Published online 2 July 2021

Nucleic Acids Research, 2021, Vol. 49, No. 17 ***e99***
https://doi.org/10.1093/nar/gkab554

Rank-in: enabling integrative analysis across microarray and RNA-seq for cancer

Kailin Tang[1], Xuejie Ji[1], Mengdi Zhou[1], Zeliang Deng[1], Yuwei Huang[1,2], Genhui Zheng[1] and Zhiwei Cao[1,*]

[1]Department of Gastroenterology, Shanghai 10th People's Hospital and School of Life Sciences and Technology, Tongji University, 1239 Siping Road, Shanghai 200092, P.R. China and [2]CAS Key Laboratory of Computational Biology, Bio-Med Big Data Center, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Science, Shanghai 200031, P.R. China

Received January 07, 2021; Revised May 10, 2021; Editorial Decision June 08, 2021; Accepted June 25, 2021

ABSTRACT

Though transcriptomics technologies evolve rapidly in the past decades, integrative analysis of mixed data between microarray and RNA-seq remains challenging due to the inherent variability difference between them. Here, Rank-In was proposed to correct the nonbiological effects across the two technologies, enabling freely blended data for consolidated analysis. Rank-In was rigorously validated via the public cell and tissue samples tested by both technologies. On the two reference samples of the SEQC project, Rank-In not only perfectly classified the 44 profiles but also achieved the best accuracy of 0.9 on predicting TaqMan-validated DEGs. More importantly, on 327 Glioblastoma (GBM) profiles and 248, 523 heterogeneous colon cancer profiles respectively, only Rank-In can successfully discriminate every single cancer profile from normal controls, while the others cannot. Further on different sizes of mixed seq-array GBM profiles, Rank-In can robustly reproduce a median range of DEG overlapping from 0.74 to 0.83 among top genes, whereas the others never exceed 0.72. Being the first effective method enabling mixed data of cross-technology analysis, Rank-In welcomes hybrid of array and seq profiles for integrative study on large/small, paired/unpaired and balanced/imbalanced samples, opening possibility to reduce sampling space of clinical cancer patients. Rank-In can be accessed at http://www.baddcao.net/rank-in/index.html.

INTRODUCTION

The past decades have witnessed the rapid development of transcriptomics technology and widely application into the cancer area, where the top two are microarray and RNA-seq platforms (1). Till October 2020, GEO hosted near 10 000 series of cancer transcriptomic data including about 5500 series from microarray and 3000 series from RNA-Seq, produced by multiple laboratories through the various version of platforms from Affymetrix, Agilent, Illumina and so on (2). To derive significant results statistically, datasets of cancer samples often need to be pooled together as many as possible for bioinformatics computation. Yet, gene expression profiles are routinely required to be compared only with controls in the same technology such as seq-cancer versus seq-control or array-caner versus array-control. Integrative analysis of mixed data remains difficult due to technology designing differences, platform variations and batch effects. Particularly, array platforms were designed by different sets of probes to detect signal sensitivity, while RNA-seq was designed to count the copy number of transcripts. For microarray data, log-expression is a continuous measurement that is approximately distributed within a certain range as a normal Gaussian random variable (3). RNA-Seq data, however, is measured in integer counts without a limited peak, which does not follow the typical normal distribution (4). The inherent design difference between them often causes the expression profiling incomparable, which differs systematically in one technology of array from that in another of RNA-seq, even to the same biological samples (5).

To explore the gaps between them, FDA started MAQC (Micro Array Quality Control) (6) and SEQC (Sequencing Quality Control) (7) projects to run both technology platforms in different labs by providing the same paired-RNA samples. The reference samples they provided included human mixed RNA samples A (Universal Human Reference RNA), human brain tissue sample B (Human Brain Reference RNA) and two mixture samples of the above two at different ratios (8). The results showed that systemic variations exist across the two technologies in the value of expression measures (9), and the datasets were not feasible to

[*]To whom correspondence should be addressed. Tel: +86 21 65980296; Email: zwcao@tongji.edu.cn

pool together for direct analysis. Also, the researchers constructed thousands of predictive models to test whether the computation results from one technology agree with that from another on the same human and rat samples, but the correlations were showed at a rugged range between 0.38 and 0.90 depending on different samples, labs and platform versions (10).

Despite that, some methods have been elaborated to correct the above nonbiological effects. The first class is to adjust the batch effects based on empirical models (11). A notable representative is ComBat, which was originally designed for microarray experiments only (12). It used the empirical Bayes method to estimate and shrink batch effects by balancing individual sets of expression results. In 2020, a parallel version of ComBat-seq was released for RNA-seq integrative analysis (13). Additional strategies aimed to filter out factors that are associated with batch effects by factorizing the input expression data and then reconstructing an adjusted matrix (14). A typical example is SVA (15), which was designed by estimating surrogate variables of the unknown effects and iteratively weighting a subset of the factors identified in the decomposition. Currently, SVA and its updated version (16) have been widely applied to bulk and single-cell RNA-seq analysis (17,18). Meanwhile, spike-in marker genes were also adopted to adjust cross-platform biases, such as housekeeping genes (19,20) or so-called ‘bias-low gene sets’ (21). Yet this hypothesis has been questionable (22), as the filtered gene signatures were reported to be highly unstable (23), depending on patient samples, disease states (24) or tissues (25). On top of the above, machine learning classifiers have also been attempted to cross-platform analysis (26), such as FSQN, which rendered RNA-seq analysis from the training of microarray data distribution (27).

So far, no methods have met well with the challenge of cross-technology data integration in this area.

In this paper, we designed a computational method, Rank-In, to make mixed datasets comparable. Our idea is to transform the raw expression into relative ranking within each profile, then weighted by the distribution of overall expression intensities among consolidated datasets. By minimizing the profiling variations between array and RNA-seq, Rank-In made it possible to combine microarray and RNA-seq data for further analysis. Independent testing on cell data and clinical samples was comprehensively done to validate the performance and robustness of Rank-In. A webserver was also set up to allow community application, with example datasets built-in for cancer subtypes.

MATERIALS AND METHODS

Design of Rank-In

Rank-In was designed as below in Figure 1. For the raw transcriptomic profiling, genes are first lining up according to the gene expression increase within each transcriptomic profile, and then the sorted genes are partitioned into 100 groups (28). Second, the ranks of gene groups are further weighted by the increasing slope of expression intensity within each group of genes for each profile. So far, a weighted internal gene ranking has been derived for each profile. Third, across consolidated profiles, singular value decomposition (SVD) is performed to reduce the nonbiological effects between different experiments or platforms. Via this, the data distribution has been unified into a similar curve across array and RNA-seq. Finally, the adjusted ranking matrix can be utilized for subsequent comparison or analysis.

Algorithm of Rank-In

Transformation of gene-level expression intensities into internal ranks. For each profile, the raw expression intensity of each gene is ranked from low to high. Then, genes are divided into 100 groups, with the same number of genes in each group (28). The continuous expression values thus have been converted to internal ranks within each profile.

