BIOINFORMATICS ORIGINAL PAPER

Vol. 31 no. 1 2015, pages 62–68
doi:10.1093/bioinformatics/btu522

Gene expression

Advance Access publication August 26, 2014

Individual-level analysis of differential expression of genes and pathways for personalized medicine

Hongwei Wang[1], Qiang Sun[2], Wenyuan Zhao[1], Lishuang Qi[1], Yunyan Gu[1], Pengfei Li[1], Mengmeng Zhang[1], Yang Li[1], Shu-Lin Liu[2,3,*] and Zheng Guo[1,4,*]

[1]College of Bioinformatics Science and Technology, [2]Genomics Research Center, Harbin Medical University, Harbin 150086, China, [3]Department of Microbiology and Infectious Diseases, University of Calgary, Calgary, AB, T2N 4N1, Canada and [4]Bioinformatics Department, Basic Medical College, Fujian Medical University, Fuzhou 350004, China

Associate Editor: Ziv Bar-Joseph

ABSTRACT

Motivation: The differential expression analysis focusing on intergroup comparison can capture only differentially expressed genes (DE genes) at the population level, which may mask the heterogeneity of differential expression in individuals. Thus, to provide patient-specific information for personalized medicine, it is necessary to conduct differential expression analysis at the individual level.

Results: We proposed a method to detect DE genes in individual disease samples by using the disrupted ordering in individual disease samples. In both simulated data and real paired cancer-normal sample data, this method showed excellent performance. It was found to be insensitive to experimental batch effects and data normalization. The landscape of stable gene pairs in a particular type of normal tissue could be predetermined using previously accumulated data, based on which dysregulated genes and pathways for any disease sample can be readily detected. The usefulness of the *RankComp* method in clinical settings was exemplified by the identification and application of prognostic markers for lung cancer.

Availability and Implementation: *RankComp* is implemented in R script that is freely available from Supplementary Materials.

Contact: guoz@ems.hrbmu.edu.cn or slliu@ucalgary.ca

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

Received on February 20, 2014; revised on July 2, 2014; accepted on July 28, 2014

1 INTRODUCTION

The differential expression analysis methods on intergroup comparison for identifying disease-related genes can be classified into two broad categories: the intensity-based methods such as T-Test (Rice, 1995), SAM (Tusher *et al*., 2001) and Limma (Smyth, 2004), and the rank-based methods such as Wilcoxon signed rank test (Wilcoxon, 1945), Wilcoxon rank sum test (Hollander and Wolfe, 1973) and RP (Breitling *et al*., 2004). Of the rank-based methods, there is a special type of method using the fold-change ordering information, such as RP (Breitling *et al*., 2004), RcoS (Navon *et al*., 2009) and FCROS (Dembele and Kastner, 2014). In general, these methods first entail ranking fold-change values within each pairwise comparison between two types of samples, then calculate a fold-change rank ordering statistic for each gene. The significance of the observed statistic could be determined by using a permutation test (e.g. RP) or probability distribution function (e.g. FCROS). The rank-based methods usually have some advantages in comparison with the intensity-based methods, such as robustness against outlier values and favorable power efficiency in some cases (Lehmann, 1975). However, both the above-mentioned intensity-based and rank-based methods are designed to detect the population-level DE genes and cannot provide patient-specific differential expression information.

Taking the heterogeneous nature of disease into account, several outlier detection methods, including COPA (Tomlins *et al*., 2005), OS (Tibshirani and Hastie, 2007), ORT (Wu, 2007), MOST (Lian, 2008) and others (de Ronde *et al*., 2013; Hu, 2008; Wang *et al*., 2012), are developed to detect genes that are dysregulated in subsets of disease samples. However, they are more sensitive to various technical artifacts, especially experimental batch effects caused by differences in laboratory conditions, reagent lots and personnel (Leek *et al*., 2010). In general, batch effects can introduce serious problems in translating experimental findings to clinical settings. For example, because of this problem, an optimized threshold value of the risk score summarizing the expression levels of signature genes determined from a set of training samples cannot be directly applied to other samples. Usually, in validation studies, data normalization is required for independently extracted samples to make all data having nearly the same scale as the training samples. However, current normalization methods that are used to adjust for systematic technical artifacts are usually not able to remove batch effects (Lazar *et al*., 2012). They may even distort biological signals (Wang *et al*., 2011). More fundamentally, large interindividual variation in gene expression will exacerbate the problem of threshold setting for risk stratification. To tackle this difficult problem, researchers have proposed to make use of the relative ordering information of gene expression within each sample, considering that the relative ordering of gene expression within each sample would be rather robust against batch effects and insensitive to data normalization (Geman *et al*., 2004; Tan *et al*., 2005).

The relative ordering of gene expression is overall stable in a particular type of normal human tissue but widely disturbed in diseased tissue. Taking this into account, a method of detecting

*To whom correspondence should be addressed.

DE genes in individual disease samples is here proposed. Using both simulated data and real paired cancer–normal data, this method was shown to have excellent performance for individual-level analyses of dysregulated genes and pathways. The use of this method in clinical contexts was exemplified by the identification and application of prognostic markers to risk stratification of lung cancer patients according to the dysregulation status of signature in each patient rather than a predefined risk threshold value.

