BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 10 2010, pages 1316–1323
doi:10.1093/bioinformatics/btq148

Gene expression

Advance Access publication April 21, 2010

# Meta-analysis for pathway enrichment analysis when combining multiple genomic studies

Kui Shen[1] and George C. Tseng[1,2,3,*]

[1]Department of Computational Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA 15213, [2]Department of Biostatistics and [3]Department of Human Genetics, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA 15261, USA

Associate Editor: Martin Bishop

## ABSTRACT

**Motivation:** Many pathway analysis (or gene set enrichment analysis) methods have been developed to identify enriched pathways under different biological states within a genomic study. As more and more microarray datasets accumulate, meta-analysis methods have also been developed to integrate information among multiple studies. Currently, most meta-analysis methods for combining genomic studies focus on biomarker detection and meta-analysis for pathway analysis has not been systematically pursued.

**Results:** We investigated two approaches of meta-analysis for pathway enrichment (MAPE) by combining statistical significance across studies at the gene level (MAPE_G) or at the pathway level (MAPE_P). Simulation results showed increased statistical power of meta-analysis approaches compared to a single study analysis and showed complementary advantages of MAPE_G and MAPE_P under different scenarios. We also developed an integrated method (MAPE_I) that incorporates advantages of both approaches. Comprehensive simulations and applications to real data on drug response of breast cancer cell lines and lung cancer tissues were evaluated to compare the performance of three MAPE variations. MAPE_P has the advantage of not requiring gene matching across studies. When MAPE_G and MAPE_P show complementary advantages, the hybrid version of MAPE_I is generally recommended.

**Availability:** http://www.biostat.pitt.edu/bioinfo/

**Contact:** ctseng@pitt.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on 16 November 2009; revised on 13 March 2010; accepted on 5 April 2010

## 1 INTRODUCTION

Microarray technology provides the ability to detect genome-wide gene expression activities with thousands of probes printed on each high-density chip. It has evolved rapidly in the past decade and has gradually become a standard tool for many biomedical studies. The high-throughput nature of the technology requires development of suitable statistical and bioinformatic methods to analyze the data. Pathway analysis (a.k.a gene set analysis) was developed to correlate the identified gene list from microarray data with a priori defined gene sets usually from biological pathway databases. As shown in Figure 1A, pathway analysis usually has three main steps. The first step is to calculate the association of each gene's expression pattern with phenotype, which is often evaluated by $t$-statistics or correlations. The second step is to map the genes into a predefined pathway database and compute the enrichment evidence score of a pathway, which summarizes association scores of genes in the pathway. Finally, the $P$-value and $Q$-value of each pathway are calculated based on an appropriate permutation test. The first naïve way for pathway analysis is to compare the number of differentially expressed (DE) genes in and outside a pathway using Fisher's exact test (Hosack *et al.*, 2003). This method was improved by gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005), an innovative method that consider the gene order and is based on association strength with phenotype instead of 0–1 indicator of being a DE gene or not. Despite the popularity of GSEA, later studies have reported its low statistical power. Tian (Tian *et al.*, 2005) and Geoman (Goeman and Buhlmann, 2007) discussed the statistical framework of pathway analysis and Tian proposed two powerful procedures: one permuting genes and the other permuting samples. Newton *et al.* (2007) introduced a random set approach. Efron and Tibshirani (Efron and Tibshirani, 2007) further improved GSEA by introducing a max–mean procedure and a re-standardization procedure.

The wide applications of microarray technology have led to an explosion of gene expression profiling studies publicly available. However, the noisy nature of microarray data, together with the relatively small sample size in each study, often results in inconsistent biological conclusions (Ein-Dor *et al.*, 2005; Tan *et al.*, 2003). Therefore, methods for synthesizing multiple microarray studies are in tremendous need. Meta-analysis, a set of statistical techniques to combine results from several studies, has been recently applied to microarray analysis to increase the reliability and robustness of results from individual studies. In traditional statistics literature, Fisher's statistic (Mosteller and Fisher, 1948) (sum of log-transformed $P$-values), minimum $P$-value statistic (Tippett, 1931) and maximum $P$-value statistic (Wilkinson, 1951) have been proposed and compared. For combining microarray studies, current 'meta-analysis' in the biological literature contains a widespread use of naive intersection/union operations or simple counting of appearances in the DE gene lists obtained from individual studies under certain criteria (e.g. false discovery rate = 0.05) (Borovecki *et al.*, 2005; Cardoso *et al.*, 2007; Pirooznia *et al.*, 2007; Segal *et al.*, 2004) and many more. One can quickly note that intersections are

[*]To whom correspondence should be addressed.

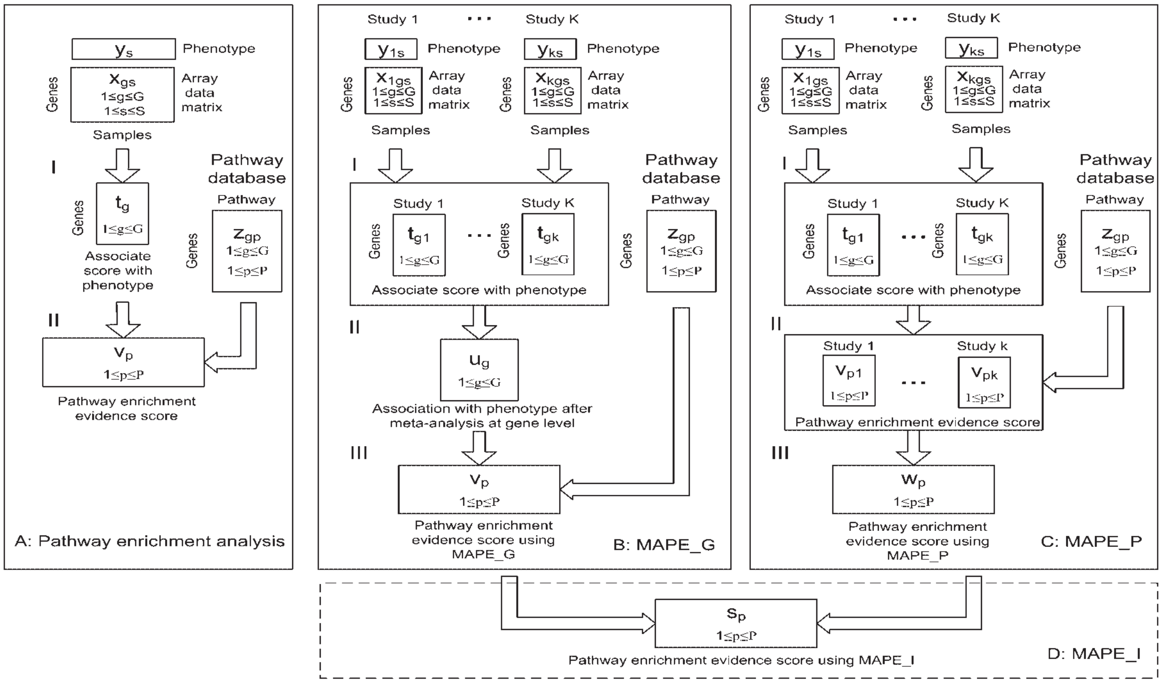
y_s Phenotype
Genes
x_gs 1≤g≤G 1≤s≤S
Array data matrix
Samples
I
Pathway database
Pathway
t_g 1≤g≤G
Associate score with phenotype
z_gp 1≤g≤G 1≤p≤P
II
v_p 1≤p≤P
Pathway enrichment evidence score
A: Pathway enrichment analysis
Study 1 ... Study K
y_1s Phenotype
y_ks Phenotype
x_1gs 1≤g≤G 1≤s≤S
x_kgs 1≤g≤G 1≤s≤S
t_g1 1≤g≤G
t_gk 1≤g≤G
Associate score with phenotype
u_g 1≤g≤G
Association with phenotype after meta-analysis at gene level
III
v_p 1≤p≤P
Pathway enrichment evidence score using MAPE_G
B: MAPE_G
Study 1
Study k
v_p1 1≤p≤P
v_pk 1≤p≤P
Pathway enrichment evidence score
w_p 1≤p≤P
Pathway enrichment evidence score using MAPE_P
C: MAPE_P
s_p 1≤p≤P
Pathway enrichment evidence score using MAPE_I
D: MAPE_I

