Cai et al. BMC Genomics (2019) 20:134
https://doi.org/10.1186/s12864-019-5502-y

BMC Genomics

RESEARCH ARTICLE Open Access

Identification and characterization of genes with absolute mRNA abundances changes in tumor cells with varied transcriptome sizes

Hao Cai[1†], Xiangyu Li[2†], Jun He[2], Wenbin Zhou[3], Kai Song[3], You Guo[1], Huaping Liu[1], Qingzhou Guan[2], Haidan Yan[2], Xianlong Wang[2*] and Zheng Guo[2,3*]

Abstract

Background: The amount of RNA per cell, namely the transcriptome size, may vary under many biological conditions including tumor. If the transcriptome size of two cells is different, direct comparison of the expression measurements on the same amount of total RNA for two samples can only identify genes with changes in the relative mRNA abundances, i.e., cellular mRNA concentration, rather than genes with changes in the absolute mRNA abundances.

Results: Our recently proposed RankCompV2 algorithm identify differentially expressed genes (DEGs) through comparing the relative expression orderings (REOs) of disease samples with that of normal samples. We reasoned that both the mRNA concentration and the absolute abundances of these DEGs must have changes in disease samples. In simulation experiments, this method showed excellent performance for identifying DEGs between normal and disease samples with different transcriptome sizes. Through analyzing data for ten cancer types, we found that a significantly higher proportion of the DEGs with absolute mRNA abundance changes overlapped or directly interacted with known cancer driver genes and anti-cancer drug targets than that of the DEGs only with mRNA concentration changes alone identified by the traditional methods. The DEGs with increased absolute mRNA abundances were enriched in DNA damage-related pathways, while DEGs with decreased absolute mRNA abundances were enriched in immune and metabolism associated pathways.

Conclusions: Both the mRNA concentration and the absolute abundances of the DEGs identified through REOs comparison change in disease samples in comparison with normal samples. In cancers these genes might play more important upstream roles in carcinogenesis.

Keywords: Differentially expressed gene, Relative expression ordering, Cellular mRNA concentration, Absolute mRNA abundances

Background

It is a common practice to identify differentially expressed genes (DEGs) between two phenotypes through comparing the gene expression profiles measured with the same amount of RNA (or mRNA) extracted from two-phenotype samples, based on the assumption that different types of cells have approximately the same amount of total RNA per cell (transcriptome size) [1]. However, this assumption does not hold under many biological conditions. For example, high expression level of c-Myc can induce global transcriptional amplification of cancer cells [2] and many cancer cells are aneuploid and/or polyploid [3], both of which may cause a change in the transcriptome size [4, 5]. Consequently, if two samples being compared are different in transcriptome sizes but still the same amounts of RNA are used which will result in different numbers of cells between the two measured samples, a direct comparison of the measurement values of the two samples can only identify transcripts with changed cellular concentration which might have no changes in absolute mRNA abundances [6]. Although

* Correspondence: wang.xianlong@139.com; guoz@ems.hrbmu.edu.cn
†Hao Cai and Xiangyu Li contributed equally to this work.
[2]Department of Bioinformatics, Fujian Key Laboratory of Medical Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou 350122, Fujian, China
Full list of author information is available at the end of the article

it could be argued that the concentrations of macromolecules are relevant parameters governing biochemical reactions inside cells, inappropriate interpretation of mRNA concentration changes might lead to incorrect conclusions for a range of biological questions, including the transcriptional characteristics of cancer cells [7].

Experimental methods have been proposed to identify genes with differential absolute mRNA abundances between two cells with different transcriptome sizes [1, 7–9]. However, none has been commonly accepted as a reliable standard approach. For example, if the number of cells used for RNA extraction can be determined, such as in experiments for cell lines or microdissected solid tumor tissues, external control RNA could be spiked into each RNA sample in proportion to the numbers of cells for data normalization [1]. However, large technical variations in spike-in control enrichment and amplification during library preparation challenge the use of spike-in controls [10, 11]. The biological scaling normalization approach proposed by Aanes et al. [8] for adjusting the variation in transcriptome sizes between samples cannot be employed if the number of cells is unknown, which is often the case in experiments for undissected solid tissues. Besides the above-mentioned difficulties, it must be pointed out that all of these strategies are not useful for previously published data, where none of the information about external RNA controls or the cell numbers could be available.

Ranking all genes according to their measured expression levels in a descending (or ascending) order in a sample, the within-sample relative expression ordering (REO) of a gene pair (G_A and G_B) represents whether the expression level of G_A is higher or lower than that of G_B in the sample. Previously, we have found that the within-sample REOs of gene pairs are highly stable in a particular type of normal tissues but widely disturbed in tumor tissues. Based on this finding, an algorithm, RankComp [12], was proposed to detect DEGs through analyzing reversal REOs pattern in an individual disease sample, taking the highly stable REOs in normal samples as the background. Recently, we adjusted this algorithm slightly to fit case-control cohort data, named RankCompV2 [13, 14]. The RankComp and RankCompV2 algorithm detect the genes with expression changes that disrupt the gene correlation structures and change the REOs of the gene pairs from one phenotype to the other. Here, we reasoned that DEGs identified through REOs comparison must change in both mRNA concentration and absolute abundances through theoretical reasoning and simulation experiment. Then, RankCompV2 was applied to ten cancer datasets. Finally, we provided preliminary evidence that the DEGs with changes in both absolute mRNA abundances and concentration are more likely to be closely related with cancer driver genes and drug targets than the DEGs which may change only in mRNA concentration exclusively identified by the popular SAM or edgeR algorithm. RankCompV2 is implemented in C language on Linux and is available on GitHub (https://github.com/pathint/reoa).

Methods

Data and processing

All expression datasets, as summarized in Table 1, were collected from the Gene Expression Omnibus (GEO) database. For microarray and beadchip datasets, quantile normalized values were used in both SAM [15] and RankCompV2. For the RNAseq data, edgeR uses raw counts as input to identify DEGs [16]. When applying the edgeR package, we employed the default TMM (trimmed mean of M-values) [17] to normalize the raw count for sequencing depth and RNA composition. Because TMM does not deal with the transcript length bias of sequencing data, the data normalized with this algorithm are not suitable to rank expression levels of genes with different transcript lengths. Thus, RankCompV2 uses $\log_2$ RPKM data where the transcript length bias has been normalized, as input to identify DEGs. See supplementary method for details (Additional file 1).

RankCompV2 algorithm

The RankCompV2 algorithm was proposed for identifying DEGs with large expression changes lead REOs within diseased samples reversed, comparing with the stable REOs within the normal samples [13].

First, gene pairs with significantly stable REOs are identified in the normal samples. Stable gene pairs, defined as gene pairs with identical REO pattern in significantly more samples for one phenotype than expected by chance, were identified by a binomial test. For a given gene pair (G_i, G_j), let s denote the number of samples in which gene i has a higher (or lower) expression level than gene j in a total of n samples, the significance of the REO pattern is determined by a binomial test as follows,

$$P = 1-\sum_{i=0}^{s-1}\binom{n}{i}(p_0)^i(1-p_0)^{n-i} \qquad (1)$$

where p_0 is the probability of observing a certain REO pattern ($G_i > G_j$ or $G_i < G_j$) in a sample by chance ($p_0 = 0.5$). Current approaches for adjusting the p-values in discrete statistics are still arguable [18–23]. Here, we used the Benjamini and Hochberg method [24] for this purpose, though the method tends to have insufficient power for discrete data [25].