2 MATERIALS AND METHODS

2.1 Data and preprocessing

Multiple microarray datasets of gene expression, generated with the Affymetrix platform (HG-U133 Plus2.0), were downloaded from Gene Expression Omnibus (GEO; Edgar *et al.*, 2002). Notably, normal samples of each type of tissue (lung or breast) were collected from datasets for studying different disorders on this tissue, as described in detail in Table 1. Paired cancer–normal samples taken from the same patients were used to evaluate the performance of *RankComp*. The cancer sample data with survival information were used for survival analysis. More detailed clinical information on survival data can be found in the Supplementary Material.

Table 1. Description of normal sample data, paired cancer–normal sample data and survival data used in this study

Tissue	GEO Acc	Number of Normal[a]	Number of Cancer	Reference
The normal sample data used for identifying stable gene pairs				
Lung	GSE18842	45		(Sanchez-Palencia *et al.*, 2010)
	GSE19188	65		(Hou *et al.*, 2010)
	GSE31210	20		(Lu *et al.*, 2010)
	GSE19804	60		(Okayama *et al.*, 2011)
	GSE37768	20		–
Breast	GSE3744	7		(Alimonti *et al.*, 2010)
	GSE7307	2		–
	GSE7904	7		(Richardson *et al.*, 2006)
	GSE10780	60		(Chen *et al.*, 2009)
	GSE10810	27		(Pedraza *et al.*, 2009)
	GSE20711	2		(Dedeurwaerder *et al.*, 2011)
	GSE21422	5		(Kretschmer *et al.*, 2011)
	GSE22544	4		(Hawthorn *et al.*, 2010)
	GSE26457	113		(Russo *et al.*, 2011)
	GSE29431	12		–
	GSE30010	107		–
	GSE42568	17		(Clarke *et al.*, 2013)
The paired cancer–normal sample data used for evaluating the performance of *RankComp*				
Lung	GSE27262	25	25	(Wei *et al.*, 2012)
Breast	GSE10780	11	11	(Chen *et al.*, 2009)
The data with survival information used for survival analysis				
Lung	GSE31210		204	(Okayama *et al.*, 2011)
	GSE29013		8	(Xie *et al.*, 2011)
	GSE30219		83	(Rousseaux *et al.*, 2013)
	GSE31546		13	–
	GSE37745		78	(Botling *et al.*, 2012)

Note: [a]To determine stable gene pairs on a particular type of normal tissue, only normal samples were collected from datasets for studying different disorders on this tissue. Thus, the information of disease samples in each dataset was not presented.

The use of relative expression obviates the need of between-chip normalization because all direct comparisons between genes occur within individual samples and inter-chip normalization can preserve order (Heinaniemi *et al.*, 2013). For this reason, the raw data (.CEL files) for each dataset was processed using the RMA algorithm for background adjustment without quantile normalization (Irizarry *et al.*, 2003). Then, each probeset ID was mapped to Entrez gene ID with the custom CDF file. If multiple probesets were mapped to the same gene, the expression value for the gene was summarized as the arithmetic mean of the values of multiple probesets (on the log2 scale).

2.2 The MSigDB pathway

The 674 canonical Reactome pathways were downloaded from the C2-CP collection of the Molecular Signatures Database (MSigDB version 4.0, updated May 31, 2013; Subramanian *et al.*, 2005). These pathways covered 6025 unique genes for pathway enrichment analysis.

2.3 Rank comparison

First, each gene expression value is converted to its rank within each sample (the smallest expression value corresponding to the minimum rank, and the largest expression value corresponding to the maximum rank). Then, pairwise comparisons are performed for all genes to identify gene pairs with stable ordering in accumulated normal samples for a particular type of tissue from different data sources. For each gene pair (G_i, G_j), being viewed as an event with only two possible outcomes (G_i>G_j or G_i<G_j), the frequency of samples in normal samples for which the rank of G_i is greater (or less) than that of G_j is estimated as follows:

$$P_{norm}(G_i > G_j) = \frac{1}{n_1}\sum_{t=1}^{n_1} I[G_{it} > G_{jt}]$$

$$P_{norm}(G_i < G_j) = \frac{1}{n_1}\sum_{t=1}^{n_1} I[G_{it} < G_{jt}]$$

where n_1 is the total number of normal samples and I is the indicator function. Stable gene pairs are defined as gene pairs with $P_{norm}(G_i > G_j)$>0.99 or $P_{norm}(G_i < G_j) > 0.99$. Next, reversal gene pairs are defined for each disease sample as gene pairs with reversal ordering in comparison with their stable ordering in normal samples (G_i>G_j→G_i<G_j or G_i<G_j→G_i>G_j). Afterwards, the Fisher's exact test is used to determine whether a given gene (G_i) is differentially expressed in a given disease sample (k) by testing the null hypothesis that the numbers of reversal gene pairs supporting its upregulation and downregulation are equal. For G_i, if its ordering is consistently lower (or higher) than that of G_j in normal samples but the opposite in the disease sample k, then this reversal gene pair is considered to support upregulation (or downregulation) of G_i in that sample. Let G denote the set of stable gene pairs including G_i in normal samples, a and b denote the numbers of gene pairs belonging to G with ordering patterns $\{G_i>G_j\}$ and $\{G_i<G_{j'}\}$ and c and d denote the corresponding numbers of gene pairs belonging to G with ordering patterns $\{G_i>G_j\}$ and $\{G_i<G_{j'}\}$ in the disease sample k. To this end, the proportions of gene pairs with ordering patterns $\{G_i>G_j\}$ and $\{G_i<G_{j'}\}$ are $\mathrm{Obs}_{11}/\mathrm{Obs}_{12}$ = a/b in normal samples and $\mathrm{Obs}_{21}/\mathrm{Obs}_{22}$ = c/d in the disease sample k, respectively. Under the null hypothesis, we will expect that $\mathrm{Obs}_{11}/\mathrm{Obs}_{12} = \mathrm{Obs}_{21}/\mathrm{Obs}_{22}$, which can be tested by the Fisher's exact test, wherein G_i is defined as upregulated if c/d > a/b, downregulated if c/d < a/b and stable expression if c/d = a/b. Finally, a filtering process is adopted to minimize the potential effect of the expression changes of other genes on the upward or downward shift in the rank of a gene detected as an up- or downregulated DE gene: if and only if it is still significant after excluding the downregulated (or upregulated) partner genes involved in the reversal gene pairs supporting its upregulation (or downregulation), it will be retained.

Notably, besides its major application for detecting DE gene at the individual level, the *RankComp* method could also be applied to identify DE genes at the subpopulation level by using the binomial test to find a non-randomly high percentage (*f*) of disease samples sharing certain DE genes. Here, the minimum value of f ($f = k/n_2$) that meets a prespecified significance level for the binomial test can be determined as follows:

$$P = \sum_{k}^{n_2} C_{n_2}^{k} {p_0}^{k} (1 - p_0)^{(n_2 - k)}$$

where k is the number of the observed dysregulated samples for each gene, n_2 is the total number of disease samples and p_0 is the probability of observing a gene being differentially expressed in a disease sample by chance, calculated as the average of the frequencies of DE genes among all the genes on the array for individual disease samples.

2.4 Evaluation of performance

First, a simulation is performed to evaluate the performance of the *RankComp* method. To retain the intrinsic structure of real microarray data, the simulations were conducted based on real microarray dataset rather than beginning by assuming normal distribution (see Section 3 for detailed description of simulation experiments). The simulation experiment enables us to know both the DE genes and non-DE genes and facilitate the calculation of sensitivity, specificity and F-score. For each simulation scenario, the sensitivity, specificity and F-score of the *RankComp* method for detection of DE genes are estimated at different FDR control levels. Here, the sensitivity is defined as the ratio of correctly identified DE genes to all DE genes and the specificity is defined as the ratio of correctly identified non-DE genes to all non-DE genes. The F-score, a harmonic mean of sensitivity and specificity, is calculated as follows:

$$F - score = \frac{2(sensitivity \times specificity)}{(sensitivity + specificity)}$$

Then, the real microarray data of paired cancer–normal samples, as described in Table 1, is used as a benchmark to evaluate the performance of the *RankComp* method. For a paired samples of cancer and normal tissues, a gene is defined as up-regulated if its expression level in the cancer sample is larger than that in the normal sample, and defined as down-regulated if its expression level in the cancer sample is smaller than that in the normal sample, regardless of whether the changes in its expression level are significant or not. Taking the change directions (up- or down-regulation) of genes observed in the paired cancer-normal samples as the golden standard, we then define a DE gene, detected by the *RankComp* method in the same cancer sample, whose change direction is consistent with the golden standard as a consistent DE gene. For each cancer sample, the consistency score, as representative of the precision of DE detection, is calculated as the ratio of the consistent DE genes to all DE genes, and the number of DE genes is used to access the detection power of the method. Notably, the corresponding paired normal tissue samples are not used to determine stable gene pairs in normal tissues.

2.5 Pathway analysis at the individual level

The hypergeometric distribution model was used to determine the significance of biological pathways enriched with up- and downregulated DE genes separately (Hong *et al.*, 2013). The *P*-values were adjusted using the Benjamini and Hochberg procedure, controlling FDR at the 5% level (Hochberg and Benjamini, 1995).

2.6 Survival analysis and pathway signature selection

The overall survival was calculated from the date of surgery until death or last follow-up contact. To avoid the bias of patient follow-up duration, patients with >120 months of follow-up were truncated at 120 months. The survival curves were estimated by the Kaplan–Meier method (Meier, 1958) and were compared using the log-rank test (Gray, 1988). The Cox proportional hazard model was used for univariate and multivariate survival analysis (Cox, 1972).

The forward stepwise algorithm was used to identify an optimal pathway predictor of survival. Beginning with the pathway with the highest concordance index (C-index; Harrell *et al.*, 1996) as the initial signature, candidate pathways were added one at a time to the signature until the next one to be added did not improve prognostic performance. At each step, predictive performance was gauged for all possible additions and evaluated using the c-index to select the optimal addition yielding the largest increase in the c-index value. The C-index is a measure of the probability that, given a pair of randomly selected patients, the model correctly predicts which patient will have a better outcome. This measure quantifies discriminatory ability, ranging from 0.5 (indicating random chance) to 1 (indicating perfect discrimination).

