high among prokaryotic organisms), and inter-specific recombination. These events result in *networks* of relationships, rather than trees of relationships.

In this work, we review the intra- and inter-species factors that cause gene tree incongruence and discuss current approaches for resolving these phenomena, with focus on non-treelike evolution. Further, we address extensions to the *coalescent* model to address non-treelike evolution. The rest of the chapter is organized as follows. In

Section 12.2 we illustrate some of the processes that lead to incongruence gene trees. In Section 12.3 we review existing methods for addressing gene tree incongruence caused by gene loss and duplication (intra-species factors). In Section 12.4, we describe the *phylogenetic network* model and discuss the problem of reconciling gene trees into species networks. In Section 12.5 we propose approaches for extending the coalescent model to incorporate non-treelike evolutionary processes. We conclude the chapter in Section 12.6.

12.2 Gene Tree Incongruence

A **gene tree** is a model of how a gene evolves through duplication, loss, and nucleotide substitution. As a gene at a locus in the genome replicates and its copies are passed on to more than one offspring, branching points are generated in the gene tree. Because the gene has a single ancestral copy, barring recombination, the resulting history is a branching tree (**?**). Sexual reproduction and meiotic recombination within populations break up the genomic history into many small pieces, each of which has a strictly treelike pattern of descent (**???**). Thus, within a species, many tangled gene trees can be found, one for each nonrecombined locus in the genome. A **species tree** depicts the pattern of branching of species lineages via the process of speciation. When reproductive communities are split by speciation, the gene copies within these communities likewise are split into separate bundles of descent. Within each bundle, the gene trees continue branching and descending through time. Thus, the gene trees are contained within the branches of the species phylogeny (**?**).

Gene trees can differ from one another as well as from the species tree. Disagreements (incongruence) among gene trees may be an artifact of the data and/or methods used (stochastic factors). Various studies show the effects of stochastic factors on the performance of phylogenetic tree reconstruction methods (e.g., **???????**). Stochastic factors confound the accurate reconstruction of evolutionary relationships, and must be handled in the first stage of a phylogenetic analysis. Incongruence among gene trees due to intra- or inter-species processes, on the other hand, is a reflection of true biological processes, and must be handled as such.

Whereas eukaryotes evolve mainly though lineal descent and mutations, bacteria obtain a large proportion of their genetic diversity through the acquisition of sequences from distantly related organisms, via horizontal gene transfer (HGT) or recombination (**?**). Views as to the extent of HGT and recombination in bacteria vary between the two extremes, with most views being in the middle (**???????**). However, there is a common belief that recombination and HGT, among other processes, form the essence of prokaryotic evolution. Further, these two are the main processes (in addition to random mutations) by which bacteria develop resistance to antibiotics (e.g., **????**). Gene transfer and exchange are considered a primary explanation of incongruence among bacterial gene phylogenies and a significant obstacle to reconstructing the prokaryotic branch of the Tree of Life (**?**).

We illustrate some of the scenarios that may lead to gene tree incongruence in Figure

(a) (b)

(c) (d)

(e)

Figure 12.1: (a) Gene tree that agrees with the species tree. (b) Gene tree that disagrees with the species tree due to gene loss and duplication. (c) Gene tree that disagrees with the species tree due to HGT. (d) An inter-specific recombination event in which genetic material is exchanged between species B and C. (e) A hybrid speciation event that leads to two incongruent gene trees.

12.1. The species (or, organismal) tree is represented by the "tubes"; it has A and B as sister taxa whose most recent common ancestor (MRCA) is a sister taxon of C. Figure 12.1(a) shows a gene evolving within the branches of the same species tree; in this case, the topologies of the gene and species trees agree (the topology of this gene tree is shown in Figure 12.2(a)). In Figure 12.1(b) we show an example of

(a) (b) (c) (d)

Figure 12.2: (a) The tree of the gene whose evolution is shown in Figure 12.1(a), and Figure 12.1(e). (b) The tree of the genes whose evolution is shown in Figures 12.1(b) and 12.1(c). (c) The tree of the gene involved in the recombination event shown in Figure 12.1(d). (e) The tree of the gene involved in the hybrid speciation event shown in Figure 12.1(e).

how intra-species processes may lead to incongruent gene trees. The figure shows a gene evolving within the branches a species tree with one duplication event and three losses. Note that the species tree differs from the gene tree; based on this gene, B and C are sister taxa and their MRCA is a sister of taxon A. This gene tree is shown in Figure 12.2(b).

Another event that may cause incongruence between the species tree and the gene tree is HGT. In the case of HGT, shown in Figure 12.1(c), genetic material is transferred from one lineage to another. Sites that are not involved in a horizontal transfer are inherited from the parent (as in Figure 12.2(a)), while other sites are horizontally transferred from another species (as in Figure 12.2(b)).

In the case of inter-specific recombination, as illustrated in Figure 12.1(d), some genetic material is exchanged between pairs of species; in this example, species B and C exchange genetic material. The genes involved in this exchange have an evolutionary history (shown in Figure 12.2(c)) that is incongruent with that of the species. In hybrid speciation, two lineages recombine to create a new species. We can distinguish *diploid hybridization*, in which the new species inherits one of the two homologs for each chromosome from each of its two parents—so that the new species has the same number of chromosomes as its parents, and *polyploid hybridization*, in which the new species inherits the two homologs of each chromosome from both parents—so that the new species has the sum of the numbers of chromosomes of its parents. Under this last heading, we can further distinguish *allopolyploidization*, in which two lineages hybridize to create a new species whose ploidy level is the sum of the ploidy levels of its two parents (the expected result), and *auto-polyploidization*, a regular speciation event that does not involve hybridization, but which doubles the ploidy level of the newly created lineage. Prior to hybridization, each site on each homolog has evolved in a tree-like fashion, although, due to meiotic recombination, different strings of sites may have different histories. Thus, each site in the homologs of the parents of the hybrid evolved in a tree-like fashion on one of the trees induced by (contained inside) the network representing the hybridization event. Figure

12.1(e) shows a network with one hybrid. Each site evolves down exactly one of the two trees shown in Figures 12.2(a) and 12.2(d).

Notice that in the case of intra-species processes (gene loss and duplication), inferring the species tree from a set of potentially conflicting gene trees is a problem of *reconciling* the gene trees and explaining their differences through duplications and losses of genes. Therefore, in this case, despite the potential incongruence among the species and gene trees, the species phylogeny is still a tree (**????**). However, in the case of recombination, HGT, and hybrid speciation, the evolutionary history of the organismal genomes cannot be represented by phylogenetic trees; rather, *phylogenetic networks* are the appropriate model (**??**).

12.3 Trees Within Trees: The Gene Tree Species Tree Problem

Various reports of instances and effects of gene loss and duplication exist in the literature (e.g., **???**). When losses and duplications are the only processes acting on the genes, a mathematical formulation of the gene tree reconciliation problem is as follows:

Definition 12.1 *(The Gene Tree Reconciliation Problem)*

> ***Input:*** *Set $\mathcal{T}$ of rooted gene trees, a cost w_D for duplications, and a cost w_L for losses.*
>
> ***Output:*** *Rooted tree T with each gene tree $t \in \mathcal{T}$ mapped onto T, so as to minimize the sum $w_D n_D + w_L n_L$, where n_D is the total number of duplications and n_L is the total number of losses, over all genes.*

This problem was shown to be NP-hard by **?** and **?**. Heuristics for the problem exist, but these do not solve the optimization problem (see **??**). Various fixed-parameter approaches have been proposed by **??** and some variants can be approximated to within a factor of 2 and shown by **?**.

When loss and duplication are the only processes acting on the genes, two different questions can be posed, depending on the input data:

1. Gene tree reconciliation problem—when the gene trees are known and the species tree is known, what is the best set of duplication and loss events that reconcile each gene tree with the species tree?
2. Species tree construction problem—when the gene trees are known, but the evolutionary relationships among the species involved is not known, can the gene trees provide the information necessary to derive an estimate of the species tree?

Both of these questions require the assumption of a certain model of gene duplication and loss. The complexity of the gene-tree reconciliation problem is determined by the model chosen, whereas the general species tree construction problem is NP-hard under all commonly used models of gene duplication and loss.

The simplest version of either problem uses a duplication-only model (i.e., losses do not occur). During the period between years 1995 and 2000, this was a commonly used model (**????????**). Under the duplication-only model, the gene tree reconciliation problem has linear-time solutions (**??**), as well as other polynomial-time algorithms that report better performance on real biological datasets (**?**). The species tree construction problem is NP-hard , as was shown by **?**. Different approaches have been taken to solving the species tree construction problem including heuristics (**?**), approximation algorithms (**?**), and fixed parameter tractable algorithms obtained by parameterizing by the number of gene duplications separating a gene tree from the species tree (**?**).

The other common model used is the more general duplication-loss model, which admits both duplication and loss events within gene trees. The gene tree reconciliation problem has been shown to be polynomial-time under conditions where the evolution of the sequences themselves are not considered (**???**); if this is taken into account, the problem becomes NP-hard (**??**). Various efficient heuristics for the problem are currently available (**??**). Early work on the gene tree reconciliation problem under this model borrowed techniques from biogeography and host/parasite evolution (**??**).

12.4 Trees Within Networks: The Gene Tree Species Network Problem

As described in Section 12.2, when events such as horizontal gene transfer, hybrid speciation, or recombination occur, the evolutionary history can no longer be modeled by a tree; rather, *phylogenetic networks* are the appropriate model in this case. In this section, we describe the phylogenetic network model and approaches for reconstructing networks from gene trees.

12.4.1 Terminology and notation

Given a (directed) graph G, let $E(G)$ denote the set of (directed) edges of G and $V(G)$ denote the set of nodes of G. Let (u, v) denote a directed edge from node u to node v; u is the *tail* and v the *head* of the edge and u is a *parent* of v. The *indegree* of a node v is the number of edges whose head is v, while the *outdegree* of v is the number of edges whose tail is v. A node of indegree 0 is a *leaf* (often called a *tip* by systematists). A directed path of length k from u to v in G is a sequence $u_0u_1 \cdots u_k$ of nodes with $u = u_0$, $v = u_k$, and $\forall i$, $1 \leq i \leq k$, $(u_{i-1}, u_i) \in E(G)$; we say that u is the tail of p and v is the head of p. Node v is *reachable* from u in G, denoted $u \rightsquigarrow v$, if there is a directed path in G from u to v; we then also say that u is an *ancestor* of v. A *cycle* in a graph is a directed path from a vertex back to itself; trees never contain cycles: in a tree, there is always a unique path between two distinct vertices. Directed acyclic graphs (or DAGs) play an important role on our model; note that every DAG contains at least one vertex of indegree 0. A *rooted directed acyclic graph*, in the context of this paper, is then a DAG with a single node of indegree 0, the *root*; note that all all other nodes are reachable from the root by a (directed) path of graph edges. We denote by $r(T)$ the root of tree T and by $L(T)$ the leaf set of T. Let T be a rooted phylogenetic tree over set S of taxa, and let $S' \subseteq S$. We denote by $T(S')$ the minimal rooted subtree of T that connects all the element of S'. Furthermore, the restriction of T to S', denote $T|S'$ is the rooted subtree that is obtained from $T(S')$ by suppressing all vertices (except for the root) whose number of incident edges is 2. Let S' be a maximum-cardinality set of leaves such that $T_1|S' = T_2|S'$, for two trees T_1 and T_2; we call $T_1|S'$ (equivalently, $T_2|S'$) the maximum agreement subtree of the two trees, denoted $MAST(T_1, T_2)$. A *clade* of a tree T is a complete subtree of T. Let $T' = MAST(T_1, T_2)$; then, $T_1 - T'$ is the set of all maximal clades whose pruning from T_1 yields T' (we define $T_2 - T'$ similarly). In other words, there do not exist two clades u and u' in $T_1 - T'$ such that either u is a clade in u', or u' is a clade in u.

We say that node x reaches node y in tree T if there is a directed path from x to y in T. We denote the root of a clade t by $r(t)$. We say that clade t_1 reaches clade t_2 (both in tree T) if $r(t_1)$ reaches $r(t_2)$. The sibling of node x in tree T is node y, denoted $sibling_T(x) = y$ whenever x and y are children of the same node in T. We denote by T_x the clade rooted at node x in T. The least common ancestor of a set X of taxa in tree T, denoted $lca_T(X)$ is the root of the minimal subtree of T that contains the leaves of X. The edge incoming into node x in tree T is denoted by $inedge_T(x)$.

12.4.2 Phylogenetic networks

? modeled phylogenetic networks using directed acyclic graphs (DAGs), and differentiated between "model" networks and "reconstructible" ones.

Model networks A phylogenetic network $N = (V, E)$ is a rooted DAG obeying certain constraints. We begin with a few definitions.

Definition 12.2 *A node $v \in V$ is a* tree node *if and only if one of these three conditions holds:*

- $indegree(v) = 0$ *and* $outdegree(v) = 2$*:* root*;*
- $indegree(v) = 1$ *and* $outdegree(v) = 0$*:* leaf*; or*
- $indegree(v) = 1$ *and* $outdegree(v) = 2$*:* internal tree node.

A node v is a network node *if and only if we have* $indegree(v) = 2$ *and* $outdegree(v) = 1$.

Tree nodes correspond to regular speciation or extinction events, whereas network nodes correspond to reticulation events (such as hybrid speciation and horizontal gene transfer). We clearly have $V_T \cap V_N = \emptyset$ and can easily verify that we have $V_T \cup V_N = V$.

Definition 12.3 *An edge $e = (u, v) \in E$ is a* tree edge *if and only if v is a tree node; it is a* network edge *if and only if v is a network node.*

The tree edges are directed from the root of the network towards the leaves and the network edges are directed from their tree-node endpoint towards their network-node endpoint.

A phylogenetic network $N = (V, E)$ defines a partial order on the set V of nodes. We can also assign times to the nodes of N, associating time $t(u)$ with node u; such an assignment, however, must be consistent with the partial order. Call a directed path p from node u to node v that contains at least one tree edge a *positive-time directed path*. If there exists a positive-time directed path from u to v, then we must have $t(u) < t(v)$. Moreover, if $e = (u, v)$ is a network edge, then we must have $t(u) = t(v)$, because a reticulation event is effectively instantaneous at the scale of evolution; thus reticulation events act as synchronization points between lineages.

