ABSTRACT

Title of Thesis:	BIOMARKER CATEGORIZATION IN TRANSCRIPTOME META-ANALYSIS BY STATISTICAL SIGNIFICANCE, BIOLOGICAL SIGNIFICANCE AND CONCORDANCE
	Zhenyao Ye, Master of Public Health, 2020
Thesis Directed By:	Assistant professor, Dr. Tianzhou (Charles) Ma, Department of Epidemiology and Biostatistics

With the advancement of high-throughput technology, transcriptomic studies have been accumulated in the public domain. Meta-analysis combines multiple studies on a related hypothesis and improves the statistical power and reproducibility of single studies. However, a majority of existing meta-analysis methods only consider the statistical significance. We propose a novel method to categorize biomarkers by simultaneously considering statistical significance, biological significance (large effect size) and concordance patterns across studies, accounting for the complex study heterogeneity that exists in most meta-analysis problems. We conducted simulation studies and applied our method to Gynecologic and breast cancer RNA-seq data from The Cancer Genome Atlas to show its strength as compared to adaptively-weighted Fisher's method. We found several major biomarker categories according to their cross-study patterns, and these categories are enriched in very different sets of pathways, offering different biological functions for future precision medicine.

BIOMARKER CATEGORIZATION IN TRANSCRIPTOME META-ANALYSIS BY STATISTICAL SIGNIFICANCE, BIOLOGICAL SIGNIFICANCE AND CONCORDANCE

by

Zhenyao Ye

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Public Health in Biostatistics
2020

Advisory Committee:
Dr. [Tianzhou (Charles) Ma], Chair
Dr. [Raul Cruz-Cano]
Dr. [Shuo Chen]
Dr. [Xin He]

Table of Contents

Chapter 1: Introduction

Review of Relevant Literature

Transcriptomics studies the complete set of RNA transcripts (both coding and non-coding genes) in individuals under specific circumstances using high-throughput technologies such as microarray and RNA sequencing (RNA-seq). The most common use of transcriptome profiling is to search for differentially expressed (DE) genes which show differences in expression level between two or more conditions (Soneson and Delorenzi, 2013). Over the years, with the advancement and more affordable price of high-throughput genomic technologies, plenty of data sets have been accumulated in public domains such as The Cancer Genome Atlas (TCGA) from NIH National Cancer Institute (https://portal.gdc.cancer.gov/), Gene Expression Omnibus (GEO) from NCBI (http://www.ncbi.nlm.nih.gov/geo/) and Sequence Read Archive (SRA) from NCBI (https://www.ncbi.nlm.nih.gov/sra).

Findings from single transcriptomic study are usually not reproducible because of limited sample size relative to large numbers of potential predictors (Ramasamy et al., 2008). Meta-analysis is a set of statistical techniques and tools to combine information from multiple and related research studies and will potentially increase reproducibility and validity of single studies. Horizontal omics meta-analysis aims to combine different sample cohorts of the same omics data types (e.g. gene expression, genetic variants) (Tseng et al., 2012). For DE gene detection, there are three major categories of horizontal meta-analysis methods: combining P-values, combining effect sizes and combining ranks (Tseng et al., 2012). Combining p-values from

individual study results is relatively simple approach and can accommodate different outcome types. For example, the famous Fisher's method (Fisher,1932) sums up log-transformed P-values obtained from differential expression analysis in individual studies. Instead of log-transformation, Stouffer's method (1949) alternatively adopts an inverse normal transformation. Li and Tseng (2011) extended Fisher's method and introduced an adaptively weighted Fisher's method (AW-Fisher) to indicate which studies contribute to the evidence aggregation and elucidates heterogeneity in meta-analysis. AW-Fisher searches all possible 0 or 1 weight for K individual study (a total of 2K -1 possibilities) to find the best adaptive weight with the smallest derived p-value of the statistics.

One limitation of combining p-values in meta-analysis method is that they only consider statistical significance (p-values) without considering the absolute magnitudes of effect size (a.k.a. fold change in genomic studies) in each study or the directionality of effect size across studies (i.e. concordance pattern). Fixed effects model (FEM) and random effects model (REM) are two most popular meta-analysis methods in combining effect sizes. Fixed effects model combines the estimated effect sizes from multiple studies assuming they share the same underlying true effect size plus measurement error in each study. Random effects model extends fixed effects model by allowing random effects for the inter-study heterogeneity in the model (Choi et al., 2003). Although random effects model can incorporate unknown inter-study heterogeneities, both fixed effects model and random effects model only allow one single concordant pattern (either up or down-regulated) for one gene across all studies and is usually underpowered when discordance exists.

Motivation

For genomics data, biological significance (large absolute value of effect sizes) and up-/down-regulated concordance are as important as the statistical significance (small p-values). Statistical significance is determined by magnitudes of effect size, variance of gene expression and sample size. On one hand, biological significance does not necessarily imply statistical significance when sample size is small and gene variation is large. For example, Gynecologic cancer is the fourth most common cancer in women and it has four different subtypes (National Cervical Cancer Coalition, 2019). One subtype, uterine carcinosarcoma (UCS), is a rare cancer that has relatively small sample size with limited statistical power. If we only consider statistical significance, many important genes would almost for sure be overlooked. On the other hand, statistical significance does not reflect biological significance when p-value is small simply due to large sample size or small variance. For example, large cohorts find many housekeeping genes to be statistically significant but those findings have minimal practical usage when the actual magnitude of effect size is too small.

One big challenge in genomic studies is how to interpret a large pool of DE gene findings. A common practice is to perform a pathway analysis by using Fisher's exact test to see in what functional domains these DE genes are enriched (Hoasck et al., 2003). However, for meta-analysis, top genes identified are of varying differential expression patterns across studies and biomarker categorization by their cross-study patterns become an important task. Ma et al. (2017) showed the biomarkers categorized by their differential expression patterns across studies are enriched in

different pathways, facilitating improved interpretation and biological hypothesis generation.

Overview of the Thesis

To address these gaps in knowledge, we develop a new approach to categorize the biomarkers by simultaneously considering statistical significance, biological significance and concordance patterns across the studies. We propose two measures in this approach for both up and down-regulated patterns in the same gene thus allow the existence of possible discordant patterns among all studies. We conduct two simulation scenarios assessing both weight patterns and gene ranking to demonstrate the strength of our method. As compared to AW-Fisher, the proposed method identifies the correct weight patterns for those genes with possible discordant and rank higher the DE genes with possible biological significance but marginal statistical significance. Our pilot study focuses on meta-analyzing the RNA-seq data from the following five TCGA tumor types (Pan-Gyn): 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). Since the underlying truth is not known in real data, we performed a pathway enrichment analysis to evaluate our approach and provide biological insights to the future precision medicine in cancer treatment. A total of 26 weight categories of different biological significance, statistical significance and concordance patterns across studies were identified by our method. Each weight category implied different functional domains of biological interest via pathway enrichment analysis.