3 RESULTS

3.1 Stable expression ordering of gene pairs in normal human tissues

To determine whether relative expression ordering of gene pairs is stable in normal samples for a particular tissue, normal tissue samples from different datasets were collected for the study of different disorders. For normal lung tissue, a total of 210 samples were collected from multiple datasets of several lung diseases, including lung cancer and chronic obstructive pulmonary disease. They were then divided into two groups: one group of 65 samples was derived from the GSE19188 dataset and another group of 145 samples was derived from four other datasets (Table 1). In each group, pairwise comparisons were performed for all genes to identify stable gene pairs present in at least 99% of normal samples. More than 95% (112 970 616 over 118 725 563) of the stable gene pairs identified in the second group were included in the 135 020 175 stable gene pairs identified in the first group. Remarkably, all of the 112 970 616 overlapping gene pairs had the same ordering patterns across the two groups of samples. This was highly unlikely to happen by chance (binomial test, $P < 1.0\text{e-}16$), indicating that stable gene pairs are highly reproducible in different sets of normal tissue samples. Similar results were observed in normal breast tissue (Table 1 and Supplementary Table S1). Notably, all genes on the array were involved in stable gene pairs for each of the two types of tissue named above, indicating that stable gene pairs widely exist in normal human tissues and that stable gene pairs are an inherent feature of gene expression in normal human tissues. These results provide a basis for individualized differential analysis using gene ranking information.

In order to further show the inherent advantage of stable expression ordering of gene pairs over distribution variations of gene expression, a pair of genes (*GCKR* and *GATA1*) with stable expression ordering obtained from normal lung tissue samples was taken as an example here. As illustrated in Figure 1, even after quantile normalization, the expression distributions of the two genes in normal samples from different datasets are

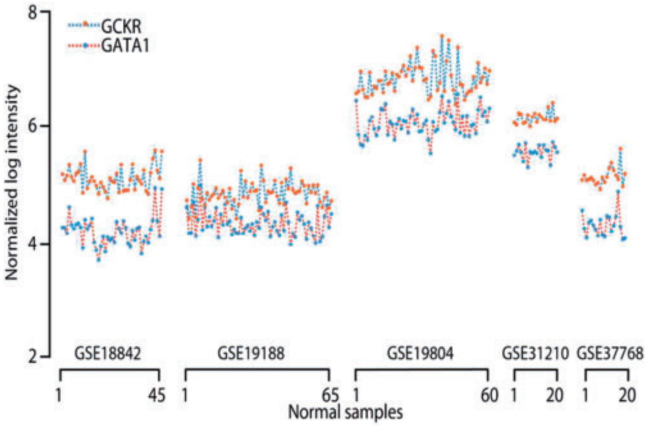

Fig. 1. An example of a pair of genes showing stable relative expression orderings in normal samples of lung tissue. The plot illustrates that the expression distributions of two genes (GCKR and GATA1) in normal samples from different datasets are still different even after quantile normalization. For example, the expression intensities of these two genes in all normal samples of GSE19804 are always higher than those in GSE19188. However, the relative expression ordering of these two genes within each sample is insensitive to the distribution variations from different datasets: the expression intensity of gene GCKR keeps larger than that of gene GATA1 within each sample in any of the datasets

still different so that the expression intensities of these genes in different datasets are incomparable. These distribution variations hamper data integration from different data sources. However, the relative ordering of gene expression within each sample is insensitive to these distribution variations and thus would facilitate multi-source data integration.

3.2 Evaluation of performance

First, the performance of the *RankComp* method was evaluated using a simulation. To retain the intrinsic structure of the data, data were simulated for 60 disease samples and 100 disease samples on the basis of the expression profiles of 15 000 genes for 60 normal lung tissue samples and 100 normal breast tissue samples extracted from the GSE19804 and GSE26457 datasets, respectively. For each normal sample, 1500 upregulated and 1500 downregulated genes were randomly generated and used to produce a disease sample by setting these genes with different magnitudes of differential expression (e.g. $\log_2FC = \pm 0.8, \pm 1.0$ and ± 1.2). Here, the fold-change ($\log_2FC$) was used to quantify the difference in expression levels of each gene between the disease sample and the corresponding normal sample. For the large dataset of 100 disease samples and 100 normal samples, when the magnitude of differential expression ($|\log_2FC|$) was increased from 0.8 to 1.2, the sensitivity increased from 79 to 96% at the cost of a slight decrease in specificity, from 94 to 85%. As shown in Table 2, the method exhibited rather good performance with sensitivity >90%, specificity >85% and F-score >90% when the magnitude of differential expression was not too small ($|\log_2FC| \geq 1.0$). To determine the effect of normal sample size, the performance of the method was studied in the small dataset of 60 disease samples and 60 normal samples. As expected, a slight decline in sensitivity, specificity and F-score was observed for each scenario. Similar results were observed for the scenarios with 5% FDR level.

