Chapter 6

Genomic meta-analysis and information integration

Research synnthesis (a.k.a. meta-analysis) has a long history in the applications of medical research and social science (e.g. education and psychology). Multiple studies investigating a similar research hypothesis may have been published in the literature and data are available in public domain. Each study has small sample size and generates weak statistical conclusion. Integrating information in the studies can potentially increase statistical power and generate a more consensus conclusion. In genetic and genomic data analysis, meta-analysis methods have been extended and applied for combinined linkage analyses, genome-wide association studies (GWAS) and microarray studies (Guerra and Goldstein, 2010; Tseng et al., 2012; Begum et al., 2012). As the technology and prevalence of high-throughput genomic experiments continue to grow, information integration and meta-analysis will undoubtedly gain popularity in genomic research. In this chapter, we introduce some basics of classical meta-analysis and then overview some genomic meta-analysis examples.

6.1 Methods for univariate meta-analysis

Traditional meta-analysis methods have two major categories for information integration: combine effect sizes and combine p-values. We will briefly introduce methods in both categories and discuss their pros and cons.

6.1.1 Combine effect sizes

Statistics to describe effect size

Depending on the data structure and biological hypothesis behind, different statistics can be used to describe the effect size within a study. To perform research synthesis, we need to select an adequate effect size to combine information across studies.

Raw mean difference (continuous observation) Suppose two groups of observations are available and the question is to compare their difference. A simple statistic is to use raw mean difference: $D = \bar{Y_1} - \bar{Y_2}$. One can assume equal standard deviation of the two groups and caculate a pooled variance $Var(D) \approx \frac{n_1+n_2}{n_1 \cdot n_2} \cdot S^2$, where $S^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$ and S_1 and S_2 are the sample variances of group 1 and 2. If we allow the standard deviations of two groups to be unequal, the variance can be estimated as $Var(D) \approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$.

Standardized mean difference (continuous observation) The raw mean difference D does not consider variation of each group. A standardized mean difference is often preferred. An intuitive and popular standardized mean difference is Cohen's d: $d = \frac{\bar{Y_1}-\bar{Y_2}}{S}$, where pooled estimate of S often used. It can be shown that $Var(d) \approx \frac{n_1+n_2}{n_1 n_2} + \frac{d^2}{2(n_1+n_2)}$. Note that the first term describes uncertainty of $\bar{Y_1}-\bar{Y_2}$ and the second term describes uncertainty of S.

Cohen's d is known to be slightly biased to overestimate the true parameter when sample size is small. The Hedge's d provides a simple correction: $d_{Hedge} = (1 - \frac{3}{4 \cdot df - 1}) \cdot d$, where $df = n_1 + n_2 = 2$. The variance can be calculated as $Var(d_{Hedge}) = (1 - \frac{3}{4 \cdot df - 1})^2 \cdot Var(d)$.

Correlation (continuous or binary observation) The sample (Pearson) correlation coefficient r is often used. It can be shown that $Var(r) \approx \frac{(1-r^2)^2}{n-1}$. Since the variance of r is too dependent on r itself (e.g. $Var(r) \approx 0$ when r close to 1), r is normally not used for research synthesis. If the underlying variables have a bivariate normal distribution, then $t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}$

under the null hypothesis (i.e. zero correlation). A third popular choice is Fisher's z-transformation: $z = 0.5 \cdot \log(\frac{1+r}{1-r})$. It can be shown that $z \sim N(tanh^{-1}(r), 1/(n-3))$.

log odds ratio (binary observation) When the observations are binary, a 2×2 table is established for each study with observations n_{11}, n_{12}, n_{21} and n_{22}. Log odds ratio is often used as the effect size to combine: $\log(o) = \log(\frac{n_{11} \cdot n_{22}}{n_{12} \cdot n_{21}})$. The asymptotic variance when sample size large can be calculated: $Var(log(o)) = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$.

