Meta-analysis in Gene Expression Studies

Levi Waldron and Markus Riester

Running title: Gene Expression Meta-Analysis

Summary
This chapter introduces methods to synthesize experimental results from independent high-throughput genomic experiments, with a focus on adaptation of traditional methods from systematic review of clinical trials and epidemiological studies. First, it reviews methods for identifying, acquiring, and preparing individual patient data for meta-analysis. It then reviews methodology for synthesizing results across studies and assessing heterogeneity, first through outlining of methods and then through a step-by-step case study in identifying genes associated with survival in high-grade serous ovarian cancer.

Keywords
Biomarkers, Microarray Analysis, Gene Expression Profiling, Ovarian Neoplasms, Meta-Analysis, Computational Molecular Biology

1. Introduction

This chapter introduces methods to synthesize experimental results from independent high-throughput genomic experiments, with a focus on adaptation of traditional methods from systematic review of clinical trials and epidemiological studies. We focus on differential gene expression because the public availability of data from gene expression microarrays far surpasses any other genomic assay; however, these methods are flexible and applicable to other genomic data types and study objectives.

The traditional systematic review and meta-analysis attempts to resolve inconsistency and uncertainty in a literature of, for example, clinical trials on effectiveness of a treatment or observational studies of association between a risk factor and health outcome. The analysis is of association between the outcome of interest and a single exposure or treatment. Great care must be taken in this situation to select and studies and exclude incomparable studies in order to avoid bias in the analysis, for example by adherence to the PRISMA guidelines [1]. Systematic review and meta-analysis has been well described in the primary literature, reviews [2, 3], and monographs [4, 5]. Although the methodology we summarize in this chapter is substantially similar, the setting presents different challenges and priorities.

Unlike traditional studies of a single exposure or treatment, in this setting thousands of variables are observed - one per gene or transcript. Association measures, most commonly differential expression of each transcript, are almost never consistently reported, and instead must be calculated for each study using Individual Patient Data (IPD). It is hard to envision that synthesized results could be affected by inadvertent bias of the meta-analyst, but different challenges are present. The bioinformatic challenges can be substantial: locating and standardizing raw gene expression data and clinical data, handling large datasets, and

repeating and interpreting thousands of meta-analyses. This chapter provides advice and recommends approaches for dealing with these challenges, and reviews relevant methods from traditional meta-analysis. Using straightforward and statistically well-established methods, meta-analysis makes it possible to overcome some of the limitations of high-dimensionality and batch effects that are inherent to high-throughput biology, and to develop extremely robust biomarkers.

2. Materials

2.1. Microarray dataset identification

If high coverage of available data is important, the most thorough approach to dataset identification is systematic literature review. Starting points for search terms of Pubmed for numerous cancer types are provided in the GeneSigDB database [6]. However the large majority of publicly available gene expression data are re-distributed through the Gene Expression Omnibus (GEO) [7] or ArrayExpress [8], and data from these resources is significantly easier to access and is more stably available than data from the websites of authors or their institutions. Other alternatives providing greater curation but much lower coverage are InSilicoDB [9], Oncomine [10], and Bioconductor (BiocViews: ExperimentData, RNAExpressionData).

2.1.1. The Gene Expression Omnibus

Experiments in GEO are represented by Series (GSE codes), which are occasionally curated as Datasets (GDS codes). Series may be composed of a single platform (GPL) or multiple platform. Platforms (GPL) annotate platform-specific identifiers and usually provide maps to standard genes identifiers. However it should be noted that the GPL annotations are generally author-provided and unstandardized, so different platforms provide different annotations or the same annotations based on different genome builds, and can even contain spreadsheet-introduced gene symbol errors [11]. When possible, it is safer to use annotations from Bioconductor [12] *.db* packages (BiocViews term *AnnotationData*) or from BioMart [13]. When manufacturer-specific annotations are not available, Bioconductor or Biomart can still be used to map stable identifiers such as Entrez Gene or Refseq to other annotations, rather than using unstable and potentially outdated identifiers, such as gene symbols, directly from the GPL annotations.

GEO is well-supported in Bioconductor by the GEOmetadb package [14] for searching meta-data and the GEOquery package [15] for downloading expression and platform data.

