Bioinformatics, 36(15), 2020, 4283–4290
doi: 10.1093/bioinformatics/btaa523
Advance Access Publication Date: 19 May 2020
Original Paper

OXFORD

Structural bioinformatics

Identification of population-level differentially expressed genes in one-phenotype data

Jiajing Xie[1,2,†], Yang Xu[1,2,†], Haifeng Chen[3,†], Meirong Chi[1,2], Jun He[1,2], Meifeng Li[1,2], Hui Liu[1,2], Jie Xia[1,2], Qingzhou Guan[1,2], Zheng Guo[1,2,*] and Haidan Yan[1,2,*]

[1]Department of Bioinformatics, Key Laboratory of Ministry of Education for Gastrointestinal Cancer, The School of Basic Medical Sciences, Fujian Medical University, Fuzhou 350122, China, [2]Key Laboratory of Medical Bioinformatics, Fujian Province, Fuzhou 350122, China and [3]Department of General Surgery, Fuzhou Second Hospital Affiliated to Xiamen University, Fuzhou 350007, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first three authors should be regarded as Joint First Authors.
Associate Editor: Arne Elofsson

Received on January 2, 2020; revised on April 15, 2020; editorial decision on May 11, 2020; accepted on May 14, 2020

Abstract

Motivation: For some specific tissues, such as the heart and brain, normal controls are difficult to obtain. Thus, studies with only a particular type of disease samples (one phenotype) cannot be analyzed using common methods, such as significance analysis of microarrays, edgeR and limma. The RankComp algorithm, which was mainly developed to identify individual-level differentially expressed genes (DEGs), can be applied to identify population-level DEGs for the one-phenotype data but cannot identify the dysregulation directions of DEGs.

Results: Here, we optimized the RankComp algorithm, termed PhenoComp. Compared with RankComp, PhenoComp provided the dysregulation directions of DEGs and had more robust detection power in both simulated and real one-phenotype data. Moreover, using the DEGs detected by common methods as the 'gold standard', the results showed that the DEGs detected by PhenoComp using only one-phenotype data were comparable to those identified by common methods using case-control samples, independent of the measurement platform. PhenoComp also exhibited good performance for weakly differential expression signal data.

Availability and implementation: The PhenoComp algorithm is available on the web at https://github.com/XJJ-student/PhenoComp.

Contact: Joyan168@126.com or guoz@ems.hrbmu.edu.cn

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

1 Introduction

Significance analysis of microarrays (SAM) (Tusher *et al.*, 2001), edgeR (Robinson *et al.*, 2010) and limma (Law *et al.*, 2014; Ritchie *et al.*, 2015), are commonly used methods for identifying differentially expressed genes (DEGs) between disease samples and normal controls. However, some specific normal controls, such as the heart and brain, are difficult to obtain for the corresponding disease research. For example, healthy cardiac tissue is usually sampled from donors who have died of chronic diseases or accidents (Molina-Navarro *et al.*, 2013). Thus, some heart disease studies have very few or even no normal control (Ein-Dor *et al.*, 2006). When a study has no normal control, common methods needing case–control samples cannot be used. Moreover, due to the existence of batch effects (Leek *et al.*, 2010; Rung and Brazma, 2013), the normal controls from different datasets cannot be directly used to identify DEGs for datasets without normal controls (one-phenotype data). Although there are some batch effect adjustment methods, such as distance-weighted discrimination (Benito *et al.*, 2004), ComBat (Johnson *et al.*, 2007), SVA (Leek *et al.*, 2012), RUV (Peixoto *et al.*, 2015) and svaseq (Leek, 2014), these methods may distort the true biological signals (Loven *et al.*, 2012; Wang *et al.*, 2012) and even produce spurious group differences, especially when two groups are unevenly distributed across different batches (Cai *et al.*, 2018; Lazar *et al.*, 2013; Nygaard *et al.*, 2016).

Previously, based on the finding that the within-sample relative expression orderings (REOs) of genes in a particular type of normal tissues are highly stable but are widely reversed in the corresponding cancer tissues, we developed a method, termed RankComp, to identify DEGs for each disease sample (Wang *et al.*, 2015). Notably, due to the normal background, stable normal REOs in a particular type of normal tissue are predetermined in accumulated normal controls

from different sources, RankComp can be used to identify individual-level DEGs for datasets without normal controls (or one-phenotype data) through finding genes whose dysregulation may lead to reversed REOs within the disease sample compared with stable normal REOs (Wang *et al.*, 2015). After identifying individual-level DEGs in a group of one-phenotype samples, RankComp can detect population-level DEGs using the binomial test to identify a nonrandom high percentage of samples sharing certain DEGs (Wang *et al.*, 2015). Because the REOs of genes within a sample are robust against batch effects and insensitive to data normalization (Geman *et al.*, 2004; Tan *et al.*, 2005), the REO-based method, RankComp, can directly use data from different sources (Wang *et al.*, 2015). However, the dysregulation directions of DEGs are not provided by the RankComp algorithm, which is important for differential expression analysis.