Similarly, gene pairs with significantly stable REOs in the disease samples are identified. Focusing on the overlaps of the two lists of gene pairs, the gene pairs with stable REOs in the normal samples are defined as the normal background stable REOs while the gene

Table 1 Twenty datasets for ten cancers analyzed in this study

Cancer Type	GEO series	Platform		Normal	Cancer	# of Genes
LIHC	GSE57957	GPL10558	Illumina beadchip	39	39	30,500
	GSE45267	GPL570	Affymetrix array	39	48	20,486
KIRC	GSE46699	GPL570	Affymetrix array	42	42	20,486
	GSE53757	GPL570	Affymetrix array	72	72	20,486
HNSC	GSE33205	GPL5175	Affymetrix array	25	44	14,963
	GSE6631	GPL8300	Affymetrix array	22	22	8592
LUSC	GSE19188	GPL570	Affymetrix array	65	27	20,486
	GSE18842	GPL570	Affymetrix array	32	32	20,486
STAD	GSE13911	GPL570	Affymetrix array	31	31	20,486
	GSE29998	GPL6947	Illumina beadchip	49	50	24,384
COAD	GSE23878	GPL570	Affymetrix array	24	35	20,486
	GSE44076	GPL13667	Affymetrix array	98	98	19,040
LUAD	GSE27262	GPL570	Affymetrix array	25	25	20,486
	GSE87340	GPL11154	Illumina HiSeq	27	27	19,471
BRCA	GSE10780	GPL570	Affymetrix array	70	30	20,486
	GSE10810	GPL570	Affymetrix array	27	31	20,486
PAAD	GSE15471	GPL570	Affymetrix array	36	36	20,486
	GSE16515	GPL570	Affymetrix array	16	36	20,486
ESCA	GSE23400	GPL96	Affymetrix array	53	53	12,432
	GSE38129	GPL571	Affymetrix array	30	30	12,432

Abbreviation: *LIHC* Liver hepatocellular carcinoma, *KIRC* Kidney renal clear cell carcinoma, *HNSC* Head and Neck squamous cell carcinoma, *LUSC* Lung squamous cell carcinoma, *STAD* Stomach adenocarcinoma, *COAD* Colon adenocarcinoma, *LUAD* Lung adenocarcinoma, *BRCA* Breast invasive carcinoma, *PAAD* Pancreatic adenocarcinoma, *ESCA* Esophageal carcinoma

pairs with reversely stable REOs in the disease samples compared with the normal samples are defined as the reversal REOs of the disease group. For a given gene *G*, we counted the numbers of gene pairs with $G > G_i$ and gene pairs with $G < G_i$ in normal and disease samples, respectively, and listed the four-cell contingency table through analyzing the *G*-background REOs and the reversal REOs. Then the Fisher's exact test is used to test whether gene *G* expresses differentially in the disease group. After the identification of all candidate DEGs, the gene pairs including candidate DEGs as partner genes are excluded from the construction of the contingency table. And the Fisher's exact test is performed again to minimize the confound effects of the expression changes of the partner genes. This filtering process is conducted iteratively until the number of DEGs keeps stable in two successive iterations. The details of the RankCompV2 algorithm were described in our previous work [13].

RankCompV2 is an empirical algorithm, where the default FDR parameter (FDR < 0.05) for the determination of significantly stable REOs can control false discoveries in simulation experiments, as demonstrated on nine datasets in our previous study [13] and on the twenty datasets in this study (Additional file 1, Additional file 2: Figure S1 and Additional file 3: Table S1).

Reproducibility evaluation of DEGs

We used the POG (Percentage of Overlapping Genes) score [26] and the concordance score to evaluate the reproducibility of DEGs identified from two independent datasets. If two lists of DEGs with length L_1 and L_2, have *n* overlaps, among which *s* have the same dysregulation directions (up- or down-regulation), then the POG score from list 1 (or 2) to list 2 (or 1), denoted as POG_{12} (or POG_{21}), is calculated as s/L_1 (or s/L_2), and the concordance score is calculated as *s/n*. We evaluated whether a concordance score is higher than what expected by chance using the binomial distribution as described above, where p_0 is the probability of a gene having the concordant dysregulation direction in the two lists by chance.

Enrichment analysis

The hypergeometric distribution was used to determine the biological pathways significantly enriched with up- and down-regulated DEGs [27], respectively, based on the Kyoto Encyclopedia of Genes and Genomes database (downloaded on May 16, 2016) [28].

Results

Theoretical basis for identifying DEGs with changes in absolute mRNA abundances

The absolute mRNA abundance of a given gene in a cell is defined as the transcript number of the gene in the cell, and the mRNA concentration is defined as the proportion of mRNA of a given gene in the total mRNA of the cell. Because the mRNA concentration of a gene in a cell is equal to the mRNA concentration of the gene in the corresponding sample including many identical cells, a direct comparison of the measurement values of two samples can identify DEGs with changes in cellular mRNA concentration. However, when the transcriptome size of a tumor cell is different from that of a normal cell, direct comparison of the measurement values of two samples cannot identify genes with changes in absolute mRNA abundances at the single cell level. Here, we reasoned that DEGs identified through REOs comparison must change in both mRNA concentration and absolute abundances at the single cell level.

Let T_k represent the amount of total mRNA in a cell of sample k ($k = 1, 2$) and S represent the same amount of total mRNA extracted from the two samples. Then the number of cells in sample k can be represented as

$$n_k = S/T_k \tag{2}$$

Under the ideal condition that mRNA is extracted from the pure normal epithelial cells (sample 1) and pure tumor epithelial cells (sample 2), the measured expression level of gene i in sample k is,

$$\begin{aligned} M_{ki} &= r_k{}^*N_{ki}{}^*n_k = r_k{}^*N_{ki}{}^*S/T_k = r_k{}^*N_{ki}/T_k{}^*S \\ &= r_k{}^*C_{ki}{}^*S \end{aligned} \tag{3}$$

where r_k is the linear correlation coefficient between the measured expression value and the transcript number of gene i in sample k with n_k cells, N_{ki} represent the transcript number of gene i ($i = 1,...,m$) in a cell of sample k. $C_{ki} = N_{ki}/T_k$ is proportional to the cellular concentration of the transcript of gene i in sample k. Here, we assume that the normalized count values of RNA-sequencing platforms and the fluorescence intensity values of microarray platforms are approximately linearly correlated with the transcript number in a sample within a certain range of gene expression level [29–32].

Since S are the same for two samples and r_k are comparable between two samples after data normalization, direct comparison of the normalized measurements (M_{ki}) between the two samples is equivalent to the comparison of the concentrations (C_{ki}) between the two samples. Consequently, the DEGs identified by using traditional methods, such as SAM for microarray data or edgeR for RNA-sequencing data, are the genes with changes in mRNA concentration (C_{ki}) between the two samples.