Relevance to Public Health and Biostatistics

Heterogeneity is an essential factor to study and develop treatments for many diseases. For example, the risk of infecting COVID-19 is dependent on the characteristics of the infectious host, the susceptible host, and the environment. Similarly, the heterogeneity exists in transcriptomic data as a marked challenge of aggregating multiple studies. Simple aggregation of the data would overlook some critical information for investigating human diseases. Even if people know the existence of population heterogeneity, it is still commonly seen in real applications of precision medicine that patients having same type and stage of cancer are treated with the same treatment. In order to obtain more useful and credible information from meta-analysis while considering between study variation, we proposed sophisticated measures to categorize inter-study heterogeneous biomarkers by simultaneously considering biological significance (absolute magnitude of effect sizes), concordance (up-/down-regulated), and statistical significance (p-values) across multiple transcriptomic studies, to facilitate the development of precision medicine.

Chapter 2: Method

Popular meta-analysis methods

Microarray meta-analysis for DE gene detection is a commonly encountered application for integrating multiple studies on a related hypothesis and improving the statistical power and reproducibility of single studies. We will discuss two major types of statistical meta-analysis methods as follow: combining p-values and combining effect sizes.

1. *Combining p-values*

1A. Fisher's method

The combined Fisher's statistic $T_{Fisher} = -2\sum_{k=1}^{K}\log(P_k)$ follows a $\chi 2$ distribution with 2K degrees of freedom under the null hypothesis (i.e. genes are not differentially expressed in all studies), where K studies are combined and P_k is the p-value of study k, $1 \leq k \leq K$. Smaller p-values contribute larger scores to the Fisher's statistic.

1B. Stouffer's method

The Stouffer's statistics $T_{Stouffer} = \sum_{k=1}^{K} z_i/\sqrt{k}$, $(z_i = \Phi^{-1}(p_i))$,where $\Phi^{-1}(x)$ is the inverse cumulative distribution function of standard normal distribution. Similar to Fisher's method, smaller p-values contribute more to the Stouffer's score, but in a smaller magnitude.

1C. Adaptively-Weighted Fisher (with biomarker categorization)

Li and Tseng (2011) introduced an adaptively weighted Fisher's method (AW-Fisher) that characterizes effective studies under a null hypothesis for each gene g is considered as H0: $\theta_{g1} = \cdots = \theta_{gk} = 0$, where θ_{gk}denotes the gene effect of gene g

and study k, against an alternative hypothesis H_A: at least one $\theta_{gk} \neq 0$ in $1 \leq k \leq K$ study when integrating multiple genomic studies.

To uncover inter-study heterogeneous gene expression patterns across studies, they started with the following weighted statistic:

$$U_{g(w_g)} = -\sum_{k=1}^{K} w_{gk} log(p_{gk}),$$

where p_{gk} is the p-value of gene g in study k,w_k is the weight assigned to the kth study and $w_g = (w_{g1}, \ldots, w_{gK})$. Under the null hypothesis that $\theta_{gk}=0$ $\forall k$, the p-value of the observed weighted statistic, $p_{U(u_g(w_g))}$, can be obtained for a given gene g and weight w_g. The adaptively-weight statistic was defined as the minimal p-value among all possible weights:

$$V_g^{AW} = \min_{w_g \in W} p_U(u_g(w_g)),$$

where $u_g(w)$ is the observed statistic for $U_g(w)$, and W is a prespecified search space, the choice of search space is $W=\{w|w_i \epsilon \{0,1\}\}$, which results in an affordable computation of $O(2^K - 1)$ based on the norm of $K \leq 10$ in a microarray meta-analysis. The resulting weight reflects a natural biological interpretation of if a study contributes to the statistical significance of a gene.

2. Combining effect size

2A. Fixed effect model (FEM)

The fixed effect model is given as $T_{FEM} = \mu + \varepsilon_k, \varepsilon_k \sim N(0, s_k^2)$,where μ is the overall mean and ε_k is a random error following a normal distribution with a variance s_k^2 within certain study k.

2B. Random effect model (REM)

The random effect model is given as $T_{REM} = \mu + \delta_k + \varepsilon_k, \varepsilon_k \sim N(0, s_k^2), \delta_k \sim N(0, \tau^2)$, where μ is overall mean, ε_k is a random error following a normal distribution with a variance s_k^2 within certain study k and δ_k is a second source of error following a normal distribution with a variance τ^2 between studies.

Our proposed method for biomarker categorization in meta-analysis

Suppose there are K transcriptomic studies, each study k (1≤k≤K) measures the gene expression of n_k sample and G genes. In the simplest scenario, we compare the case and control samples in each study and identify the differentially expressed genes using popular methods such as limma (Ritchie et al. (2015)). After DE analysis, we obtain the effect size estimates ($log2FC_{gk}$) and p-values (p_{gk}) for each gene g (1≤g≤G) in each study k. We proposed the following two weighted statistics, one for up-regulated pattern and the other for down-regulated pattern, to look for the combination of studies that provides the best concordant DE evidence across all studies:

$$T^{+}_{g(w_g)} = \frac{\sum_{k\neq k', g\epsilon G, log_2FC_{gk}>0, log_2FC_{gk'}>0} w^{+}_{gk} * log_2FC_{gk} * w^{+}_{gk'} * log_2FC_{gk'} * |(log_{10}p_{gk} + log_{10}p_{gk'})|}{\sum_k w^{+}_{gk}}$$

$$T^{-}_{g(w_g)} = \frac{\sum_{k\neq k', g\epsilon G, log_2FC_{gk}<0, log_2FC_{gk'}<0} w^{-}_{gk} * log_2FC_{gk} * w^{-}_{gk'} * log_2FC_{gk'} * |(log_{10}p_{gk} + log_{10}p_{gk'})|}{\sum_k w^{-}_{gk}}$$

where w^{+}_{gk} *a*nd w^{-}_{gk} is a weight of 0 or 1 assigned to the kth study for gth gene. $\log2FC_{gk}$ is the calculated log2 fold change for gene g in study k. By default, $w^{+}_{gk} = 0$ for those studies with $\log_2FC_{gk}<0$ and $w^{-}_{gk} = 0$ for $\log_2FC_{gk}>0$. p_{gk} is the p-value for gene g in study k outputted from the differential expression analysis. $T^{+}_{g(w_g)}$ and $T^{-}_{g(w_g)}$ are the statistics for up-regulated pattern and down-regulated pattern,

respectively. In this measure, we account for statistical significance, biological significance and concordance patterns. Note that for genes with concordant directionality across all studies (either up-regulated or down-regulated), we define the corresponding $T^{+}_{g(w_g)}$or $T^{-}_{g(w_g)}$ in the other direction to be all zero. The weights in the other direction are also suppressed to zero for all studies. When only one study has an opposite direction, the statistics for the corresponding direction and the weight of the corresponding study are also zero.