**Fig. 1.** Diagram for MAPE analysis. (**A**) Pathway enrichment analysis for an individual study; (**B**) MAPE_G; (**C**) MAPE_P; (**D**): MAPE_I.

often too conservative and unions are anticonservative, especially when the number of studies increases. Rhode *et al.* (2002) was the first to apply Fisher's method to microarray data for a real sense of meta-analysis. Many other approaches have been proposed since then, including random effects model (Choi *et al.*, 2003), latent variable approach (Choi *et al.*, 2007) and Bayesian approaches (Shen *et al.*, 2004). Currently, meta-analysis methods for microarray studies are mostly aimed at combining different studies to identify DE genes, an analysis at the gene level. However, DE gene analysis has two main shortcomings. First, the identified DE genes often do not directly relate to the phenotype of interest. Second, a gene set from an important pathway may act in concert with moderate activities, which cannot be detected by DE gene analysis, while the pathway may have important biological effects on the phenotype of concern. Pathway analysis has an inherent advantage for work with meta-analysis. It is well known that the lists of DE genes from independent studies associated with the same phenotype often have little overlap (Ein-Dor *et al.*, 2005; Tan *et al.*, 2003), while pathway analysis often generates improved consistency (Manoli *et al.*, 2006). This situation motivates us to develop systematic approaches of meta-analysis for pathway enrichment (MAPE), which provides a more robust and powerful tool than standard gene set analysis. In this article, we investigate two natural approaches of MAPE by combining statistical significance across studies at the gene level (MAPE_G) before pathway analysis or at the pathway level (MAPE_P) after pathway analysis (Fig. 1B and C). Simulation results and applications to real datasets show their complementary advantages under different scenarios. We, thus, introduce an integrated method (MAPE_I) that incorporates the advantages of both MAPE_G and MAPE_P. MAPE_G and MAPE_P have been implicitly or explicitly used in a few publications (Setlur *et al.*, 2007; Thomassen *et al.*, 2008) while a systematic methodology and a rigorous evaluation have not been investigated. To our knowledge, this is the first article to systematically develop and evaluate meta-analysis for pathway analysis in microarray studies.

## 2 METHODS

### 2.1 Framework of MAPE_G and MAPE_P

Figure 1 presents the general framework of pathway enrichment analysis for single study (Fig. 1A) and three variations of MAPE methods proposed in this article (Fig. 1B–D). MAPE_G combines $P$-value information from multiple studies at the gene level in Step II of Figure 1B. In contrast, MAPE_P integrates $P$-values at the pathway level in Step III of Figure 1C. Since MAPE_G and MAPE_P have complementary advantages under different scenarios that will be demonstrated by extensive simulations in a later section, a hybrid form of MAPE_I is proposed in Figure 1D to incorporate merits of both methods (Fig. 1D).

Mathematical notations and data structure of Figure 1 are described below. Figure 1A describes framework for pathway analysis in a single study. Suppose a data matrix $\{x_{gs}\}$ $(1 \le g \le G, 1 \le s \le S)$ represents the gene expression intensity of gene $g$ and sample $s$. A binary phenotype label is available for each sample: $\{y_s\}$ $(1 \le s \le S)$ and $y_s \in \{0,1\}$ (e.g. 0 represents normal patients and 1 represents tumor patients). A pathway database matrix $\{z_{gp}\}$ $(1 \le g \le G, 1 \le p \le P)$ represents the pathway information of $P$ pathways where $z_{gp} = 1$ when gene $g$ belongs to pathway $p$ and $z_{gp} = 0$ otherwise. In Step I of Figure 1A, the association score with phenotype in each gene $g$ is first calculated as $t_g$ (usually by $t$-statistics or a correlation measure). In Step II, a pathway enrichment evidence score $v_p$ is calculated for each pathway $p$ [e.g. Kolmogorov—Smirnov (KS) statistics or mean of $t$-statistics]. A gene-wise and/or sample-wise permutation test is then performed to assess the statistical significance of $v_p$. Normally $Q$-values $[q(v_p)]$ are evaluated and the false discovery rate is controlled at 5% (meaning among detected pathways, on average 5% are false discoveries) or sometimes at relaxed 10 or 15%. In final conclusion, all pathways with a $Q$-value <5% are reported as enriched pathways [i.e. $\{p: q(v_p) \leq 5\%\}$].