Because both T_k and S in eq. (3) are constant for a particular sample k, the within-sample REOs ranked according to the concentration (C_{ki}) are the same with the REOs ranked according to the transcript number (N_{ki}). Therefore, the observed reversal REOs in sample 2 compared with sample 1 must be the reversal REOs of both the mRNA concentration and the transcript number (absolute mRNA abundances). Thus, the DEGs identified by the REO-based RankCompV2 algorithm must have changes in both mRNA concentration and absolute abundances. Consequently, they should be included in the DEGs with concentration changes detected by traditional quantitative-based methods such as SAM or edgeR, given that the later can achieve sufficient power in the data under analysis.

Evaluation of performance

We assumed that the number of reads mapping to a transcript sequence is roughly proportional to the RNA amount of the transcript and the sum of the read counts of all transcripts (total mapped reads) was used to represent the total RNA amount of a sample. Thus, we performed a simulation experiment based on the RNA-sequencing data of the GSE87340 dataset to evaluate the performance of RankCompV2 in data with global transcriptome size changes. For the 19,471 genes measured for the 27 normal samples, after removing genes with a count of 0 in more than 75% of the samples, we simulated disease samples by randomly selecting 6000 genes to produce 3000, 4000 and 5000 up-regulated DEGs and correspondingly 3000, 2000 and 1000 down-regulated DEGs, respectively. For each simulation experiment, the up-regulated genes were equally divided into four groups and the fold change (FC) levels of the genes in the four groups were assigned as 2, 3, 4 and 5, respectively. Similarly, the down-regulated genes were equally divided into four groups and the FC levels of the genes in the four groups were assigned as 1/2, 1/3, 1/4 and 1/5, respectively.

When simulating more up-regulated DEGs than down-regulated DEGs, the simulated disease samples tend to have more total transcript counts than the normal samples, which mean that the transcriptome size of a disease cell is larger than that of a normal cell. In order to simulate the same amount of total RNA extracted from two samples, the read counts of each transcript in simulated disease samples were multiplied by a transcriptome size factor to make the total counts of simulated disease samples keep the same with that of normal samples. The factor is the fold change of the transcriptome size (the amount of total RNA per cell) between the tumor cell and the normal cell.

Read counts were used in edgeR and the RPKM values calculated from the counts were used in RankCompV2 to identify DEGs. Each simulation experiment was repeated 100 times. The sensitivity (the ratio of correctly identified DEGs to all true DEGs), the specificity (the ratio of correctly identified non-DEGs to all true non-DEGs), the F-score (a harmonic mean of the sensitivity and the specificity) and the false discovery rate (FDR, the ratio of true non-DEGs to all identified DEGs) were employed to evaluate the performance of different algorithms.

As shown in Table 2, when the number of up-regulated DEGs increased from 3000 to 5000, the ratio of the total counts of the normal sample to that of the simulated disease samples decreased from 0.7570 to 0.5972. The TMM normalization [17] can estimate a scale factor to adjust the different total RNA output between samples. When the numbers of up- and down-regulated DEGs were equal, edgeR which can incorporate these factors into DEGs analysis exhibited higher average sensitivity and F-score than RankCompV2. However, the FDR of edgeR was up to 54.89% when the up-regulated DEGs were more than the down-regulated DEGs. Instead, RankCompV2 exhibited rather good performance with sensitivity > 95%, specificity > 99% and FDR < 0.15%. For each simulation experiment the DEGs identified by RankCompV2 were completely included in the DEGs identified by edgeR. The simulation results confirmed the above mathematical reasoning and demonstrated RankCompV2 can identify genes with expression change in both mRNA concentration and absolute abundances.

To assess the strength of the methodology, we performed another simulation experiment based on the RNA-sequencing data of the 27 normal samples in the GSE87340 dataset. We randomly generated 4000 up-regulated and 2000 down-regulated genes by changing their measured values in each samples with FC levels of 1.5 to 3.5 to produce 27 disease samples. In each simulation experiment, all the selected genes were set at the same FC level. Each simulation experiment was repeated 100 times. The simulated disease samples were also multiplied by a transcriptome size factor so as to simulate the same amount of total RNA extracted from two samples. The average transcriptome size factors for each 100 simulated experiments were listed in Table 3. Then edge R and RankCompV2 were performed to identify DEGs. The sensitivity, specificity, F-score and FDR were calculated.

As shown in Table 3, the average sensitivity of RankCompV2 was only 56.10% for DEGs with FC of 1.5 and up to 98.84% with FC of 3.5, which suggested that RankCompV2 performed well for DEGs with large expression changes. In general, RankCompV2 showed a very high specificity and a very low FDR when the FC level increased from 1.5 to 3.5. Notably, the FDR of edgeR rises as the FC level increases. The underlying reason is as follows. When up-regulated DEGs with a larger FC level was introduced in the simulation, it leads to a bigger global transcriptome size of a tumor cell than that of a normal cell, as shown by the decreased transcriptome size factor (Table 3), the ratio between the normal transcriptome size and the simulated tumor cell size. The edgeR algorithm identifies DEGs through comparing the read counts of a gene between the two samples. Given the same amount of total input RNA for the two samples, many genes without differences in mRNA absolute abundances would have lower read counts in the tumor sample than in the normal sample, thus be identified as down-regulated genes, which leads to a higher FDR of edgeR if taking DEGs with absolute mRNA abundance changes as the reference.

We also performed a simulation experiment on the genome size changes leading to the global transcriptome size variations. The simulation experiment also demonstrated that RankCompV2 could identify genes with expression changes in absolute abundances and performed well for DEGs with large expression changes (Additional file 4).

Reproducible DEGs with changes in absolute mRNA abundances in ten cancers

We collected two datasets of gene expression profiles for each of ten cancer types (Table 1). For each dataset, we compared the DEGs between the normal and cancer samples identified by RankCompV2 with the DEGs identified by SAM for microarray data or by edgeR for RNA-sequencing data. In GSE57957 measured by microarray for liver hepatocellular carcinoma (LIHC), SAM identified 11,497 DEGs with false discovery rate (FDR) < 0.05, which included 3603 of the 3715 RankCompV2

Table 2 Simulation evaluation with different transcriptome sizes for edgeR and RankCompV2

Up/Down	Factor	edgeR				RankCompV2			
		Sen	Spe	F-score	FDR	Sen	Spe	F-score	FDR
3000/3000	0.7570	99.31%	100.00%	99.65%	0.00%	95.24%	100.00%	98.53%	0.00%
4000/2000	0.6682	99.55%	89.40%	94.18%	18.95%	95.47%	99.98%	98.58%	0.05%
5000/1000	0.5972	99.41%	45.62%	62.51%	54.89%	95.96%	99.94%	98.71%	0.13%

Note: Up (or Down) indexes the number of simulated up-regulated (or down-regulated) DEGs; Factor is defined as the fold change of the transcriptome sizes between the simulated tumor cell and the normal cell; Sen represents sensitivity defined as the ratio of correctly identified DEGs to all true DEGs; Spe represents specificity defined as the ratio of correctly identified non-DEGs to all true non-DEGs); F-score is a harmonic mean of the sensitivity and the specificity; FDR is the abbreviation of false discovery rate defined as the ratio of true non-DEGs to all identified DEGs