Table 2. Sensitivity, specificity and F-score for the *RankComp* method in simulated data

Dataset		100 versus 100		60 versus 60	
$\lvert\log_2FC\rvert$	FDR	1%	5%	1%	5%
0.8	Sensitivity	0.7915	0.8329	0.7098	0.7539
	Specificity	0.9439	0.9311	0.9304	0.9120
	F-score	0.8610	0.8793	0.8053	0.8254
1.0	Sensitivity	0.9153	0.9384	0.8520	0.8836
	Specificity	0.9078	0.8742	0.8793	0.8345
	F-score	0.9115	0.9052	0.8654	0.8583
1.2	Sensitivity	0.9633	0.9748	0.9248	0.9452
	Specificity	0.8527	0.7859	0.7970	0.7163
	F-score	0.9046	0.8702	0.8562	0.8150

Then, the real microarray data of paired cancer-normal samples was used as a benchmark to evaluate the performance of the method. Using the stable gene pairs obtained from the 210 normal lung tissue samples, with 5% FDR control, DE genes for each lung cancer sample from an independent paired cancer-normal sample dataset (GSE27262) were detected using the *RankComp* method described in Section 2.3. To ensure the association between the individualized DE genes and cancer, the analysis was restricted to genes that were dysregulated in a non-randomly high percentage of cancer samples. For each cancer sample, average 2722 DE genes showing dysregulated in at least 44% of samples were identified with an average precision of 90.35%, as indicated by the consistency between the detected dysregulation directions of the DE genes and their actual dysregulation directions observed in the paired samples. In particular, when focusing on the analysis of the top 100 ranked DE genes for each cancer sample, the average precision reached 99.80%. Similar evaluation results were observed in an independent paired cancer–normal sample dataset for breast cancer (GSE10780; Supplementary Table S2). These results indicated that the *RankComp* method can reliably capture differential expression signals, especially the top-ranked signals, in cancer patients based on the landscape of stable gene pairs obtained in advance using previously accumulated data.

Notably, the criteria for identifying stable gene pairs in normal samples may affect the results of differential expression analysis. When the criteria were decreased from strict control (99%) to loose control (95%), the average number of DE genes for lung cancer increased by about 44% and the average precision decreased by about 5%. Similar results were observed for breast cancer (Supplementary Table S3). A strict control of 99% would minimize false positives.

In principle, the *RankComp* method is an incremental learning process, during which its performance can be improved along with the accumulation of normal samples for a particular type of tissue by setting an increasingly reliable landscape of stable gene pairs. This advantage here was demonstrated by investigating the impact of normal samples on the performance. From the

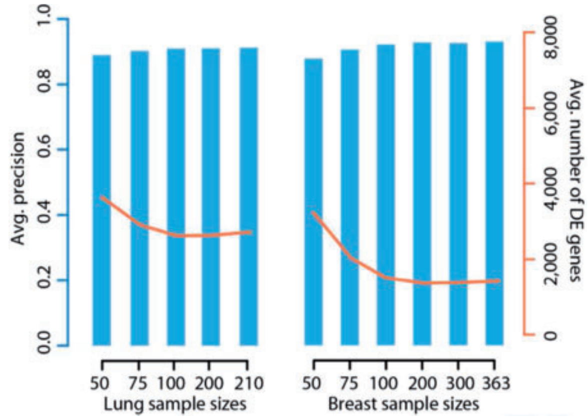

Fig. 2. Impact of normal sample size on the performance of the *RankComp* method. Blue bars represent the average precision, and orange line represents the average number of DE genes per sample

210 normal lung tissue samples, subsets of different sample sizes, ranging from 50 samples per group to the highest sample size, were randomly extracted. As shown in Figure 2, when the normal sample size increased from 50 to 210, the average precision increased from 88.78 to 90.35% at the cost of a decline in average number of DE genes, from 3726 to 2722. However, the rate of decline gradually decreased, indicating a stably improving trend of DE genes detection as the normal sample size increased. The same trend was observed for breast cancer (Fig. 2).

3.3 Use of individual-level DE analysis for outlier detection

As described in the Methods, besides the individual-level analysis of DE genes, the *RankComp* method can be applied to identify DE genes at the subpopulation level, similar to outlier detection methods such as COPA, OS, ORT and MOST. Because it has been shown that COPA and MOST for outlier detection perform relatively better in comparison with several other methods such as OS and ORT (Karrila *et al.*, 2011), we compared only *RankComp* with COPA and MOST in simulation data. The detailed descriptions of simulation experiments and parameter settings were presented in the Supplementary Material, and the detailed results were presented in Supplementary Table S4. Briefly, the F-score comparison results showed that *RankComp* performed similarly with COPA and better than MOST at φ = 5%, 10% and 20%, but obviously better than COPA and slightly worse than MOST at φ = 50% and 80%. These results demonstrated that the *RankComp* method for detecting DE genes in subsets of disease samples can perform better in many scenarios, indicating a potential advantage of applying the *RankComp* method for detecting DE genes in various proportions of disease samples. In addition, we mimicked systematic batch effects in the simulated datasets to evaluate the robustness of the *RankComp* method against systematic batch effects. As shown in Supplementary Table S5, the *RankComp* method had the advantage of insensitivity to systematic batch effects (for details, see Supplementary Material).

