

RESEARCH ARTICLE

Open Access



# A regulation probability model-based meta-analysis of multiple transcriptomics data sets for cancer biomarker identification

Xin-Ping Xie<sup>1</sup>, Yu-Feng Xie<sup>1,2</sup> and Hong-Qiang Wang<sup>2,3\*</sup>

## Abstract

**Background:** Large-scale accumulation of omics data poses a pressing challenge of integrative analysis of multiple data sets in bioinformatics. An open question of such integrative analysis is how to pinpoint consistent but subtle gene activity patterns across studies. Study heterogeneity needs to be addressed carefully for this goal.

**Results:** This paper proposes a regulation probability model-based meta-analysis, *j*GRP, for identifying differentially expressed genes (DEGs). The method integrates multiple transcriptomics data sets in a gene regulatory space instead of in a gene expression space, which makes it easy to capture and manage data heterogeneity across studies from different laboratories or platforms. Specifically, we transform gene expression profiles into a united gene regulation profile across studies by mathematically defining two gene regulation events between two conditions and estimating their occurring probabilities in a sample. Finally, a novel differential expression statistic is established based on the gene regulation profiles, realizing accurate and flexible identification of DEGs in gene regulation space. We evaluated the proposed method on simulation data and real-world cancer datasets and showed the effectiveness and efficiency of *j*GRP in identifying DEGs identification in the context of meta-analysis.

**Conclusions:** Data heterogeneity largely influences the performance of meta-analysis of DEGs identification. Existing different meta-analysis methods were revealed to exhibit very different degrees of sensitivity to study heterogeneity. The proposed method, *j*GRP, can be a standalone tool due to its united framework and controllable way to deal with study heterogeneity.

**Keywords:** Cancer, Transcriptomics data, Meta-analysis, Differential expression, Regulation probability

## Background

High throughput biotechnology has become a routine tool in biological and biomedical research [1, 2]. Its extensive applications have been generating and accumulating a flood of omics data that bring unprecedented opportunity for elucidating cancer or other diseases at a molecular level [3–6]. For example, various types of omics data for nearly 10,000 tumor or normal samples have been released from the cancer genome atlas (TCGA) project. In the two famous public databases, Gene Expression Omnibus (GEO) and ArrayExpress, there are millions of assays generated in more than 30,000 studies world-wide available online [7, 8]. To reduce sample bias and increase

statistical power, one needs to reuse the flood of omics data in a meta-analysis way, gaining deeper insights into the molecular pathology of cancer or other diseases [9]. How to implement efficient meta-analysis of these data sets poses a pressing challenge for computational biologists and bioinformaticians.

Meta-analysis of transcriptomic data needs to interrogate consistent but subtle gene activity patterns across studies. Currently, there exist three categories of meta-analysis methods used for DEGs identification: *p*-value-based, effect size-based and rank-based. These methods deal with non-specific variations at different levels of data. For example, in statistics, *p*-value methods are most intuitive and allow for standardization of topic-related associations from studies to the common scale of significance. However, the performance of the *p*-value methods is stringently conditional on the estimation model of *p*-values used in individual

\* Correspondence: hqwang126@126.com

<sup>2</sup>Cancer Hospital, CAS, Hefei, Anhui 230031, China

<sup>3</sup>MICB Lab., Hefei Institutes of Physical Science, CAS, Hefei 230031, China

Full list of author information is available at the end of the article



analysis [10, 11]. To improve the situation, Li and Tseng [10] proposed an adaptively weighted strategy (AW) for  $p$ -value combination. Recently, Li et al. [12] introduced multiple test procedure and established assumption-weighting statistics, including I2, I2&direction, and mean cor, pooled cor, which are expected to settle down the heterogeneity and capture the concordance between different studies. Unlike the  $p$ -value methods, the effect size methods rely on a  $t$ -statistic-like model and can directly model the effect sizes across different studies. There are two commonly used effect size models in meta-analysis of transcriptomics data: fixed-effect model (FEM) and random effect model (REM), whose difference mainly lies in whether ignoring between-study variations or not. Compared with the  $p$ -value methods, the effect size methods are more sensitive to data distribution and noise inherent in microarray data, leading to unreliable effect size estimates [13].

As a non-parametric method, rank-based methods rely on combining the fold-change ranks, rather than combining  $p$ -values as in the  $p$ -value methods or expression levels as in the effect size methods. Compared with the effect size models, the rank-based methods make fewer or no assumptions about data structures in modeling differential expression of genes and thus runs more robust and outlier-free in performing meta-analysis for screening DEGs [14, 15]. A representative rank-based method is the *Rankprod* method proposed by Hong et al. [13]. In *Rankprod*, multiple fold changes are computed from all possible pair-wise comparisons of samples in each data set, and the rank product for each gene is then carried out by ranking the resulting fold changes within each comparison. For significance analysis, *Rankprod* assesses the null distributions of the rank product in each data set by Permutation tests. Unfortunately, *Rankprod* only work well for data sets where two categories of differential genes with two opposite directions are involved, and is less sensitive to inconsistent patterns of differential expression across studies [12, 16]. Additionally, Wang et al. proposed a matrix decomposition-based strategy for meta-analysis of transcriptomics data, which improves meta-analysis by mining differential physiological signals hidden behind multiple data sets [17].

A main issue in gene expression meta-analysis is how to deal with the study heterogeneity across data sets. The heterogeneity possibly comes from three sources: 1) Experimental environments. Gene expression datasets were often produced using different platforms and different processing facilities. Such kind of heterogeneity is often referred to as cross-lab/platform heterogeneity or batch effect [18]; 2) Incorrect gene annotations as technique mistakes, which occur when aligning target sequences or probes [19]; 3) Biological variability including various sub-subtypes of cancer or minor biological differences

(e.g. age, gender or ethnicity). These heterogeneities could deteriorate identifying DEGs in meta-analysis if they are not addressed properly. Dealing with these heterogeneities should be simultaneously removing the non-specific heterogeneity and accommodating the minor biological ones properly. We previously proposed a regulation probability-based statistic for identifying DEGs in a single experiment, referred to as GRP [20]. The GRP model estimates the probabilities of two regulation events occurring between sample groups and allows to capture and control data noise or the intra-class heterogeneity. We here extend the model to deal with study heterogeneity in the context of meta-analysis of multiple data sets. Briefly speaking, the proposed method, joint GRP (*jGRP*), maps gene expression data across studies to a regulatory space and then measures expression difference in the regulatory space. In the resulted gene regulation profile, study heterogeneity can be efficiently captured and controlled by a regulation confidence parameter. We evaluated the proposed methods on both simulation data and real-world transcriptomic data sets, and experimental results demonstrate the superior performance of *jGRP* in gene expression meta-analysis for DEGs identification.

## Methods

The main idea of the proposed method is to integrate multiple expression data sets at the level of regulation rather than at the level of expression. More specifically, we produce a united gene regulation profile across studies from independent gene expression profiles and measure differential expression by characterizing the regulation property of genes between two conditions. Biologically, two opposite regulation events possibly occur in tumor relative to normal tissue for a given gene: up-regulation ( $U$ ) and down-regulation ( $D$ ). The former means that a gene expresses higher in tumor than does in normal tissue, while the latter means that a gene expresses lower in tumor than does in normal tissue. Let  $P(U)$  and  $P(D)$  represent the estimates of the two events' probabilities, a regulation-based differential expression statistic can be defined as

$$jGRP = P(U) - P(D) \quad (1)$$

The statistic  $jGRP \in [-1, 1]$  reflects how likely the gene is regulated, whose positive value implies an up-regulation event occurring while whose negative value implies a down-regulation event occurring. A gene with a positive *jGRP* is potentially an onco-gene while the one with a negative *jGRP* is potentially a tumor suppressor.  $P(U)$  and  $P(D)$  need to be estimated in a gene regulation space. So, we first map gene expression profiles from microarrays or RNA-seq technology into a regulatory space, and the

resulting gene regulation profiles can be used to estimate the two regulation probabilities, statistically.