Definition 12.4 *Given a network N, two nodes u and v cannot* co-exist *(in time) if there exists a sequence $P = \langle p_1, p_2, \ldots, p_k \rangle$ of paths such that:*

- *p_i is a positive-time directed path, for every $1 \leq i \leq k$;*
- *u is the tail of p_1, and v is the head of p_k; and*
- *for every $1 \leq i \leq k-1$, there exists a network node whose two parents are the head of p_i and the tail of p_{i+1}.*

Obviously, if two nodes x and y cannot co-exist in time, then a reticulation event between them cannot occur.

Definition 12.5 *A* model phylogenetic network *is a rooted DAG obeying the following constraints:*

1. *Every node has indegree and outdegree defined by one of the four combinations $(0, 2)$, $(1, 0)$, $(1, 2)$, or $(2, 1)$—corresponding to, respectively, root, leaves, internal tree nodes, and network nodes.*
2. *If two nodes u and v cannot co-exist in time, then there does not exist a network node w with edges (u, w) and (v, w).*
3. *Given any edge of the network, at least one of its endpoints must be a tree node.*

Reconstructible networks Definition 12.5 of model phylogenetic networks assumes that complete information about every step in the evolutionary history is available. Such is the case in simulations and in artificial phylogenies evolved in a laboratory setting—hence our use of the term *model*. When attempting to reconstruct a phylogenetic network from sample data, however, a researcher will normally have only incomplete information, due to a combination of extinctions, incomplete sampling, and abnormal model conditions. Extinctions and incomplete sampling have the same consequences: the data do not reflect all of the various lineages that contributed to the current situation. Abnormal conditions include insufficient differentiation along edges, in which case some of the edges may not be reconstructible, leading to polytomies and thus to nodes of outdegree larger than 2. All three types of problems may lead to the reconstruction of networks that violate the constraints of Definition 12.5.

(The distinction between a model phylogeny and a reconstructible phylogeny is common with trees as well: for instance, model trees are always rooted, whereas reconstructed trees are usually unrooted. In networks, both the model network and the reconstructed network must be rooted: reticulations only make sense with directed edges.) Clearly, then, a reconstructible network will require changes from the definition of a model network. In particular, the degree constraints must be relaxed to allow arbitrary outdegrees for both network nodes and internal tree nodes. In addition, the time coexistence property must be reconsidered.

There are at least two types of problems in reconstructing phylogenetic networks. First, slow evolution may give rise to edges so short that they cannot be reconstructed, leading to polytomies. This problem cannot be resolved within the DAG framework, so we must relax the constraints on the outdegree of tree nodes. Secondly, missing data may lead methods to reconstruct networks that violate indegree constraints or time coexistence. In such cases, we can postprocess the reconstructed network to restore compliance with most of the constraints in the following three steps:

1. For each network node w with outdegree larger than 1, say with edges (w, v_1), $\ldots$, (w, v_k), add a new tree node u with edge (w, u) and, for each i, $1 \leq i \leq k$, replace edge (w, v_i) by edge (u, v_i).
2. For each network node w whose parents u and v violate time coexistence, add two tree nodes w_u and w_v and replace the two network edges (u, w) and (v, w) by four edges: the two tree edges (u, w_u) and (v, w_v) and the two network edges (w_u, w) and (w_v, w).
3. For each edge (u, v) where both u and v are network nodes, add a new tree node w and replace the edge (u, v) by the two edges (u, w) and (w, v).

The resulting network is consistent with the original reconstruction, but now satisfies the outdegree requirement for network nodes, obeys time coexistence (the introduction of tree edges on the paths to the network node allow arbitrary time delays), and no longer violates the requirement that at least one endpoint of each edge be a tree node. Moreover, this postprocessing is unique and quite simple. We can thus define a reconstructible network in terms similar to a model network.

Definition 12.6 *A* reconstructible phylogenetic network *is a rooted DAG obeying the following constraints:*

1. *Every node has indegree and outdegree defined by one of the three (indegree,outdegree) combinations $(0, x)$, $(1, y)$, or $(z, 1)$, for $x \geq 1$, $y \geq 0$, and $z \geq 2$—corresponding to, respectively, root, other tree nodes (internal nodes and leaves), and network nodes.*
2. *If two nodes u and v cannot co-exist in time, then there does not exist a network node w with edges (u, w) and (v, w).*
3. *Given any edge of the network, at least one of its endpoints must be a tree node.*

Definition 12.7 *A network N induces a tree T' if T' can be obtained from N by the following two steps:*

1. *For each network node in N, remove all but one of the edges incident into it; and*
2. *for every node v such that $indegree(v) = outdegree(v) = 1$, the parent of v is u, and the child of v is w, remove v and the two edges (u, v) and (v, w), and add new edge (u, w) (this is referred to in the literature as the* forced contraction *operation).*

For example, the network N shown in Figure 12.1(e) induces both trees shown in Figure 12.2(a) and Figure 12.2(d).

12.4.3 Reconstructing networks from gene trees

From a graph-theoretic point of view, the problem can be formulated as pure phylogenetic network reconstruction (**???**). In the case of HGT, and despite the fact the evolutionary history of the set of organisms is a network, **?** showed that an underlying species tree can still be inferred. In this case, a phylogenetic network is a pair (T, Ξ), where T is the species (organismal) tree, and Ξ is a set of HGT edges whose addition to T results in a phylogenetic network N that induces all the gene trees. The problem can be formulated as follows.

Definition 12.8 *(The HGT Reconstruction Problem)*

> **Input:** *A species tree ST and a set G of gene trees.*
>
> **Output:** *Set Ξ of minimum cardinality whose addition to ST yields phylogenetic network N that induces each of the gene trees in G.*

However, in the case of hybrid speciation, there is no underlying species tree; instead the problem is one of reconstructing a phylogenetic network N that induces a given set of gene trees.

Definition 12.9 *(The Hybrid Speciation Reconstruction Problem)*

> **Input:** *A set G of gene trees.*
>
> **Output:** *A Phylogenetic network N with minimum number of network nodes that induces each of the gene trees in G.*

The minimization criterion reflects the fact that the simplest solution is sought; in this case, the simplest solution is one with the minimum number of HGT or hybrid speciation events. We illustrate this point with the example species tree ST in Figure 12.3(a) and the gene tree GT in Figure 12.3(b). Assume that the actual HGT events that took place are the one depicted in Figure 12.3(c). Nonetheless, the scenario depicted in Figure 12.3(d) has fewer HGT events, yet induces GT. In this case, the solution in Figure 12.3(d) is the one sought by the HGT Reconstruction Problem. Although the scenarios depicted in Figure 12.3(c) and Figure 12.3(d) are very different, inferring the one in Fig 12.3(c) as the correct solution, in the absence of

(a) (b) (c) (d)

Figure 12.3: (a) A species tree ST. (b) A tree GT of a horizontally transferred gene. Both networks in (c) and (d) are formed based on ST, and both induce GT. However, even though the actual HGT scenarios that took place are described by the network in (c), the HGT Reconstruction Problem seeks the solution in (d).

any additional biological knowledge about the organisms, would be rather arbitrary. Hence, based on the species and gene tree topologies, solving the HGT Reconstruction Problem offers the “best” solution. Another serious problem that impacts the identifiability of reticulate evolution is that of extinction and incomplete taxon sampling. **?** illustrated some of the scenarios that lead to non-identifiability of reticulation events from a set of gene trees.

? gave an efficient algorithm for solving the HGT Reconstruction Problem; however, their algorithm handles limited special cases of the problem in which the number of HGT events is very small, and the number of times a gene is transferred is very low (also, their tool handles only binary trees; **?**). **?** gave efficient algorithms for solving the Hybrid Speciation Reconstruction Problem, but for constrained phylogenetic networks, referred to as *gt-networks*; further, they handled only binary trees. **?** have recently introduced RIATA-HGT, which is the first method for solving the general case of the HGT Reconstruction Problem. The method can be easily modified to yield a heuristic for solving the Hybrid Speciation Reconstruction Problem. We now describe the method and its empirical performance.

RIATA-HGT: reconstructing HGT from gene trees

We describe the algorithm underlying RIATA-HGT in terms of a species tree and a gene tree. The core of RIATA-HGT is the divide-and-conquer algorithm Compute-HGT algorithm (outlined in Figure 12.4). The algorithm starts by computing the $MAST$, T', of the species tree ST and gene tree GT; tree T' forms the basis for detecting and reconstructing the HGT events (computing T' is done in Step 1 in Figure 12.4). The algorithm then decomposes the clade sets U_1 and U_2 (whose removal from ST and GT, respectively, yields T') into maximal clades such that each maximal clade in one of the two sets is “matched” by a maximal clade on the same leaf set in the second set. The algorithm for this decomposition is outlined in Figure 12.5. The algorithm then recurses on each maximal clade and its matching maximal clade (Steps 5.c.(1) and 5.d.(5).(1) in Figure 12.4) to compute the HGT events whose recipients form sub-clades of those maximal clades. Finally, we add a single HGT event per each maximal clade to connect it to its “donor” in the ST; this is achieved through the calls to AddSingleHGT (Figure 12.6) in Steps 5.c.(2) and 5.d.(5).(3) in Figure 12.4.

```
PROCEDURE COMPUTEHGT($ST$,$GT$)
**Input:** Species tree $ST$, and gene tree $GT$, both on the same set $S$ of taxa.
**Output:** Computes the set $\Xi$ of HGT events such that the pair $(ST, \Xi)$ induces $GT$.
1. $T' = MAST(ST, GT)$;
2. **If** $T' = ST$ **then**
   (a) **Return**;
3. $U_1 = ST - T'$; $U_2 = GT - T'$;
4. $V = \emptyset$;
5. **Foreach** $u_2 \in U_2$
   (a) $Decompose(U_1, u_2, T', V)$;
6. $U_2 = V$;
7. **While** $V \neq \emptyset$
   (a) Let $u_2$ be an element of $V$;
   (b) Let $u_1 \in U_1$ be such that $L(u_2) \subseteq L(u_1)$;
   (c) $Y = \{y \in U_2 : L(y) \cap L(u_1) \neq \emptyset\}$;
   (d) $Z = \{y|(L(y) - L(u_1)) :\ y \in Y\}$;
   (e) $V = V - Y$; $V = V \cup Z$;
   (f) $X = \{u_1|L(y) : y \in Y\}$;
   (g) **Foreach** $y \in Y$
       i. Let $x \in X$ be such that $L(x) \cap L(y) \neq \emptyset$;
       ii. $ComputeHGT(x, y)$;
       iii. $AddSingleHGT(ST, GT, y, U_2, T')$;
```

Figure 12.4: The main algorithm for detecting and reconstructing HGT events based on a pair of species tree and gene tree.

Theoretically, RIATA-HGT may not compute the minimum-cardinality set of HGT events; **?** established the following properties of the method.

Theorem 12.1 *Given a species tree ST and a gene tree GT, the network N obtained by running RIATA-HGT on (ST, GT) induces GT. Further, RIATA-HGT takes $O(n^4)$ time in the worst case, where n is the number of leaves in ST.*

Moreover, experimental results show very good empirical performance on synthetic

```
PROCEDURE DECOMPOSE($U_1$,$u_2$, $T$, $U'$)
Input: Set $U_1$ of clades from $ST$, clade $u_2$ from $GT$, the backbone clade $u_2$, and
$U'$ which will contain the “refined” clades of $u_2$.
Output: Decompose $u_2$ so that no clade in $U'$ has a leaf set that is the union of
leaf sets of more than one clade in $U_1$.
1. If $\exists u_1 \in U_1$ such that $L(u_2) \subseteq L(u_1)$ then
   (a) $U' = U' \cup \{u_2\}$;
   (b) $B(u_2) = T$;
   (c) Return $u_2$;
2. Else
   (a) If $\exists u_1 \in U_1$ such that $r(u_2) = r(u_2|L(u_1))$
       i. $t = u_2|L(u_1)$;
       ii. $U' = U' \cup \{t\}$;
       iii. $B(t) = T$;
       iv. Let $X = u_2 - t$;
       v. Foreach $x \in X$
          A. $Decompose(U_1, x, t, U')$;
       vi. Return $t$;
   (b) Else
       i. Let $c_1, \ldots, c_k$ be the children of $r(u_2)$;
       ii. $x = Decompose(U_1, T_{c_1}, T, U')$;
       iii. For $i = 2$ to $k$
          A. $Decompose(U_1, T_{c_i}, x, U')$;
       iv. Return $x$;
```

Figure 12.5: The algorithm for decomposing the clades in U_1 and U_2 such that for all $u_1 \in U_1$ and $u_2 \in U_2$ we have $L(u_1) \not\subseteq L(u_2)$.

data, as illustrated in Figure 12.7. The whisker-and-box plot in Figure 12.7(a) shows the individual numbers of HGT events as predicted by RIATA-HGT versus the actual numbers. Figure 12.7(b) shows the average (of 30 runs) numbers of HGT events as predicted by RIATA-HGT versus the actual numbers (for full details of how the simulation studies were conducted and detailed analyses of the results, please refer to **?**). The plots demonstrate empirically the excellent performance of RIATA-HGT; it estimates the exact number of HGT events in a great majority of the cases, with very mild over- or under-estimation in the other cases. Over-estimation is an artifact of the heuristic nature of RIATA-HGT, whereas under-estimation is an artifact of the parsimony criterion in the definition of the problem (see the discussion above). RIATA-HGT was also applied to the bacterial gene datasets reported in **?**, and produced the results hypothesized by Lerat *et al.* In summary, RIATA-HGT performed very well on the synthetic datasets, as well as on the biological datasets.

```
PROCEDURE ADDSINGLEHGT($ST, GT, u_2, U_2, T'$)
Input: Species tree $ST$, gene tree $GT$, clade $u_2$ of $GT$, set $U_2$ of clades of $GT$, and $MAST\ T'$ of $ST$ and $GT$.
Output: Add to $\Xi$ a single HGT event whose donor is determined in this procedure and whose recipient is clade $u_2$.

1. $Q = L(u_2) \cup L(B(u_2))$;
2. $T'' = GT|Q$; $p = lca_{T''}(L(u_2))$;
3. $tq = lca_{ST}(L(u_2))$; $te = inedge_{ST}(tq)$;
4. If $p$ is a child of $r(T'')$ and $|L(B(u_2))| > 1$ then
    (a) $sq = lca_{ST}(L(B(u_2)))$;
    (b) $\Xi = \Xi \cup (sq \rightarrow te)$;
5. Else
    (a) $O = \bigcup_{\{p' : p' = sibling_{T''}(p)\}} L(T_{p'})$;
    (b) $sq = lca_{ST}(O)$; $se = inedge_{ST}(sq)$;
    (c) $\Xi = \Xi \cup (se \rightarrow te)$;
```

Figure 12.6: The algorithm for detecting and reconstructing the single HGT event in which clade u_2 is the recipient.