Weighting ranks based on expression intensities. We define the expression intensity of gene *i* in profile *j* as e_{ij}, and the internal ranking of the gene *i* in the profile *j* as r_{ij}. The weight of r_{ij} is described as $w_{ij} = 2ar_{ij} + b$, where a and b are the coefficients of the equation $e_{ij} = ar_{ij}{}^2 + br_{ij} + c$. These coefficients are calculated by the least square function in SciPy. Formally, the adjusted ranking matrix $R_{N\times M}$ was calculated as:

$$R_{ij} = r_{ij} \cdot w_{ij}$$

where *N* denoted the whole number of genes and *M* denoted the number of profiles in each class.

Using matrix decomposition to remove nonbiological effects. We assume the following model for the observed data $R_{N\times M}$:

$$R_{N\times M} = x_{N\times M} + y_{N\times M} + \alpha_j + \varepsilon_{N\times M}$$

Where $x_{N\times M}$ is a matrix of the overall gene expression, $y_{N\times M}$ is a matrix of the nonbiological batch effects of genes, α_j is the effect corresponding to experimental conditions, and the term $\varepsilon_{N\times M}$ represents random noise.

The mean of the rank of the weighted gene would be estimated close to the true rank. Here, we get the variance matrix of genes:

$$R_{real} = x_{N\times M} + \alpha_j \approx Me_{ij}$$

$$R'_{N\times M} = R_{N\times M} - Me_{ij} = y_{N\times M} + \varepsilon_{N\times M}$$

Where Me_{ij} denotes mean of the *i*-th gene of the experimental group or control group, $R_{N\times M}$ is the adjusted ranking matrix, $R'_{N\times M}$ is the variance matrix of genes including random errors and nonbiological effects.

The mean of different groups is used to represent the actual values of the experiments, to approximate the matrix of the weighted gene rank changes caused by the error, and the nonbiological effect matrix is obtained from the change matrix by the SVD method. Here, we get the nonbiological effect matrix:

$$R'_{N\times M} = R_{N\times M} - Me_{ij} = y_{N\times M} + \varepsilon_{N\times M} = U\sum V^T$$

Where $U_{N\times N}$, $\sum_{k\times k}$ and $V_{M\times M}$ are sub-matrices obtained by SVD decomposition.

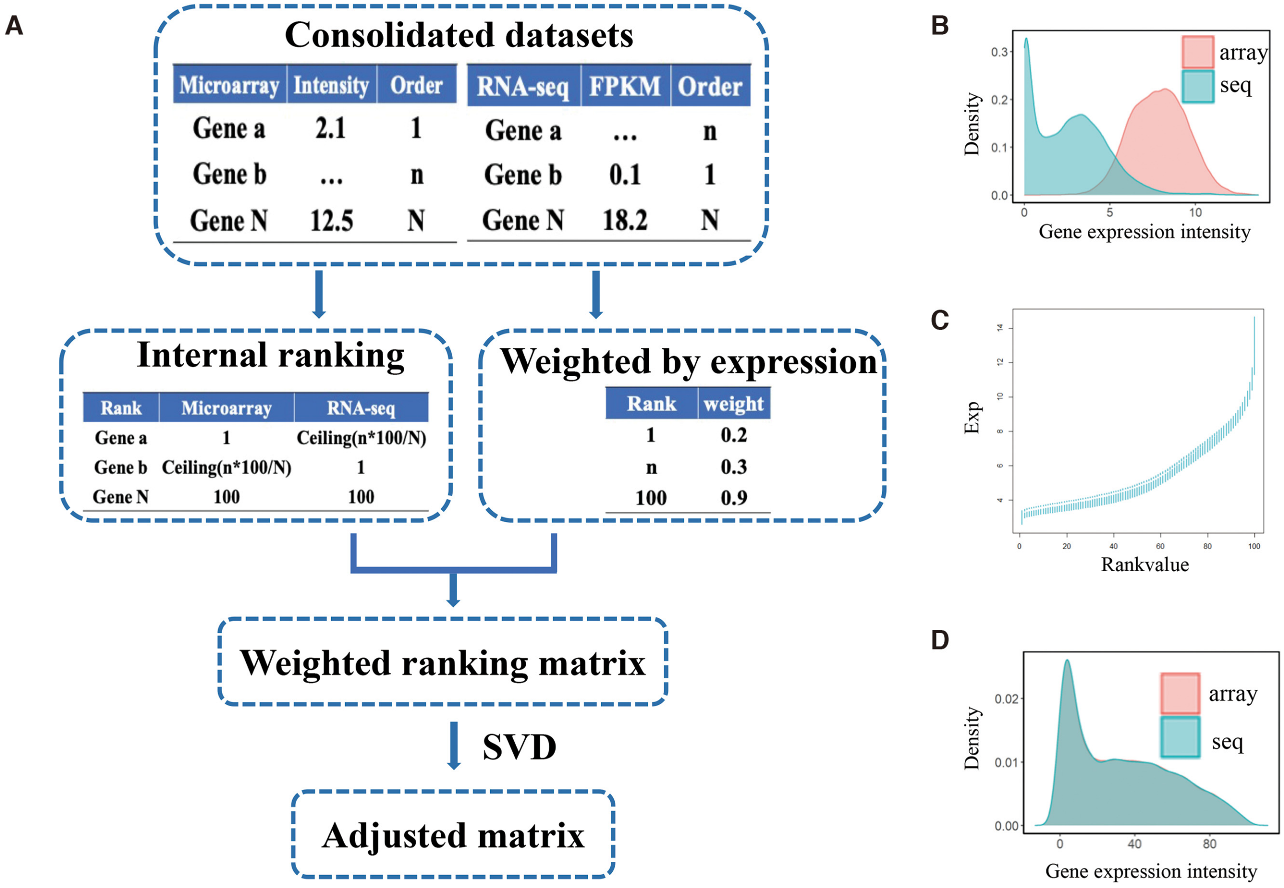


Figure 1. The workflow of Rank-In. (**A**) The Rank-In workflow. The consolidated expression profiles from microarray and RNA-seq are transformed into internal ranking and further weighted by intensity increasing, then calculated by SVD into the adjusted ranking matrix. (**B**) The data distribution difference of raw profiles between microarray and RNA-seq. (**C**) The sorted ranking according to expression. (**D**) The data distribution from microarray and RNA-seq after Rank-In.

Noted that the variance matrix of genes $R'_{N\times M}$ mostly comes from nonbiological effects, the SVD method is used to observe the magnitude of the singular value and determine the k value to maintain the random error and approximate the nonbiological effects such as platform effects.

$$R'_{N\times M} = U_{N\times k} \sum_{k\times k} V^{T}_{M\times k} + \varepsilon_{N\times M}$$

$$y_{N\times M} \approx U_{N\times k} \sum_{k\times k} V^{T}_{M\times k}$$

Where k is the main number of nonbiological variables, and $y_{N\times M}$ is the nonbiological effect matrix.

Nonbiological effects are eliminated by subtracting nonbiological effects from the original matrix. Thus, we have adjusted the nonbiological effects:

$$R^{Adjust}_{N\times M} = R_{N\times M} - y_{N\times M} = x_{N\times M} + \alpha_j + \varepsilon_{N\times M}$$

Where $R^{Adjust}_{N\times M}$ is the final adjusted matrix that removes most of the nonbiological effects.

A complete list of all software used in Rank-In is provided in Supplementary Table S1.