**Mapping gene expression data to gene regulatory space**

Suppose  $T$  studies each with two sample classes: tumor and normal tissue. For all the studies, we divide the total sample space into two subspaces: tumor subspace  $S_1$  and normal tissue subspace  $S_2$ . For a given gene, we assume three regulation statuses in a sample: up-regulated one denoted by 1, down-regulated one denoted by -1, and non-regulated one denoted by 0. Considering a study  $s$  consisting of  $n$  tumor samples and  $m$  normal samples and a gene  $g$  whose expression levels in the tumor and normal tissue samples are  $Y_1 = \{a_{11}, a_{12}, \dots, a_{1n}\}$  and  $Y_2 = \{a_{21}, a_{22}, \dots, a_{2m}\}$  respectively, we can map the expression levels of gene  $g$  into a regulatory space as follows:

1) For the  $i$ th tumor sample with expression level  $a_{1i}$ , its regulatory status can be determined as

$$r_{1i} = \begin{cases} 1 & l_i \geq \tau \\ -1 & 1 - l_i > \tau \\ 0 & \text{others} \end{cases} \quad (2)$$

where  $l_i = \sum_{k=1}^m I(a_{1i} \geq a_{2k}) / m$  represents the proportion of normal samples with an expression value not lower than  $a_{1i}$ , and  $0.5 \leq \tau \leq 1$  is a constant, referred to as regulation confidence cutoff, which controls the reliability of the inferred status.  $I(\cdot)$  is an indicator whose value is one if the condition is true and zero else.

2) For the  $i$ th normal sample with expression level  $a_{2i}$ , its regulatory status can be determined as

$$r_{2i} = \begin{cases} 1 & r_i \geq \tau \\ -1 & 1 - r_i > \tau \\ 0 & \text{others} \end{cases} \quad (3)$$

where  $r_i = \sum_{k=1}^n I(a_{2i} \leq a_{1k}) / n$  represents the proportion of tumor samples with expression values not lower than  $a_{2i}$ .

Combining Eqs.(2) and (3), the regulation profile of gene  $g$  in study  $s$  can be formulated as

$$R_s = [-1, 0, 1]^{m+n} \quad (4)$$

and then the united regulation profile across the  $T$  studies as

$$R = [R_1, R_2, \dots, R_T] \quad (5)$$

**Statistical estimation of  $jGRP$  statistic**

Given the two sample subspaces  $S_1$  and  $S_2$ , we estimate the two regulation events' probabilities based on the

regulatory statuses using the total probability theorem as follows:

$$P(U) = P(Y_1)P(U|Y_1) + P(Y_2)P(U|Y_2) \quad (6)$$

and

$$P(D) = P(Y_1)P(D|Y_1) + P(Y_2)P(D|Y_2) \quad (7)$$

where the prior probabilities of cancer and normal samples,  $P(Y_1)$  and  $P(Y_2)$ , can be assessed as the proportions of cancer and normal samples in all the  $T$  studies respectively, and the rest four conditional probabilities can be assessed as the proportions of samples with up/down-regulated statuses in the corresponding subspace. Then, the statistic  $jGRP$  can be derived as

$$jGRP = \frac{s_u - s_d}{n + m} \quad (8)$$

where  $s_u$  and  $s_d$  are the numbers of samples in which gene  $g$  is in up-regulated and down-regulated statuses, respectively. Note that the summation ( $S$ ) of  $P(U)$  and  $P(D)$  could vary around 1 depending on  $\tau$ :  $S$  will be larger than one if  $\tau \leq 0.5$  and be smaller than one else.

**Significance analysis of  $jGRP$**

We design a permutation test procedure for the significance analysis of  $jGRP$ . In the procedure, the labels of all samples across studies are randomly permuted  $B = 1000$  times, and thus  $B$  permuted  $jGRPs$  can be obtained by running the  $jGRP$  procedure on the permuted data. The  $B$  permuted  $jGRPs$  provide an approximate to the null distribution of  $jGRP$  statistic, and so the significance level of an observed  $jGRP$  can be estimated as

$$p\text{-value} = \frac{\sum_{i=1}^B I(|jGRP_i| \geq |jGRP|)}{B} \quad (9)$$

where  $jGRP_i$ ,  $i = 1, 2, \dots, B$  represents the  $i$ th permuted  $jGRP$  from the permutation experiment.

**Results**

**Evaluation on simulation data**

**Simulation data generation**

Generally, study heterogeneity could come from: (i) Difference in the fraction of studies that show significantly differential expression in all the studies; (ii) Difference in different expression directions across studies. Accordingly, we generated two types of simulation data, simulation-I and II, which focus on the two aspects of heterogeneity respectively, by revising the procedure in [21].

Assume  $T = 10$  studies each consisting of tumor and normal tissue groups of sizes randomly sampling from 4 to 15 and totally  $G = 10,000$  genes to be considered. For simulation-I where DEGs are homogeneously differentially

expressed, we simulated five categories of DEGs: differentially expressed in ten, eight, six, four and two studies, respectively. All the categories each were supposed to contain 500 genes, and the rest genes (7500) were assumed to be non-differential in any of the studies. For simulation-II, we assumed DEGs to be differentially expressed in different directions in different studies and considered two groups of categories of differential expression: The first group has differential expression in all ten studies, which consists of three categories: 1) differentially expressed in the same direction in all ten studies; 2) differentially expressed in seven of ten studies in one direction but in the rest (three) in the other direction; 3) differentially expressed in five of ten studies in one direction but in the rest (five) in the other direction. The second group have differential expression in six out of ten studies and consists of three categories: 1) differentially expressed in all six studies in the same direction; 2) differentially expressed in four of six studies in one direction, but in the rest (two) in the other direction; 3) differentially expressed in half studies in one direction, but in another half (three) in the other direction. Each of the six categories was assumed to contain 500 genes, and the rest genes (7000) were assumed to be non-differential in any of the studies. Tables 1 and 2 summarizes the details of the configuration of these simulation data.

To synthesize the expression level of genes, we assume that the expression of each gene follows a normal distribution in each group and each study, i.e., the expression level  $x_{gsic}$  of a gene  $g$  in sample  $i$  of group  $c$  in study  $s$  was randomly sampled from  $N(\mu_{gsc}, \sigma_{study}^2)$ . Specifically, for the normal tissue group, the mean of expression was designed as  $\mu_{gs0} = \mu + \alpha_g + \beta_s + (\alpha\beta)_{gs}$ , where  $\mu$  represents a constant background expression,  $\alpha_g \sim N(0, \sigma_{gene}^2)$  represents the gene bias,  $\beta_s \sim N(0, \sigma_{study}^2)$  represents the study bias, and  $(\alpha\beta)_{gs} \sim N(0, \sigma_{int}^2)$  represents the gene-study interaction. For the tumor group, the mean of expression was  $\mu_{gs1} = \mu_{gs0}$  for non-differential genes and  $\mu_{gs1} = \mu_{gs0} + \delta + v_g + \epsilon_{gs}$  for differential genes, where  $\delta$  is the pooled

**Table 1** Differential expression settings of Simulation data-I/ Simulation data-II

Category No.	Number of differential expression studies	Differential expression direction
1	10/10	Same/Same
2	8/10	Same/7:3
3	6/10	Same/5:5
4	4/6	Same/Same
5	2/6	Same/4:2
6	0/6	Same/3:3