2.2. Dataset preparation

Whereas meta-analysis for clinical trials and observational studies in epidemiology is often possible using published summary statistics and confidence intervals, thousands of rows of summary statistics are required in meta-analysis. Normally these must be calculated from Individual Patient Data (IPD) after appropriate standardizing steps. This section first describes steps to prepare datasets for the calculation of summary statistics appropriate for meta-analysis in ways that reduce the impact of unwanted technical variability between studies. However, we

note that some amount of heterogeneity between datasets is inevitable, arising both from experimental settings and from differences in patient recruitment and treatment. Heterogeneity should not be viewed as an enemy of the meta-analyst. The existence of heterogeneity provides rationale for using meta-analysis to identify robust genomic signals present independently of the heterogeneity, and to investigate the impact of heterogeneity on identified genome / patient associations.

2.2.1. Curation

Individual-patient metadata must be standardized across studies, including variable names and the values they take. This process is error-prone and it can be difficult to catch mistakes later in high-throughput analysis, therefore a template-based syntax checking recommendable. For example the curatedOvarianData package [16] published an R script for template-based curation and *regular expression* checking in R, the InSilicoDB web service [9] provides a graphical interface to curation.

2.2.2. Preprocessing

The application of different microarray pre-processing algorithms may introduce technical heterogeneity. Although in our experience (for example [17, 18]) even different microarray platforms do not necessarily contribute significant heterogeneity, when the analyst is in a position to pre-process raw data, certain pre-processing approaches will reduce the potential for heterogeneity. Common normalization and probe set summarization methods such as Robust Multi-array Average [19] use multiple arrays to estimate probe effects, and heterogeneity may be introduced by processing datasets separately and therefore using different estimates of probe effects for each dataset. One approach to avoiding these differences is to preprocess all datasets together, using a low-memory function such as justRMA from the Affymetrix Bioconductor [12] package. For supported platforms, the *frozen RMA* method [20] uses a frozen reference database of thousands of publicly available raw data files, and eliminates differences in estimated probe set effects across datasets. We emphasize, however, that it is also reasonable to use data already pre-processed by different algorithms, and assess the amount of heterogeneity *post-hoc* using the I^2 or Cochrane's Q statistic.

When different technological platforms used prevent application of comparable pre-processing methods across all studies, a next-best approach is to scale the observations for each gene (or row) to z-scores, by subtracting the mean and dividing by the standard deviation. This should be done *after* ensuring that any dataset-wide variance-stabilizing transformation, such as the log-transform, has been applied consistently in all or none of the datasets. Scaling to each variable in each dataset to unit variance ensures that fold-change or other effect-size estimates are comparable across studies, which may be adequate when synthesizing results obtained from comparable but different measurement technologies.

2.2.3. Batch effects

Anyone familiar with the problems that batch effects can cause for single studies [21, 22]will be concerned about the potential for batch effects to impact a meta-analysis. Using traditional methods of assessing heterogeneity, one can identify the extent to which heterogeneity between datasets is impacting a meta-analysis, and identify which datasets are most

responsible for the heterogeneity. One can establish whether the amount of heterogeneity warrants the potentially large amount of effort required to identify and correct for batch effects in individual studies, and if so whether batch correction actually helps. may be surprised to find that batch effects in some cases have limited practical effect on a meta-analysis.

2.2.4. Gene collapsing

A basic requirement of in gene expression meta-analysis that each study contain overlapping sets of measurements. If all studies used the same technological platform, it is possible to perform meta-analysis either using manufacturer-specific probe set identifiers, or gene-level summaries. To synthesize analysis across different platforms, however, it is necessary to map and summarize manufacturer-specific probe set identifiers to standard identifiers such as Entrez Gene or gene symbols. Miller *et al* [23] discuss and compare alternatives for merging probe set level data to gene level. We highlight one additional consideration for meta-analysis, that when using an approach that selects a single representative probe set per gene, it is preferable to select the same probe set for each study in the meta-analysis. For example, Ganzfried and Riester *et al.* [16] selected for each gene the representative probe set with maximum mean across all studies of a common platform. Approaches which do not use the dataset at hand for probe set selection, such as *Jetset* [24] or *BrainArray* [25], also avoid introducing heterogeneity that can arise from representing a gene by different probe sets in each dataset.

