frontiers in Genetics

ORIGINAL RESEARCH
published: 02 July 2021
doi: 10.3389/fgene.2021.651546

Biomarker Categorization in Transcriptomic Meta-Analysis by Concordant Patterns With Application to Pan-Cancer Studies

Zhenyao Ye[1†], Hongjie Ke[1†], Shuo Chen[2], Raul Cruz-Cano[1], Xin He[1], Jing Zhang[1], Joanne Dorgan[2], Donald K. Milton[3] and Tianzhou Ma[1]*

[1] Department of Epidemiology and Biostatistics, School of Public Health, University of Maryland, College Park, College Park, MD, United States, [2] Department of Epidemiology and Public Health, School of Medicine, University of Maryland, Baltimore, Baltimore, MD, United States, [3] Maryland Institute for Applied Environmental Health, School of Public Health, University of Maryland, College Park, College Park, MD, United States

OPEN ACCESS

Edited by:
Chao Xu,
University of Oklahoma Health Sciences Center, United States

Reviewed by:
Lan Zhang,
University of Michigan, United States
Md Ashad Alam,
Tulane University, United States

***Correspondence:**
Tianzhou Ma
tma0929@umd.edu

[†] *These authors have contributed equally to this work*

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *10 January 2021*
Accepted: *28 May 2021*
Published: *02 July 2021*

Citation:
Ye Z, Ke H, Chen S, Cruz-Cano R, He X, Zhang J, Dorgan J, Milton DK and Ma T (2021) Biomarker Categorization in Transcriptomic Meta-Analysis by Concordant Patterns With Application to Pan-Cancer Studies. Front. Genet. 12:651546. doi: 10.3389/fgene.2021.651546

With the increasing availability and dropping cost of high-throughput technology in recent years, many-omics datasets have accumulated in the public domain. Combining multiple transcriptomic studies on related hypothesis via meta-analysis can improve statistical power and reproducibility over single studies. For differential expression (DE) analysis, biomarker categorization by DE pattern across studies is a natural but critical task following biomarker detection to help explain between study heterogeneity and classify biomarkers into categories with potentially related functionality. In this paper, we propose a novel meta-analysis method to categorize biomarkers by simultaneously considering the concordant pattern and the biological and statistical significance across studies. Biomarkers with the same DE pattern can be analyzed together in downstream pathway enrichment analysis. In the presence of different types of transcripts (e.g., mRNA, miRNA, and lncRNA, etc.), integrative analysis including miRNA/lncRNA target enrichment analysis and miRNA-mRNA and lncRNA-mRNA causal regulatory network analysis can be conducted jointly on all the transcripts of the same category. We applied our method to two Pan-cancer transcriptomic study examples with single or multiple types of transcripts available. Targeted downstream analysis identified categories of biomarkers with unique functionality and regulatory relationships that motivate new hypothesis in Pan-cancer analysis.

Keywords: biomarker categorization, differential expression, meta-analysis, pan-cancer, transcriptomics

INTRODUCTION

The revolutionary advancement of high-throughput technology in recent years has generated large amounts of omics data of various kinds (e.g., genetics variants, gene expression and DNA methylation, etc.), which improves our understanding of human disease and enables the development of more effective therapies in personalized medicine (Richardson et al., 2016). As more studies are conducted on a related hypothesis, meta-analysis, by combining evidence from multiple studies, has become a popular choice in genomic research to improve upon the power,

accuracy, and reproducibility of individual studies (Ramasamy et al., 2008; Begum et al., 2012; Tseng et al., 2012). One of the main purposes of transcriptomics studies is to identify genes or RNAs that express differently between two or more conditions (e.g., diseased patients vs. healthy controls), also known as differential expression (DE) analysis or candidate biomarker detection. Many meta-analysis methods have been developed or applied to DE analysis, including combining *p*-values (Fisher, 1992) or effect sizes (Choi et al., 2003) and rank-based approaches (Hong et al., 2006). One may refer to Tseng et al. (2012) for an overview of the major meta-analysis methods in transcriptomic studies and Ma et al. (2019) for an overview of available software tools. Yet, a majority of conventional meta-analysis methods only generate a list of differentially expressed genes with strong aggregated evidence without further investigating in what studies are the genes differentially expressed.

Study or population heterogeneity always exists and has been critical to biomarker detection (Di Camillo et al., 2012). For example, The Cancer Genome Atlas (TCGA) consortium completed a Pan-Cancer Atlas of multi-platform molecular profiles spanning 33 cancer types in an effort to provide insights into the commonalities and differences across tumor lineages (Weinstein et al., 2013; Hoadley et al., 2018). When meta-analysis is performed on Pan-cancer transcriptomic studies, we expect to see both DE genes common in all tumor types as well as genes differentially expressed in some tumor types but not others. Biomarker categorization according to their DE patterns across studies is demanding in genomic studies for three reasons. First, biomarkers that share unique cross-study DE patterns are potentially involved in related functions (Berger et al., 2018). Such unique categories of genes with similar function can be used to generate new biological hypotheses. Second, biomarker categorization can make high dimensional genomic data more tractable. For example, in cancer transcriptomic studies, which frequently detect thousands of DE genes, downstream analysis methods such as pathway enrichment analysis or network analysis cannot be applied directly. By partitioning the original large set of DE genes into smaller subsets, biomarker categorization facilitates more focused downstream analysis. Third, RNA sequencing (RNA-seq) technology has led to an explosion of transcriptomic studies profiling both coding (i.e., mRNA) and noncoding RNAs (i.e., miRNA, rRNA, lncRNA, etc.) (Di Bella et al., 2020). Joint analysis of different RNA types with the same cross-study DE patterns can improve understanding of their regulatory relationships, which may lead to inferences about the underlying mechanisms of complex human diseases like cancer.

Li and Tseng (2011) first proposed an adaptively weighted Fisher (AW-Fisher) method for biomarker categorization that assigns a binary weight of 0 or 1 to each study and searches for the pattern of weights that minimizes the aggregate statistics for each gene. Though the method incorporates statistical significance by combining two-sided *p*-values across studies, it does not take into account the direction of regulation (e.g., up-regulated or down-regulated). Other methods incorporate biomarker categorization within the Bayesian framework and combine one-sided *p*-values or Bayesian posterior probabilities (Ma et al., 2017; Huo et al., 2019) but not the magnitudes of effect sizes. In practice, biological significance (i.e., large effect size) and statistical significance (i.e., small *p*-value) do not always occur in tandem (depending on sample size and variance) though they are equally important in interpreting study results (Sullivan and Feinn, 2012; Solla et al., 2018).

In this paper, we propose a novel meta-analysis method to detect and categorize biomarkers by simultaneously considering concordant pattern (i.e., direction of regulation), biological and statistical significance across studies. In addition, we develop a permutation test to assess the uncertainty of the proposed statistics and to control the false discovery rate (FDR). When only coding genes are included, after categorization we perform downstream pathway enrichment analysis with topological information on each category of genes for more biological insights (**Figure 1A**). In the presence of diverse RNAs, we jointly analyze all RNA species in the same category using miRNA/lncRNA target enrichment analysis and lncRNA-mRNA and miRNA-mRNA causal regulatory network analysis (**Figure 1B**). We show by simulation that our method detects both concordant and discordant biomarkers and assigns the correct weights. We apply our method to two Pan-cancer transcriptomic data examples: (1) Pan Gynecologic cancer (Pan-Gyn) data with coding genes only; (2) Pan Kidney cancer (Pan-Kidney) data that include mRNA, miRNA as well as lncRNA. The identified biomarker categories show unique functionality and informative regulatory relationships and could suggest new hypotheses about mechanisms underlying exclusive and shared features of different cancer types.

MATERIALS AND METHODS

Popular Meta-Analysis Methods

Tseng et al. (2012) reviewed the major types of meta-analysis methods for DE gene detection in microarrays and classified the methods into four main classes: combining *p*-values, combining effect sizes, combining ranks, and direct merging. We will discuss selected meta-analysis methods from the first two classes that are relevant to our proposed method.

Combining *P*-Values

Fisher's method (Fisher, 1992)

The conventional Fisher's method combines log transformed *p*-value from each study with the statistic $T_{\text{Fisher}} = -2\sum_{k=1}^{K}\log(p_k)$, which follows a χ^2 distribution with $2K$ degrees of freedom under the null hypothesis (i.e., genes not differentially expressed in all studies), where K is the number of studies and p_k is the *p*-value of study k, $1 \le k \le K$.

Stouffer's method (Stouffer, 1949)

The Stouffer's method proposes inverse normal transformation of *p*-value with the statistic $T_{\text{Stouffer}} \sum_{k=1}^{K} \Phi^{-1}(1-p_k)/\sqrt{K}$, which follows a standard normal distribution under the null, where $\Phi^{-1}(x)$ is the inverse cumulative distribution function of the standard normal distribution.

