

IMPUTATION OF TRUNCATED P-VALUES FOR META-ANALYSIS METHODS AND ITS GENOMIC APPLICATION

BY SHAOWU TANG, YING DING, ETIENNE SIBILLE, JEFFREY S. MOGIL, WILLIAM R.
LARIVIERE, AND GEORGE C. TSENG

University of Pittsburgh

Microarray analysis to monitor expression activities in thousands of genes simultaneously has become routine in biomedical research during the past decade. A tremendous amount of expression profiles are generated and stored in the public domain and information integration by meta-analysis to detect differentially expressed (DE) genes has become popular to obtain increased statistical power and validated findings. Methods that aggregate transformed p-value evidence have been widely used in genomic settings, among which Fisher's and Stouffer's methods are the most popular ones. In practice, raw data and p-values of DE evidence are often not available in genomic studies **that are to be combined**. Instead, only the detected DE gene lists under a certain p-value threshold (e.g. DE genes with p-value < 0.001) are reported in journal publications. The truncated p-value information makes the aforementioned meta-analysis methods inapplicable and researchers are forced to apply a less efficient vote counting method or naïvely drop the studies with incomplete information. The purpose of this paper is to develop effective meta-analysis methods for such situations with partially censored p-values. We developed and compared three imputation methods – mean imputation, single random imputation and multiple imputation – for a general class of evidence aggregation methods of which Fisher's and Stouffer's methods are special examples. The null distribution of each method was analytically derived and subsequent inference and **genomic analysis frameworks** were established. Simulations were performed to investigate the type I error, power and the control of false discovery rate (FDR) for (correlated) gene expression data. The proposed methods were applied to several genomic applications in colorectal cancer, pain and liquid association analysis of major depressive disorder (MDD). The results showed that imputation methods outperformed existing naïve approaches. Mean imputation and multiple imputation methods performed the best and are recommended for future applications.

1. Introduction and motivation. Microarray analysis to monitor expression activities in thousands of genes simultaneously has become routine in biomedical research during the past decade. The rapid development in biological high-throughput technology results in a tremendous amount of experimental data and many datasets are available from public domains such as Gene Expression Omnibus (GEO) and ArrayExpress. Since most microarray studies have relatively small sample sizes and limited statistical power, integrating information from multiple transcriptomic studies using meta-analysis techniques is becoming popular. Microarray meta-analysis usually refers to combining multiple transcriptomic studies for detecting differentially expressed (DE) genes (or candidate markers). DE gene analysis identifies genes differentially expressed across two or more

AMS 2000 subject classifications: Microarray analysis, meta-analysis, Fisher's method, Stouffer's method, missing value imputation

conditions (e.g., cases and controls) with statistical significance and/or biological significance (e.g., fold change). Microarray meta-analysis in many situations refers to performing traditional meta-analysis techniques on each gene repeatedly and then controlling the false discovery rate (FDR) to adjust p-values for multiple comparison (Borovecki et al. 2005; Cardoso et al. 2007; Pirooznia et al. 2007; Segal et al. 2004). Fisher’s method (Fisher 1931) was the first meta-analysis technique introduced in microarray data analysis in 2002 (Rhodes et al. 2002), followed by Tippett’s minimum p-value method in 2003 (Moreau et al. 2003). Subsequently many meta-analysis approaches have been used in this field, including extensions of existing meta-analysis techniques and novel methods to encompass the challenges presented in the genomic setting (Choi et al. 2003, Choi et al. 2007, Moerau et al. 2003, Owen 2009, Li and Tseng 2011, and see a review paper Tseng et al. 2012).

To combine findings from multiple research studies, one needs to know either the effect size or the p-value for each study. Since the differences in data structures and statistical hypotheses across multiple studies may make the direct combination of effect sizes impossible or the result suspicious, combining p-values from multiple studies is often more appealing. **Popular p-value combination methods (see review and comparative papers Tseng et al. 2012 and Chang et al. 2013) can be split into two major categories of evidence aggregation methods (including Fisher’s, Stouffer’s and logit methods) and order statistic methods (such as minimum p-value, maximum p-value and r-th ordered p-value by Song et al. 2014). Evidence aggregation methods utilize summation of certain transformations of p-values as their test statistics to aggregate differential expression evidence across studies.** Among evidence aggregation methods, Fisher’s method is the most well-known, in which the test statistic is defined as $T^{Fisher} = -2 \sum_{k=1}^K \log(p_k)$, where K is the number of independent studies that are to be combined and p_k is the p-value of individual study $k, 1 \leq k \leq K$. Under the null hypothesis of no effect size in all studies and assuming that studies are independent and models for assessing p-values are correctly specified, T^{Fisher} follows a chi-square distribution with degrees of freedom $2K$. **Fisher’s method has been popular** due to its simplicity and some theoretical properties, including admissibility under Gaussian assumption (Birnbaum 1954 & 1955) and asymptotically Bahadur optimality (ABO) under equal non-zero effect sizes across studies (Littel and Folk, 1971 & 1973). Some variations of Fisher’s methods were proposed by using unequal weights or a trimmed version of Fisher’s test statistic (Olkin and Saner, 2001). Another widely used evidence aggregation method is the Stouffer’s method (**Stouffer 1949**), in which the test statistic is defined as $T^{Stouffer} = \sum_{k=1}^K \Phi^{-1}(p_k)$, where $\Phi^{-1}(\cdot)$ is the inverse cumulative distribution function (CDF) of standard normal distribution.

In order to combine p-values, all p-values across studies should be known. In genomic applications, however, raw data and thus p-values are often not available and usually only a list of statistically significant DE genes (p-value less than a threshold) is provided in the publication (Griffith et al., 2006). Although many journals and funding agencies have encouraged or enforced data sharing policies, the situation has only improved moderately. Many researchers are still concerned about data ownership, and researchers whose studies are sponsored by private funding are not obligated to share data in the public domain. For example, in [Chan et. al 2007], publications of 23 colorectal cancer versus normal gene expression profiling studies were collected to perform meta-analysis to identify consistently reported candidate disease-associated genes. However only one raw dataset is available from the Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>, GSE3294) and most other papers only provided a list of DE genes (and their p-values) under a pre-specified

p-value threshold. A second motivating example comes from a microarray meta-analysis study for pain research (LaCroix-Fralish 2011), in which 19 microarray studies of pain models were collected to detect the gene signature and patterns of pain conditions. Among the 19 studies, only one raw dataset was available on the author’s website and all the other papers reported the DE gene lists under different thresholds.

In these two motivating examples (details to be shown in Section 4.1 and 4.2), the incomplete data forced researchers to either drop studies with incomplete p-values or apply the convenient vote counting method (Hedges and Olkin, 1980). Dropping studies with incomplete information greatly reduces the statistical power and is obviously not applicable in the two motivating examples since the complete data was available in only one study. The conventional vote counting procedure is well-known as flawed and low-powered (McCarley et al., 2001). Ioannidis et al., (2009) attempted to reproduce 18 microarray studies published in Nature Genetics during 2005-2006. Interestingly, only two were ”in principle” replicated, six ”partially” replicated and ten could not be reproduced. This result illustrates well the wide-spread difficulty of obtaining raw data or reproducing published results in the field. Therefore, developing methods to efficiently combine studies with truncated p-value information is an important problem in microarray meta-analysis.