When combining multiple studies, we assume genes in multiple studies are matched and no missing value exists. In Figure 1B and C, denote by $\{x_{kgs}\}$ $(1 \leq k \leq K,\ 1 \leq g \leq G,\ 1 \leq s \leq S_k)$ the expression intensity of gene $g$ and sample $s$ in study $k$. $\{y_{ks}\}$ $(1 \leq k \leq K,\ 1 \leq s \leq S_k)$ and $y_{ks} \in \{0,1\}$ represents the phenotype label for sample $s$ in study $k$. Figure 1B shows the procedure for the MAPE_G method. In Step I, the association scores with phenotype are calculated in each study [i.e. $\{t_{gk}\}$, $1 \leq g \leq G, 1 \leq k \leq K$]. In Step II, meta-analysis is performed for biomarker detection and produces a new association score after meta-analysis at the gene level [i.e. $\{u_g\}$, $1 \leq g \leq G$]. In Step III, the pathway enrichment analysis is performed as in Step II in Figure 1A. The evidence scores $\{v_p\}$, corresponding $Q$-values $\{q(v_p)\}$ and a list of enriched pathways are then generated. This method can be viewed as a natural combination of meta-analysis for biomarker detection (Steps I and II) and pathway enrichment analysis (Step III) in a sequential manner. Rhodes (Rhodes *et al.*, 2002) has implicitly performed similar analysis by first querying DE genes from meta-analysis and then assess pathway enrichment using the KEGG database (Kanehisa and Goto, 2000). For MAPE_G proposed in this article, we replace the *ad hoc* two-stage procedure with a unified evaluation by permutation test.

In Figure 1C, the framework for MAPE_P is shown. The Step I of association scores for each study is identical to that in MAPE_G. In Step II, instead of performing meta-analysis at the gene level, we performed pathway enrichment analysis in each individual study to obtain the study-wise pathway enrichment evidence scores: $\{v_{pk}\}$ $(1 \leq k \leq K, 1 \leq p \leq P)$. The meta-analysis on the pathway level was then performed in Step III to derive the combined evidence score $w_p$. Finally, the corresponding $Q$-values, $q(w_p)$, are assessed by permutation test.

## 2.2 Example pathways showing complementary advantages of MAPE_G versus MAPE_P

It is easily seen that MAPE_P has an important advantage in that the genes across multiple studies need not be matched to perform meta-analysis as needed in MAPE_G (Step II of Fig. 1B). Specifically, we can relax the data setting in Figure 1C to $\{x_{kgs}\}$ $(1 \leq k \leq K,\ 1 \leq g \leq G_k,\ 1 \leq s \leq S_k)$ and $\{t_{kg}\}$ $(1 \leq g \leq G_k, 1 \leq k \leq K)$ so that different studies may have a different number of genes and the genes are not matched across studies. The gene matching issue is particularly significant when studies from different microarray platforms are combined. Supplementary Table 3 shows summary statistics of two lung cancer studies to be combined. The Bhat study used the Affymetrix HG-U95A platform and the Beer study used Affymetrix HU6800. Only 5515 Entrez genes overlapped across the two studies and the MAPE_G method had to drop information from 3490 out of 9005 genes that appear in Bhat but not in Beer. When more studies of different array platforms are included, the number of overlapping genes will decrease dramatically. Published studies have also demonstrated weak consistency across studies at the gene level but increased consistency at the pathway level (Ein-Dor *et al.*, 2005; Manoli *et al.*, 2006; Tan *et al.*, 2003). From these points of view, MAPE_P seems to be preferable to MAPE_G in general.

When we analyzed a combination of two lung cancer studies, however, we identified some example pathways with better power by MAPE_P and others with MAPE_G. Supplementary Figure 2 show scatter plots of $Q$-values obtained from MAPE_P and MAPE_G at minus log scale. The result shows somewhat general agreement in two methods while presents many discordant pathways detected by one method but not by the other. To gain a better intuition of why and how this happens, Figure 2A and B shows two example pathways of ALCALAY_AML_NPMC_UP (AANU; genes with increased expression in acute myeloid leukemia bearing cytoplasmic nucleophosmin) and HDACI_COLON_TSABUT_UP (HCTU; genes upregulated by both butyrate and trichostatin A at any time point up to 48 h in SW260 colon carcinoma cells), based on the C2 collection of MsigDB, a biology pathway database for cancer provided by Broad institute (Subramanian *et al.*, 2005). AANU was identified as an enriched pathway by MAPE_P but not by MAPE_G (Fig. 2A and C). In contrast, HCTU was identified by MAPE_G but not by MAPE_P (Fig. 2B and D). We performed DE gene analysis by $t$-test adjusted by Benjamini–Hochberg procedure in each study separately [false discovery rate (FDR) = 5%] and displayed only biomarkers detected by either study in Figure 2A–D. In Figure 2A and C, we found that 13 genes were identified as DE genes in both studies in the AANU pathway (gene set I in Fig. 2A). Thirteen genes were DE in Beer but not in Bhat and 27 genes were DE in Bhat but not in Beer (gene sets II and III in Fig. 2A). To explain the 'MAPE_P preferred pathway' in Figure 2A and 'MAPE_G preferred pathway' in Figure 2B, we defined a simple concordance index (CI) as the ratio of common DE genes in both studies versus DE genes in at least one of the two studies. The AANU pathway was detected by MAPE_P but not by MAPE_G because the CI is as low as $13/(13+13+27)=0.245$. When we pursued meta-analysis at the gene level, fewer genes were significant in Step II of Figure 1B although the meta-analysis at the pathway level in Step III of Figure 1C is statistically significant. On the other hand, the high CI in the HCTU pathway [$\mathrm{CI}=13/(13+1+9)=0.565$] was intuitively associated with the high statistical power of MAPE_G while MAPE_P did not detect this pathway. Such high CI pathways detected only by MAPE_G are usually important because the biomarkers are repeatedly identified in multiple studies. From the two example pathways above, we conclude that although intuitively MAPE_P has the convenience of not having to match genes across studies, MAPE_G has an advantage in some other particular situations. Based on this finding, we developed extensive simulations (shown in the Section 3) to illustrate conditions when MAPE_G outperforms MAPE_P and vice versa.

## 2.3 Framework of MAPE_I

Since pathways detected by both MAPE_G and MAPE_P are of biological interest, we proposed a simple integrative method, namely MAPE_I, to incorporate the complementary advantages of both methods (Fig. 1D). Specifically, we used a minP statistic that takes the minimum $P$-value from MAPE_G and MAPE_P for each pathway [i.e. $s_p = \min(p(v_p),\ p(w_p))$]. The statistical inference and control of FDR were similarly performed by permutation analysis (see detailed algorithm in Fig. 3).

## 2.4 Detailed implementation

Numerous pathway analysis and meta-analysis methods for microarray data have been described. Most of these methods have pros and cons under different conditions and for different biological purposes. Under the general framework shown in Figure 1 for MAPE_G, MAPE_P and MAPE_I, we can virtually apply and combine any pathway analysis and meta-analysis method for implementation. There are four major considerations and choices in practice.

*2.4.1 Statistic selection for association evidence with phenotype* For simplicity, we considered $t$-statistics for a binary phenotype label in this article. For multiclass, continuous or censored survival phenotype, different test statistics, such as $F$-statistics, Pearson correlation measure or statistics from Cox proportional hazard model, may be used, respectively.