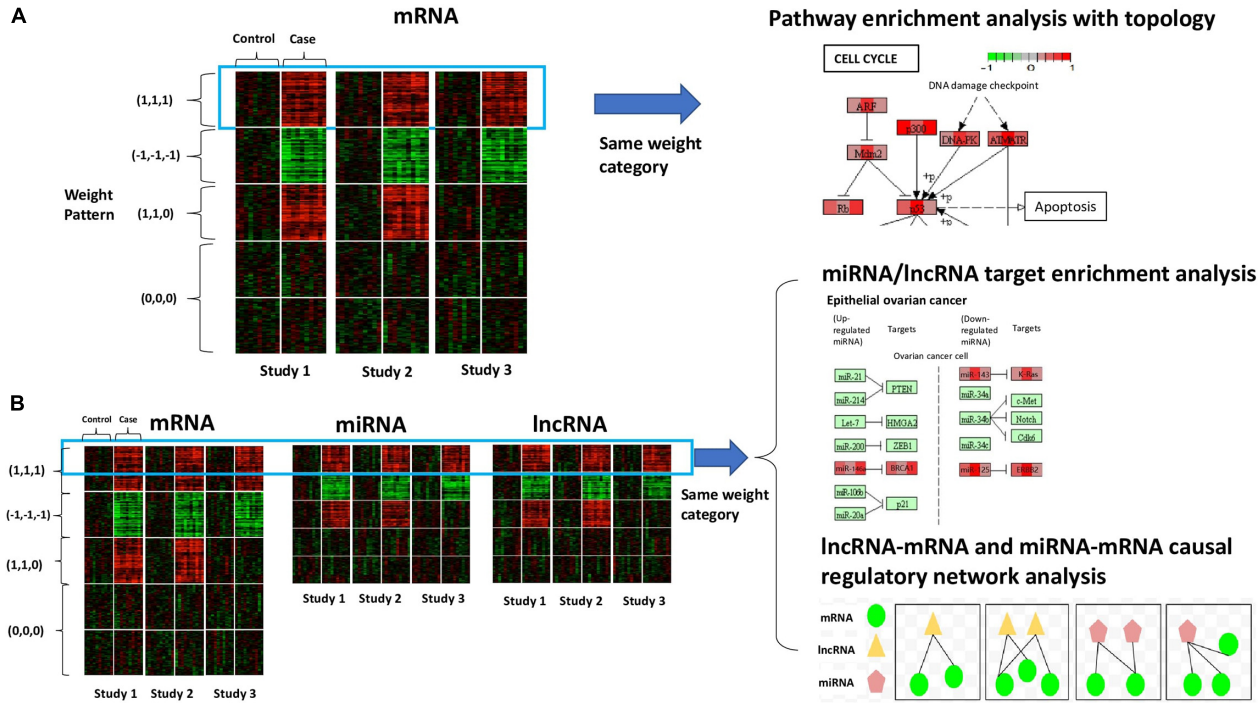

FIGURE 1 | Conceptual framework of our method. **(A)** The scenario with mRNA (or coding genes) only. The heatmap shows the gene expression of all samples from three studies. Rows refer to genes sorted by the specified weight category, columns refer to samples, and solid white lines are used to separate different conditions (control vs. case). Colors of the cells correspond to scaled expression level. The green/red indicates lower/higher expression. Pathway enrichment analysis is applied to genes belonging to the same weight category with topological information to visualize the cross-study DE patterns at the molecular level. **(B)** The scenario with diverse RNA species (e.g., mRNA, miRNA, and lncRNA). The three heatmaps show the expression of different types of transcripts of all samples from three studies, sorted by weight category. In the presence of multiple types of RNA species, we will perform integrative analysis on all the transcripts belonging to the same weight category together. Possible downstream analysis includes miRNA/lncRNA target enrichment analysis and lncRNA-mRNA and miRNA-mRNA causal regulatory network analysis.

Adaptively weighted fisher's method (AW-Fisher) (Li and Tseng, 2011)

Fisher's method does not differentiate DE in a single study or multiple studies as long as their aggregate contribution to the final statistics remains the same. To overcome this and better explain the between study heterogeneity, Li and Tseng (2011) introduced an AW-Fisher's method as a modification of the original Fisher's method. The AW-Fisher method considers $U(\overrightarrow{w}) = -2\sum_{k=1}^{K} w_k log(p_k)$ for each gene, where $\overrightarrow{w} = (w_1, \ldots, w_K)$ and each w_k is a binary weight of 0 or 1 assigned to each study k. Denote by $p\left(U(\overrightarrow{w})\right)$ the p-value when the weight $\overrightarrow{w}$ is given, the AW-Fisher statistic is defined as: $T_{AW} = \min_{\overrightarrow{w}} p\left(U(\overrightarrow{w})\right)$, where the optimal weight $(\widehat{w_1}, \ldots, \widehat{w_K})$ that minimizes the p-value indicates the subset of studies that contribute to the aggregate statistics and naturally categorizes the biomarkers. There is no closed-form distribution for AW-Fisher statistics under the null, so permutation tests and importance sampling is used to obtain the p-value and control the FDR.

Combining Effect Size

Fixed effect model (FEM) and random effect model (REM) (Choi et al., 2003)

Fixed effect model (FEM) combines effect sizes across all studies for each gene using a simple liner model: $T_k = \mu + \varepsilon_k,\ \varepsilon_k \sim N(0, s_k^2)$, where μ is the overall mean and the within-study variance s_k^2 represents the sampling error conditioned on study k. The combined point estimate of μ is a weighted average of study-specific effect sizes, where weights are equal to the inverse of s_k^2. FEM will prioritize concordant genes with the same directionality across all studies.

When strong between studies heterogeneity exists and the underlying population effect size is assumed to be unequal across studies, an REM is given hierarchically as $T_k = \theta_k + \varepsilon_k,\ \varepsilon_k \sim N\left(0, s_k^2\right);\ \theta_k = \mu + \delta_k,\ \delta_k \sim N(0, \tau^2)$, where between-study variance τ^2 represents the additional source of variability between studies. A homogeneity test can be performed to test whether τ^2 is zero or not, and determine the appropriateness of FEM or REM. Like FEM, REM also prioritizes concordant genes but with more flexibility across studies. Neither of FEM nor REM produces biomarker categorization results.

Remarks

P-value combination methods are powerful for detecting genes that have non-zero effects in at least one study (HS_B alternative hypothesis setting as in Chang et al. (2013) without considering the magnitudes and directionality of effects across studies. Thus, *p*-value methods cannot distinguish concordant genes (i.e., upregulated or downregulated in all studies) from discordant

genes (i.e., upregulated in some studies but downregulated in others). In contrast, effect size combination methods take directionality into account but favor only concordant genes. Even so, discordant genes can still be of interest in, for example Pan-cancer analysis, to understand between tumor heterogeneity. We, therefore, propose a new meta-analysis method that incorporates both *p*-value and effect size combination methods, and considers concordant pattern as well as biological and statistical significance simultaneously to assist biomarker detection and categorization. Here we will introduce our method namely BCMC (**B**iomarker **C**ategorization in **M**eta-analysis by **C**oncordance).

New Meta-Analysis Method for Biomarker Detection and Categorization

Suppose there are K transcriptomic studies, each study k ($1 \leq k \leq K$) measures the gene expression of n_k samples and G genes. We use gene expression as example to introduce our method though the method is ready to analyze other types of transcripts such as miRNA and lncRNA. Our objective in meta-analysis is to detect candidate genes differentially expressed between the case (e.g., patients diagnosed with disease) and control (e.g., healthy subjects) group in multiple studies and categorize the detected genes by their DE patterns across studies. We first perform DE analysis using popular methods such as limma (Ritchie et al., 2015) for microarray or DESeq2 (Love et al., 2014) for RNA-seq in each study and obtain the summary statistics including effect size estimates (log2 fold change or LFC_{gk}) and p-values (p_{gk}) for each gene g ($1 \leq g \leq G$) in each study k. Effect sizes and p-values represent biological and statistical significance, respectively, and can be treated as DE evidence for single studies. The smaller the p-value and the larger the magnitude of effect size, the more likely a gene will be a DE gene in the study. In meta-analysis, concordance (i.e., a gene having the same sign of effect size in different studies) is regarded as additional piece of DE evidence. We define gth gene as being up-regulated in kth study when $\mathrm{LFC}_{\mathrm{gk}} > 0$ (i.e., having higher expression in case group) and being down-regulated when $\mathrm{LFC}_{\mathrm{gk}} < 0$ (i.e., having higher expression in control group).

When integrating multiple transcriptomic studies, DE genes may be altered in study-specific patterns. For example, some genes are differentially expressed in all studies while others are only differentially expressed in specific subset of studies. Meta-analysis methods also have different groups of targeted biomarkers as reflected by different statistical hypothesis settings. The null hypothesis for each gene in meta-analysis is commonly defined as: $H_0: \theta_{\mathrm{g1}} = \cdots = \theta_{\mathrm{gK}} = 0$, where θ_{gk} represents the true effect of gene g in study k. Depending on the types of targeted biomarkers, three alternative hypotheses have been proposed in the meta-analysis literature (Birnbaum, 1954; Tseng et al., 2012; Song and Tseng, 2014). The first setting (HS_A) aims to detect DE genes that have non-zero effect in all studies, i.e., $\theta_{\mathrm{gk}} \neq 0$ for all k. The second setting (HS_B) aims to detect DE genes that have non-zero effect in at least one study, i.e., $\theta_{\mathrm{gk}} \neq 0$ for some k. The third setting (HS_r) aims to detect DE genes that have non-zero effect in at least r studies, i.e., $\sum_{\mathrm{k}=1}^{K} I\left\{\theta_{\mathrm{gk}} \neq 0\right\} \geq r$. As we show next, our method generally follows HS_r setting with specifically $r = 2$ (i.e., we detect DE genes that have non-zero effect in at least two studies).