Data preprocessing

The original expression intensity of microarray data is log-transformed based on 2. Probe IDs are mapped to gene IDs using the latest corresponding platform annotation files. Multiple probes are mapped to the same gene, the arithmetic mean of the values of the multiple probes could be used as the expression value of this gene.

For the RNA-seq data, original counts (.sra files) are downloaded and the fragments per kilobase of transcript per million fragments mapped (FPKM) is calculated by Tophat2 (29) and cufflinks (30). Transcripts per kilobase million (TPM) is calculated according to the definition (counts per length of transcript [kb] per million reads mapped), while the trimmed mean of M values (TMM) is calculated by edgeR package (31). Those genes with zero counts (or zero FPKM/TPM/TMM) in all profiles are excluded. All data are then log2(x + 1) transformed.

From the MAQC project, 1044 genes are validated by TaqMan quantitative PCR (32). About 328 genes are finally defined as true DEGs (a log2 fold change >2 and P-value <0.05 between class A and B) and 93 genes are filtered as true non-DEGs (a log2 fold change <0.2 and P-value >0.05 between class A and B).

Differentially expressed genes (DEGs) selection

DEGs are calculated by a nonparametric approach (Wilcoxon Rank Sum test). An FDR (false discovery rate) threshold of 0.05 for multiple testing is used.

Evaluation parameters

To evaluate the performance of Rank-in, parameters include three aspects: accuracy, precision and recall. Accuracy is the similarity between the measurements of a prediction and actual (true) value. The precision (or called specificity) is the probability of its positive claims being correct. The recall (or called sensitivity) is the probability of facts being claimed as true. Formally, the accuracy, precision and recall are defined as follows:

accuracy = (TP+TN)/(TP+FP+TN+FN)
precision = TP/(TP+FP)
recall = TP/(TP+FN)

where TP, FP denotes the true/false positives, and TN, FN denotes the true/false negatives.

RESULTS

Rank-In was comprehensively validated on both SEQC cell data and TCGA clinical data, where the same biological samples were tested on both microarray and RNA-seq platforms. The performance of Rank-In was evaluated via the unsupervised hierarchical clustering effects and the ability to pick up true differential expression genes (DEGs), in comparing with four representative peers, including uncorrected method without processing, Combat (12), SVA (15) and Angel's method (21).

Perfect clustering on SEQC datasets

At the level of cell samples, GEO data of MAQC/SEQC project (GSE56457 and GSE47774) were selected, covering two unique and homogenous RNA only (sample A and sample B). Each sample was sent to different platforms and labs to obtain transcriptomic profiles. Considering the data balance, 22 profiles were chosen for each biological sample, covering 14 from 3 array platforms and 8 from one RNA-seq platform (Supplementary Table S2).

The clustering results were illustrated in Figure 2A. As being shown, without adjustment, the uncorrected method failed to distinguish the profiles of the two biological samples. All the other four (Rank-In, ComBat, SVA and Angel's method) can group A and B. Theoretically, without technology or platform biases, different profiles representing the same biological condition would likely cluster in a random way. In other words, clustering according to platform or technology, instead of biological conditions, frequently indicates the existence of substantial biases. In Figure 2A, within sample A or B, profiles have been sub-clustered according to various platforms in the three peers. Only Rank-In aligned those sibling profiles in an approximately random manner regardless of their original platforms, indicating its ability to reduce the platform variation. Here, sibling profiles were tentatively referred to as those multiple profiles sequenced by different labs and platforms, but they were all derived from the same biological sample such as sample A or B. On top of that, extra RNA-seq platform was added and tested (Supplementary Table S3). Rank-In remained the best among peers (Supplementary Figure S1).

The data distribution of expression intensity processed by each method was checked and compared between array and RNA-seq technologies, as Figure 2B showed. The raw data without processing demonstrate two peaks in RNA-seq but one completely deviating peak in microarray. In contrast to the large difference in the uncorrected method, Combat and SVA have significantly shrunk the discrepancy by formulating two peaks and Angel's method by deleting all peaks. Note that, Rank-In almost removed the curve deviation between microarray and RNA-seq and merged them into one-peak distribution.

Table 1. Performance of different methods* on detecting DEGs† on SEQC data

Method	Accuracy	Precision	Recall
Rank-In	**0.90**	**0.91**	**0.97**
ComBat	0.87	0.86	0.99
SVA	0.87	0.86	0.99
Uncorrected	0.45	0.99	0.29

*Angel's method is not fit to derive DEGs and thus not shown.
†Genes with FDR value <0.05 are grouped as predicted DEGs and those ≥0.05 as predicted non-DEGs.

High DEG detecting rate on SEQC data

In addition to the overall clustering ability, the performance of detecting DEGs was tested for different methods. In the SEQC project, 1044 genes were validated by TaqMan quantitative PCR (32), where 328 genes were finally defined as true positive DEGs and 93 genes were filtered as true negative non-DEGs between sample A and B. Since Angel's method is not fit to pick DEGs, only the left peers were put under test on predicting the DEG list between sample A and B. In line with the true positives and true negatives from the above PCR results, the prediction accuracy, precision and recall were summarized in Table 1. Accuracy represents the overall predictive performance, while precision indicates the ability that the predicted DEG becomes true positives, and recall means how many of the true positives can be predicted out (33). It can be seen from Table 1 that, without correction, though high precision (0.99) can be made, only a very minor portion of true DEGs can be picked with a recall rate of 0.29, leading to the lowest accuracy of 0.45. ComBat and SVA gave almost the same accuracy of 0.87, with the precision of 0.86 and the highest recall of 0.99, indicating a slightly higher false-positive rate in detecting DEG. Among the peers, Rank-In obtained a balanced precision of 0.91 and a recall rate of 0.97. Due to the trade-off between precision and recall, Rank-In achieved the best accuracy of 0.90 in terms of overall DEG-detecting ability. Small variation was noted on SEQC data among peer performance. The reason may come from the apparent DEG difference between homogenous brain RNA and the homogenous universal RNA samples, which would be easily detected by all methods.