12.5 The Coalescent and Reticulate Evolution

12.5.1 The coalescent and lineage sorting in ancestral populations

Intra-species events (*i.e.*, gene duplication and loss) occur because of random contribution of each individual to the next generation. Some fail to have offsprings (gene loss) while some happen to have multiple offsprings (gene duplication). This means a number of duplication and loss events occur every generation. In population genetics, this process was first modeled by R. A. Fisher and S. Wright, in which each gene of the population at a particular generation is chosen independently from the gene pool of the previous generation, regardless of whether the genes are in the same individual or in different individuals.

Under the Wright-Fisher model, "the coalescent" considers the process backward in time (**???**). That is, the ancestral lineages of genes of interest are traced from offsprings to parents. A coalescent event occurs when two (or sometimes more) genes are originated from the same parent, which is called the most recent common ancestor (MRCA) of the two genes. This event corresponds to gene duplication when the process is considered forward in time. Gene loss events can be ignored in the coalescent process, because we are not interested in the lineages that do not exist at present.

The basic process can be treated as follows. Consider a pair of genes at time τ_1 in a random mating haploid population. The population size at time τ is denoted by

(a) (b)

Figure 12.7: The results of RIATA-HGT on synthetic datasets. (a) A box-and-whisker plot for the predictions of HGT event numbers made by RIATA-HGT. (b) The averages of HGT event numbers estimated by RIATA-HGT vs. the actual number of HGT events. Each point is the average of 30 runs of RIATA-HGT.

Figure 12.8: An illustration of the coalescent process in a three species model with discrete generations. The process is considered backward in time from present, T_0, to past. Circles represent haploid individuals. We are interested in the gene tree of the three genes (haploids) from the three species. Their ancestral lineages are represented by closed circles connected by lines. A coalescent event occurs when a pair of lineages happen to share a single parental gene (haploid).

$N(\tau)$. The probability that the pair are from the same parental gene at the previous generation (time τ_1+1) is $1/N(\tau_1+1)$. Therefore, starting at τ_1, the probability that the coalescence between the pair occurs at τ_2 is given by

$$Prob(\tau_2) = \frac{1}{N(\tau_2)} \sum_{\tau=\tau_1+1}^{\tau_2-1} \left(\frac{1}{N(\tau)} \right). \tag{12.1}$$

When $N(\tau)$ is constant, the probability density distribution (pdf) of the coalescent time (*i.e.*, $t=\tau_2-\tau_1$) is given by a geometric distribution, and can be approximated by an exponential distribution for a large N:

$$Prob(t) = \frac{1}{N} e^{-t/N}. \tag{12.2}$$

The coalescent process is usually ignored in phylogenetic analysis, but has a significant effect (causing lineage sorting) when closely related species are considered

(a) (b)

Figure 12.9: (a) The probabilities of the three types of gene tree, (AB)C, (AC)B, and A(BC), as functions of $(T_2 - T_1)/N$. (b) The probabilities that the gene tree is resolved from DNA sequence data. The probabilities are given functions of the mutation rate for the three types of tree, (AB)C, (AC)B, and A(BC), when $(T_2 - T_1)/N = 0.5$. The white regions represent the probabilities that the gene tree is not resolved.

(**???**). The situation of Figure 12.1(b) is reconsidered under the framework of the coalescent in Figure 12.8. Here, it is assumed that species A and B split $T_1 = 5$ generations ago, and the ancestral species of A and B and species C split $T_2 = 19$ generation ago. The ancestral lineage of a gene from species A and that from B meet in their ancestral population at time $\tau = 6$, and they coalesce at $\tau = 35$, which predates T_2, the speciation time between (A, B) and C. The ancestral lineage of B enters in the ancestral population of the three species at time $\tau = 20$, and first coalesces with the lineage of C. Therefore, the gene tree is represented by $A(BC)$ while the species tree is $(AB)C$. That is, the gene tree and species tree are “incongruent”. Under the model in Figure 12.8, the probability that the gene tree is congruent with the species tree is 0.85, which is one minus the product of the probability that the ancestral lineages of A and B do not coalesce between $\tau = 6$ and $\tau = 9$, and the probability that the first coalescence in the ancestral population of the three species occur between (A and C) or (B and C). The former probability is $\frac{14}{15}\frac{12}{13}\frac{11}{12}...\frac{7}{8}\frac{7}{8} = 0.22$ and the latter is $\frac{2}{3}$.

Under the three-species model (Figure 12.8), there are three possible types of gene tree, $(AB)C$, $(AC)B$ and $A(BC)$. Let $Prob$[(AB)C], $Prob$[(AC)B] and $Prob$[A(BC)] be the probabilities of the three types of gene tree. These three probabilities are simply expressed with a continuous time approximation when all populations have equal and constant population sizes, N, where N is large:

$$Prob[\text{(AB)C}] = 1 - \frac{2}{3}e^{-(T_2-T_1)/N}, \tag{12.3}$$

and

$$Prob[\text{(AC)B}] = Prob[\text{A(BC)}] = \frac{1}{3}e^{-(T_2-T_1)/N}. \tag{12.4}$$

Figure 12.9(a) shows the three probabilities as functions of $(T_2 - T_1)/N$.

An interesting application of this three species problem is in hominoids; A: human, B: chimpanzee and C: gorilla. It is believed that the species three is $(AB)C$. **?** investigated DNA sequences from 88 autosomal intergenic regions, and the gene tree is estimated for each region. They found that 36 regions support the species tree, $(AB)C$, while 10 estimated trees are $(AC)B$ and 6 are $A(BC)$. No resolution is obtained for the remaining 36 regions (see below). It is possible to estimate the time between two speciation events, $T_2 - T_1$, assuming all populations have equal and constant diploid population sizes, N (**?**). Since 36 out of 52 gene trees are congruent with the species tree, $T_2 - T_1$ is estimated to be $-\text{ln}[(3/2)(36/52)] = 0.77$ times $2N$

(a) (b)

Figure 12.10: (a) A three species model with a HGT event. A demonstration that a congruent tree could be observed even with HGT. (b) The probabilities of the three types of gene tree, (ab)c', (ac')b, and a(bc'), as functions of T_h/N. $T_1 = 2N$ and $T_2 = 3N$ are assumed.

generations. It should be noted that $2N$ is used for the coalescent time scale instead of N because hominoids are diploids. If we assume N to be $5 \times 10^4 - 1 \times 10^5$ (**??**), the time between two speciation events is $7.7 - 15.5 \times 10^4$ generations, which is roughly $1 - 3$ million years assuming a generation time of $15 - 20$ years.

It is important to notice that the estimation of the gene tree from DNA sequence data is based on the nucleotide differences between sequences, and that the gene tree is sometimes unresolved. One of the reasons for that is a lack of nucleotide differences such that DNA sequence data are not informative enough to resolve the gene tree. This possibility strongly depends on the mutation rate. Let μ be the mutation rate per region per generation, and consider the effect of mutation on the estimation of the gene tree. We consider the simplest model of mutations on DNA sequences, the infinite site model (**?**), in which mutation rate per site is so small that no multiple mutations at a single site are allowed. Consider a gene tree, $(AB)C$, and suppose that we have a reasonable outgroup sequence such that we know the sequence of the MRCA of the three sequences. It is obvious that mutations on the internal branch between the MRCA of the three and the MRCA of A and B are informative. If at least one mutation occurred on this branch, the gene tree can be resolved from the DNA sequence alignment. This effect is investigated by assuming that the number of mutations on a branch with length t follows a Poisson distribution with mean μt. Figure 12.9(b) shows the probability that the gene tree is resolved; $T_2 - T_1 = 0.5N$ generations is assumed so that the probability that the gene tree is $(AB)C$ is about 0.6. As expected, as the mutation rate increases, the probability that the gene tree is resolved from the sequence alignment increases, and this probability exceeds 90% when $N\mu > 1.52$. Similar results are obtained for the other two types of trees, $(AC)B$ and $A(BC)$, that appears with probability 0.2 for each (see also Figure 12.9(b)).

12.5.2 Gene trees, species trees and reticulate evolution

In the previous section, we have shown that the gene tree is not always identical to the species tree. With keeping this in mind, let us consider the effect of horizontal gene transfer (HGT) on gene tree under the framework of the coalescent.

The application of the coalescent theory to bacteria is straightforward. Bacterial evolution is better described by the Moran model rather than the Wright-Fisher model because bacteria do not fit a discrete generation model. Suppose that each haploid individual in a bacterial population with size N has a lifespan that follows an exponential distribution with mean l. When an individual dies, another individual randomly chosen from the population replaces it to keep the population size constant. In other words, one of the $N-1$ alive lineages is duplicated to replace the dead one. Under the Moran model, the ancestral lineages of individuals of interest can be traced backward in time, and the coalescent time between a pair of individuals follows an exponential distribution with mean $lN/2$ (**??**). This means that one half of the mean lifetime in the Moran model corresponds to one generation in the Wright-Fisher model.

It may usually be thought that HGT can be detected when the gene tree and species tree are incongruent (see Section 12.4). However, the situation is complicated when lineage sorting is also involved. Consider a model with three species, A, B, and C, in which an HGT event occurs from species B to C. Suppose the ancient circular genome has a single copy of a gene as illustrated in Figure 12.10(a). Let a, b and c be the focal orthologous genes in the three species, respectively. At time T_h, a gene escaped from species B and was inserted in a genome in species C at T_i, which is denoted by c'. Following the HGT event, c was physically deleted from the genome, so that each of the three species currently has a single copy of the focal gene.

If there is no lineage sorting, the gene tree should be $a(bc')$. Since this tree is incongruent with the species tree, $(AB)C$, we could consider it as an evidence for HGT. However, as demonstrated in Section 12.2, lineage sorting could also produce the incongruence between the gene tree and species tree without HGT. It is also important to note that lineage sorting, coupled with HGT, could produce congruent gene tree, as illustrated in Figure 12.10(a). Although b and c' have more chance to coalesce first, the probability that the first coalescence occurs between a and b or between a and c' may not be negligible especially when $T_1 - T_h$ is short.

The probabilities of the three types of gene tree can be formulated under this trispecies model with HGT as illustrated in Figure 12.10(a). Here, T_h could exceed T_1, in such a case it can be considered that HGT occurred before the speciation between A and B. Assuming that all populations have equal and constant population sizes, N, the three probability can be obtained modifying (12.3) and (12.4):

$$Prob[\text{(AB)C}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \le T_1 \\ 1-\frac{2}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}, \qquad (12.5)$$

$$Prob[\text{(AC)B}] = \begin{cases} \frac{1}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \le T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}, \qquad (12.6)$$

and

$$Prob[\text{A(BC)}] = \begin{cases} 1-\frac{2}{3}e^{-(T_1-T_h)/N} & \text{if } T_h \le T_1 \\ \frac{1}{3}e^{-(T_h-T_1)/N} & \text{if } T_h > T_1 \end{cases}. \qquad (12.7)$$

Figure 12.10(b) shows the three probability assuming $T_1 = 2N$ and $T_2 = 3N$.

Thus, lineage sorting due to the coalescent process works as a noise for detecting and reconstructing HGT based on gene tree, sometimes mimicking the evidence for HGT and sometimes creating a false positive evidence for HGT. Therefore, to distinguish HGT and lineage sorting, statistics based on the theory introduced in this chapter is needed. We only considered very simple cases with three species here, but it is straightforward to extend the theory to more complicated models.

12.6 Summary

In this chapter, we have reconsidered the gene tree species tree problem in the context of reticulate evolution. In particular, we discussed gene tree incongruence due to reticulate evolution and presented our recent heuristic, RIATA-HGT, for resolving this type of incongruence. Further, we have addressed extensions of the coalescent model to incorporate non-treelike evolutionary events, such as horizontal gene transfer. Gene tree incongruence is both an obstacle impeding accurate phylogeny reconstruction and a tool for detecting and reconstructing evolutionary events such as HGT and hybrid speciation. Future directions for further research include:

1. Testing the performance of existing methods for resolving gene tree incongruence in the context of intra- and inter-species evolutionary events.
2. Developing and testing accurate and fast methods for reconstructing phylogenetic networks from gene trees under the conditions of incomplete taxon sampling and missing orthologs.
3. Extending our initial progress on the coalescent model beyond three species and to incorporate hybrid speciation and meiotic recombination.

CHAPTER 13

From genomes to organisms: integrating genomic data

Cristian I. Castillo-Davis
***

13.1 Introduction

Biology has gone from being a data-poor science to a data-rich one and thus presents an exciting challenge not only for biologists but also for statisticians, computer scientists and other quantitative workers. The prodigious and ever-growing bounty of “-omic” data generated by technologies enabled by whole genome DNA sequencing projects is quickly out-pacing our ability to digest and meaningfully synthesize it. These data include transciptional, proteomic, and phenotypic data, to name but a few.

However, recent work has shown that a biologically and statistically thoughtful combination of different data types in either a hypothesis-driven or data-mining framework can lead to a deeper, more comprehensive understanding of biology. Post- genomic analysis, the interpretation and synthesis of thousands of data points from a chemical, clinical, evolutionary, or other perspective thus promises to be an area of great methodological and scientific development in this century.

For many genes, something is known about their molecular and biological function, pathway membership, physical chromosomal location, level of polymorphism, RNAi phenotype, disease phenotype, and rate of molecular evolution. For non-coding regions, data are often available concerning the presence of known or putative transcription factor binding sites, levels of DNA methylation or acetylation, and GC content. While freely available through public databases, these different kinds of biological data are often unexamined with respect to each other. One reason for this situation is a lack of conceptual and methodological tools for their analysis. The continual release of new genomic and proteomic datasets insures that this situation will only be exacerbated in the coming decades. At the same time, this problem offers an unprecedented opportunity for innovation and scientific discovery not only for biologist but for statisticians, computer scientists, and others.