Notably, we could also apply the *RankComp* method to identify the population-level DE genes, defined as the combination of subpopulation-level DE genes, considering that the population-level DE genes might not be dysregulated in all samples. In this sense, we compared *RankComp* with the Wilcoxon signed-rank test and Limma. As shown in Supplementary Table S4, the F-score comparison showed that *RankComp* performed better than both the Wilcoxon signed-rank test and Limma for all the scenarios with φ = 5% and still better than Limma for the scenarios with φ = 10%, but worse than both the Wilcoxon signed-rank test and Limma for all the scenarios with φ = 20, 50 and 80%. These results suggested that *RankComp* is generally uncompetitive for detecting the population-level DE genes. Thus, we do not recommend using the *RankComp* method to detect DE genes at the population level.

3.4 Use of individual-level DE analysis in survival analysis

Individualized DE gene analysis provides a basis for individual-level pathway analysis, which could be readily applied in clinical contexts. To exemplify this application, a publicly available microarray dataset of 204 early-stage lung adenocarcinoma samples with survival data was analyzed (GSE31210). Using the landscape of stable gene pairs obtained from the 235 normal lung tissue samples, dysregulated genes were detected for each of 204 cancer samples at the FDR level of 0.05. Focusing on these cancer-associated genes for each cancer sample, biological pathways that were significantly enriched with up- and down-DE genes in this sample were detected at the FDR level of 0.05. The results indicated that the dysregulation of biological pathways in lung cancer patients is heterogeneous (Supplementary Fig. S1). For example, the 'deposition of new CENPA-containing nucleosomes at the centromere pathway' was significantly dysregulated in 50% of cancer patients but not in the others. Dividing the cohort of 204 patients into two groups according to the dysregulation status of this pathway showed that the dysregulation of this pathway was significantly associated with poor overall survival in lung cancer patients (Log-rank test, P = 1.29e-05). Multivariate analysis showed that high-risk and low-risk designation remained statistically significant after adjustment for age, gender and smoking status (Supplementary Fig. S2A and Supplementary Table S6). This result was validated in an independent pooled cohort of 182 lung cancer samples (Supplementary Fig. S2B and Supplementary Table S6).

Many other pathways whose dysregulation status was predictive for survival of lung cancer patients were also found (Supplementary Table S7). To look into this further, a forward stepwise algorithm was used to select an optimal set of pathways with the highest predictive power for prognosis. A two-pathway signature consisting of the 'deposition of new CENPA-containing nucleosomes at the centromere pathway' and 'DNA replication pathway' was generated in the training cohort (Fig. 3A), with a poor prognosis for patients with simultaneous dysregulation of these two pathways confirmed in the validation cohort (Fig. 3B). Multivariate analysis showed that the two-pathway signature remained significantly associated with overall survival after adjustment for age, gender and stage (Supplementary Table S8). It has been found that dysregulation of the former pathway can cause ectopic formation resulting in multicentric chromosomes and consequential genome instability that drives tumor progression (Amato *et al.*, 2009). For instance, the

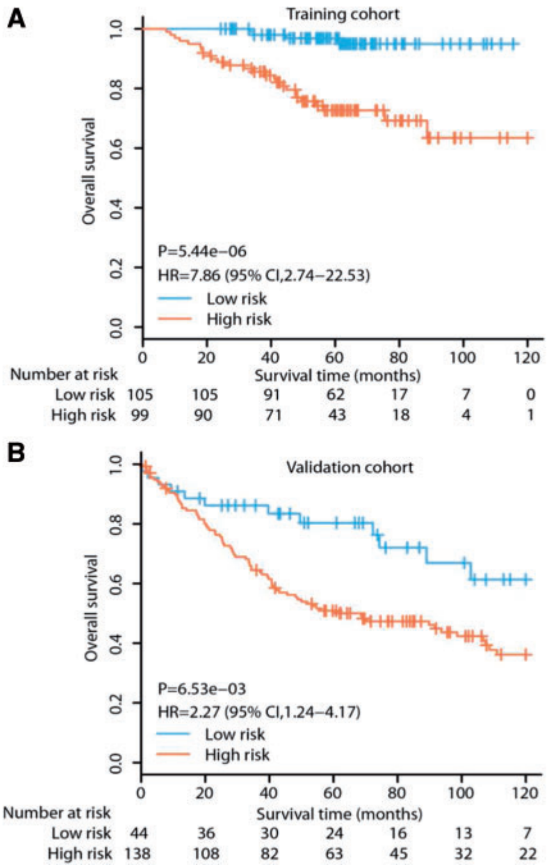

Fig. 3. Kaplan–Meier estimates of overall survival according to the optimal two-pathway signature in the training and validation cohorts. (**A**) Overall survival curves in the training cohort. (**B**) Overall survival curves in the validation cohort

overexpression of CENPA gene involved in this pathway promotes cancer progression by increasing genome instability (Amato *et al*., 2009). Dysregulation of the latter pathway can trigger replicative stress and replication-associated DNA damage, favoring the accumulation of genetic alterations in cancer cells (Allera-Moreau *et al*., 2012; Bartkova *et al*., 2006). For instance, the overexpression of PLK1 gene involved in this pathway promotes cancer progression by increasing resistance to replication stress respectively (Allera-Moreau *et al*., 2012). These studies provided the additional evidence to support our findings from the aspect of biological importance.