To detect DE genes and categorize them by cross-study DE patterns, we propose the following two aggregate statistics for each gene that combines DE evidence across up-regulated studies or down-regulated studies, respectively:

$$T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})} = \frac{\sum_{LFC_{\mathrm{gk}}>0;\ LFC_{\mathrm{gk'}}>0;\ k\neq k'} (w^{+}_{\mathrm{gk}} w^{+}_{\mathrm{gk'}} LFC_{\mathrm{gk}} LFC_{\mathrm{gk'}} |log_{10} p_{\mathrm{gk}} + log_{10} p_{\mathrm{gk'}}|)}{\sum_k w^{+}_{\mathrm{gk}}}$$

$$T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})} = \frac{\sum_{LFC_{\mathrm{gk}}<0;\ LFC_{\mathrm{gk'}}<0;\ k\neq k'} (w^{-}_{\mathrm{gk}} w^{-}_{\mathrm{gk'}} LFC_{\mathrm{gk}} LFC_{\mathrm{gk'}} |log_{10} p_{\mathrm{gk}} + log_{10} p_{\mathrm{gk'}}|)}{\sum_k w^{-}_{\mathrm{gk}}},$$

where w^{+}_{gk} and w^{-}_{gk} are binary weights of 0 or 1 assigned to the kth study for gth gene, indicating whether a study is selected for inclusion in aggregate statistics or not, +/− indicate upregulation or downregulation part, $\overrightarrow{w}^{+}_{g} = \left(w^{+}_{\mathrm{g1}}, \ldots, w^{+}_{\mathrm{gK}}\right)$ and $\overrightarrow{w}^{-}_{g} = \left(w^{-}_{\mathrm{g1}}, \ldots, w^{-}_{\mathrm{gK}}\right)$. LFC_{gk} is the $\log_2$ fold change and p_{gk} the corresponding p-value for gene g in study k obtained from single study DE analysis.

For gth gene, $T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})}$ aggregates the information of single study summary statistics (including both p-value and effect size) over up-regulated studies (i.e., those studies with $LFC_{\mathrm{gk}} > 0$), while $T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})}$ aggregates that over down-regulated studies (i.e., those studies with $LFC_{\mathrm{gk}} < 0$). The binary weights are used to indicate what studies to include to the aggregate statistics and the optimal weights that maximize the statistics will be searched for each gene. In the proposed aggregate statistics, we simultaneously account for concordant patterns (where LFC_{gk} and $LFC_{\mathrm{gk'}}$ have the same sign), biological significance (estimated as the product of LFC_{gk}) and statistical significance [estimated as the sum of $\log_{10}(p_{\mathrm{gk}})$]. This will encourage combining studies with the same directionality to find the best evidence for DE, which is consistent with the purpose of meta-analysis to identify more reproducible genes in multiple studies. Similar statistics have been proposed for concordant and discordant analysis of orthologous genes between a pair of species (Domaszewska et al., 2017). From the formula, we can see that the proposed statistic is essentially a weighted average of all study pairs with effect sizes in the same direction. A weighted average of all studies instead of study pairs is an alternative approach but it tends to exclude studies with moderate effect sizes or p-values (see a toy example in **Supplementary Table 1**).

By default, we assume $w^{+}_{\mathrm{gk}} = 0$ for studies with $LFC_{\mathrm{gk}} < 0$ and $w^{-}_{\mathrm{gk}} = 0$ for $LFC_{\mathrm{gk}} > 0$ to avoid conflict between the two statistics. When no studies are up-regulated or down-regulated for a particular gene, we suppress the corresponding $T^{+}_{\mathrm{g}(\overrightarrow{w}^{+}_{\mathrm{g}})}$ or $T^{-}_{\mathrm{g}(\overrightarrow{w}^{-}_{\mathrm{g}})}$ to zero and assign zero weights. The statistics aggregates over study pairs so we need to choose at least two studies to

make it meaningful. When only one study is up-regulated or down-regulated, we also suppress the corresponding $T^{+}_{g(\vec{w}^{+}_{g})}$ or $T^{-}_{g(\vec{w}^{-}_{g})}$ to zero.

We then search for the optimal weights to identify the subset of studies that maximize each of the two aggregate statistics. Such optimal weights describe the DE patterns of each gene across studies and provide natural categorization of all genes with potential biological interpretation. The corresponding maximum statistics are defined as:

$$R^{+}_{g} = \max_{\vec{w}^{+}_{g} \in W} T^{+}_{g(\vec{w}^{+}_{g})}; R^{-}_{g} = \max_{\vec{w}^{-}_{g} \in W} T^{-}_{g(\vec{w}^{-}_{g})},$$

where W is the pre-defined searching space of weights with aforementioned restrictions. The resulting optimal weights are denoted as $\vec{w}^{+*}_{g}$ and $\vec{w}^{-*}_{g}$. The biomarkers are then categorized according to the distribution of optimal weights among studies by merging the information of w^{+*}_{g} and w^{-*}_{g}, i.e., the final weights $\vec{w}^{*}_{g} = \vec{1} \circ \vec{w}^{+*}_{g} + \vec{-1} \circ \vec{w}^{-*}_{g}$ For example, concordantly up-regulated genes with $\vec{w}^{+*}_{g}$ = (0, 0, 1, 1, 1) and $\vec{w}^{-*}_{g}$ = (0,0,0,0,0) will be in one category [$\vec{w}^{*}_{g}$ = (0, 0, 1, 1, 1)], while concordantly down-regulated genes with $\vec{w}^{+*}_{g}$ = (0, 0, 0, 0, 0) and $\vec{w}^{-*}_{g}$ = (0,0,1,1,1) will be in the other category [$\vec{w}^{*}_{g}$ = (0, 0, −1, −1, −1)]. Note that the proposed statistics can describe both up-regulated and down-regulated patterns in the same gene, thus also allowing the detection of discordant genes. In cases both patterns exist and we want to find a dominant pattern in the discordant gene, we can further define R_g = max (R^{+}_{g}, R^{-}_{g}) and use the corresponding $\vec{w}^{+*}_{g}$ or $\vec{w}^{-*}_{g}$ for biomarker categorization.

To assess the uncertainty of R^{+}_{g} and R^{-}_{g} and determine DE in meta-analysis, we develop a permutation-based test to calculate the p-value and FDR adjusted p-value (also known as q-value) of the statistics. We permute group labels (i.e., case or control group) in each study B times and calculate the maximum statistics in each permuted dataset. For each gene, we obtain two p-values corresponding to R^{+}_{g} and R^{-}_{g}, respectively:

$$p^{+}_{g_{(R^{+}_{g})}} = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left\{R^{+(b)}_{g'} \geq R^{+}_{g}\right\} + 1}{B * G + 1};$$

$$p^{-}_{g_{(R^{-}_{g})}} = \frac{\sum_{b=1}^{B} \sum_{g'=1}^{G} I\left\{R^{-(b)}_{g'} \geq R^{-}_{g}\right\} + 1}{B * G + 1},$$

where $R^{+(b)}_{g'}$ and $R^{-(b)}_{g'}$ are the maximum statistics for gth gene in bth ($1 \leq b \leq B$) permutation. The value of one is added to both numerator and denominator to avoid zero p-values. After p-values are generated, we further estimate the proportion of null genes π_0 as:

$$\hat{\pi}^{+}_{0} = \frac{\sum_{g=1}^{G} I\{p^{+}_{g_{(R^{+}_{g})}} \epsilon A\}}{G * \ell(A)}; \hat{\pi}^{-}_{0} = \frac{\sum_{g=1}^{G} I\{p^{-}_{g_{(R^{-}_{g})}} \epsilon A\}}{G * \ell(A)},$$

normally we choose $A = [0.5, 1]$ and $\ell(A) = 0.5$ to estimate the null proportion, following the guidance in the previous methods and the literature of FDR (Storey, 2002; Storey and Tibshirani, 2003; Li and Tseng, 2011). In most cases, the density of p-values beyond 0.5 is fairly flat, implying most null p-values are located in this region. In practice, depending on the problem, other common choices of A = [0.05,1] or A = [0.025,1] can also be applied. The optimal A can be empirically determined by minimizing some loss function, we do not discuss further here and refer readers to Storey (2002), Storey and Tibshirani (2003) for more details.

Then, q-values can be calculated as

$$q^{+}_{g_{(R^{+}_{g})}} = \frac{\hat{\pi}^{+}_{0} \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left\{R^{+(b)}_{g'} \geq R^{+}_{g}\right\} + 1}{B * \sum_{g'=1}^{G} I\left\{R^{+}_{g'} \geq R^{+}_{g}\right\} + 1},$$

$$q^{-}_{g_{(R^{-}_{g})}} = \frac{\hat{\pi}^{-}_{0} \sum_{b=1}^{B} \sum_{g'=1}^{G} I\left\{R^{-(b)}_{g'} \geq R^{-}_{g}\right\} + 1}{B * \sum_{g'=1}^{G} I\left\{R^{-}_{g'} \geq R^{-}_{g}\right\} + 1}$$

Likewise, p-value and q-value of the dominant pattern statistics R_g (i.e., $p_{g(R_g)}$ and $q_{g(R_g)}$) can be obtained in the same way. In real data application, we determine DE in meta-analysis using the permuted p-value or q-value for the dominant pattern. Note that p-values and q-values of a zero R^{+}_{g} or R^{-}_{g} are equal to one.