Converting effect sizes from one to another

In research synthesis, it is common that different studies adopt different study designs and different reporting measure (effect sizes). It is important to adequately convert to comparable effect sizes before research synthesis. Below we describe methods to convert between standardized mean difference and log odds ratio and between standardized mean difference and correlation. (see p. 231 of Cooper et al., 2009)

Converting between d and $\log(o)$ It can be shown that $\log(o) \approx \frac{\pi d}{\sqrt{3}}$ and $Var(\log(o)) \approx \frac{\pi^2}{3} Var(d)$. Conversely, $d = \frac{\sqrt{3}}{\pi} \log(o)$ and $Var(d) = \frac{3}{\pi^2} Var(log(o))$.

Converting between d and r It can be shown that $r \approx \frac{d}{\sqrt{d^2+a}}$, where $a = \frac{(n_1+n_2)^2}{n_1 n_2}$ and $Var(r) \approx \frac{a^2}{(d^2+a)^3} Var(d)$. When $n_1 = n_2$, $a = 4$ and $Var(r) \approx \frac{16}{(d^2+4)^3} Var(d)$. Conversely, $d = \frac{2r}{\sqrt{1-r^2}}$ and $Var(d) = \frac{4}{(1-r^2)^3} Var(r)$.

Fixed effects model and random effects model

Consider the example of Table 14.1 (p259 of Cooper et al., 2009). How to combine correlation (or odds ratio) across studies?

Fixed effects model Suppose the observed effect sizes T_i has the underlying true population effect size θ_i and variance v_i (i.e. $T_i = \theta_i + \epsilon_i$, $var(\epsilon_i) = v_i$). In practice, T_i and estimate of v_i can be obtained from single study analysis of study i. Fixed effects model assumes that $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$. An weighted estimation of θ can be obtained by $\bar{T} = \frac{\sum_{i=1}^K w_i T_i}{\sum_{i=1}^K w_i}$. To choose the weights w_i, it can be shown that $w_i = 1/v_i$ minimizes the varaiance of $\bar{T}$ (Exercise 2). Consequently, $Var(\bar{T}) = \frac{1}{\sum_{i=1}^K (1/v_i)}$. To draw conclusion on θ, we may perform the hypothesis testing $H_0 : \theta = 0$ versus $H_A : \theta \neq 0$ and apply the statistic

$z = \frac{|\bar{T}|}{\sqrt{Var(\bar{T})}}$. When K is large, z is asymptotically a standard normal distribution.

To perform fixed effects model, it is important to test whether the assumptioin $\theta_1 = \theta_2 = \cdots = \theta_k = \theta$ is true or not. Such a homogeneity test can be done using the Q statistic: $Q = \sum_{i=1}^{K} \frac{(T_i - \bar{T})^2}{v_i} = \sum_{i=1}^{K} w_i (T_i - \bar{T})^2$. Under null hypothesis that the effect sizes are equal, Q follows chi-squared distribution with degree of freedom K-1 asymptotically. The index $I^2 = \max(0, \frac{Q-(K-1)}{Q})$ is often used to quantify degree of heterogeneity. When $I^2 < 25\%$, it is considered small heterogeneity. When $25\% \leq I^2 < 50\%$, it is considered medium heterogeneity and $50\% \leq I^2$ means large heterogeneity.

Random effects model The assumption of fixed effects model is often violated in practice. There may be many factors in the studies that can affect the effect sizes and make them uncomparable. For example, one study may apply a daily dose of 10mg of a given drug (or vitamin) while another study may apply 30mg. Other factors such as demographic variables (such as age, gender and race) or survey methods can also have impact on the effect sizes. The random effects model assumes that the underlying population effects sizes θ_i can be non-eqaual. They come from one underlying true effect size θ but with added study-specific variability σ_θ^2. The model can be described as: $T_i = \theta_i + \epsilon_i$, $Var(\epsilon_i) = v_i$ and $\theta_i = \theta + \delta_i$, $Var(\delta_i) = \sigma_\theta^2$. As a result, $Var(T_i) = \sigma_\theta^2 + v_i$.