**Table 2** Top 20 KEGG pathways enriched in the DEG list of  $jGRP(\tau = 0.7)$

Term	P-value	BH-adjusted p-value
hsa04610:Complement and coagulation cascades	1.55E-07	4.61E-05
hsa04110:Cell cycle	4.60E-07	6.85E-05
hsa05150:Staphylococcus aureus infection	4.69E-07	4.66E-05
hsa05200:Pathways in cancer	7.69E-07	5.73E-05
hsa01130:Biosynthesis of antibiotics	1.28E-05	7.62E-04
hsa05222:Small cell lung cancer	4.42E-05	0.002192532
hsa05166:HTLV-I infection	4.90E-05	0.002081948
hsa04512:ECM-receptor interaction	8.49E-05	0.003157108
hsa04510:Focal adhesion	1.53E-04	0.005064416
hsa04640:Hematopoietic cell lineage	2.60E-04	0.007713087
hsa04514:Cell adhesion molecules (CAMs)	3.22E-04	0.008693226
hsa05133:Pertussis	3.93E-04	0.009705856
hsa04115:p53 signaling pathway	4.52E-04	0.01031831
hsa04668:TNF signaling pathway	6.53E-04	0.013813372
hsa05416:Viral myocarditis	6.59E-04	0.01300695
hsa05144:Malaria	7.11E-04	0.013154554
hsa05202:Transcriptional misregulation in cancer	7.72E-04	0.013454985
hsa05323:Rheumatoid arthritis	0.00129	0.021136644
hsa00051:Fructose and mannose metabolism	0.001356	0.021051205
hsa00480:Glutathione metabolism	0.00145	0.021393467

mean expression difference,  $v_g \sim N(0, \sigma_{diff}^2)$  is the gene bias of the expression difference, and  $\epsilon_{gs} \sim N(0, \sigma_{derr}^2)$  is the gene-study interaction of the expression difference. We used two sets of the parameters  $(\mu, \sigma_{gene}^2, \sigma_{study}^2, \sigma_{int}^2, \sigma_{err}^2, \delta, \sigma_{diff}^2, \sigma_{derr}^2)$ : A) (5, 1.25, 0.49, 0.25, 0.16, 0.8, 0.0016, 0.256) and B) (5, 6.25, 0.49, 0.25, 0.16, 0.8, 0.0016, 0.256). Compared with A, B increases only the gene effect but retain other effects for investigating the influence of gene effect. In summary, four data scenarios were synthesized: Simulation I with parameter setting A (Simulation-IA) or parameter setting B (Simulation-IB), Simulation II with parameter setting A (Simulation-IIA) or parameter setting B (Simulation-IIB). For each data scenario, twenty data sets were randomly generated in the experiment and average results over them were used for algorithm evaluation.

**Simulation data analysis**

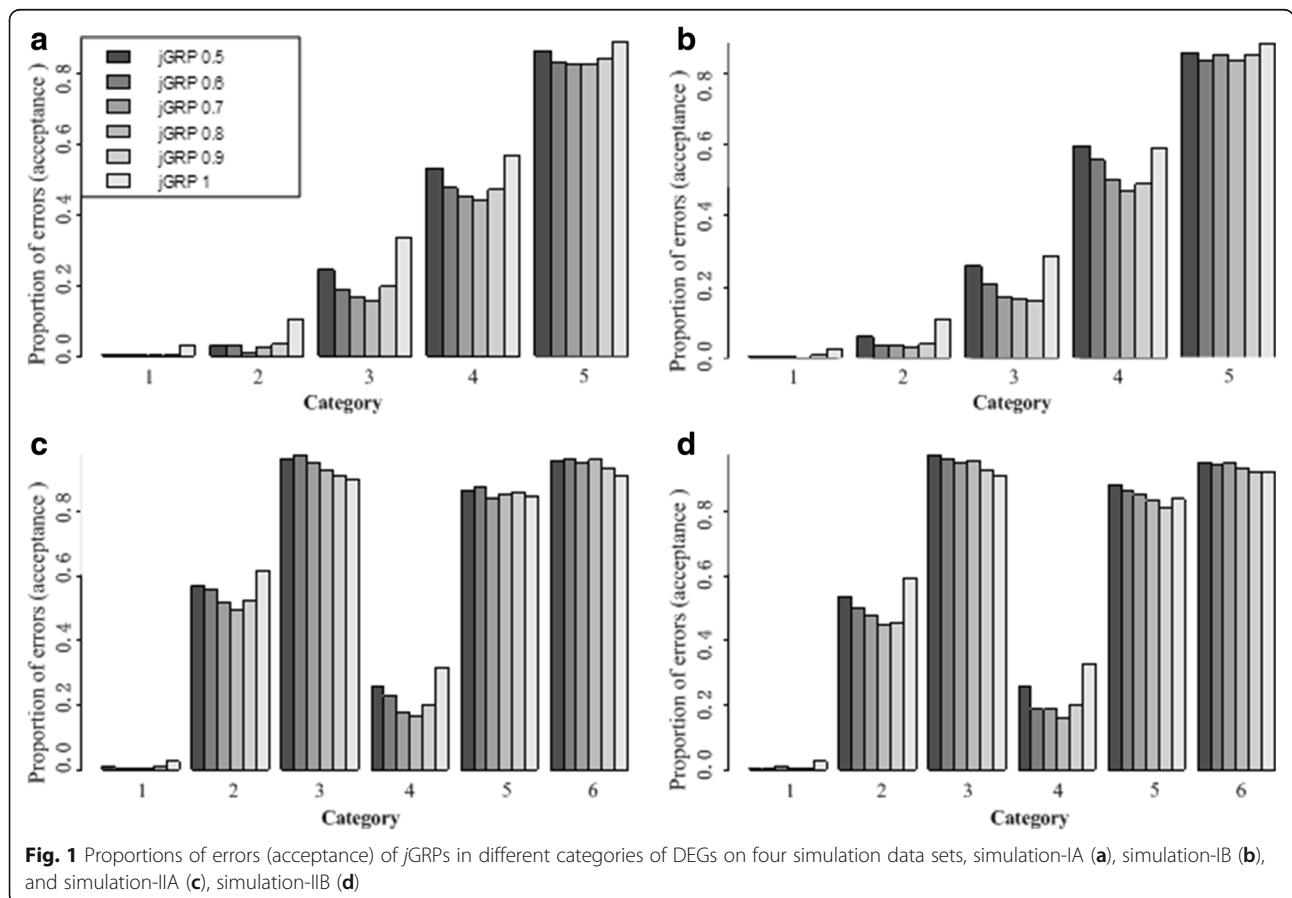
Considering the importance of the regulation confidence cutoff parameter  $\tau$  to the performance of  $jGRP$ , we varied  $\tau = 0.5, 0.6, 0.7, 0.8, 0.9, 1$  and repeatedly applied  $jGRP$  to analyze the simulation data. To control