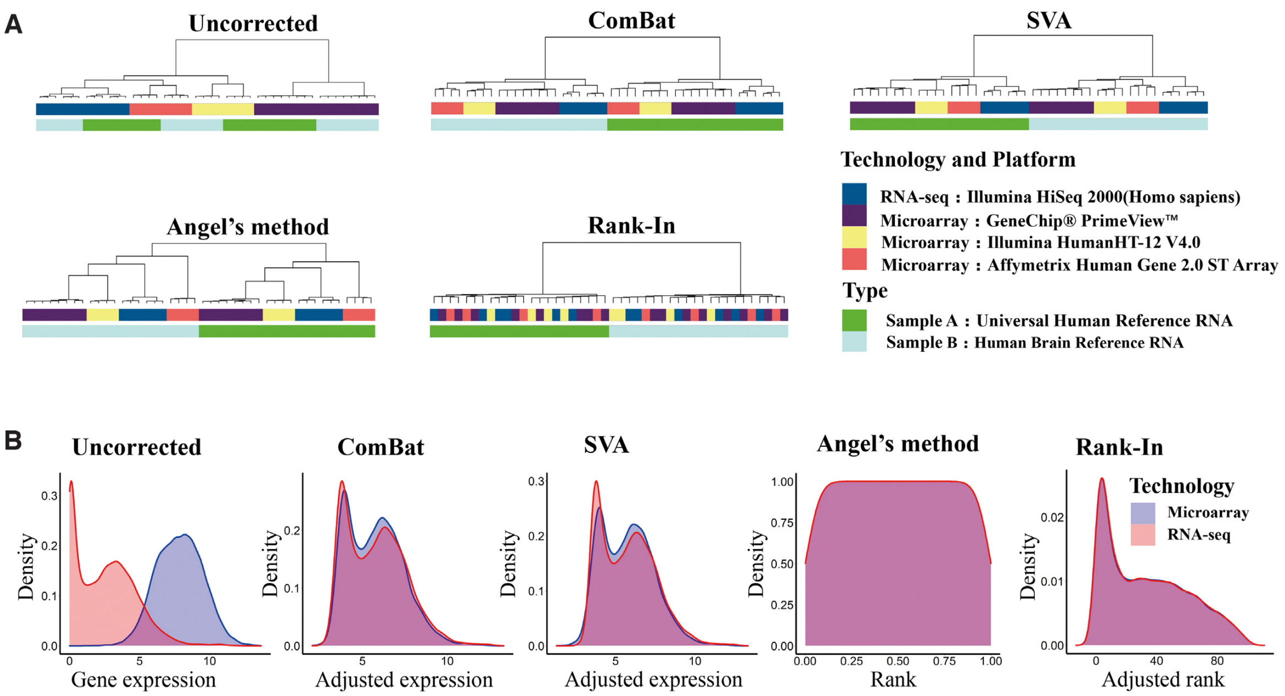


Figure 2. Performance and distribution comparison of Rank-In and other methods on SEQC data. (**A**) Unsupervised clustering of SEQC data. The upper horizontal bars illustrate four different platforms, and the bottom bars illustrate the sample type of A and B. (**B**) The data distribution of profiles from microarray and RNA-seq before and after adjustment.

Best performance on clustering clinical samples

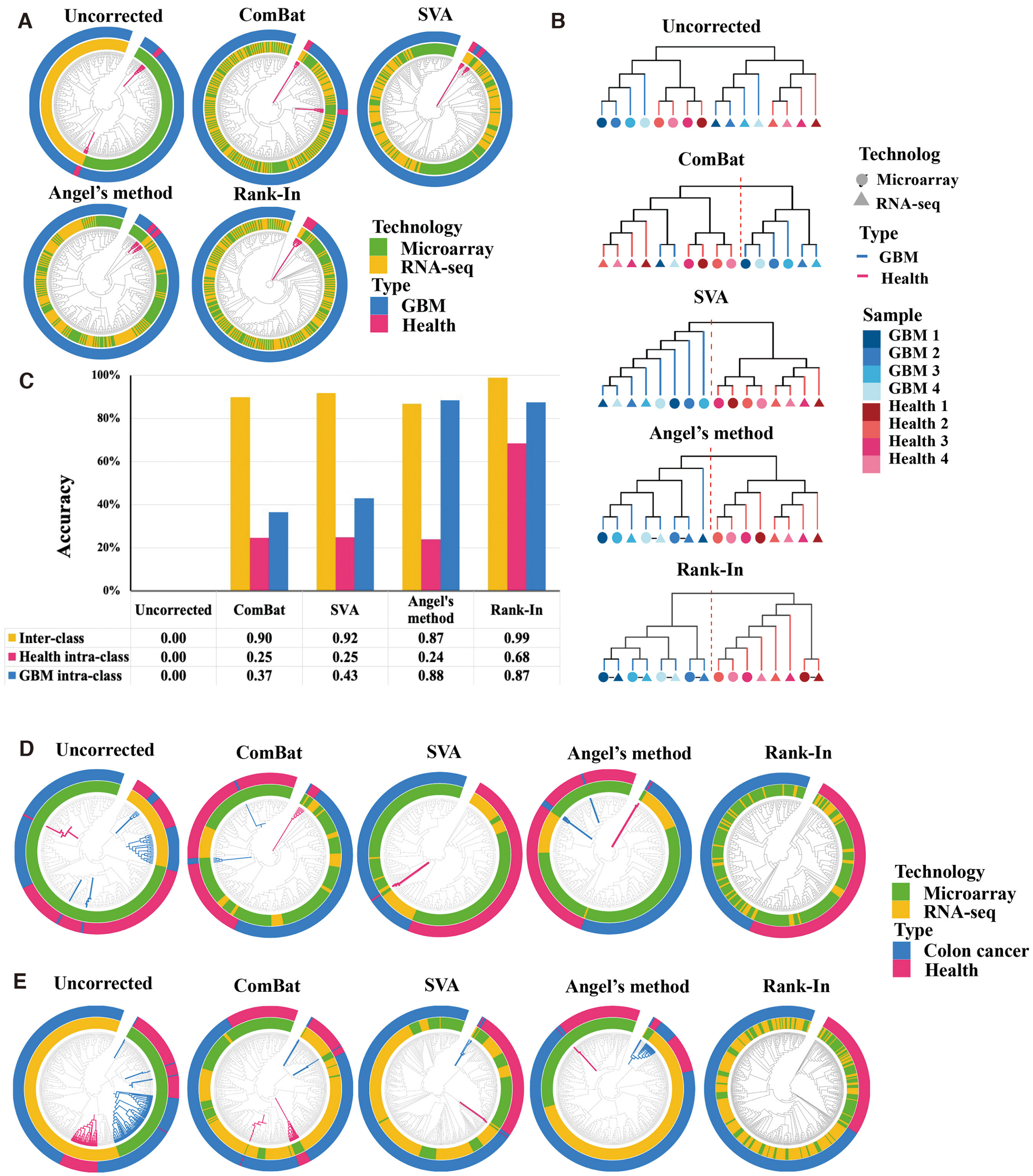
In reality of clinical samples, the major challenge lies in heterogeneous variations from personal backgrounds to tumor stages and tissue differentiation etc. The Glioblastoma (GBM) data were collected from the TCGA database where each clinical sample was tested by both microarray and RNA-seq (34). A total of 327 expression profiles were incorporated, covering 157 GBM patients and 4 healthy controls (Supplementary Table S4).

Unsupervised hierarchical clustering was made on the above profiles for each method (Figure 3A). It is shown that the uncorrected method clustered profiles into two categories fully according to the technologies of RNA-seq and microarray. Whereas ComBat and SVA clustered four heath samples into disease groups, Angel's method mixed the health with GBM profiling, though health samples were grouped at a closer distance. Only Rank-In perfectly clustered these health samples into one branch out of GBM samples. Further, the GBM expression datasets were approximately uniformly aligned regardless of their initial technology or platform, suggesting the lowest bias after Rank-In processing.

Further, as ComBat was reported not to work well if sample classes were not properly balanced across batches (35), resampling was made to construct a balanced dataset between tumor and health. To make a fair play, paired microarray and RNA-seq data of four GBM patients were randomly selected out of 157 GBM patients 1000 times to match the four health controls. A representative clustering tree is illustrated for each method in Figure 3B. Subsequent statistics was made on clustering effects from two perspectives: inter-class differentiation between tumor and health profiles, and intra-class clustering effects of individual samples within tumor and health (Figure 3C). It can be seen that, on this small dataset of 16 profiles with balanced data, the inter-class performance of ComBat, SVA and Angel's indeed got increased. At 90% probability, ComBat can make correct classification between tumor and health profiles. Rank-In achieved the highest inter-class differentiation accuracy at 99%, and the second was obtained by SVA.