Given the model, the question is how to estimate the variance of random effects σ_θ^2? In the first approach, one can consider $s^2(T) = \sum_{i=1}^{K} \frac{(T_i - \bar{T})^2)}{K-1}$. We can show that $E(s^2(T)) = \sigma_\theta^2 + \frac{1}{K} \sum_{i=1}^{K} \sigma^2(T_i|\theta_i)$. As a result, we can estimate the random effect variance by $\hat{\sigma_\theta}^2 = s^2(T) - \frac{1}{K} \sum_{i=1}^{K} v_i$. This estimator is, however, problematic. It can obtain negative estimation even when the homogeneity statistic Q rejects null hypothesis. A better alternative comes from $Q = \sum_{i=1}^{K} \frac{(T_i - \bar{T})^2}{v_i}$. We can show that $E(Q) = c\sigma_\theta^2 + (K-1)$, where $c = \sum_{i=1}^{K} w_i - \frac{\sum w_i^2}{\sum w_i}$. The variance of random effects is estimated as $\hat{\sigma_\theta}^2 = \frac{Q-(K-1)}{c}$. If Q rejects the null hypothesis of homogeneity test, $Q > K-1$ and $\hat{\sigma_\theta}^2 > 0$. For more details, refer to Cooper et al. (2009).

6.1.2 Combine p-values

Combining effect sizes are statistically efficient if the estimate of effect size is the goal and data fit well with the parametric model. In many

situations, estimation of effect size is not possible (e.g. survival or time series outcome is considered rather than two-sample comparison) and combining p-values provide a more flexible choice. This is especially true in many genomic meta-analysis. Below we introduce a few classical as well as recently developed methods for combining p-values, many of which have been widely used in genomic meta-analysis.

Evidence aggregation methods

In the evidence aggregation methods below, we combine p-values $p_1, \cdots, p_K$ from K studies. The p-values are transformed into an evidence score and the evidences scores of multiples studies are summed as the test statistic.

Fisher's method Fisher's method (Fisher, 1948) sums up the log-transformed p-values: $T^{Fisher} = -2\sum_{k=1}^{K} \log(p_k)$. Smaller p-values contributes larger score to the Fisher's statistic. Under the null hypothesis assuming that the studies are independent and the statistical models to assess the p-values are correct, T^{Fisher} follows a chi-squared distribution with degree of freedom $2K$.

Adaptively weighted Fisher's method One potential problem of Fisher's method is that an extremely small p-value can dominantly contribute to the large Fisher score and claim statistical significance. This cannot be distinguished with situations when marginal p-values from many studies are combined. For example, the Fisher's statistic can not distinguish from combining $\overrightarrow{p}^{(1)} = (p_1, p_2, p_3, p_4) = (0.0001, 1, 1, 1)$ and $\overrightarrow{p}^{(2)} = (p_1, p_2, p_3, p_4) = (0.1, 0.1, 0.1, 0.1)$. In both cases, $T^{Fisher} = 18.42$ that generates p=0.018. Li and Tseng (2011) proposed an adaptively weighted (AW) modification for Fisher's method to solve the problem. The AW method considers $\tilde{T}(w_1, \cdots, w_K) = -2\sum_{k=1}^{K} w_k \cdot \log(p_k)$, where w_k are weights selected from either 0 or 1. Denote by $p(\tilde{T}(w_1, \cdots, w_K))$ the p-value of the statistic when weights $(w_1, \cdots, w_K)$ are given and the optimal weight is selected as

$$(\hat{w}_1, \cdots, \hat{w}_K) = \arg \min_{\{\text{all possible 0-1 weights}\}} p(\tilde{T}(w_1, \cdots, w_K)).$$

The AW statistic is defined as

$$\begin{aligned} T^{AW} &= p(\tilde{T}(\hat{w}_1, \cdots, \hat{w}_K)) = \min_{\{\text{all possible 0-1 weights}\}} p(\tilde{T}(w_1, \cdots, w_K)) \\ &= \min_{\{\text{all possible 0-1 weights}\}} 1 - F_{\chi^2(2\sum w_k)}(-2\sum w_k \cdot log(p_k)). \end{aligned} \tag{6.1}$$

Intuitively, the 0-1 weights of $\hat{w}_1, \cdots, \hat{w}_K$ indicate the subset of studies that contributes to the statistical significance of the meta-analysis. For example, the estimated weights for $\overrightarrow{p}^{(1)}$ will generate $(\hat{w}_1, \cdots, \hat{w}_4) = (1, 0, 0, 0)$ and that of $\overrightarrow{p}^{(2)}$ will be $(\hat{w}_1, \cdots, \hat{w}_4) = (1, 1, 1, 1)$.

