BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 4 2015, pages 572–580
doi:10.1093/bioinformatics/btu679

Systems biology

Advance Access publication October 16, 2014

jNMFMA: a joint non-negative matrix factorization meta-analysis of transcriptomics data

Hong-Qiang Wang[1,*], Chun-Hou Zheng[2] and Xing-Ming Zhao[3]

[1]Machine Intelligence and Computational Biology Lab, Hefei Institutes of Physical Science, Chinese Academy of Science, Hefei 230031, China, [2]College of Electrical Engineering and Automation, Anhui University, Hefei 230031, China and [3]Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Tremendous amount of omics data being accumulated poses a pressing challenge of meta-analyzing the heterogeneous data for mining new biological knowledge. Most existing methods deal with each gene independently, thus often resulting in high false positive rates in detecting differentially expressed genes (DEG). To our knowledge, no or little effort has been devoted to methods that consider dependence structures underlying transcriptomics data for DEG identification in meta-analysis context.

Results: This article proposes a new meta-analysis method for identification of DEGs based on joint non-negative matrix factorization (*j*NMFMA). We mathematically extend non-negative matrix factorization (NMF) to a joint version (*j*NMF), which is used to simultaneously decompose multiple transcriptomics data matrices into one common submatrix plus multiple individual submatrices. By the *j*NMF, the dependence structures underlying transcriptomics data can be interrogated and utilized, while the high-dimensional transcriptomics data are mapped into a low-dimensional space spanned by metagenes that represent hidden biological signals. *j*NMFMA finally identifies DEGs as genes that are associated with differentially expressed metagenes. The ability of extracting dependence structures makes *j*NMFMA more efficient and robust to identify DEGs in meta-analysis context. Furthermore, *j*NMFMA is also flexible to identify DEGs that are consistent among various types of omics data, e.g. gene expression and DNA methylation. Experimental results on both simulation data and real-world cancer data demonstrate the effectiveness of *j*NMFMA and its superior performance over other popular approaches.

Availability and implementation: R code for *j*NMFMA is available for non-commercial use *via* http://micblab.iim.ac.cn/Download/.

Contact: hqwang@ustc.edu

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

Received on July 10, 2014; revised on September 26, 2014; accepted on October 10, 2014

1 INTRODUCTION

As high throughput biotechnologies have become routine tools in biological and biomedical researches, tremendous amounts of omics data have been generated that provide great opportunity for deciphering molecular mechanisms of cancer or other diseases (Jiao *et al.*, 2014; Natrajan and Wilkerson, 2013; TCGA, 2012; Zhang *et al.*, 2013). Two famous public gene expression databases, GEO (www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (www.ebi.ac.uk/arrayexpress/), have deposited transcriptomic data with more than a million assays from more than 30 000 studies. Another valuable resource, the TCGA project (http://cancergenome.nih.gov/), has released various types of omics data for nearly 10 000 cancer patient samples. Reusing the flood of transcriptomics data with meta-analysis can reduce sample bias and increase statistical power, and thus allow for indepth understanding of pathology of cancer or other diseases at molecular level (Rung and Brazma, 2013). However, the key issue of meta-analysis, i.e. capturing consistent but subtle patterns of gene activity across multiple transcriptomics datasets, still remains challenging both theoretically and practically.

Differentially expressed genes (DEG) across studies could reflect subtle but consistent biological effects and might be false negatives in individual analysis (Xia *et al.*, 2013). To efficiently identify DEGs, meta-analysis methods need to overcome a variety of biological or non-biological variations introduced by distinct protocols and data platforms used in individual studies (Rung and Brazma, 2013). From the aspect of information to be combined, existing meta-analysis methods can be categorized into three classes: *P*-value-based, effect size-based and rank-based, which each deal with non-specific variations at different levels of data. Among them, the *P*-value-based method is statistically most intuitive but allows for standardization of topic-related associations from studies to the common scale of significance (Li and Tseng, 2011). However, the performance of *P* value-based methods heavily depends on the underlying method used for *P* value calculation in individual analysis (Tseng *et al.*, 2012). Compared with *P* value-based methods, the effect size-based methods estimate and directly synthesize effect sizes across studies by using a *t*-statistic-like model. Because the effect size quantity provides a direct measure of differential expression, effect size methods tend to be more efficient in detecting DEGs than the *P* value-based methods (Hong and Breitling, 2008). There are two types of effect size models that can be used for meta-analysis of transcriptomics data: fixed-effect model (FEM) and random effect model (REM), which differ in whether between-study variation is ignorable. Generally, effect size-based methods suffer from unreliable error estimates due to improper distribution assumption

*To whom correspondence should be addressed.

and large noise inherent in microarray data (Hong and Breitling, 2008). Comparatively, the rank-based methods combine the ranks of fold-change, instead of expression values as in effect size models and have an advantage of fewer or no assumptions about data structures (Breitling and Herzyk, 2005). These two features make the rank-based methods more robust and outlier-free in ranking and assessing genes (Xia *et al*., 2013). A representative rank-based method is the RankProd method, proposed by Hong *et al*. (2006), which has been extensively demonstrated to be more reliable and robust than many other methods, especially in low sample number and/or large noise settings (Chang *et al*., 2013; Hong and Breitling, 2008).

Statistical hypothesis about variations of differential expression is another non-trivial factor in meta-analysis. Two complementary hypotheses are generally assumed behind meta-analysis methods: one (HSa) assuming that DEGs are differentially expressed (DE) in all the studies and the other (HSb) assuming that DEGs are DE in one or more studies. The former is more desirable when all the studies are homogeneous, while the latter is more useful and efficient when heterogeneity is expected across datasets. To reconcile the two complementary hypotheses, Song and Tseng (2014) recently presented a new type of hypothesis, HSr, which allows users to specify in what fraction of studies genes are expected to be DE.

It is well-known that gene correlations are ubiquitous in transcriptomics data and can exert substantial impact on the performance of microarray data analysis (Wang *et al*., 2011). Without exception, the dependence structures also considerably influence the meta-analysis of multiple transcriptomics datasets and need to be dealt with carefully (Choi *et al*., 2003). To our knowledge, no or little effort has been devoted to the data dependency issue in meta-analysis of transcriptomics data. Generally, the dependency can be addressed by transforming the high-dimensional data into a low-dimensional space. In computer science, many dimensional reduction or projection methods (Gan *et al*., 2014; Gaujoux and Seoighe, 2012; Zeng *et al*., 2008) can be used for the task. For example, Lê Cao *et al*. applied canonical correlation (CC) analysis to pursuit low-dimensional projections that maximize the associations between two omics variables (Lê Cao *et al*., 2009). Other researchers proposed to use coinertia analyses to explore the relationships between two different types of omics datasets (Fagan *et al*., 2007; Jeffery *et al*., 2007).