In this study, we optimized the RankComp algorithm to improve the detection power of DEGs and provide the dysregulation directions of DEGs. The optimized algorithm was named PhenoComp. For a particular type of disease, we clearly pointed out that the probabilities of a gene being upregulated and downregulated should be estimated separately in each of the disease datasets. The performance of PhenoComp was evaluated by concordant analysis of DEGs identified by it and the common methods in independent datasets for ischemic cardiomyopathy (ICM) and glioma. We also compared the detection power of PhenoComp with that of RankComp in simulated and real one-phenotype data and evaluated whether PhenoComp could be applied to identify DEGs between drug response and non-response groups for estrogen receptor (ER)-negative breast cancer patients with neoadjuvant chemotherapy response information.

Table 1. Datasets used in this study

Tissues	Datasets	Platform	Control[a]	Case[b]
Heart	GSE42955	GPL6244	5	12
	GSE22253	GPL6244	108	—
	GSE57338	GPL11532	136	95
	GSE141910	RNA-seq	166	—
	GSE116250	RNA-seq	14	13
	GSE26887[c]	GPL6244	5	12
	GSE46224[c]	RNA-seq	8	8
Brain	GSE35493	GPL570	9	—
	GSE94349	GPL570	27	10
	GSE68015	GPL570	16	11
	GSE86574	GPL570	10	15
	GSE26966	GPL570	9	—
	GTEx	RNA-seq	406	—
	GSE50161[c]	GPL570	13	34
	TCGA[c]	RNA-seq	5	156
Breast	GSE20194	GPL96	27	37
	GSE23988	GPL96	13	16
	GSE20271	GPL96	8	26

[a]Tissues sampled from normal heart, normal brain or pCR ER-negative breast cancer.

[b]Tissues sampled from ICM, glioma or RD ER-negative breast cancer.

[c]Datasets were independent datasets used to evaluate the performance of the algorithms.

2 Materials and methods

2.1 Data and preprocessing

As shown in Table 1, seven datasets for heart tissues, eight datasets for brain tissues and three datasets for ER-negative breast cancer tissues were downloaded from Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) (Barrett *et al.*, 2013), The Cancer Genome Atlas (TCGA) data portal (http://tcga-data.nci.nih.gov/tcga/) (International Cancer Genome Consortium *et al.*, 2010) and the Genotype-Tissue Expression data portal (https://www.gtexportal.org/home/index.html) (GTEx Consortium *et al.*, 2017). Normal and ICM tissues were collected for heart tissues, and normal and glioma tissues were collected for brain tissues. For breast tissues, ER-negative breast cancer patients with neoadjuvant chemotherapy response information were used in this study. The neoadjuvant chemotherapy included paclitaxel, 5-fluorouracil, cyclophosphamide and doxorubicin. Responses to preoperative chemotherapy were categorized as pathological complete response (pCR) or residual invasive disease (RD).

Here, gene expression profiles measured by the Affymetrix microarray platform were processed by the Robust Multi-array Average algorithm for background adjustment without quantile normalization (Irizarry *et al.*, 2003). For the Illumina microarray platform, the processed data were directly downloaded from GEO for subsequent analysis. For the RNA-sequencing (RNA-seq) platform, the raw sequencing data were downloaded and processed by Trimmomatic (Bolger *et al.*, 2014). We aligned the reads to the reference genome (GRCh37) using hisat2 (Kim *et al.*, 2015). Both count and reads per kilobase million (RPKM) values were calculated for each gene using StringTie (Pertea *et al.*, 2015). We used the RPKM values of genes in the REO-based methods, such as RankComp and PhenoComp, because the RPKM values could represent the actual expression abundance of genes by normalizing both the gene length and sequencing depth (Cai *et al.*, 2018; Mortazavi *et al.*, 2008; Trapnell *et al.*, 2010; Wagner *et al.*, 2012). The count values were used for edgeR (Robinson *et al.*, 2010) and limma (Law *et al.*, 2014).