The adaptive weighting method considers all combinatorial subsets of studies and finds the best subest to describe the meta-analysis contribution. This concept has received increasing attention and has been modified for fixed effects model for GWAS meta-analysis using frequentist (Bhattacharjee et al., 2012) or Bayesian approaches (Han and Askin, 2012). Section 6.2.2 discusses further the statistical properties of AW-Fisher. Conceptually, the AW concept can be applied to methods other than Fisher (e.g. AW-FEM, AW-Stouffer, etc).

?? inference: closed form solution for null distribution or permutation.

Stouffer's method Stouffer's method (Stouffer et al, 1949) adopts a different p-value transformation: $T^{Stouffer} = -\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Phi^{-1}(p_k)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution. Similar to Fisher's statistic, smaller p-values contributes more to the score. Under the null hypothesis, $T^{Stouffer} \sim N(0, 1)$. Note that if p_k's are obtained from two-sided tests, only the right-side rejection region for $T^{Stouffer}$ should be used. That is, the p-value for Stouffer's score is calculated as $p(T^{Stouffer}) = 1 - \Phi(T^{Stouffer})$. But if p_k's are from one-sided tests (e.g. $p_k \approx 0$ means significant up-regulation and $p_k \approx 1$ means significant down-regulation for a given gene), both sides of rejection regions should be considered: $p(T^{Stouffer}) = 2 \cdot \min\{\Phi(T^{Stouffer}), 1 - \Phi(T^{Stouffer})\}$. In this case, Stouffer automatically detects genes with concordant DE direction across studies (see Section 6.2.3).

Vote counting method The vote counting method counts the number of studies that have p-values smaller than a threshold α (e.g. $\alpha = 0.05$): $T^{VC} = \sum_{k=1}^{K} 1_{\{p_k < \alpha\}}$. Denote by $\pi = E(T^{VC})/K$. Binomial test can be used to test $H_0 : \pi = \pi_0$ vs $H_A : \pi > \pi_0$ for a given π_0 (e.g. π_0=0.5). The vote counting method has been found powerless asymptotically when the effect size in each study is moderate (Hedges and Olkins, 1980).

Other methods of this category Other transformations, including logit (Lancaster, 1961) and inverse chi-square transformation with varying degrees of freedom (George, 1977), have also been proposed. Different weights or variations of Fishers statistic have also been considered. Good (1955) suggested using unequal weights for individual studies in which weights are determined by subject experts. Olkin and Saner (2001) have proposed a trimmed version of Fishers statistic to remove the potential effects of

aberrant extremes. For comprehensive reviews and comparisons of various meta-analysis approaches, see Hedges and Olkin (1985) and Cousins (2007).

Order statistics methods

Another category of p-value combination method is to use the order statistic of the observed p-valued as the test statistic.

maximum p-value (maxP) The maxP method takes the maximum p-value of all studies as the test statistic: $T^{maxP} = \max_{1\leq k\leq K} p_k$. Under null hypothesis, T^{maxP} follows a beta distribution with parameters K and 1. Intuitively, maxP method requires all studies to have small p-values to reject the null hypothesis.

minimum p-value (minP) In contrast to maxP, the minP method takes the minimum p-value of all studies as the test statistic: $T^{minP} = \min_{1\leq k\leq K} p_k$. Under null hypothesis, T^{minP} follows a beta distribution with parameters 1 and K. Intuitively, minP method can reject the null hypothesis if a small enough p-value is obtained from any of the K studdies.

rth order p-value (rOP) The rOP method takes the order statistic as the test statistic: $T^{rOP} = p_{(r)}$, where $p_{(1)}, \cdots, p_K$ are the sorted p-values.Under the null hypothesis, T^{rOP} follows a beta distribution with parameter r and $K - r + 1$. The minP and maxP methods are special cases of rOP when r=1 and K, respectively. The rOP statistic is shown to be an inverse function of vote counting. Denote by $T^{VC} = r = f(\alpha) = \sum_{k=1}^{K} 1_{\{p_k<\alpha\}}$. It can be shown that $T^{rOP} = \alpha = f^{-1}(r) = p_{(r)}$. Although rOP is closely connected to vote countint, it does not have the undesirable powerless property as vote counting has (see section 4.2 in Song and Tseng; 2013).