Table 3 Simulation evaluation with different FC levels for edgeR and RankCompV2

FC level	Factor	edgeR				RankCompV2			
		Sen	Spe	F-score	FDR	Sen	Spe	F-score	FDR
1.5	0.9360	85.83%	100.00%	92.37%	0.00%	56.10%	100.00%	71.88%	0.00%
2	0.8647	98.76%	99.76%	99.26%	0.51%	82.50%	100.00%	90.41%	0.00%
2.5	0.7996	99.52%	97.82%	98.63%	4.23%	93.50%	100.00%	96.64%	0.00%
3	0.7448	99.78%	94.85%	97.21%	9.83%	97.79%	100.00%	98.88%	0.01%
3.5	0.6965	99.88%	92.61%	96.08%	13.87%	98.84%	99.98%	99.40%	0.05%

Note: FC is the abbreviation of Fold change; Factor is defined as the fold change of the transcriptome sizes between the simulated tumor cell and the normal cell; Sen represents sensitivity defined as the ratio of correctly identified DEGs to all true DEGs; Spe represents specificity defined as the ratio of correctly identified non-DEGs to all true non-DEGs); F-score is a harmonic mean of the sensitivity and the specificity; FDR is the abbreviation of false discovery rate defined as the ratio of true non-DEGs to all identified DEGs

DEGs. In GSE45267 for LIHC, the 14,192 DEGs selected by SAM included 4712 of the 4728 DEGs detected by RankCompV2. The concordance scores of the dysregulation directions of the overlaps between DEGs detected by RankCompV2 and DEGs detected by SAM in the two datasets were all 100%. Similar results were observed in the 18 datasets for other nine cancer types (Fig. 1a and Additional file 5: Table S2). As expected, POG_{21} is lower than POG_{12} because the DEGs identified by RankCompV2 should be included in the DEGs identified by

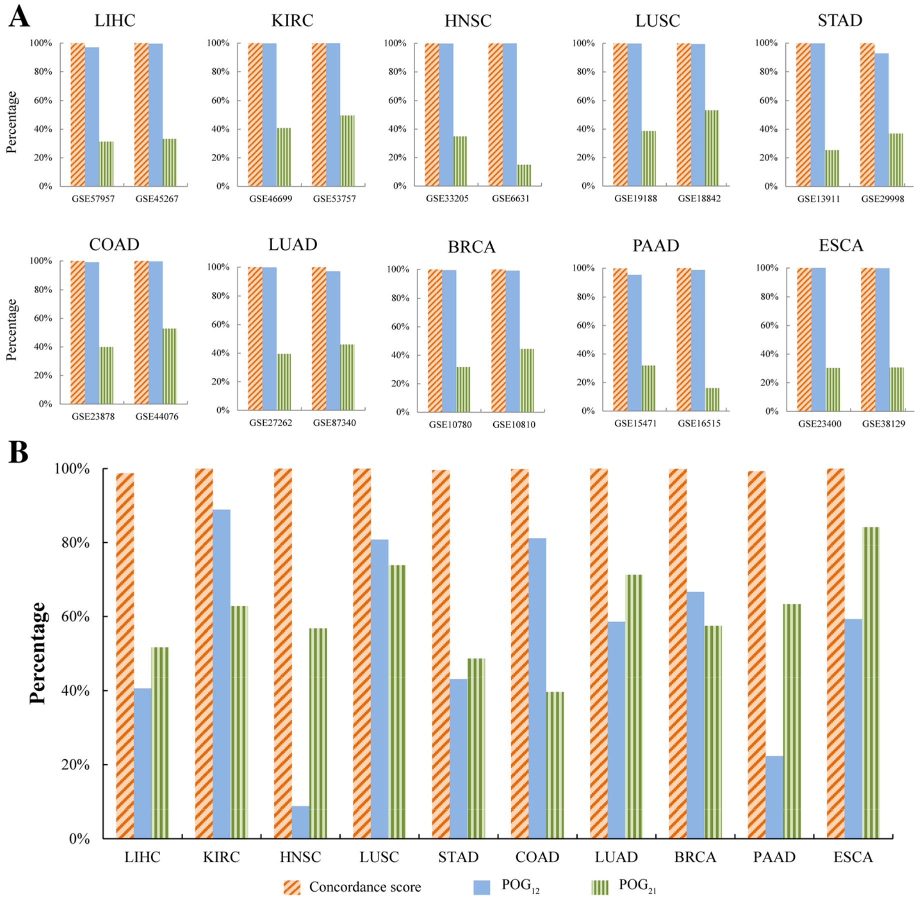

Fig. 1 Reproducibility evaluation of DEGs identified by RankCompV2. **a** Comparison of DEGs identified by RankCompV2 with DEGs identified by SAM or edgeR in the same dataset. POG_{12} (or POG_{21}) denotes the ratio of consistently detected DEGs by the two methods to DEGs identified by RankCompV2 (or by SAM or edgeR) for the same dataset. **b** Comparison of RankCompV2 DEGs identified from different datasets. POG_{12} (or POG_{21}) denotes the ratio of consistently detected DEGs from two datasets to DEGs detected from the first dataset (or the second dataset) for each cancer type. The concordance score denotes the percentage of consistently detected DEGs that display the same dysregulation directions to the overlapped DEGs between two DEG lists

Table 4 Numbers of identified absolute and relative DEGs in ten cancers

Cancer type	Absolute DEGs		Relative DEGs	
	Up	Down	Up	Down
LIHC	3525	2947	5424	6625
KIRC	5295	4420	2882	6279
HNSC	841	1000	2142	1790
LUSC	3523	4795	7067	2363
STAD	2946	2003	4611	6042
COAD	3922	4692	4799	2877
LUAD	2905	3868	6186	2121
BRCA	2185	3251	6458	1771
PAAD	4261	2294	3658	8408
ESCA	1901	1334	2459	4438

Abbreviation: *LIHC* Liver hepatocellular carcinoma, *KIRC* Kidney renal clear cell carcinoma, *HNSC* Head and Neck squamous cell carcinoma, *LUSC* Lung squamous cell carcinoma, *STAD* Stomach adenocarcinoma, *COAD* Colon adenocarcinoma, *LUAD* Lung adenocarcinoma, *BRCA* Breast invasive carcinoma, *PAAD* Pancreatic adenocarcinoma, *ESCA* Esophageal carcinoma

SAM or edgeR. The results confirmed the above mathematical reasoning, which also provided circumstantial evidence of the high accuracy of the RankCompV2 method.

The DEGs identified by RankCompV2 were highly reproducible in independent datasets. For LIHC, RankCompV2 identified 3715 DEGs from the GSE57957 dataset and 51.71% of them were included in the 4728 DEGs detected from the GSE45267 dataset. The concordance score of the overlapped 1946 DEGs was 98.72% which was unlikely to be observed by chance (binomial test, $p < 1.0E\text{-}16$). The highly reproducibility of RankCompV2 DEGs identified from independent datasets were also observed in the other nine cancer types (Fig. 1b).