In this paper, we assume that $K = K_1 + K_2$ studies are combined. In K_1 studies, the raw gene expression data matrix and sample annotations are available and the complete p-values p_{gi} ($1 \leq g \leq G$ for genes and $1 \leq i \leq K_1$) can be reproduced for meta-analysis. For the remaining K_2 studies, either the raw data or annotation are not available. Only incomplete information of a DE gene list (under p-value threshold α_i for study i) is provided in the journal publication. In this situation, the available information is an indicator function $\mathbb{1}_{\{p_{gi} < \alpha_i\}}$ to represent whether the p-value of gene g in study i is smaller than α_i or not. We outline the paper structure as the following. In Section 2, a general class of evidence aggregation meta-analysis methods under a univariate scenario were investigated for the mean imputation, the single random imputation and the multiple imputation methods respectively, in which the exact or approximate null distributions were derived under the null hypotheses and the results are shown for the Fisher’s and the Stouffer’s methods. In Section 3.1 simulations of expression profile were performed to compare performance of different methods. Simulations were further performed in Section 3.2 using 8 major depressive disorder (MDD) and 7 prostate cancer studies where raw data were completely available and the true best performance (complete case) could be obtained. In Section 4 the proposed methods were applied to the two motivating examples. In Section 4.1, the methods were applied to 7 colorectal cancer studies, where the raw data were available only in 3 studies. In Section 4.2, the proposed methods were applied to 11 microarray studies of pain conditions, where no raw data was available. In Section 4.3, we developed an unconventional application of the proposed methods to facilitate the large computational and data storage needs in a liquid association meta-analysis. Discussions and conclusions are included in Section 5 and all proof is left in the Appendix.

2. Methods and inferences.

2.1. *Evidence aggregation meta-analysis methods.* **Here we consider** a general class of univariate evidence aggregation meta-analysis methods (for gene g fixed), in which the test statistics are defined as the sum of selected transformations of p-values for each individual study. Without loss of generality, assuming that $F_X(\cdot)$ is the cumulative distribution function (CDF) of **a continuous**

random variable X , the test statistic T is defined as

$$(2.1) \quad T = \sum_{i=1}^K T_k := \sum_{k=1}^K F_X^{-1}(p_k),$$

where p_k is the p-value from the k th study. Theoretically X can be any continuous random variable. However, in practice, X is usually selected such that the test statistic T follows a simple distribution. For instance, when $X \sim \chi_2^2$, it holds $T \sim \chi_{2K}^2$ (Fisher's method) and $T \sim N(0, K)$ holds, provided $X \sim N(0, 1)$ (Souffer's method).

The hypothesis that corresponds to testing the homogeneous effect sizes of K studies by evidence aggregation methods is a union-intersection test (UIT)(Roy 1953):

$$(2.2) \quad H_0 : \bigcap_{k=1}^K \{\theta_k = 0\} \text{ versus } H_A : \bigcup_{k=1}^K \{\theta_k \neq 0\}.$$

In this paper, we focus on two popular special cases:

1. Fisher's method (Fisher 1931): When $X \sim \chi_2^2$, $T_k = F_X^{-1}(p_k) = -2 \log(p_k)$.
2. Stouffer's method (Stouffer 1949): When $X \sim N(0, 1)$, $T_k = F_X^{-1}(p_k) = \Phi^{-1}(p_k)$.

Another example is the logit method (Hedges and Olkin, 1985), where $T_k = -\log(\frac{p_k}{1-p_k})$. But since this method is rarely used in practice, we will not examine further in this paper. To apply the evidence aggregation meta-analysis methods mentioned above, all the p-values should be observed. However, in genomic applications, it often happens that p-values of some studies are truncated and only their ranges are reported. Two naïve methods are commonly used to overcome this situation: vote counting method or the available-case method which only combines studies with observed p-values. The available-case method discards rich information contained in the studies with truncated p-values, and therefore the statistical power is reduced. Hedges and Olkin (1980) showed that the power of vote counting converges to 0 when many studies of moderate effect sizes are combined and therefore the vote counting method should be avoided whenever possible. In this section, three imputation methods - mean imputation, single random imputation and multiple imputation method - are proposed and investigated to combine studies with truncated p-values and the corresponding null distributions are derived analytically, respectively. **We first define some notations.**

Assume that K independent studies are to be combined and p_1, \dots, p_K are the corresponding p-values. Without loss of generality, assume that all the p-values are available in the the first K_1 studies and only the indicator function of DE evidence are reported in the other K_2 studies.

Define a pair (c_i, x_i) , $i = 1, \dots, K$ for each study, in which c_i is the "censoring" indicator satisfying

$$(2.3) \quad c_i := \begin{cases} 0, & \text{if } p_i \text{ is observed (i.e., } 1 \leq i \leq K_1), \\ 1, & \text{if } p_i \text{ is censored (i.e., } K_1 + 1 \leq i \leq K), \end{cases}$$

and x_i is the final observed values which is defined as

$$(2.4) \quad x_i := \begin{cases} p_i, & \text{if } c_i = 0, \\ \mathbb{1}_{\{p_i < \alpha_i\}}, & \text{if } c_i = 1, \end{cases}$$

where α_i is the p-value threshold for study i ($K_1 + 1 \leq i \leq K_1 + K_2 = K$). For each $i = 1, 2, \dots, K$, one can impute the missing value by \tilde{p}_i :

$$\tilde{p}_i = p_i \cdot \mathbb{1}_{\{c_i=0\}} + [q_i \cdot \mathbb{1}_{\{x_i=1\}} + r_i \cdot \mathbb{1}_{\{x_i=0\}}] \cdot \mathbb{1}_{\{c_i=1\}}$$

with $q_i \in (0, \alpha_i)$, and $r_i \in [\alpha_i, 1)$. Section 2.2-2.4 develop three imputation methods for selection of q_i and r_i .

2.2. Mean imputation method. The simplest imputation method is the mean imputation method, in which $q_i = \frac{\alpha_i}{2}$ and $r_i = \frac{1+\alpha_i}{2}$. Then the test statistic \tilde{T} for truncated data satisfies

$$(2.5) \quad \tilde{T} = \sum_{i=1}^K \tilde{T}_i = \sum_{i=1}^K F_X^{-1}(\tilde{p}_i) = \sum_{i=1}^{K_1} F_X^{-1}(p_i) + \sum_{j=1}^{K_2} F_X^{-1}(\tilde{p}_{K_1+j}) = A + \sum_{j=1}^{K_2} B_j,$$

with

$$(2.6) \quad A = \sum_{i=1}^{K_1} F_X^{-1}(p_i) \text{ and } B_j = F_X^{-1}(\tilde{p}_{K_1+j}) = F^{-1}\left(\frac{\alpha_{K_1+j}}{2}\right) \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + F^{-1}\left(\frac{1 + \alpha_{K_1+j}}{2}\right) \cdot \mathbb{1}_{\{p_{K_1+j} \geq \alpha_{K_1+j}\}}$$

for $j = 1, \dots, K_2$. Recall that under null hypothesis, the random variable A satisfies $A \sim \chi_{2K_1}^2$ for the Fisher's method and $A \sim N(0, K_1)$ for the Stouffer's method. Obviously B_j follows a Bernoulli distribution.

The results can be summarized into the following theorem (proof left to online supplementary file):

THEOREM 1. For $j = 1, 2, \dots, K_2$ and given t , by defining

$$(2.7) \quad b_j = F_X^{-1}\left(\frac{\alpha_{K_1+j}}{2}\right) - F_X^{-1}\left(\frac{1 + \alpha_{K_1+j}}{2}\right) \text{ and } c = \sum_{j=1}^{K_2} F_X^{-1}\left(\frac{1 + \alpha_{K_1+j}}{2}\right),$$

it holds

$$(2.8) \quad \mathbb{P}(\tilde{T} \leq t) = \sum_{(j_1, \dots, j_{K_2}) \in \{0,1\}^{K_2}} \prod_{i=1}^{K_2} \alpha_{K_1+i}^{j_i} (1 - \alpha_{K_1+i})^{1-j_i} F_A\left(t - c - \sum_{i=1}^{K_2} j_i b_i\right),$$

where $F_A(\cdot)$ is the CDF of A . Given the CDF, the expected values of test statistic \tilde{T} under null distributions can be calculated as follows.

1. For the Fisher's method, it holds

$$\mathbb{E}(\tilde{T}) = 2K_1 - 2 \sum_{j=1}^{K_2} \left[\alpha_{K_1+j} \log\left(\frac{\alpha_{K_1+j}}{2}\right) + (1 - \alpha_{K_1+j}) \log\left(\frac{1 + \alpha_{K_1+j}}{2}\right) \right],$$

while the expectation of the original T is $\mathbb{E}(T) = 2K_1 + 2K_2 = 2K$.