Table 6.1 shows four hypothetical genes to compare different p-value combination methods. Gene A and B shows that Fisher, , AW, Stouffer and minP detect markers that have small p-values in "one or more" studies. Specifically, gene B has a very small p-value in only one study and the four methods all detect it. (??discuss gene C and gene D??) The result clearly shows the pros and cons of different hypothesis settings and methods and indicate the different hypothesis settings behind the methods (??improve the table, add AW). See discussion of formal hypothesis settings in Section 6.2.

Table 6.1: Four hypothetical genes to compare different mea-analysis methods and to illustrate the motivation of rOP (*: p-values smaller than 0.05)

	gene A	gene B	gene C	gene D
Study 1	0.1	1E-20	0.25	0.15
Study 2	0.1	0.9	0.25	0.15
Study 3	0.1	0.9	0.25	0.15
Study 4	0.1	0.9	0.25	0.15
Study 5	0.1	0.9	0.25	0.9
Fisher (HS_B)	0.01*	1E-15*	0.18	0.12
Stouffer (HS_B)	0.002*	0.03*	0.07	0.10
minP (HS_B)	0.41	5E-20*	0.76	0.56
maxP (HS_A)	1E-5*	0.59	0.001*	0.59
rOP ($r = 4$) (HS_r)	5E-4*	0.92	0.015*	0.002*

6.1.3 Simulation and power comparison of different methods

Power comparison between FEM, REM and Fisher

Power comparison between Fisher, minP and AW

6.2 Hypothesis settings and statistical properties

6.2.1 Different hypothesis settings

Following the convention of Birnbaum (1954) and Li and Tseng (2011), the hypothesis setting of different meta-analysis methods can be categorized into two extreme situations:

$$HS_A : \left\{ H_0 : \bigcap_{k=1}^{K} \{\theta_k = 0\} \text{ versus } H_A : \bigcap_{k=1}^{K} \{\theta_k \neq 0\} \right\}$$

$$HS_B : \left\{ H_0 : \bigcap_{k=1}^{K} \{\theta_k = 0\} \text{ versus } H_A : \bigcup_{k=1}^{K} \{\theta_k \neq 0\} \right\}.$$

In the first hypothesis setting (HS_A), the testing will be rejected only when the effect sizes of "all" studies are non-zero (e.g. gene A in Table 6.1). For the second hypothesis setting (HS_B), it is rejected whenever

"one or more" studies are non-zero (gene A-D in Table 6.1). Among methods described in this Chapter, all evidence aggregation methods (Fisher, AW, Stouffer and logit) and minP belong to HS_B. The maxP method and fix effects model belong to HS_A. Random effects models does not belong to HS_A but is very close (since when the overall effect size is non-zero, effect sizes of some studies may be zero due to random effects). We note that HS_B is identical to the a classical union-intersection test (UIT) (Roy, 1953) but HS_A is different from intersection-union test (IUT) (Berger, 1982; Berger and Hsu, 1996). In IUT, the statistical hypothesis is in complementary form between null and alternative hypothesis:

$$HS_{IUT} : \left\{ H_0 : \bigcup_{k=1}^{K} \{\theta_k = 0\} \text{ versus } H_A : \bigcap_{k=1}^{K} \{\theta_k \neq 0\} \right\}.$$

According to Song and Tseng (2013a), intermediate hypothesis settings in between HS_A and HS_B can be developed. Define HS_r $(1 \leq r \leq K)$ as

$$HS_r : \left\{ H_0 : \bigcap_{k=1}^{K} \{\theta_k = 0\} \text{ versus } H_A : \sum_{k=1}^{K} I\{\theta_k \neq 0\} \geq r \right\}$$

The rOP method is designed for HS_r hypothesis setting. In Song and Tseng (2013), the parameter r in rOP was restricted to $[(K+1)/2] \leq r \leq K$ so that rOP is considered as a relaxed (or robust) form of HS_A.