Non-negative matrix factorization (NMF) is a recently developed projection method, originally proposed for learning natural parts of faces or semantic features of text (Lee and Seung, 1999). Under the constraint of non-negativity, NMF specifically factorizes data matrix to recover natural parts that are integral to the whole. It has been shown that NMF and its sparse versions perform well in microarray data analysis for pattern discovery and classification (Brunet *et al*., 2004; Kim and Park, 2007; Zhang *et al*., 2012; Zheng *et al*., 2011).

In this study, we propose a joint NMF transcriptomics data meta-analysis method (*j*NMFMA) for DEG identification. Biologically, a certain number of independent biological signals underlie a phenotype and collectively shape gene activities that are associated with the phenotype. Relative to real gene entity, these biological signals can be referred to as metagenes. One metagene might dominate the expression of a group of genes and account for the dependence structures between these genes. In return, we can recover these metagenes from transcriptomics data and identify DEGs by associating them with metagenes that are responsible for the phenotype. Following this, this article develops a joint NMF algorithm (*j*NMF) to simultaneously decompose multiple transcriptomics datasets into a low-dimensional space spanned by metagenes. Then, we employ a regulation probability model to extract DE metagenes and formulate a new metagene-based statistic for measuring differential expression of a gene in meta-analysis context. In summary, *j*NMFMA can interrogate gene correlations in a joint-decomposition way for efficient meta-analysis of transcriptomics data. The use of metagenes as an intermediate step also makes *j*NMFMA flexible in identifying various types of DEGs by confining metagenes. For example, *j*NMFMA can be used to meta-analyze transcriptomics data and DNA methylation data of a same or similar topic for identifying methylation-driven DEGs (as shown in Results section). Compared with other dimensional reduction methods, *j*NMFMA, as an extension of NMF, derives a parts-based representation, favors a sparse matrix decomposition and thus is more suitable to analyze large-scale omics data that are sparse in nature.

To evaluate the performance of *j*NMFMA, we first applied it to simulation data, where *j*NMFMA successfully identified DEGs with high accuracies. Considering that lung cancer is one of the most malignant tumors worldwide, *j*NMFMA was then employed to identify gene signatures for lung adenocarcinoma (LUAD) based on real-world gene expression and methylation data collected from GEO database. In addition, we downloaded from TCGA (http://cancergenome.nih.gov/) another two expression and methylation datasets of LUAD as independent evaluation datasets. The results on both simulation and real-world cancer data show that *j*NMFMA is able to efficiently identify DEGs with biological significance and outperforms other popular approaches.

2 METHODS

Given two or more microarray datasets X_i consisting of same genes, *j*NMFMA first jointly factorizes them into a common submatrix W for all the datasets, named loading coefficient matrix (LCM), plus individual submatrix H_i for data set i, named metagene matrix (MGM), as shown in Figure 1A. All the H_i can be horizontally stacked to form an overall metagene matrix H, each row corresponding to a metagene that represents a hidden biological signal behind the datasets, as shown in Figure 1B. Each row of W reflects the relationships of a gene with the metagenes that hold in all the data. In other worlds, all the data X_i are a result driven by the hidden metagenes and represent as a linear combination by the common loading coefficients in W. We use $H(H_i)$ to identify DE metagenes that are associated with a phenotype of interest, as shown in Figure 1B, and then DEGs as genes that are associated with DE metagenes based on W.

2.1 Mathematics of joint non-negative matrix factorization (*j*NMF)

Assume S datasets X_i, $i = 1, 2, \ldots, S$, with G common genes to be meta-analyzed. Let x_{gj} be the expression level of gene g in sample j, the ith data set be denoted by $X_i = \{x_{gj}, g = 1, 2, \ldots, G, j = 1, 2, \ldots, n_i\}$, where n_i is the number of samples in the dataset. *j*NMF pursues a

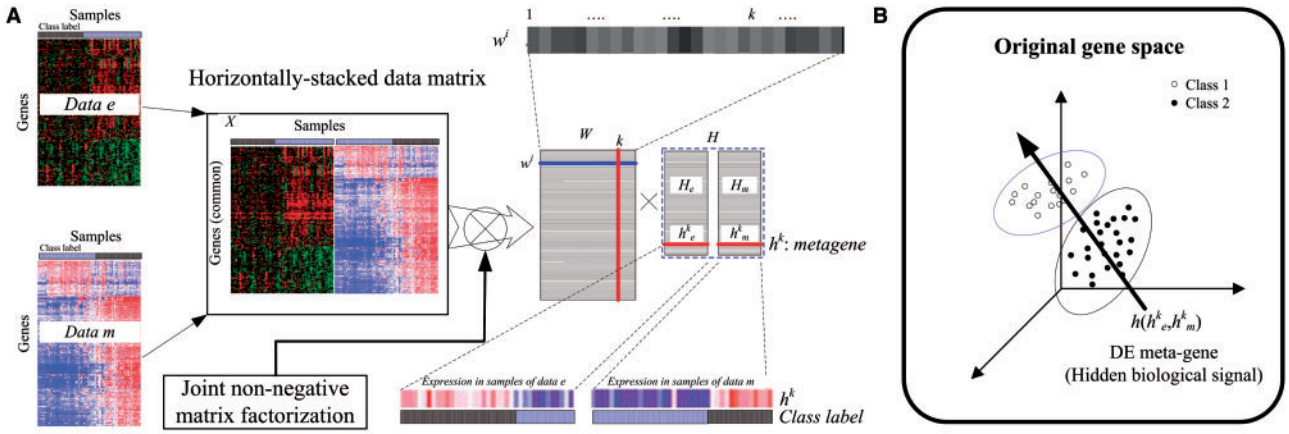

Fig. 1. Framework of *j*NMF (**A**) and resulted DE metagenes (**B**)

simultaneous decomposition of these datasets as follows:

$$X_i = WH_i + E_i,\ i = 1, 2, \ldots, S \tag{1}$$

where W is the sub-matrix LCM of size $G \times k$ (k is an integer constant), H_i is the sub-matrix MGM of size $k \times n_i$ for dataset i, and E_i represents an error matrix for dataset i for accommodating data heterogeneity and noise. Biologically, the decomposition means to recover k hidden biological signals (metagenes) behind the datasets. Mathematically, the decomposition can be formulated as an optimization problem:

$$\begin{aligned} &\min \Gamma(W, H_i) = \sum_{i=1,2,\ldots,S} \|X_i - WH_i\|^2 \\ &s.t. \quad w_{gv} \geq 0, h^i_{vj} \geq 0, g = 1, 2, \ldots, G,\ v = 1, 2, \ldots, k, \\ &j = 1, 2, \ldots, n_i,\ i = 1, 2, \ldots, S \end{aligned} \tag{2}$$

Because the objective function is not convex on both W and H together, no standard algorithm exists for an immediate solution of Equation (2). Practically, it is desirable to find local minima for such optimization problems. In this article, we develop a two-stage multiplicative updating algorithm (see Supplementary Material) for solving Equation (2). A proof for the convergence of the optimization algorithm is given in Supplementary Material.