Taken together, these analyses demonstrated the capacity of the *RankComp* method to add value in a prognostic setting and provide patient-specific information for personalized medicine.

4 DISCUSSION

The overall stable ordering of gene expression in a normal human tissue may reflect the biological reality that the normal state should be robust against various perturbations (Shiraishi *et al*., 2010). During the transition from the normal state to a disease state, the ordering of gene expression residing in a diseased tissue may be subject to extensive changes, which could be sufficient to reveal patient-specific differential expression information, as demonstrated by the results based on both simulated data and real paired cancer–normal sample data. One unique advantage of the present relative ordering-based method is that it is insensitive to batch effects and data normalization and thus can directly use microarray data from different data sources. In particular, the landscape of stable gene pairs in a particular type of normal tissue can be pre-determined using previously accumulated data of normal samples collected for the study of different disorders on the tissue and could become increasingly stable and reliable along with data accumulation of normal samples. Based on the landscape, dysregulated genes and pathways for an individual disease sample of this tissue can be readily detected.

Analysis of individualized DE genes could have important applications. First, it can be of value in the practice of personalized medicine. For prognostic risk stratification, the present method can directly stratify patients at the individual level based on the dysregulation status of signature in each patient. Optimal risk threshold value that is determined from a set of training samples using a risk-scoring method usually needs to be redetermined in independent samples because of technical and biological variations. Another possible application of the *RankComp* method is in addressing the scarcity of normal tissue samples, which are often rare because of the invasive nature of sample collection. Fortunately, normal samples for a particular type of tissue, generated in different laboratories for the study of different disorders, are often collected in public repositories. Using these normal samples, this method can be used to predetermine the landscape of stable expression ordering of gene pairs for that particular type of tissue. Then, based on this landscape, dysregulated genes and pathways for any disease sample of this tissue can be readily detected. In this way, the present method makes it possible to maximize the reuse of accumulated data for normal tissue samples and facilitate the research of human disease.

Nevertheless, the present method also has several limitations. First, the *RankComp* method may have insufficient power to detect genes whose differential expression causes minor changes in the gene ranking profile. Fortunately, even if a certain number of DE genes go undetected in a given disease sample, shifting the focus of the analysis from individual genes to pathways tends to produce relatively robust results despite insufficient power (Yang *et al*., 2008; Zou *et al*., 2012). Second, the analysis presented here was performed only on microarray data from the same platform because the ordering of gene expression is sensitive to microarray platforms to some degree. One possible way of addressing this limitation is to filter out gene pairs with unstable ordering in datasets produced by different platforms.

Funding: The Natural Science Foundation of China (91029717, 81071646, 81372213, 81201822 and 81030029) and Research Fund for the Doctoral Program of Higher Education of China (20112307110011).

Conflict of interest: none declared.

REFERENCES

Alimonti,A. *et al.* (2010) Subtle variations in Pten dose determine cancer susceptibility. *Nat. Genet.*, **42**, 454–458.

Allera-Moreau,C. *et al.* (2012) DNA replication stress response involving PLK1, CDC6, POLQ, RAD51 and CLASPIN upregulation prognoses the outcome of early/mid-stage non-small cell lung cancer patients. *Oncogenesis*, **1**, e30.

Amato,A. *et al.* (2009) CENPA overexpression promotes genome instability in pRb-depleted human cells. *Mol. Cancer*, **8**, 119.

Bartkova,J. *et al.* (2006) Oncogene-induced senescence is part of the tumorigenesis barrier imposed by DNA damage checkpoints. *Nature*, **444**, 633–637.

Botling,J. *et al.* (2012) Biomarker discovery in non-small cell lung cancer: integrating gene expression profiling, meta-analysis, and tissue microarray validation. *Clin. Cancer Res.*, **19**, 194–204.

Breitling,R. *et al.* (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Chen,D.T. *et al.* (2009) Proliferative genes dominate malignancy-risk gene signature in histologically-normal breast tissue. *Breast Cancer Res. Treat.*, **119**, 335–346.

Clarke,C. *et al.* (2013) Correlating transcriptional networks to breast cancer survival: a large-scale coexpression analysis. *Carcinogenesis*, **34**, 2300–2308.

Cox,D.R. (1972) Regression Models and Life-Tables. *J. R. Stat. Soc. Ser. B Methodol.*, **34**, 187–220.

de Ronde,J.J. *et al.* (2013) Identifying subgroup markers in heterogeneous populations. *Nucleic Acids Res.*, **41**, e200.

Dedeurwaerder,S. *et al.* (2011) DNA methylation profiling reveals a predominant immune component in breast cancers. *EMBO Mol. Med*, **3**, 726–741.

Dembele,D. and Kastner,P. (2014) Fold change rank ordering statistics: a new method for detecting differentially expressed genes. *BMC Bioinformatics*, **15**, 14.