Downstream Analysis on Each Identified Categories of Biomarkers

Each transcriptomic study was carefully assessed for inclusion to meta-analysis using objective criteria or systematic quality control methods (Kang et al., 2012). When only expression of mRNA data is available for the K selected transcriptomic studies, we applied our meta-analysis and identified multiple categories of mRNAs at certain BCMC p-value or q-value cutoffs, each with a unique DE pattern across the studies. DE analysis is useful to narrow down targets but focusing on single gene change across datasets is not sufficient. We still need to conduct further investigation on whether mRNAs belonging to the same category contain unifying biological theme. For each unique category of mRNAs, we then performed pathway enrichment analysis to gain more insights into their unique functions (section "Pathway Enrichment Analysis of mRNA Expression"). When expression data of mRNA, miRNA and lncRNA are all available, we applied our meta-analysis method to each type of transcripts separately and then analyzed each unique category of differentially expressed mRNA, miRNA, and lncRNA (those with the same weight or same cross-study DE pattern) together. Specifically, we performed miRNAs/lncRNAs target gene enrichment analysis (section "miRNAs/lncRNAs Target Gene Enrichment Analysis") and LncRNA-mRNA and miRNA-mRNA causal regulatory network analysis (section "LncRNA-mRNA and miRNA-mRNA Causal Regulatory Network Analysis").

Pathway Enrichment Analysis of mRNA Expression

For each category of mRNAs with unique DE pattern across the studies, we looked for biological pathways that are enriched in each category of genes more than would be expected by chance. The enriched pathways for each category can infer the unique biological functions only associated with specific study subsets and help generate new hypotheses. The *p*-value for the enrichment of a pathway was calculated using Fisher's exact test (Upton, 1992) and multiple testing was corrected by Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995). Multiple popular pathway databases were used including Gene Ontology (GO) (Ashburner et al., 2000), Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2017), Oncogenic signaling Pathways (Sanchez-Vega et al., 2018) and Reactome (Fabregat et al., 2016). Pathways in each pathway database was carefully selected for their relatedness to the problem of interest and small pathways (e.g., pathway size <10) were filtered out for the lack of power. For pathways with topological information available (e.g., pathways in KEGG), we apply the R package *"Pathview"* (Luo and Brouwer, 2013), to display the study-specific information (e.g., weights, effect sizes, etc.) on relevant pathway topology graphs.

miRNAs/lncRNAs Target Gene Enrichment Analysis

Going beyond the traditional central dogma, non-coding RNAs such as micro-RNA (or miRNA) and long non-coding RNAs (lncRNA) play important regulatory roles in mRNAs expression (Bartel, 2004; Hubé and Francastel, 2018). To understand whether miRNA/lncRNA target at mRNAs in the same category with unique cross-study DE pattern, we analyzed each unique category of mRNA, miRNA and lncRNA of the same cross-study DE pattern together and performed miRNA/lncRNAs target gene enrichment analysis on each category. Specifically, for each unique category, we first used the miRTarBase database (Chou et al., 2018) and LncRNA2Target v2.0 database (Cheng et al., 2019) to obtain common target genes of each miRNA and lncRNA in this category. We then looked for miRNA/lncRNA with target genes enriched in the gene list falling in the same category more than would be expected by chance. The *p*-value for the enrichment of miRNA/lncRNA was calculated using Fisher's exact test (Upton, 1992) and multiple testing was corrected by BH procedure (Benjamini and Hochberg, 1995).

LncRNA-mRNA and miRNA-mRNA Causal Regulatory Network Analysis

In addition to target gene enrichment analysis, we are also interested in investigating the causal regulatory relationship among the various types of transcripts in the same category using network analysis. For each unique category of mRNA and lncRNA with the same cross-study DE pattern, we followed the MSLCRN pipeline to perform module-specific lncRNA-mRNA regulatory network analysis (Zhang et al., 2019). The MSLCRN pipeline starts by using WGCNA (Langfelder and Horvath, 2008) to construct lncRNA-mRNA co-expression networks and identify modules that contain both lncRNA and mRNA. For each lncRNA-mRNA module, parallel IDA (Le et al., 2016) is then applied to learn the causal structure and estimate the causal effect of lncRNA on mRNA. IDA consists of two main steps. It first uses a parallel version of the PC algorithm (Spirtes et al., 2000; Kalisch and Bühlman, 2007; Le et al., 2016), commonly used approach for learning the causal structure of a Bayesian network, to obtain the directed acyclic graphs (DAGs) for each module. Then, the causal effect of lncRNAs on mRNAs (i.e., the lncRNA $\geq$ mRNA directed edges in the DAG) are estimated by applying do-calculus (Pearl, 2000), causal calculus that uses Bayesian conditioning to generate probabilistic formulas for the causal effect. Lastly, the module-specific causal regulatory networks are integrated to form the global lncRNA-mRNA causal regulatory network and visualized using Cytoscape (Shannon et al., 2003). In constructing the regulatory network, we use absolute values of the causal effects cutoffs to assess the regulatory strengths and confirm the regulatory relationships. More details on the use of MSLCRN to infer causal regulatory network can be found in Zhang et al. (2019). Module-specific miRNA-mRNA causal regulatory networks can be obtained in a similar way using the same tool.

SIMULATION

We conduct simulation studies to evaluate the performance of our method in biomarker detection and categorization when compared to AW-Fisher (Li and Tseng, 2011), FEM and REM methods (Choi et al., 2003). Only power is assessed for FEM and REM methods since they do not categorize biomarkers. We assume a total of $G = 2000$ genes expressed in $K = 5$ studies, each study has a total sample size of $n = 100$, evenly split into control and case groups $\left(n_{\text{case}} = n_{\text{control}} = \frac{n}{2} = 50\right)$. The details on how data are simulated are described below:

1. We generate 800 genes with 40 gene clusters (20 genes in each cluster) and another 1,200 genes that do not belong to any cluster. The cluster indexes for each gene g $\left(1 \leq g \leq 2000\right)$ is randomly sampled.
2. For genes in cluster c $(1 \leq c \leq 40)$ and study k $(1 \leq k \leq 5)$, we first generate a covariance matrix according to inverse Wishart distribution $\Sigma'_{ck} \sim W^{-1}(\Psi, 60)$, where $\Psi = 0.5I_{20\times 20} + 0.5J_{20\times 20}$, I is the identity matrix and J is the matrix with all elements equal to one. Then, we standardized Σ'_{ck} into Σ_{ck} to make sure all the diagonal elements are one.
3. We sample baseline gene expression levels of the 20 genes in cluster c for sample i in study k by $\left(X'_{\text{gc1ik}}, \ldots, X'_{\text{gc20ik}}\right)^T \sim MVN(0, \Sigma_{ck})$, where $1 \leq i \leq n$ and $1 \leq k \leq K$. For those 1200 genes that are not in any cluster, we sample the baseline gene expression level independently from $N\left(0, \sigma_k^2\right)$, where $1 \leq k \leq 5$ and $\sigma_k \sim Unif(\sigma - 0.2, \sigma + 0.2)$ with $\sigma = 2$.
4. Denote by $\delta_{gk} \in \{0, 1, -1\}$ that gene g is non-DE, up-regulated or down-regulated in study k. We assume the first 800 genes to be DE genes divided into four mutually exclusive parts:

   (1) Concordantly up-regulated genes ($N = 225$): randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k I_{\{\delta_{gk}=1\}} \geq 2$ and $\sum_k I_{\{\delta_{gk}=-1\}} \leq 1$.

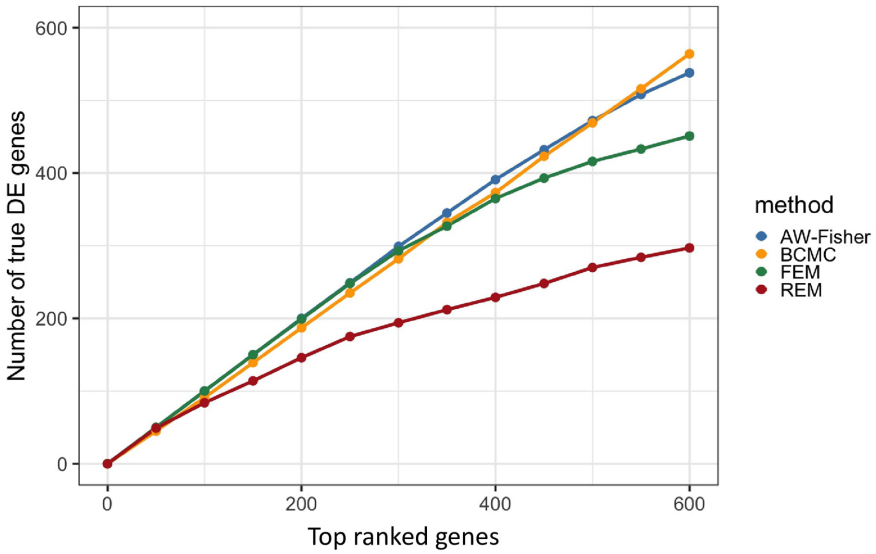

FIGURE 2 | Plot of the number of true DE genes vs. top ranked genes by *p*-value of each method.