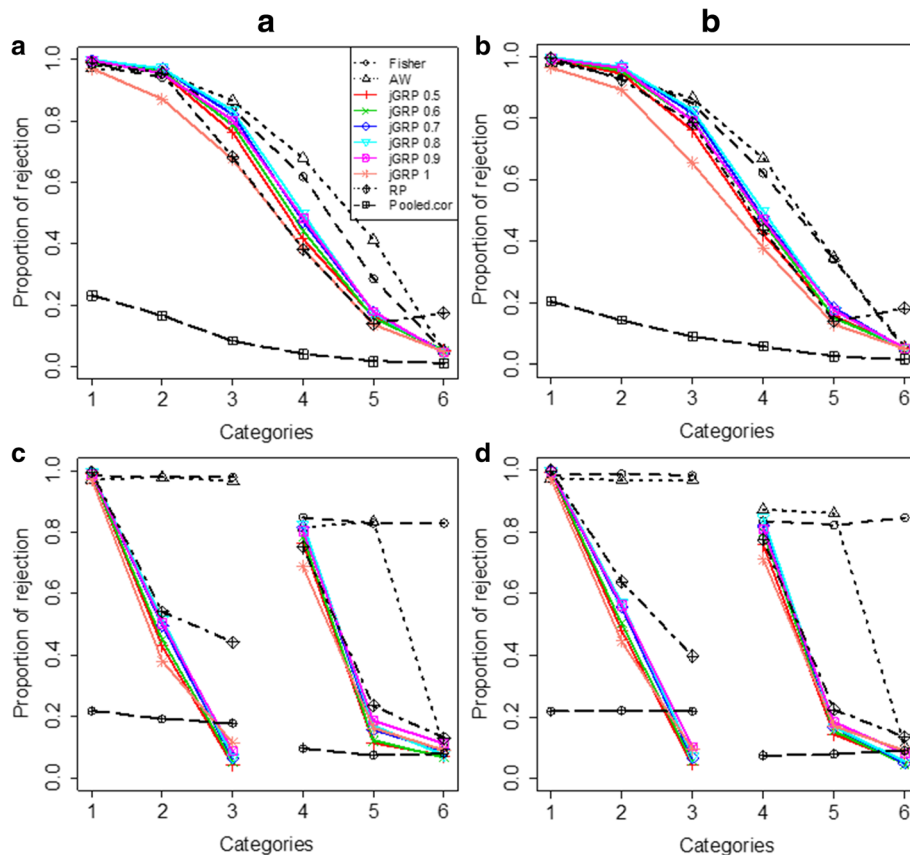
false positive rates (FPR), the resulted  $p$ -values were corrected using the Benjamini-Hochberg (BH) procedure [22, 23]. Figure 1 summarizes the proportions of errors (acceptance) in each category of genes at an ad hoc BH-adjusted- $p$ -value cutoff of 0.05 in the four data scenarios. From this figure, it can be found that, generally, too large or too small values of  $\tau$  led to large errors, irrespective of any of the four data scenarios, as expected. The parameter  $\tau$  directly controls the regulation confidence and captures the variation of differential expression across studies. Theoretically, too small  $\tau$  can not filter out noise or non-specific heterogeneity such that DEGs will be recognized in a low confidence, leading to spurious DEGs, while too large  $\tau$  means a too stringent control of study heterogeneity such that intra-class biological heterogeneity per se is excluded, missing true DEGs with complex patterns of differential expression. Relative to Simulation-IA, Simulation-IB have an increased gene effect, which led to slightly larger  $\tau$  (around 0.8), at which the errors reach to the lowest, than that for simulation-IB (around 0.7) as shown in Fig. 1a-b. Similar results were observed between the two scenarios of Simulation-II, as shown in Fig. 1c-d.

Results also revealed that the error proportion gradually increases from Category 1 to 5 in both data scenarios of

Simulation-I, as shown in Fig. 1a-b. This is consistent with the increasing heterogeneity of differential expression from Category 1 to 5. Similar phenomena were observed for Simulation-II (Fig. 1c-d). In Simulation-II, genes could be differentially expressed in different directions across studies, which produces additional heterogeneity for DEGs identification. Specifically, the heterogeneity increases from Category 1 to 3 and from Category 4 to 6. From Fig. 1c-d, we can clearly see that the error proportion gradually increases in a corresponding way across these categories, irrespective of Simulation-IIA or Simulation-IB. In summary, these results show that the proposed method can deal with various types of data heterogeneity across studies in a controllable way.

For comparison evaluation, we also applied previous methods, Fisher's [24], AW [10], RankProd (RP) [25] and pooled cor [21], to analyze the simulation data. Two R packages, MetaDE and RankProd, were called to implement the two previous methods, AW and RP, respectively. For AW, the modt model was set (as default) to calculate the  $p$ -values for individual study and the fudge parameter to be the median variability estimator. Figure 2 compares the proportions of rejection (DEGs called) by  $j$ GRP at a BH-adjusted- $p$ -value cutoff of 0.05 with those by the four previous method in the four data





**Fig. 2** Comparison of the rejection proportions of jGRPs with those of previous methods on four simulation data sets, simulation-IA (a), simulation-IB (b), and simulation-IIA (c), simulation-IIB (d)

scenarios. As described above, study heterogeneity gradually increases from Categories 1 to 5 in the two scenarios of Simulation-I and from Categories 1(4) to 3(6) in the two scenarios of Simulation-II. It is expected that a reasonable method should be sensitive to the change of heterogeneity and have the proportions of rejection gradually drop as the heterogeneity increases across the categories in all the four data scenarios accordingly. From Fig. 2, we can clearly see that although jGRPs as well as the previous methods all are sensitive to the change of heterogeneity, they have different degrees of sensitivity in different simulation scenarios. Generally, the *p*-value-based methods led to the two extremes among these methods: Fisher’s and AW are least sensitive, while pooled cor is most sensitive. Especially, pooled cor seems too stringent to miss some DEGs that are even consistently differentially expressed across all the ten studies (Category 1) in all the four data scenarios. Lying in between the two extremes, jGRPs seems to be reasonably sensitive with a mild result in all the four data scenarios, and the sensitivity changes with the regulation confidence parameter in a controllable way: the larger or smaller the parameter the more sensitive jGRP. Results also reveals

that RP is less sensitive to inconsistent expression patterns (Fig. 2c-d), which is consistent with the observations in [12]. In summary, jGRP shows a superior power of dealing with various types of study heterogeneity.

**Application to real microarray expression data**

Considering that lung cancer is one of the most malignant tumors worldwide, we collected three real microarray lung adenocarcinoma (LUAD) cancer datasets from the GEO database: Landi’s data (GSE10072) [26], Selamat’s data (GSE32863) [27], and Su’s data (GSE7670) [28], in which all samples were divided into lung adenocarcinoma and normal (NTL). The Landi’s data consist of the expression levels of ~13,000 probes in total 107 (58 LUAD and 49 NTL) samples; The Selamat’s data consist of the expression levels of ~25,000 probes in total 117 (58 LUAD and 59 NTL) samples; The Su’s data consist of the expression levels of ~13,000 probes in 54 (27 paired LUAD/NTL) samples. During generating these datasets, different microarray platforms were used to measure gene expression levels in parallel: Illumina Human WG-6 v3.0 Expression BeadChips for Landi’s data, HG-U133A Affymetrix chips for Selamat’s data, and Affymetrix Human Genome

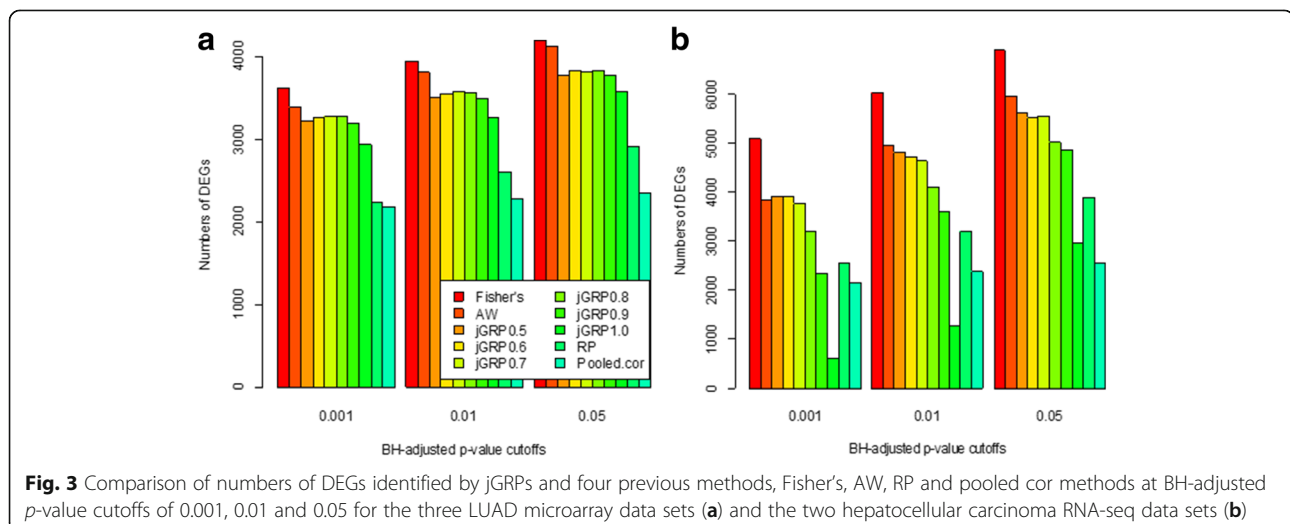
U133A array for Su's data, which complicated data heterogeneity across these studies. We preprocessed the three datasets according to the following procedure: Averaging the intensities of multiple probes matching a same Entrez ID as the expression levels of the corresponding gene, and filtering out non-specific or noise genes by a CV filter (setting the CV cutoff as 0.05) [29]. As a result, the expression levels of 4728 common genes were used for meta-analysis for detecting LUAD-related DEGs.