Enrichment of cancer driver genes and drug targets in DEGs with absolute abundance changes

For each cancer type, the two lists of DEGs identified from the two datasets by SAM for the microarray data or by edgeR for the RNA-sequencing data were combined, excluding those with contradictory dysregulation directions. Similar combination processes were performed on DEGs selected by RankCompV2. The DEGs identified by RankCompV2, with expression change in both mRNA concentration and absolute abundances, were termed as absolute DEGs, while the DEGs solely detected by SAM or edgeR were termed as relative DEGs with changes in concentration only. The numbers of the absolute DEGs and relative DEGs for the ten cancer types were listed in Table 4. Then, we explored the biological significance of the absolute DEGs.

With the 616 cancer driver genes downloaded from the Catalogue Of Somatic Mutations (COSMIC, version81, updated 9th May 2017) database [33], we found 25.79% of the 6472 absolute DEGs of LIHC overlapped or directly interacted with known cancer driver genes based on the protein–protein interaction data downloaded from the STRING v10 database [34], which was significantly higher than the corresponding ratio (18.35%) for the 12,049 relative DEGs (Fisher's exact test, $p < 1.0E\text{-}16$). Similar results were observed for the remained nine cancer types (Fig. 2a). Based on the cancer driver genes downloaded from the DriverDBv2 database [35], where the driver genes for each cancer type were identified by at least two algorithms from the mutation data in the TCGA database, we also observed significantly higher ratios of cancer driver genes and interaction genes in the absolute DEGs than in the relative DEGs for all the ten cancer types (Fig. 2b). The results indicate that DEGs changing in absolute abundances are more likely to be related with upstream events of carcinogenesis than DEGs changing in concentration only.

With 116 targets of 148 anti-cancer drugs documented in CancerDR [36], we found that 16.56% of the 6472 absolute DEGs for LIHC overlapped or directly interacted with known anti-cancer drug targets in the STRING network, which was significantly higher than the corresponding ratio 10.91% for the 12,049 relative DEGs (Fisher's exact test, $p < 1.0E\text{-}16$). Similar results were observed for the remained nine cancer types (Fig. 2c).

Functional analysis of DEGs with absolute abundance changes

Pathway enrichment analysis was performed for the absolute and relative DEGs, respectively. As shown in Additional file 6: Table S3, for each of the ten cancers, the pathways enriched with the absolute DEGs were much more than and quite different from the pathways enriched with the relative DEGs. As summarized in Fig. 3, the pathways enriched with absolute DEGs for at least five cancer types were very different from the pathways enriched with relative DEGs for at least five cancer types. The up-regulated absolute DEGs were enriched in many pathways related with response to DNA damages, including "mismatch repair", "base excision repair", "nucleotide excision repair", "homologous recombination" and "Fanconi anemia pathway" [37]. These genes were also enriched in "p53 signaling", "cell cycle", "DNA replication", "pyrimidine metabolism" and "purine metabolism". The pathways enriched by relative DEGs included "proteasome" and "protein processing in endoplasmic reticulum" besides "RNA transport" and "spliceosome" which were also enriched by up-regulated absolute DEGs.

The down-regulated absolute DEGs were commonly enriched in many metabolism pathways, including amino acid, carbohydrate and lipid metabolism, and in immune associated pathways, including "chemokine signaling", "complement and coagulation cascades" and

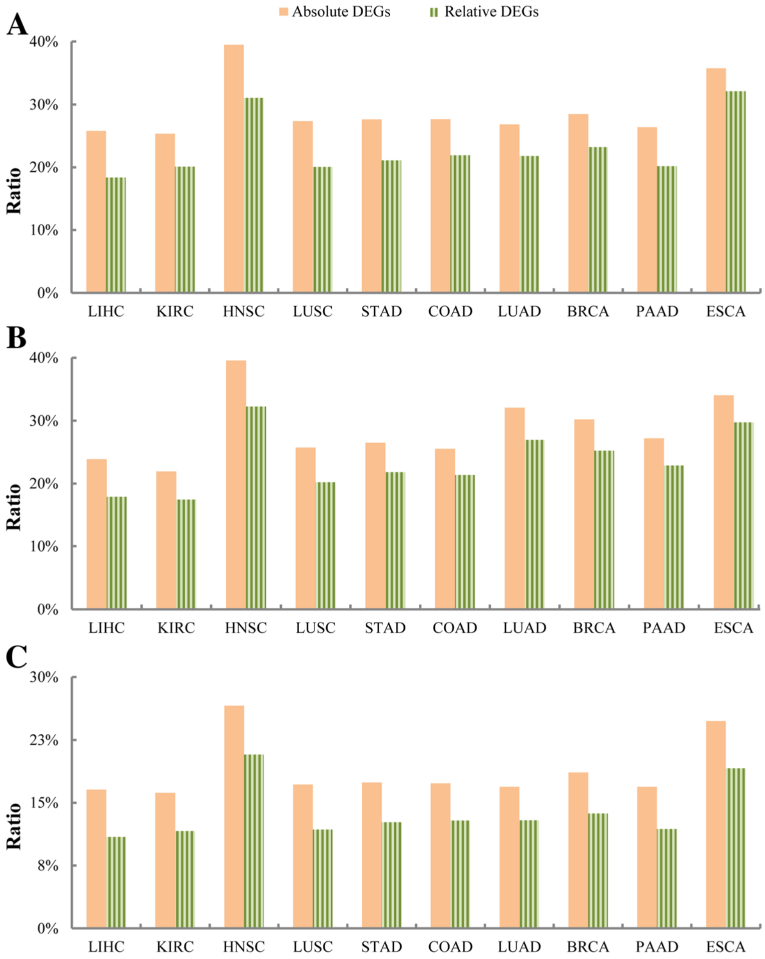

Fig. 2 Association between the identified absolute and relative DEGs with known cancer driver genes or drug targets. **a** 616 cancer driver genes from COSMIC; **b** cancer driver genes from DriverDBv2; **c** anti-cancer drug targets. Statistically significant differences ($p < 0.05$) were found in all the ratios between absolute DEGs and relative DEGs

"leukocyte transendothelial migration". In contrast, the down-regulated relative DEGs were enriched in signaling pathways including "calcium signaling", "cAMP signaling", "neuroactive ligand-receptor interaction" and two sensory system pathways of "olfactory transduction" and "taste transduction". The difference of pathways enriched by two type DEGs is an issue worth further analysis.

Discussion

We demonstrated that the REO-based RankCompV2 algorithm can identify DEGs with changes in both mRNA concentration and absolute abundances (the absolute DEGs), while the quantitative-based algorithms can identify only those with changes in mRNA concentration (the relative DEGs). Through studies for all the ten cancers, the absolute DEGs have a higher probability associated with both known cancer driver genes and drug targets than the relative DEGs. Thus, we speculate that the absolute DEGs might play a more important upstream role in carcinogenesis. In addition pathway enrichment analysis showed that up-regulated absolute DEGs are significantly enriched in DNA damage-related pathways and down-regulated absolute DEGs are significantly enriched in immune and metabolism associated pathways. The genome instability including DNA damages and tumor-promoting inflammation driven by immune cells are two enabling characteristics of tumor and instrumental for tumorigenesis and progression [38], and energy metabolism dysregulation is a fundamental hallmark to fuel cancer cell growth and division [38].

