Meta-analysis and Combining Information in Genetics

Rudy Guerra
&
David Allison

Contents

CHAPTER 1

Combining Information Across Genome-wide Scans

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

1.1 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.

1.2 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.

1.2.1 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

$$X_t^2 = -2 \sum_{s=1}^{k} \ln(p_{st}) \tag{1.1}$$

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

$$P_t = \mathrm{P}(\chi^2_{2k} > X_t^2), \tag{1.2}$$

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 (1.2) 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

$$p^*_{st} = Cp_{st} + 2\lambda GZ(p_{st})\phi(\Phi^{-1}(p_{st}))V[\Phi^{-1}(p_{st})\sqrt{4\lambda\Delta}] \tag{1.3}$$

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 (1.3) 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)

$$p_{st} = 1 - \Phi[sign(LOD_{st})\sqrt{2\ln(10)|LOD|}], \tag{1.4}$$

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 (1.4) 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

$$V_t = \sum_{s=1}^{m} R(X_{st}) \tag{1.5}$$

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 (1.5) 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 (1.5) 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 (1.5), 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.

1.2.2 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

$$Z_{MA_t} = \frac{\sum_{s=1}^{k} w_{st} Z_{st}}{\sqrt{\sum_{s=1}^{k} w_{st}^2}} \tag{1.6}$$

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.

1.3 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.

1.3.1 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.

1.3.2 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.

1.3.3 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.

1.3.4 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.

1.4 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.

1.5 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]

1.6 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: $Z_{st} = \Phi^{-1}(p_{st})$

CHAPTER 2

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

2.1 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).

2.2 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.

2.3 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.

In the GSMA, the genome is divided into bins of approximately equal cM width. 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.*. 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.

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

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

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. For each study, the ranks within a study are randomly re-assigned to bins, and then the summed rank calculated for each bin. For d replicates, dn summed rank values are obtained, and the p-value for the observed summed rank is calculated from the number of simulated bins with summed rank greater than the observed summed rank $(= r)$. The p-value is then $p_{SR} = (r + 1)/(dn + 1)$, where n is the number of simulated bins (North et al., 2003). Calculating critical values from simulations is particularly appropriate where the assigned ranks depart from the integer values $1, 2, \ldots, n$ assumed in the distribution function above, through tied ranks or missing values (see Table 2.1).

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 each bin and independently of other bins, and 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 (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, and p = 0.0083 (= 1/120) for suggestive evidence of linkage.

For a genome-wide assessment of linkage, the ordered rank (OR) p-value (p_{OR}) may be used (Levinson et al., 2003). This uses simulations of the complete GSMA

Table 2.1: 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.

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. So, in a simulation of 5000 complete GSMAs, the bin with the highest 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).

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

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)). 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.

2.4 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 2.1, 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.

2.5 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 $(n + 1) - R/m$.

2.6 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 2.2). Although narrow bins may intuitively provide more information (see Figure 2.1), localisation through linkage information is broad. Adjacent bins may show evidence for linkage (see, for example, rheumatoid arthritis (Fisher et al., 2003) and 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 2.3.

2.7 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 2.2) 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.

2.8 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

Table 2.2: 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)

Disease	Publication	# studies	# families	# bins	Weights	# bins with SR Nom./Sugg./Gen.	$p_{SR} < 0.05$, $p_{OR} < 0.05$
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 2.3: 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

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 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.

2.9 Power to detect linkage using the GSMA

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 2.2).

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 $p < 0.01$. 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).

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 2.1. 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. 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. An average rank of 32 gives nominal evidence for linkage with 4 studies, but provides genome-wide evidence for linkage with 20 studies. With 60 bins, lower average ranks are required for linkage, so that the evidence must be stronger in linked bins where wider bins are used. 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 20 gives genome-wide evidence for linkage if this is obtained using 120 bins, but only nominal significance with 60 bins. 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.

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 genotype data from each study. 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.