We will search all possible 0 or 1 weight for K individual study in each of the two weighted statistics and define the best concordant statistics R^{+}_{g} and R^{-}_{g} as the maximum $T^{+}_{g(w_g)}$and $T^{-}_{g(w_g)}$among all possible weights for each gene:

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

$W^{+}\ and\ W^{-}$ are pre-defined searching space for up- and down-regulated studies with aforementioned restrictions. The resulting weights in R^{+}_{g} and R^{-}_{g} can then be used to categorize the biomarkers. We denote the best weights as $w^{+*}_{g}$and w^{-*}_{g}, respectively. The biomarkers are categorized according to different best weight distribution among studies (i.e. by merging the information of $w^{+*}_{g}$and w^{-*}_{g}). Comparing to the AW Fisher's method for gene ranking and biomarker categorization, our method has three significant advantages. Firstly, both biological significance and statistical significance are considered in our statistics. Secondly, the statistics prioritizes concordant genes across multiple studies and provides ranking for both up and down-regulated patterns. Finally, for discordant genes with both up and down-regulated patterns, we can use R^{+}_{g} and R^{-}_{g} to interpret and rank genes for both patterns or instead define a

dominating concordant pattern for each gene by taking the maximum of two R_g= max (R_g^+, R_g^-).

Chapter 3: Results

Simulation

Simulation setting

We conducted simulations to demonstrate the strength of the proposed best concordance statistics in weight categorization and gene ranking compared to AW Fisher method under two simulation scenarios.

Scenario 1. Assess weight patterns of both concordant and discordant genes

We generated n=100 observations consisting of 50 cases and 50 controls in each of K=5 studies. We first sampled the baseline gene expression level X'_{gik}independently for each observation i ($1 \leq i \leq 100$) from N(0, σ^2_{gk}), where σ^2_{gk} was randomly drawn from {0.2, 0.3, 0.4, 0.5, 0.6} for each gene g ($1 \leq g \leq 2000$) in study k ($1 \leq k \leq 5$). We then sampled a total of 800 DE genes of eight differential expression patterns across studies (100 genes for each pattern) by using an indicator $\delta_{gk} \in \{-1,0,1\}$ for all G=2000 genes (See Table 4 for nine true differential expression gene patterns), where δ_{gk} =1 or -1 indicates gene g in study k is an up- or down-regulated DE gene, δ_{gk} =0 indicates a non-DE gene. The rest 1200 genes are non-DE genes that have δ_{gk} =0 in all studies. When δ_{gk} =1 or -1, we sampled absolute effect size $\mu_{gk}$from a uniform distribution in the range of [0.7,1] for gene g in study k. The gene expression of control samples is $X_{gik} = X'_{gik}$, and the expression of case samples is $Y_{gik} = X'_{gik} + \mu_{gk} * \delta_{gk}$, for $1 \leq g \leq 2000$, $1 \leq i \leq 50$ and $1 \leq k \leq 5$. For non-DE genes, $Y_{gik} = X_{gik} = X'_{gik}$. We applied both our method and AW-Fisher method to the simulated data for biomarker categorization and compared their assigned weight patterns to the true

weight patterns (δ_{gk}). In addition, we also defined concordant DE genes as those up-regulated or down-regulated in at least $\lfloor k/2 \rfloor$ studies (one DE gene can be concordant up and down in the same) and plotted the true number of concordant DE genes versus the top ranked genes by the statistics of each method (R_g^+/R_g^- in our method and p-value in AW-Fisher's method).

Scenario 2. Assess gene ranking

In this scenario, we assessed the ranking of genes by our method and AW-Fisher method considering statistically significance, biological significance and concordance patterns across studies. We followed from scenario 1 and performed a second simulation by increasing the variance of gene expression σ_{gk}^2 to be randomly drawn from {4.3, 4.5, 4.4, 4.7, 4.2} for each gene g ($1 \leq g \leq 2000$) in study k ($1 \leq k \leq 5$). For our method, we used $R_g = \max(R_g^+, R_g^-)$ to rank the genes. For assessment, we plotted the true number of DE genes versus the top ranked genes by the statistics of each method.

Simulation results

For simulation scenario 1, Table 1 shows number of genes with correct weight patterns by our method and AW-Fisher's method. Both adaptively weighted Fisher's method and our proposed measure perform equally well by almost perfectly assigning the correct weights to studies with indicator $\delta_{gk} = 1$ or -1 in all five studies ($\delta_g^{(k=1,2,3,4,5)}$ =(1,1,1,1,1) or (-1,-1,-1,-1,-1))or first three studies k=1, 2, and 3 ($\delta_g^{(k=1,2,3,4,5)}$ =(1,1,1,0,0) or (-1,-1,-1,0,0)). When discordance exists in the true DE gene patterns (e.g. $\delta_g^{(k=1,2,3,4,5)}$ = (1,1,1,1,-1)), the adaptively weighted Fisher's method performed worse and incorrectly assigned weights of 1 to all five studies. On

the contrary, our measures not only correctly assigned weights to DE genes, but also reflected up- and down-regulated concordance.

Table 1 Number of corrected identified weight patterns by our method and AW-Fisher's method

True weight pattern	Number of genes with correct weight pattern by AW-Fisher's method	Number of genes with correct weight pattern by our method	
		Up-regulated part	Down-regulated part
1,1,1,1,1	100	97	NA
-1,-1,-1,-1,-1	100	NA	96
1,1,1,1,-1	0	100	NA
-1,-1,-1,-1,1	0	NA	98
1,1,1,-1,-1	0	100	100
-1,-1,-1,1,1	0	100	100
1,1,1,0,0	99	100	NA
-1,-1,-1,0,0	99	NA	100

Among all the DE genes with the eight true weight patterns, there are a total of 600 genes with concordant up-regulated pattern and 600 genes with concordant down-regulated pattern. Figure 1 and 2 show that our method identified more true concordant (up or down-regulated) genes among the top ranked genes by R_g^+ or R_g^- as compared to AW-Fisher method which ranked genes by p-values.

Figure 1. True number of concordant DE genes versus the top ranked genes by the statistics of R_g^+ in our method and p-value in AW-Fisher's method.

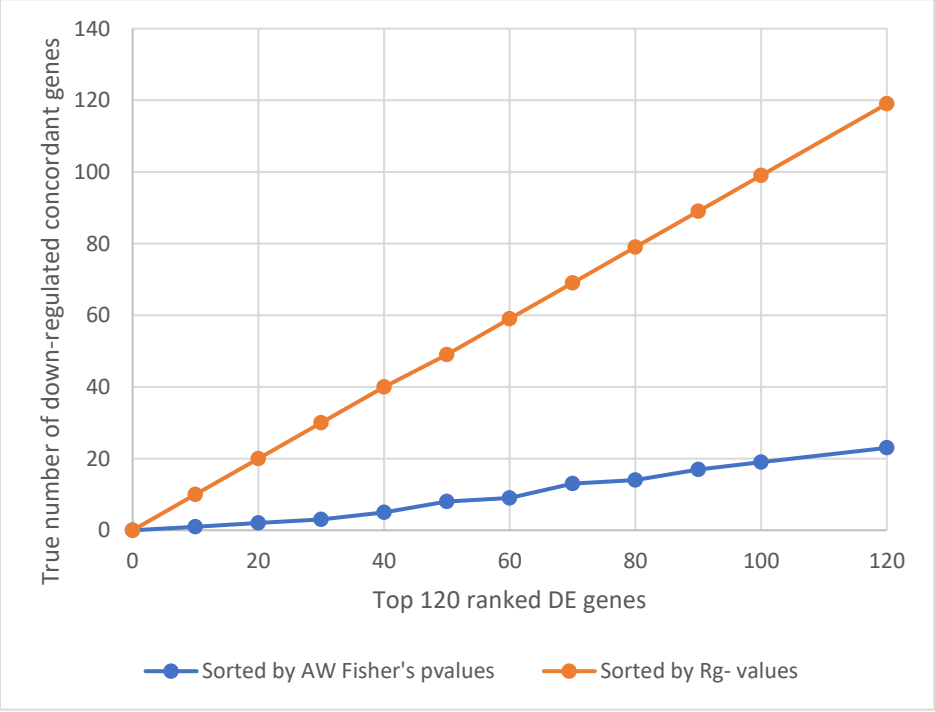


Figure 2. True number of concordant DE genes versus the top ranked genes by the statistics of R_g^- in our method and p-value in AW-Fisher's method.