We applied *j*GRPs with varying  $\tau = 0.5, 0.6, 0.7, 0.8, 0.9$  and 1 to analyze the three data sets simultaneously. To control false positive rates (FPRs), the resulting *p*-values for each gene were corrected using Benjamini-Hochberg (BH) procedure [22, 23]. For comparison, four previous methods, Fisher's [24], AW [10], RP [25] and Pooled cor [21], were also applied to re-analyze these data sets. Figure 3a shows the numbers of DEGs by *j*GRP and the previous methods at three BH-corrected *p*-value cutoffs of 0.001, 0.01 and 0.05. From this figure, it can be clearly seen that our *j*GRP methods obtained a moderate result between the two previous methods, which is consistent with those on the simulation data above. Furthermore, for *j*GRPs, varying  $\tau$  resulted in a similar changing pattern of the number of identified DEGs to those for the simulation data above, and  $\tau = 0.7$  obtained the largest and seemingly more reasonable number of DEGs.

Results show that 3281 genes were called significantly differentially expressed between normal and LUAD tissues by *j*GRP ( $\tau = 0.7$ ) at an ad hoc BH-adjusted *p*-value cutoff of 0.001. Literature survey shows that many of these DEGs were previously reported to be related to lung cancer. For example, the gene with the largest value of *j*GRP (1), EPAS1, plays important roles in cancer progression and has been widely reported to be over-expressed in non-small cell lung cancer (NSCLC) tissues

as a significant marker for poor prognosis [30, 31]. Other researchers have evidenced that in murine models of lung cancer, high expression levels of EPAS1 relate to tumor of large size, invasion and angiogenesis [32, 33].

One unique feature of *j*GRP is to automatically label DEGs with up-regulation or down-regulation in cancer. As a result, the 3281 DEGs were further divided by *j*GRP into two categories with different regulatory directions: 1655 (Additional file 1: Table S1) were with a negative *j*GRP statistic meaning a down-regulation in LUAD tissues relative to normal tissues, and 1626 (Additional file 1: Table S2) with a positive *j*GRP statistic meaning an up-regulation in LUAD. Among the 1655 down-regulated genes, many have been previously reported to be lowly expressed in lung tumor. For example, gene MTRR, which was missed by all the four previous method, Fisher's, AW, RP and Pooled.cor, at an ad hoc BH-adjusted *p*-value cutoff of 0.001, was found with *j*GRP = -0.36 (*p*-value <  $3 \times 10^{-5}$ , BH-corrected *p*-value <  $5 \times 10^{-5}$ ) to significantly down-regulated in LUAD. For this gene, Aksoy-Sagirli et al. [34] recently reported that its single-nucleotide polymorphism, MTRR 66 A > G, is significantly associated with lung cancer risk. Another gene, FAM107A with a large value of *j*GRP = -0.99 (*p*-value <  $10^{-16}$ , BH-corrected *p*-value <  $10^{-16}$ ), also named DRR1 and TU3A, is the member A of the family with sequence similarity 107, localized in chromosomal region 3p21.1 and ~10 kb long. Biologically, the protein that FAM107A encodes is involved in cell cycle regulation via apoptosis induction. It has been evidenced that FAM107A is frequently lost in various types of cancer, including ovarian cancer, cell carcinoma (RCC), prostate cancer and lung cancer cell lines [35, 36]. Recently, Pastuszak-Lewandoska et al. [37] observed that FAM107A was dramatically down-regulated in NSCLC samples relative to in tumor adjacent normal



**Fig. 3** Comparison of numbers of DEGs identified by *j*GRPs and four previous methods, Fisher's, AW, RP and pooled cor methods at BH-adjusted *p*-value cutoffs of 0.001, 0.01 and 0.05 for the three LUAD microarray data sets (a) and the two hepatocellular carcinoma RNA-seq data sets (b)

tissues. Gene TCF21 with a large value of  $jGRP$  ( $jGRP = -0.99$ ,  $p$ -value  $< 10^{-16}$ , BH-corrected  $p$ -value  $< 10^{-16}$ ), which encodes a transcription factor of the basic helix-loop-helix family, was extensively observed as tumor suppressor to under-express in human malignancies. Especially, Wang et al. [17] reported that the underrepresentation of TCF21 in LUAD tissues may be driven by its hypermethylation. The epigenetic inactivation in lung cancer was experimentally observed by Smith et al. [38] using restriction landmark genomic scanning. Recently, Shivapurkar et al. [39] adopted DNA sequencing technique to narrow down the sequence of TCF21 and pinpointed a short CpG-rich segment in the CpG island within exon 1 that is predominantly methylated in lung cancer cell lines but unmethylated in normal epithelial cells of lung. The short segment may account for the TCF21 expression abnormality in lung cancer. A more evidence reported by Richards et al. is that the association between hypermethylation and under-expression of TCF21 is specific to tumor tissues and occurs very frequently in various types of non-small cell lung cancer (NSCLC), even in the early-stage of NSCLC [40]. Taken together, these evidences confirm the down-regulation pattern of TCF21 in LUAD and suggest that it may be driven by its hypermethylation.

Among the 1626 up-regulated genes, many have also been previously reported to be under-expressed in lung cancer. For example, gene STRN3, which was missed by RP and Pooled.cor at an ad hoc BH-adjusted  $p$ -value cutoff of 0.001, was found to be up-regulated in LUAD with  $jGRP = 0.32$  ( $p$ -value  $< 5 \times 10^{-4}$ , BH-corrected  $p$ -value  $< 7 \times 10^{-4}$ ). As a single marker, STRN3 efficiently distinguished 100 NSCLC patients from 147 control subjects with a sensitivity of 84% and a specificity of 81%, and was included into a membrane array-based assay for non-invasive diagnosis of patients with NSCLC [41]. Another gene COL11A1 with a large value of  $jGRP$  ( $jGRP = 0.97$ ,  $p$ -value  $< 10^{-16}$ , BH-corrected  $p$ -value  $< 10^{-16}$ ) has been previously reported to take part as a minor fibrillar collagen in cell proliferation, migration and the tumorigenesis of many human malignancies. For example, Shen et al. [42] experimentally observed that the gene was significantly up-regulated in recurrent NSCLC tissues and in NSCLC with lymph node metastasis. It has been revealed that Smad signaling functionally mediates the overexpression of COL11A1 in NSCLC cells during the cell proliferation, migration and invasion of NSCLC cell lines in vitro. COL11A1 can act as a biomarker for clinical diagnosis of metastatic NSCLC [42]. For gene HMGA1 ( $jGRP = 0.97$ ,  $p$ -value  $< 10^{-16}$ , BH-corrected  $p$ -value  $< 10^{-16}$ ), the two previous methods, pooled cor and Fisher's, ranked it at 183th and after 1000, respectively. Biologically, HMGA1 encodes a protein that is functionally associated with chromatin, which is involved in

the metastatic progression of cancer cells. Previous studies reported that HMGA1 is widely over-expressed in a variety of aggressive tumors, suggesting that HMGA1 may act as a convictive biomarker for NSCLC prognostic prediction [43]. Especially, using immunohistochemistry, Zhang et al. [44] found that increased protein levels of HMGA1 are positively correlated with the status of clinical stage, classification of T, N and M, and differentiated degree in NSCLC.