2.10 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

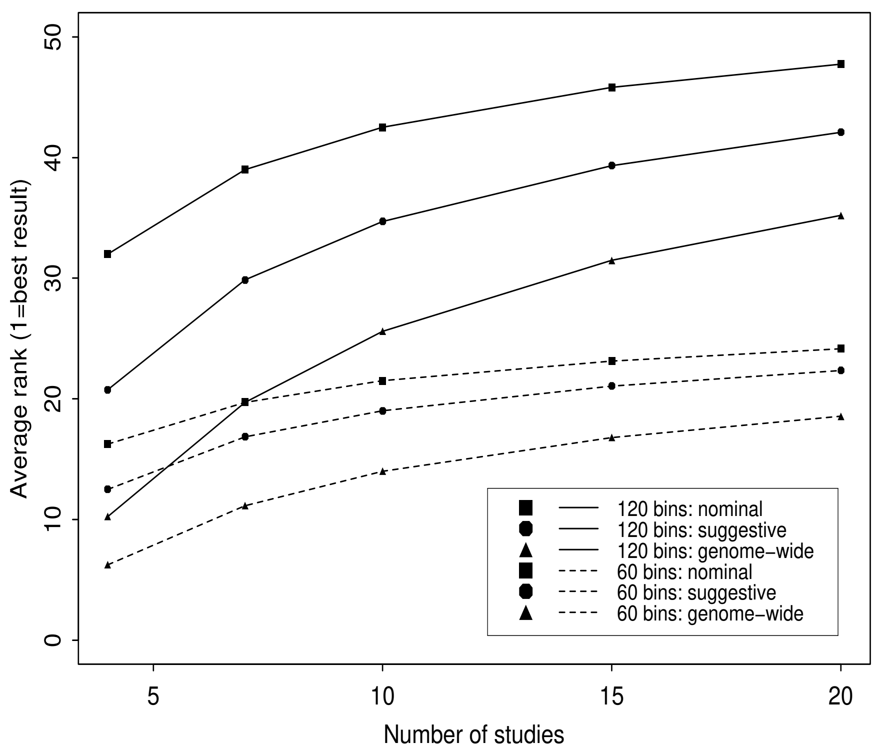

Figure 2.1: Critical values of the average rank required for genome-wide, suggestive, and nominal evidence for linkage, by number of bins.

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, ... are pooled in the first analysis, and 2+3, 4+5 ... 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 2.1 shows, reducing the total number of bins reduces the power to detect linkage.

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).

2.11 Limitations of the GSMA

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

2.11.1 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.

2.11.2 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.

2.11.3 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.

2.12 Disease studies using the GSMA

The GSMA has been applied in 14 studies of complex diseases, summarised in Table 2.2 (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 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. 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.

2.13 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.

2.14 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.

2.15 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 3

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

3.1 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.

3.2 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 publicly 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 publicly 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 publicly-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. (2004), 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; Sørlie 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 standardized 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.

3.3 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 partitions 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.

3.4 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 3.1 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.

3.5 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 critical assessment 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

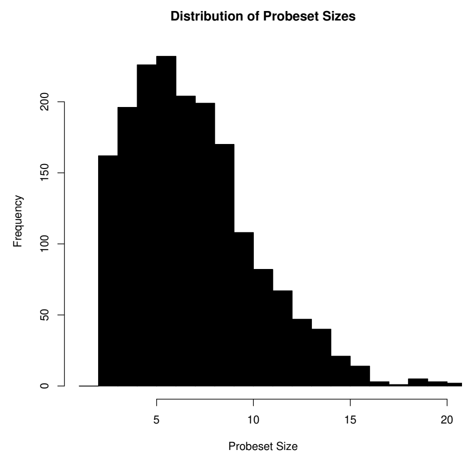

Figure 3.1: Histogram of number of probes in each "partial probeset."

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).

3.5.1 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.

3.5.2 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 3.2 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

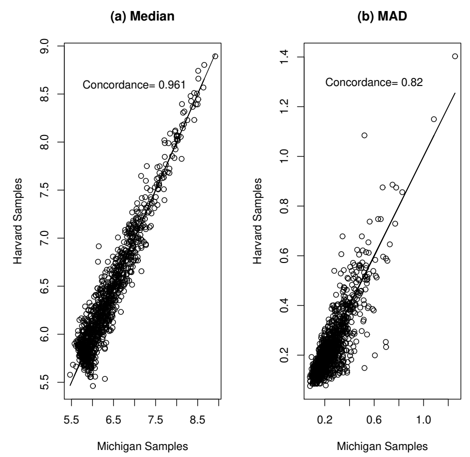

Figure 3.2: Median (a) and median absolute deviation (b) expression levels for each partial probeset based on the Harvard samples run on the U95Av2 chips vs. the Michigan samples run on the HuGeneFL chip. The high concordance in these measures suggests we obtain reasonably comparable expression levels 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 3.3. 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.