2.2 Identification of DEGs

We define two events of gene regulation: upregulation (UR) event and downregulation (DR) event in treatment (T) relative to control (C). Following the regulation probability model (RPM) (Wang and Huang, 2006), the difference between probabilities of events UR and DR for each metagene, referred to as absolute regulation probability (ARP), is calculated based on H_i and is used to select DE metagenes. This results in two types of DE metagenes: *uMG*, in which metagenes are upregulated in T relative to C for all H_i, and *dMG*, in which metagenes are downregulated in T relative to C for all H_i. These DE metagenes potentially represent two types of hidden biological signals that are responsible for the distinction between T and C. Considering that each gene can be viewed as a linear combination of the metagenes weighted by row of W (Equation (1)), we formulate the following statistic to measure the differential expression of a gene:

$$d = \sum_{j \in uMG} \alpha_j w_j - \sum_{j \in dMG} \alpha_j w_j,\ \alpha_j = \sum_{i=1}^{S} \frac{n_i}{n} P_i^j \tag{3}$$

where P_i^j is the ARP of DE metagene j estimated by RPM from H_i and n is the total number of samples in the datasets. The two terms at the right side of Equation (3) represent the weighted contributions of *uMG* and *dMG* to the expression of the gene, respectively. Since some metagenes are more discriminative of sample classification than others, we weight their contributions using a quantity of α_j calculated by the ARP and data confidences (n_i/n) of each dataset. Larger difference between the two contributions means a higher likelihood that the gene is differentially expressed in T relative to C. The sign of d indicates the regulation direction.

2.3 Significance estimation

We employ a permutation test to estimate the significance of DEGs. A critical step in the permutation test is to calculate permuted ds that are required to sample from null hypothesis. One possible way for this may be to randomly shuffle sample labels in the original datasets and rerun *j*NMFMA. However, this is computationally infeasible for 1000 or more replicates in permutation test setting. Alternatively, we define the null hypothesis as one that W matrix is non-discriminative of the sample classification, and then randomly shuffle the elements of W $B = 1000$ times for calculating B permutated ds. Let d_m^b be the permuted d for gene m in the bth permutation, the P value for an observed d can be calculated as

$$p = \frac{1}{GB} \sum_{b=1}^{B} \sum_{m=1}^{G} I(|d| < |d_m{}^b|) \tag{4}$$

where $I(\cdot)$ is an indicator function, yielding 1 if the condition is true and 0 otherwise.

2.4 Meta-analysis of gene expression data and DNA methylation data

*j*NMFMA can also meta-analyze transcriptomics data and DNA methylation data with common gene entities for identifying DEGs with negatively correlated expression and methylation patterns (mDEG). For such mDEGs, the differential expression may be dominantly driven by its own altered methylation status. This is useful to detect methylation-driven cancer driver genes as defined in (Bock and Lengauer, 2008; Das and Singal, 2004). To this end, we consider two alternative molecular events, one is simultaneous hypomethylation and upregulation in T relative to C and another simultaneous hypermethylation and downregulation in T relative to C. With the two alternative molecular events, another two types of metagenes can be extracted using RPM: *umMG*, in which metagenes are hypomethylated and upregulated in T relative to C, and *dmMG*, in which metagenes are hypermethylated and downregulated in

T relative to C. Similar to Equation (3), we formulate the following statistic to identify mDEGs:

$$d = \sum_{j \in umMG} \alpha_j w_j - \sum_{j \in dmMG} \alpha_j w_j,\ \alpha_j = \sum_{i=1}^{S} \frac{n_i}{n} P_i^j \quad (5)$$

and create a permutation test similar to that in Section 2.3 for estimating the significance of mDEGs.

2.5 Simulation data generation

We generated Simulation Data I by revising the procedure in (Wang *et al.*, 2011). Assume two studies, I and II, and that samples in each study come from condition A (n) or B (n). Total $G = 900$ simulation genes are divided into nine groups, each group containing 100 genes whose expression follows a same regulation mode, as shown in Supplementary Table S1. Of these groups, G1 is upregulated in A relative to B in both studies and G2 is downregulated in A relative to B in both studies, which both are target DEGs to identify in the experiment. The 'expression' data in each study were synthesized as follows: First, a correlation background matrix X $[G \times 2n]$ was generated as hidden dependence structures by (i) randomly selecting clump size m from $\{1, 2, 3, \ldots, 100\}$ and clump-wise correlation ρ from a uniform distribution $U(0.5, 1)$. For a given (m,ρ) pair, we (ii) generated noise vectors $e_{\cdot j}$ of dimension $m \times 1$ from a multivariate normal distribution $N(0_m,(1-\rho)I_m + \rho 1_m 1'_m)$ for sample j and (iii) set $x_{\cdot j} = \mu + diag(\omega)e_{\cdot j}$, where μ and ω are an $m \times 1$ vector of elements $\mu_g \sim 1000\chi_5^2$ and of elements $\omega_g = e^{\beta 0/2}\mu_g^{\beta 1/2}$ (β_0 and β_1 are two constants) respectively, and $diag$ is a diagonalization function, as the background expression values of the m genes in the clump at samples $j = 1, 2, \ldots, 2n$. In the experiment, we set $\beta_0 = -5$ and $\beta_1 = 2$. Then, based on the correlation background, for the first eight groups, we added (or subtracted) a term $2^{-0.5}\delta_g\omega_g$, $\delta_g \sim U(5,10)$, to (from) the samples in condition B according to the regulation modes (Supplementary Table S1) as final expression levels, and left the background as final expression levels for the ninth group. The true expression ratios for the first eight groups are $1 + 2^{-1/2}e^{\beta_0/2}\delta_g \sim U(1.29, 1.58)$. To examine the effect of sample size, we varied n among {6, 20, 50, 100, 200} to generate six simulation data scenarios.