*2.4.2 Statistic selection for meta-analysis* Various meta-analysis statistics, including Fisher's statistic, minimum $P$-value statistic (minP) and maximum $P$-value statistic (maxP), have been discussed in the Section 1. The best choice of meta-analysis statistics depends on the particular biological goal of interest. Following the convention of Birnbaum (Birnbaum, 1954), two different hypothesis settings may be considered:

$$\mathrm{HS}_A : \left\{ H_0 : \theta_{1g} = \cdots = \theta_{Kg} = 0 \text{ versus } H_A : \theta_{kg} \neq 0, \forall 1 \leq k \leq K \right\}$$

$$\mathrm{HS}_B : \left\{ H_0 : \theta_{1g} = \cdots = \theta_{Kg} = 0 \text{ versus } H_A : \text{ at least one } \theta_{kg} \neq 0 \right\}$$

where $\theta_{kg}$ represents the effect size of gene $g$ in study $k$. $\mathrm{HS}_A$ corresponds to the biological question: 'which genes are consistently differentially expressed in all studies?'. In contrast, $\mathrm{HS}_B$ detects genes if they are DE

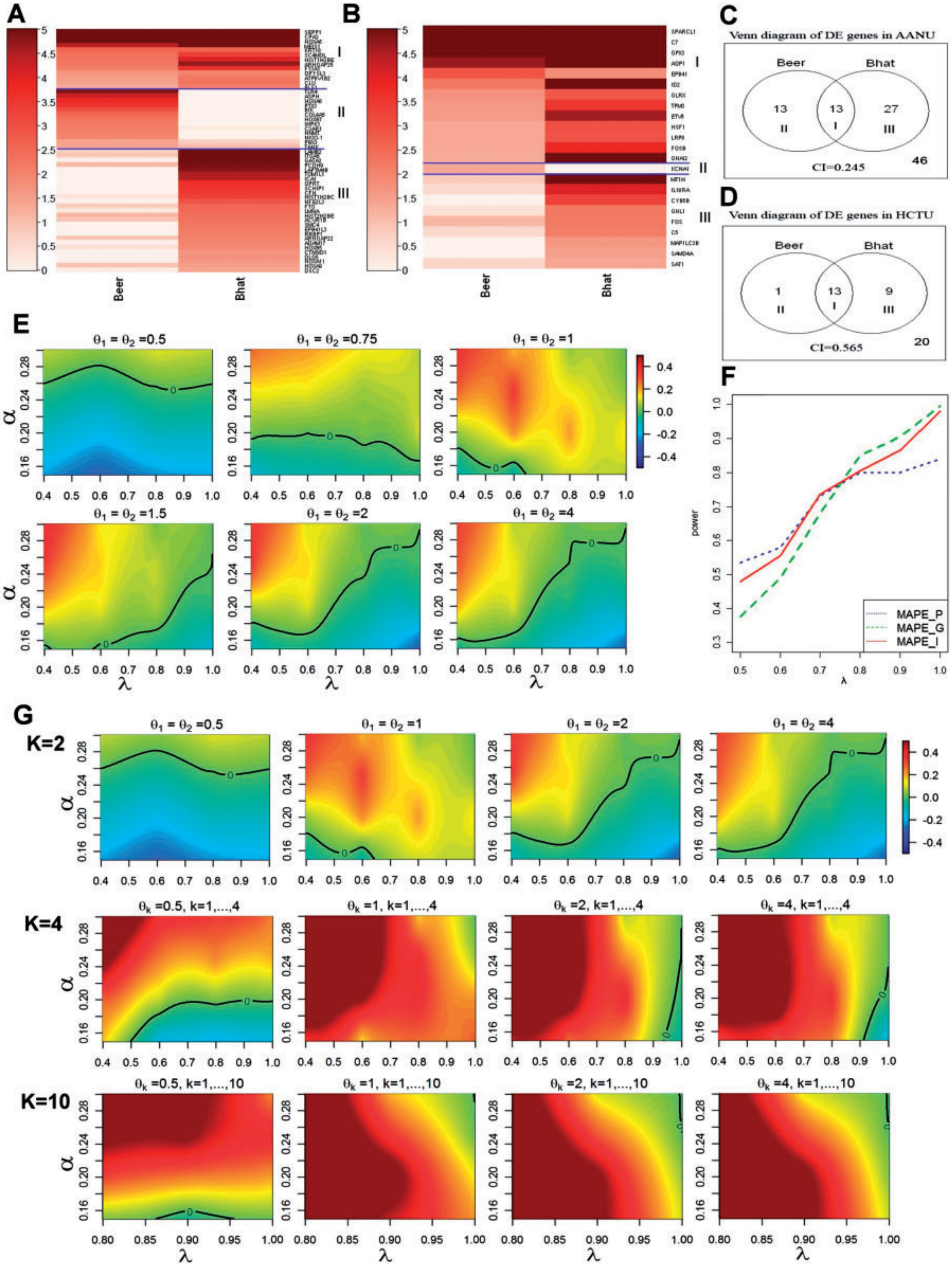

**Fig. 2.** (**A** and **B**) Heatmaps of genes in two example pathways identified by MAPE_P alone and by MAPE_G alone, respectively, in lung cancer studies. AANU in 2A is detected by MAPE_P ($q = 0.007$) but not by MAPE_G ($q = 0.073$) whereas HCTU in 2B is detected by MAPE_G ($q = 0.016$) but not MAPE_P ($q = 0.071$). The $Q$-values of each individual gene (on the row) and study (on the column) are shown by gradient color in $-\log$ (base 10) scale. Gene set I contains biomarkers with $Q$-value $<0.05$ for both studies and gene sets II and III contain significant biomarkers ($q < 0.05$) in one of the studies. (**C** and **D**) Venn diagram of biomarkers detected by each individual study (Beer and Bhat) in AANU and HCTU. (**E**) Power difference between MAPE_P and MAPE_G for various $\alpha$ and $\lambda$ when $\theta_1 = \theta_2 = 0.5$, 0.75, 1, 1.5, 2 and 4. Yellow/red color shows better power of MAPE_P over MAPE_G and blue color vice versa. Solid lines show contours of equal power between MAPE_P and MAPE_G. (**F**) The blue, green and red lines indicates the power of MAPE_P, MAPE_G and MAPE_I, respectively, when $\theta = 4$, $\alpha = 0.2$. (**G**) Power difference between MAPE_P and MAPE_G for various $\alpha$ and $\lambda$ when combining 2, 4 or 10 studies.

**Procedures of MAPE_G**

I. For a given study $k$, compute $P$-values of differential expression:

1. Compute the t-statistic, $t_{gk}$, of gene $g$ in study $k$, where $1 \le g \le G$, $1 \le k \le K$.

2. Permute group labels in each study $B$ times, and calculate the permuted statistics, $t_{gk}^{(b)}$, where $1 \le b \le B$.