To further assess the DEGs identified by different methods, we also performed pathway enrichment analysis using the commonly used online DAVID tool (<http://david.abcc.ncifcrf.gov/home.jsp>). As a result, DAVID reported 42, 57, 53, 40, 20 KEGG pathways (Additional file 1: Table S3-S7) significantly enriched in the DEG lists of  $jGRP$  ( $\tau = 0.7$ ) and four previous methods, Fisher's, AW, RP and Pooled cor, at an ad hoc  $p$ -value cutoff of 0.05, respectively. Compared with the previous methods,  $jGRP$  gave higher ranks to pathways that are related to cancer progression, including cell cycle (Rank 2) comprised of a series of cellular events that leads to the division and duplication of DNA (DNA replication) of a cell, and small cell lung cancer (Rank 6), as shown in Table 2. Especially, the Complement and coagulation cascades pathway ranked at 1 was recently reported to dysfunction in lung cancer [45, 46].  $jGRP$  also called another two lung cancer-related pathways, NF-kappa B signaling pathway and PI3K-Akt signaling pathway, but pooled cor did not. In NF-kappa B signaling pathway, nuclear factor- $\kappa$ B (NF $\kappa$ B) is a family of transcription factors that regulate the expression of genes that are involved in cell proliferation, differentiation and inflammatory responses. It has been widely evidenced that activating F $\kappa$ B can induce tumorigenesis of normal cells [47–49].

#### Application to RNA-seq expression data

We also evaluated the performance of the proposed method on RNA-seq expression data. Hepatocellular carcinoma (HCC) is the third leading cause of cancer-related deaths. Two HCC RNA-seq data sets were collected from the GEO database: Liu's data (GSE77314) [50] and Dong's data (GSE77509) [51], both of which were measured using Illumina Hiseq 2000, and jointly analyzed them for identifying HCC biomarkers. The former consists of mRNA profiles of 50 paired normal and HCC samples, and the latter consists of mRNA profiles of 40 matched HCC patients and adjacent normal tissues. For quality control, we preprocessed the two datasets by averaging the FPKM values with a same Entrez ID as the expression levels of the corresponding gene and filtering out non-specific or noise genes based on a CV filter [29]. As a result, two HCC expression data sets containing 4945 common genes were jointly analyzed for identifying HCC-related DEGs.



Similar to the three LUAD microarray data sets, we applied  $j$ GRPs with varying  $\tau = 0.5, 0.6, 0.7, 0.8, 0.9, 1$  and the four previous methods, Fisher's [24], AW [10], RP [25] and Pooled cor [21], to jointly analyze the two RNA-seq data sets, respectively, and corrected  $p$ -values using Benjamini-Hochberg (BH) procedure [22, 23] for controlling false positive rates. Figure 3b shows the numbers of DEGs called by  $j$ GRPs and the previous methods at three BH-corrected  $p$ -value cutoffs of 0.001, 0.01 and 0.05. Similar to Fig. 3a, b reveals that most of  $j$ GRPs obtained an intermediate result between those by the previous methods, Fisher's, AW and Pooled cor, for the HCC RNA-seq data. Among the  $j$ GRPs, the one with  $\tau = 0.6$ , which is smaller than 0.7 for the LUAD data sets above, led to a more reasonable result, implying that it is more heterogeneous across the two HCC data sets than that across the three LUAD data sets. The high heterogeneity may be the reason for the unusually large numbers of DEGs by RP which is less sensitive to inconsistent patterns of expression [12].

Totally, there were 1724 genes called significantly differential expressed between normal and HCC tissues by  $j$ GRP ( $\tau = 0.6$ ) at a BH-adjusted  $p$ -value cutoff of 0.001. Among them, 1206 (Additional file 1: Table S8) were with a negative  $j$ GRP statistic, i.e., a down-regulation in HCC tissues relative to normal tissue, and 518 (Additional file 1: Table S9) with a positive  $j$ GRP statistic, i.e., an up-regulation in HCC. The imbalance of up- and down-regulated genes informed a higher degree of heterogeneity across the two HCC data sets compared with that across the three LUAD data sets (1655 down-regulated DEGs and 1626 up-regulated DEGs), which is in concordance with the unusually larger numbers of DEGs by RP. Then, we examined the biological functions of the two sets of DEGs. Literature survey shows that many of them have been previously reported to relate to HCC or cancer. For example, one of down-regulated DEGs, *Nat2*, with  $j$ GRP = -1,  $p$ -value  $< 10^{-16}$  and BH-corrected  $p$ -value  $< 10^{-16}$ , can both activate and deactivate arylamine and hydrazine drugs and carcinogens. Some polymorphisms in *Nat2* have been previously reported to increase the risk of HCC and drug toxicity [52, 53]. Recently, it has been widely observed that *Nat2* are consistently and stably down-regulated in more than three hundred HCC patients [54]. One of up-regulated DEGs, *CDC20*, with  $j$ GRP = 1,  $p$ -value  $< 10^{-16}$ , and BH-corrected  $p$ -value  $< 10^{-16}$ , biologically acts as a regulatory unit in cell cycle that interacts with several proteins at multiple points of cell cycle. Li et al. [55] reported that high expression of *CDC20* is associated with development and progression of hepatocellular carcinoma. Recently, *CDC20* has been suggested to be a potential novel cancer therapeutic target [56]. We also conducted pathway analysis using the DAVID tool on the 1724 DEGs. As a result,

39 KEGG pathways (Additional file 1: Table S10) were called to be significantly enriched in the DEG list at an ad hoc  $p$ -value cutoff of 0.05, many of which were previously found to be involved in tumorigenesis, e.g., cell cycle and p53 signaling pathway. Especially, a new pathway, i.e., Bile secretion pathway, was found to be significantly enriched and relate to HCC ( $p$ -value =  $2.5 \times 10^{-8}$ ), which though needs to be further investigated by pathologists. Biologically, Bile is a vital secretion, which is essential in digesting and absorbing fats and fat-soluble vitamins in the small intestine. There are two mechanisms that influence Bile secretion: membrane transport systems in hepatocytes and cholangiocytes and the structural and functional integrity of the biliary tree. The dysfunction of the two mechanisms may cause the signaling abnormality of the Bile secretion pathway in HCC.

## Discussion

The central problem in transcriptomics data meta-analysis is how to deal with study heterogeneity. The heterogeneity complicates the distribution of gene expression and thus hinders accurately pinpointing the concordance of differential expression across studies. Two intuitive alternative approaches for data integration could be 1) Directly use the information contained in several data-sets; and 2) Cluster higher/lower expressed genes in each data-set and then zoom in on the interesting genes. However, they both ignore or inappropriately deal with the gene expression heterogeneity problem between studies. Currently, most methods for meta-analysis of differential expression directly operate in gene expression space, which are based on either  $p$ -values, ranks, or hierarchical  $t$ -statistic models. The proposed method,  $j$ GRP, at the first time establishes a universal and flexible integrative framework that operates in gene regulation space instead of in gene expression space, in which individual samples from different sources are more compatible. The regulation profile for a sample is derived from its expression profile based on probabilistic theory, where biological variability and noise inherent in gene expression data are modeled efficiently in combination with an adjustable parameter. It is also intuitive and simple to implement and easy to use in practice. We expect that this work can promote a research interest in borrowing gene regulation knowledge for integrative identification of DEGs.