The reasoning for that the REOs-based RankCompV2 algorithm can identify DEGs with changes in absolute mRNA abundances is based on the ideal condition that the gene expression measurements are well correlated with the transcript numbers. For tumor tissues, the ideal condition could be violated due to variations of the tumor epithelial cell proportions in tissues sampled from different sites of a tumor and partial RNA degradation

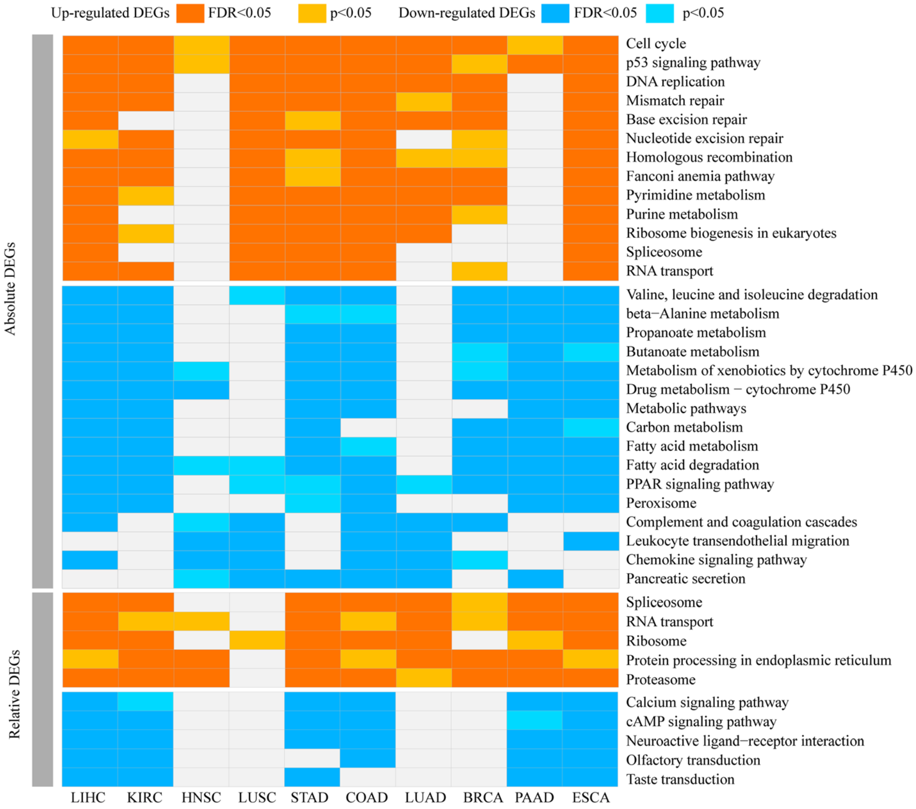

Fig. 3 The pathways significantly enriched with absolute and relative DEGs, respectively, in each cancer type and commonly enriched in at least five cancer types

during sample preparation. However, the qualitative nature of REOs lends them the advantage being robust against the above-mentioned confounding factors [39–41]. As demonstrated in our recent study, the stromal cells in tumor tissues have similar REOs with those of epithelial cells in tumor tissues [39, 41]. More than 96% REOs in the tumor tissues with above 70% of proportion of epithelial cells are consistent with the REOs in tumor epithelial cells, and about 90% REOs in tumor epithelial cells are kept in tumor tissues even when the proportion of epithelial cells decreases to 30% [41]. Therefore, the REOs-based RankCompV2 algorithm would be largely applicable to real tumor data of macro-dissected cancer tissues.

In order to simplify the reasoning, we assumed that genes have similar measurement biases and the correlation coefficients (r) between the measured expression values and the transcript numbers are similar for different genes in eq. (3). However, different measurement biases exist among different samples and among different genes within samples in actual measurement [42–44]. Thus, it seems unreasonable to compare the measured expression levels of different genes within a sample. However, it has been shown that the within-sample REOs are robust against systematic biases of measurements, experimental batch effects and data normalization [45]. Moreover, the gene pairs with large rank difference tend to retain the same REO patterns in samples measured with different platforms [46]. The high robustness of within-sample REOs indicate that the influence of measurement biases on the REOs is small.

Here, we further analyzed the influence of measurement biases on the within-sample REO-based algorithm. Considering two genes, i and j which have the expression levels E_i and E_j, respectively, their measured values can be written as $M_i = E_i r_i$ and $M_j = E_j r_j$. The ordering between the two measured values can be judged by the ratio of $M_i/M_j = (r_i/r_j)\ (E_i/E_j)$. If there is no bias, $r_i = r_j$, then M_i/M_j will be the same as E_i/E_j. If r_i and r_j are not the same but remain constant in the measurement range, M_i/M_j is proportional to the ground truth ratio, therefore the observed REOs may not reflect the true REOs, which will reduce the statistical power of the RankCompV2 algorithm but will not introduce false discoveries (Additional file 7). Furthermore, if the bias is not systematic, the misjudged REOs will distribute randomly in the four cells of the contingency table of the counts of numbers of gene pairs with $M_i > M_j$ or $M_i < M_j$. Therefore, the detection power is expected to be reduced. If the dynamic range is not linear, the situation will be more complicated. But we also expect that the main influence to our method is to reduce the detection power slightly. It means that the RankCompV2 algorithm can detect at least a part of the DEGs with absolute mRNA abundances changes, which still have biological significances. We believe that the power of the RankCompV2 algorithm will increase along with the improvement of gene expression measurement technologies.

Genomic copy number aberrations (CNAs) could also account for a substantial portion of gene expression changes. Currently, CNAs in cancer genomes are determined by comparing the measurements for the same amounts of DNA extracted from cancer and normal tissues, based on the assumption that the overall yields of DNA per cell (genome sizes) of different cell types are approximately the same. However, this assumption is least likely to hold because many cancers are aneuploid and/or polyploid [3]. The wrong assumption might lead to a serious consequence because CNAs are often used to determine cancer driver genes [47, 48]. Although several methods such as FREEC [49] and CNAnorm [50] have been proposed to correct the issues associated with cancer genome sizes in estimating copy number alterations based on deep sequencing data, there still exist some limitations. For example, FREEC cannot deal with patients' tumor samples due to the need of providing the ploidy of the most abundant copy number; CNAnorm is based on the assumption that tumor cells are largely monoclonal or polyclonal in a similar way, which could produce misleading results in tumors with large clonal variations [49, 50]. Furthermore, these methods cannot analyze the vast amount of microarray CNA data. The REOs comparison algorithm cannot be used to detect CNAs because, theoretically, the DNA intensity signals in normal cells should be equal. Due to the same problem of cancer cell aneuploid and/or polyploid [3], DNA methylation analyses using bisulfite-Seq data based on the same amount of DNA are also problematic [9]. Notably, the average beta value of a given locus measured by the Illumina beadarray can be interpreted as an estimate on the proportion of methylated cells to all measured cells [51–53]. Thus, when comparing two samples with different average DNA yields per cell, the differentially methylated loci will not be affected when similar amounts of DNA are extracted from different number of cells.