Since there is no one solution to the problem of integrating high-throughput genomic data, and since the types of data available will undoubtedly change over time, I will concentrate on familiarizing the reader with specific examples where integrative post-genomic analysis has been successfully applied, and highlight key areas of investigation that are especially fertile for future contribution. In doing so, out of familiarity, I will use examples largely from my own work. My goal is not a comprehensive review of the literature but an illustration of some of the applications, challenging problems, and exciting possibilities of combining different types of genomic data toward biological ends.

13.2 Case-study I — Intron evolution

To illustrate a relatively straightforward case of hypothesis-driven post-genomic analysis that uses disparate data-types in its execution we will begin with an example involving the evolution of gene structure Castillo-Davis et al. (2002).

13.2.1 Background

Introns are intervening sections of DNA within protein-coding regions of genes that do not encode amino-acids (Figure 13.1) and are primarily made up of non-functional "junk DNA." These sections of DNA are nonetheless transcribed (copied) by the cell along with the protein-coding sections (known as exons) into messenger RNA (mRNA) as one long transcript. The introns in an mRNA transcript are subsequently cut out of the transcript (literally) and the exons spliced together (literally) to form the usable mRNA message.

This mRNA transcript is later translated into an amino-acid chain which then folds to make a protein. Some of the largest introns are found in the human genome, where the total length of intron sequences in a gene often reaches tens of thousands of nucleotides such that transcription of a single gene requires several minutes and thousands of ATP molecules (the energy currency of the cell).

13.2.2 Hypothesis

Because transcription is a slow and metabolically expensive process in eukaryotes, it was hypothesized that, at least for highly expressed genes, transcription of long introns, might be energetically costly. If so, in genes that are highly expressed, it is predicted that natural selection will favor shorter introns. To test this hypothesis requires at least two sets of data: 1) data on gene structure detailing the sizes of exons and introns making up all the genes in a genome, and 2) estimates of the expression level of each gene.

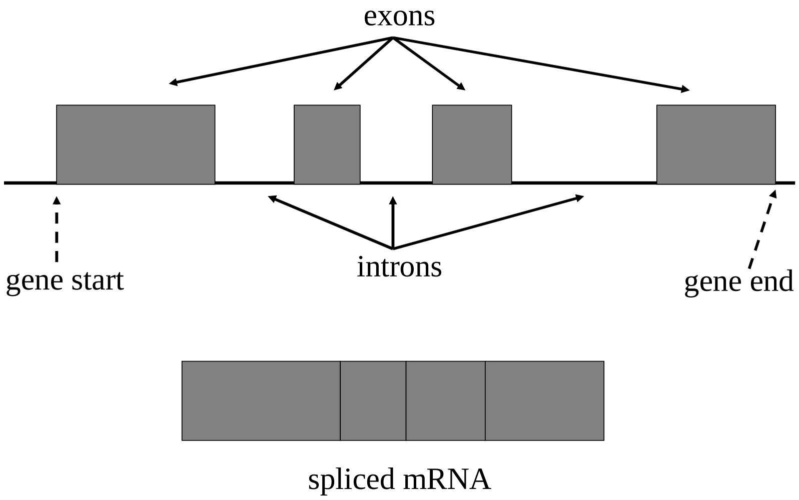


Figure 13.1: Exon-intron structure of a gene

13.2.3 Methods

At the time of this study, sufficient information on both exon-intron structure and gene expression data were available only for two species: the nematode *Caenorhabditis elegans* and human. Gene structure information for each species was available through genome databases and consisted of coordinates listing exon and intron boundaries. In terms of expression data, for *C. elegans*, Affymetrix microarray expression data collected over development was available that provided absolute transcript abundance measures for each gene. Unfortunately, such microarray experiments were not available for human, and gene expression was instead estimated by expressed sequence tag abundance.

Expressed sequence tags (ESTs) are short stretches of DNA, randomly sequenced after reverse transcribing a pool of mRNA that is typically extracted from a tissue or organ. Since some mRNA transcripts are more abundant than others, these will be sequenced more often, and in turn will end up making up the bulk of sequences in an EST database. By aligning the known DNA sequence of a given gene with EST sequences in an EST database and counting the number of significant matches, one can estimate the expression level of that gene Bortoluzzi and Danieli (1999). This was the approach taken in this study to estimate gene expression level in human, using BLAST Altschul et al. (1997) for sequence alignment and all available human EST sequences in GenBank Benson et al. (2005).

13.2.4 Result

By combining information on intron size and the two types of expression data discussed above, it was found that introns in highly expressed genes were indeed substantially shorter than those of genes expressed at low levels in both in human and *C. elegans* (Figure 13.2).

13.2.5 Discussion

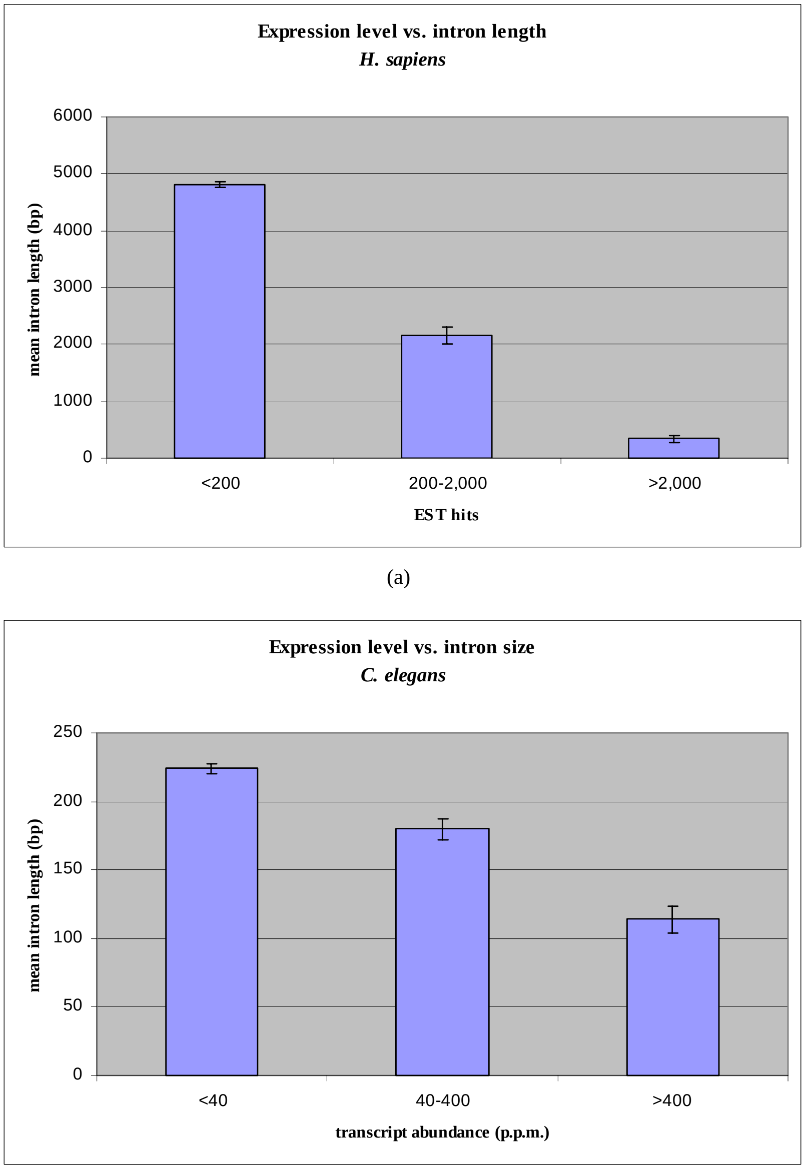
In this case study, the authors had a very specific hypothesis in mind and attempted to test its predictions using available data. No sophisticated modeling was used nor were high-level statistics necessary to obtain the biological results. This case study shows that the evaluation of important biological hypotheses is possible with a minimal amount of disparate genomic data (in this case, three types) if the data are combined in a biologically and statistically thoughtful manner. Many important biological questions remain unanswered even in the wake of an abundance of genomic data. I hope this inspires workers outside and on the periphery of biology to apply novel tools and fresh perspectives to genomic investigation. The opportunity for methodological and scientific contribution are great.

13.3 Case Study II — Functional genomics and protein evolution

To illustrate a case of post-genomic analysis that is more data-mining in spirit and that utilizes a number of different data types, we now turn to a study on protein conservation and function Castillo-Davis et al. (2004). This study is largely aimed at answering three basic questions: “What are the slowest evolving (most conserved) proteins in animal genomes and what do they do?” and “What are the fastest evolving (least conserved) proteins in animal genomes and what do they do?” And finally, “Are fast and slow evolving genes the same types of genes in different animals?”

13.3.1 Background

An important question in biology is how selective forces act on the genome in the evolution of different species. For example, does natural selection act similarly on proteins across lineages as distinct as phyla? Since most multicellular organisms contain a similar complement of genes and gene families owing, in part, to a common cellular biology, it might be expected that natural selection acts homogeneously across functionally similar genes in widely disparate taxa. However, this is not certain and there are many reasons why inhomogeneous levels of conservation across the proteome might be expected in different animals for example strong lineage-specific adaptation. To address this question we need first to determine the rate of evolution of all genes in two different animals and second, integrate this information with data on gene function.



(a)

(b)

Figure 13.2: Mean expression level versus intron size in (a) *H. sapiens* and (b) *C. elegans*

13.3.2 Methods

Rates of evolution for two species pairs in two different animal phyla, Chordata (mouse/human) and Nematoda (*C. elegans*/*C. briggsae*) (Figure 13.3) were estimated by the maximum likelihood method of Yang and colleagues Goldman and Yang (1994) Nielsen and Yang (1998). This method calculates the estimated rate of nonsynonymous (amino-acid changing) substitutions between proteins d_N and the synonymous (non amino-acid changing) rate of substitution d_S between proteins.

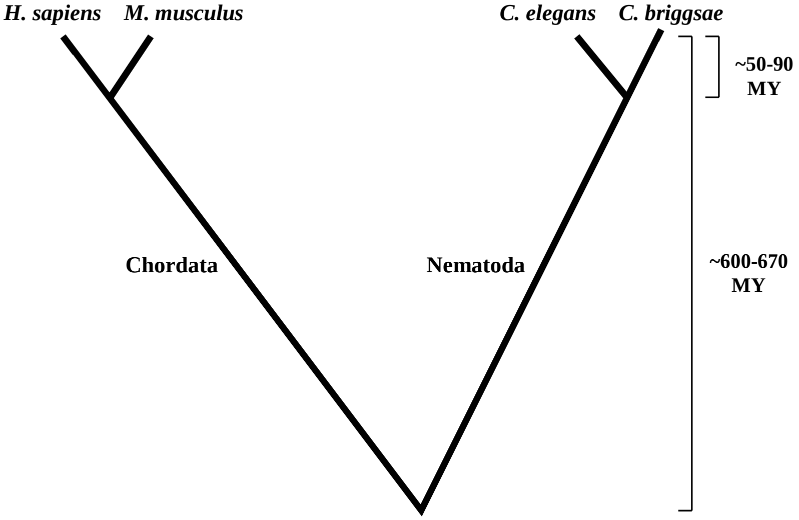


Figure 13.3: Evolutionary relationships and divergence times of species studied in case study II. MY = million years.

These data were subsequently examined with respect to gene function using two complementary approaches. In the first approach, a list of the top 10% fastest and slowest evolving proteins in each species pair (in terms of d_N were compared with known gene functions from the Gene Ontology (GO) database Ashburner et al. (2000b) (http://www.geneontology.org) and tested for statistical enrichment. In the second, tissue-specific expression of all genes in the mammalian dataset was estimated based on hits to EST sequence libraries and then rates of evolution for genes expressed in each tissue type were calculated.

The statistical enrichment of various functional classes among slow and fast evolving genes was evaluated using GeneMerge Castillo-Davis and Hartl (2003) (http://www.oeb.harvard.edu/hartl/lab/publications/ Annotated gene functions from the Gene Ontology Consortium Ashburner et al. (2000b) for human and *C. elegans* were used as input for GeneMerge. GeneMerge related methods will be discussed in greater depth later in the chapter.

EST data were obtained from cDNA libraries available in GenBank (http://www.ncbi.nlm.nih.gov).

More than 450,000 ESTs from 12 normal adult mouse tissues were collected and alignments evaluated against each mouse gene using BLASTN Altschul et al. (1997). Genes with significant hits to ESTs were then normalized and clustered into tissue-specific groups by means of a Self-Organizing Tree Algorithm (SOTA) Herrero et al. (2001). Clusters represent genes that have similar expression patterns across tissues (Figure 13.4). Mean divergence estimates were then calculated for each cluster with confidence intervals estimated by means of nonparametric bootstrap re-sampling with 1,000 replicates.

13.3.3 Results

The 10% fastest evolving genes in mammals, according to the GO annotations, were largely involved in reproduction, immunity, and signal transduction (Table 2), whereas transcription factors were over-represented among fast evolving nematode proteins (Table 3).

GO Description	*Fraction*	*P-value*	*GO ID*
Immune response	100/577	3.77E-040	GO:0006955
Response to pest/pathogen/parasite	61/577	2.76E-023	GO:0009613
Antimicrobial humoral response	24/577	4.84E-013	GO:0019730
Response to wounding	27/577	2.68E-006	GO:0009611
Innate immune response	20/577	0.000357	GO:0045087
Inflammatory response	19/577	0.001230	GO:0006954
Lymphocyte activation	7/577	0.008820	GO:0046649
Pregnancy	8/577	0.009790	GO:0007565

Table 2. Functional overrepresentation of fast evolving mammal genes.

GO Description	*Fraction*	*P-value*	*GO ID*
DNA-dependent regul. of transcription	45/753	4.27E-5	GO:0006355
Regulation of transcription	45/753	5.72E-5	GO:0045449
Nucleic acid metabolism	53/753	0.03675	GO:0006139

Table 3. Functional overrepresentation of fast evolving worm genes.

Corroborating these results, the EST data (Figure 13.4) showed that genes co-expressed in the thymus and spleen (immune organs) in mouse evolved the fastest among all tissues d_N = 0.142. Additionally, an accelerated mean rate of evolution was seen in genes co-expressed in the ovary and uterus d_N = 0.122, organs with a reproductive role.