For intra-class performance, clustering trees of the eight profiles from four tumor/health samples may present in different formats depending on the relative distance between them after data processing. According to the best and worst clustering illustrated in Supplementary Figure S2, the distance between paired profiles from the same sample may vary between 0 and 7. We chose the distance of 4 as a moderate stringent cutoff and calculated the averaged percentage of clustered samples with paired profiles being ≤4 in all the 4000 simulations (Figure 3C). It can be seen that Rank-In and Angel's method performs similarly well on clustering GBM tumor profiles. However, on clustering health profiles, Angel's only gives 24%, similar to SVA and Combat. While Rank-In archived a high performance of 68%, demonstrating a significant advantage over other peers.

To further test the performance on more comprehensive clinical application, paired/unpaired and balanced/imbalanced conditions of larger colon cancer datasets from https://xenabrowser.net/datapages and literature (36) were curated with cancer versus normal of 124:124 and 384:139 respectively (Supplementary Table S5). Rank-In appears as the only one to correctly classify every single cancer from normal profiles on different



	Uncorrected	ComBat	SVA	Angel's method	Rank-In
Inter-class	0.00	0.90	0.92	0.87	0.99
Health intra-class	0.00	0.25	0.25	0.24	0.68
GBM intra-class	0.00	0.37	0.43	0.88	0.87

Figure 3. Clustering performance on clinical data of GBM and colon cancer. (**A**) Unsupervised hierarchical clustering on 319 GBM and 8 health profiles. (**B**) Representative clustering tree from the 1000 randomizations. (**C**) Clustering accuracy of inter-class of GBM versus health, intra-class of GBM and intra-class of health from 1000 randomizations. (**D**) Unsupervised hierarchical clustering results on 124 colon cancer and 124 normal profiles in a paired and balanced condition. (**E**) Unsupervised hierarchical clustering results on 384 colon cancer and 139 normal profiles in unpaired and imbalanced condition.

scenarios, while others cannot (Figure 3D,E). Collectively, Rank-In demonstrated the best performance in discriminating cancer from normal profiles, also in clustering sibling profiles of the same biological sample, indicating its unique ability in reducing nonbiological effects across microarray and RNA-seq.

High robustness in tolerating different sample size

The sample size to reproduce DEGs was also tested on both simulated samples and GBM samples respectively on different scenarios. Theoretically, the DEG list between the batches, platforms or technologies would be highly overlapped once the technology and platform effects being properly corrected.

In simulated condition, both model performance on different sample size and model tolerance to imbalanced sample size were tested with 200 predefined DEGs among 10 000 genes for each sample. The sample size varied from 14 to 500, containing an equal number of disease and control in each batch, and gene expression profiles were randomly created 100 times for each sample and divided into parallel batches or platforms by an R package ('madsim' package) (37). Since small datasets may not necessarily lead to reproducible DEG, the overlapping rates between the top 200 DEGs from the parallel platforms were shown in Figure 4A. Rank-In gave a rate of 0.61 at the smallest sample size of 14 in each platform, further 0.82 at the sample size of 40, and 0.9 at a sample size of 150. Model tolerance to imbalanced sample size was done within a fixed number of 500 samples in each platform. The samples of the control group ranged from 10 to 250, corresponding to the disease group from 490 to 250. As shown in Figure 4B, under different imbalance ratios, Rank-In gave consistently high overlapping DEG ranging from 0.81 to 0.92. Particularly, under extreme imbalanced condition (control 10 versus disease 490), the lowest overlapping rate remained around 0.81, indicating the tolerance of our model to imbalanced sample size.

On clinical data, GBM samples were randomly drawn to pool together with health samples into similar size with simulated data. As being shown in Figure 4C, the Rank-In got the overlapping rate of 0.74 at the smallest sample size of 14 in each platform, then gradually grew to 0.78 when the sample size increased to 40. After that, the rate stabilized around 0.79 with similar variation despite the increasing sample size. Thus, Rank-In seems to work robustly on relatively small consolidated datasets though a larger sample size may give a better result.

Furthermore, we investigated the ability of Rank-In to detect DEG in GBM case. For the 157 GBM and 4 health tissue samples, their paired array or RNA-seq profiles are randomly split into two sets, with each containing the same list of tissue samples. DEGs were derived for individual set between 157 GBM and 4 health control, and further overlapping between the two sets. Thus, the higher overlapping rate indicates the better ability of the method in removing technology effects. After 100 times of split randomization, the overlapped rate (*y*-axis) was plotted into Figure 4D according to different cutoffs of top DEG lists for peering methods. It can be seen that the percentage of overlapping genes tends to increase with the increasing cutoff of the top DEGs list, excepting the uncorrected data. Combat and SVA showed similar results, with a median range from 0.58 to 0.72 and 0.54 to 0.68 respectively. Interestingly, Rank-In maintained the top rate among peers at different cutoffs, with a median range from 0.74 to 0.83, demonstrating header-and-shoulder higher performance than peers. On top of that, the standard deviations of Rank-In were kept the lowest, indicating the robustness of this method.

Online platform

To promote the community application, Rank-In was established as an online platform at http://www.badd-cao.net/rank-in/. Rank-In requires two files as input, the gene expression profiles of all consolidated samples and the class label of each sample (Figure 5A). The expression matrix should be uploaded as gene expression intensities in the log-transformed format of raw value for microarray or FPKM/TPM/TMM for RNA-seq. The class label file needs to contain the same sample descriptors as the data file, as well as the class identifies. Additional information about the experimental design is encouraged to upload for better results, such as platforms/batches and cancer types. Three files will be given by Rank-In, including adjusted matrix, DEG list if two classes are labeled, and an unsupervised clustering tree (Figure 5B). More instructions can be found on the online help page.

DISCUSSIONS

The rapid upgrading of RNA profiling technologies makes the consolidated analysis of cancer transcriptomics a demanding task due to the inherent difference of platform design and data distribution. Existing computational tools are recommended to compare within the same type of transcriptomic data. Here, Rank-In was designed enabling direct integrative analysis for mixed data of array and RNA-seq. To achieve the above, Rank-In transformed the raw expression intensity of overlapping genes between consolidated samples into relative ranking within each gene expression profile, and further partitioned into gene groups to avoid oversensitivity, as the relative levels of gene expression were found to be roughly comparable across Array and RNA-seq technologies, but never the absolute expression intensity (7). Considering that purely internal ranking may lead to evenly distributed genes, the ranks were further weighted by the intensity increasing slope within gene groups. Be noted that the slope may display smoothly in the middle groups, but become extremely sharp at both top and bottom ends, particularly to RNA-Seq data (38,39), as Figure 2B showed. As such, the ranking of middle-group genes may be adjusted into lower value, while those in extreme top or bottom remain similar after weighted processing. At last, the technology, platform and batch variances can be further reduced by matrix decomposition through adjusting those volatile genes across different samples. So far, the evenly distributed internal ranking matrix has been normalized into an un-even one from 0 to 100, to enable cross-sample analysis.