The RankCompV2 algorithm can only identify DEGs with sufficiently large expression changes that widely change the REOs of the genes from one phenotype to the other. On one hand, such DEGs might be of special biological significance, because functionally related genes tend to express coordinately in a stable state of physiological or pathological condition [54]. On the other hand, many genes with small absolute abundance changes would be determined as relative DEGs, which would blur the differences between the absolute DEGs and the relative DEGs. To learn more about the DEGs with changes in absolute mRNA abundances, it needs to develop new biological techniques and/or bioinformatics algorithms. Finally, we note that the studies on expression comparisons of microRNA and long non-coding RNA between two phenotypes are also based on the wrong assumption of similar overall yields of RNA molecules among different cells [9, 55], where the REO-based methods are applicable [56, 57]. It would be also interesting to study whether these RNA molecules with absolute abundances changes might have specific biological significances.

Conclusions

REO-based algorithm identified DEGs with changes in both mRNA concentration and absolute abundances. Through studies for all the ten cancers, we found DEGs with absolute mRNA abundance changes are more likely to be closely related with cancer driver genes and drug targets and enriched in DNA damage, metabolism and immune associated pathways. The genes with absolute mRNA abundances changes in cancers might play more important upstream roles in carcinogenesis.

Additional files

Additional file 1: Supplementary Method including Data and preprocessing, The SAM and edgeR algorithms and simulation experiments on null datasets. (DOCX 28 kb)

Additional file 2: Figure S1. Stacked bar chart for the distribution of the numbers of DEGs identified from the simulated null datasets among 100 repeated experiments. (TIF 1234 kb)

Additional file 3: Table S1. Numbers of DEGs identified from the null datasets. (XLSX 17 kb)

Additional file 4: The simulation experiments in data with global transcriptome size changes due to the genome size changes; **Table S4.** Simulation evaluation with different levels of copy number variations for RankCompV2. (DOCX 18 kb)

Additional file 5: Table S2. The numbers of DEGs identified by RankCompV2 and SAM or edgeR for each dataset. (DOCX 18 kb)

Additional file 6: Table S3. Pathway enrichment analysis of the absolute DEGs and relative DEGs for ten cancers. (XLSX 58 kb)

Additional file 7: The influence of measurement biases on RankCompV2. (DOCX 17 kb)

Abbreviations

BRCA: Breast invasive carcinoma; CNAs: Copy number aberrations; COAD: Colon adenocarcinoma; COSMIC: Catalogue of Somatic Mutations; DEGs: Differentially expressed genes; ESCA: Esophageal carcinoma; FDR: False discovery rate; GEO: Gene Expression Omnibus; HNSC: Head and Neck squamous cell carcinoma; KIRC: Kidney renal clear cell carcinoma; LIHC: Liver hepatocellular carcinoma; LUAD: Lung adenocarcinoma; LUSC: Lung squamous cell carcinoma; PAAD: Pancreatic adenocarcinoma; POG: Percentage of Overlapping Genes; REOs: Relative expression orderings; STAD: Stomach adenocarcinoma

Acknowledgements

Thank you to all the individuals who participated in this study. We would also like to acknowledge the resources at GEO and TCGA that facilitated this research.

Funding
This work was supported by the National Natural Science Foundation of China [81572935, 21534008, 81602738 and 61602119] and the Joint Scientific and Technology Innovation Fund of Fujian Province [2016Y9044].

Availability of data and materials
Previously data analyzed in this study should be requested from the authors of the original publications. Please see methods cohort description, for references to these publications.

Authors' contributions
ZG and XLW conceived the project. HC and XYL designed data analyses. HC, LXY, HJ and ZWB performed experiments. KS, YG, HPL, QZG and HDY interpreted data. HC, ZG and XLW wrote the manuscript. All authors contributed to the preparation of the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

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

Author details
[1]Medical Big Data and Bioinformatics Research Centre, First Affiliated Hospital of Gannan Medical University, Ganzhou 341000, Jiangxi, China. [2]Department of Bioinformatics, Fujian Key Laboratory of Medical Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, Fujian Medical University, Fuzhou 350122, Fujian, China. [3]Department of Systems Biology, College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150086, Fujian, China.

Received: 13 October 2018 Accepted: 31 January 2019
Published online: 13 February 2019

References
1. Loven J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. Revisiting global gene expression analysis. Cell. 2012; 151(3):476–82.
2. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, Wang R, Green DR, Tessarollo L, Casellas R, et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell. 2012;151(1):68–79.
3. Weaver DA, Nestor-Kalinoski AL, Craig K, Gorris M, Parikh T, Mabry H, Allison DC. Corrections for mRNA extraction and sample normalization errors find increased mRNA levels may compensate for cancer haplo-insufficiency. Genes Chromosomes Cancer. 2014;53(2):194–210.
4. Birchler JA. Facts and artifacts in studies of gene expression in aneuploids and sex chromosomes. Chromosoma. 2014;123(5):459–69.
5. Stevens JB, Horne SD, Abdallah BY, Ye CJ, Heng HH. Chromosomal instability and transcriptome dynamics in cancer. Cancer Metastasis Rev. 2013;32(3–4):391–402.
6. Coate JE, Doyle JJ. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol Evol. 2010;2:534–46.
7. Coate JE, Doyle JJ. Variation in transcriptome size: are we getting the message? Chromosoma. 2015;124(1):27–43.
8. Aanes H, Winata C, Moen LF, Ostrup O, Mathavan S, Collas P, Rognes T, Alestrom P. Normalization of RNA-sequencing data from samples with varying mRNA levels. PLoS One. 2014;9(2):e89158.
9. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. Mol Cell Biol. 2015;36(5):662–7.
10. Qing T, Yu Y, Du T, Shi L. mRNA enrichment protocols determine the quantification characteristics of external RNA spike-in controls in RNA-Seq studies. Sci China Life Sci. 2013;56(2):134–42.
11. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014; 32(9):896–902.
12. Wang H, Sun Q, Zhao W, Qi L, Gu Y, Li P, Zhang M, Li Y, Liu SL, Guo Z. Individual-level analysis of differential expression of genes and pathways for personalized medicine. Bioinformatics. 2015;31(1):62–8.
13. Cai H, Li X, Li J, Liang Q, Zheng W, Guan Q, Guo Z, Wang X. Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. Int J Biol Sci. 2018;14(8):892–900.
14. Li X, Cai H, Wang X, Ao L, Guo Y, He J, Gu Y, Qi L, Guan Q, Lin X, et al. A rank-based algorithm of differential expression analysis for small cell line data with statistical control. Brief Bioinform. 2017. https://doi.org/10.1093/bib/bbx135.
15. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–21.
16. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
17. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11(3):R25.
18. Gilbert PB. A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. J R Stat Soc Ser C. 2005;54(1):143–58.
19. Tarone RE. A modified Bonferroni method for discrete data. Biometrics. 1990;46(2):515–22.
20. Fellows I. The minimaxity of the mid *p*-value under linear and squared loss functions. Commun Stat Theory Methods. 2010;40(2):244–54.
21. Austin SR, Dialsingh I, Altman N. Multiple hypothesis testing: a review. Indian Soc Agric Stat. 2014;68:303–14.
22. Muralidharan O. An empirical Bayes mixture method for effect size and false discovery rate estimation. Ann Appl Stat. 2010;4(1):422-38.
23. Martin R, Tokdar ST. A nonparametric empirical Bayes framework for large-scale multiple testing. Biostatistics. 2012;13(3):427–39.
24. Hochberg YBY. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol. 1995;57(1):289–300.
25. Heller R, Gur H. False discovery rate controlling procedures for discrete tests. arXiv preprint arXiv; 2011. p. 1112. 4627
26. Zhang M, Yao C, Guo Z, Zou J, Zhang L, Xiao H, Wang D, Yang D, Gong X, Zhu J, et al. Apparently low reproducibility of true differential expression discoveries in microarray studies. Bioinformatics. 2008;24(18): 2057–63.
27. Hong G, Zhang W, Li H, Shen X, Guo Z. Separate enrichment analysis of pathways for up- and downregulated genes. J R Soc Interface. 2014;11(92): 20130950.
28. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. Nucleic Acids Res. 2002;30(1):42–6.
29. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundances using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci. 2012;131(4):281–5.
30. Chudin E, Walker R, Kosaka A, Wu SX, Rabert D, Chang TK, Kreder DE. Assessment of the relationship between signal intensities and transcript concentration for Affymetrix GeneChip arraysGenome Biol. 2002;3(1): RESEARCH0005.
31. Skvortsov D, Abdueva D, Curtis C, Schaub B, Tavare S. Explaining differences in saturation levels for Affymetrix GeneChip arrays. Nucleic Acids Res. 2007; 35(12):4154–63.
32. Gharaibeh RZ, Fodor AA, Gibas CJ. Accurate estimates of microarray target concentration from a simple sequence-independent Langmuir model. PLoS One. 2010;5(12):e14464.
33. Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004;4(3):177–83.
34. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2015;43(Database issue):D447–52.