(2) Concordantly down-regulated genes ($N = 225$): randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k I_{\{\delta_{gk}=-1\}} \geq 2$ and $\sum_k I_{\{\delta_{gk}=1\}} \leq 1$.

(3) Discordant genes with both up-regulated and down-regulated patterns ($N = 150$): randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k I_{\{\delta_{gk}=1\}} \geq 2$ and $\sum_k I_{\{\delta_{gk}=-1\}} \geq 2$.

(4) Other genes that are DE in only one study without any concordant patterns ($N = 200$): we randomly sample $\delta_{gk} \in \{0, 1, -1\}$ such that $\sum_k |\delta_{gk}| = 1$.

5. To simulate effect size for DE genes in each study (when $\delta_{gk} \neq 0$), we sample from a uniform distribution $\mu_{gk} \sim Unif(1, 3)$. The gene expression level X_{gik} are assumed to be X'_{gik} for control samples and $X_{gik} = X'_{g(i+n/2)k} + \mu_{gk} \cdot \delta_{gk}$ for case samples, where $1 \leq g \leq 2000$, $1 \leq i \leq n/2$, and $1 \leq k \leq 5$.

To assess power and biomarker categorization performance, we focus on DE genes in the first three categories of genes with concordant patterns in at least two studies ($N = 600$). We also simulate additional scenario with smaller sample size and variance: $n = 20$ & $\sigma = 1$, results are included in the Supplement (**Supplementary Figure 1** and **Supplementary Table 2**).

Figure 2 shows the number of true DE genes detected among the top genes ranked by *p*-value for each method. BCMC is more powerful than AW-Fisher and FEM/REM by detecting more true DE genes among the top ranked genes. **Table 1** summarizes the number of true DE genes detected as well as with correct weight pattern in each of the three categories of DE genes identified by each method. BCMC and FEM detect more true DE genes than AW-Fisher for concordant genes. Due to the model restriction, FEM and REM fail to detect most discordant genes. AW-Fisher is equally powerful as BCMC in detecting discordant genes, however, it ignores the directionality of effects, and thus assigns the incorrect weights to genes with both up-regulated and down-regulated patterns (basically they fail to distinguish $w = -1$ from $w = 1$). Our method detects these discordant DE genes while at the same time assigns the correct weights categorizing these genes.

REAL DATA APPLICATION

Gene Expression Analysis in Pan-Gynecologic (Pan-Gyn) Studies

We applied our method to the gene expression data of TCGA Pan-Gyn studies including high-grade serous ovarian cystadenocarcinoma (OV), uterine corpus endometrial carcinoma (UCEC), cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), uterine carcinosarcoma (UCS), and invasive breast carcinoma (BRCA) (Berger et al., 2018). Berger et al. (2018) identified 23 genes (e.g., BRCA1, PTEN, TP53, etc.) that were mutated at higher frequency across all Pan-Gyn cancers than non-Gyn cancers, highlighting the similarities across Pan-Gyn cohort. We focused on 19 of these genes and split samples in each study into a mutation "carrier" group and a mutation "non-carrier" group depending on whether subjects gained mutations in at least one of the genes (**Supplementary Figure 2**). Since no or very few samples were assigned to the mutation carrier group for UCS ($N_{mutation} = 0$) and UCEC ($N_{mutation} = 8$), we excluded those two studies and restricted our meta-analysis to only three gynecologic cancer types (i.e., number of studies $K = 3$) including OV (mutation carrier vs. non-carrier: 217/90), BRCA (692/408) and CESC (109/197). The purpose is to detect differentially expressed genes

TABLE 1 | Summary of number of true DE genes detected and with correct weight patterns by the four methods in each of the three categories of DE genes described in the simulation setting.

Methods	BCMC		AW-Fisher		FEM	REM
DE Gene categories	Number of true DE genes	Number of true DE genes with correct weight	Number of true DE genes	Number of true DE genes with correct weight		
Concordant up (N = 225)	206	116	195	106	203	151
Concordant down (N = 225)	210	119	195	108	201	144
Discordant (N = 150)	148	135	148	0	47	2
Total (N = 600)	564	370	538	214	451	297

between mutation carrier and non-carrier groups and categorize them according to their cross-study DE patterns. We found the overall survival differed significantly between the two groups for each cancer type (**Supplementary Figures 3–5**). This implied the differentially expressed biomarkers between these two groups can have potential prognostic values related to mutational processes and serve as optimal therapeutic intervention targets (Helleday et al., 2014; Lawrence et al., 2014).

The RNA-seq data in Transcripts Per Million (TPM) values of each cancer type were downloaded from LinkedOmics (Vasaikar et al., 2018). We first merged the three datasets by matching the gene symbols and removed genes with mean TPM < 5. A total of 9,900 mRNAs remained and were $\log_2$ transformed for analysis. We performed DE analysis by limma (Ritchie et al., 2015) and obtained the p-value and LFC from each of the three studies. We then performed meta-analysis using BCMC and the other methods.

All methods detected thousands of DE genes at both q-value cutoffs (for BCMC, q-value for dominant pattern was used so we focused on concordant genes only), which is common in Pan-cancer studies (**Table 2**). It becomes imperative task to partition these DE genes into smaller subsets by cross-study DE patterns before performing downstream analysis. BCMC categorized these DE biomarkers ($q < 0.05$) into eight groups according to the optimal weight assignments, each displaying a unique expression pattern across the different studies (**Figure 3** and **Supplementary Table 3**). We then merged genes with equal $|\vec{w}_g^*|$ into the same group (i.e., genes with $\vec{w}_g^* = (0, 1, 1)$ and those with $\vec{w}_g^* = (0, -1, -1)$ are merged into the same group, allowing both up-regulated and down-regulated genes in the same pathway) and performed pathway enrichment analysis on each of the four merged groups using four pathway databases: GO (Ashburner et al., 2000), KEGG (Kanehisa et al., 2017), Oncogenic (Sanchez-Vega et al., 2018) and Reactome (Fabregat et al., 2016). The top 100 pathways enriched by each category have little overlap partly validating our speculation in motivation that the different categories of biomarkers may play different functional roles (**Figure 4**). For example, top pathways for $|\vec{w}_g^*| = (1, 0, 1)$ (i.e., DE in OV and CESC but not in BRCA) are mainly involved in cell junction and adhesion related functions (**Supplementary Table 4** in **Supplemental File 1**). Top pathways for $|\vec{w}_g^*| = (1, 1, 0)$ (i.e., DE in OV and BRCA but not in CESC) are mainly involved in immune and defense response. **Figure 5** shows the topology of one example KEGG pathway "Antigen processing and presentation" enriched by the genes with $|\vec{w}_g^*| = (1, 1, 0)$. The highlighted DE genes showed strong DE signals (signed LFC) in OV and BRAC but not in CESC. These genes colocalized and interacted with each other as a functional unit inside the pathway.

These unique gene sets of different cross-cancer DE patterns and the associated pathways enriched help gain more insights into the homogeneous and heterogenous molecular mechanism of different Gynecologic cancer and assist the development of useful diagnostic and therapeutic strategies common or specific to cancer types. Understanding commonality and difference in drug targets can also guide the drug repurposing strategy in cancer drug development (Li et al., 2021).

TABLE 2 | Summary of numbers of DE genes detected by each method at different cutoffs for the Pan-Gyn study example. For BCMC, q-values for the dominant pattern are used.

Methods				
q-value	BCMC	AW-Fisher	FEM	REM
$q < 0.05$	1,345	3,113	2,866	983
$q < 0.15$	3,931	4,743	4,342	1,641

Integrative Analysis of mRNA, lncRNA, and miRNA in Pan-Kidney Studies

We also used BCMC to perform integrative analysis of three different types of transcripts (mRNA, lncRNA, and miRNA) in the TCGA Pan-Kidney cohort including kidney chromophobe (KICH), kidney renal clear cell carcinoma (KIRC), and kidney renal papillary cell carcinoma (KIRP). LncRNA and miRNA have been found playing important regulatory roles on gene expression in kidney cancers (Linehan et al., 2010; Linehan, 2012; Ricketts et al., 2018). The integrative analysis of these multi-omics data provides additional insights into the biological mechanism underlying the multiple histologic subtypes of kidney cancers. We aimed to detect the differentially expressed biomarkers (mRNA, miRNA, or lncRNA) that drive the progression of kidney cancer by comparing samples from early pathologic stage (stage I and II) to late stage (stage III and stage IV) for three kidney cancer types (i.e., number of studies $K = 3$) and investigating the regulatory relationships among these biomarkers. Number of subjects in the two pathologic stages of each kidney cancer available in mRNA, miRNA and lncRNA expression data were summarized in **Supplementary Table 5**.