Among the peers, Combat and SVA are both great methods being widely adopted within microarray or RNA-seq

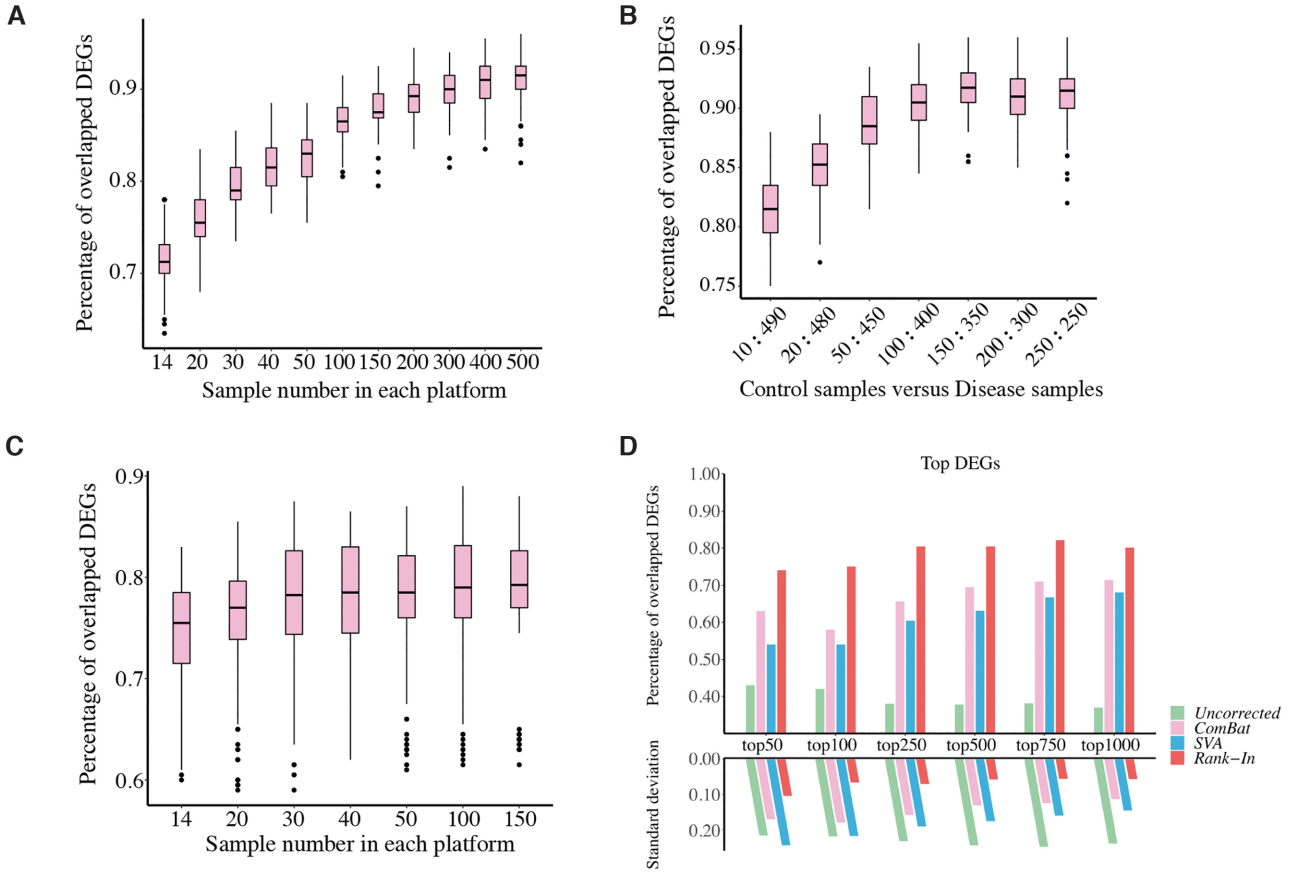
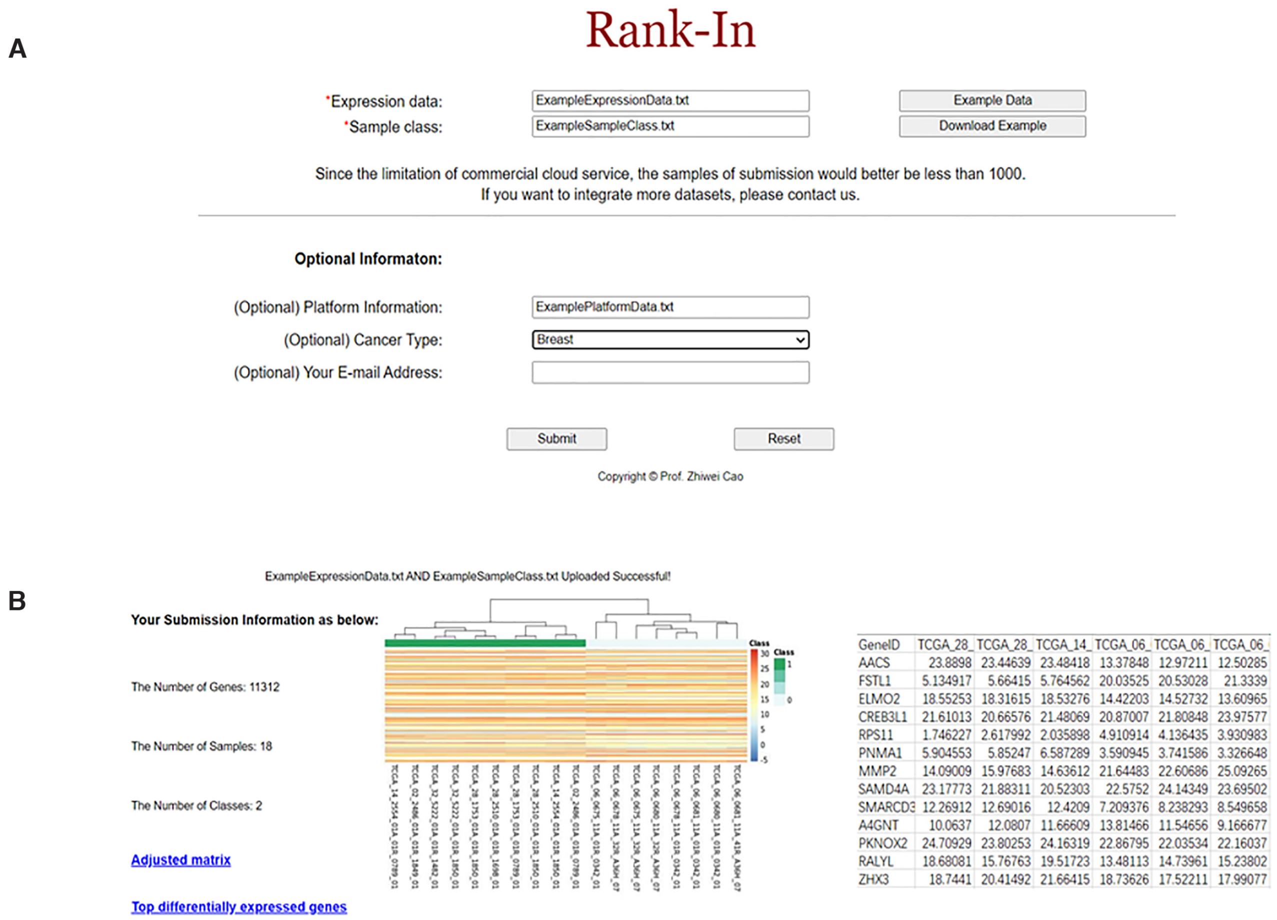