Simulation data II were designed to mimic hidden biological signals. Assume $G = 600$ genes and that their expression is a linear combination of six hidden biological signals. Of the six hidden biological signals, two are assumed to be true DE signals and four noise signals. Similar to Simulation Data I, we assumed two studies and that samples in each study belong to condition A (n) or B (n). The two true DE signals were simulated by equations $h_{1i} = 5.1(6.1) + |\delta|$ in sample i of condition A (B) and $h_{2i} = 5.3(4.3) + |\delta|$ in sample i of condition A (B), respectively. δ was randomly sampled as noise from a normal distribution $N(0,0.5)$. The four noise signals are divided into two types: discordant DE signals, which are differentially expressed but have discordant differential expression patterns in the two studies, and non-DE signals, which are not differentially expressed in any study at all. The former two noise signals were synthesized as discordant DE signals by equations $h_{3i} = 5.2(6.2) + |\delta|$ in sample i of condition A (B) in one study and $h_{3i} = 6.2(5.2) + |\delta|$ in sample i of condition A (B) in another study, or by equations $h_{4i} = 5.4(4.4) + |\delta|$ in sample i of condition A (B) in one study and h_{4i} 4.4(5.4) + $|\delta|$ in sample i of condition A (B) in another study, while the latter two noise signals as non-DE signals by equations $h_{5i} = 5.7 + |\delta|$ and $h_{6i} = 5.8 + |\delta|$ in all samples of conditions A and B in the two studies, respectively. The six hidden signals comprised hidden matrices $H_s = [h_{ji}^S]$, $j = 1, 2, \ldots, 6$, $i = 1, 2, \ldots, n$, for study s, $s = 1, 2$. Next, we formed a matrix W $[600 \times 6]$ whose rows represent the combination coefficients of the six hidden signals for the $G = 600$ genes. We assumed the first 200 genes to be true DEGs, of which the first half are dominated in expression by the first hidden signal and the second half by the second hidden signal, and the last 400 genes to be non-DEGs, of which the first to third sets of 50 genes are dominated in expression by the third to fifth hidden (noise) signals respectively and the last 250 not by any hidden signal. Accordingly, the matrix W was formed by (i) Sampling all elements w_{ij}, $i = 1, 2, \ldots, 600$, $j = 1, 2, \ldots, 6$ from $U(0,0.5)$ and (ii) Replacing the elements w_{i1}, $i = 1, \ldots, 100$, w_{i2}, $i = 101, \ldots, 200$, w_{i3}, $i = 201, \ldots, 250$, w_{i4}, $i = 251, \ldots, 300$, w_{i5}, $i = 301, \ldots, 350$, with random numbers from $U(0.5,1)$. Finally, we synthesized the simulation data by $X_s = WH_s$, $s = 1, 2$. Similar to Simulation Data I, we varied $n = 6, 20, 50, 100, 200$ to generate five data scenarios for examining the influence of sample size.

3 RESULTS

3.1 Evaluation on simulation data I

We first evaluated our method using Simulation Data I. Consider that the parameter k potentially influences the stability of jNMF and thus the performance of jNMFMA. We varied $k = 10, 50, 100, 300$, and ran jNMFMA with random initialization twice on a dataset and calculated the similarity (Pearson correlation coefficient) of the resulted two vectors of ds as a measure of stability. For unbiased evaluation, 100 random simulation datasets were generated in each of the five n data scenarios. We observed the changes of the similarity with k in the five n scenarios (Supplementary Fig. S1). We found that the similarity steadily increases as k increases, regardless of the n scenario, suggesting that large ks enhance the stability of jNMF and thus the reproducibility of jNMFMA. This should benefit from the fact that larger k results in sparser decomposition and increases the reliability of capturing intrinsic biological signals underlying the datasets. To tradeoff the decomposition stability and computational cost, we set $k = 300$ in subsequent analyses.

It is also observed that true DEGs prefer a large d but non-DE genes not in all the five data scenarios (Supplementary Fig. S2), indicating the ability of jNMFMA to detect DEGs in meta-analysis context. The ability is also demonstrated by the changing trend of false discovery rates (FDR) with rank thresholds of d (Supplementary Fig. S3 (A–E)). As the rank threshold increases, more genes are selected, and thus FDR increases and the curves (1-FDR) go down gradually. It is also revealed (Supplementary Fig. S3 (A–E)) that large ks led to low FDR, regardless of the threshold used, which is in agreement with the changing pattern of reproducibility (Supplementary Fig. S1). The comparison of the receiver operation characteristic (ROC) curves between jNMFMA and two straightforward meta-analysis methods, the intersection (Indiv-Inters) and union (Indivi-Union) sets of DEGs from analyses on individual datasets, confirms the competent power of jNMFMA with highest average area under ROC (AUC) of ~0.95 over the five n data scenarios (Supplementary Fig. S3F).

3.2 Evaluation on simulation data II

Based on the Simulation Data II, we further compared jNMFMA with three popular methods, AW (Li and Tseng, 2011), REM (Choi *et al.*, 2003), RankProd (Hong *et al.*, 2006), as well as the two straightforward methods, Indiv-Uinon and Indiv-Inters, using the following measures: AUC; False positive rate (FPR), FPR = FP/(FP + TN); False negative rate (FNR), FNR = FN/(FN + TP); Positive predictive value (PPV),

PPV = TP/(TP + FP); Accuracy (ACC), ACC = (TP + TN)/(TP + FP + TN + FN); Matthews coefficient constant (MCC),

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

where TP, FP, TN and FN are the numbers of true positives, false positives, true negatives and false negatives, respectively. Note that the three previous methods, AW, REM and RankProd were implemented using the R packages, MetaDE, GeneMeta and RankProd, respectively. Specifically, for AW, the *P* values for individual study were calculated using the modt method (as default) and the fudge parameter was chosen to be the median variability estimator in the genome (as default). Table 1 lists the results of *j*NMFMA and the five previous methods in different *n* scenarios at an *ad hoc P* value cutoff of 0.01. From this table, we can clearly see that the performance of *j*NMFMA is superior to the previous methods with highest ACC, highest PPV, highest MCC, highest AUC and lowest FPR–FNR disparities in almost all five *n* data scenarios. The advantage becomes more conspicuous as *n* increases.

3.3 Application to real-world lung cancer gene expression datasets

In this analysis, three real-world lung cancer microarray datasets collected from GEO (http://www.ncbi.nlm.nih.gov/geo/) were used: Selamat's data (GSE32863) (Selamat *et al*., 2012), Landi's data (GSE10072) (Landi *et al*., 2008) and Su's data (GSE7670) (Su *et al*., 2007). These datasets used different microarray platforms to monitor gene expression: HG-U133A Affymetrix chips for Selamat's data, Illumina Human WG-6 v3.0 Expression BeadChips for Landi's data and Affymetrix Human Genome U133A array for Su's data. Samples in the three datasets were divided into LUAD and normal (NTL). In the Selamat's data, total 117 (58 LUAD and 59 NTL) samples were monitored with the expression levels of ~25 441 genes. In the Landi's data, 107 (58 LUAD and 49 NTL) samples were monitored with the expression levels of ~13 267 genes, and 54 (27 paired LUAD/NTL) samples were monitored with the expression levels of ~13 212 genes in the Su's data. We preprocessed the three datasets as follows: The intensities of multiple probes matching a same Entrez ID were averaged as the expression values of the gene, and non-specific or noise genes were filtered out using a coefficient of variation (CV) filter (Li and Li, 2008) with a CV cutoff of 0.05. Finally, 4728 common genes were left for meta-analysis for identifying LUAD-related DEGs. For extensive evaluation, we selected two datasets at a time [$C(3, 2) = 3$ times] for meta-analysis. For presentation, the three meta-analysis scenarios, Selamat data and Landi data, Landi data and Su data and Su data and Selamat data, are denoted as SeL, SuL and SuSe, respectively. The average outcomes from the three meta-analysis scenarios were used for comparison.