2.2.5. Pathway or gene set collapsing

While pathway or gene set approaches are routinely used to test for enrichments in gene rankings, for example via cutoff-free methods such as GSEA [26] or via methods for analysing gene lists such as implemented the DAVID webservice [27], the value of collapsing features to pathways is appreciated only recently. The idea of these approaches is to calculate for each sample and pathway (or gene set) a single pathway activation score, utilizing the expression values of all measured genes in this pathway. The gene-by-samples expression matrix is transformed into a pathway-by-samples expression matrix. Such a pathway activation score calculation is a potential noise reduction step and can be used to more robustly compare expression data across very different assays, for example even when the data was obtained from different species [28].

The first step is selecting appropriate gene sets for the problem at hand. A commonly used resource is MSigDB [26], which provides curated sets of genes in different categories, for example canonical pathways such as KEGG [29] or REACTOME [30, 31], downstream targets of gene regulators such as transcription factors or miRNAs, or genes associated with gene ontology (GO) terms [32]. The expected expression direction (“up” vs. “down”) of genes within a set of genes representing an active pathway provides important information and high activity of both up- and down-regulated genes may cancel each other out. It is thus recommended to split pathways into up- and down-regulated gene sets when this information is available [33]. Final pathway scores can be calculated by subtracting the activation scores of down-regulated genes from the ones of the up-regulated activation scores.

Various methods for collapsing genes to gene sets have been proposed [34], most notably ssGSEA [35] or GSVA [33]. Since these methods need to distinguish, for all genes in the gene sets, activated or inhibited gene expression from normal expression levels, these methods work better the larger the dataset is [33]. This limitation could be in theory avoided when the future methods support platform-specific databases of gene expression ranges, as for example utilized in the fRMA approach [20] discussed in the "Preprocessing" section. In a meta-analysis setting, gene set collapsing methods have been used for example to robustly classify samples by subtype [36] or comparing *in vivo*, *in vitro* and murine data [37].

2.2.6. Duplicate checking

The methods discussed here assume independence of studies and samples. This assumption can be violated by re-use of clinical tissue specimens by a research group in subsequent studies, or sharing of specimens between different research groups and in consortial studies. We developed the *doppelgangR* R package (https://github.com/lwaldron/doppelgangr) to facilitate identification of duplicates from gene expression profiles. Duplicates may also be identified by patient identifiers and inspection of published papers, and subsequent papers by the same research group deserve extra attention to duplicate-checking.

2.2.7. Gene pre-filtering

Once the data for all studies has been preprocessed so that features (probe sets, genes, pathway activation scores etc.) are comparable across studies, it is further advisable to investigate whether the features indeed measure the same biological signal, especially when data was obtained from different platforms. The integrative correlation technique proposed by [38] can be used to select "reproducible" genes. The basic idea behind this approach is that genes should be co-expressed with the same set of other genes across platforms and studies. Therefore, the correlation in expression of a given gene *G* is calculated between *G* and every other gene in a study, i.e., to identify the "neighborhood" of *G*. If this "neighborhood" is very different across datasets, the average correlation of correlation profiles across all pair of studies would be low; only if the average correlation of correlations exceeds a certain threshold, gene *G* is thus called reproducible and is included in the meta-analysis.

3. Methods

Although several methods have been proposed for meta-analysis of gene expression microarrays, traditional methods developed for synthesis of clinical trials and epidemiological studies remain highly relevant and are the most well-understood and implemented. Several implementations are available in the R environment, but our favorite for maturity and documentation is *Metafor* [39]. We refer to discussion therein for references to alternative software packages. The classic methods of fixed and random-effects synthesis are simply be applied for each gene or feature, with the main challenge being repetition of the methods for thousands of rows of a gene expression matrix, and summary and interpretation of thousands of results.