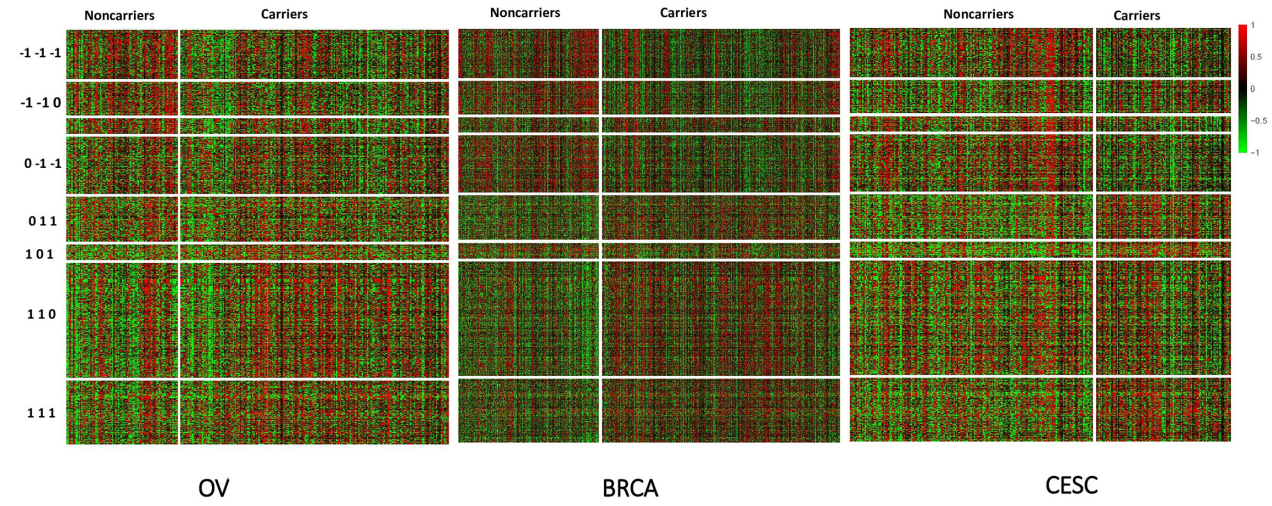

FIGURE 3 | Heatmap of standardized expression values of differentially expressed genes (BCMC $q < 0.05$) sorted by weight patterns for the Pan-Gyn cancer example.

We downloaded mRNA (in Reads Per Kilobase of transcript per Million mapped reads or RPKM) and miRNA (in Reads Per Million mapped reads or RPM) sequencing data from LinkedOmics (Vasaikar et al., 2018) and lncRNA sequencing data (in RPM) from The Atlas of Noncoding RNAs in Cancer (TANRIC) (Li et al., 2015) for all the three kidney cancer subtypes. We first merged the three subtypes by matching RNA symbols/IDs. We then separately filtered each of the three types of biomarkers by removing mRNAs with mean RPKM < 5, lncRNAs with mean RPM < 0.1, and miRNAs with mean RPM $= 0$, followed by $\log_2$ transformation. A total of 15,332 mRNAs, 2,415 lncRNAs and 719 miRNAs remained for analysis. We performed DE analysis by limma (Ritchie et al., 2015) in each study and then meta-analysis to categorize biomarkers according to cross-study DE patterns for each RNA species. For different types of RNA belonging to the same category, we further performed miRNA target gene enrichment analysis and lncRNA-mRNA causal regulatory network analysis to understand their complex interacting relationships in kidney cancer.

Both BCMC and AW-Fisher methods detected thousands of differentially expressed biomarkers (including mRNA, lncRNA, and miRNA) at both q-value cutoffs with high proportion of overlap (**Table 3**). Biomarkers detected by BCMC tend to have both significant p-values and large effect sizes in the studies indicated by optimal weights (**Supplementary Figure 6**). These biomarkers ($q < 0.05$) were partitioned into eight categories by different weight patterns (**Supplementary Table 6**). We merged biomarkers with the same $|\vec{w}_g^*|$ into the same group. We focused on the group with $|\vec{w}_g^*| = (1, 1, 1)$ to understand the common multi-omics regulatory among all histologic subtypes of kidney cancer and performed downstream analysis. In miRNA target gene enrichment analysis, we found the target gene sets of two DE miRNAs "miR-655" and "miR-326" were enriched in the DE gene list in the same group ($p < 0.05$; **Supplementary Table 7** in the **Supplementary File 1**), implying the potential regulatory relationship between different biomarker types consistent in all kidney cancer subtypes. The gene *ATAD2* targeted by miR-655 was reported as a prognostic marker for kidney disease (Chen et al., 2017). In causal network analysis, we identified two lncRNA-mRNA regulatory networks (**Supplementary Figure 8** and **Supplementary Table 8**). **Figure 6** shows the network with two hub lnRNAs, the hub lncRNA ENSG00000267449 and several mRNAs belonging to the ribosomal protein family in the same network were found consistently differentially expressed in all three subtypes, implying their potentially joint role in promoting the development of kidney cancers (Zhou et al., 2015; Dolezal et al., 2018).

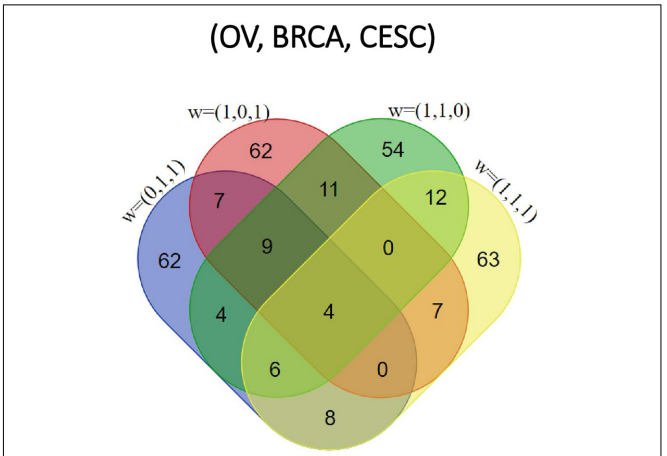

FIGURE 4 | Venn diagram of top 100 pathways enriched by each of the four categories [$|w_g^*| = (0, 1, 1), (1, 0, 1), (1, 1, 0)$, and $(1, 1, 1)$; corresponding to OV, BRCA and CESC, respectively] for the Pan-Gyn study example.

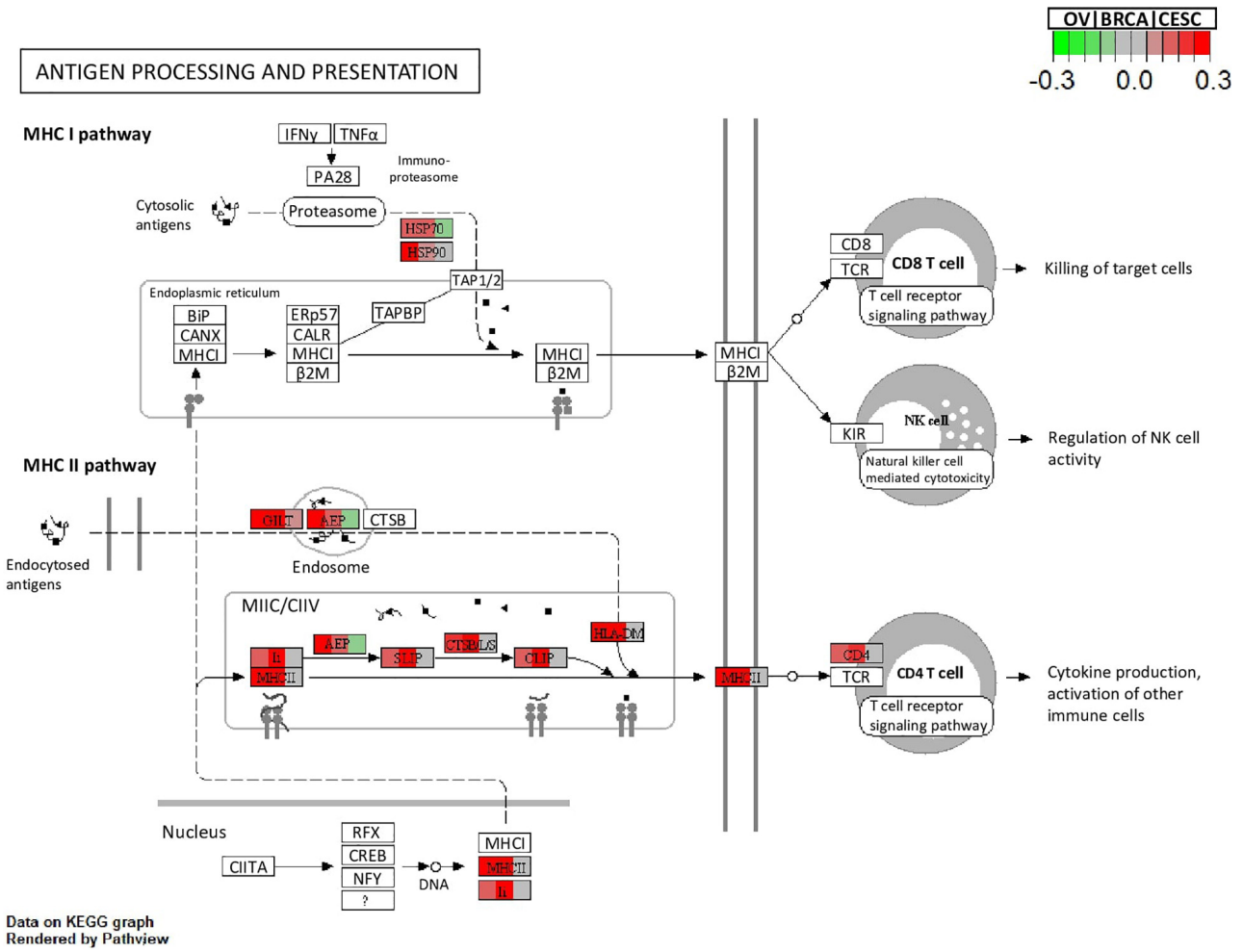

FIGURE 5 | Visualization of the topology plot of a KEGG pathway "Antigen processing and presentation" enriched by the genes with $|w_g^*| = (1, 1, 0)$ (corresponding to OV, BRCA and CESC, respectively) for the Pan-Gyn example. Each box that represents a gene is split into three parts to represent the three studies. Colors indicate the signed LFC of the mapped DE genes in the three studies.