2.2 The RankComp and PhenoComp algorithms

For a particular type of disease, our previously developed RankComp was mainly used to identify DEGs for each of the disease samples (Wang *et al.*, 2015). Using RankComp, we first identified highly stable REOs as the normal background (Wang *et al.*, 2015). For each gene pair, g_j and g_k ($j = 1.m$, $k = 1.m$, and $j \neq k$), and let e_j and e_k represented the expression level of g_j and g_k, respectively, and m represented the total number of genes. Then, the REO of g_j and g_k in a sample was $e_j > e_k$ or $e_j < e_k$. For RankComp, an REO, keeping at least 99% of previously accumulated normal samples, was defined as highly stable. Then, for each disease sample, if a highly stable REO $e_j > e_k$ or $e_j < e_k$ in normal samples was reversed ($e_j < e_k$ or $e_j > e_k$) in the disease sample, the REO was defined as reversal. Then, for a disease sample, through testing the null hypothesis that the proportion of reversal REOs supporting the upregulation of a gene was equal to the proportion of reversal REOs supporting the downregulation of the gene, the Fisher's exact test was used to identify whether the gene was differentially expressed in the disease sample (Wang *et al.*, 2015). After rejecting the null hypothesis, the gene in the disease sample was defined as upregulated if the number of reversal REOs supporting the upregulation of the gene was larger than the number of reversal REOs supporting the downregulation of the gene; conversely, the gene was defined as downregulated (Wang *et al.*, 2015).

Here, we used a dataset with only n ICM samples as an example to illustrate how the RankComp and PhenoComp algorithms identify population-level DEGs for one-phenotype data. As shown in Figure 1, after identifying individual-level DEGs for each of the n ICM samples using RankComp, we showed the differential expression statuses of all m genes in each of the n samples. The differential expression statuses included upregulation, downregulation, and non-differential expression (non-DE). Then, RankComp transformed the upregulation and downregulation statuses of m genes in n samples to the DE statuses, ignoring the dysregulation directions. Finally, the binomial test was used to identify DEGs shared by a nonrandom high percentage of ICM samples, namely, population-level DEGs for the dataset. For each gene, the null hypothesis was that $p = p_{de}$ and the alternative hypothesis was $p > p_{de}$. The P value to determine whether a gene was population-level differentially expressed was calculated as follows:

$$p = \sum_{i=k}^{n} \binom{n}{i} p_{de}^{i} (1 - p_{de})^{n-i}, \qquad (1)$$

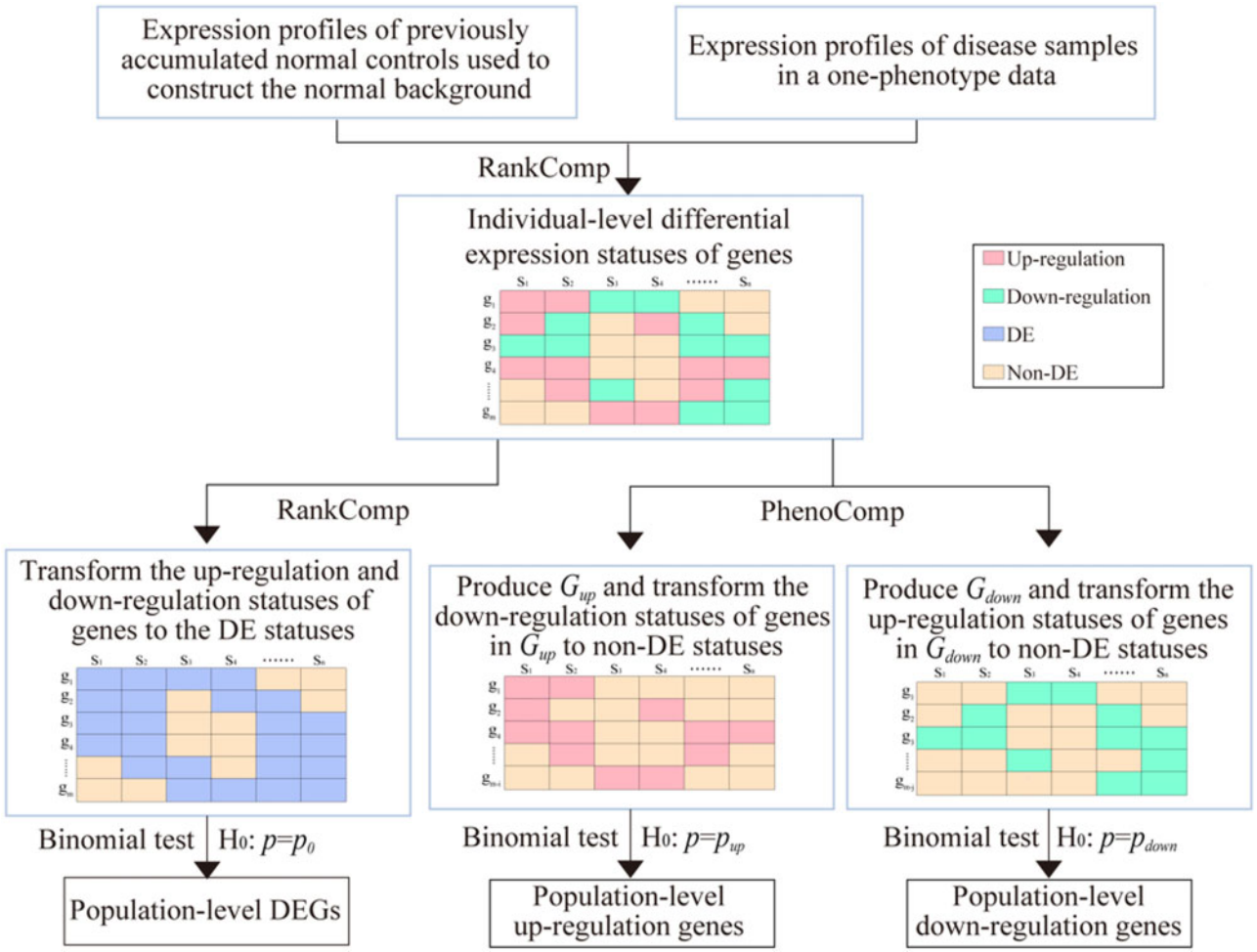

Fig. 1. RankComp and PhenoComp algorithms to identify population-level DEGs. For a dataset with only disease samples, we first identified the individual-level DEGs for a particular type of disease sample using the stable REOs determined by previously accumulated normal samples as normal background. Then, based on the individual-level differential expression statuses of genes, RankComp and PhenoComp developed different methods to infer population-level DEGs

where n and k are the total number of ICM samples in the dataset and the number of ICM samples in the dataset with the gene being DE status, respectively. The p_{de} is the probability of observing a gene being differentially expressed in an ICM sample by chance. For RankComp, the p_{de} was calculated as the average of the frequencies of DEGs among the m genes for individual ICM samples (Wang *et al.*, 2015).

For differential expression analysis, the dysregulation directions of DEGs were important information, but this key point was not considered by RankComp. Here, we optimized the RankComp algorithm to improve the detection power and provide the dysregulation directions of DEGs. The optimized algorithm was named PhenoComp. As shown in Figure 1, based on the individual-level differential expression statuses (upregulation, downregulation and non-DE) of m genes in n ICM samples identified by RankComp, PhenoComp respectively produced two sets of genes, denoted G_{up} and G_{down}. A gene that was upregulated in at least one sample was in the set of G_{up}. Similarly, a gene that was downregulated in at least one sample was in G_{down}. Then a gene that was upregulated in parts of n samples and downregulated in some of the other samples were in both G_{up} and G_{down}. When the gene was in G_{up}, PhenoComp transformed the downregulation status of the gene to non-DE status. The upregulation status of the gene was transformed to the status of non-DE when the gene was in G_{down}. After transformation, the genes in G_{up} only had two statuses of upregulation and non-DE, and the genes in G_{down} only had two statuses of downregulation and non-DE. Then, using the binomial test, PhenoComp identified the population-level upregulation and downregulation genes in G_{up} and G_{down} (Fig. 1). For each gene in G_{up} (G_{down}), the null hypothesis was that $p = p_{up}$ ($p = p_{down}$), and the alternative hypothesis was $p > p_{up}$ ($p > p_{down}$). The probability of a gene being upregulated, p_{up}, was calculated as follows:

$$p_{up} = p_{de} \times p_{up|de}. \qquad (2)$$

The p_{de} was estimated as the median or average of the frequencies of DEGs among m genes for individual ICM samples. The $p_{up|de}$ was estimated as the median or average of the frequencies of upregulated genes among DEGs for each of the n ICM samples. The probability of a gene being downregulated, p_{down}, was estimated in the same way. The $p_{down|de}$ was estimated as the median or average of the frequencies of downregulated genes among DEGs for each of the n ICM samples. The Benjamini–Hochberg procedure was used to control the false discovery rate (FDR) in the multiple tests. Notably, when a gene was identified as both upregulated and downregulated, the gene was deleted.