3.1. Fixed effects meta-analysis

Here we summarize the methods described by DerSimonian and Laird [2] for synthesis of effect size estimates across studies. Although their paper is concerned with risk ratio or risk difference in case-control studies, the methods are applicable to the synthesis of other statistics or *effect sizes* as long as they are accompanied by a standard error for each study. These methods are parametric in that a distribution of effect sizes across studies is assumed: constant in the fixed effects model, or normal in the random effects model. Let $\boldsymbol{\theta}_i$ be a per-gene estimate of interest that is assumed to normally distributed, such as log fold-change for differential expression from a Limma analysis [40] or log hazard ratio in a univariate Cox proportional hazards model, where *i* indexes each independent study, *i=1, …, K*. It can represent the coefficient of a gene in a generalized linear model with non-normally distributed residuals and linear or non-linear link function, or a simple differential expression analysis. The coefficient can be corrected for clinical covariates in a multivariate regression model. What matters is that the method produces the estimates of interest, and accompanying standard errors, for each genomic feature in each study.

The fixed-effects model is developed under the assumption of one true effect size, with differences between studies attributed to individual-level sampling variation, *i.e.*:

$$\theta_i = \theta_F \quad (1)$$

where θ_F is the common true effect size. Under a fixed-effects model, the synthesized estimate, $\hat{\theta}_F$, is a weighted average of estimates from each study:

$$\hat{\theta}_F = \frac{\sum_{i=1}^{K} w_i \hat{\theta}_i}{\sum_{i=1}^{K} w_i} \quad (2)$$

These weights are commonly taken to be the inverse squared standard error of the effect estimates from each study:

$$w_i = \frac{1}{\sigma i^2} \quad (3)$$

The standard error of the fixed-effects estimate is the inverse mean of the study-specific weights:

$$S.E.(\hat{\theta}_F) = \sqrt{\frac{1}{\sum_{i=1}^{K} w_i}} \quad (4)$$

Equations (2-4) are sufficient to calculate a fixed-effects meta-analytical estimate of log fold-change, log hazard ratio, etc. We note that some software and R functions produce confidence intervals rather than standard errors; these are converted to standard error in the standard way, for example to convert a 95% interval to Standard Error:

$$S.E. = \frac{C.I._{upper}^{95\%} - C.I._{lower}^{95\%}}{2 \times 1.96} \quad (5)$$

a

Although this calculation can be performed by meta-analysis packages such as *metafor*, we present it to highlight the simplicity of the model used and to distinguish it from the a random-effects model.

3.2. Random effects meta-analysis

Under the random-effects model, the synthesized estimate is given by equation 2 but with altered weights:

$$w_i^* = 1/(\hat{\tau}^2 + \hat{\sigma}_i^2) \quad (6)$$

where $\boldsymbol{\tau}^2$ is a heterogeneity parameter and standard output of meta-analysis software, $\widehat{\tau}^2$ is estimated from the data, and $\widehat{\sigma_i^2}$ are the variances of each study's effect size estimate $\widehat{\theta}_i$. As $\boldsymbol{\tau}^2 \rightarrow 0$ (no heterogeneity) the random-effects solution converges to the fixed-effects solutions, and that as $\boldsymbol{\tau}^2 \rightarrow \infty$ (very large heterogeneity) the random-effects solution is a simple average of the per-study effects, regardless of sample sizes or standard errors of each study. The null hypothesis of no heterogeneity ($\boldsymbol{\tau}^2 = 0$) can be tested by the *Cochrane's Q statistic*, which is just the summed product of study weights by squared residuals of per-study effect sizes from the fixed-effects estimate:

$$Q = \sum_{i=1}^{K} w_i(\hat{\theta}_i - \hat{\theta}_F)^2 \quad (7)$$

The Q statistic is $\boldsymbol{\chi}^2_{k-1}$ distributed under the null hypothesis of no true between-study heterogeneity in effect sizes (e.g. as in Equation (1)), and is a standard test of heterogeneity. $\boldsymbol{\tau}^2$ can be understood as an estimate of total amount of heterogeneity present, in the same units as variance of the effect size. Another commonly reported measure of heterogeneity is I^2:

$$I^2 = \frac{\hat{\tau}^2}{\hat{\tau}^s + \hat{\sigma}^2_{typical}} \quad , (8)$$

where $\widehat{\sigma}^2_{typical}$ is the "typical" single-study variance in effect size estimate. I^2 can be understood as fraction of total variability in the estimate of a single study's effect size that is due to heterogeneity.