In contrast, the slowest-evolving genes in both nematodes and mammals were involved in *the same* basic cellular processes including protein biosynthesis, cell growth and GTP-mediated signal transduction (Table 4,5).

GO Description	*Fraction*	*P-value*	*GO ID*
Protein metabolism	140/699	5.76E-10	GO:0019538
Intracellular protein transport	44/699	5.89E-9	GO:0006886
Small GTPase mediated sign. transd.	30/699	1.85E-7	GO:0007264
Ubiquitin-dependent protein catabolism	25/699	4.81E-6	GO:0006511
Biosynthesis	69/699	0.000284	GO:0009058
Nucleocytoplasmic transport	13/699	0.000461	GO:0006913
Metabolism	265/699	0.001011	GO:0008152
mRNA splicing	10/699	0.039416	GO:0006371

Table 4. Functional overrepresentation of slow evolving mammal genes.

GO Description	*Fraction*	*P-value*	*GO ID*
Physiological processes	268/753	4.04E-12	GO:0007582
Protein biosynthesis	48/753	1.82E-11	GO:0006412
Cellular process	132/753	4.89E-11	GO:0009987
Biosynthesis	63/753	5.25E-11	GO:0009058
Small GTPase mediated sign. transd.	23/753	7.89E-11	GO:0007264
Metabolism	189/753	2.17E-5	GO:0008152
Protein metabolism	92/753	3.01E-5	GO:0019538
mRNA splicing	5/753	0.016747	GO:0006371

Table 5. Functional overrepresentation of slow evolving worm genes.

Thus it appears that while fast-evolving genes tend to be lineage-specific, highly-conserved genes are the same in different types of animals and are mainly involved in core cellular functions.

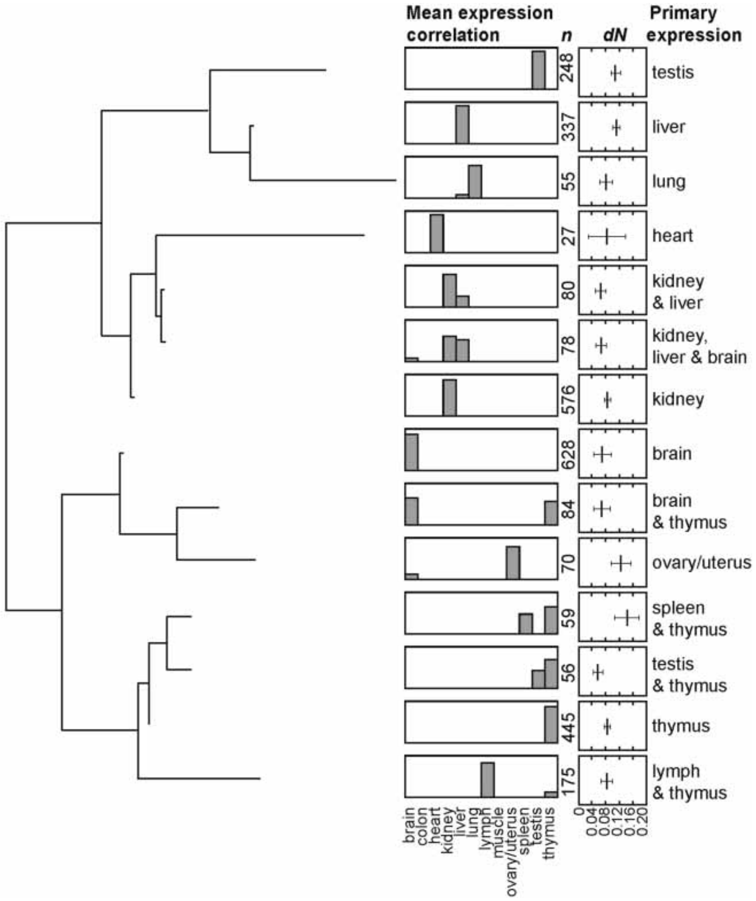


Figure 13.4: Tissue-specific gene expression and protein divergence. Histograms show the mean correlation coefficient for gene expressed in a cluster. Reproduced with permission from Cold Spring Harbor Laboratory Press Castillo-Davis et al. (2004).

13.3.4 Discussion

Leaving aside the biological implications of the study we will concentrate on the methods used to integrate the comparative and functional genomic data. Firstly, note that a two-pronged and complementary approach was taken to establish gene function. Database annotations are currently incomplete with upwards of 50% of genes having unknown function, even in model organisms. Thus, it was important in this study to complement the database annotation data with a method using all genes, even those with unknown functions. The EST-based tissue-specific expression analysis satisfied this goal. In general when combing genomic data— which are often noisy or incomplete— similar strategies of data complementation are often useful since certain data types can bolster deficiencies in others.

13.4 Toward general methods for data-combination and exploration

Having reviewed two case-studies involving the combination of disparate data types, we now turn to a more general discussion of methods to combine and analyze genomic data. Data associated with genes are many and varied and will undoubtedly grow as genomic and proteomic investigations accelerate. To deal with this explosion of data requires 1) a clear analytical framework and 2) the flexibility to examine new data as soon as they become available. To date, there are very few approaches that meet both these criteria.

However, one approach that has been quite fruitful is the so-called over-representation framework where investigators examine the overlap of particular attributes in a sample of genes drawn from a larger set of genes, often a genome. By far, the most common application of this approach is the examination of a list of genes that are found to be highly expressed in, say, breast cancer tissue versus normal breast tissue, for statistical over-representation of gene functions within the list. There are several programs that implement this general algorithm Martin et al. (2004) (Table 6) using functions provided by the Gene Ontology Consortium; the most commonly implemented statistic to assess overrepresentation is based on the hypergeometric distribution Martin et al. (2004).

$$\Pr(r|n,p,k) = \frac{\binom{pn}{r}\binom{(1-p)n}{k-r}}{\binom{n}{k}} \tag{13.1}$$

The hypergeometric distribution describes the discrete probability of selecting r items of one kind in a sample of size k from a population of size n, where p is equal to the proportion of r type items in the population, and sampling is without replacement. The hypergeometric thus gives quantification of the level of ones surprise at finding overrepresentation for a particular item in a given sample of size k drawn from the larger population, size n. k is typically a set of genes that are highly or lowly expressed and n is the population set, the set from which k is drawn, usually all genes on a particular microarray. *P*-values can be calculated by summing over the tail of the distribution for all less-likely cases.

$$\sum_{i=r}^{k} \Pr(i|n,p,k) \tag{13.2}$$

Since several to hundreds of gene attributes are usually tested for overrepresentation in a given analysis, correction for multiple hypothesis testing is important. For example, if we were to test whether a set of genes involved in brain cancer were found disproportionately on a particular chromosome by testing for overrepresentation on each of the 22 non-sex human chromosomes, we would have carried out 22 separate hypothesis tests. Thus, some kind of *P*-value correction must be made. The traditional Bonferroni correction is the most popular Martin et al. (2004), but, less severe corrections, such as those based on the False Discovery Rate (FDR) Benjamini and Hochberg (1995b) Storey (2002a), are becoming more common.

Program	*Stat.*	*Mult. Test. Corr.*
CLENCH	Hypergeometric*	NA
FatiGO	Fisher exact test	FDR
FuncAssociate	Fisher exact test	*P*-value adjus.
FuncSpec	Hypergeometric	Bonferroni
GeneMerge	Hypergeometric	Bonferroni
GFINDer	Hypergeometric*	Bonferroni
GoMiner	Fisher exact test	NA
Gostat	Fisher exact test	Holm/FDR/Yekutieli
GO Term-Finder (CPAN)	Hypergeometric	Bonferroni/FDR
GOTM	Hypergeometric	NA
GOToolBox	Hypergeometric*	Bonferroni

Table 6. Overrepresentation tools that use Gene Ontology annotations; from Martin et al. (2004); see references therein for associated publications). * indicates software is capable of other statistical tests as well.

The first general-purpose implementation of the overrepresentation approach to genomic data, GeneMerge Castillo-Davis and Hartl (2003) was designed with the express purpose of combining many different types of data related to genes, and thus will be our focus here. In GeneMerge the study set k may be genes found to be significantly up or down-regulated in a microarray experiment or a list of genes deemed interesting *for any another reason*. Genes in the sample k are associated with particular identifiers, for example functions, processes, or states. The number of genes with a particular identifier is r. p is the fraction of genes in the population n associated with the particular identifier under investigation.

GeneMerge returns both descriptive information regarding the genes under investigation and Bonferroni corrected and uncorrected rank scores regarding overrepresentation of any number of different descriptors in a given set of genes. Functional or categorical descriptive data is associated with genes in *gene-association* files. These text files link each gene in a genome with a particular datum of information. For example, the name of a gene and its chromosomal location, its sensitivity to a particular small-molecule, or its identity as over-expressed in a particular type of cancer.

The use of overrepresentation techniques has been quite useful when applied to microarray data using GO gene functions (for example Ranz et al. (2003) Pletcher et al. (2002) and many, many more) and genetic pathway membership (for example Cavalieri et al. (2000) and many, many more). Interestingly however, this method has been less often used for data exploration and hypothesis testing of more diverse gene-association data, for example, mutation phenotypes, microarray expression outcomes, and genetic interactions. A partial list of gene-association data potentially useful for different genomic analyses is given in Table 7. Unfortunately, most software implementations do not allow users to generate and utilize a wide range of gene-association data. One advantage of GeneMerge over other similar programs is that its gene-association files are simply tab-delimited text files that can be prepared using any spreadsheet program.

Gene-association Data
knock-out phenotype
disease phenotype
polymorphic / non-polymorphic locus
local recombination rate
expression phenotype under influence of chemical X
publication mention
transcription factor binding sites
protein-protein network connectivity (degree)
viability/inviability if deleted
acetylated under condition X
GC content
sex-specific expression
tissue-specific expression
has ortholog in clade X
rate of molecular evolution
genetic interactions with other genes
over/under-expressed in experiment X
alternatively spliced
RNAi phenotype
expressed in anatomical region X

Table 7. A partial list of gene-association data.

To illustrate how the overrepresentation approach can be applied to data beyond the traditional microarray expression/GO function paradigm, two examples are given below. These are followed by a discussion of the limitations and possible extensions of the method as well as potential new approaches for the combination and examination of disparate genomic data.

13.4.1 Overrepresentation methods — beyond microarray data

To illustrate the flexibility of overrepresentation techniques as applied to genomic data, I will present two unpublished examples, one involving a population genomics question and another involving literature mining.

Population genomics

Recent analysis of genomic data suggests that protein evolution is related to protein effects on organism fitness; specifically, it has been shown that proteins that cause lowered fitness when deleted in yeast, so called “non-dispensable” genes, tend to evolve more slowly Hirsh and Fraser (2001). While amino acid substitution rates tend to be higher in dispensable genes over long evolutionary distances Hirsh and Fraser (2001), it is not known whether in natural populations these genes also tend to be more *polymorphic*, that is, show more inter-individual variation. Given that selection against deleterious mutations is also expected to operate at the population level, coupled with the observation that variation among natural populations is ultimately transformed into variation among species, we may predict that dispensable genes will be more polymorphic within populations. In other words, non-essential genes will show more variation than essential genes in a population.

Polymorphic genes were identified using genomic hybridizations to Affymetrix arrays containing 126,645 unique 25mer yeast probes among 14 strains of laboratory and wild yeast Winzeler et al. (2003). These arrays are sensitive to the detection of single nucleotide polymorphisms. Unfortunately, distinction between synonymous and nonsynonymous substitutions is not possible with these arrays. Genes with at least one detected polymorphism among the 14 strains were considered polymorphic. Genes with no polymorphism were considered non-polymorphic. Among the 2991 genes probed on the chip, 1874 (63%) were polymorphic by this criterion. To create a deletion viability gene association file, lists of genes that result in inviablity or are viable when deleted were obtained from the Saccharomyces Genome Database (http://genome-www.stanford.edu/Saccharomyces/) based on the data of Winzeler et al. (1999) and Giaever et al. (2002). 4713 genes were listed as having a deletion viable phenotype and 1115 genes an inviable deletion phenotype. 413 genes had no data available concerning deletion phenotype. The hypothesis that population level polymorphism is more likely to occur in dispensable genes appears to be supported by the data.

Among *S. cerevisiae* genes categorized as polymorphic, more are viable upon deletion than is expected by chance. Of the 1874 genes categorized as polymorphic, 1454 (77%) were deletion viable, representing an enrichment in this class of genes ($P < 0.006$) (Table 8).

Description	*Fraction*	*P-value*	*corr. P-value*	*ID*
deletion inviable	336/1874	0.3880	0.7760	del_inv
deletion viable	1454/1874	0.0025	0.0051	del_via

Table 8. Polymorphism and deletion viability in yeast.

Thus selection against deleterious mutations in potentially more important genes appears to result in visibly lower levels of polymorphism at the population level. While this result is preliminary, it provides one example of how overrepresentation approaches can be used to explore genomic hypotheses efficiently in data beyond the microarray/GO function paradigm.

Literature mining

Using word frequency to extract meaning from a corpus of literature is a mainstay of text-mining techniques. In terms of genomic analysis, for example, Jenssen and colleagues Jenssen et al. (2001) used the frequency of co-occurrence of gene names in scientific abstracts to generate a gene-to-gene co-citation network that can be used in the analysis of microarray data. Conversely, others have used literature-mining techniques to asses whether clusters of particular genes share a common biological function Raychaudhuri et al. (2002). Since literature also constitutes a type of gene-association data, albeit of a more complex kind, it is possible to use overrepresentation-based approaches to mine literature as well.

One example of this strategy uses abstracts from scientific publications and extracted keywords to look for overrepresentation of keywords among publications associated with gene lists (Hong, Liu, Wong, Castillo-Davis, unpublished). In this work, approximately 10 million literature abstracts associated with all genes under analysis were extracted from PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi) and filtered for keywords by excluding all non-technical words using a generic dictionary word list. Next, the overrepresentation of keywords in papers associated with the sample set of genes versus the population set was assessed using the hypergeometric distribution. This simple approach was effective when used to examine a set of genes with known enrichment in developmental functions in human (Table 9). This literature-based method generated much more detailed information on gene function and biological sub-processes, than, for instance, GO annotations (data not shown). These included specific gene names (BMP, NOTCH, MYC, NOGGIN), functional regions (enhancer, homeobox), and potentially interesting disease relationships (Parkinsonism).