Song and Tseng (2013b) have discovered that HS_A and HS_r can be anti-conservative in genomic applications since the null hypothesis assumes differential expression in zero study and it is not complement to the alternative hypothesis (??explain more later??). For HS_A, HS_{IUT} is a better hypothesis setting. For HS_r, $HS_{r'}$ below shoud be considered instead:

$$HS_{r'} : \left\{ H_0 : \sum_{k=1}^{K} I\{\theta_k \neq 0\} < r \text{ versus } H_A : \sum_{k=1}^{K} I\{\theta_k \neq 0\} \geq r \right\}$$

They developed a semi-parametric Bayesian mixture model to accommodate the composite null hypothesis in HS_{IUT} and $HS_{r'}$ and provided a Bayes factor for comparing two competing null and alternative hypotheses for decision making.

6.2.2 Statistical properties of the methods

Statistical power and admissibility

– Compare power between Fisher, minP and AW-Fisher When h=1, minP is more powerful than Fisher. AW-Fisher is only slightly lower power than minP. When $h > 4$, Fisher is much more powerful than minP. AW-Fisher is again only slightly less powerful than Fisher. In both extremes, AW-Fisher is always near the more powerful method.

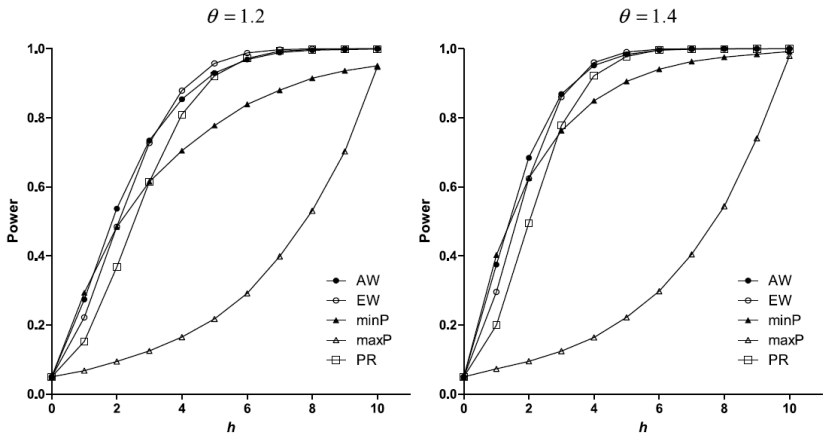

Figure 6.1: Statistical power of Fisher, minP and AW under different alternative hypotheses. EW: Fisher; AW: AW-Fisher. Figure comes from Li and Tseng, 2011.

– No UMP test; Fisher, Stouffer, minP and AW-Fisher are all admissible.

Performance of different meta-analysis methods are difficult to compare as it relates to the hypothesis setting and underlying data distributions. For performance of different hypothesis testing methods, we often ask two questions: (1) whether there exists a uniformly most powerful (UMP) test, and (2) whether a method is admissible. For UMP test, one is interested in whether there exists a best method that has better or no less statistical power than any other method under all alternative hypothesis scenarios. Birnbaum (1954, 1955??) has shown that no p-value combination meta-analysis method is UMP under HS_B even in the most simplified situation. He then established general conditions for evaluation different methods, including monotonicity and admissibility. A test