In performing meta-analysis for thousands of gene expression features, these estimates are produced for *each* feature. A global picture of heterogeneity can be developed from histograms of I^2, $\boldsymbol{\tau}^2$, and Cochrane's Q-test p-values. Some features will exhibit more heterogeneity than others, and one can consider whether heterogeneity is likely technical or biological. For example, features can be ranked by evidence of heterogeneity, and gene set analysis can be performed on the basis of this ranking.

3.3. Fixed vs. random effects meta-analysis

It is important to understand heterogeneity in the data. In the context of gene expression meta-analysis, a fixed-effects meta-analysis will identify the genes with strongest effect in the studies used as training data, while a random-effects model attempts to identify the genes with

strongest average effect in a hypothetical population of studies. While the latter is theoretically preferable, it might not be optimal if, for example a source of heterogeneity is not expected to occur again in future data. For example, different microarray platforms might measure particular transcripts with different accuracy. If heterogeneity is due to technical issues in a single study or platform, then a random-effects analysis might remove genes with strong effect from a ranked list, because the random-effects model may unnecessarily down-weigh a useful predictor as a result of a problematic study or platform. In this context it is difficult to a priori decide whether fixed- or random-effects will work better. A sound approach is to try both, compare them, try to identify and understand the sources of heterogeneity, and choose the simpler approach if results are very similar.

3.4. Rank-based meta-analysis

The rank products method [41] was developed in the early days of microarray data analysis for identifying differentially expressed genes. Datasets in this era were typically small and noisy, which made a method free of distributional assumptions particularly useful. The method was soon extended for meta-analyses [42], and became a popular choice for microarray meta-analyses, mainly because of its simplicity, shown robustness [43], and for its more straightforward support for expression direction (“up” vs. “down”), compared to other simple and then commonly used methods such as synthesizing p-values via Fisher’s or Stouffer’s method. Since rank products weighs datasets by sample size, the results are in general expected to be more similar to a fixed-effects than to a random-effects meta-analysis [17]. A major practical disadvantage of the method, the computational cost of the permutation tests required to estimate the rank product statistic, was recently mitigated by the implementation of a fast approximation [44]. Rank products can give very different results compared to marginal tests for example when genes are highly correlated, clustered in so-called gene modules, since it breaks ties randomly. This will add random noise to the ranks of large gene modules, thus lowering their significance. This can be an advantage when final lists of differentially expressed genes comprise only highly correlated genes after adjusting for multiple testing.

3.5. Other approaches to synthesis

Several other methods have been frequently used in the literature, and we refer to the excellent review by Tseng et al. [45] for a comprehensive overview of those. These methods include “vote counting” approaches, probabilistic combining of p-values from different studies, and merging studies during data pre-processing. Vote counting approaches, where genes are considered significant or validated when they reach a p-value cutoff in a minimum number of datasets, are statistically inefficient but popular due to their simplicity. Combining p-values requires only minimal information about studies and is frequently used for example when estimating standard errors of effect sizes is not possible. For an overview and recent improvements in this class of methods, see [46]. Direct merging of datasets will create batch effects, and any imbalance of the outcome of interest between datasets will result in confounding with these batch effects. This method should only be used when transcriptome differences between datasets is be expected to be larger than batch effects, and the meta-analysis approach described above is

infeasible due to extreme imbalance between datasets, and confounding is unavoidable. When broad transcriptome modeling occurs between cases and controls, such as between oral squamous cell carcinoma and normal tissue [47], such merging of datasets can be acceptable even when datasets are unbalanced in case/control prevalence.

3.6. Case Study

In this section we work through a case study in identifying differentially expressed genes and assessing the extent and potential sources of heterogeneity. We focus on identifying genes associated with overall survival in ovarian cancer, using the curatedOvarianData Bioconductor package [16]. Full code and output for performing the analysis in this section are available from http://lwaldron.github.io/GeneExpressionMetaAnalysis/.