2. For the Stouffer's method, it holds

$$\mathbb{E}(\tilde{T}) = \sum_{j=1}^{K_2} \left[\alpha_{K_1+j} \Phi^{-1}\left(\frac{\alpha_{K_1+j}}{2}\right) + (1 - \alpha_{K_1+j}) \Phi^{-1}\left(\frac{1 + \alpha_{K_1+j}}{2}\right) \right],$$

while the expectation of the original T is $\mathbb{E}(T) = 0$.

Note that there are 2^{K_2} terms summation in the right hand side of Equ. (2.8), which may cause severe computing problem when K_2 is large. However, when some α_i are equal, the formula can be simplified. Without loss of generality, assume there are $r \geq 1$ different p-value thresholds $\{\beta_1, \dots, \beta_r\}$ such that

$$(2.9) \quad \sum_{j=1}^{K_2} \mathbb{1}_{\{\alpha_{K_1+j}=\beta_1\}} = n_1, \dots, \sum_{j=1}^{K_2} \mathbb{1}_{\{\alpha_{K_1+j}=\beta_r\}} = n_r \text{ and } \sum_{l=1}^r n_l = K_2,$$

then by defining $f(j; n_l, \beta_l) := \frac{n_l!}{j!(n_l-j)!} \beta_l^j (1-\beta_l)^{n_l-j}$ for $j = 0, \dots, n_l$ and $l = 1, \dots, r$, the formula can be simplified as

$$(2.10) \quad \mathbb{P}(\tilde{T} \leq t) = \sum_{j_1=0}^{n_1} \dots \sum_{j_r=0}^{n_r} \prod_{l=1}^r f(j_l; n_l, \beta_l) F_A(t - c - \sum_{l=1}^r j_l (F_X^{-1}(\frac{\beta_l}{2}) - F_X^{-1}(\frac{1+\beta_l}{2}))).$$

Therefore, the summation is reduced from 2^{K_2} terms to $\prod_{l=1}^r (n_l + 1)$ terms.

From the above theorem one concludes that \tilde{T} is a biased estimator of the original T . This motivates the following two stochastic imputation methods.

2.3. Single random imputation method. It is well-known that the mean imputation method will underestimate the variance of $\{p_{K_1+j}\}_{j=1}^{K_2}$ (Little and Rubin 2002). Furthermore, Theorem 1 indicates that the test statistic \tilde{T} from the mean imputation method is a biased estimator of the original T . To avoid this problem, one can replace the mean by randomly simulating q_i and r_i from $\text{Uniform}(0, \alpha_i)$ and $\text{Uniform}(\alpha_i, 1)$ respectively.

Recall that for $j = 1, \dots, K_2$, $B_j = F_X^{-1}(\tilde{p}_{K_1+j})$. The next theorem (proof left to online supplementary file) states that $B_j \sim X$ holds under the null hypothesis, **i.e., B_j and X follow the same distribution.**

THEOREM 2. For $j = 1, 2, \dots, K_2$, it holds

$$(2.11) \quad B_j \sim X.$$

The following corollary is a simple consequence of the above theorem.

COROLLARY. For the single random imputation method, the following facts hold for \tilde{T} :

1. For Fisher's method, it holds $B_j \sim \chi_2^2$ and therefore $\tilde{T} \sim \chi_{2K}^2$.
2. For Stouffer method, it holds $B_j \sim N(0, 1)$ and therefore $\tilde{T} \sim N(0, K)$.

Therefore, in this case, \tilde{T} is a unbiased estimator of T defined in Equ. (2.1).

2.4. Multiple imputation method. Although the single random imputation method allows the use of standard complete-data meta-analysis methods, it cannot reflect the sampling variability from one random sample. The multiple imputation method (MI) overcomes this disadvantage (Little and Rubin 2002). In MI, each missing value is imputed D times. Therefore $\{\tilde{T}^l\}_{l=1}^D$ is a sequence of test statistics which are defined as

$$(2.12) \quad \tilde{T}^l = \sum_{i=1}^K F_X^{-1}(\tilde{p}_i^l) = A + \sum_{j=1}^{K_2} B_j^l, \text{ for } l = 1, \dots, D$$

with

$$(2.13) \quad q_i^l \sim \text{Uniform}(0, \alpha_i) \text{ and } r_i^l \sim \text{Uniform}(\alpha_i, 1).$$

The test statistic is defined as $\bar{T} = \frac{1}{D} \sum_{l=1}^D \tilde{T}^l$, which satisfies,

$$\begin{aligned} \bar{T} &= A + \sum_{j=1}^{K_2} \left[\left(\frac{1}{D} \sum_{l=1}^D F_X^{-1}(q_{K_1+j}^l) \right) \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + \left(\frac{1}{D} \sum_{l=1}^D F_X^{-1}(r_{K_1+j}^l) \right) \cdot \mathbb{1}_{\{p_{K_1+j} \geq \alpha_{K_1+j}\}} \right] \\ &= A + \sum_{j=1}^{K_2} \left[\left(\frac{1}{D} \sum_{l=1}^D W_j^l \right) \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + \left(\frac{1}{D} \sum_{l=1}^D V_j^l \right) \cdot \mathbb{1}_{\{p_{K_1+j} \geq \alpha_{K_1+j}\}} \right] \\ &= A + \sum_{j=1}^{K_2} [\bar{W}_j \cdot \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}} + \bar{V}_j \cdot (1 - \mathbb{1}_{\{p_{K_1+j} < \alpha_{K_1+j}\}})] = A + \sum_{j=1}^{K_2} Z_j. \end{aligned}$$

Since $Z_j = \bar{W}_j$ with probability α_{K_1+j} and $Z_j = \bar{V}_j$ with probability $1 - \alpha_{K_1+j}$, Z_j is a mixture distribution of \bar{W}_j and \bar{V}_j and therefore $\bar{T} - A$ is a mixture distribution of $\{\bar{W}_j, \bar{V}_j, j = 1, \dots, K_2\}$. Note that W_j^l and V_j^l are independent and identically distributed (i.i.d) for fixed j . Denoting by $(\mu_{W_j}, \sigma_{W_j}^2), (\mu_{V_j}, \sigma_{V_j}^2)$ the mean and variance of W_j^l and V_j^l respectively, then by the central limit theorem one concludes that for large enough $D > 0$ it holds

$$\bar{W}_j = \left(\frac{1}{D} \sum_{l=1}^D W_j^l \right) \sim N(\mu_{W_j}, \frac{\sigma_{W_j}^2}{D}), \text{ and } \bar{V}_j = \left(\frac{1}{D} \sum_{l=1}^D V_j^l \right) \sim N(\mu_{V_j}, \frac{\sigma_{V_j}^2}{D}).$$

Then the following theorem holds.

THEOREM 3. For $(j_1, \dots, j_{K_2}) \in \{0, 1\}^{K_2}$, by defining $U(j_1, \dots, j_{K_2}) = \sum_{i=1}^{K_2} (j_i \bar{W}_i + (1 - j_i) \bar{V}_i)$ which satisfies

$$(2.14) \quad U(j_1, \dots, j_{K_2}) \sim N\left[\sum_{i=1}^{K_2} (j_i \mu_{W_i} + (1 - j_i) \mu_{V_i}), \frac{1}{D} \sum_{i=1}^{K_2} (j_i \sigma_{W_i}^2 + (1 - j_i) \sigma_{V_i}^2)\right],$$

then for sufficiently large D , it holds approximately that

$$(2.15) \quad \mathbb{P}(\bar{T} \leq t) = \sum_{(j_1, \dots, j_{K_2}) \in \{0, 1\}^{K_2}} \prod_{i=1}^{K_2} \alpha_i^{j_i} (1 - \alpha_i)^{1 - j_i} \mathbb{P}(A + U(j_1, \dots, j_{K_2}) \leq t).$$

The detailed notations are left to online supplementary file.