35. Chung IF, Chen CY, Su SC, Li CY, Wu KJ, Wang HW, Cheng WC. DriverDBv2: a database for human cancer driver gene research. Nucleic Acids Res. 2016; 44(D1):D975–9.
36. Kumar R, Chaudhary K, Gupta S, Singh H, Kumar S, Gautam A, Kapoor P, Raghava GP. CancerDR: cancer drug resistance database. Sci Rep. 2013;3:1445.
37. Kaschutnig P, Bogeska R, Walter D, Lier A, Huntscha S, Milsom MD. The Fanconi anemia pathway is required for efficient repair of stress-induced DNA damage in haematopoietic stem cells. Cell Cycle. 2015; 14(17):2734–42.
38. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144(5):646–74.
39. Xu H, Guo X, Sun Q, Zhang M, Qi L, Li Y, Chen L, Gu Y, Guo Z, Zhao W. The influence of cancer tissue sampling on the identification of cancer characteristics. Sci Rep. 2015;5:15474.
40. Chen R, Guan Q, Cheng J, He J, Liu H, Cai H, Hong G, Zhang J, Li N, Ao L, et al. Robust transcriptional tumor signatures applicable to both formalin-fixed paraffin-embedded and fresh-frozen samples. Oncotarget. 2017;8(4):6652–62.
41. Cheng J, Guo Y, Gao Q, Li H, Yan H, Li M, Cai H, Zheng W, Li X, Jiang W, et al. Circumvent the uncertainty in the applications of transcriptional signatures to tumor tissues sampled from different tumor sites. Oncotarget. 2017;8(18):30265–75.
42. Zheng W, Chung LM, Zhao H. Bias detection and correction in RNA-sequencing data. BMC Bioinf. 2011;12:290.
43. Ekblom R, Smeds L, Ellegren H. Patterns of sequencing coverage bias revealed by ultra-deep sequencing of vertebrate mitochondria. BMC Genomics. 2014;15:467.
44. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.
45. Qi L, Chen L, Li Y, Qin Y, Pan R, Zhao W, Gu Y, Wang H, Wang R, Chen X, et al. Critical limitations of prognostic signatures based on risk scores summarized from gene expression levels: a case study for resected stage I non-small-cell lung cancer. Brief Bioinform. 2016;17(2):233–42.
46. Guan Q, Yan H, Chen Y, Zheng B, Cai H, He J, Song K, Guo Y, Ao L, Liu H, et al. Quantitative or qualitative transcriptional diagnostic signatures? A case study for colorectal cancer. BMC Genomics. 2018;19(1):99.
47. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Pe'er D. An integrated approach to uncover drivers of cancer. Cell. 2010;143(6):1005–17.
48. Wang K, Lim HY, Shi S, Lee J, Deng S, Xie T, Zhu Z, Wang Y, Pocalyko D, Yang WJ, et al. Genomic landscape of copy number aberrations enables the identification of oncogenic drivers in hepatocellular carcinoma. Hepatology. 2013;58(2):706–17.
49. Boeva V, Zinovyev A, Bleakley K, Vert JP, Janoueix-Lerosey I, Delattre O, Barillot E. Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. Bioinformatics. 2011; 27(2):268–9.
50. Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. Bioinformatics. 2012;28(1):40–7.
51. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E, et al. High-throughput DNA methylation profiling using universal bead arrays. Genome Res. 2006;16(3):383–93.
52. Houseman EA, Christensen BC, Karagas MR, Wrensch MR, Nelson HH, Wiemels JL, Zheng S, Wiencke JK, Kelsey KT, Marsit CJ. Copy number variation has little impact on bead-array-based measures of DNA methylation. Bioinformatics. 2009;25(16):1999–2005.
53. Feber A, Guilhamon P, Lechner M, Fenton T, Wilson GA, Thirlwell C, Morris TJ, Flanagan AM, Teschendorff AE, Kelly JD, et al. Using high-density DNA methylation arrays to profile copy number alterations. Genome Biol. 2014; 15(2):R30.
54. Wang D, Cheng L, Zhang Y, Wu R, Wang M, Gu Y, Zhao W, Li P, Li B, Zhang Y, et al. Extensive up-regulation of gene expression in cancer: the normalised use of microarray data. Mol BioSyst. 2012;8(3):818–27.
55. Wu D, Hu Y, Tong S, Williams BR, Smyth GK, Gantier MP. The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. Rna. 2013;19(7):876–88.
56. Yan H, Cai H, Guan Q, He J, Zhang J, Guo Y, Huang H, Li X, Li Y, Gu Y, et al. Individualized analysis of differentially expressed miRNAs with application to the identification of miRNAs deregulated commonly in lung cancer tissues. Brief Bioinform. 2017. https://doi.org/10.1093/bib/bbx015.
57. Peng F, Wang R, Zhang Y, Zhao Z, Zhou W, Chang Z, Liang H, Zhao W, Qi L, Guo Z, et al. Differential expression analysis at the individual level reveals a lncRNA prognostic signature for lung adenocarcinoma. Mol Cancer. 2017; 16(1):98.