For simulation scenario 2, Figure 3 demonstrates that our method detects a larger number of true DE genes than AW-Fisher method among the top ranked genes. Table 2 gives some example genes in each of the cross-study DE patterns to compare the two methods. For example, a DE gene 77 was assigned (1,1,1,1,1) as the best weight in our method but (1,1,0,0,1) in AW-Fisher method. It has large effect sizes (0.95 and 1.05) and also relatively large p-values (0.28 and 0.26) in the third and fourth study, thus ranks high in our method but ranks low in AW Fisher. This shows the advantage of our method in equally considering both the statistical and biological significance. DE gene 366 ranks higher in our method but ranks low in AW-Fisher mainly because AW-Fisher ignores the directionality of effect size.

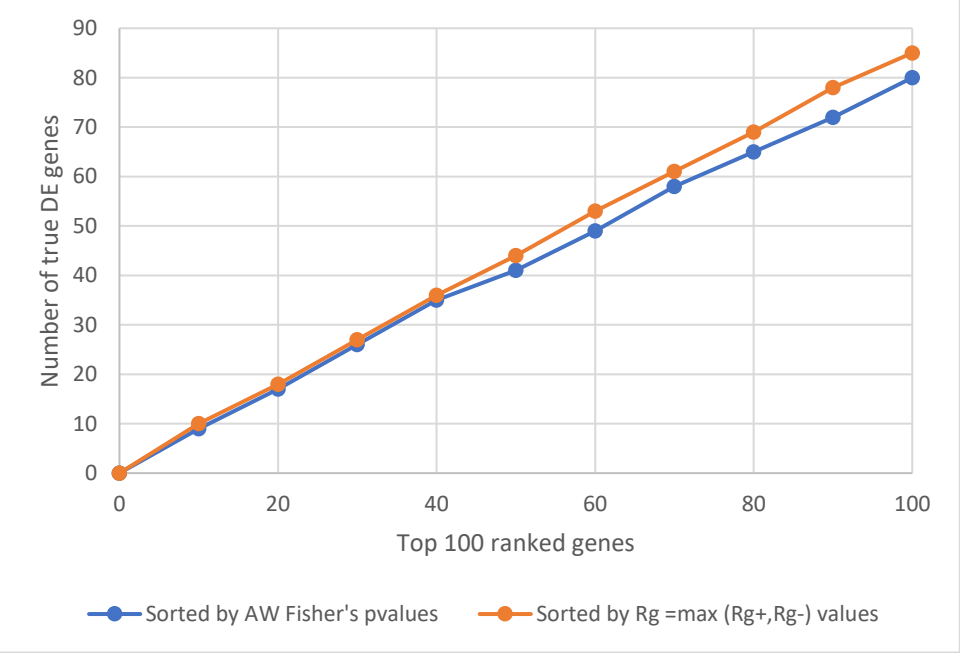


Figure 3. Number of true DE genes versus the top ranked genes by the statistics of R_g =max (R_g^+, R_g^-) in our method and p-value in AW-Fisher's method.

Table 2. For each of nine true gene patterns, concluding rank positions and statistical information of example genes in adaptively weighted Fisher's method and our measures (R_g =max (R_g^+, R_g^-)).

Gene	Truth	AW Fisher	Proposed measures	Lfc1	Lfc2	Lfc3	Lfc4	Lfc5	P1	P2	P3	P4	P5
		Rank position/ weight pattern	Rank position/ weight pattern										
77	1,1,1,1,1	101/ 1,1,0,0,1	51/ 1,1,1,1,1	1.91	2.14	0.95	1.05	1.83	0.03	0.02	0.28	0.26	0.03
125	-1,-1,-1,-1,-1	2/ 1,0,1,1,0	1/ 1,1,1,1,1	-2.23	-1.37	-3.27	-2.48	-1.24	0.01	0.13	0.00	0.01	0.14
270	1,1,1,1,-1	15/ 1,0,1,1,1	7/ 1,0,1,1,0	1.33	0.42	2.39	3.15	-1.45	0.12	0.64	0.01	0.00	0.08
366	-1,-1,-1,-1,1	31/ 1,1,1,0,1	17/ 1,1,1,0,0	-1.80	-2.22	-2.64	0.01	1.25	0.04	0.01	0.00	0.99	0.14
435	1,1,1,-1,-1	52/ 1,0,1,1,1	56/ 1,1,1,0,0	2.80	1.09	1.84	-1.75	-1.26	0.00	0.22	0.04	0.06	0.13
564	-1,-1,-1,1,1	37/ 0,1,1,1,1	34/ 0,0,0,1,1	0.36	-1.46	-1.31	3.13	1.98	0.68	0.10	0.14	0.00	0.02
668	1,1,1,0,0	47/ 0,1,1,0,0	12/ 1,1,1,0,0	1.27	2.88	2.16	-0.76	0.08	0.14	0.00	0.01	0.42	0.92
701	-1,-1,-1,0,0	58/ 1,1,0,0,0	45/ 1,1,1,0,0	-1.84	-3.03	-1.01	0.34	-0.13	0.03	0.00	0.25	0.72	0.88

Real application

Data description

The Pan-Gynecologic (Pan-Gyn) molecular data from The Cancer Genome Atlas (TCGA) project consists of 2579 tumors representing four gynecological types plus breast and aims to provide insight into the commonalities and differences across multiple tumor lineages (Berger et al. (2018), Weinstein et al. (2013)). We collected the data from 1620 cancer patient samples of the "Pan-Gyn" cohort including 299 OV, 263 UCEC, 306 CESC, 57 UCS, and 695 BRCA samples, with both the RNA-seq data and the clinical outcomes available. We first merged the RNA-seq data of the five cancer types (regarded as studies in our meta-analysis framework) by matching the gene symbols and implementing quantile normalization to force the distributions to be the same across all five datasets. Genes with mean expression levels less or equal to 1 and variance expression levels less or equal to 2nd qualities

were filtered out, and 7127 genes remained. We applied a log2 transformation for the data for statistical analysis.