Similar to the mean imputation method, the formula can be simplified when some p-value thresholds are equal, i.e.,

$$(2.16) \quad \mathbb{P}(\bar{T} \leq t) = \sum_{j_1=0}^{n_1} \dots \sum_{j_r=0}^{n_r} \prod_{l=1}^r f(j_l; n_l, \beta_l) \mathbb{P}(A + U(j_1, \dots, j_r) \leq t),$$

with $U(j_1, \dots, j_r) = \sum_{l=1}^r (j_l F_X^{-1}(q_l) + (n_l - j_l) F_X^{-1}(r_l))$, $q_l \sim \text{Uniform}(0, \beta_l)$ and $r_l \sim \text{Uniform}(\beta_l, 1)$.

3. Simulation results.

3.1. *Simulated expression profiles.* To evaluate performance of the proposed imputation methods in the genomic setting, we simulated expression profiles with correlated gene structure and variable effect sizes as follows.

Simulate gene correlation structure for $G = 10,000$ genes, $N = 100$ samples in each study, and $K = 10$ studies. In each study, 4,000 of the 10,000 genes belong to $C = 200$ independent clusters..

- Step 1 Randomly sample gene cluster labels of 10,000 genes ($C_g \in \{0, 1, 2, \dots, C\}$ and $1 \leq g \leq G$), such that $C = 200$ clusters each containing 20 genes are generated ($\sum_g \mathbb{1}(C_g = c) = 20, \forall 1 \leq c \leq C = 200$) and the remaining 6,000 genes are unclustered genes ($\sum_g \mathbb{1}(C_g = 0) = 6,000$).
- Step 2 For any cluster c ($1 \leq c \leq C$) in study k ($1 \leq k \leq K$), sample $\Sigma'_{ck} \sim W^{-1}(\Psi, 60)$, where $\Psi = 0.5I_{20 \times 20} + 0.5J_{20 \times 20}$, W^{-1} denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all the entries being 1. Set vector σ_{ck} as the square roots of the diagonal elements in Σ'_{ck} . Calculate Σ_{ck} such that $\sigma_{ck}\Sigma_{ck}\sigma_{ck}^T = \Sigma'_{gk}$.
- Step 3 Denote by $g_1^{(c)}, \dots, g_{20}^{(c)}$ as the indices for genes in cluster c . In other words, $C_{g_j^{(c)}} = c$, where $1 \leq c \leq 200$ and $1 \leq j \leq 20$. Sample expression of clustered genes by $(X'_{g_1^{(c)}nk}, \dots, X'_{g_{20}^{(c)}nk})^T \sim MVN(0, \Sigma_{ck})$, where $1 \leq n \leq N = 100$ and $1 \leq k \leq K = 10$. Sample expression for unclustered genes $X'_{gnk} \sim N(0, 1)$ for $1 \leq n \leq N$ and $1 \leq k \leq K$ if $C_g = 0$.

Simulate differential expression pattern..

- Step 4 Sample effect sizes μ_{gk} from $\text{Unif}(0.1, 0.5)$ for $1 \leq g \leq 1,000$ as DE genes and set $\mu_{gk} = 0$ for $1,001 \leq g \leq G$ as non-DE genes.
- Step 5 For the first 50 control samples, $X_{gnk} = X'_{gnk}$ ($1 \leq g \leq G, 1 \leq n \leq N/2 = 50, 1 \leq k \leq K$). For cases, $Y_{gnk} = X'_{g(n+50)k} + \mu_{gk}$ ($1 \leq g \leq G, 1 \leq n \leq N/2 = 50, 1 \leq k \leq K$).

In the simulated datasets, $K = 10$ studies with $G = 10,000$ genes were simulated. Within each study, there were $\frac{N}{2} = 50$ cases and 50 controls. The first 1,000 genes were DE in all 10 studies with effect sizes randomly simulated **from a uniform distribution on (0.1, 0.5)** respectively, and the remaining 9,000 were non-DE genes. **We chose this effect size range to produce an averaged standardized effect size at $\frac{0.3}{1 \cdot \sqrt{50}} = 0.1414$ so that the DE analysis generates $\sim 500 - 600$ candidate DE genes (Table 1), a commonly seen range in real applications.** In each study, 200 gene clusters existed, each containing 20 genes. The correlation structure within each cluster was simulated from an inverse Wishart distribution

In the simulations, we performed a two sample t-test for each gene in each study and then combined the p-values using the imputation methods proposed in this paper. For simplicity, we viewed the p-values from the last 5 studies as truncated with thresholds $(\alpha_1, \dots, \alpha_5) = (0.001, 0.001, 0.01, 0.01, 0.05)$ respectively. In most genomic meta-analysis, researchers often use conventional permutation analysis by permuting sample labels to compute the p-values to preserve gene correlation structure. However, such a nonparametric approach is not applicable in our situation, since raw data are not available in some studies. In order to control the false discovery rate (FDR), we examined Benjamini-Hochberg (B-H) method (Benjamini and Hochberg, 1995) and Benjamini-Yekutieli (B-Y) method (Benjamini and Yekutieli, 2001) separately. The number of DE genes detected at nominal FDR rate 5% were recorded and the true FDR rates were computed for each meta-analysis method by

$$\text{FDR} = \frac{\sum_g \mathbb{1}(\text{gene } g \text{ detected with } g \geq 1001)}{\#\{\text{genes detected}\}}.$$

In the multiple imputation method, $D = 50$ was selected. Simulations were repeated for 50 times and the mean and standard errors of numbers of DE genes controlled by BH and BY methods and their true FDR are reported in Table 1. The results showed that the FDRs were controlled well for B-H correction but rather conservative for B-Y correction (the true FDR of B-Y is only 1/10 of B-H at nominal $FDR = 5\%$). This is consistent with the previous observation that the B-Y adjustment tends to be over-conservative since it guards against any type of correlation structure (Benjamini and Yekutieli, 2001). As a result, the BH correction will be used for all applications hereafter. The simulation results showed consistently that imputation methods had higher statistical power than the available-case method, and the mean imputation and multiple imputation methods outperform single random imputation method with similar performance. Surprisingly, the ratio of detected DE genes compared to complete case increased from 41.6% in available case (263.5/632.9) to 80.4% in mean imputation (508.6/632.9) using Fisher’s method. The improvement is even more significant using Stouffer’s method (from 41.8% to 86.7%), while at the same time the true FDRs were controlled at similar level for all methods. The result shows that imputation methods successfully utilize the incomplete p-value information to greatly recover the detection power.

TABLE 1
Simulation results for correlated data matrix at nominal FDR=5%

		Fisher		Stouffer	
	Method/Mean(s.e.)	No. DE	True FDR	No. DE	True FDR
BH	Complete cases	632.9(32.5)	0.043(0.0013)	518.6(36.2)	0.046(0.0015)
	available-case	263.5(37.4)	0.048(0.0076)	216.8(35.3)	0.064(0.022)
	Mean imputation	508.6(35.1)	0.046(0.0016)	449.8(36.2)	0.047(0.0022)
	Single imputation	408.9(35.7)	0.043(0.0018)	293.9(32.6)	0.045(0.0027)
	Multiple imputation	509.2(35.0)	0.045(0.0015)	463.8(35.7)	0.050(0.0019)
BY	Complete cases	354.0(34.4)	0.0041(0.00083)	261.7(33.9)	0.0036(0.00097)
	available-case	102.4(21.9)	0.0047(0.0012)	82.8(20.6)	0.0029(0.00096)
	Mean imputation	234.5(32.1)	0.0037(0.00074)	203.8(30.8)	0.0034(0.00073)
	Single imputation	164.0(27.3)	0.0057(0.0014)	113.5(22.3)	0.0039(0.0015)
	Multiple imputation	235.3(32.0)	0.0037(0.00075)	216.1(30.9)	0.0050(0.0010)

We further examined the situation when gene dependence structure does not exist (i.e. Steps 1-3 were skipped and $X'_{gnk} \sim N(0, 1)$). Table 2 shows the true Type I error control under nominal significance level 5% (i.e. True type I error = $\frac{\sum_{g=1,001}^{10,000} \mathbb{1}(\text{gene } g \text{ is detected at significance level } 0.05)}{9,000}$). The result shows adequate type I error control and confirms the validity of the closed form or approximated formula of different imputation methods in Section 2.