Figure 4. The overlapping rate of top DEGs between microarray and RNA-seq. Angel's method is not fit to derive DEGs, and thus not shown. (**A**) The boxplot of overlapped DEGs on different simulated sample sizes. (**B**) The boxplot of overlapped DEGs on different imbalanced ratios. (**C**) The boxplot of overlapped DEGs on different GBM sample sizes. (**D**) The overlapped top DEGs between microarray and RNA-seq on GBM.

separately. Until this year of 2020, Angels' method was released claiming for cross-platform transcriptomics analysis between array and RNA-seq for blood samples. Instead of all overlapping genes, it can auto-screen and select a subset of genes with low platform variance for subsequent cell-type clustering. Though being designed for blood samples, Angel's method obtained similar performance with Combat and SVA in our tests on SEQC cell samples. Yet on clinical samples of both GBM and colon cancer, its performance shows fluctuation, with general achievement better than Combat, but worse than SVA. This may be related to the inherent limitation of the 'marker genes' strategy, where the personal variation in health tissues might be not enough to become the 'prominent features' in data collection (21). While in Rank-In, both relative and absolute intensity of gene expression are considered for all overlapping genes between datasets. This, coupled with SVD to further reduce nonbiological bias, may have rendered Rank-In to surpass the other peers.

Being validated on different scenarios, Rank-In showed unique robustness not only on tumor and health controls but also on balanced and imbalanced datasets, even the size of normal control is relatively small. Thus Rank-In could provide practical means to revitalize and reutilize those obsolete datasets for a new analysis of refreshing results, particularly to those precious tissue samples difficult to obtain, such as brain or thyroid tissues. In this regard, Rank-In may help to save the number of samples that need to be collected in clinical research. In terms of application boundary, Rank-In is suggested to be used with caution for those array profiles with probes targeting only a minor portion of whole genomes. As RNA-seq is designed to detect all genes expressed, a partial and biased overlapping list may cause a global shift in the expression ranking curve, which has broken the requirement of Rank-In that the overall gene expression levels are comparable from a view of the whole genomic portfolio. In the future, this method would be extended to multiple tumor subtypes, and to combine public data for cancer transcriptome atlas across multiple platforms and studies.

To summarize, we developed a method enabling integrative transcriptomics analysis of mixed data across microarray and RNA-seq for cancer, tolerating small/large, paired/unpaired and balanced/imbalanced samples. With continuous updating, Rank-in would be particularly useful to analyzing blended data across different transcriptome technologies, platforms or batches, as well as consolidating limited cancer samples for large-scale bioinformatics analysis.



Rank-In

*Expression data: ExampleExpressionData.txt

*Sample class: ExampleSampleClass.txt

Example Data

Download Example

Since the limitation of commercial cloud service, the samples of submission would better be less than 1000.
If you want to integrate more datasets, please contact us.

Optional Informaton:

(Optional) Platform Information: ExamplePlatformData.txt

(Optional) Cancer Type: Breast

(Optional) Your E-mail Address:

Submit Reset

Copyright © Prof. Zhiwei Cao

ExampleExpressionData.txt AND ExampleSampleClass.txt Uploaded Successful!

Your Submission Information as below:

The Number of Genes: 11312

The Number of Samples: 18

The Number of Classes: 2

Adjusted matrix

Top differentially expressed genes

Class
30
25
20
15
10
5
0
-5
Class
1
0
TCGA_14_2554_01A_01R_0789_01
TCGA_02_2486_01A_01R_1849_01
TCGA_32_5222_01A_01R_1482_01
TCGA_32_5222_01A_01R_1850_01
TCGA_28_1753_01A_01R_1850_01
TCGA_28_2510_01A_01R_1698_01
TCGA_28_1753_01A_01R_0789_01
TCGA_28_2510_01A_01R_1850_01
TCGA_14_2554_01A_01R_1850_01
TCGA_02_2486_01A_01R_0789_01
TCGA_06_0675_11A_01R_0342_01
TCGA_06_0678_11A_32R_A36H_07
TCGA_06_0675_11A_32R_A36H_07
TCGA_06_0680_11A_32R_A36H_07
TCGA_06_0678_11A_01R_0342_01
TCGA_06_0681_11A_01R_0342_01
TCGA_06_0680_11A_01R_0342_01
TCGA_06_0681_11A_41R_A36H_07

GeneID	TCGA_28_	TCGA_28_	TCGA_14_	TCGA_06_	TCGA_06_	TCGA_06_
AACS	23.8898	23.44639	23.48418	13.37848	12.97211	12.50285
FSTL1	5.134917	5.66415	5.764562	20.03525	20.53028	21.3339
ELMO2	18.55253	18.31615	18.53276	14.42203	14.52732	13.60965
CREB3L1	21.61013	20.66576	21.48069	20.87007	21.80848	23.97577
RPS11	1.746227	2.617992	2.035898	4.910914	4.136435	3.930983
PNMA1	5.904553	5.85247	6.587289	3.590945	3.741586	3.326648
MMP2	14.09009	15.97683	14.63612	21.64483	22.60686	25.09265
SAMD4A	23.17773	21.88311	20.52303	22.5752	24.14349	23.69502
SMARCD3	12.26912	12.69016	12.4209	7.209376	8.238293	8.549658
A4GNT	10.0637	12.0807	11.66609	13.81466	11.54656	9.166677
PKNOX2	24.70929	23.80253	24.16319	22.86795	22.03534	22.16037
RALYL	18.68081	15.76763	19.51723	13.48113	14.73961	15.23802
ZHX3	18.7441	20.41492	21.66415	18.73626	17.52211	17.99077

Figure 5. The online platform of Rank-In. (**A**) The interface of uploading consolidated data. (**B**) The interface of downloading the results such as adjusted matrix and sorted DEGs.

DATA AVAILABILITY

The data used in this study are listed in the supplementary tables.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Prof. Leming Shi and Dr. Ying Yu for valuable discussions of SEQC data.
Author Contributions: K.L.T. performed analysis, and wrote the manuscript; M.D.Z. and Z.L.D. performed the analysis; X.J.J. and Y.W.H. collected samples and performed the analysis; G.H.Z constructed the web server; Z.W.C edited the manuscript and supervised the study.

FUNDING

National Key R&D Program of China [2017YFC1700200, 2019YFA0905900, 2017YFC0908405]; National Natural Science Foundation of China [32070657, 81830080]. Funding for open access charge: National Key R&D Program of China [2017YFC1700200].

Conflict of interest statement. None declared.