Results

Table 3 showed the number of genes, total sample size and the number of dead/alive samples (defined from the overall survival data) of each cancer type.

Table 3. Descriptive analysis of real data set

TCGA tumor types	OV	UCEC	CESC	UCS	BRCA
Mean>1 and Variance>Q2	10093	10069	10186	10243	10043
Total sample size (Dead/Alive)	299(182/117)	263(38/225)	306(73/233)	57(35/22)	695(89/606)
Censoring rate	39.13%	85.56%	76.14%	62.86%	87.19%

A total of 26 different weight categories were determined by our measures (Table 4).

Table 4. Weight categories gained by R_g =max (R_g^+, R_g^-)) in our measures.

Weight pattern	Count	Weight pattern	Count
0 0 0 1 1	656	0 0 1 1 1	717
0 0 1 0 1	371	0 1 0 1 1	250
0 0 1 1 0	805	0 1 1 1 0	339
0 1 0 0 1	238	1 0 0 1 1	367
0 1 0 1 0	262	1 0 1 0 1	216
0 1 1 0 0	158	1 0 1 1 0	339
0 1 1 0 1	268	1 1 0 0 1	186
1 0 0 0 1	328	1 1 0 1 0	154
1 0 0 1 0	388	1 1 1 0 0	141
1 0 1 0 0	146	0 1 1 1 1	136
1 1 0 0 0	110	1 0 1 1 1	201
1 1 0 1 1	71	1 1 1 1 0	105
1 1 1 0 1	110	1 1 1 1 1	65

The heatmaps of standardized log2-fold change show the representative weight categories of genes with adaptively weighted Fisher's p-values less than 0.05 by using R_g^+ (Figure 4) and R_g^- (Figure 5). Figure 6 and 7 zoomed in to two specific weight categories. We can see weights were assigned to the studies: BRCA, CESC and UCEC which are having small p-values. Our measures also assign weight to the study UCS which having relatively larger p-values but with large effect size. Table 3 shows the sample size of UCS is 57 that limits statistical power, partly explained why UCS usually have large effect size but not significance p-values. Our measure considered absolute magnitude of effect size to avoid overlooking some important genes in study UCS. For the 35 genes shown in figure 6, the direction of effect size of all studies except for OV are up-regulated. Figure 7 shows the direction of effect size of the 9 genes are down-regulated with large absolute magnitude values.

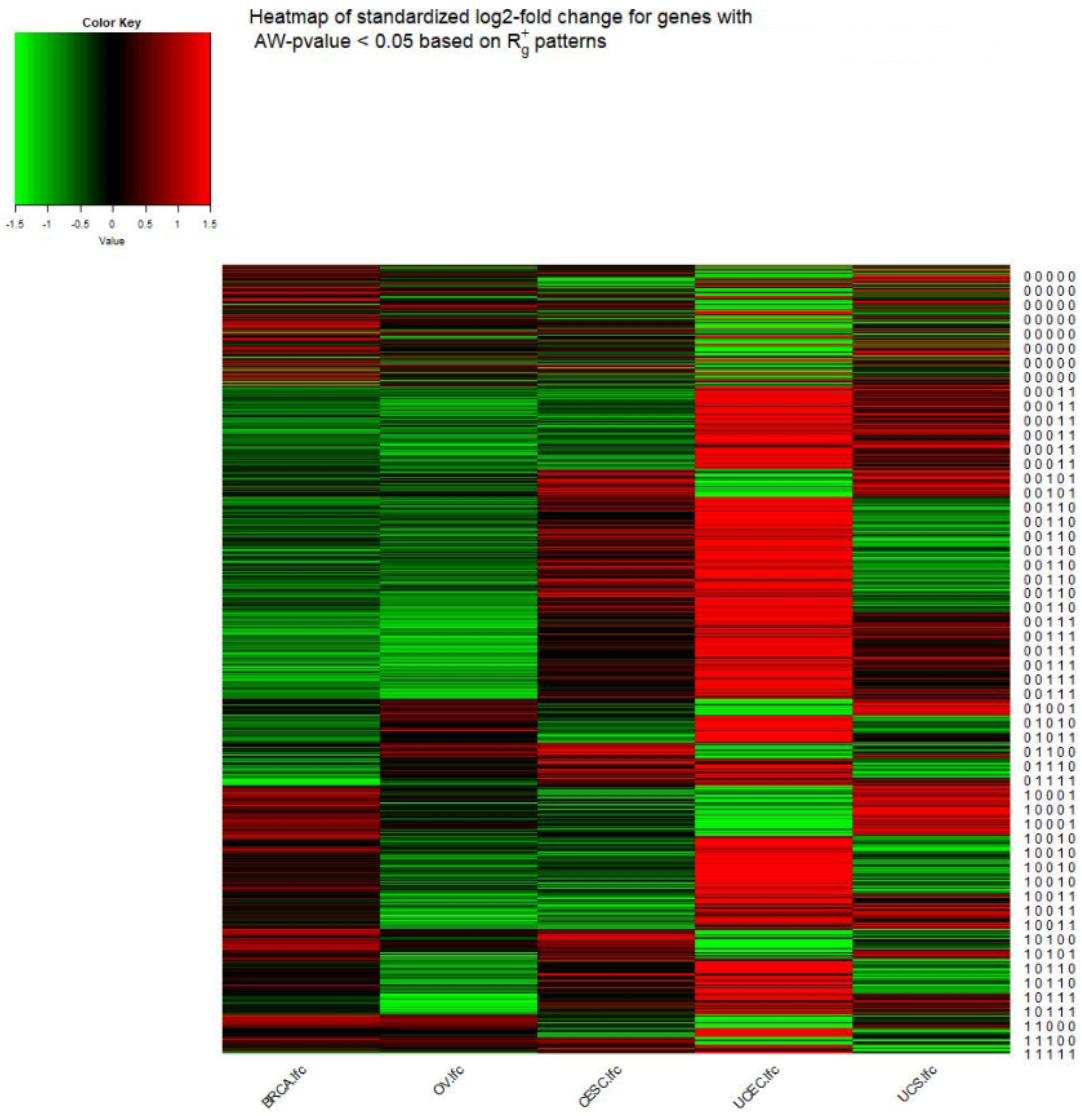


Figure 4. Heatmap of standardized log2-fold change values showing representative weight categories of genes with adaptively weighted Fisher's p-values less than 0.05 by using R_g^+.