TABLE 2
Type I error control for independent data matrix at nominal significance level 5%

	Fisher	Stouffer
Complete cases	0.050(0.00031)	0.050(0.00037)
available-case	0.050(0.00035)	0.050(0.00033)
Mean imputation	0.050(0.00031)	0.050(0.00033)
Single imputation	0.050(0.00032)	0.051(0.00032)
Multiple imputation	0.050(0.00031)	0.051(0.00031)

To investigate the impact of D on the performance of multiple imputation method, simulations were performed for $D \in \{20, 30, 50, 100, 150, 200, 250, 300, 500\}$. The result is shown in Supplement Figure 1 which demonstrates that the performance of multiple imputation method is quite robust

for different number of imputation D . We use $D = 50$ throughout this paper.

3.2. Simulation from complete real datasets. In this subsection, the proposed methods were applied to two real microarray datasets, including 7 prostate cancer studies (Gorlov 2009) and 8 major depressive disorder (MDD) studies (Wang et al., 2012)). The details are summarized in Supplement Table 1. For each dataset, about half of the studies (four for MDD and three for prostate cancer) were randomly selected with p-value truncation threshold 0.05. Five methods including complete data, available-case, single random imputation, mean imputation and multiple imputation methods were applied to the datasets with the simulated incomplete data to impute by Stouffer’s and Fisher’s methods respectively. The generated p-values were corrected by the B-H method and the simulation was repeated for 50 times. Figure 1 shows boxplots of the numbers of differentially expressed (DE) genes at $FDR = 1\%$ for different methods in MDD and $FDR = 0.5\%$ for prostate cancer data. **Figure 1 indicates similar conclusions** that the multiple imputation and the mean imputation methods detect more DE genes than the available-case method and single random imputation method. In the MDD example, very few DE genes (average of 16 and 83 for Fisher and Stouffer respectively) were detected using the available-case method if half of the studies have truncated p-values. The mean and multiple imputation methods greatly improved the detection sensitivity. About 95.2% (Fisher) and 96.3% (Stouffer) of DE genes detected by the mean imputation method overlapped with DE genes detected by complete data analysis in MDD and about 94.7% (Fisher) and 88.1% (Stouffer) of DE genes detected by the mean imputation method overlapped with DE genes detected by complete data analysis in prostate cancer, showing the ability of imputation methods to recover DE gene detection power.

4. Applications.

4.1. Application to colorectal cancer. In the first motivating example, we followed Chan et al. (2007) and attempted to collect 23 colorectal cancer versus normal gene expression profiling studies. Raw data were available in only one study (Bianchini 2006) and 4 of the other 22 studies containing more than 100 DE genes at different p-value thresholds were included in our analysis. We searched the GEO database and identified two additional new studies (Jiang et al. 2008 and Bellot et al. 2012). The seven studies under analysis were summarized in Table 3. After gene-matching, 6,361 genes overlapped in all three studies with raw data. The available-case method, the mean imputation method, the single random imputation method and the multiple imputation method were applied for the seven studies for the Fisher and Stouffer methods respectively and the results were reported in Table 4. For the single random imputation method and multiple imputation method, the analyses were repeated 50 times and the mean and standard error of the number of DE genes detected were reported under FDR control by the BH method. The results demonstrate that for various FDR thresholds, the mean imputation method and the multiple imputation method detected more DE genes than the available-case method and the single random imputation method, which was consistent with previous findings in simulations. Under $FDR = 0.01\%$ control, Fisher and Stouffer mean imputation detected 2.07 (1183/571) and 10.35 (383/37) times of DE genes than those by available-case method, respectively.

4.2. Application to pain research. The second motivating example comes from the meta-analysis of 20 microarray studies of pain to detect the patterns of pain (LaCroix-Fralish, 2011). The original meta-analysis utilized DE gene lists from each study under different threshold criteria from

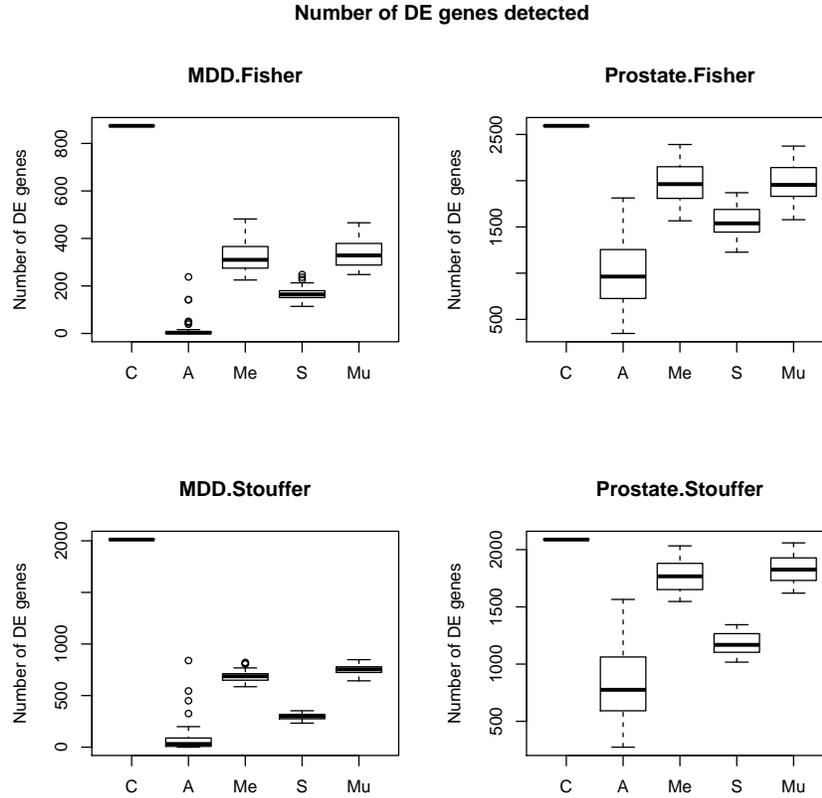


FIG 1. Number of DE genes detected by Fisher's or Stouffer's method. C: complete data; A: available-case; Me: mean-imputation; S: single-imputation; Mu: multiple imputation

p-value, FDR or fold change and identified 79 "statistically significant" genes that appeared in the DE gene lists of four or more studies. The vote counting method essentially lost tremendous amount of information with flawed statistical inference. When we attempted to repeat the meta-analysis, raw data of only one of the 20 studies (*Barr.2005*) could be found. The old platform used in that study, however, contained only 792 genes and had to be excluded from further meta-analysis. In the remaining 19 studies, 11 studies contained DE gene lists under various p-value thresholds (marked bold in Supplement Table 2) and were included in our application. In other words, this example contained exclusively only studies with truncated p-values. Table 5 shows the result of three imputation methods. Fisher and Stouffer identified 280 and 45 genes under 5% FDR control, respectively. Note that the original meta-analysis tested the 79 genes using an overall binomial test and the statistical significance was controlled at an overall p-value level, not at a gene-specific FDR level. As a result, DE gene lists from the new imputation methods are theoretically more powerful and accurate.