Keyword	*P*-value
hif	6.14E-63
myc	8.86E-36
parkin	3.59E-33
gata	1.11E-20
bmp	2.71E-19
morphogenetic	2.46E-17
notch	2.46E-17
eph	3.18E-17
ephrin	8.99E-16
transcription	1.82E-15
malformation	3.90E-15
parkinsonism	8.12E-14
twist	3.96E-12
homeodomain	1.22E-10
hox	1.31E-10
differentiation	1.47E-09
noggin	1.64E-09
developmental	3.90E-09
homeobox	9.25E-09
enhancer	1.44E-08
morphogenesis	5.86E-08

Table 9. Literature-based over-representation results for developmental genes.

Different implementations of literature-based overrepresentation methods along these lines and others Muller et al. (2004) are likely to be increasingly useful for extracting biological meaning from genomic data.

13.5 Limitations and possible solutions

A major drawback of most over-representation methods is the discretization of data into binary categories. In case study II for example, a decision had to be made about what constituted slow and fast evolving genes. The authors chose to test for functional overrepresentation among the 10% slowest and then the 10% fastest evolving genes. While the results of the study were not particularly sensitive to the 10% cut-off (data not shown) such a cut-off may not be desired in other contexts. For example, for other purposes it might be interesting to comprehensively examine the relationship between evolutionary rate and gene function. One could choose *a priori* a set of biological functions and some level of granularity in the GO hierarchy and then calculate rates of evolution of genes in each category. Going a step further one could calculate rates of evolution for genes in every functional category at all levels in the GO hierarchy. While informative, note that interrogation of the data in this manner has moved us from a hypothesis-testing mode to a data-mining mode.

The visualization of the results of such a comprehensive analysis also present a difficult problem. How can one visualize the discrete and structured functional relationships inherent in the GO hierarchy and the continuous evolutionary rate information or other such variables, perhaps several, all at the same time? A bubble graph illustrated in Figure 13.5 is one possible solution. In this figure, each node represents a particular GO function and edges connect functions in accordance with relationships of the GO function tree (a so-called directed acyclic graph, or DAG). The size of each node indicates the mean rate of amino-acid substitution (d_N) for genes within the node— that is, the rate of evolution of genes with a particular function. In this hypothetical example, genes with known transcription factor (TF) activity exhibit a faster rate of evolution than other types genes.

Note that the number of genes within each node may be different and genes may belong to multiple nodes. Thus, while we have improved our understanding of relationship between evolutionary rate and gene function in the data, we have done so at the expense of statistical power; all possible relationships between evolutionary rate and function have been explored. Such trade-offs are likely to be common and must be weighed by the aims of the study, specifically, whether the ultimate goal is hypothesis-testing or data-mining.

While some gene attributes are discrete, such as on which chromosome a gene resides, others are continuous, such as the expression level of a gene in a particular tissue or its relative level of evolutionary conservation. In these cases the usefulness of categorical statistical tests such as those based on the hypergeometric distribution are called into question. Suppose, for example, that instead of rigidly assigning gene function in a boolean manner, one could assign probabilities concerning gene function to genes. How might one design a test for overrepresentation in this case? While regression-based techniques have been applied on a case-by-case basis to particular problems (for instance Liu et al. (2002) there is as of yet no general algorithm available for the interrogation and comparison of disparate data types with continuous values or with a mixture of continuous and categorical values.

The development of such a framework will be challenging, in particular because the categorical structure of some types of biological data can be complex, such as the GO DAG. Interestingly however, directed and undirected graphs are often extremely

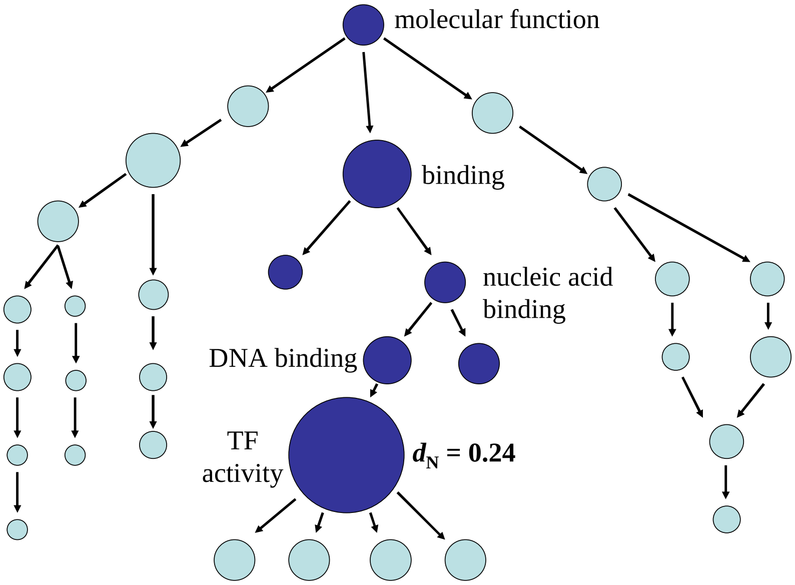


Figure 13.5: Bubble graph representation of the relationship between a continuous variate (rate of evolution, d_N) and a graphical structure (the GO functional hierarchy).

natural representations of gene functions and gene interactions, for example, protein-protein interaction networks Uetz et al. (2000a) Ito et al. (2001) Li et al. (2004) Giot et al. (2003) and metabolic and developmental pathways Kanehisa et al. (2002). The addition of weights to graph edges or variance measurements for individual nodes will only increase the complexity of analyzing such data. The development of statistical tests and data-exploration methods, perhaps akin to overrepresentation techniques, will be critical in exploiting these types of data.

Equally important in the analysis of disparate genomic data is data visualization. How best to visualize several dimensions of the data simultaneously, some of which may have complex structures? Some overrepresentation-based tools have begun to address this issue by creating dynamic output that maps, for instance, overrepresentation *P*-values onto the GO hierarchy, for example, the“GO Term Finder” of the Saccharomyces Genome Database Boyle et al. (2004) (Figure 13.6, or the up- or down-regulation of genes onto a metabolic pathway using the KEGG database Kanehisa et al. (2002), for example, Pathway Processor Grosu et al. (2002). Unfortunately, these visualization solutions are species- or gene-association-specific and have not yet been generalized.

A particularly important challenge will be the analysis of high-throughput phenotypic data in combination with genomic data. Phenotypic data, including anatomical sections of Medicine (U.S.), three-dimensional CT scans, and MR images have

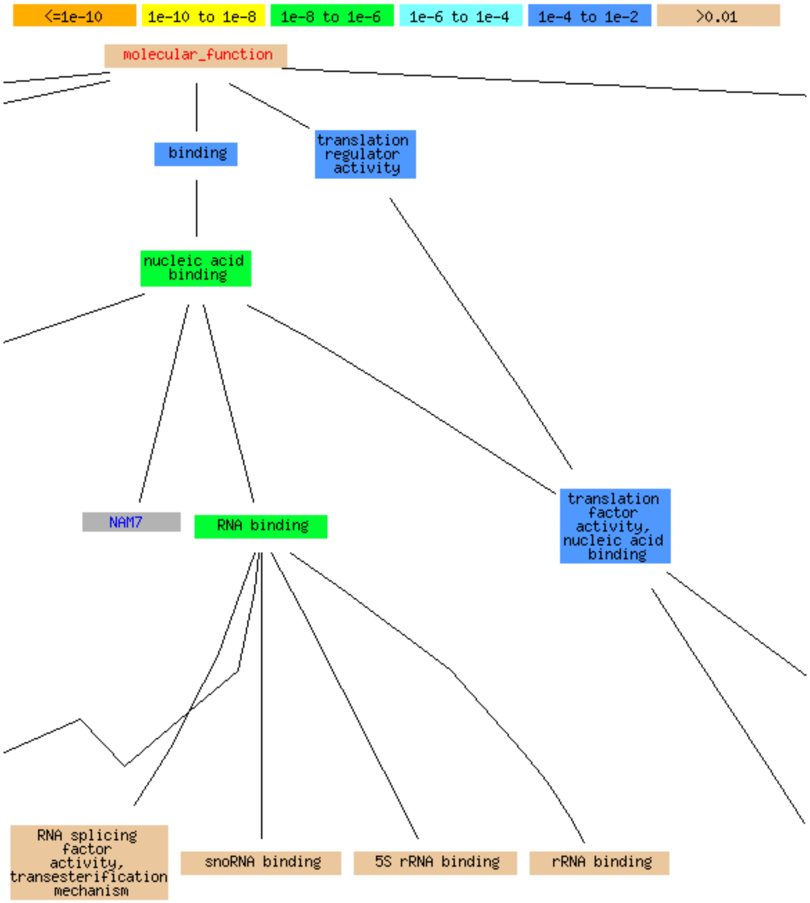


Figure 13.6: Partial graphical output of the SGD GoTermFinder Boyle et al. (2004).

even more complex structures than graphs and network diagrams. Incorporating genomic and proteomic data with the mixture of continuous and discontinuous spatial information inherent in morphological data will be challenging. Flexible visualization and statistical techniques that allow for the input and processing of standardized structural information (in the form of graphs, network diagrams, or 3-dimensional volumes) along with the requisite gene lists and relevant gene-association data is desperately needed. Contribution from many different disciplines, including computer imaging, scientific visualization, and biostatistics, especially areas related to morphometrics, will be required to achieve a comprehensive understanding of how genotype and organism phenotype are related.

13.6 Summary

Despite an abundance of genomic, proteomic, and increasingly, gross phenotypic data, many straightforward biological questions remain difficult to answer due to the complex and varied nature of these data. As we have seen, overrepresentation techniques and related methods, when applied creatively and critically, hold some promise in helping shed light on this tangled surfeit of biological information. In particular, the use of more and varied gene-association data with these methods promises to be quite powerful for data-mining and cursory hypothesis testing applications. At the same time, the limitations of these approaches are many; highly structured data in the form of gene networks, morphological data, protein-protein interactions, and simply the growing dimensionality of biological measurements in genome-wide studies strain the conceptual and statistical limits of the overrepresentation framework. Many challenges remain in assimilating complex biological data structures into current statistical and data-mining approaches. Data visualization will be an additional challenge. Progress will likely require heavy cross-disciplinary collaboration amongst statisticians, biologists, and computer scientists, among others. The expansion and application of statistical and graphical approaches to the analysis of genomic data presents numerous, rich opportunities for intellectual contribution. With luck, these advances will help expedite the larger goal of deciphering nature's profound complexity.

13.7 Acknowledgements

I wish to thank Jun S. Liu for support during my post-doctoral fellowship at Harvard, Pengyu Hong for collaborating on the literature mining work discussed, Bjarki Eldon, Gordon Kindlmann, and Rima Izem for help with LaTeX, and Ping Ma, Lei Shen, Gordon Kindlmann, Christian Landry, and Rima Izem for valuable comments on the manuscript.

References

Affymetrix, *Affymetrix: Microarray Suite User's Guide, version 5.0*, Santa Clara, CA, 2001.

Al-Sharour, F., R. Diaz-Uriarte, and J. Dopazo, FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes, *Bioinformatics*, 20:578–580, 2004.

Allison, D.B. and M. Heo, Meta-analysis of linkage data under worst-case conditions: a demonstration using the human OB region, *Genetics*, 148:859–865, 1998.

Alter, O., P.O. Brown, and D. Botstein, Singular value decomposition for genome-wide expression data processing and modeling, *Proceedings of the National Academy of Sciences USA*, 97:10101–10106, 2000.

Altmuller, J. et al., Genomewide scans of complex human diseases: True linkage is hard to find, *American Journal of Human Genetics*, 69:936–950, 2001.

Altschul, S.F. et al., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res*, 25:3389–3402, 1997.

An, P. et al., Genome-wide linkage scans for fasting glucose, insulin, and insulin resistance in the National Heart, Lung, and Blood Institute Family Blood Pressure Program: evidence of linkages to chromosome 7q36 and 19q13 from meta-analysis, *Diabetes*, 54:909–914, 2005.

Ashburner, M. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genetics*, 25:25–29, 2000a.

Ashburner, M. et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nat Genet*, 25:25–29, 2000b.

Babron, M.C. et al., Meta and pooled analysis of European coeliac disease data, *European Journal of Human Genetics*, 11:828–834, 2003.

Badner, J.A. and E.S. Gershon, Meta-analysis of whole-genome linkage scans of bipolar disorder and schizophrenia, *Molecular Psychiatry*, 7:405–411, 2002a.

Badner, J.A. and E.S. Gershon, Regional meta-analysis of published data supports linkage of autism with markers on chromosome 7, *Molecular Psychiatry*, 7:56–66, 2002b.

Bailey, K.R., Inter-study differences: how should they influence the interpretation and analysis of results?, *Statistics in Medicine*, 6:351–358, 1987.

Balasubramanian, R. et al., A graph-theoretic approach to testing associations between disparate sources of functional genomics data, *Bioinformatics*, 20:3353–3362, 2004.

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

Barrett, T. et al., NCBI GEO: mining millions of expression profiles-database and tools, *Nucleic Acids Research*, 33:D562–D566, 2005.

Beer, D.G. et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine*, 9:816–824, 2002.

Beissbarth, T. and T.P. Speed, GOstat: find statistically overrepresented Gene Ontologies within a group of genes, *Bioinformatics*, 20:1464–14655, 2004.

Benito, M. et al., Adjustment of systematic microarray data biases, *Bioinformatics*, 20:105–114, 2004.

Benjamini, Y. and Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society B*, 57:289–300, 1995a.

Benjamini, Y. and Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Statist. Soc. B*, 57:289–300, 1995b.

Benjamini, Y. and W. Liu, A step-down multiple hypothesis testing procedure that controls the false discovery rate under independence, *Journal of Statistical Planning and Inference*, 82:163–170, 1999.

Benjamini, Y. and D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *Annals of Statistics*, 29:1165–1188, 2001.

Benson, D.A. et al., GenBank, *Nucleic Acids Res*, 33:34–38, 2005.

Bhattacharjee, A. et al., Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses, *Proceedings of the National Academy of Sciences USA*, 98:13790–13795, 2001.

Bolstad, B., *affyPLM: affyPLM - Probe Level Models*, 2004, URL `http://www.stat.berkeley.edu/users/bolstad/AffyExtensions`, r package version 1.2.5.