The regulation confidence cutoff parameter  $\tau$  reflects a tradeoff between regulation confidence and noise accommodation and is of importance to the performance of  $j$ GRP. How to properly choose the parameter is still an open question. The choice should be conditional on the study heterogeneity at hand. Here, we would like to recommend 0.7 as default for the parameter for simplicity or to try different values among 0.5 and 1 and then choose a proper value, depending on a particular data condition.

## Conclusions

We have presented a novel transcriptomic data meta-analysis method, *jGRP*, for identifying differentially expressed genes. The method integrates multiple gene expression data sets in a gene regulatory space instead of in the original gene expression space, which makes it easy to relieve the data heterogeneity between cross-lab or cross-platform studies. To produce the regulatory space, two gene regulation events between two conditions were mathematically defined, whose occurring probabilities were estimated using the total probabilistic theorem. Based on the resulting gene regulation profiles, a novel statistic, *jGRP*, was established to measure the differential expression of a gene in the regulatory space. *jGRP* introduces a parameter ( $\tau$ ) for users to conveniently adjust to fit into various levels of study heterogeneity in practice. Compared with existing methods, *jGRP* provides a united and flexible framework for DEGs identification in a meta-analysis context and is intuitive and simple to implement in practice, which can be a stand-alone tool due to the superior power of dealing with study heterogeneity. We evaluated the proposed method on simulation data and real-world microarray and RNA-seq gene expression data sets, and experimental results demonstrate the effectiveness and efficiency of *jGRP* for DEGs identification in gene expression data meta-analysis. Future work will be focused on guidelines for the choice of the regulation confidence cutoff parameter and biological verification of the new DEGs identified in the real applications.

## Additional file

**Additional file 1: Table S1.** List of 1655 genes with a negative *jGRP* statistic meaning a down-regulation in LUAD tissues relative to normal tissues on the three LUAD data sets. **Table S2.** List of 1626 genes with a positive *jGRP* statistic meaning a up-regulation in LUAD tissues relative to normal tissues on the three LUAD data sets. **Table S3.** List of 42 KEGG pathways significantly enriched in the DEG lists of *jGRP* ( $\tau = 0.7$ ) by DAVID. **Table S4.** List of 57 KEGG pathways significantly enriched in the DEG lists of Fisher's by DAVID. **Table S5.** List of 53 KEGG pathways significantly enriched in the DEG lists of AW by DAVID. **Table S6.** List of 40 KEGG pathways significantly enriched in the DEG lists of RP by DAVID. **Table S7.** List of 20 KEGG pathways significantly enriched in the DEG lists of Pooled cor by DAVID. (RAR 259 kb)

## Acknowledgements

None.

## Funding

This work was supported in part by the National Natural Science Foundation of China (Nos. 61374181, 61402010); the Anhui Province Natural Science Foundation (1408085MF133); K. C. Wong education foundation.

## Availability of data and materials

The three LUAD data sets can be downloaded from <http://www.ncbi.nlm.nih.gov/geo/>; Tables S1–7 are available in the supplemental files. R code for *jGRP* is available for non-commercial use at the URL <http://miclab.iim.ac.cn/>.

## Authors' contributions

XPX and YFX performed data analysis experiments and XPX and HQW designed the study and drafted the manuscript. All authors read and approved the final manuscript.

## Ethics approval and consent to participate

Not applicable

## Consent for publication

Not applicable

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>School of Mathematics and Physics, Anhui Jianzhu University, Hefei, Anhui 230022, China. <sup>2</sup>Cancer Hospital, CAS, Hefei, Anhui 230031, China. <sup>3</sup>MICB Lab, Hefei Institutes of Physical Science, CAS, Hefei 230031, China.

Received: 28 January 2017 Accepted: 15 August 2017

Published online: 23 August 2017

## References

- Ghazani AA, Oliver NM, St. Pierre JP, Garofalo A, Rainville IR, Hiller E, Treacy DJ, Rojas-Rudilla V, Wood S, Bair E, et al. Assigning clinical meaning to somatic and germ-line whole-exome sequencing data in a prospective cancer precision medicine study. *Genet Med*. 2017;19(7):787–95.
- H-j S, Chen J, Ni B, Yang X, Wu Y-Z. Recent advances and current issues in single-cell sequencing of tumors. *Cancer Lett*. 2015;365(1):1–10.
- Jiao Y, Widschwendter M, Teschendorff AE. A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*. 2014;30(16):2360–6.
- Neapolitan R, Horvath C, Jiang X. Pan-Cancer analysis of TCGA data reveals notable signaling pathways. *BMC Cancer*. 2015;15(1):516.
- TCGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 2012;487(7407):330–7.
- Natrajan R, Wilkerson P. From integrative genomics to therapeutic targets. *Cancer Res*. 2013;73(12):3483–8.
- Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res*. 2013;41(Database issue):D991.
- Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, et al. Array express update—simplifying data submissions. *Nucleic Acids Res*. 2015;43(Database issue):D1113–6.
- Rung J, Brazma A. Reuse of public genome-wide gene expression data. *Nat Rev Genet*. 2013;14(2):89–99.
- Li J, Tseng GC. An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *Ann Appl Stat*. 2011;5(2A):994–1019.
- Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res*. 2012;40(9):3785–99.
- Li Y, Ghosh D. Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data. *Bioinformatics*. 2012;28(6):807–14.
- Hong F, Breitling R. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008;24(3):374–82.
- Breitling FHaR. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*. 2008;24:374–82.
- Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock REW. INMEX: a web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res*. 2013;41(W1):W63–70.
- Chang L-C, Lin H-M, Sibille E, Tseng G. Meta-analysis methods for combining multiple expression profiles: comparisons, statistical