To validate the finding, we used the Gene Functional Annotation tool from the DAVID Bioinformatics Resources website (<http://david.abcc.ncifcrf.gov>). DAVID applied a modified Fisher's exact test to evaluate the association between the DE gene lists and pathways. Functional annotation of the 280 DE genes from the Fisher's mean imputation method identified 208 pathways at FDR= 5%,

TABLE 3
Seven colorectal cancer versus normal tissue expression profiling studies included in analysis

Study	No. of samples	No. of genes	Raw data availability	No. of DE genes	No. of overlapped DE genes	p-value threshold
<i>Bianchini_2006</i>	24	7403	GSE3294	-	-	-
<i>Bellot_2012</i>	17	18191	GSE24993	-	-	-
<i>Jiang_2008</i>	48	18197	GSE10950	-	-	-
<i>Grade_2007</i>	103	21543	-	1950	635	1e-7
<i>Croner_2005</i>	33	22283	-	130	47	0.006
<i>Kim_2004</i>	32	18861	-	448	143	0.001
<i>Bertucci_2004</i>	50	8074	-	245	97	0.009

TABLE 4
Summary of results for colorectal cancer

FDR	Fisher				Stouffer			
	Available	Mean	Single	Multiple	Available	Mean	Single	Multiple
1%	2587	2855	2172.4(2.90)	2785.4(2.93)	1318	1675	668.4(3.96)	1616.0(2.10)
0.1%	1472	1874	1265.6(2.34)	1805.7(1.50)	299	709	252.7(1.93)	680.5(1.12)
0.01%	571	1183	748.4(1.89)	1138.6(2.00)	37	383	102.5(1.65)	366.7(0.69)

among which selected important pain-related pathways were grouped into five major biological categories and displayed in Table 6. In contrast, the 79 genes from vote counting identified only 14 pathways, of which the expected pain-related pathways under the categories of inflammation and of differentiation, development and projection are missing (see Table 6). The pathway enrichment q-values after multiple comparison control of the "280 gene list" were very significant, while those of the "79 gene list" were not. Since the p-value calculation from Fisher's exact test can be impacted by the DE gene size, we further compared the enrichment odd-ratios of genes in the pathway versus in the DE gene list. Still the enrichment odds-ratios of the "280 gene list" were generally much higher than those for the "79 gene list", showing stronger pain functional association from the Fisher's mean imputation method.

TABLE 5
Summary of results for patterns of pain

	Fisher	Stouffer
Mean	280	45
Single	57.04 (1.6228)	16.44(0.8605)
Multiple	280.36(0.8105)	77.56(0.6616)

TABLE 6
Summary of pathway analysis by DAVID

Category	Pathway ID	280 DE (Fisher's mean imputation)			79 DE (Vote counting)		
		pval	qval	odds ratio	pval	qval	odds ratio
Differentiation, development and projection	GO : 0030182 ~ neuron differentiation	5.6e-6	0.0006	3.1	0.26	0.95	1.6
	GO : 0045664 ~ regulation of neuron differentiation	1.6e-5	0.0011	4.7	0.37	0.98	1.9
	GO : 0048666 ~ neuron development	2.5e-6	0.0003	3.6	0.24	0.94	1.7
	GO : 0051960 ~ regulation of nervous system development	6.5e-6	0.0006	4.2	0.29	0.96	1.9
	GO : 0031175 ~ neuron projection development	1.6e-5	0.0012	3.7	0.27	0.96	1.8
	GO : 0042995 ~ cell projection	3.6e-11	3.2e-9	3.5	0.033	0.47	1.9
	GO : 0043005 ~ neuron projection	3.0e-11	3.4e-9	4.3	0.043	0.51	2.0
	GO : 0030030 ~ cell projection organization	1.6e-5	0.0012	3.3	0.24	0.94	1.7
Response to stimuli	GO : 0009611 ~ response to wounding	3.8e-10	2.8e-7	4.3	2.7e-5	0.016	3.6
	GO : 0009719 ~ response to endogenous stimulus	3.2e-8	1.7e-5	3.4	0.35	0.97	1.3
	GO : 0048584 ~ positive regulation of response to stimulus	7.9e-8	2.5e-5	4.9	0.0049	0.34	3.6
	GO : 0032101 ~ regulation of response to external stimulus	1.1e-5	0.001	4.8	0.043	0.71	2.8
Immune	GO : 0050778 ~ positive regulation of immune response	4.2e-7	7.6e-5	5.9	0.018	0.57	4.0
	GO : 0002684 ~ positive regulation of immune system process	1.9e-6	0.0003	4.4	0.0009	0.13	4.2
	GO : 0006956 ~ complement activation	3.0e-5	0.0016	11.5	0.011	0.46	8.4
	GO : 0002478 ~ antigen processing and presentation of exogenous peptide antigen	1.3e-6	0.00022	19.0	0.00098	0.12	10.64
Inflammation	GO : 0002673 ~ regulation of acute inflammatory response	1.4e-6	0.0002	14.1	0.19	0.93	3.8
	GO : 0002526 ~ acute inflammatory response	7.1e-06	0.0007	6.7	0.012	0.48	4.4
	GO : 0050727 ~ regulation of inflammatory response	1.9e-5	0.0012	6.9	0.17	0.92	2.8
	GO : 0006954 ~ inflammatory response	1.5e-5	0.0012	4.1	0.001	0.11	3.8
Regulation of Transmission	GO : 0051969 ~ regulation of transmission of nerve impulse	6.0e-6	0.0006	4.8	0.057	0.80	2.4

4.3. *Application to a three-way association method (liquid association).* So far the proposed imputation methods were applied successfully to two real microarray datasets of colorectal cancer and pain research in which the actual p-values of some genes were not reported in a subset of studies. In this section we show that the proposed imputation methods can be useful in the meta-analysis of "big data" such as GWAS or eQTL, where the main computational problem is often the data storage.

In the literature, it has long been argued that positively correlated expression profiles are likely to encode functionally related proteins. Liquid association (LA) analysis (Li 2002) is an advanced three-way co-expression analysis beyond the traditional pairwise correlations. For any triplet of genes X, Y and Z , the LA score $LA(X, Y|Z)$ measures the effect that expression of Z to control on and off of the co-expression between X and Y . For example, high expression of Z turns on positive correlation between X and Y while when expression of Z is low, X and Y are negatively or non-correlated. Theory in Li (2002) simplified calculation of the LA score to a linear order of sample size and made the genome-wide computation barely feasible. Suppose we want to combine K studies of the liquid association, liquid association p-values of all triplets in all $K = 10$ studies have to be stored for meta-analysis. When the number of genes $G = 1,000$, the number of p-values to be stored is $G \cdot C_2^{G-1} \cdot K = 4.985GB$. For a reasonable $G = 20,000$ genome-wide analysis, storage size for all p-values quickly increases to $39.994TB$. One may argue that univariate (i.e. triplet by triplet) meta-analysis may be applied repeatedly to avoid the need of storing all p-value results. There are many other genomic meta-analysis situations when this may not be feasible. For example, in GWAS meta-analysis under a consortium collaboration, raw genotyping data cannot be shared for privacy reasons and only the derived statistics or p-values can be transferred for meta-analysis. Below we describe how imputation methods can help circumvent the tremendous data storage problem.

We performed a small scale of analysis on 566 DE genes previously reported from the meta-analysis of the eight MDD studies used in Section 3.2 (Wang et al., 2012). The total number of possible triplets $(X, Y|Z)$ was 90,180,780. By setting up p-value threshold at 0.001, we only needed to store exact p-values for 2,094,123 ($\sim 2.32\%$) triplets and the remaining were truncated as considered in this paper. Since we also needed to store the truncation index information, we only needed to store $2 \times 2.32\% = 4.64\%$ of the information and the compression ratio was 95.36%. To investigate the loss of information by the truncation, Figure 2 shows meta-analysis p-values (at $-\log(p)$ scale) from Fisher's method using full data and Fisher mean imputation method using truncated data. The result shows high concordance in the top significant triplets, which are the major targets of this exploratory analysis. Among the top 1000 triplets detected by Fisher's method using complete p-value information, 83.7% of them were also identified by the top 1000 by Fisher mean imputation. The remaining 163 triplets were still in top ranks (rank between 1199 and 4763) using truncated data in the result of Fisher mean imputation. This result suggests good potential of applying data truncation to preserve the most informative information and performing imputation to approximate the finding of the top targets when meta-analysis of "big data" is needed. The compression ratio may further increase by a more stringent truncation threshold but the performance may somewhat decline as a trade-off.