Table 2 lists the results by *j*NMFMA and the five previous methods. The *q*-values were calculated using SLIM (Wang *et al*., 2011) for controlling FDR. From Table 2, we can see that *j*NMFMA called 27% genes significantly differentially expressed at a *p*-value cutoff of 0.01 and 36% genes at a *q*-value cutoff of 0.1. As expected, AW identified the most proportions of DEGs among the methods except for Indiv-Union, irrespective of the *P* value or *q*-value cutoffs used. RankProd identified smaller proportions of DEGs than those by REM, which is in agreement with the observations in (Hong and Breitling, 2008; Tseng *et al*., 2012). Compared with these previous methods, *j*NMFMA has two unique features: removing dependence structures and alleviating data heterogeneity. As described in Methods, *j*NMFMA simultaneously transforms the datasets into a low-dimensional space for removing gene correlations, and by introducing decomposition error terms (Equation (1)), *j*NMFMA can also handle data heterogeneity and study biases under a unified framework. These features increase the reliability and robustness of *j*NMFMA in detecting DEGs.

Because the truly DEGs are unknown, we employed Correspondence At the Top (CAT) plots (Irizarry *et al*., 2005) to evaluate the results of these methods. CAT plots are based on

Table 1. Performance comparison (%, mean ± SD) of *j*NMFMA and previous methods on simulation data II

	FPR	FNR	ACC	PPV	MCC	AUC
$n = 6$						
AW	2.0 ± 4.5	6161 ± 12	66 ± 3.3	49 ± 6.3	2.0 ± 9.8	62 ± 11
REM	0.4 ± 0.2	70 ± 13	74 ± 2.6	**78 ± 4.1**	36 ± 7.8	75 ± 6.6
RankProd	12 ± 1.8	58 ± 8.2	73 ± 2.4	63 ± .4.1	34 ± 7.0	75 ± 2.8
Indiv-Inters	3.3 ± 2.3	91 ± 6.4	68 ± 2.6	60 ± 23	12 ± 12	59 ± 5.8
Indiv-Union	37 ± 7.1	45 ± 9.8	60 ± 7.1	43 ± 7.4	17 ± 14	61 ± 12
*j*NMFMA	**12 ± 8.9**	**44 ± 2.0**	**77 ± 1.8**	74 ± 12	**48 ± 5.2**	**83 ± 2.2**
$n = 20$						
AW	55 ± 8.4	22 ± 6.1	56 ± 7.1	42 ± 5.1	22 ± 12	65 ± 6.8
REM	10 ± 0.6	53 ± 9.8	76 ± 3.1	**70 ± 3.8**	42 ± 8.5	82 ± 3.3
RankProd	**29 ± 1.7**	**33 ± 2**	70 ± 0.7	54 ± 1.0	37 ± 1.1	76 ± 1.1
Indiv-Inters	16 ± 3.6	68 ± 12	67 ± 2.7	49 ± 5.7	18 ± 9.2	62 ± 4.5
Indiv-Union	61 ± 7.8	20 ± 5	53 ± 6.3	40 ± 4	19 ± 11	65 ± 7.9
*j*NMFMA	18 ± 0.5	30 ± 1.6	**78 ± 0.8**	65 ± 1.2	**51 ± 2**	**84 ± 1.4**
$n = 50$						
AW	85 ± 8.9	2.8 ± 3.4	42 ± 4.9	37 ± 1.8	19 ± 4.2	53 ± 12
REM	14 ± 2.6	47 ± 9.1	75 ± 1.6	66 ± 1.7	42 ± 5.0	83 ± 0.7
RankProd	45 ± 1.8	21 ± 2.5	63 ± 1.5	47 ± 1.3	33 ± 3.1	74 ± 2
Indiv-Inters	46 ± 13	33 ± 19	58 ± 3.2	42 ± 2.2	20 ± 7.2	65 ± 6.6
Indiv-Union	88 ± 8.4	2.2 ± 2.6	41 ± 4.9	36 ± 1.8	17 ± 5.1	68 ± 2.6
*j*NMFMA	**18 ± 1.6**	**30 ± 6.1**	**78 ± 1.1**	**66 ± 0.6**	**51 ± 3.6**	**84 ± 1.9**
$n = 100$						
AW	89 ± 5.2	2.3 ± 1.5	40 ± 3.1	35 ± 1.1	15 ± 4	49 ± 4.3
REM	16 ± 2.5	41 ± 4.9	76 ± 2.2	66 ± 4.1	45 ± 5.2	83 ± 1.9
RankProd	51 ± 1.5	14 ± 2.2	61 ± 1.0	46 ± 0.7	34 ± 2.1	72 ± 1.6
Indiv-Inters	39 ± 21	37 ± 23	61 ± 6.6	49 ± 12	25 ± 3.8	68 ± 3
Indiv-Union	80 ± 24	10 ± 20	43 ± 9.1	37 ± 3.1	15 ± 2.6	67 ± 3.9
*j*NMFMA	**19 ± 1.8**	**28 ± 2.5**	**78 ± 1.9**	**66 ± 2.7**	**52 ± 3.9**	**85 ± 1.5**
$n = 200$						
AW	98 ± 0.3	0.4 ± 0.2	35 ± 0.2	34 ± 0.1	6.7 ± 0.4	42 ± 1.3
REM	17 ± 1.8	40 ± 3.4	75 ± 2	64 ± 3.3	44 ± 4.7	82 ± 2.4
RP	58 ± 2.5	10 ± 0.2	58 ± 1.6	44 ± 1.0	33 ± 1.8	71 ± 1.3
Indiv-Inters	74 ± 3.0	11 ± 3	47 ± 1.1	38 ± 0.2	18 ± 1.1	67 ± 1.9
Indiv-Union	98 ± 0.4	0.4 ± 0.2	34 ± 0.3	34 ± 0.1	6.0 ± 1.7	68 ± 4.1
jNMFMA	**18 ± 0.7**	**30 ± 2.4**	**78 ± 0.6**	**66 ± 0. 7**	**51 ± 1.8**	**85 ± 1.1**

Note: Bold indicates the best values.