Bolstad, B.M. et al., A comparison of normalization methods for high density oligonucleotide array data based on bias and variance, *Bioinformatics*, 19:185–193, 2003.

Bortoluzzi, S. and G.A. Danieli, Towards an in silico analysis of transcription patterns, *Trends Genet*, 15:118–119, 1999.

Boyle, E.I. et al., GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, 20:3710–3715, 2004.

Brazma, A. et al., Minimum information about a microarray experiment (MIAME) – toward standards for microarray data, *Nature Genetics*, 29:373, 2001.

Bussemaker, H.J. and E.D. Siggia, Regulatory element detection using correlation with expression, *Nature Genetics*, 27:167–171, 2001.

Castillo-Davis, C.I. and D.L. Hartl, GeneMerge–post-genomic analysis, data mining, and hypothesis testing, *Bioinformatics*, 19:891–892, 2003.

Castillo-Davis, C.I. et al., The functional genomic distribution of protein divergence in two animal phyla: coevolution, genomic conflict, and constraint, *Genome Res*, 14:802–811, 2004.

Castillo-Davis, C.I. et al., Selection for short introns in highly expressed genes, *Nat Genet*, 31:415–418, 2002.

Cavalieri, D., J.P. Townsend, and D.L. Hartl, Manifold anomalies in gene expression in a vineyard isolate of Saccharomyces cerevisiae revealed by DNA microarray analysis, *Proc Natl Acad Sci U S A*, 97:12369–12374, 2000.

Chan, E.Y. et al., Increased huntingtin protein length reduces the number of polyglutamine-induced gene expression changes in mouse models of Huntington's disease, *Human Molecular Genetics*, 11:1939–1951, 2002.

Chiodini, B.D. and C.M. Lewis, Meta-analysis of 4 coronary heart disease genome-wide linkage studies confirms a susceptibility locus on chromosome 3q, *Arteriosclerosis Thrombosis and Vascular Biology*, 23:1863–1868, 2003.

Chipping Forecast, The chipping forecast, *Nature Genetics*, 21, 1999, URL `http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v21/n1s/index.html`.

Chipping Forecast, The chipping forecast II, *Nature Genetics*, 32, 2002, URL `http://www.nature.com/cgi-taf/DynaPage.taf?file=/ng/journal/v32/n4s/index.html`.

Choi, J.K. et al., Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, 19 Suppl 1:i84–i90, 2003.

Cochran, W.G., The combination of estimates from different experiments, *Biometrics*, 10:101–129, 1954.

Collin, F., Analysis of oligonucleotide data with a view to data quality assessment, Ph.D. thesis, Department of Statistics, University of California, Berkeley, 2004.

Conlon, E. M., L.X.S.L.J. and J.S. Liu, Integrating regulatory motif discovery and genome-wide expression analysis, *Proceedings of the National Academy of Sciences USA*, 100:3339–3344, 2003.

Cooper, H.M. and L.V. Hedges, *The Handbook of Research Synthesis*, Russell Sage Foundation, 1994.

Cooperative", T.T.M.S.G., A meta-analysis of genomic screens in multiple sclerosis, *Multiple Sclerosis*, 7:3–11, 2001.

Cope, L.M. et al., A benchmark for Affymetrix GeneChip expression measures, *Bioinformatics*, 20:323–331, 2004.

Cordell, H.J., Sample size requirements to control for stochastic variation in magnitude and location of allele-sharing linkage statistics in affected sibling pairs, *Annals of Human Genetics*, 65:491–502, 2001.

Cui, X. and G.A. Churchill, Statistical tests for differential expression in cDNA microarray experiments, *Genome Biology*, 4:210, 2003.

Cui, X. et al., Improved statistical tests for differential gene expression by shrinking variance components estimates, *Biostatistics*, 6:59–75, 2005.

Dabney, A. and J.D.S. with assistance from Gregory R. Warnes, *qvalue: Q-value estimation for false discovery rate control*, , r package version 1.1.

Demenais, F. et al., A meta-analysis of four european genome screens (GIFT consortium) shows evidence for a novel region on chromosome 17p11.2-q22 linked to type 2 diabetes, *Human Molecular Genetics*, 12:1865–1873, 2003.

Dempfle, A. and S. Loesgen, Meta-analysis of linkage studies for complex diseases: An overview of methods and a simulation study, *Annals of Human Genetics*, 68:69–83, 2004.

Diehn, M. et al., SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data, *Nucleic Acids Research*, 31:219–223, 2003.

Dobbin, K.K. et al., Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays, *Clinical Cancer Research*, 11:565–572, 2005.

Draghici, S. et al., Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate, *Nucleic Acids Research*, 31:3775–3781, 2003.

Dudoit, S., J.P. Shaffer, and J.C. Boldrick, Multiple hypothesis testing in microarray experiments, *Statistical Science*, 18:71–103, 2003.

Dudoit, S., M.J. van der Laan, and K.S. Pollard, Multiple testing, part i. Single-step procedures for control of general Type I error rates, *Statistical Applications in Genetics and Molecular Biology*, 3:Article 13, 2004.

Edgar, R., M. Domrachev, and A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Research*, 30:207–210, 2002.

Elkon, R. et al., Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells, *Genome Research*, 13:773–780, 2003.

Etzel, C. and R. Guerra, Meta-analysis of genetic-linkage analysis of quantitative-trait loci, *American Journal of Human Genetics*, 71:56–65, 2002.

Etzel, C.J., M. Liu, and T.J. Costello, An updated meta-analysis approach for genetic linkage, *BMC Genetics*, 6 Suppl 1:S43, 2005.

Everitt, B.S., L. Landau, and M. Leese, *Cluster Analysis*, Oxford University Press, 4th ed., 2001.

Faraway, J.J., Distribution of the admixture test for the detection of linkage under heterogeneity, *Genetic Epidemiology*, 10:75–83, 1993.

Feingold, E., P.O. Brown, and D. Siegmund, Gaussian models for genetic linkage analysis using complete high-resolution maps of identity by descent, *American Journal of Human Genetics*, 53:234–251, 1993.

Fisher, R.A., *Statistical methods for research workers*, London: Oliver & Lloyd, 1925.

Fisher, R.A., *Statistical Methods for Research Workers*, Oxford University Press, 4th ed., 1932.

Fisher, S.A. et al., Meta-analysis of genome scans of age-related macular degeneration, *Human Molecular Genetics*, 14:2257–2264, 2005.

Fisher, S.A., J.S. Lanchbury, and C.M. Lewis, Meta-analysis of four rheumatoid arthritis genome-wide linkage studies - confirmation of a susceptibility locus on chromosome 16, *Arthritis and Rheumatism*, 48:1200–1206, 2003.

Folks, L.J., Combination of independent tests, in Krishnaiah, P.R. and P.K. Sen, eds., *Handbook of Statistics*, vol. 4, pp. 113–121, New York: North-Holland, 1984.

Fox, J., *car: Companion to Applied Regression*, 2005, URL `http://www.r-project.org,http://socserv.socsci.mcmaster.ca/jfox/`, r package version 1.0-17.

GAMES and T.T.M.S.G. Cooperative", A meta-analysis of whole genome linkage screens in multiple sclerosis, *Journal of Neuroimmunology*, 143:39–46, 2003.

Ge, Y., S. Dudoit, and T.P. Speed, Resampling-based multiple testing for microarray data analysis (with discussion), *Test*, 12:1–77, 2003.

Gentleman, R.C. et al., BioConductor: Open software development for computational biology and bioinformatics, *Genome Biology*, 5:R80, 2004, URL `http://genomebiology.com/2004/5/10/R80`.

Ghosh, D., Mixture models for assessing differential expression in complex tissues using microarray data, *Bioinformatics*, 20:1663–1669, 2004.

Ghosh, D. et al., Statistical issues and methods for meta-analysis of microarray data: a case study in prostate cancer, *Functional and Integrative Genomics*, 3:180–188, 2003.

Giaever, G. et al., Functional profiling of the Saccharomyces cerevisiae genome, *Nature*, 418:387–391, 2002.

Giot, L. et al., A protein interaction map of Drosophila melanogaster, *Science*, 302:1727–1736, 2003.

Goldman, N. and Z. Yang, A codon-based model of nucleotide substitution for protein-coding DNA sequences, *Mol Biol Evol*, 11:725–736, 1994.

Goldstein, D.R. and M. Delorenzi, Statistical design and data analysis for microarray experiments, in Berger, A. and M.A. Roberts, eds., *Unravelling Lipid Metabolism with Microarrays*, New York: Dekker, 2004.

Grosu, P. et al., Pathway Processor: a tool for integrating whole-genome expression results into metabolic networks, *Genome Res*, 12:1121–1126, 2002.

Gu, C. et al., Meta-analysis methodology for combining non-parametric sibpair linkage results: genetic homogeneity and identical markers, *Genetic Epidemiology*, 15:609–626, 1998.

Gu, C., M.A. Province, and D.C. Rao, Meta-analysis of genetic studies, in Rao, D.C.

and M.A. Province, eds., *Genetic Dissection of Complex Traits: Challenges for the Next Millennium*, pp. 255–272, San Diego: Academic Press, 2001.

Guerra, R. et al., Meta-analysis by combining p-values: simulated linkage studies, *Genetic Epidemiology*, 17 Suppl 1:S605–S609, 1999.

Hanahan, D. and R.A. Weinberg, The hallmarks of cancer, *Cell*, 100:57–70, 2000.

Haseman, J.K. and R.C. Elston, The investigation of linkage between a quantitative trait and a marker locus, *Behavior Genetics*, 2:3–19, 1972.

Hasselblad, V., Meta-analysis of environmental health data, *Science of the Total Environment*, 160:545–558, 1995.

Heo, M. et al., A meta-analytic investigation of linkage and association of common leptin receptor (LEPR) polymorphisms with body mass index and waist circumference, *International Journal of Obesity and Related Metabolic Disorders*, 26:640–646, 2002.

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

Hirsh, A.E. and H.B. Fraser, Protein dispensability and rate of evolution, *Nature*, 411:1046–1049, 2001.

Huang, E. et al., Gene expression phenotypic models that predict the activity of oncogenic pathways, *Nature Genetics*, 34:226–230, 2003.

Huang, X. and W. Pan, Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays, *Functional and Integrative Genomics*, 2:126–183, 2002.

Ideker, T., T. Galitski, and L. Hood, A new approach to decoding life: systems biology, *Annual Review of Genomics and Human Genetics*, 2:343–372, 2001.

Ihaka, R. and R. Gentleman, R: A language for data analysis and graphics, *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, *Nature*, 409:860–921, 2001.

Irizarry, R.A. et al., Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, 31:e15, 2003a.

Irizarry, R.A. et al., *affy: Methods for Affymetrix Oligonucleotide Arrays*, 2004, r package version 1.5.8.

Irizarry, R.A. et al., Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, 4:249–264, 2003b.

Irizarry, R.A. et al., Multiple-laboratory comparison of microarray platforms, *Nature Methods*, 2:345–50, 2005.

Ito, T. et al., A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci U S A*, 98:4569–4574, 2001.

Iyengar, S. and J. Greenhouse, Selection models and the file drawer problem, *Statis-*

tical Science, 3:109–117, 1988.

Iyengar, S.K. et al., Improved evidence for linkage on 6p and 5p with retrospective pooling of data from three asthma genome screens, *Genetic Epidemiology*, 21 Suppl 1:S130–S135, 2001.

Jain, N. et al., Local pooled error test for identifying differentially expressed genes with a small number of replicated microarrays, *Bioinformatics*, 19:1945–1951, 2003.

Jansen, R. et al., A Bayesian networks approach for predicting protein-protein interactions from genomic data, *Science*, 302:449–453, 2003.

Jensen, F.V., *Bayesian Networks and Decision Graphs*, Springer-Verlag, 2001.

Jenssen, T.K. et al., A literature network of human genes for high-throughput analysis of gene expression, *Nat Genet*, 28:21–28, 2001.

Ji, Y. et al., RefSeq refinements of UniGene-based gene matching improves the correlation between microarray platforms, Tech. rep., MD Anderson Cancer Center, Department of Biostatistics and Applied Mathematics, 2005.

Jiang, H. et al., Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes, *BMC Bioinformatics*, 5:81, 2004.

Johnson, L. et al., Meta-analysis of five genome-wide linkage studies for body mass index reveals significant evidence for linkage to chromosome 8p, *International Journal of Obesity*, 29:413–419, 2005.

Kanehisa, M. et al., The KEGG databases at GenomeNet, *Nucleic Acids Res*, 30:42–46, 2002.

Kaufman, L. and P. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*, Wiley, 1990.

Koivukoski, L. et al., Meta-analysis of genome-wide scans for hypertension and blood pressure in Caucasians shows evidence of susceptibility regions on chromosomes 2 and 3, *Human Molecular Genetics*, 13:2325–2332, 2004.

Kooperberg, C. et al., Significance testing for small sample microarray experiments, *Statistics in Medicine*, 24:2281–2298, 2005.

Koziol, J.A. and A.C. Feng, A note on the genome scan meta-analysis statistic, *Annals of Human Genetics*, 68:376–380, 2004.

Kuo, W.P. et al., Analysis of matched mRNA measurements from two different microarray technologies, *Bioinformatics*, 18:405–412, 2002.

Lamb, J. et al., A mechanism of cyclin d1 action encoded in the patterns of gene expression in human cancer, *Cell*, 114:323–334, 2003.

Lander, E. and L. Kruglyak, Genetic dissection of complex traits – guidelines for interpreting and reporting linkage results, *Nature Genetics*, 11:241–247, 1995.

Lauritzen, S., *Graphical Models*, Oxford University Press, 1996.

Lee, H.K. et al., Coexpression analysis of human genes across many microarray data sets, *Genome Research*, 14:1085–1094, 2004.

Lee, T.I. et al., Transcriptional regulatory networks in Saccharomyces cerevisiae, *Science*, 298:799–804, 2002.

Levinson, D.F. et al., No major schizophrenia locus detected on chromosome 1q in a large multicenter sample, *Science*, 296:739–741, 2002.

Levinson, D.F. et al., Genome scan meta-analysis of schizophrenia and bipolar disorder, part I: Methods and power analysis, *American Journal of Human Genetics*, 73:17–33, 2003.

Lewis, C.M. et al., Genome scan meta-analysis of schizophrenia and bipolar disorder, part II: Schizophrenia, *American Journal of Human Genetics*, 73:34–48, 2003.