5. Discussion and conclusion. When combining multiple genomic studies by p-value combination methods, the raw data are often not available and only the ranges of p-values are reported for some studies in genomic applications. This is especially true for microarray meta-analysis since owners of many microarray studies tend not to publish their data in the public domain. This

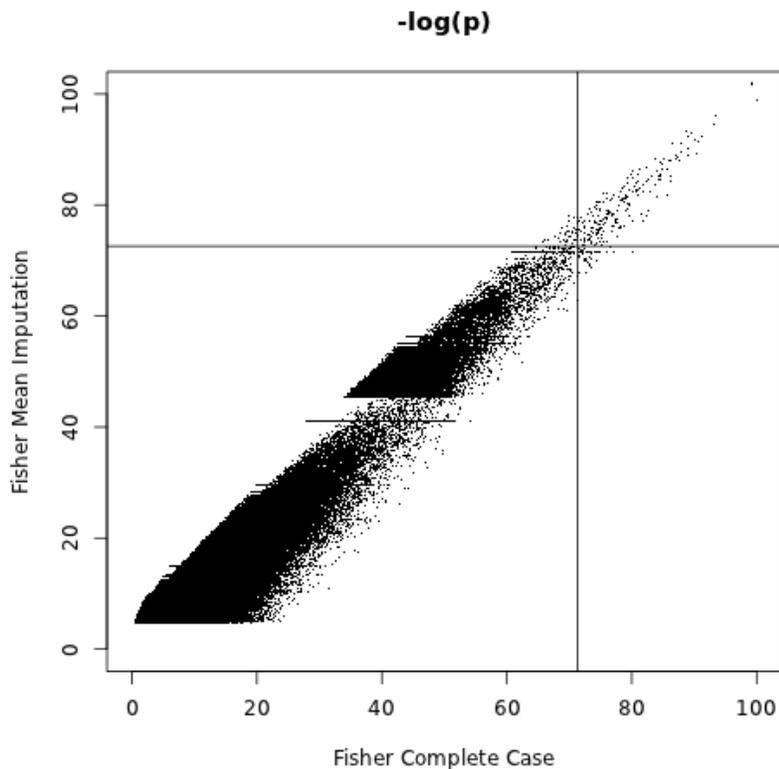


FIG 2. $-\log(p)$ comparison of the mean imputation method using truncated data with the complete case method using complete data. Vertical line: $x = 71.3$. Horizontal line: $y = 72.58$. Points right to vertical line are top 1,000 triplets detected by Fisher's complete case method, and points above to horizontal line are top 1,000 triplets detected by Fisher's mean imputation method

incomplete data issue is often encountered when one attempts to perform a large-scale microarray meta-analysis. If raw data are not available, two naïve methods - vote counting method and available-case method - are commonly used. Since these two methods completely or largely neglect the information contained in the truncated p-values, and their statistical power is generally low. In this paper, we proposed three imputation methods for a general class of evidence aggregation meta-analysis methods to combine independent studies with truncated p-values: mean imputation, single random imputation and multiple imputation methods. For each proposed imputation method, the null distribution was derived analytically for the Fisher and Stouffer methods. Theoretical results showed that the test statistics from the single random imputation and the multiple imputation methods were unbiased, while those for mean imputation method were biased. Simulations were performed for the imputed Fisher method and imputed Stouffer method. The simulation results showed that type I errors were well-controlled for all methods, which was consistent with our theoretical derivation. Compared to the naïve available-case method, all the imputation methods achieved higher statistical powers, and the mean imputation and the multiple imputation methods recovered much of the power that the complete cases method achieved even when half of the studies had truncated p-values. Furthermore, Supplementary Figure 1 showed that the power of the multiple imputation method was robust to the number of imputation D . Although small to moderate

D provided good results, we recommend choosing D being larger than 50 to guarantee that central limit theorem can approximate well. Applications to two motivating examples in colorectal cancer and pain conditions showed that both mean imputation and multiple imputation performed among the best in terms of detection sensitivity and biological validation by pathway analysis.

In regression-type missing-data imputation methods, the null distribution of the error term is unknown and is assumed to be normally distributed with equal variance, a setting in which multiple imputation method usually outperforms mean imputation in practice and in theory (Little and Rubin 2002), particularly because mean imputation underestimates the true variance. However, our simulation results demonstrated that the power of the two methods were quite similar. Two reasons may contribute to this result. First, although the test statistic from the mean imputation method is biased and neglects the variation of truncated p-values, its p-value can be computed accurately when the null distribution is derived analytically. Second and more importantly, we find that the test statistic of mean imputation is in fact $F_X^{-1}(\mathbb{E}(p))$, while for sufficiently large D , the test statistic of multiple imputation converges to $\mathbb{E}(F_X^{-1}(p))$ in distribution. It is easy to show that these two quantities are very close to each other for a small range of p , provided $F_X^{-1}(\cdot)$ is smooth. Since $F_X^{-1}(\cdot)$ is infinitely differentiable for the Fisher and Stouffer methods, and the small p-value range in $(0, \alpha)$ are particularly of interest to us, it is not surprising that the mean imputation method and multiple imputation method perform similarly. Since the mean imputation method achieved almost the same power as the multiple imputation method with less computational complexity, it is more appealing and is recommended for microarray meta-analysis, where the imputed meta-analysis method is performed repeatedly for thousands of genes. In this paper only the evidence aggregation meta-analysis methods are investigated and further work will be needed to extended these results to order statistic based methods such as minP and maxP.

Note that although the truncated p-value issue discussed in this paper may appear similar to the problem of "publication bias", it is fundamentally different. Publication bias refers to the fact that a study with a large positive treatment effect is more likely to be published than a study with a relatively small treatment effect, resulting in bias if one only considers published studies. Denote by p_1, p_2, \dots, p_N the p-values of all conducted studies that should have been collected. Only a subset of likely more significant p-values p_1, p_2, \dots, p_n are observed. Under this setting, N is unknown and p_{n+1}, \dots, p_N are unknown as well. Since the number of missing publications is unknown, Duval and Tweedie proposed the "Trim and Fill" method to identify and correct for funnel plot asymmetry arising from publication bias (Duval and Tweedie, 2000a and 2000b), in which an estimate of the number of missing studies is provided and an adjusted treatment effect is estimated by performing a meta-analysis including the imputed studies. For the truncated p-value problem we consider here, the total number of studies, the number of studies with truncated p-values and the p-value truncation thresholds are all known. Therefore, investigation of the imputation of truncated p-values in meta-analysis is different from the traditional "publication bias" problem and has not been studied in the meta-analysis literature, to the best of our knowledge.

In this paper, the methods we developed mainly target on microarray meta-analysis but the issue can happen frequently in other types of genomic meta-analysis (e.g. GWAS; Begurn et. al. 2012). In Section 4.3, we demonstrated an unconventional application of our methods to meta-analysis of liquid association. Due to the large number of triplets tested in the three-way association, the needed p-value storage is huge. By preserving only the most informative data by truncation, the

storage burden is greatly alleviated and our imputation methods help approximate and recover the top meta-analysis targets with little power loss. In an on-going project, we also attempt to combine multiple genome-wide eQTL results via meta-analysis. In eQTL, regression analysis is used to investigate the association of a SNP genotyping and a gene expression. It is impractical to store all genome-wide eQTL p-values as the storage space required is too large (25,000 genes \times 2,000,000 SNPs = 5×10^{10} p-values). A practical solution is to record only the eQTL p-values smaller than a threshold (say 10^{-4}) for meta-analysis, which leads to the same statistical setting as discussed in this paper. In another project, we combine results from multiple ChIP-seq peak calling algorithms to develop a meta-caller. Since each peak caller algorithm can only report the top peaks with p-values smaller than a certain p-value threshold, we again encounter the same truncated p-value problem in meta-analysis. As more and more complex genomic data are generated and the need for meta-analysis increases, we expect the imputation methods we propose in this paper will find even more applications in the future.

Acknowledgement. This study was supported by NIH R21MH094862. The authors would like to thank S.C. Morton for discussion. **We wish to express our sincere thanks to the associate editor and two reviewers for their valuable comments to significantly improve this paper.**

References.