These results demonstrate the power of our method to detect biomarkers of different types in Pan-cancer meta-analysis and to categorize them into functionally relevant biomarkers by DE patterns, which could suggest commonalities and differences in underlying mechanisms of multiple cancer types.

DISCUSSION

In this paper, we proposed a novel meta-analysis method for candidate biomarker detection in multiple transcriptomic studies that further categorizes biomarkers by concordant patterns as well as by biological and statistical significance across studies.

TABLE 3 | Summary of number of differentially expressed biomarkers among each of the three RNA species detected by each method at different cutoffs for the Pan-Kidney study example. For BCMC, *q*-values for the dominant pattern are used.

Type of biomarkers	**mRNA**		**lncRNA**		**miRNA**	
***q*-value**	**BCMC**	**AW-Fisher**	**BCMC**	**AW-Fisher**	**BCMC**	**AW-Fisher**
$q < 0.05$	7,317	9,472	764	1,281	239	283
Intersection	6,391		622		206	
$q < 0.15$	11,810	11,440	1,468	1,464	358	358
Intersection	10,057		1,244		292	

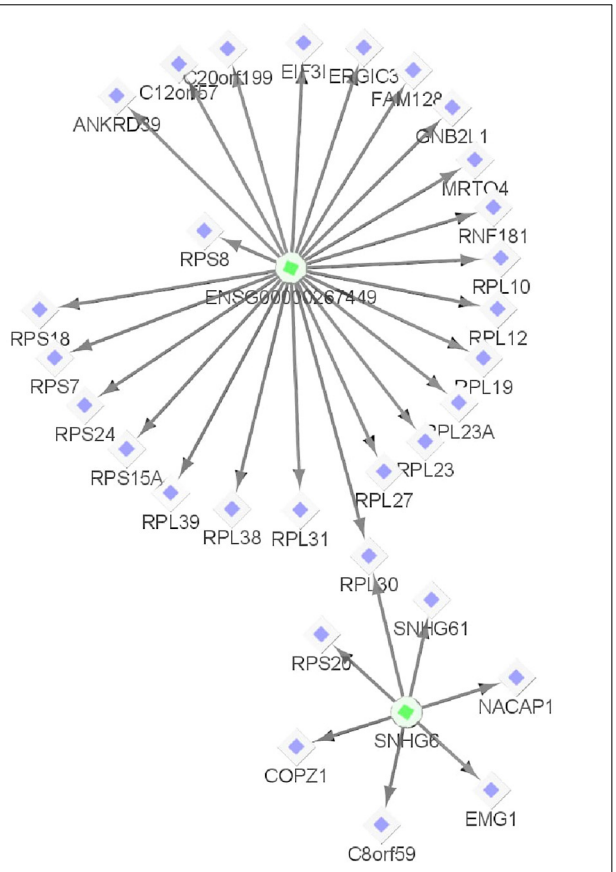

FIGURE 6 | One example lncRNA-mRNA regulatory network identified from biomarkers with $|w_g^*| = (1, 1, 1)$ (corresponding to KICH, KIRC, and KIRP, respectively) for the Pan-Kidney example. The circle shapes represent lncRNAs highlighted in green and diamond shapes represent mRNAs highlighted in purple. The arrows indicate the network relationships between lncRNAs and mRNAs.

Numerous downstream analysis tools including pathway analysis and causal network analysis are applied to each category of biomarkers with either single or multiple types of RNA species. Simulations and real data application to two Pan-cancer multi-omics studies showed the advantage of our method in classifying differentially expressed biomarkers into classes with unique biological functions and relationships that can be further investigated in future studies.

Meta-analysis is a set of statistical analytical methods and tools that combine multiple related studies to improve power and reproducibility over a single study. In recent years, we have witnessed the development of many useful meta-analysis methods applied to genomic studies for different biological purposes (Choi et al., 2003; Shen and Tseng, 2010; Li and Tseng, 2011; Huo et al., 2016, 2020; Kim et al., 2016, 2018; Zhu et al., 2017; Ma et al., 2019; Zeng et al., 2020). Genomic data is usually of high dimension and the between study heterogeneity is large due to both technological and cohort effects. In addition to improving power, post-hoc categorization of biomarkers into smaller subsets by cross-study patterns for subsequent analysis is important in genomic meta-analysis. Our meta-analysis method that aggregates over both *p*-value and effect size is a fast and intuitive solution for this purpose. Compared to other popular meta-analysis methods that include biomarker categorization, our method considers concordant pattern, and biological and statistical significance simultaneously. By calculating statistics separately for up-regulated and down-regulated parts, we can detect both concordant genes that have consistent patterns across all studies and discordant genes that are up/down regulated in some studies while down/up regulated in others. Both of these kinds of genes can be of interest in Pan-cancer analysis. For example, high expression of some genes might worsen the prognosis of all cancer types, while high expression of other genes might worsen prognosis for some cancers but be beneficial to other cancer types.

Our method also applies to the scenario when there is more than one RNA species present and proposes to jointly analyze different types of biomarkers under the same category for more biological insights. As more omics data are accumulated in the public domain, similar strategies can be applied for integrative analysis, for example with epigenomic (e.g., DNA methylation, histone modification), proteomic and metabolomic data. Unique features of each omics data type need to be addressed and will be considered as a future direction to extend our method.

Like most other two-stage meta-analysis methods, our method is based on summary measures such as *p*-values and $\log_2$ fold changes from each study. In addition, the method assigns a single optimal weight to each gene without quantifying the uncertainty in weight assignment. A more comprehensive Bayesian hierarchical model can be applied to raw data and summary measures to better capture the stochasticity and provide soft weight assignment. Our method requires the DE genes to be concordant in at least two studies to be detected, consistent with the purpose of meta-analysis in prioritizing more reproducible biomarkers. As the number of studies becomes large, the likelihood of being differentially expressed in only one study decreases. Thus, we expect the method to perform well as the number of studies increases. Since the method relies on summary measures, increasing the number of studies will not materially increase the computational burden. Additionally, use of more sophisticated parallel computing techniques will improve the speed of permutation tests. An R package called "BCMC" is available at https://github.com/kehongjie/BCMC to implement our method.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://github.com/kehongjie/BCMC.

AUTHOR CONTRIBUTIONS

ZY and HK developed the method, performed the analysis, and wrote the manuscript. TM supervised the project and took

the lead in editing the manuscript. SC, RC-C, XH, JZ, JD, and DM contributed to manuscript writing and polishing. All authors provided critical feedback and helped shape the research, analysis and manuscript.

FUNDING

Research reported in this publication was supported by the National Institute on Drug Abuse (NIDA) of National Institute of Health under award number 1DP1DA048968-01 to SC and TM, and the University of Maryland Faculty-Student Research Award (FSRA) to ZY, HK, and TM.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.651546/full#supplementary-material

REFERENCES

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297.

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res.* 40, 3777–3784. doi: 10.1093/nar/gkr1255

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B (Methodol.)* 57, 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x

Berger, A. C., Korkut, A., Kanchi, R. S., Hegde, A. M., Lenoir, W., Liu, W., et al. (2018). A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* 33, 690–705.e9.

Birnbaum, A. (1954). Combining independent tests of significance. *J. Am. Stat. Assoc.* 49, 559–574. doi: 10.2307/2281130

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 Bioinform.* 14:368. doi: 10.1186/1471-2105-14-368

Chen, D., Maruschke, M., Hakenberg, O., Zimmermann, W., Stief, C. G., and Buchner, A. (2017). TOP2A, HELLS, ATAD2, and TET3 are novel prognostic markers in renal cell carcinoma. *Urology* 102:265.e1–265.e7.

Cheng, L., Wang, P., Tian, R., Wang, S., Guo, Q., Luo, M., et al. (2019). LncRNA2Target v2. 0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* 47, D140–D144.

Choi, J. K., Yu, U., Kim, S., and Yoo, O. J. (2003). Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* 19, i84–i90.