Edgar,R. *et al.* (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Geman,D. *et al.* (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article19.

Gray,R.J. (1988) A class of K-sample tests for comparing the cumulative incidence of a competing risk. *Inst. Math. Stat.*, **16**, 1141–1154.

Harrell,F.E. Jr *et al.* (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat. Med*, **15**, 361–387.

Hawthorn,L. *et al.* (2010) Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast. *BMC Cancer*, **10**, 460.

Heinaniemi,M. *et al.* (2013) Gene-pair expression signatures reveal lineage control. *Nat. Methods*, **10**, 577–583.

Hochberg,Y. and Benjamini,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.*, **57**, 289–300.

Hollander,M. and Wolfe,D.A. (1973) *Nonparametric Statistical Methods*. John Wiley & Sons, New York.

Hong,G. *et al.* (2013) Separate enrichment analysis of pathways for up- and down-regulated genes. *J. R. Soc. Interface*, **11**, 20130950.

Hou,J. *et al.* (2010) Gene expression-based classification of non-small cell lung carcinomas and survival prediction. *PLoS One*, **5**, e10312.

Hu,J. (2008) Cancer outlier detection based on likelihood ratio test. *Bioinformatics*, **24**, 2193–2199.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Karrila,S. *et al.* (2011) A comparison of methods for data-driven cancer outlier discovery, and an application scheme to semisupervised predictive biomarker discovery. *Cancer Inform.*, **10**, 109–120.

Kretschmer,C. *et al.* (2011) Identification of early molecular markers for breast cancer. *Mol. Cancer*, **10**, 15.

Lazar,C. *et al.* (2012) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform.*, **14**, 469–490.

Leek,J.T. *et al.* (2010) Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.*, **11**, 733–739.

Lehmann,E.L. (1975) *Nonparametrics: Statistical Methods Based on Ranks*. Holden-Day, San Francisco.

Lian,H. (2008) MOST: detecting cancer differential gene expression. *Biostatistics*, **9**, 411–418.

Lu,T.P. *et al.* (2010) Identification of a novel biomarker, SEMA5A, for non-small cell lung carcinoma in nonsmoking women. *Cancer Epidemiol. Biomarkers Prev.*, **19**, 2590–2597.

Meier,P. and Kalpan,E.L. (1958) Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53**, 457–481.

Navon,R. *et al.* (2009) Novel rank-based statistical methods reveal microRNAs with differential expression in multiple cancer types. *PLoS One*, **4**, e8003.

Okayama,H. *et al.* (2011) Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res.*, **72**, 100–111.

Pedraza,V. *et al.* (2009) Gene expression signatures in breast cancer distinguish phenotype characteristics, histologic subtypes, and tumor invasiveness. *Cancer*, **116**, 486–496.

Rice,J.A. (1995) *Mathematical Statistics and Data Analysis*, second edition. Belmont, CA: Duxbury Press.

Richardson,A.L. *et al.* (2006) X chromosomal abnormalities in basal-like human breast cancer. *Cancer Cell*, **9**, 121–132.

Rousseaux,S. *et al.* (2013) Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci. Transl. Med.*, **5**, 186ra166.

Russo,J. *et al.* (2011) Pregnancy-induced chromatin remodeling in the breast of postmenopausal women. *Int. J. Cancer*, **131**, 1059–1070.

Sanchez-Palencia,A. *et al.* (2010) Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *Int. J. Cancer*, **129**, 355–364.

Shiraishi,T. *et al.* (2010) Large-scale analysis of network bistability for human cancers. *PLoS Comput. Biol.*, **6**, e1000851.

Smyth,G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article3.

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

Tan,A.C. *et al.* (2005) Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, **21**, 3896–3904.

Tibshirani,R. and Hastie,T. (2007) Outlier sums for differential gene expression analysis. *Biostatistics*, **8**, 2–8.

Tomlins,S.A. *et al.* (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science*, **310**, 644–648.

Tusher,V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.

Wang,D. *et al.* (2011) Extensive increase of microarray signals in cancers calls for novel normalization assumptions. *Comput. Biol. Chem.*, **35**, 126–130.

Wang,Y. *et al.* (2012) Weighted change-point method for detecting differential gene expression in breast cancer microarray data. *PLoS One*, **7**, e29860.

Wei,T.Y. *et al.* (2012) Protein arginine methyltransferase 5 is a potential oncoprotein that upregulates G1 cyclins/cyclin-dependent kinases and the phosphoinositide 3-kinase/AKT signaling cascade. *Cancer Sci.*, **103**, 1640–1650.

Wilcoxon,F. (1945) Individual comparisons by ranking methods. *Biometr. Bull.*, **1**, 80–83.

Wu,B. (2007) Cancer outlier differential gene expression detection. *Biostatistics*, **8**, 566–575.

Xie,Y. *et al.* (2011) Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of non-small-cell lung cancer patients. *Clin. Cancer Res.*, **17**, 5705–5714.

Yang,D. *et al.* (2008) Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, **24**, 265–271.

Zou,J. *et al.* (2012) Revealing weak differential gene expressions and their reproducible functions associated with breast cancer metastasis. *Comput. Biol. Chem.*, **39**, 1–5.