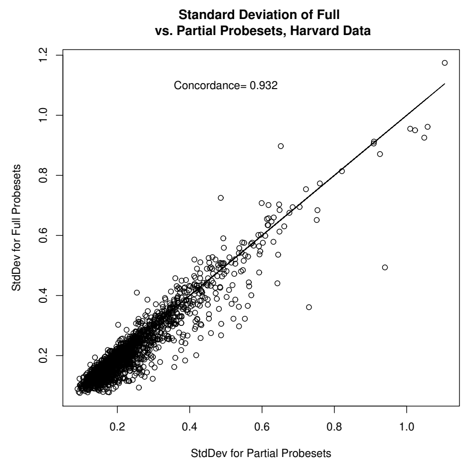

Figure 3.3: Standard deviation across Harvard samples for each gene based on full and partial probesets. A "full probeset" contains all probes on the U95Av2 chip mapping to a unique UniGene ID, while the corresponding "partial probeset."

3.5.3 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, 2000)) 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.

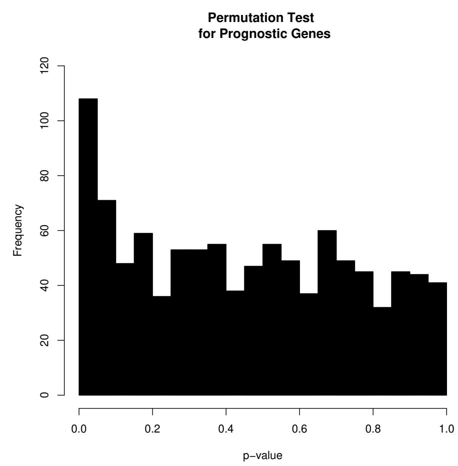

Figure 3.4: Histogram of p-values from permutation test on gene coefficient in Cox model containing clinical covariates and each one of the 1036 candidate genes. The corresponding histogram for the LRT is nearly identical.

Figure 3.4 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 3.1 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 3.1: 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.

3.6 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 3.5 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/).

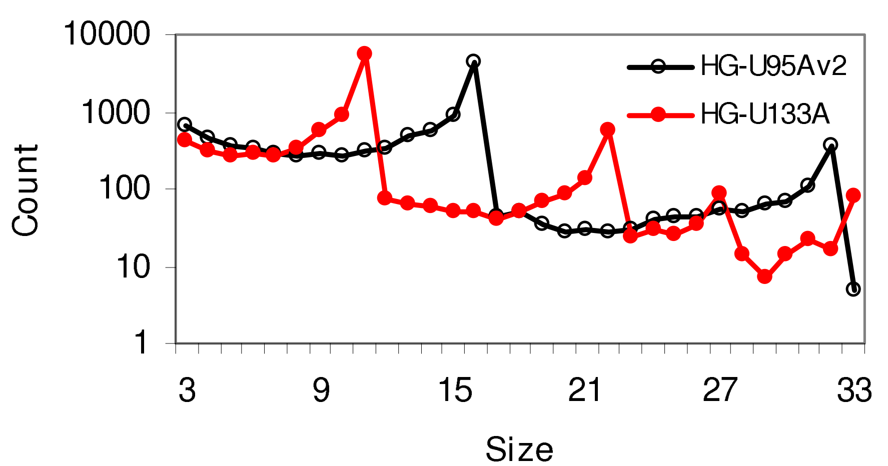

Figure 3.5: Histogram of number of probes per FLTBP.

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.

3.7 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.

3.7.1 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., 2003), 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).

3.7.2 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 3.6 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 3.6A clearly demonstrates that, when using the PDNN quantification method, the FLTBP mapping tends to yield better correlations than the Affymetrix mapping ($p < 0.00001$,