3. Estimate the $P$-value of $t_{gk}$ as $p(t_{gk}) = \sum_{b=1}^{B} \sum_{g'=1}^{G} I(|t_{g'k}^{(b)}| \ge |t_{gk}|) \big/ (B \cdot G)$ and $P$-value of $t_{gk}^{(b)}$ as $p(t_{gk}^{(b)}) = \sum_{b'=1}^{B} \sum_{g'=1}^{G} I(|t_{g'k}^{(b')}| \ge |t_{gk}^{(b)}|) \big/ (B \cdot G)$.

II. Meta-analysis:

1. The maximum $P$-value statistic (maxP), $u_g = \max_{1 \le k \le K} p(t_{gk})$, is applied for the meta analysis. Similarly, $u_g^{(b)} = \max_{1 \le k \le K} p(t_{gk}^{(b)})$.

2. Estimate the $P$-value of maxP statistics as $p(u_g) = \sum_{b=1}^{B} \sum_{g'=1}^{G} I(u_{g'}^{(b)} \le u_g) \big/ (B \cdot G)$.

III. Enrichment analysis:

1. Given a pathway $p$, compute $v_p$, the KS statistic for gene set enrichment based on $p(u_g)$.

2. Permute gene labels $B$ times, and calculate the permuted statistics, $v_p^{(b)}$, $1 \le b \le B$.

3. Estimate $P$-value of pathway $p$ as $p(v_p) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(v_{p'}^{(b)} \ge v_p) \big/ (B \cdot P)$ and similarly calculate $p(v_p^{(b)}) = \sum_{b'=1}^{B} \sum_{p'=1}^{P} I(v_{p'}^{(b')} \ge v_p^{(b)}) \big/ (B \cdot P)$.

4. Estimate $Q$-value of pathway $p$ as $q(v_p) = \hat{\pi}_0 \cdot \sum_{b=1}^{B} \sum_{p'=1}^{P} I(v_{p'}^{(b)} \le v_p) \big/ \left(B \cdot \sum_{p'=1}^{P} I(v_{p'} \le v_p)\right)$, where $\hat{\pi}_0$ is an estimate of the proportion of non-enriched pathways. Following Storey (2002), we may estimate $\hat{\pi}_0 = \sum_{p=1}^{P} I(p(v_p) \in A) \big/ P \cdot l(A)$, where $A$=[0.5, 1] and $l(A)$=0.5. In practice, many reports have indicated the intrinsic difficulty in estimating $\pi_0$ and a poor estimate of $\pi_0$ can greatly deteriorate the FDR estimation. A conservative suggestion is to always set $\hat{\pi}_0 = 1$ and it is adopted throughout this paper. $\mathrm{P}_{MAPE\_G} = \{p : q(v_p) \le 0.05\}$ is the enriched pathways obtained by MAPE_G.

**Procedures of MAPE_P**

I. Pathway enrichment analysis:

1. For each study $k$, Calculate $p(t_{gk})$, the p-value of gene $g$, by Student t-test, $1 \le g \le G$.

2. Given a pathway $p$, compute the KS statistic $v_{pk}$ that compares the $P$-values ($p(t_{gk})$) inside and outside the pathway.

3. Permute gene labels B times, and calculate the permuted statistics, $v_{pk}^{(b)}$, $1 \le b \le B$.

4. Estimate the $P$-value of KS statistic in pathway $p$ and study $k$ as $p(v_{pk}) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(v_{p'k}^{(b)} \ge v_{pk}) \big/ (B \cdot P)$ and similarly calculate $p(v_{pk}^{(b)}) = \sum_{b'=1}^{B} \sum_{p'=1}^{P} I(v_{p'k}^{(b')} \ge v_{pk}^{(b)}) \big/ (B \cdot P)$.

II. Meta-analysis:

1. The maximum $P$-value statistic (maxP) is applied for meta-analysis: $w_p = \max_{1 \le k \le K} p(v_{pk})$ and $w_p^{(b)} = \max_{1 \le k \le K} p(v_{pk}^{(b)})$.

2. Estimate $P$-value of $w_p$ as $p(w_p) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(w_{p'}^{(b)} \le w_p) \big/ (B \cdot P)$. Similarly $p(w_p^{(b)}) = \sum_{b'=1}^{B} \sum_{p'=1}^{P} I(w_{p'}^{(b')} \le w_p^{(b)}) \big/ (B \cdot P)$.

3. Estimate $Q$-value as $q(w_p) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(w_{p'}^{(b)} \le w_p) \big/ \left(B \cdot \sum_{p'=1}^{P} I(w_{p'} \le w_p)\right)$. $\mathrm{P}_{MAPE\_P} = \{p : q(w_p) \le 0.05\}$ are enriched pathways from MAPE_P.

**Procedures of MAPE_I**

1. Let $s_p = \min(p(v_p), p(w_p))$ and $s_p^{(b)} = \min(p(v_p^{(b)}), p(w_p^{(b)}))$ from Procedures in MAPE_G and MAPE_P in Figure 3 and 4.

2. Estimate the $P$-value as $p(s_p) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(s_{p'}^{(b)} \le s_p) \big/ (B \cdot P)$.

3. Estimate $Q$-value as $q(s_p) = \sum_{b=1}^{B} \sum_{p'=1}^{P} I(s_{p'}^{(b)} \le s_p) \big/ \left(B \cdot \sum_{p'=1}^{P} I(s_{p'} \le s_p)\right)$. $\mathrm{P}_{MAPE\_I} = \{p : q(s_p) \le 0.05\}$ are enriched pathways identified by MAPE_I.

**Fig. 3.** Detailed algorithms of MAPE_G, MAPE_P and MAPE_I.

in one or more studies. It can be seen that maxP corresponds to $\mathrm{HS}_A$, and Fisher's statistic and minP correspond to $\mathrm{HS}_B$. In this article, we focus on the conservative maxP statistic to identify consistent biomarkers across all microarray studies. Specifically, we will first calculate the $P$-values of evidence scores at the gene level in Step II of Figure 1B or at the pathway level in Step III of Figure 1C. The maxP statistic for meta-analysis at the gene level is $u_g = \max_{1 \le k \le K} p(t_{gk})$ and the pathway level is $w_p = \max_{1 \le k \le K} p(v_{kp})$.

*2.4.3 Statistic selection for the pathway enrichment analysis method* The goal of pathway analysis is to test whether genes in a pathway are coherently associated with the phenotype of interest. The first widely used enrichment evidence score was calculated by Fisher's exact test to determine whether the ratio of DE genes in a pathway is higher than the ratio outside the pathway with statistical significance. The shortcoming of Fisher's exact test is that it loses information by only counting the number of DE genes instead of considering the gene order, which can be overcome by using averaged $t$-statistics or KS statistics. In this article, we apply KS statistics (described in Supplementary Material) for the pathway enrichment analysis.