Table 2. Proportions of DEGs by different methods at varying *P* value/*q*-value cutoffs

Cutoff	1E–04	0.001	0.01	0.1
Indiv-Union	0.61/0.58	0.71/0.68	0.83/0.80	0.94/0.92
Indiv-Inters	0.18/0.14	0.25/0.21	0.36/0.32	0.53/0.49
AW	0.42/0.38	0.62/0.60	0.72/0.70	0.84/0.83
REM	0.41/0.39	0.48/0.46	0.57/0.55	0.72/0.69
RankProd	0.37/0.34	0.43/0.40	0.52/0.48	0.65/0.61
*j*NMFMA	0.08/0.04	0.14/0.07	0.27/0.16	0.49/0.36

the fact that genes identified in multiple independent studies are likely to be truly significant, and high reproducibility among independent studies suggests a high reliability (Hong and Breitling, 2008). In a CAT plot, the proportion of top *m* genes identified in one analysis that are re-discovered in top *m* genes based on another independent dataset is plotted against the number *m*. As described above, in each of the three meta-analysis scenarios (SeL, SuL and SuSe), two of the three LUAD datasets were used for meta-analysis. To draw CAT plots in each meta-analysis scenario, the dataset that was not included in the meta-analysis was used as an independent dataset on which individual analysis for DEG identification was done. The results from the individual analysis were compared with those from the meta-analysis to produce CAT plots. For extensive comparison, we included Indiv-Inters and another two previous methods, i.e. Fisher's and Stoufer's (Tseng *et al.*, 2012), both similar to AW but in different weighted ways. Figure 2(A–C) shows the resulted CAT plots by the methods in the three meta-analysis scenarios, respectively. From these figures, it can be found that *j*NMFMA led to highest overlapping proportions (except for $m = 50$) among these methods for all the three meta-analyses, indicating the superior reproducibility of *j*NMFMA. In contrast, the three *P* value-based methods performed worst among these methods. Another three methods, REM, RankProd and Indiv-Inters, obtained very similar CAT curves as an intermediate between *j*NMFMA and the three *P* value methods. RankProd had slightly higher overlapping rates than REM and Indiv-Inters, which is in agreement with the observations in (Chang *et al.*, 2013; Hong and Breitling, 2008). More interestingly, the reproducibility clearly divides these methods into three groups: Group I with high reproducibility, including only *j*NMFMA, group II with moderate reproducibility, including REM and RankProd, and group III with low reproducibility, consisting of the three *P* value-based methods, AW, Fisher's and Stoufer's, as shown in Figure 2(A–C). The grouping is likely related to the ways that the two factors, data dependency and data heterogeneity, are exploited in the meta-analysis methods: Both in Group I (*j*NMFMA), only data heterogeneity in group II, and none or little information in group III.

As expected, methods with higher proportions of identified DEGs (Table 2) usually have a lower reproducibility, e.g. RankProd, REM, AW and *j*NMFMA, as shown in Figure 2(A–C). However, although it called more DEGs than *j*NMFMA (Table 2), Indiv-Inters did not produce a higher reproducibility. This seemly suggests that *j*NMFMA obtained more reliable DEGs while Indiv-Inters produced more false positives.

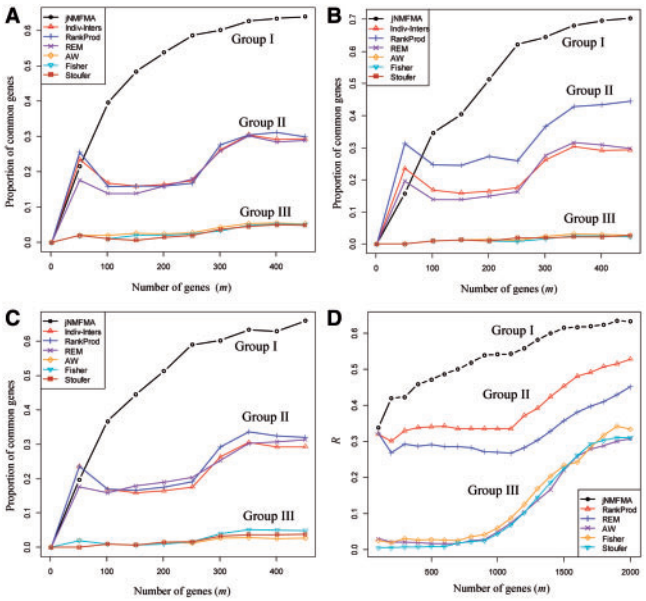

Fig. 2. CAT plots of different methods in meta-analysis scenarios of SeL (**A**), SuL (**B**) and SuSe (**C**) and changing curves of R (**D**) with top *m* genes among the three meta-analysis scenarios

We further examined how consistent the results by each method are between the three meta-analysis scenarios. Similar to the CAT plots, we calculated the ratio of intersection to union genes (*R*) in top *m* genes from each meta-analysis scenario, as shown in Figure 2D. From Figure 2D, it can be found that *j*NMFMA obtained highest values of *R* among the methods for almost all *ms*, confirming the best reproducibility of *j*NMFMA. Figure 2D also witnessed the reproducibility-based stratification of the meta-analysis methods revealed by CAT plots (Fig. 2(A–C)).

3.4 Joint analysis of gene expression and methylation datasets of lung cancer

Biologically, DNA methylation can alter the activity of cells by regulating gene expression, and so it is critical to find cancer-related mDEGs whose abnormal expression is driven by their own aberrant DNA methylation for understanding of cancer pathology. To this end, we additionally downloaded an LUAD DNA methylation dataset (Selamat Meth), originally published in (Selamat *et al.*, 2012), from GEO (GSE32867). The methylation data were measured using the Illumina HumanMethylation27 BeadChip platform. The platform was designed to be limited to the 5' promoter region (Novakovic *et al.*, 2011), and only 1576 of 27 578 probes on it targeted non-promoter regions (>1 kb from TSS). Because promoter and intragenic DNA methylation tend to have opposite influence on gene expression, we only considered ~27 000 promoter probes in the analysis. For the genes measured using multiple probes, their methylation levels were taken as those of the probes with

the maximum average beta-values over all samples. The dataset contains 59 LUAD and matched adjacent normal (NTL) tissues.

We applied *j*NMFMA to meta-analyze the methylation dataset and the above three expression datasets for identification of LUAD-related mDEGs. Among these datasets, there are totally 3598 common genes, and we meta-analyzed the four datasets based on the 3598 genes using *j*NMFMA. By *j*NMF, 14 *dmMG* and 11 *umMG* meta-genes were extracted at a RPM *P* value cutoff of 0.05. Based on the two sets of DE metagenes, we calculated *d* for each gene and their significances by the permutation test. Resulted *P* values were adjusted using SLIM to obtain *q*-values. We summarized the numbers of mDEGs called significant at different cutoffs of *p*-value or *q*-value (Supplementary Fig. S4). With an *ad hoc q*-value cutoff of 0.05, we identified 260 mDEGs, of which 122 are hypomethylated oncogenes-like genes with positive *ds* and 138 hypermethylated suppressors-like genes with negative *ds*, as listed in Supplementary Table S2.