- [1] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**: 289-300.
- [2] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, **29**: 1165-1188.
- [3] Berger R.L. (1982). Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, **24**: 295-300.
- [4] Bertucci F., Salas S., Eysteries S., et al. (2004). Gene expression profiling of colon cancer by DNA microarrays and correlation with histoclinical parameters. *Oncogene*, **23**: 1377-91.
- [5] Bianchini M., Levy E., Zucchini C., et al. (2006). Comparative study of gene expression by cDNA microarray in human colorectal cancer tissues and normal mucosa. *Int J Oncol*, **29**: 83-94.
- [6] Birnbaum A. (1954). Combining independent tests of significance. *Journal of the American Statistical Association*, **49**: 559-574.
- [7] Birnbaum A. (1955). Characterizations of complete classes of tests of some multiparametric hypothesis, with applications to likelihood ratio tests. *ANN. Math. Statist.*, **26**: 21-36.
- [8] Bellot G.L., Tan W.H., Tay L.L., Koh D. et al. (2012). Reliability of tumor primary cultures as a model for drug response prediction: expression profiles comparison of tissues versus primary cultures from colorectal cancer patients. *J Cancer Res Clin Oncol*, **138(3)**: 463-482.
- [9] Borovecki F. et al (2005). Genome-wide expression profiling of human blood reveals biomarkers for huntingtons disease. *Proceedings of the National Academy of Sciences*, **102**: 11023-11028.
- [10] Cardoso J. et al. (2007). Expression and genomic profiling of colorectal cancer. *Biochimica et Biophysica Acta-Reviews on Cancer*, **1775**: 103-137.
- [11] Chan S.K., Griffith O.L., Tai I.T. and Jones S.J.M. (2008). Meta-analysis of colorectal cancer gene expression profiling studies identifies consistently reported candidate biomarkers. *Cancer Epidemiol Biomarkers Prev*, **17(3)**: 543-552.
- [12] Chang L.C., Lin H.M., Sibille E. and Tseng G.C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, **14**: 368.
- [13] Choi J.K. et al. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**: 84-90.
- [14] Choi H. et al. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, **8**: 364-383.
- [15] Cooper H.M. and Hedges L.V. (Eds.) (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.
- [16] Croner RS, Foertsch T, Brueckl WM, et al. (2005). Common denominator genes that distinguish colorectal carcinoma from normal mucosa. *Int J Colorectal Dis*, **20**: 353-362.

- [17] Duval S. and Tweedie R.L. (2000a). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, **56(2)**: 455-463.
- [18] Duval S. and Tweedie R.L. (2000b). A nonparametric "Trim and Fill" method of accounting for publication bias in meta-analysis. *JASA*, **95(1)**: 89-98.
- [19] Fisher R.A. (1932). *Statistical methods for research workers*. Edinburgh, Oliver & Boyd, 4th edition.
- [20] Fleiss J.L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260), New York: Russell Sage Foundation.
- [21] Folks J.L. (1984). Combination of independent tests. In *Handbook of statistics 4. Nonparametric methods*, P.R. Krishnaiah and P.K. Sen (eds):New York, North-Holland
- [22] Gorlov I.P. et al. (2009). Candidate pathways and genes for prostate cancer: a meta-analysis of gene expression data. *BMC Medical Genomics*, **2**:48.
- [23] Grade M., Hormann P., Becker S., et al. (2007). Gene expression profiling reveals a massive, aneuploidy-dependent transcriptional deregulation and distinct differences between lymph node-negative and lymph node-positive colon carcinomas. *Cancer Res*, **67**:41-56.
- [24] Griffith O.L., Jones S.J.M. and Wiseman S.M. (2006). Meta-analysis and meta-review of thyroid cancer gene expression profiling studies identifies important diagnostic biomarkers. *J. Clin. Oncol.*, **24**:5043-51.
- [25] Hedges L.V. and Olkin I. (1980). *Vote-counting methods in research synthesis*. *Psychological Bulletin*, **88**:359.
- [26] Hedges L.V. and Olkin I. (1985). *Statistical methods for meta-analysis*. Academic Press Inc.: Orlando, Florida.
- [27] Hedges L.V. and Vevea J.L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, **3**:486-504.
- [28] Hedges L.V. (2007). Meta-analysis. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp.919-953), New York: Russell Sage Foundation.
- [29] Hunter J.E. and Schmidt F.L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, **8**:275-292.
- [30] Ioannidis J.P.A., Allison D.B., Ball C.A. and et. al. (2009). Repeatability of published microarray gene expression analysis. *Nature Genetics*, **41**:149-155.
- [31] Jiang X., Tan J., Li J., Kivimäe S. et al. (2008). DACT3 is an epigenetic regulator of Wnt/beta-catenin signaling in colorectal cancer and is a therapeutic target of histone modifications. *Cancer Cell*, **13(6)**:529-41.
- [32] Kim H., Nam S.W., Rhee H., et al. (2004). Different gene expression profiles between microsatellite instability-high and microsatellite stable colorectal carcinomas. *Oncogene*, **23**: 6219-25.
- [33] Kwon H.C., Kim S.H., Roh M.S., et al. (2004). Gene expression profiling in lymph node-positive and lymph node-negative colorectal cancer. *Dis Colon Rectum*, **47**: 141-152.
- [34] LaCroix-Fralish M.L. et. al.. (2011). Patterns of pain: Meta-analysis of microarray studies of pain. *Pain*, **152**: 1888-1898.
- [35] Lancaster H. (1961). The combination of probabilities: an application of orthonormal functions. *Australian Journal of Statistics*, **3**: 20-33.
- [36] Littell C.L. and Floks J.L.. (1971). Asymptotic Optimality of Fisher's Method of Combining Independent Tests. *Journal of the American Statistical Association*, **66(336)**: 802-805.
- [37] Littell C.L. and Floks J.L.. (1973). Asymptotic Optimality of Fisher's Method of Combining Independent Tests II. *Journal of the American Statistical Association*, **68(341)**: 193-194.
- [38] Lau J., Antman E.M. and Jimenez-Silva J. et. al. (1992). Cumulative meta-analysis of therapeutic trials for Myocardial infarction. *The new England Journal of Medicine*, **327**:248-254.
- [39] McCarley R.W., Wible C.G., Frumin M., Hirayasu Y., Levitt J.J., Shenton M.E. (2001). Why vote-count reviews don't count [letter to the editor]. *Biological Psychiatry*, **49**: 161-163.
- [40] Li J. and Tseng G.C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Annals of Applied Statistics*, **5(2A)**:994-1019.
- [41] Littell R.C. and Folks J.L.(1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, **66**: 802-806.
- [42] Littell R.C. and Folks J.L. (1973). Asymptotic optimality of Fisher's method of combining independent tests ii. *Journal of the American Statistical Association*, **68**: 193-194.
- [43] Moreau Y. et al. (2003). Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends in Genetics*, **19**: 570-577.
- [44] Olkin I. and Saner H. (2001). Approximations for trimmed Fisher procedures in research synthesis. *Statist. Methods Med. Res.*, **10**: 267-276.
- [45] Owen A.B. (2009). Karl pearson's meta-analysis revisited. *Annals of Statistics*, **37**: 3867-3892.
- [46] Pirooznia M., Nagarajan V. and Deng Y. (2007). Gene venn - a web application for comparing gene lists using venn diagram. *Bioinformatics*, **1**: 420-422.

- [47] Rhodes D. et al. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer research*, **62**: 4427-4433.
- [48] Rosenthal R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231-244), New York: Russell Sage Foundation.
- [49] Roy S.N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, **24(2)**: 220-238.
- [50] Little R. and Rubin D. (2002). *Statistical analysis with missing data, second edition*. John Wiley & Sons, Inc.: Hoboken, New Jersey.
- [51] Segal E. et al. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, **3**: 1090-1098.
- [52] Song C. and Tseng G.C. (2014). Order statistic for robust genomic meta-analysis. *Annals of Applied Statistics*. Accepted.
- [53] Stouffer S. et al. (1949). *The American soldier, volume I: adjustment during army life*. Princeton University press.
- [54] Sterne J. (editor) (2009). *Meta-analysis in Stata: an updated collection from the Stata Journal*. Stata press.
- [55] Tippett L.H.C. (1931). *The methods in statistics*. Williams and Norgate, LTD., 1st edition.
- [56] Tseng G.C. et al. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Research*, **40(9)**:3785-99.
- [57] Wang X., Lin Y., Song C., Sibille E. and Tseng G.C. (2012). Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder. *BMC Bioinformatics*, **13**: 52.
- [58] Wilkinson B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, **48**: 156-157.