- characterization and an application guideline. *BMC Bioinformatics*. 2013; 14(1):368.
17. Wang H-Q, Zheng C-H, Zhao X-M. jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data. *Bioinformatics*. 2015; 31(4):572–80.
  18. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733–9.
  19. Dai M, Wang P, Boyd A, Kostov G, Athey B, Jones E, Bunney W, Myers R, Speed T, Akil H. Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*. 2005;33(20):e175.
  20. Wang H-Q, Huang D-S. Regulation probability method for gene selection. *Pattern Recogn Lett*. 2006;27(2):116–22.
  21. Li Y, Ghosh D. Assumption weighting for incorporating heterogeneity into meta-analysis of genomic data. *Bioinformatics*. 2012;28:807–14.
  22. Kim KI, van de Wiel M. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*. 2008;9(1):114.
  23. Wang H-Q, Tuominen LK, Tsai C-J. SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*. 2011;27(2):225–31.
  24. Song C, Tseng GC. Hypothesis setting and order statistics for robust genomic meta-analysis. *Ann Appl Stat*. 2014;8(2):777–800.
  25. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*. 2006;22(22):2825–7.
  26. Landi MT, Dracheva T, Rotunno M, Figueroa JD, Liu H, Dasgupta A, Mann FE, Fukuoka J, Hames M, Bergen AW, et al. Gene expression signature of cigarette smoking and its role in lung Adenocarcinoma development and survival. *PLoS One*. 2008;3(2):e1651.
  27. Selamat SA, Chung BS, Girard L, Zhang W, Zhang Y, Campan M, Siegmund KD, Koss MN, Hagen JA, Lam WL, et al. Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res*. 2012;22(7):1197–211.
  28. Su L-J, Chang C-W, Wu Y-C, Chen K-C, Lin C-J, Liang S-C, Lin C-H, Whang-Peng J, Hsu S-L, Chen C-H, et al. Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*. 2007;8(1):140.
  29. Li S, Li D. DNA microarray technology and data analysis in cancer research. Singapore: World Scientific Publishing; 2008.
  30. Giatromanolaki A, Koukourakis MI, Sivridis E, Turley H, Talks K, Pezzella F, Gatter KC, Harris AL. Relation of hypoxia inducible factor 1[agr] and 2[agr] in operable non-small cell lung cancer to angiogenic/molecular profile of tumours and survival. *Br J Cancer*. 2001;85(6):881–90.
  31. Han SS, Yeager M, Moore LE, Wei M-H, Pfeiffer R, Toure O, Purdue MP, Johansson M, Scelo G, Chung CC, et al. The chromosome 2p21 region harbors a complex genetic architecture for association with risk for renal cell carcinoma. *Hum Mol Genet*. 2011;21(5):1190–200.
  32. Kim WY, Perera S, Zhou B, Carretero J, Yeh JJ, Heathcote SA, Jackson AL, Nikolinos P, Ospina B, Naumov G. HIF2alpha cooperates with RAS to promote lung tumorigenesis in mice. *J Clin Invest*. 2009;119(8):2160.
  33. Mazumdar J, Hickey MM, Pant DK, Durham AC, Sweet-Cordero A, Vachani A, Jacks T, Chodosh LA, Kissil JL, Simon MC. HIF-2alpha deletion promotes Kras-driven lung tumor development. *Proc Natl Acad Sci U S A*. 2010; 107(32):14182.
  34. Aksoy-Sagirlı P, Erdenay A, Kaytan-Saglam E, Kizir A. Association of Three Single-Nucleotide Polymorphisms in MTR and MTRR genes with lung cancer in a Turkish population. *Genet Test Mol Biomarkers*. 2017; 21(7):428–32.
  35. Liu Q, Zhao XY, Bai RZ, Liang SF, Nie CL, Yuan Z, Wang CT, Wu Y, Chen LJ, Wei YQ. Induction of tumor inhibition and apoptosis by a candidate tumor suppressor gene DRR1 on 3p21.1. *Oncol Rep*. 2009;22(5):1069–75.
  36. Zhao X-Y, Liang S-F, Yao S-H, Ma F-X, Hu Z-G, Yan F, Yuan Z, Ruan X-Z, Yang H-S, Zhou Q, et al. Identification and preliminary function study of Xenopus laevis DRR1 gene. *Biochem Biophys Res Commun*. 2007;361(1):74–8.
  37. Pastuszek-Lewandoska D, Czarnecka KH, Migdalska-Sk M, Nawrot E, Domaska D, Kiszakiewicz J, Kordiak J, Antczak A, Grski P, Brzezińska-Lasota E. Decreased FAM107A expression in patients with non-small cell lung cancer. In: Pokorski M, editor. *Respiratory Carcinogenesis*, vol. 852. Cham: Springer International Publishing; 2015. p. 39–48.
  38. Smith LT, Lin M, Brena RM, Lang JC, Schuller DE, Otterson GA, Morrison CD, Smiraglia DJ, Plass C. Epigenetic regulation of the tumor suppressor gene TCF21 on 6q23-q24 in lung and head and neck cancer. *Pro Natl Acad Sci USA*. 2006;103(4):982–7.
  39. Shivapurkar N, Stastny V, Xie Y, Prinsen C, Frenkel E, Czerniak B, Thunnissen FB, Minna JD, Gazdar AF. Differential Methylation of a short CpG-rich sequence within exon 1 of TCF21 gene: a promising cancer biomarker assay. *Cancer Epidemiol Biomark Prev*. 2008;17(4):995–1000.
  40. Richards KL, Zhang B, Sun M, Dong W, Churchill J, Bachinski LL, Wilson CD, Baggerly KA, Yin G, Hayes DN, et al. Methylation of the candidate biomarker TCF21 is very frequent across a spectrum of early-stage non-small cell lung cancers. *Cancer*. 2010;117(3):606–17.
  41. Sheu C-C, Yu Y-P, Tsai J-R, Chang M-Y, Lin S-R, Hwang J-J, Chong I-W. Development of a membrane array-based multimarker assay for detection of circulating cancer cells in patients with non-small cell lung cancer. *Int J Cancer*. 2006;119(6):1419–26.
  42. Shen L, Yang M, Lin Q, Zhang Z, Zhu B, Miao C. COL11A1 is overexpressed in recurrent non-small cell lung cancer and promotes cell proliferation, migration, invasion and drug resistance. *Oncol Rep*. 2016;36(2):877–85.
  43. Huso TH, Resar LMS. The high mobility group A1 molecular switch: turning on cancer - can we turn it off? *Expert Opin Ther Targets*. 2014;18(5):541–53.
  44. Zhang Z, Wang Q, Chen F, Liu J. Elevated expression of HMGA1 correlates with the malignant status and prognosis of non-small cell lung cancer. *Tumor Biol*. 2014;36(2):1213–9.
  45. Wang L, Zhao L, Qiao Y. Identification of potential therapeutic targets for lung cancer by bioinformatics analysis. *Mol Med Rep*. 2016;13(3):1975–82.
  46. Zhang W, Zhang Q, Zhang M, Zhang Y, Li F, Lei P. Network analysis in the identification of special mechanisms between small cell lung cancer and non-small cell lung cancer. *Thorac Cancer*. 2014;5(6):556–4.
  47. Baldwin JAS. THE NF- $\kappa$ B AND I $\kappa$ B PROTEINS: new discoveries and insights. *Ann Rev Immunol*. 1996;14(1):649–83.
  48. Hamdane M, Davidcordonnier MH, D'Halluin JC. Activation of p65 NF- $\kappa$ B protein by p210BCR-ABL in a myeloid cell line (P210BCR-ABL activates p65 NF- $\kappa$ B). *Oncogene*. 1997;15(19):2267–75.
  49. Kirchner D, Duyster J, Ottmann O, Schmid RM, Bergmann L, Munzert G. Mechanisms of Bcr-Abl-mediated NF- $\kappa$ B/Rel activation. *Exp Hematol*. 2003; 31(6):504–11.
  50. Gang L, Hou G, Liang L, Li Y, Zhou W, Lei L. Potential diagnostic and prognostic marker dimethylglycine dehydrogenase (DMGDH) suppresses hepatocellular carcinoma metastasis in vitro and in vivo. *Oncotarget*. 2016; 7(22):32607–16.
  51. Yang Y, Chen L, Gu J, Zhang H, Yuan J, Lian Q, Lv G, Wang S, Wu Y, Yang YT. Recurrently deregulated lncRNAs in hepatocellular carcinoma. *Nat Commun*. 2017;8:14421.
  52. Zhang X, Lu X, Shi Q, Xu X-Q, Leung H-C, Harris L, Iglehart J, Miron A, Liu J, Wong W. Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*. 2006;7(1):197.
  53. Farker K, Schotte U, Scheele J, Hoffmann A. Impact of N-acetyltransferase polymorphism (NAT2) in hepatocellular carcinoma (HCC) – an investigation in a department of surgical medicine. *Exp Toxicol Pathol*. 2003;54(5):387–91.
  54. Yin F, Shu L, Xia L, Li T, Tao P, Nan Y, Shu L, Zeng X, Qiu X. Microarray-based identification of genes associated with cancer progression and prognosis in hepatocellular carcinoma. *J Exp Clin Cancer Res*. 2016;35(1):127.
  55. Li J, Gao J, Du J, Huang Z, Wei L. Increased CDC20 expression is associated with development and progression of hepatocellular carcinoma. *Int J Oncol*. 2014;45(4):1547–55.
  56. Greenman C, Stephens P, Smith R, Dalgleish GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, et al. Patterns of somatic mutation in human cancer genomes. *Nature*. 2007;446(7132):153–8.