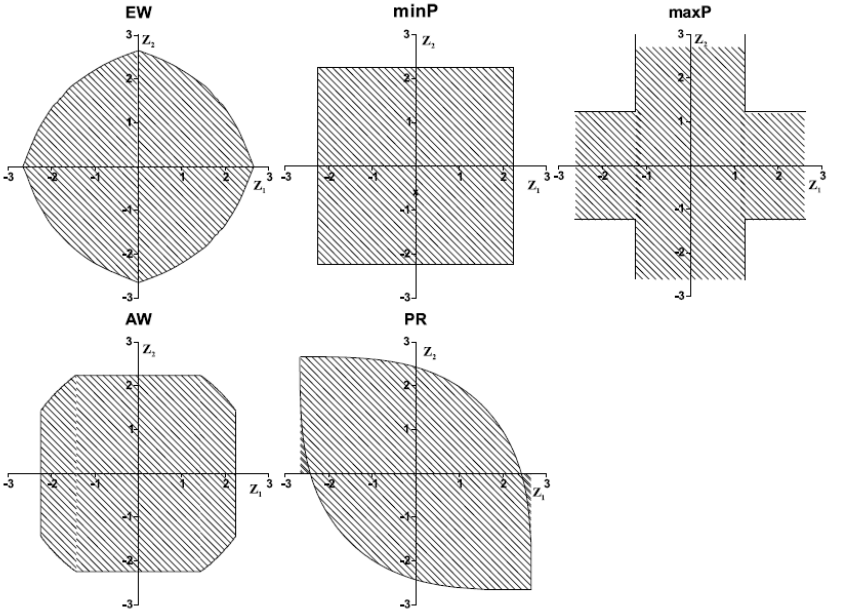

Figure 6.2: Rejection regions of different meta-analysis methods when combining two studies under Gaussian assumption. Figure comes from Li and Tseng, 2011.

is considered admissible if it cannot be uniformly improved by any other test. Consider a two-sample test of the mean of a Gaussian distribution with known variance (see Section 5 of Li and Tseng 2011). It has been shown that a combination method is admissible under the simplified situation if and only if the acceptance region is convex. Figure 6.2 shows that Fisher, minP, Stouffer and AW are all admissible. The maxP method is not admissible as it does not target on HS_B.

Asymptotic properties

–Fisher is ABO when all studies have equal non-zero effect sizes. AW-Fisher is ABO in a more general class (to be proved in Shaowu's thesis). Although Fishers method is not the most uniformly powerful, it does exhibit good power for a wide range of conditions. It is also recognized for its asymptotically Bahadur optimal (ABO) characteristic, when the studies have the same effect size for alternative hypotheses [Littell and

Folks (1971, 1973)].

– vote counting asymptotically powerless.

– rOP does not have the problem of vote couting.

6.2.3 Concordance of effect sizes

Most p-value combination methods combine two-sided p-values by default when two-sample comparison is performed. Such a practice has a potential issue that the hypothesis may be rejected with positive effect size in one study but negative in the other. To avoid this problem, Owen (2009) and Pearson (1934) applied a one-sided test form of Fishers method to address the possible discordance issue. Two Fisher scores are first obtained from left and right one-sided p-values: $S^{Fisher;L} = -2\sum_{k=1}^{K} log(\tilde{p}_k)$ and $S^{Fisher;R} = -2\sum_{k=1}^{K} log(1-\tilde{p}_k)$, where $\tilde{p}_k$ is the left-sided p-value of study k. The one-sided corrected Fisher score is defined as $S^{Fisher;C} = \max(S^{Fisher;L}, S^{Fisher;R})$. Similar modification can be applied to minP, maxP and rOP as well (see Song and Tseng, 2013a). For Stouffer's method, either one-sided p-value or a similar z-transformation of two-sided p-value ($T = \sum_{k=1}^{K} \Phi^{-1}(1-p_k/2) \cdot \{\text{sign of effect size}\}$) considering effect size direction have been widely used in GWAS meta-analysis (Tseng et al., 2011).

6.3 Genomic meta-analysis

6.3.1 Microarray meta-analysis

The NAR review paper. Tseng et al. (2012)
combine effect size; combine p-value, combine ranks and directly merge

A few R packages are available to implement one or a few microarray meta-analysis methods described above: MetaArray, metaMA, GeneMeta. The MetaDE (Wang et al., 2012) package provides a comprehensive coverage of ¿12 microarray meta-analysis methods.

6.3.2 GWAS meta-analysis

The NAR reivew paper. Begum et al. (2012).

6.4 Issues and other types of genomic meta-analysis

Inclusion/exclusion criteria and quality assessment

One of the most important obstacles to successful meta-analysis is dataset quality (Eysenck, 1994). Inclusion of a poor quality or outlying study in the information integration can greatly dilute information contained, weaken statistical power or even distort final biological conclusions. In the meta-analysis literature, many quality assessment protocols and systems have been proposed. The procedure, however, still inevitably involve expert opinion and human intervention. For microarray meta-analysis, most existing studies either apply subjective expert opinion or ad hoc quality control criteria (e.g. large enough sample size or good quality array platform). A set of methodology called MetaQC (Kang et al., 2011) has been developed for objective quality control and alleviate such potential pitfalls in meta-analysis. The method considers homogeneity and reproducibility of combined studies and biological validation from external pathway information.