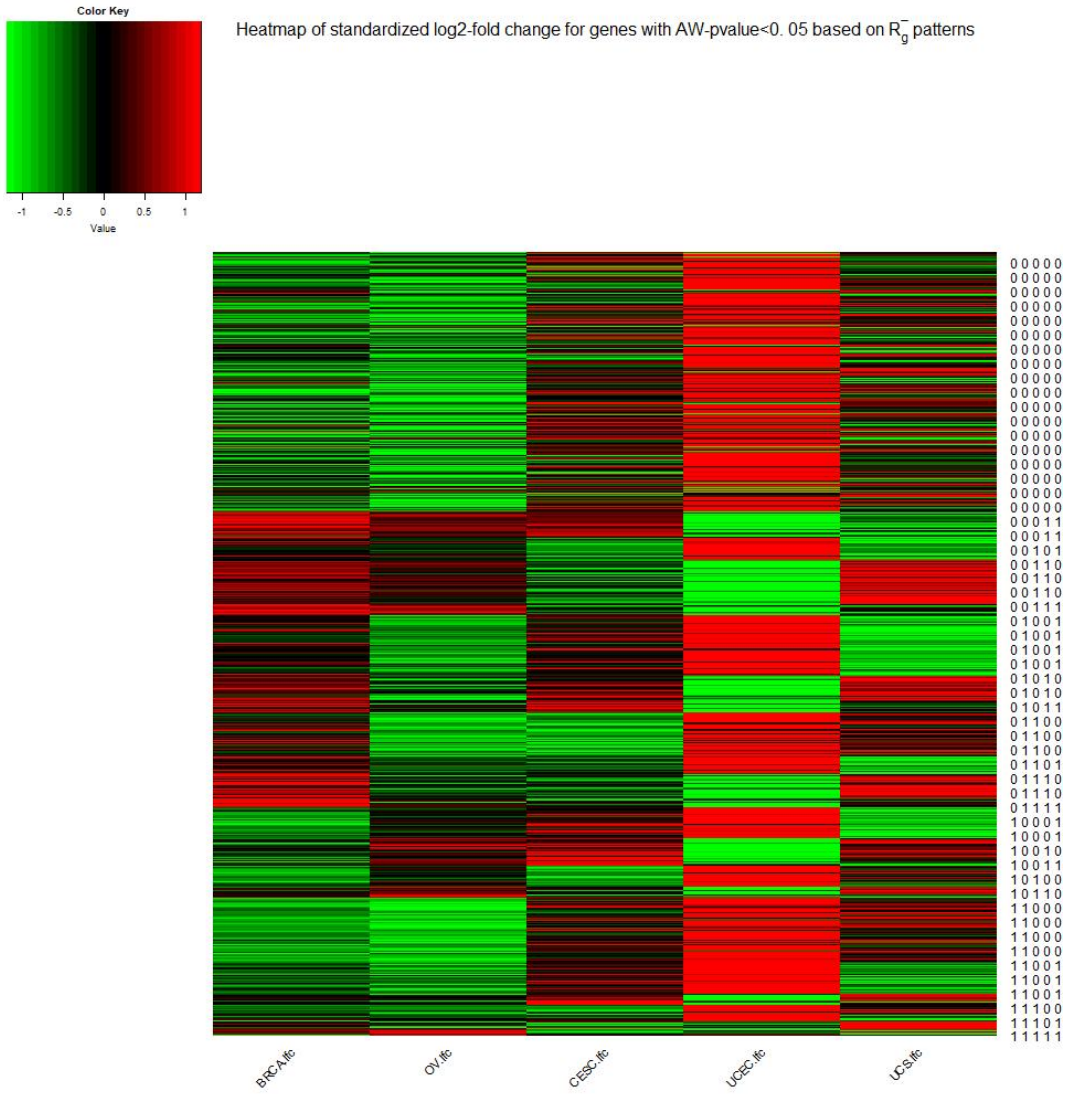


Figure 5. Heatmap of standardized log2-fold change values showing representative weight categories of genes with adaptively weighted Fisher's p-values less than 0.05 by using R_g^-.

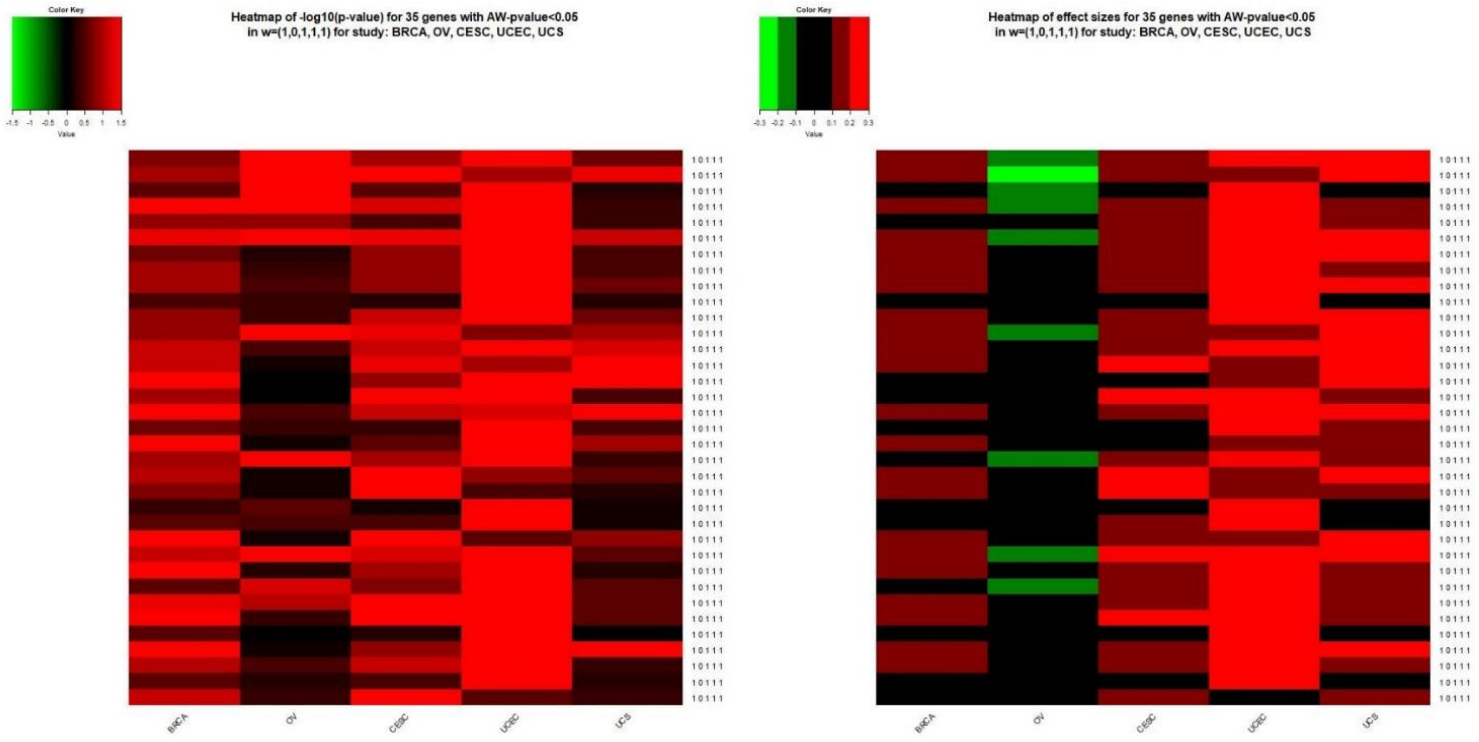


Figure 6. Heatmap of -log10(p-values) and effect sizes (log2-fold change) for 35 genes identified as a weight pattern (1,0,1,1,1) for study BRCA, OV, CESC, UCEC, UCS by using R_g^+.