Chou, C.-H., Shrestha, S., Yang, C.-D., Chang, N.-W., Lin, Y.-L., Liao, K.-W., et al. (2018). miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res.* 46, D296–D302.

Di Bella, S., La Ferlita, A., Carapezza, G., Alaimo, S., Isacchi, A., Ferro, A., et al. (2020). A benchmarking of pipelines for detecting ncRNAs from RNA-Seq data. *Brief. Bioinform.* 21, 1987–1998. doi: 10.1093/bib/bbz110

Di Camillo, B., Sanavia, T., Martini, M., Jurman, G., Sambo, F., Barla, A., et al. (2012). Effect of size and heterogeneity of samples on biomarker discovery: synthetic and real data assessment. *PLoS One* 7:e32200. doi: 10.1371/journal.pone.0032200

Dolezal, J. M., Dash, A. P., and Prochownik, E. V. (2018). Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. *BMC Cancer* 18:275. doi: 10.1186/s12885-018-4178-z

Domaszewska, T., Scheuermann, L., Hahnke, K., Mollenkopf, H., Dorhoi, A., Kaufmann, S. H., et al. (2017). Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis. *Sci. Rep.* 7:12094.

Fabregat, A., Sidiropoulos, K., Garapati, P., Gillespie, M., Hausmann, K., Haw, R., et al. (2016). The reactome pathway knowledgebase. *Nucleic Acids Res.* 44, D481–D487.

Fisher, R. A. (1992). "Statistical methods for research workers," in *Breakthroughs in Statistics*, eds S. Kotz and N. L. Johnson (New York, NY: Springer), 66–70.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nat. Rev. Genet.* 15, 585–598. doi: 10.1038/nrg3729

Hoadley, K. A., Yau, C., Hinoue, T., Wolf, D. M., Lazar, A. J., Drill, E., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell* 173, 291–304.e6.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics* 22, 2825–2827. doi: 10.1093/bioinformatics/btl476

Hubé, F., and Francastel, C. (2018). Coding and non-coding RNAs, the frontier has never been so blurred. *Front. Genet.* 9:140. doi: 10.3389/fgene.2018.00140

Huo, Z., Ding, Y., Liu, S., Oesterreich, S., and Tseng, G. (2016). Meta-analytic framework for sparse k-means to identify disease subtypes in multiple transcriptomic studies. *J. Am. Stat. Assoc.* 111, 27–42. doi: 10.1080/01621459.2015.1086354

Huo, Z., Song, C., and Tseng, G. (2019). Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. *Ann. Appl. Stat.* 13:340.

Huo, Z., Tang, S., Park, Y., and Tseng, G. (2020). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics* 36, 524–532. doi: 10.1093/bioinformatics/btz589

Kalisch, M., and Bühlman, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.* 8, 613–636.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 45, D353–D361.

Kang, D. D., Sibille, E., Kaminski, N., and Tseng, G. C. (2012). MetaQC: objective quality control and inclusion/exclusion criteria for genomic meta-analysis. *Nucleic Acids Res.* 40:e15. doi: 10.1093/nar/gkr1071

Kim, S., Kang, D., Huo, Z., Park, Y., and Tseng, G. C. (2018). Meta-analytic principal component analysis in integrative omics application. *Bioinformatics* 34, 1321–1328. doi: 10.1093/bioinformatics/btx765

Kim, S., Lin, C.-W., and Tseng, G. C. (2016). MetaKTSP: a meta-analytic top scoring pair method for robust cross-study validation of omics prediction analysis. *Bioinformatics* 32, 1966–1973. doi: 10.1093/bioinformatics/btw115

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 9:559. doi: 10.1186/1471-2105-9-559

Lawrence, M. S., Stojanov, P., Mermel, C. H., Robinson, J. T., Garraway, L. A., Golub, T. R., et al. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* 505, 495–501. doi: 10.1038/nature12912

Le, T., Hoang, T., Li, J., Liu, L., Liu, H., and Hu, S. (2016). "A fast PC algorithm for high dimensional causal discovery with multi-core PCs," in *Proceedings of the IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 16, (New York, NY: IEEE), 1483–1495. doi: 10.1109/tcbb.2016.2591526

Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., et al. (2015). TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.* 75, 3728–3737. doi: 10.1158/0008-5472.can-15-0273

Li, J., and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann. Appl. Stat.* 5, 994–1019.

Li, Y., Dong, Y.-P., Qian, Y.-W., Yu, L.-X., Wen, W., Cui, X.-L., et al. (2021). Identification of important genes and drug repurposing based on clinical-centered analysis across human cancers. *Acta Pharmacol. Sin.* 42, 282–289. doi: 10.1038/s41401-020-0451-1

Linehan, W. M. (2012). Genetic basis of kidney cancer: role of genomics for the development of disease-based therapeutics. *Genome Res.* 22, 2089–2100. doi: 10.1101/gr.131110.111

Linehan, W. M., Srinivasan, R., and Schmidt, L. S. (2010). The genetic basis of kidney cancer: a metabolic disease. *Nat. Rev. Urol.* 7:277. doi: 10.1038/nrurol.2010.47

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550.

Luo, W., and Brouwer, C. (2013). Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics* 29, 1830–1831. doi: 10.1093/bioinformatics/btt285

Ma, T., Huo, Z., Kuo, A., Zhu, L., Fang, Z., Zeng, X., et al. (2019). MetaOmics: analysis pipeline and browser-based software suite for transcriptomic meta-analysis. *Bioinformatics* 35, 1597–1599. doi: 10.1093/bioinformatics/bty825

Ma, T., Liang, F., and Tseng, G. (2017). Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using bayesian hierarchical models. *J. R. Stat. Soc. Ser. C Appl. Stat.* 66:847. doi: 10.1111/rssc.12199

Pearl, J. (2000). *Causality: Models, Reasoning and Inference*, Vol. 9. Cambridge, MA: Cambridge university press, 10–11.

Ramasamy, A., Mondry, A., Holmes, C. C., and Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med.* 5:e184. doi: 10.1371/journal.pmed.0050184

Richardson, S., Tseng, G. C., and Sun, W. (2016). Statistical methods in integrative genomics. *Annu. Rev. Stat. Appl.* 3, 181–209. doi: 10.1146/annurev-statistics-041715-033506

Ricketts, C. J., De Cubas, A. A., Fan, H., Smith, C. C., Lang, M., Reznik, E., et al. (2018). The cancer genome atlas comprehensive molecular characterization of renal cell carcinoma. *Cell Rep.* 23, 313–326.e5.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W. K., Luna, A., La, K. C., et al. (2018). Oncogenic signaling pathways in the cancer genome atlas. *Cell* 173, 321–337.e10.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shen, K., and Tseng, G. C. (2010). Meta-analysis for pathway enrichment analysis when combining multiple genomic studies. *Bioinformatics* 26, 1316–1323. doi: 10.1093/bioinformatics/btq148

Solla, F., Tran, A., Bertoncelli, D., Musoff, C., and Bertoncelli, C. M. (2018). Why a p-value is not enough. *Clin. Spine Surg.* 31, 385–388.

Song, C., and Tseng, G. C. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *Ann. Appl. Stat.* 8:777.

Spirtes, P., Glymour, C. N., Scheines, R., and Heckerman, D. (2000). *Causation, Prediction, and Search*. Cambridge, MA: MIT press.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* 64, 479–498. doi: 10.1111/1467-9868.00346

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100, 9440–9445. doi: 10.1073/pnas.1530509100

Stouffer, S. (1949). A study of attitudes. *Sci. Am.* 180, 11–15.

Sullivan, G. M., and Feinn, R. (2012). Using effect size-or why the P value is not enough. *J. Graduate Med. Educ.* 4, 279–282. doi: 10.4300/jgme-d-12-00156.1

Tseng, G. C., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.* 40, 3785–3799. doi: 10.1093/nar/gkr1265

Upton, G. J. (1992). Fisher's exact test. *J. R. Stat. Soc. Ser. A (Stat. Soc.)* 155, 395–402.

Vasaikar, S. V., Straub, P., Wang, J., and Zhang, B. (2018). LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res.* 46, D956–D963.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45:1113. doi: 10.1038/ng.2764

Zeng, X., Zong, W., Lin, C.-W., Fang, Z., Ma, T., Lewis, D. A., et al. (2020). Comparative pathway integrator: a framework of meta-analytic integration of multiple transcriptomic studies for consensual and differential pathway analysis. *Genes* 11:696. doi: 10.3390/genes11060696

Zhang, J., Le, T. D., Liu, L., and Li, J. (2019). Inferring and analyzing module-specific lncRNA–mRNA causal regulatory networks in human cancer. *Brief. Bioinform.* 20, 1403–1419. doi: 10.1093/bib/bby008

Zhou, X., Liao, W.-J., Liao, J.-M., Liao, P., and Lu, H. (2015). Ribosomal proteins: functions beyond the ribosome. *J. Mol. Cell Biol.* 7, 92–104. doi: 10.1093/jmcb/mjv014

Zhu, L., Ding, Y., Chen, C.-Y., Wang, L., Huo, Z., Kim, S., et al. (2017). MetaDCN: meta-analysis framework for differential co-expression network detection with an application in breast cancer. *Bioinformatics* 33, 1121–1129.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.