REFERENCES

1. Stark,R., Grzelak,M. and Hadfield,J. (2019) RNA sequencing: the teenage years. *Nat. Rev. Genet.*, **20**, 631–656.
2. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.*, **41**, D991–D995.
3. Hoyle,D.C., Rattray,M., Jupp,R. and Brass,A. (2002) Making sense of microarray data distributions. *Bioinformatics*, **18**, 576–584.
4. Shahjaman,M., Manir Hossain Mollah,M., Rezanur Rahman,M., Islam,S.M.S. and Nurul Haque Mollah,M. (2020) Robust identification of differentially expressed genes from RNA-seq data. *Genomics*, **112**, 2000–2010.
5. Bradford,J.R., Hey,Y., Yates,T., Li,Y., Pepper,S.D. and Miller,C.J. (2010) A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC Genomics*, **11**, 282.
6. Wen,Z.N., Su,Z.Q., Liu,J., Ning,B.T., Guo,L., Tong,W.D. and Shi,L.M. (2011) The MicroArray Quality Control (MAQC) project and cross-platform analysis of microarray data. In: Lu,H.H-.S., Schölkopf,B. and Zhao,H. (eds). *Handbook of Statistical Bioinformatics*, Springer, Berlin, Heidelberg, pp. 171–192.
7. Xu,J.S., Gong,B.S., Wu,L.H., Thakkar,S., Hong,H.X. and Tong,W.D. (2016) Comprehensive assessments of RNA-seq by the SEQC consortium: FDA-Led efforts advance precision medicine. *Pharmaceutics*, **8**, 8.

8. Wang,H., He,X., Band,M., Wilson,C. and Liu,L. (2005) A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics*, **6**, 71.
9. Su,Z.Q., Labaj,P.P., Li,S., Thierry-Mieg,J., Thierry-Mieg,D., Shi,W., Wang,C., Schroth,G.P., Setterquist,R.A., Thompson,J.F. *et al.* (2014) A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.*, **32**, 903–914.
10. Su,Z., Fang,H., Hong,H., Shi,L., Zhang,W., Zhang,W., Zhang,Y., Dong,Z., Lancashire,L.J., Bessarabova,M. *et al.* (2014) An investigation of biomarkers derived from legacy microarray data for their utility in the RNA-seq era. *Genome Biol.*, **15**, 523.
11. Lazar,C., Meganck,S., Taminau,J., Steenhoff,D., Coletta,A., Molter,C., Weiss-Solis,D.Y., Duque,R., Bersini,H. and Nowe,A. (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief. Bioinform.*, **14**, 469–490.
12. Johnson,W.E., Li,C. and Rabinovic,A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.
13. Zhang,Y., Parmigiani,G. and Johnson,W.E. (2020) ComBat-seq: batch effect adjustment for RNA-seq count data. *NAR Genom. Bioinform.*, **2**, lqaa078.
14. Leek,J.T. and Storey,J.D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLos Genet.*, **3**, 1724–1735.
15. Leek,J.T., Johnson,W.E., Parker,H.S., Jaffe,A.E. and Storey,J.D. (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.
16. Leek,J.T. (2014) svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, **42**, e161.
17. Chen,W., Zhang,S., Williams,J., Ju,B., Shaner,B., Easton,J., Wu,G. and Chen,X. (2020) A comparison of methods accounting for batch effects in differential expression analysis of UMI count based single cell RNA sequencing. *Comput. Struct. Biotechnol. J.*, **18**, 861–873.
18. Li,S., Labaj,P.P., Zumbo,P., Sykacek,P., Shi,W., Shi,L., Phan,J., Wu,P.Y., Wang,M., Wang,C. *et al.* (2014) Detecting and correcting systematic variation in large-scale RNA sequencing data. *Nat. Biotechnol.*, **32**, 888–895.
19. Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.
20. Risso,D., Ngai,J., Speed,T.P. and Dudoit,S. (2014) Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.*, **32**, 896–902.
21. Angel,P.W., Rajab,N., Deng,Y., Pacheco,C.M., Chen,T., Le Cao,K.A., Choi,J. and Wells,C.A. (2020) A simple, scalable approach to building a cross-platform transcriptome atlas. *PLoS Comput. Biol.*, **16**, e1008219.
22. Jaffe,A.E., Hyde,T., Kleinman,J., Weinbergern,D.R., Chenoweth,J.G., McKay,R.D., Leek,J.T. and Colantuoni,C. (2015) Practical impacts of genomic data “cleaning” on biological discovery using surrogate variable analysis. *BMC Bioinform.*, **16**, 372.
23. Tang,Z.Q., Han,L.Y., Lin,H.H., Cui,J., Jia,J., Low,B.C., Li,B.W. and Chen,Y.Z. (2007) Derivation of stable microarray cancer-differentiating signatures using consensus scoring of multiple random sampling and gene-ranking consistency evaluation. *Cancer Res.*, **67**, 9996–10003.
24. Xu,L., Luo,H., Wang,R., Wu,W.W., Phue,J.N., Shen,R.F., Juhl,H., Wu,L., Alterovitz,W.L., Simonyan,V. *et al.* (2019) Novel reference genes in colorectal cancer identify a distinct subset of high stage tumors and their associated histologically normal colonic tissues. *BMC Med. Genet.*, **20**, 138.
25. Caracausi,M., Piovesan,A., Antonaros,F., Strippoli,P., Vitale,L. and Pelleri,M.C. (2017) Systematic identification of human housekeeping genes possibly useful as references in gene expression studies. *Mol. Med. Rep.*, **16**, 2397–2410.
26. Thompson,J.A., Tan,J. and Greene,C.S. (2016) Cross-platform normalization of microarray and RNA-seq data for machine learning applications. *PeerJ.*, **4**, e1621.
27. Franks,J.M., Cai,G. and Whitfield,M.L. (2018) Feature specific quantile normalization enables cross-platform classification of molecular subtypes using gene expression data. *Bioinformatics*, **34**, 1868–1874.
28. Wang,P., Yang,Y., Han,W. and Ma,D. (2015) ImmuSort, a database on gene plasticity and electronic sorting for immune cells. *Sci. Rep.*, **5**, 10370.
29. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
30. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
31. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
32. Canales,R.D., Luo,Y., Willey,J.C., Austermiller,B., Barbacioru,C.C., Boysen,C., Hunkapiller,K., Jensen,R.V., Knight,C.R., Lee,K.Y. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.
33. Chawla,N.V. (2010) In: Maimon,O. and Rokach,L. (eds). *Data Mining and Knowledge Discovery Handbook*. Springer US, Boston, MA, pp. 875–886.
34. Zhu,Y., Qiu,P. and Ji,Y. (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat. Methods*, **11**, 599–600.
35. Goh,W.W.B., Wang,W. and Wong,L. (2017) Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends Biotechnol.*, **35**, 498–507.
36. Cordero,D., Sole,X., Crous-Bou,M., Sanz-Pamplona,R., Pare-Brunet,L., Guino,E., Olivares,D., Berenguer,A., Santos,C., Salazar,R. *et al.* (2014) Large differences in global transcriptional regulatory programs of normal and tumor colon cells. *BMC Cancer*, **14**, 708.
37. Dembele,D. (2013) A flexible microarray data simulation model. *Microarrays (Basel)*, **2**, 115–130.
38. Wang,C., Gong,B.S., Bushel,P.R., Thierry-Mieg,J., Thierry-Mieg,D., Xu,J.S., Fang,H., Hong,H.X., Shen,J., Su,Z.Q. *et al.* (2014) The concordance between RNA-seq and microarray data depends on chemical treatment and transcript abundance. *Nat. Biotechnol.*, **32**, 926–932.
39. Zhao,S., Fung-Leung,W.P., Bittner,A., Ngo,K. and Liu,X. (2014) Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLoS One*, **9**, e78644.