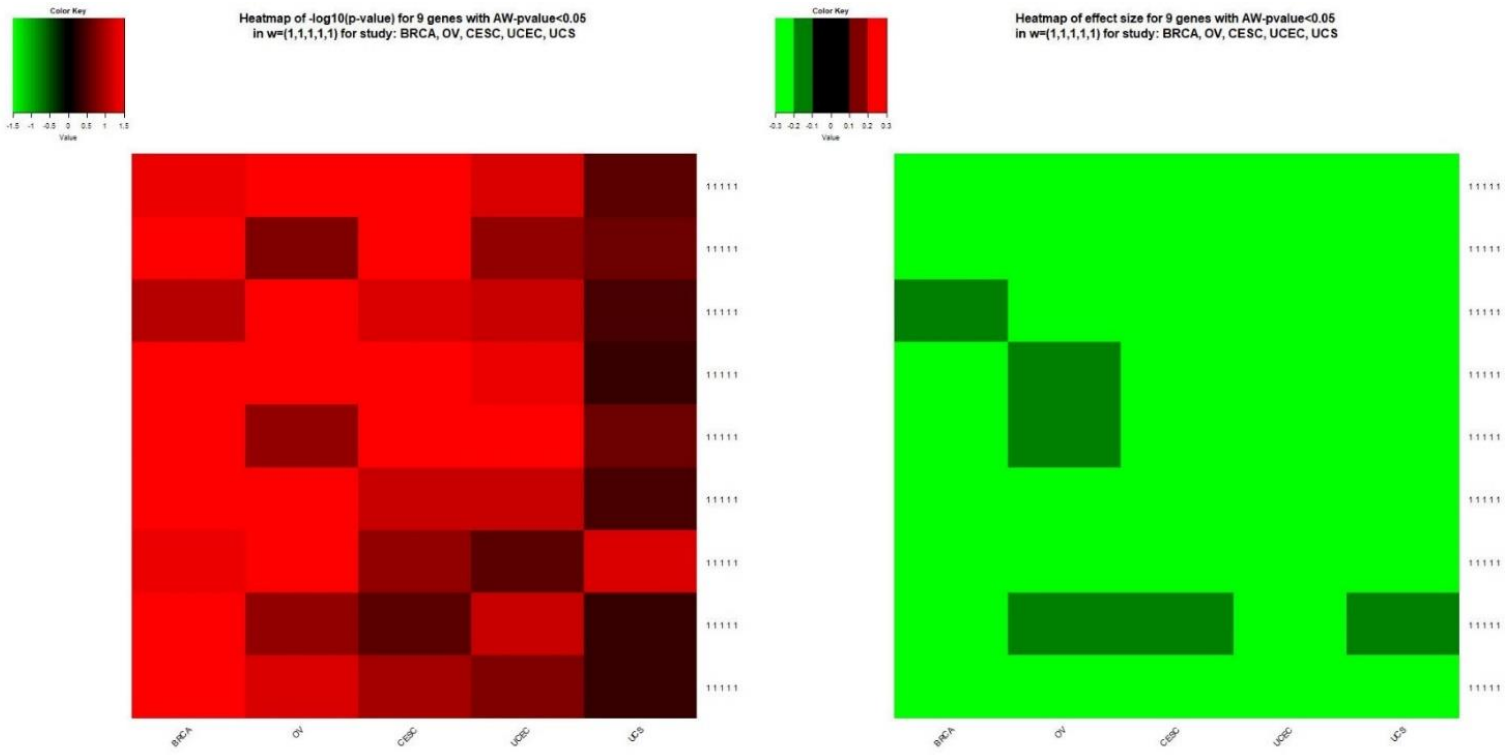
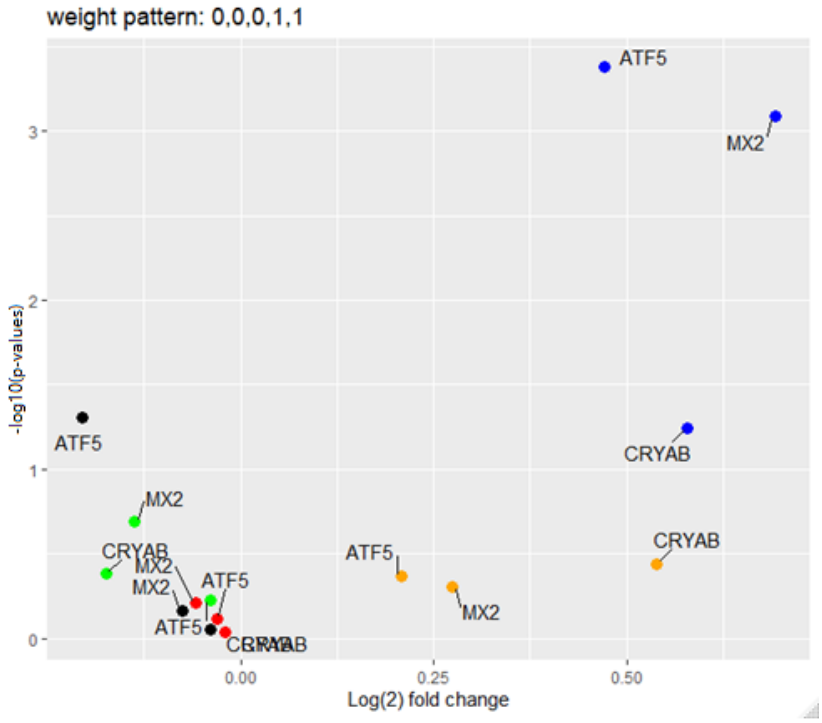


Figure 7. Heatmap of -log10(p-values) and effect sizes (log2-fold change) for 9 genes identified as a weight pattern (1,1,1,1,1) for study BRCA, OV, CESC, UCEC, USC.

Figure 8 shows a volcano plot of three example genes “ATF5”, “MX2” and “CRYAB”, all have (0,0,0,1,1) as the best estimated weight by our method. It can be clearly seen that all of these three genes in the UCS (orange color) and UCEC (blue color) have relatively large absolute magnitude of effect sizes and also relatively small p-values compared to the other three studies.



BRCA-green | OV-red | CESC-black | UCEC-blue | UCS-orange

Figure 8. The volcano plot of three example genes "ATF5", "MX2" , and "CRYAB" identified as a weight pattern: (1,0,0,0,1) by R_g=max(R_g^+, R_g^-) for study BRCA, OV, CESC, UCEC, UCS.

Pathway analysis

After obtaining different weight categories, we performed pathway enrichment analysis on each weight category separately using Fisher's exact test, focusing on pathway sizes larger than 10 from four pathway database: Biocarta Pathway (http://www.biocarta.com/), Gene Ontology (GO; www.geneontology.org) , Kyoto encyclopedia of genes and genomes (KEGG; http://www.genome.jp/kegg/), and Reactome (https://reactome.org/). We used a Venn diagram to compare the top 100 pathways enriched by each of the four weight categories. The overlapping part has less than five pathways among the four different weight patterns, which indicates that

different categories of biomarkers identified by our method may imply different functional domains of biological interest.