*2.4.4 Control of false discovery and evaluation of Q-value* The $P$-values and $Q$-values of the pathway enrichment evidence score are usually computed by permutation test, considering that the null distribution is difficult to obtain analytically. Two basic permutation procedures, gene permutation and phenotype permutation, have been proposed based on two related but not equivalent null hypotheses, Q1 and Q2, respectively, as follows:

Q1: the genes in a gene set have the same pattern associated with the phenotype of interest as the genes outside of the gene set.

Q2: no genes in the gene set having expression patterns associated with the phenotype.

Details about these two null hypotheses have been discussed by Tian (Tian *et al.*, 2005) and Geoman (Goeman and Buhlmann, 2007), and both approaches have their limitations. Briefly, Q1 with gene permutation takes the background information (the expression of genes outside of the pathway) into consideration while destroys the gene correlation structure in the data. Q2 with sample permutation preserves the gene correlation structure while ignores the background information. How to explain and integrate the different test results by Q1 and Q2 is still an open question. Tian proposed a heuristic gene set ranking by the average rank from phenotype permutation and gene permutation tests. The GSEA adopted a hybrid method that generates null distribution under Q2 and used KS statistics to test Q1. Efron (Efron and Tibshirani, 2007) proposed Gene Set Analysis (GSA) analysis by using a max-mean statistics and re-standardization procedure. A bootstrap method and new null were introduced by Barry *et al.* (2008) to overcome the weakness of both gene and sample permutation. Conceptually, any gene set analysis method described above can be adopted into our general framework depicted in Figure 1. For simplicity and fast computation, we used the KS statistic for Q1 (i.e. gene permutation) to illustrate our enrichment procedure in this article. Detailed algorithms for MAPE_P, MAPE_G and MAPE_I are given in Figure 3, respectively.

## 3 RESULTS

### 3.1 Simulation results

We applied a one-pathway simple simulation model to compare the power of MAPE_G, MAPE_P and MAPE_I and to identify conditions (parameter subspaces) in which one method outperforms the other. We define the type of biologically relevant pathways with statistical power of MAPE_G greater than that of MAPE_P as 'MAPE_G preferred pathways' and define 'MAPE_P preferred pathways' similarly. The simulation results below will gives us insight into the unique advantages of MAPE_G and MAPE_P, respectively, and argue the necessity of MAPE_I when both MAPE_G preferred and MAPE_P preferred pathways exist in the data and we are interested in detecting both types of biologically relevant pathways.

Suppose $G=500$ genes are contained in the genome. The first 100 genes belong to a pathway. Our pathway database has only one pathway ($p$ = 1): $\{z_{gp}\}$, $z_{gp}=1$ when $1\leq g\leq 100$ and $z_{gp}=0$ when $101\leq g\leq 500$. We generate a random binary vector $D=\{d_1,\dots,d_G\}$ to indicate whether gene $g$ is a DE gene or not. In the first 100 genes, a total of $100\times\alpha$ genes are DE genes (i.e. $\sum_{g=1}^{100}d_g=100\cdot\alpha$). In the next 400 genes, a total of $100\times\alpha_0$ genes are DE genes (i.e. $\sum_{g=101}^{500}d_g=400\cdot\alpha_0$). We fix $\alpha_0=0.1$ in our simulation. Intuitively, there is no pathway enrichment if $\alpha=0.1$ and pathway enrichment exists if $\alpha>0.1$.

Given the DE gene indicators, two independent array studies are subsequently simulated for meta-analysis. We assume each study contains $S=40$ samples. The first 20 samples are controls and the next 20 samples are cases (i.e. $y_s=0$ if $1\leq s\leq 20$ and $y_s=1$ if $21\leq s\leq 40$). When gene $g$ is a DE gene ($d_g=1$) and for all $k$, the expression intensities are simulated from $x_{kgs}\sim$N $(\theta,1)$ if $1\leq s\leq 20$ and $x_{kgs}\sim$N (0,1) if $21\leq s\leq 40$. For a non-DE gene $g$ ($d_g=0$), the expression intensities are simulated from $x_{kgs}\sim$N (0,1) $\forall s$ and $k$. We further assume that the two array studies adopt different array platforms and each of them only covers a portion of genes in the genome. We assume the chance of each gene to be covered by study $k$ is randomly generated with a sampling rate $\lambda_k$. The sampled indicator vectors for gene $g$ in study $k$ is denoted by $h_{gk}$, where $h_{gk}=1$ if gene g appears in study $k$ and $h_{gk}=0$ otherwise. In the following, we set $\lambda=Pr(h_{gk}=1)=\lambda_k$ $(1\leq g\leq G=500$ and $1\leq k\leq K=2)$. As a result, study $k$ contains $G_k=\sum_{g=1}^{G}h_{gk}$ genes in the data matrix, which is a random variable and may be different in each simulation. The overlapped gene set of the two studies contains $G'=\sum_{g=1}^{G}h_{g1}\cdot h_{g2}$. In the implementation of MAPE_P, the original data in both studies with $G_1$ and $G_2$ genes can be used. For MAPE_G, the method requires only matched genes and the overlapped gene set $G'$ in each study will be applied.

We perform $\alpha=\{0.15, 0.2, 0.25, 0.3\}$ and $\lambda=\{0.4, 0.6, 0.8, 1\}$ in this article. A total of $B=200$ independent simulations are performed for each parameter setting. Intuitively, $\theta$ represents the effect size of the DE genes in the data, $\alpha$ represents the strength of pathway enrichment and $\lambda$ represents the coverage of an array platform on the genome. The power calculation of MAPE procedures is calculated as the proportion of occurrences that the pathway is claimed as an enriched pathway (details shown in the Supplementary Material). Five simulation scenarios have been extensively explored below.

Scenario 1 (different degrees of effect sizes): $\theta_1=\theta_2=$0.5, 0.75, 1, 1.5, 2 and 4.

Scenario 2 (unequal effect size or equivalently unequal sample size across studies): $(\theta_1=2, \theta_2=3)$ and $(\theta_1=2, \theta_2=4)$.

Scenario 3 (increasing number of studies): $\theta_1=\dots=\theta_4=0.5$, 1, 2 and 4; $\theta_1=\dots=\theta_{10}=0.5$, 1, 2 and 4.

Scenario 4 (varying effect sizes across genes): $\theta_{1g}=\theta_{2g}\sim$N (1.5, 0.5), N (2, 0.5) and N (4, 0.5).