Literature survey shows that many of the 122 hypomethylated oncogenes-like mDEGs were previously reported to be related to lung cancer. Take NIMA-related kinase 2 (NEK2) ($d = 2.4$, P value = 2E–5 and q-value = 0.0047). The gene has been recently identified to be one of several genes most associated with tumor growth in the lungs (Cappello *et al*., 2014). Biologically, NEK2 is a cell cycle-related protein kinase located at a cell's centrosome, which regulates centrosome cohesion and separation through phosphorylation of structural components of the centrosome. Supplementary Fig. S5A barplots the expression and methylation levels of NEK2 (probe id: cg12820481; CpG island location: 1:209915094-209916382) in NTL and LUAD in the four datasets used, showing that NEK2 is significantly (P values $<$ 1E–7 by Kruskal–Wallis test) over-expressed in LUAD in all the three expression datasets (Selamat, Landi and Su) and is significantly (P value = 0.06) hypomethylated in LUAD in the DNA methylation dataset (Selamat Meth). NEK2 expression in non-involved lung tissue was found to be associated with a 3-fold increased risk of mortality from LUAD (Landi *et al*., 2008). Zhong *et al*. (2014) demonstrated that NEK2 can be a proliferation marker in non-small cell lung cancer (NSCLC) prognosis. Recent biological experiments showed the overexpression of NEK2 activates the Akt pathway and increases the levels of β-catenin protein, thus leading to abnormal proliferation of cancer cells (Das *et al*., 2013). By *j*NMFMA, the abnormal over-expression of NEK2 is found to be likely driven by its DNA hypomethylation in LUAD (Supplementary Fig. S5A). Furthermore, we examined the associations of the expression and methylation of NEK2 with LUAD patient's survival time on an independent data set from TCGA (Downloaded on December 1, 2013). The TCGA data contain 155 expression (76) or methylation (78) samples of LUAD with disease-free survival time available. Supplementary Fig. S6A shows the changes of the expression and methylation of NEK2 in LUAD over four survival time intervals, <1, 1–3, 3–5 and >5 years. From this figure, it can be seen that survival time is negatively correlated with the expression of NEK2 but positively correlated with the methylation of NEK2, suggesting the malignance of NEK2 expression and its alleviation by hypomethylation. The result is consistent with the increased expression and decreased methylation in LUAD relative to NTL (Supplementary Fig. S5A).

Among the 138 hypermethylated suppressors-like mDEGs, many of them were also previously reported to be associated with lung cancer. For example, the TCF21 gene ($d = -1.8$, P value = 4.8E–5 and q-value = 0.005) among them is well-known to be frequently lost in human malignancies as tumor suppressor. Supplementary Fig. S5B barplots the expression and methylation (probe id: cg24215443; CpG island location: NA) levels of TCF21 in LUAD and NTL in the four datasets, showing that TCF21 is significantly hypermethylated and under-represented in LUAD. These patterns are consistent with the changing trends of expression and methylation levels over the four survival time intervals in the independent TCGA dataset (Supplementary Fig. S6B). For TCF21, using restriction landmark genomic scanning, Smith *et al*. (2006) experimentally observed the epigenetic inactivation in lung and head and neck cancers. Furthermore, using DNA sequencing technique, Shivapurkar *et al*. (2008) narrowed down a short CpG-rich segment (eight specific CpG sites in the CpG island within exon 1) in the sequence of TCF21, which was observed to be unmethylated in normal lung epithelial cells but to be predominantly methylated in lung cancer cell lines. The short segment accounts for the abnormality of TCF21 in lung cancer. The hypermethylation and under-expression patterns of TCF21 are tumor specific and very frequent in all types of NSCLCs, even in early-stage disease (Richards *et al*., 2010). With these evidences, Richards *et al*. (2010) suggested that TCF21 can be a potential candidate methylation biomarker for NSCLC screening.

4 CONCLUSIONS AND DISCUSSIONS

We have proposed a new computational method (*j*NMFMA) for transcriptomics data meta-analysis for detection of DEGs. The method jointly factorizes multiple transcriptomics data matrices into a low-dimensional metagene space. The joint factorization can interrogate hidden dependence structures and reduce data heterogeneity in omics data. Based on the extracted DE metagenes by RPM, a new statistic *d* was formulated for measuring differential expression of genes in meta-analysis context. Experimental results on simulation data and real-world datasets demonstrated the effectiveness and efficiency of *j*NMFMA in transcriptomics data meta-analysis.

Despite the difference in combined information, most of existing methods treat each gene independently in estimating differential expression. However, plenty of dependence structures inherent in transcriptomics data complicate the meta-analysis and often lead to high FPRs of DEGs. *j*NMFMA uses joint non-negative matrix factorization to address the data dependency. CAT plots and other performance examinations (Table 1; Fig. 2) confirm the effectiveness of *j*NMFMA in dealing with data dependency. *j*NMFMA also explicitly formulates data heterogeneity and noise in terms of decomposition error as in Equation (1) and thus allows for an immediate removal. The CAT plots (Fig. 2) revealed three groups of meta-analysis methods with different levels of reproducibility, which seems to be related to the way to dealing with data dependency and heterogeneity.

Another advantage of *j*NMFMA is the flexibility in detecting various types of DEGs, e.g. DEGs and mDEGs. The flexibility is

especially useful for systems biology where molecular activities at different levels could be positively or negatively correlated in their co-functioning. Experiment on four lung cancer expression and methylation datasets demonstrated the utility of the flexibility in detecting potential cancer driver signatures (mDEGs) that are distinguished from 'passengers' that do not biologically contribute to tumorigenesis (Akavia *et al*., 2010; D'Antonio and Ciccarelli, 2013; Forde *et al*., 2014). The resulted genes, e.g. NEK2 and TCF21, provide potential epigenetic strategy for cancer treatment in the clinic.

*j*NMFMA has a parameter, the number (k) of decomposition dimensions, to be preset in practice. Numeric experiments (Supplementary Fig. S2) showed that large ks (e.g., $k = 300$) generally lead to a high reproducibility of results. More favorably, we would like to recommend to examine the similarity of ds from multiple random runs of *j*NMFMA for a range of k and choose the one with highest similarity.

We noticed that there are several possible directions to improve *j*NMFMA. First, due to the non-negativity requirement of NMF, *j*NMFMA is only applicable to non-negative datasets. Second, the proposed method depends upon a relatively large number of samples. Thirdly, considering that sparse NMF provides an implicit way to control sparse matrix decomposition (Hoyer, 2004), it is needed to explore the utility of sparse versions of *j*NMFMA for better generality and interpretability of *j*NMFMA results. Future works will be done to deal with these challenges.

ACKNOWLEDGEMENTS

HQW thanks Dr. Junwen Wang of the Hong Kong Univerity (HKU), for his constructive suggestions about omics data integration, and Dr. Maria Wong of HKU for her explanation about the lung cancer data, and other Dr. Wang lab members who had numerous discussion with HQW.