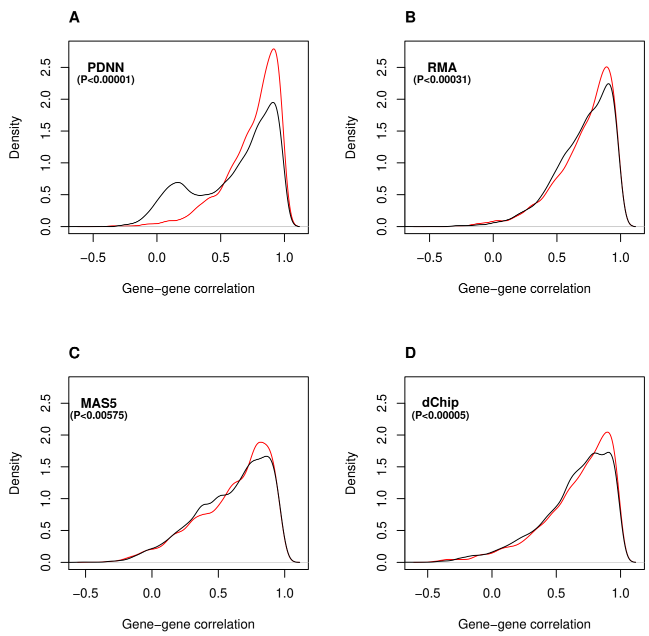

Figure 3.6: Distribution of gene-to-gene correlation between probesets on two U95Av2 and U133A arrays, combining information over all samples, using both Affymetrix-defined probesets and FLTBPs. The correlations were computed using four different quantification methods, (A) PDNN, (B) RMA, (C) MAS5.0, 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 3.6B-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 3.6A, the distributions in Figure 3.6B-D are shifted more toward 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 3.7. This increased correlation observed from the PDNN method may reflect the man-

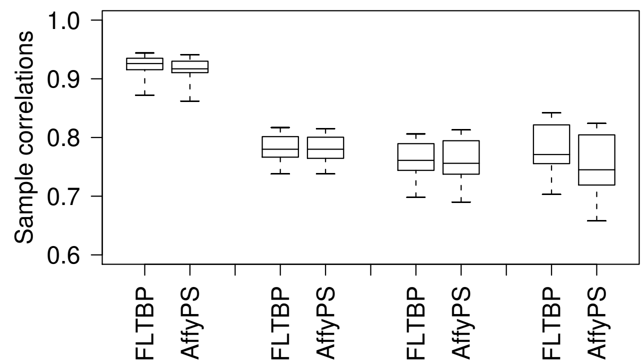

Figure 3.7: Distribution of sample-to-sample correlation between probesets on two U95Av2 and U133A arrays, combining information over all genes, using both Affymetrix-defined probesets and FLTBPs. The correlations were computed using four different quantification methods, PDNN, RMA, MAS5.0, and dChip, 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.

3.8 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.

References

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.

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.

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.

Beer, D.G. et al., Gene-expression profiles predict survival of patients with lung adenocarcinoma, *Nature Medicine*, 9:816–824, 2002.

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, Series B*, 57:289–300, 2000.

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.

Brazma, A. et al., Minimum information about a microarray experiment (MIAME) – toward standards for microarray data, *Nature Genetics*, 29:373, 2001.

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, *Arte-*

riosclerosis Thrombosis and Vascular Biology, 23:1863–1868, 2003.

Choi, J.K. et al., Combining multiple microarray studies and modeling interstudy variation, *Bioinformatics*, 19 Suppl 1:i84–i90, 2003.

Cooperative", T.T.M.S.G., A meta-analysis of genomic screens in multiple sclerosis, *Multiple Sclerosis*, 7:3–11, 2001.

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.

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.

Dobbin, K.K. et al., Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays, *Clinical Cancer Research*, 11:565–572, 2005.

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.

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, 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.

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.

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.

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.

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.

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.

Irizarry, R.A. et al., Summaries of Affymetrix GeneChip probe level data, *Nucleic Acids Research*, 31:e15, 2003.

Irizarry, R.A. et al., *affy: Methods for Affymetrix Oligonucleotide Arrays*, 2004.

Iyengar, S. and J. Greenhouse, Selection models and the file drawer problem, *Statistical 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.

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.

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.

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.

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.

Lander, E. and L. Kruglyak, Genetic dissection of complex traits – guidelines for interpreting and reporting linkage results, *Nature Genetics*, 11:241–247, 1995.

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, 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.

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.

Mah, N. et al., A comparison of oligonucleotide and cDNA-based microarray systems, *Physiological Genomics*, 16:361–370, 2004.

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.

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.

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.

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.

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 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.

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.

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.

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., 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, 2004.

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.

Sawcer, S. et al., Empirical genomewide significance levels established by whole genome simulations, *Genetic Epidemiology*, 14:223–229, 1990.

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.

Sørlie, 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.

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.

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.

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.

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.

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.

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.

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.