Scenario 5 (inclusion of one weak signal study): $(\theta_1=\theta_2=\theta_3=2, \theta_4=0.1)$ and $(\theta_1=\theta_2=\theta_3)$ $\sim$ N (2, 0.05), $\theta_4\sim$N (2, 0.2).

Results of Scenario 1 of varying effect sizes and Scenario 3 of varying number of studies are displayed in Figure 2E and G. The power difference of MAPE_P and MAPE_G are displayed by gradient colors under different $\alpha$ and $\lambda$ conditions. The smooth contour plots are performed with a surface smoothing technique using the R package 'field' (Fields Development Team, 2006) and the solid lines demonstrating equal statistical power between MAPE_G and MAPE_P are shown in each plot. In the weak signal situation in Figure 2E $(\theta_1=\theta_2=0.5)$, MAPE_G outperforms MAPE_P. For medium signal conditions $(\theta_1=\theta_2=0.75$, 1 and 1.5), MAPE_P generally outperforms MAPE_G. Finally, when the effect sizes are large $(\theta_1=\theta_2=2$ and 4), MAPE_G is more powerful when the array coverage rate $\lambda$ $(0.8\leq\lambda\leq 1)$ is high or the pathway enrichment strength $\alpha$ is low $(0.15\leq\lambda\leq 0.2)$. On the other hand, for a low array coverage rate $\lambda$ $(0.4\leq\lambda\leq 0.7)$, the advantage of MAPE_P of not requiring gene matching across studies becomes evident and MAPE_P is more powerful than MAPE_G. In Figure 2G, MAPE_P has increasingly better statistical power than MAPE_G when the number of studies increases from 2 to 10. The detailed results of statistical power of MAPE_G, MAPE_P and the corresponding power difference are shown in Supplementary Figures 3, 5.1 and 5.2.

Our simulation examines the power of a single pathway. In a real application, hundreds to thousands of pathways are analyzed in the pathway database. Both types of pathways for which MAPE_G or MAPE_P have better power will co-exist in an analysis. This motivates our development of an integrated method MAPE_I to incorporate the advantages of the two methods. Figure 2F shows the power curves of the three methods when $\theta=4$, $\alpha=0.2$ and varying $\lambda$. MAPE_I clearly has more robust and better performance over MAPE_G and MAPE_P on the two extremes of small and large $\lambda$. The power curves in Supplementary Figures 3B, 5.1B and 5.2B show that MAPE_I generally has the best or near the best statistical power among the three MAPE variations.

Results of Scenario 2 (unequal effect size across studies), Scenario 4 (varying effect sizes across genes) and Scenario 5 (inclusion of a weak signal study) are shown in Supplementary

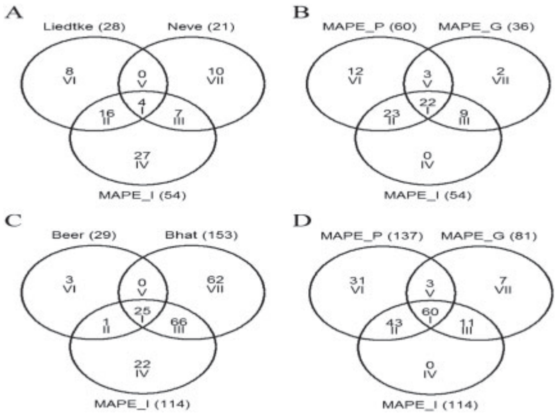

**Fig. 4.** Venn diagram comparing pathways detected by individual studies and by the three MAPE methods. (**A** and **B**) Application in combining Liedtke and Neve breast cancer drug response studies. (**C** and **D**) Application in combining Beer and Bhat lung cancer studies.

Material (Supplementary Figs 4, 6 and 7). The conclusions did not alter those from the basic simulation of Scenario 1. For Scenario 5, inclusion of a weak signal study does not affect the conclusion, showing robustness of meta-analysis when majority of studies contain concordant information.

## 3.2 Two applications in breast cancer cell line's drug response studies and lung cancer studies

We applied the proposed methods to two applications: drug response in breast cancer cell lines and lung cancer studies. The pathway database used was the C2 collection of MsigDB described earlier. There were 1892 pathways in total in the C2 collection. We only tested pathways with pathway size (# of genes) larger than 15 and less than 500 in the data. Under this criterion, 970 pathways are tested in the lung example and 704 pathways in the breast cancer example. In the first application, microarray datasets from two independent breast cancer cell line studies were downloaded from public websites (details see Supplementary Material). The breast cell lines' sensitivity or resistance to paclitaxel was determined by 50% growth inhibitory concentrations. We applied MAPE approaches to identify enriched pathways that are related to drug response to paclitaxel in breast cancer cells. Drug response-related pathways provide important information for studying mechanisms of drug resistance and will shed light to new pharmaceutical targets or techniques to overcome resistance. Figure 4A shows a venn diagram of pathways detected by each individual study and by MAPE_I when $Q$-value cutoff was set to 0.15. Clearly, meta-analysis provides better statistical power to identify more pathways that individual analyses cannot discover (54 pathways identified by MAPE_I versus 28 and 21 pathways identified by each individual study). Among the 27 pathways detected by MAPE_I but not by either individual study analysis (category IV in Fig. 4A), many are known drug response-related pathways, including MYC- and EGF-related pathways. In the 8 and 10 pathways identified by individual studies but not by MAPE_I (category VI and VII in Fig. 4A), no known drug – response-related pathway was found. In Figure 3B, 60, 36 and 54 pathways were identified by MAPE_P, MAPE_G and MAPE_I, respectively. Among pathways detected by MAPE_P but not by MAPE_G (see categories II and VI in Supplementary Table 2), many are cell proliferation, gene regulation and estrogen-related pathways. Similarly, pathways detected by MAPE_G but not by MAPE_P (categories III and VII in Supplementary Table 2) also include many cancer and drug response-related pathways. MAPE_I identified top-ranked pathways detected by either methods with the price of ignoring low-ranked pathways (categories V, VI and VII) at $q = 0.15$. If we relax the $Q$-value cutoff of MAPE_I to 0.2, majority (10 out of 17) of pathways identified by MAPE_P or MAPE_G but not by MAPE_I at cutoff 0.15 were identified by MAPE_I, showing that MAPE_I is a good way to incorporate, summarize and prioritize resulting pathways from MAPE_P and MAPE_G. Details of all enriched pathway results are listed in Supplementary Table 2. These pathways are predominantly related to cell proliferation, oncogenic pathways and estrogen receptor-associated pathways. In the literature, intersection or union methods are commonly used due to their simplicity. From a statistical point of view, the intersection method is clearly over conservative and the union method is anticonservative by losing control of FDR. From Figure 4A, intersection method identifies only four pathways and is too conservative. The union method identifies 45 pathways, which is fewer than 54 pathways identified by MAPE_I and loses FDR control.