Although our original meta-analysis using the curatedOvarianData database was limited to late-stage, high-grade, serous ovarian cancer, here we include all patients for which overall survival is known, to demonstrate investigation of sources of heterogeneity. This analysis follows the basic steps:

1. Scale all genes to unit variance. In analyses where large numbers of samples are removed from a study, it is preferable to do scaling with the full number of samples. Scaling is critical when studies use different microarray platforms, so that coefficients have the same scale when synthesizing across studies.
2. Apply inclusion and exclusion criteria: microarray studies of primary tumors only, studies with at least 40 patients and 15 deaths, including only patients where censored overall survival is known.
3. Remove duplicate samples and studies that are subsets of another study. Steps 2 and 3 reduce the database from 30 studies and 4,411 samples to 15 studies and 2,271 samples. With a curated database, these steps are performed automatically and can include Regular Expression filters on patient meta-data.
4. Restrict analysis to genes available on every platform. This is not necessary, but was a convenience for this analysis.
5. Perform meta-analysis for each gene, using both a random-effects and a fixed-effects model. In this example we fit a univariate Cox Proportional Hazards model and use the coefficient of the continuous gene expression variable for synthesis.

At this stage we can assess the presence of heterogeneity by plotting a histogram of p-values from Cochrane's Q-test (Figure 1) for each gene. In the absence of any heterogeneity, these p-values would be uniformly distributed between 0 and 1. However we observe that many genes exhibit some heterogeneity across studies in association with with overall survival. We then identify NUAK1 as the gene with strongest evidence of prognostic association with overall survival (Figure 2). This gene exhibits no evidence of heterogeneity, with a non-significant Cochrane's Q test and identical synthesized log hazard ratios by fixed or random-effects meta-analysis. We identify *KALRN* as the gene with greatest evidence for heterogeneity in log hazard ratio (Figure 3), with the proportion of patients whose tumors were suboptimally debulked being one likely source of this heterogeneity (Figure 4). Overall, although some global evidence of heterogeneity exists, its impact on the synthesized estimate even of the gene

exhibiting greatest heterogeneity (*KALRN*) is not large, with the random-effects estimate differing from the fixed-effects mainly in having a somewhat larger confidence interval. This is consistent with recent independent analysis using newly proposed methods for identifying homogeneous and heterogeneous variables in pooled cohort studies that found none of 15 genes with known prognostic significance exhibited significant heterogeneity in prognostic association in the curatedOvarianData database [48].

Figure 1: Assessing heterogeneity of all genes by Cochrane's Q-test. The test was performed for all genes that present in every study, with p-values obtained from output of the rma.uni() function from the metafor library. In the absence of any heterogeneity a uniform distribution of p-values between 0 and 1 would be observed; this histogram indicates evidence of heterogeneity since small p-values are more frequent than larger p-values. This does not however imply that the magnitude of the heterogeneity is large, or that differentially-expressed genes identified are invalid.

Figure 2: Forest plot for the gene with the strongest evidence of association with overall survival in ovarian cancer. This gene, *NUAK1*, is the top-ranked gene according to synthesized p-values by both fixed and random-effects meta-analysis; in fact, it shows no evidence of heterogeneity and the synthesized estimates. Fifteen rows with study names on the left show the log Hazard Ratio of a Cox Proportional Hazards model for each study, with the point estimate and 95% confidence interval, with box sizes proportional to the inverse square of the standard error of each study (or roughly to the number of deaths). The right-hand column provides the numeric point estimate and confidence interval. Two diamonds at the bottom show the point estimate and 95% confidence interval of the synthesized log Hazard Ratio by fixed and random-effects estimate, which are identical. In other words, the study-to-study variation seen here is consistent with homogeneous studies and sampling variation only at the level of individual patients.

Figure 3: Forest plot for *KALRN*, which demonstrates the strongest evidence of heterogeneity between studies. The p-value from Cochrane's Q-test is marginally significant after Bonferroni correction (FWER = 0.06). Heterogeneity is apparent as studies with both significantly positive and negative Cox coefficients are observed. Note however that the synthesized point estimates are similar by fixed and random effects, but the standard error and confidence interval are smaller in the fixed-effects model.

Figure 4: Association between log Hazard Ratio and percentage of suboptimally debulked patients. Six covariates with prognostic relevance were considered as potential sources of heterogeneity between studies: suboptimal debulking of tumors, tumor histology, grade (high/low), stage (early/late), and age (greater or less than 70 years). For each of these covariates, a linear regression was performed between per-study covariate prevalence and per-study Cox coefficient (log Hazard Ratio), with points weighted by study size. Only the prevalence of suboptimal debulking was significantly associated with Cox coefficient (P=0.006).