Funding: This work was supported by the National Natural Science Foundation of China (61374181, 61300058, 61272339, 91130032, 61103075, 61402010); the Anhui Province Natural Science Foundation (1408085MF133); Research Grants Council, Hong Kong SAR, China (grant number 781511M), and HKU genomics SRT, Innovation Program of Shanghai Municipal Education Commission (13ZZ072); Shanghai Pujiang Program (13PJD032); K. C. Wong education foundation.

Conflict of interest: none declared.

REFERENCES

Akavia,U.D. *et al.* (2010) An integrated approach to uncover drivers of cancer. *Cell*, **143**, 1005–1017.

Bock,C. and Lengauer,T. (2008) Computational epigenetics. *Bioinformatics*, **24**, 1–10.

Breitling,R. and Herzyk,P. (2005) Rank-based methods as a non-parametric alternative of the T-statistic for the analysis of biological microarray data. *J. Bioinform. Comput. Biol.*, **3**, 1171–1189.

Brunet,J.-P. *et al.* (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, **101**, 4164–4169.

Cappello,P. *et al.* (2014) Role of Nek2 on centrosome duplication and aneuploidy in breast cancer cells. *Oncogene*, **33**, 2375–2384.

Chang,L.-C. *et al.* (2013) Meta-analysis methods for combining multiple expression profiles: comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, **14**, 368.

Choi,J. *et al.* (2003) Combining multiple microarray studies and modeling inter-study variation. *Bioinformatics*, **19**, i84–90.

D'Antonio,M. and Ciccarelli,F. (2013) Integrated analysis of recurrent properties of cancer genes to identify novel drivers. *Genome Biol.*, **14**, R52.

Das,P.M. and Singal,R. (2004) DNA methylation and cancer. *J. Clin. Oncol.*, **22**, 4632–4642.

Das,T.K. *et al.* (2013) Centrosomal kinase Nek2 cooperates with oncogenic pathways to promote metastasis. *Oncogenesis*, **2**, e69.

Fagan,A. *et al.* (2007) A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, **7**, 2162–2171.

Forde,P.M. *et al.* (2014) New strategies in lung cancer: epigenetic therapy for non-small-cell lung cancer. *Clin. Cancer Res.*, **20**, 2244–2248.

Gan,B. *et al.* (2014) Sparse representation for tumor classification based on feature extraction using latent low-rank representation. *BioMed Res. Int.*, **2014**, 7.

Gaujoux,R. and Seoighe,C. (2012) Semi-supervised nonnegative matrix factorization for gene expression deconvolution: a case study. *Infection Genet. Evol.*, **12**, 913–921.

Hong,F. and Breitling,R. (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **24**, 374–382.

Hong,F. *et al.* (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, **22**, 2825–2827.

Hoyer,P.O. (2004) Non-negative matrix factorization with sparseness constraints. *J. Mach. Learn. Res.*, **5**, 1457–1469.

Irizarry,R.A. *et al.* (2005) Multiple-laboratory comparison of microarray platforms. *Nat. Methods*, **2**, 345–350.

Jeffery,I.B. *et al.* (2007) Integrating transcription factor binding site information with gene expression datasets. *Bioinformatics*, **23**, 298–305.

Jiao,Y. *et al.* (2014) A systems-level integrative framework for genome-wide DNA methylation and gene expression data identifies differential gene expression modules under epigenetic control. *Bioinformatics*, **30**, 2360–2366.

Kim,H. and Park,H. (2007) Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics*, **23**, 1495–1502.

Landi,M.T. *et al.* (2008) Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. *PLoS One*, **3**, e1651.

Lê Cao,K.-A. *et al.* (2009) integrOmics: an R package to unravel relationships between two omics datasets. *Bioinformatics*, **25**, 2855–2856.

Lee,D.D. and Seung,H.S. (1999) Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

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

Li,S. and Li,D. (2008) *DNA Microarray Technology and Data Analysis in Dancer Research*. World Scientific Publishing, Singapore.

Natrajan,R. and Wilkerson,P. (2013) From integrative genomics to therapeutic targets. *Cancer Res.*, **73**, 3483–3488.

Novakovic,B. *et al.* (2011) Evidence for widespread changes in promoter methylation profile in human placenta in response to increasing gestational age and environmental/stochastic factors. *BMC Genomics*, **12**, 529.

Richards,K.L. *et al.* (2010) Methylation of the candidate biomarker TCF21 is very frequent across a spectrum of early-stage nonsmall cell lung cancers. *Cancer*, **117**, 606–617.

Rung,J. and Brazma,A. (2013) Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.*, **14**, 89–99.

Selamat,S.A. *et al.* (2012) Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression. *Genome Res.*, **22**, 1197–1211.

Shivapurkar,N. *et al.* (2008) Differential methylation of a short CpG-rich sequence within exon 1 of TCF21 gene: a promising cancer biomarker assay. *Cancer Epidemiol. Biomarkers Prev.*, **17**, 995–1000.

Smith,L.T. *et al.* (2006) Epigenetic regulation of the tumor suppressor gene TCF21 on 6q23-q24 in lung and head and neck cancer. *Proc. Natl Acad. Sci. USA*, **103**, 982–987.

Song,C. and Tseng,G.C. (2014) Hypothesis setting and Order statistics for robust genomic meta-analysis. *Ann. Appl. Stat*, **8**, 777–800.

Su,L.-J. *et al.* (2007) Selection of DDX5 as a novel internal control for Q-RT-PCR from microarray data using a block bootstrap re-sampling scheme. *BMC Genomics*, **8**, 140.

TCGA. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.

Tseng,G. *et al.* (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, **40**, 3785–3799.

Wang,H.-Q. and Huang,D.-S. (2006) Regulation probability method for gene selection, *Patt. Recogn. Lett.*, **27**, 116–122.

Wang,H.-Q. *et al.* (2011) SLIM: a sliding linear model for estimating the proportion of true null hypotheses in datasets with dependence structures. *Bioinformatics*, **27**, 225–231.

Xia,J. *et al.* (2013) INMEX: A web-based tool for integrative meta-analysis of expression data. *Nucleic Acids Res.*, **41**, W63–W70.

Zeng,X.-Q. *et al.* (2008) Dimension reduction with redundant gene elimination for tumor classification. *BMC Bioinformatics*, **9**, S8.

Zhang,S. *et al.* (2012) Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Res.*, **40**, 9379–9391.

Zhang,W. *et al.* (2013) Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer. *Cell Reports*, **4**, 542–553.

Zheng,C.-H. *et al.* (2011) Tumor classification based on non-negative matrix factorization using gene expression data. *IEEE Trans. Nanobiosci.*, **10**, 86–93.

Zhong,X. *et al.* (2014) Examining Nek2 as a better proliferation marker in non-small cell lung cancer prognosis. *Tumor Biol.*, **35**, 7155–7162.