In the second example, we performed similar analysis to the application of combining two lung cancer studies when comparing normal versus adenocarcinoma tissues. Similar conclusions are found in Figure 4C and D. The detailed results are shown in Supplementary Table 4.

## 4 CONCLUSIONS AND DISCUSSIONS

In this article, we discussed framework and evaluation of two meta-analysis approaches for pathway enrichment analysis, namely MAPE_G and MAPE_P, which combine statistical significance at the gene level and at the pathway level respectively. In general, MAPE_P has the advantage of not requiring gene matching across studies and is often statistically more powerful. MAPE_G is, however, found more powerful when the majority of genes across studies can be properly matched or when the effect sizes are small. We proposed an integrated approach, namely MAPE_I, to accommodate the advantages of MAPE_G and MAPE_P and to capture pathways of potential biological interest from both methods. Our simulation study characterized conditions of when and how MAPE_G and MAPE_P outperform one another and verified the robust performance of MAPE_I. Applications to breast cancer cell line drug response and lung cancer demonstrated similar conclusions and identified previously verified pathways related to drug response and carcinogenesis. Meta-analysis identified more pathways than individual studies. The MAPE_I procedure integrated results from MAPE_P and MAPE_G and is generally recommended in practice.

To our knowledge, this is the first article to systematically investigate and develop meta-analysis approaches for pathway enrichment analysis. This article provides an initial investigation of a unified framework. Conceptually, any meta-analysis technique and pathway enrichment method can be combined under the proposed framework. This article has several limitations and future directions. First of all, the conclusions are drawn in two class comparisons using $t$-statistics. Our software package has allowed correlation, $F$-statistics and Cox proportional hazard model to allow continuous, multiclass and censored survival clinical variables while more extensive evaluation is needed for such situations. Secondly, we require no gene missingness across all studies for MAPE_G in this article. This requirement could be relaxed by modifying MAPE_G algorithm allowing gene missingness in the inference of null distributions. Thirdly, our algorithm integrates information across studies by taking maximum of *P*-values, requiring DE signal in all studies. When many studies are combined, the maxP method becomes too stringent and a robust *r*-th rank method (e.g. 70 percentile statistics across all studies) is more desirable. Finally, the permutation test scheme in this article does not rigorously consider gene dependency structure and potential hierarchical dependency in pathway database (e.g. Gene Ontology). Although detailed discussion is beyond the scope of this article, many recent methods have been developed to answer or alleviate this problem (Benjamini and Yekutieli, 2001; Farcomeni, 2006; Romano *et al.*, 2008; Tsai *et al.*, 2003) and incorporation of these methods is a future direction.

As was discussed in text, many meta-analysis techniques and pathway enrichment analysis methods have been developed in the past few years. Another valuable future direction would involve incorporating and evaluating the many available methods in both areas and assess their performance to choose the best method. As the next-generation sequencing technology is getting popular and will become affordable for expression analysis in the near future, the proposed methodology can be extended to such RNA-seq data.

## ACKNOWLEDGEMENTS

The authors would like to thank the reviewers and the editor for insightful suggestions and critiques.

*Funding*: National Institutes of Health (KL2 RR024154-04 to G.C.T.); University of Pittsburgh (Central Research Development Fund, CRDF; Competitive Medical Research Fund, CMRF).

*Conflict of Interest*: none declared.

## REFERENCES

Barry,T. *et al.* (2008) A statistical framework for testing functional categories in microarray data. *Ann. Appl. Stat.*, **2**, 286–315.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.*, **29**, 1165–1188.

Birnbaum,A. (1954) Combining independent tests of significance. *J. Am. Stat. Assoc.*, **49**, 559–574.

Borovecki,F. *et al.* (2005) Genome-wide expression profiling of human blood reveals biomarkers for Huntington's disease. *Proc. Natl Acad. Sci. USA*, **102**, 11023–11028.

Cardoso,J. *et al.* (2007) Expression and genomic profiling of colorectal cancer. *Biochim. Biophy. Acta Rev Cancer*, **1775**, 103–137.

Choi,H. *et al.* (2007) A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, **8**, 364.

Choi,J. *et al.* (2003) Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, **19**, i84–i90.

Efron,B. and Tibshirani,R. (2007) On testing the significance of sets of genes. *Ann. Appl. Stat.*, **1**, 107–129.

Ein-Dor,L. *et al.* (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21**, 171–178.

Farcomeni,A. (2006) More powerful control of the false discovery rate under dependence. *Stat. Meth. Appl.*, **15**, 43–73.

Fields Development Team (2006) *Fields: Tools for Spatial Data*. National Center for Atmospheric Research, Boulder, CO.

Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

Hosack,D. *et al.* (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.*, **28**, 27–30.

Manoli,T. *et al.* (2006) Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics*, **22**, 2500–2506.

Mosteller,F. and Fisher,R.A. (1948) Questions and answers. *Am. Stat.*, **2**, 30–31.

Newton,M. *et al.* (2007) Random-set methods identify distinct aspects of the enrichment signal in gene-set analysis. *Ann. Appl. Stat.*, **1**, 85–106.

Pirooznia,M. *et al.* (2007) GeneVenn - a web application for comparing gene lists using Venn diagrams. *Bioinformation*, **1**, 420–422.

Rhodes,D.R. *et al.* (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.

Romano,J.P. *et al.* (2008) Control of the false discovery rate under dependence using the bootstrap and subsampling. *Test*, **17**, 417–442.

Segal,E. *et al.* (2004) A module map showing conditional activity of expression modules in cancer. *Nat. Genet.*, **36**, 1090–1098.

Setlur,S.R. *et al.* (2007) Integrative microarray analysis of pathways dysregulated in metastatic prostate cancer. *Cancer Res.*, **67**, 10296–10303.

Shen,R. *et al.* (2004) Prognostic meta-signature of breast cancer developed by two-stage mixture modeling of microarray data. *BMC Genomics*, **5**, 94.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

Tan,P.K. *et al.* (2003) Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.*, **31**, 5676–5684.

Thomassen,M. *et al.* (2008) Gene expression meta-analysis identifies metastatic pathways and transcription factors in breast cancer. *BMC Cancer*, **8**, 394.

Tian,L. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proc. Natl Acad. Sci. USA*, **102**, 13544–13549.

Tippett,L.H.C. (1931) *The Methods in Statistics*. Williams and Norgate, Ltd, London.

Tsai,C. *et al.* (2003) Estimation of false discovery rates in multiple testing: application to gene microarray data. *Biometrics*, **59**, 1071–1081.

Wilkinson,B. (1951) A statistical consideration in psychological research. *Psychol. Bull.*, **48**, 156–158.