The area of data points is proportional to study sample size and weighting in the linear regression. Higher proportions of suboptimally debulked patients are associated with greater association between *KARLN* expression and survival; in other words, KARLN is more strongly associated with bad prognosis in studies with more suboptimally debulked patients.

3.7. Extensions to predictive modeling

The most likely objectives of meta-analysis are to identify candidate differentially expressed genes, or to develop and validate a predictive model. The former objective has been reviewed also by [49]; the latter is a more recent application of genomic meta-analysis that has not to our knowledge been reviewed. Bernau *et al.* [50] recently proposed *leave-one-dataset-in cross-study validation* for validating of prediction models and comparing prediction algorithms using a collection of independent studies. In this approach, each dataset is used in turn for model training, and all other datasets for validation. The resulting matrix of independent validation statistics is analyzed for evidence of outlying studies, and for estimation of cross-study validation accuracy. Although this method is very useful for comparing algorithms and identifying outlying studies, a more accurate model can be developed using all available studies for training. Riester *et al.* [17] proposed a related leave-one-dataset-*out* cross-validation, whereby each study is used in turn for validation and the remaining n-1 studies are used for training. Log fold-change or Cox regression coefficients are synthesized across the training studies by meta-analysis as described in this chapter, and synthesized coefficients are directly used as the coefficients of a linear prediction score.

4. Notes

Meta-analysis of genomic datasets, even using basic statistical methods adapted from unrelated fields, is a powerful tool for overcoming dimensionality and batch effects to develop robust biomarkers and prediction rules. Specialized statistical methodologies are still needed, for example, for aggregating evidence of heterogeneity across all gene expression measurements and appropriately re-weighting studies by their concordance with other studies, and for missing data imputation that leverages independent datasets. The most significant barriers to successful use of genomic meta-analysis are public availability of data and annotations, and the tedious work of standardizing datasets from disparate sources to enable straightforward analysis of individual patient data. As RNA sequencing catches up with and potentially overtakes microarray technology as the dominant method of transcriptome profiling, the necessity of patient privacy presents new challenges for the open sharing of data that makes meta-analysis possible. We hope that the methods outlined in this chapter, and the successes of recent gene expression meta-analyses, will help provide the necessary motivation for researchers, journal editors, reviewers, and funding agencies to continue the tradition of open data sharing that was developed for the microarray.

5. References

1. Moher D, Liberati A, Tetzlaff J, et al. (2010) Preferred reporting items for systematic reviews

and meta-analyses: the PRISMA statement. Int J Surg 8:336–341.

2. DerSimonian R, Laird N (1986) Meta-analysis in clinical trials. Control Clin Trials 7:177–188.

3. Moher D, Olkin I (1995) Meta-analysis of Randomized Controlled Trials: A Concern for Standards. JAMA 274:1962–1964.

4. Lipsey MW, Wilson DB (2001) Practical meta-analysis. Sage publications Thousand Oaks, CA

5. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2011) Introduction to meta-analysis. John Wiley & Sons

6. Culhane AC, Schröder MS, Sultana R, et al. (2011) GeneSigDB: a manually curated database and resource for analysis of gene expression signatures. Nucleic Acids Res 40:D1060–6.

7. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30:207–210.

8. Kolesnikov N, Hastings E, Keays M, et al. (2015) ArrayExpress update-simplifying data submissions. Nucleic Acids Res 43:D1113–6.

9. Taminau J, Steenhoff D, Coletta A, et al. (2011) inSilicoDb: an R/Bioconductor package for accessing human Affymetrix expert-curated datasets from GEO. Bioinformatics 27:3204–3205.

10. Rhodes DR, Kalyana-Sundaram S, Mahavisno V, et al. (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. Neoplasia 9:166–180.

11. Zeeberg BR, Riss J, Kane DW, et al. (2004) Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. BMC Bioinformatics 5:80.

12. Gentleman RC, Carey VJ, Bates DM, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5:R80.

13. Haider S, Ballester B, Smedley D, et al. (2009) BioMart Central Portal--unified access to biological data. Nucleic Acids Res 37:W23–7.

14. Zhu Y, Davis S, Stephens R, et al. (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. Bioinformatics 24:2798–2800.

15. Davis S, Meltzer PS (2007) GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics 23:1846–1847.

16. Ganzfried BF, Riester M, Haibe-Kains B, et al. (2013) curatedOvarianData: clinically annotated data for the ovarian cancer transcriptome. Database 2013:bat013.