Li, C. and W.H. Wong, Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection, *Proceedings of the National Academy of Sciences USA*, 98:31–36, 2001.

Li, S. et al., A map of the interactome network of the metazoan C. elegans, *Science*, 303:540–543, 2004.

Li, Z. and D.C. Rao, Random effects model for meta-analysis of multiple quantitative sibpair linkage studies, *Genetic Epidemiology*, 13:377–383, 1996.

Liu, W., W. Zhao, and G.A. Chase, Genome scan meta-analysis for hypertension, *American Journal of Hypertension*, 17:1100–1106, 2004.

Liu, X.S., D.L. Brutlag, and J.S. Liu, An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments, *Nat Biotechnol*, 20:835–839, 2002.

Lockhart, D.J. et al., Expression monitoring by hybridization to high-density oligonucleotide arrays, *Nature Biotechnology*, 14:1675–1680, 1996.

Loesgen, S. et al., Weighting schemes in pooled linkage analysis, *Genetic Epidemiology*, 21 Suppl 1:S142–S147, 2001.

Lönnstedt, I. et al., Microarray analysis of two interacting treatments: a linear model', Tech. rep., Uppsala University, Sweden, Department of Mathematics, 2001.

Lönnstedt, I. and T.P. Speed, Replicated microarray data, *Statistica Sinica*, 12:31–46, 2002.

Lumley, T., *rmeta: Meta-analysis*, 2004, r package version 2.12.

Luthi-Carter, R. et al., Dysregulation of gene expression in the R6/2 model of polyglutamine disease: parallel changes in muscle and brain, *Human Molecular Genetics*, 11:1911–1926, 2002a.

Luthi-Carter, R. et al., Decreased expression of striatal signaling genes in a mouse model of Huntington´s disease, *Human Molecular Genetics*, 9:1259–1271, 2000.

Luthi-Carter, R. et al., Polyglutamine and transcription: gene expression changes shared by DRPLA and Huntington's disease mouse models reveal context-independent effects, *Human Molecular Genetics*, 11:1927–1937, 2002b.

Mah, N. et al., A comparison of oligonucleotide and cDNA-based microarray systems, *Physiological Genomics*, 16:361–370, 2004.

Mangiarini, L. et al., Exon 1 of the HD gene with an expanded CAG repeat is sufficient to cause a progressive neurological phenotype in transgenic mice, *Cell*, 87:493–506, 1996.

Marazita, M.L. et al., Meta-analysis of 13 genome scans reveals multiple cleft lip/palate genes with novel loci on 9q21 and 2q32-35, *American Journal of Human Genetics*, 75:161–173, 2004.

Marshall, E., Getting the noise out of gene arrays, *Science*, 306:630–631, 2004.

Martin, D. et al., GOToolBox: functional analysis of gene datasets based on Gene Ontology, *Genome Biol*, 5:R101, 2004.

Mecham, B.H. et al., Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements, *Nucleic Acids Research*, 32:e74, 2004a.

Mecham, B.H. et al., Increased measurement accuracy for sequence-verified microarray probes, *Physiological Genomics*, 18:308–315, 2004b.

Mok, S.C. et al., Prostasin, a potential serum marker for ovarian cancer: identification through microarray technology, *Journal of the National Cancer Institute*, 93:1458–1464, 2001.

Morris, J.S. et al., Pooling information across different studies and oligonucleotide microarray chip types to identify prognostic genes for lung cancer, in Shoemaker, J. and S.M. Lin, eds., *Methods of Microarray Data Analysis IV*, pp. 51–66, New York: Springer-Verlag, 2005.

Morton, N.E., Sequential tests for the detection of linkage, *American Journal of Human Genetics*, 7:277–318, 1955.

Muller, H.M., E.E. Kenny, and P.W. Sternberg, Textpresso: an ontology-based information retrieval and extraction system for biological literature, *PLoS Biol*, 2:e309, 2004.

Nielsen, R. and Z. Yang, Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene, *Genetics*, 148:929–936, 1998.

Nielsen, T.O. et al., Molecular characterization of soft tissue tumours: a gene expression study, *Lancet*, 359:1301–1307, 2002.

Normand, S.L., Meta-analysis: formulating, evaluating, combining and reporting, *Statistics in Medicine*, 18:321–359, 1999.

North, B.V., D. Curtis, and P.C. Sham, A note on the calculation of empirical P values from Monte Carlo procedures, *American Journal of Human Genetics*, 72:498–499, 2003.

O'Donovan, M.C., N.M. Williams, and M.J. Owen, Recent advances in the genetics of schizophrenia, *Human Molecular Genetics*, 12:R125–R133, 2003.

of Medicine (U.S.) Board of Regents, N.L., Electronic imaging: Report of the board of regents. U.S. department of health and human services, public health service, national institutes of health, NIH Publication 90-219, 1990.

Ott, J., *Analysis of Human Genetic Linkage*, Baltimore, MD: Johns Hopkins University Press, 3rd ed., 1999.

Pardi, F., D.F. Levinson, and C.M. Lewis, GSMA: software implementation of the genome search meta-analysis method, *Bioinformatics*, 21:4430–4431, 2005.

Parmigiani, G. et al., A statistical framework for molecular-based classification in cancer, *Journal of the Royal Statistical Society, Series B*, 64:717–736, 2002.

Parmigiani, G. et al., A cross-study comparison of gene expression studies for the molecular classification of lung cancer, *Clinical Cancer Research*, 10:2922–2927, 2004.

Pearson, K., On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random, *Biometrika*, 25:379–410, 1933.

Peri, S. et al., Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Research*, 13:2363–2371, 2003.

Pletcher, S.D. et al., Genome-wide transcript profiles in aging and calorically restricted Drosophila melanogaster, *Curr Biol*, 12:712–723, 2002.

Province, M.A., The significance of not finding a gene, *American Journal of Human Genetics*, 69:660–663, 2001.

Province, M.A. et al., A meta-analysis of genome-wide linkage scans for hypertension: the National Heart, Lung and Blood Institute Family Blood Pressure Program, *Biometrika*, 16:144–147, 2003.

R Development Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 200, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Ranz, J.M. et al., Sex-dependent gene expression and evolution of the Drosophila transcriptome, *Science*, 300:1742–1745, 2003.

Rao, D.C., CAT scans, PET scans, and genomic scans, *Genetic Epidemiology*, 15:1–18, 1998.

Rao, D.C. and C. Gu, False positives and false negatives in genome scans, *Advances in Genetics*, 42:487–498, 2001.

Raychaudhuri, S., H. Schutze, and R.B. Altman, Using text analysis to identify functionally coherent gene groups, *Genome Res*, 12:1582–1590, 2002.

Reiner, A., D. Yekutieli, and Y. Benjamini, Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics*, 19:368–375, 2003.

Rhodes, D.R. et al., Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer, *Cancer Research*, 62:4427–4433, 2002.

Rhodes, D.R. et al., Mining for regulatory programs in the cancer transcriptome, *Nature Genetics*, 37:579–583, 2005a.

Rhodes, D.R. et al., Probabilistic model of the human protein-protein interaction network, *Nature Biotechnology*, 23:951–959, 2005b.

Rhodes, D.R. et al., Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression, *Proceedings of the National Academy of Sciences USA*, 101:9309–9314, 2004a.

Rhodes, D.R. et al., ONCOMINE: a cancer microarray database and integrated data-mining platform, *Neoplasia*, 6:1–6, 2004b.

Rice, W.R., A consensus combined p-value test and the family-wide significance of component tests, *Biometrics*, 46:303–308, 1990.

Risch, N. and K. Merikangas, The future of genetic studies of complex human diseases, *Science*, 273:1516–1517, 1996.

Rosenthal, R., The ”file drawer problem” and tolerance for null results, *Psychological Bulletin*, 86:638–641, 1979.

Rosenthal, R., *Meta-Analytic Procedures for Social Research*, Beverly Hills, CA: Sage Press, 1984.

Sagoo, G.S. et al., Meta-analysis of genome-wide studies of psoriasis susceptibility reveals linkage to chromosomes 6p21 and 4q28-q31 in Caucasian and Chinese Hans population, *Journal of Investigative Dermatology*, 122:1401–1405, 2004.

Sarkar, S.K., Some results on false discovery rates in multiple testing procedures, *Annals of Statistics*, 30:239–257, 2002.

Sawcer, S. et al., Empirical genomewide significance levels established by whole genome simulations, *Genetic Epidemiology*, 14:223–229, 1990.

Segal, E. et al., A module map showing conditional activity of expression modules in cancer, *Nature Genetics*, 36:1090–1098, 2004.

Segurado, R. et al., Genome scan meta-analysis of schizophrenia and bipolar disorder, part III: Bipolar disorder, *American Journal of Human Genetics*, 73:49–62, 2003.

Shen, R., D. Ghosh, and A.M. Chinnaiyan, Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data, *BMC Genomics*, 5:94, 2004.

Simpson, E.H., The interpretation of interaction in contingency tables, *Journal of the Royal Statistical Society, Series B*, 13:238–241, 1951.

Smyth, G.K., Linear models and empirical Bayes methods for assessing differential expression in microarray experiments, *Statistical Applications in Genetics and Molecular Biology*, 3:Article 3, 2004.

Sorlie, T. et al., Repeated observation of breast tumor subtypes in independent gene expression data sets, *Proceedings of the National Academy of Sciences USA*, 100:8418–8423, 2003.

Stec, J. et al., Comparison of the predictive accuracy of DNA array based multigene classifiers across cDNA arrays and Affymetrix GeneChips, *Journal of Molecular Diagnosis*, 7:357–367, 2005.

Stevens, J.R. and R.W. Doerge, Combining Affymetrix microarray results, *BMC Bioinformatics*, 6:57, 2005.

Storey, J.D., A direct approach to false discovery rates, *J. R. Statist. Soc. B*, 64:479–498, 2002a.

Storey, J.D., A direct approach to false discovery rates, *Journal of the Royal Statistical Society, Series B*, 64:479–498, 2002b.

Storey, J.D. and R. Tibshirani, Statistical significance for genome-wide experiments, *Proceedings of the National Academy of Sciences USA*, 100:9440–9445, 2003.

Sutton, A.J. et al., *Methods for Meta-Analysis in Medical Research*, John Wiley & Sons, 2000.

Sweet-Cordero, A. et al., An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis, *Nature Genetics*, 37:48–55, 2005.

Tan, P.K. et al., Evaluation of gene expression measurements from commercial microarray platforms, *Nucleic Acids Research*, 31:5676–5684, 2003.

Tippett, L.H.C., *The Methods of Statistics*, London: Williams & Norgate, 1st ed., 1931.

Tukey, J.W., *Exploratory Data Analysis*, Addison-Wesley, 1977.

Tusher, V.G., R. Tibshirani, and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences USA*, 98:5116–5121, 2001.

Uetz, P. et al., A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, 403:623–627, 2000a.

Uetz, P. et al., A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae, *Nature*, 403:623–627, 2000b.

van der Laan, M.J., S. Dudoit, and K.S. Pollard, Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives, *Statistical Applications in Genetics and Molecular Biology*, 3:Article 15, 2004a.

van der Laan, M.J., S. Dudoit, and K.S. Pollard, Multiple testing. Part II. step-down procedures for control of the family-wise error rate, *Statistical Applications in Genetics and Molecular Biology*, 3:Article 14, 2004b.

van Heel, D.A. et al., Inflammatory bowel disease susceptibility loci defined by genome scan meta-analysis of 1952 affected relative pairs, *Human Molecular Genetics*, 13:763–770, 2004.

Varambally, S. et al., The polycomb group protein EZH2 is involved in progression of prostate cancer, *Nature*, 419:624–629, 2002.

Venter, J.C. et al., The sequence of the human genome, *Science*, 291:1304–1351, 2001.

Wang, J. et al., Differences in gene expression between b-cell chronic lymphocytic leukemia and normal b cells: a meta-analysis of three microarray studies, *Bioinformatics*, 20:3166–3178, 2004.

Wasserman, S. and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge University Press, 1994.

Welch, B.L., The significance of the difference between two means when the population variances are unequal, *Biometrika*, 29:350–362, 1938.

Williams, C.N. et al., Using a genome-wide scan and meta-analysis to identify a novel IBD locus and confirm previously identified IBD loci, *Inflammatory Bowel Diseases*, 8:375–381, 2002.

Winzeler, E.A. et al., Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays, *Genetics*, 163:79–89, 2003.

Winzeler, E.A. et al., Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis, *Science*, 285:901–906, 1999.

Wise, L.H., Inclusion of candidate region studies in meta-analysis using the genome screen meta-analysis method: application to asthma data, *Genetic Epidemiology*, 21 Suppl 1:S160–S165, 2001.

Wise, L.H., J.S. Lanchbury, and C.M. Lewis, Meta-analysis of genome searches, *Annals of Human Genetics*, 63:263–272, 1999.

Wright, G. et al., A gene expression-based method to diagnose clinically distinct subgroups of diffuse large B cell lymphoma, *Proceedings of the National Academy of Sciences USA*, 100:10585–10587, 2003.

Wright, G.W. and R.M. Simon, A random variance model for detection of differential gene expression in small microarray experiments, *Bioinformatics*, 19:2448–2455, 2003.

Wu, C. et al., A probe-to-transcripts mapping method for cross-platform comparisons of microarray data, Tech. rep., BEPress, 2005.

Wu, X. et al., A combined analysis of genomewide linkage scans for body mass index from the National Heart, Lung, and Blood Institute Family Blood Pressure Program, *American Journal of Human Genetics*, 70:1247–1256, 2002.

Yan, P.S. et al., Dissecting complex epigenetic alterations in breast cancer using CpG island microarrays, *Cancer Research*, 61:8375–8380, 2001.

Zhang, L., M.F. Miles, and K.D. Aldape, A model of molecular interactions on short oligonucleotide microarrays, *Nature Biotechnology*, 21:818–821, 2003.

Zintzaras, E. and J.P. Ioannidis, HEGESMA: genome search meta-analysis and heterogeneity testing, *Bioinformatics*, 21:3672–3673, 2005a.

Zintzaras, E. and J.P.A. Ioannidis, Heterogeneity testing in meta-analysis of genome searches, *Genetic Epidemiology*, 28:123–137, 2005b.