Publication bias

Related reading:

- Larry V. Hedges and Ingram Olkin. Statistical Methods for Meta-Analysis. 1985.
- Stouffer, S. A., Suchman, E. A , DeVinney, L.C., Star, S.A., Williams, R.M. Jr (1949). Adjustment During Army Life. Princeton, NJ, Princeton University Press.
- Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd (Edinburgh).
- Harris M. Cooper,Larry V. Hedges,Jeff C. Valentine. The handbook of research synthesis and meta-analysis. 2009. 2nd Edition.
- Rudy Guerra, Darlene Renee Goldstein. Meta-analysis and combining information in genetics and genomics. 2010. Chapman & Hall/CRC.
- George C. Tseng, Debashis Ghosh and Eleanor Feingold. (2012) Comprehensive literature review and statistical considerations for

microarray meta-analysis. Nucleic Acids Research 40 (9): 3785-3799.

- Ferdouse Begum, Debashis Ghosh, George C. Tseng, Eleanor Feingold. (2012) Comprehensive literature review and statistical considerations for GWAS meta-analysis. Nucleic Acids Research 40 (9): 3777-3784.
- Jia Li and George C. Tseng. (2011) An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. Annals of Applied Statistics. 5:994-1019.
- Song Chi and George C. Tseng. (2013a) Hypothesis setting and order statistic for robust genomic meta-analysis. Annals of Applied Statistics. In revision.
- Song Chi and George C. Tseng. (2013b) Semi-parametric Bayesian Approach for a Hierarchical Mixture Model in Genomic Meta-analysis.
- Eysenck,H. (1994) Systematic reviews: meta-analysis and its problems. BMJ, 309, 789.
- Dongwan D. Kang, Etienne Sibille, Naftali Kaminski, and George C. Tseng. (2011) MetaQC: Objective Quality Control and Inclusion/Exclusion Criteria for Genomic Meta-Analysis. Nucleic Acids Research. 40(2):e15.
- Xingbin Wang, Dongwan Kang, Kui Shen, Chi Song, Shuya Lu, Lunching Chang, Serena G. Liao, Zhiguang Huo, Naftali Kaminski, Etienne Sibille, Yan Lin, Jia Li and George C. Tseng. (2012) A Suite of R Packages for Quality Control, Differentially Expressed Gene and Enriched Pathway Detection in Microarray Meta-analysis. Bioinformatics. 28:2534-2536.
- Samsiddhi Bhattacharjee, Preetha Rajaraman, Kevin B. Jacobs, William A. Wheeler, Beatrice S. Melin, Patricia Hartge, GliomaScan Consortium, Meredith Yeager, Charles C. Chung, Stephen J. Chanock, and Nilanjan Chatterjee. (2012) A Subset-Based Approach Improves Power and Interpretation for the Combined Analysis of Genetic Association Studies of Heterogeneous Traits. The American Journal of Human Genetics 90, 821835.
- Buhm Han, Eleazar Eskin. (2012) Interpreting Meta-Analyses of Genome-Wide Association Studies. PLOS Genetics.

Exercise:

1. (an exercise from pathway analysis)
   (1) Simulate two groups $x_1, \cdots, x_{50} \sim N(0,1)$ and $y_1, \cdots, y_{50} \sim N(1,1)$. Designate the first 15 observations of group 1 and the first 5 observations of group 2 (i.e. $x_1, \cdots, x_{15}, y_1, \cdots, y_5$) as in the pathway and the other 80 observations as outside the pathway.
   (2) Write an R function (without using "ks.test" function in R) to perform KS-test for this simultated data set: (a) First calculate the observated KS-statistics. (b) Perform permutation analysis for $B = 100,000$ times to generate null distribution of the KS-statistic and derive the p-value.
   (3) Apply the "ks.test" function in R to derive the p-value and compare the result.

2. Show that the selection weights $w_i = 1/v_i$ is optimal in the sense that the variance $\bar{T}$ is minimized. Since $\bar{T}$ is unbiased, such a selection is optimal.

3. Prove the null distributions of Fisher's method, maxP, minP and rOP.