17. Riester M, Wei W, Waldron L, et al. (2014) Risk prediction for late-stage ovarian cancer by

meta-analysis of 1525 patient samples. J Natl Cancer Inst. doi: 10.1093/jnci/dju048

18. Waldron L, Haibe-Kains B, Culhane AC, et al. (2014) Comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. J Natl Cancer Inst. doi: 10.1093/jnci/dju049

19. Irizarry RA, Hobbs B, Collin F, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 4:249–264.

20. McCall MN, Bolstad BM, Irizarry RA (2010) Frozen robust multiarray analysis (fRMA). Biostatistics 11:242–253.

21. Leek JT, Scharpf RB, Bravo HC, et al. (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11:733–739.

22. Johnson WE, Li C, Rabinovic A (2006) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8:118–127.

23. Miller JA, Cai C, Langfelder P, et al. (2011) Strategies for aggregating gene expression data: the collapseRows R function. BMC Bioinformatics 12:322.

24. Li Q, Birkbak NJ, Gyorffy B, et al. (2011) Jetset: selecting the optimal microarray probe set to represent a gene. BMC Bioinformatics 12:474.

25. Dai M, Wang P, Boyd AD, et al. (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. Nucleic Acids Res 33:e175.

26. Subramanian A, Tamayo P, Mootha VK, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A 102:15545–15550.

27. Huang DW, Sherman BT, Tan Q, et al. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Biol 8:R183.

28. Altschuler GM, Hofmann O, Kalatskaya I, et al. (2013) Pathprinting: An integrative approach to understand the functional basis of disease. Genome Med 5:68.

29. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28:27–30.

30. Croft D, Mundo AF, Haw R, et al. (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42:D472–7.

31. Milacic M, Haw R, Rothfels K, et al. (2012) Annotating cancer variants and anti-cancer therapeutics in reactome. Cancers 4:1180–1211.

32. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

33. Hänzelmann S, Castelo R, Guinney J (2013) GSVA: gene set variation analysis for

microarray and RNA-seq data. BMC Bioinformatics 14:7.

34. Tarca AL, Bhatti G, Romero R (2013) A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. PLoS One 8:e79217.

35. Barbie DA, Tamayo P, Boehm JS, et al. (2009) Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature 462:108–112.

36. Verhaak RGW, Tamayo P, Yang J-Y, et al. (2013) Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J Clin Invest 123:517–525.

37. Ozawa T, Riester M, Cheng Y-K, et al. (2014) Most Human Non-GCIMP Glioblastoma Subtypes Evolve from a Common Proneural-like Precursor Glioma. Cancer Cell 26:288–300.

38. Parmigiani G, Garrett-Mayer ES, Anbazhagan R, Gabrielson E (2004) A cross-study comparison of gene expression studies for the molecular classification of lung cancer. Clin Cancer Res 10:2922–2927.

39. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J. Stat. Softw.

40. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3.

41. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573:83–92.

42. Hong F, Breitling R, McEntee CW, et al. (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22:2825–2827.

43. Hong F, Breitling R (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. Bioinformatics 24:374–382.

44. Heskes T, Eisinga R, Breitling R (2014) A fast algorithm for determining bounds and accurate approximate p -values of the rank product statistic for replicate experiments. BMC Bioinformatics 15:367.

45. Tseng GC, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40:3785–3799.

46. Li Y, Ghosh D (2014) Meta-analysis based on weighted ordered P-values for genomic data with heterogeneity. BMC Bioinformatics 15:226.

47. Reis PP, Waldron L, Perez-Ordonez B, et al. (2011) A gene signature in histologically normal surgical margins is predictive of oral carcinoma recurrence. BMC Cancer 11:437.

48. Cheng X, Lu W, Liu M (2015) Identification of homogeneous and heterogeneous variables in pooled cohort studies. Biometrics. doi: 10.1111/biom.12285

49. Ramasamy A, Mondry A, Holmes CC, Altman DG (2008) Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS Med 5:e184.

50. Bernau C, Riester M, Boulesteix A-L, et al. (2014) Cross-study validation for the assessment of prediction algorithms. Bioinformatics 30:i105–12.