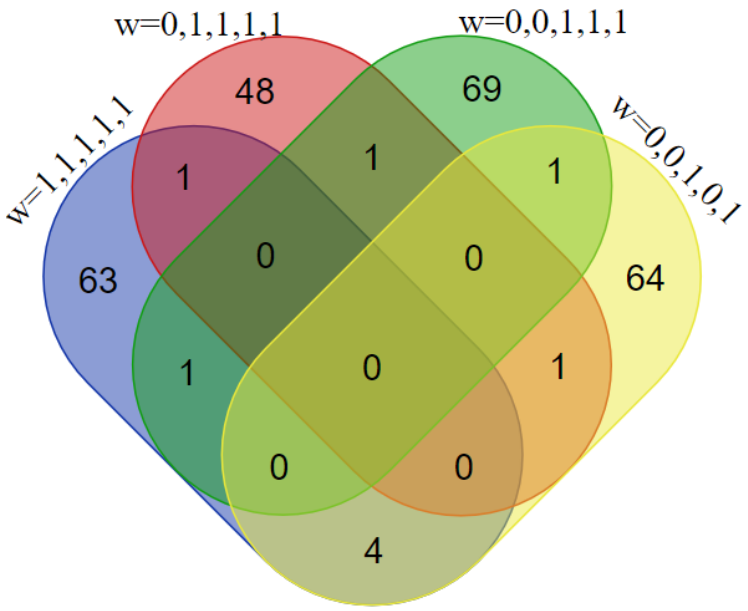


Figure 9. The Venn diagram of pathway intersections with pathway sizes larger than 10 among top 100 pathways enriched by each of the four weight categories ((1,1,1,1,1), (0,1,1,1,1), (0,0,1,1,1) and (0,0,1,0,1)) identified by using $R_g=\max(R_g^+, R_g^-)$ for study BRCA, OV, CESC, UCEC, USC.

Chapter 4: Discussion and Conclusion

The proposed research is highly responsive to a comprehensive method to categorize inter-study heterogenous biomarkers by simultaneously considering statistical significance (p-values), biological significance (absolute magnitude of effect size) and up-/down-regulated concordance (direction of effect size). Our main findings from the simulation and real data analysis shows that more true concordant (up or down-regulated) genes were identified and ranked higher in our method as compared to the AW Fisher method. In the real application, small sample size studies having both large absolute magnitude of effect size and relatively large p-values in most genes were also assigned weights in our method as compared to the AW Fisher method. The new biomarker categories found in our method may be used for predicting the properties or classes of new samples and offer biologically meaningful information for future precision medicine.

There are three limitations of our measures. First, our method cannot distinguish whether a gene is DE gene or non-DE gene. In our study, the DE genes were determined by adaptively weighted Fisher's p-values (<0.05). We will derive a p-value for our concordance statistics by applying empirical procedures like permutation or bootstrapping for better inference and gene ranking in the future. Multiple testing issue will also need to be addressed once we have a p-value in the method. In real application, we have 7127 hypotheses to test, and a significance level of 0.05. The probability of observing at least one significant result is 100% ($1-(1-0.05)^{7127}$), meaning we have a 100% chance of observing at least one significant result, even if all of the tests are actually not significant. FDR procedure will need to

be applied to correct for the multiplicity. Secondly, we assume the simplest scenario where genes are independent from each other. In the future, we will consider dependence among the genes in the weight categorization. Thirdly, is the number of weigh category is impacted by the number of studies in our measure due to 2^K-1 possibilities of w_{gk}. The more studies are included in our measure, the more possibilities of weight patterns that we would have, and the heavier the computation. In the future, we will develop efficient algorithm for a fast selection of best weight among all possibilities.

Reference

1. Soneson, C., Delorenzi, M. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14, 91 (2013). https://doi.org/10.1186/1471-2105-14-91
2. Ramasamy, A., Mondry, A., Holmes, C. C., & Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. PLoS medicine, 5(9), e184. https://doi.org/10.1371/journal.pmed.0050184
3. Tseng G, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40(9):3785–3799
4. Fisher, R. A. (1925). Statistical methods for research workers. Genesis Publishing Pvt Ltd.
5. Stouffer SA, Suchman EA, DeVinnery L, Star S, Williams RMJr. , The American Soldier, Volume I: Adjustement during Army Life, 1949 Princeton, NJPrinceton University Press
6. Li J, Tseng GC. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies, Ann. App. Stat., 2011, vol. 5 (pg. 994-1019)
7. Choi JK, Yu U, Kim S, Yoo OJ. Combining multiple microarray studies and modeling interstudy variation, Bioinformatics, 2003, vol. 19 Suppl. 1(pg. i84-i90)

8. Domaszewska, T., Scheuermann, L., Hahnke, K. *et al.* Concordant and discordant gene expression patterns in mouse strains identify best-fit animal model for human tuberculosis. *Sci Rep* **7,** 12094 (2017). https://doi.org/10.1038/s41598-017-11812-x
9. Hosack D, et al. Identifying biological themes within lists of genes with EASE, Genome Biol., 2003, vol. 4 pg. R70
10. Berger AC, Korkut A, Kanchi RS, Hegde AM, Lenoir W, Liu W, Liu Y, Fan H, Shen H, Ravikumar V, et al. A comprehensive Pan-Cancer molecular study of gynecologic and breast cancers. Cancer Cell. 2018;33 https://doi.org/10.1016/j.ccell.2018.03.014.
11. Weinstein J.N. et al. (2013) The cancer genome atlas pan-cancer analysis project. Nat. Genet., 45, 1113–1120.
12. National Cervical Cancer Coalition (2019). Gynecological Cancers. Retrieved from https://www.nccc-online.org/hpvcervical-cancer/gynecological-cancers/.
13. Chang, L.-C., Lin, H.-M., Sibille, E., & Tseng, G. C. (2013). Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. BMC Bioinformatics, 14(1), 368. doi: 10.1186/1471-2105-14-368
14. Ma, T., Liang, F. and Tseng, GC. (2017). Biomarker detection and categorization in ribonucleic acid sequencing meta-analysis using Bayesian hierarchical model. Journal of the Royal Statistical Society: Series C. 66(4): 847-867.

15. Tianzhou, Ma, et al. “A Joint Bayesian Model for Integrating Microarray and RNA Sequencing Transcriptomic Data.” Journal of Computational Biology, vol. 24, no. 7, 2017, pp. 647–662., doi:10.1089/cmb.2017.0056.
16. Tianzhou Ma, Zhiguang Huo, Anche Kuo, Li Zhu, Fang Zhou, Xiangrui Zeng, Chien-Wei Lin, Silvia Liu, Lin Wang, Peng Liu, Tanbin Rahman, Lun-Ching Chang, Sunghwan Kim, Jia Li, Yongseok Park, Chi Song and George C. Tseng. (2018). MetaOmics - Comprehensive Analysis Pipeline and Web-based Software Suite for Transcriptomic Meta-Analysis. Bioinformatics.