2.3 Performance evaluation

Here, two independent datasets for ICM (GSE46224 measured by RNA-seq and GSE26887 measured by microarray) and two independent datasets for glioma (TCGA measured by RNA-seq and GSE50161 measured by microarray) were used to evaluate the performance of PhenoComp (Table 1). Notably, to illustrate the power of PhenoComp and RankComp in detecting population-level DEGs

for one-phenotype data, the normal tissues in the independent datasets were not used when we identified population-level DEGs using PhenoComp and RankComp. Here, a simulated experiment was first used to evaluate the performance of the two methods. The one-phenotype data, namely, the data with only the expression profiles of a particular type of disease sample, were produced based on the expression profiles of normal tissues in independent datasets (see Section 3 for a detailed description of simulation experiments). According to the simulated experiments, we determined the real DEGs and non-DEGs and then calculated the sensitivity, specificity, and F-score. The sensitivity (specificity) was calculated as the proportion of correctly identified DEGs (non-DEGs) to all DEGs (non-DEGs) and the *F*-score was calculated as follows:

$$F-\text{score}=\frac{2\ (\text{sensitivity}\times\ \text{specificity})}{(\text{sensitivity}+\ \text{specificity})}\quad(3)$$

However, we also used the DEGs identified by common methods as the 'gold standard' to evaluate the performance of PhenoComp. Then, we calculated the concordance score and the percentage of overlapping genes (POG) score (Cai *et al.*, 2018; Zhang *et al.*, 2008) to evaluate the consistency between DEGs identified by PhenoComp and those identified by common methods. The number of overlapping genes between the two lists of DEGs identified by two methods was denoted as *o*, and the number of genes that had the same dysregulation directions among the *o* genes was denoted as *s*. Then the concordance score was calculated as *s*/*o*. Here, the binomial distribution test was performed to evaluate whether the concordance score of *s*/*o* occurred by chance. For two lists of DEGs with length l_1 and l_2, the POG score from list 1 (or 2) to list 2 (or 1), denoted as POG_{12} (or POG_{21}), was calculated as s/l_1 (or s/l_2). Common methods including edgeR, limma and SAM were used to detect population-level DEGs using case–control samples in the independent datasets. EdgeR and limma were used for datasets measured by RNA-seq. Limma and SAM were used for datasets measured by microarray. However, because the DEGs were only significantly enriched in pathways that contained a sufficient number of real DEGs, we also evaluated the performance of PhenoComp by performing enrichment analysis for DEGs detected by it.

2.4 Functional enrichment analysis

We downloaded 244 non-disease pathways with 6934 genes from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2012) in April 2019. The pathways with at least 10 genes remained in this study (Jung *et al.*, 2019) and a metabolic pathway with 1266 genes was discarded. Then the remaining 230 pathways with genes ranging from 10 to 448 were used in this study. The hypergeometric distribution model was used to test whether a set of genes observed in a pathway was significantly more than that expected by chance.

3 Results

3.1 Estimation of the probability of a gene being upregulated and downregulated

First, RankComp was used to identify DEGs for each of the ICM samples. We used 429 normal heart samples, including 180 samples derived from GSE116250 and GSE141910 measured by RNA-seq and 249 samples derived from three datasets (GSE42955, GSE22253 and GSE57338) measured by microarray, to identify highly stable REOs as the normal background. A total of 66 539 760 and 101 187 275 stable REOs were identified for datasets measured by RNA-seq and microarray, respectively (Supplementary Table S1). Among the 35 000 826 overlapping stable REOs in RNA-seq and microarray, 99.36% (34 775 527) showed the same REO patterns (binomial test, $P<1.0E{-}16$). Then, based on these across-platform stable REOs, individual-level differential expression statuses of genes, including upregulation, downregulation and non-DE, were identified for 140 ICM samples from five datasets with ICM tissues using RankComp. Then, the frequencies of DEGs among all of the genes, and upregulated and downregulated genes among all of the DEGs, denoted as f_{de}, $f_{up|de}$ and $f_{down|de}$, were calculated in each ICM sample. The results showed that the distributions of the f_{de}, $f_{up|de}$ and $f_{down|de}$ values varied widely across different datasets, which suggests that the p_{de}, $p_{up|de}$ and $p_{down|de}$, should be separately estimated in each ICM dataset. For RankComp, the p_{de} value in ICM also needed to be estimated in each dataset when it was used to infer population-level DEGs. However, RankComp did not clearly illustrate that the p_{de} value should be estimated in per dataset, or only merged all analyzed datasets to calculate a mean value of f_{de} as the p_{de} value. On the other hand, RankComp only used the mean value of f_{de} as the p_{de} value but did not evaluate whether this method was reasonable. For PhenoComp, the median (or mean) values of f_{de}, $f_{up|de}$ and $f_{down|de}$ were 11.35% (14.91%), 59.65% (59.71%) and 40.35% (40.29%) for independent dataset GSE46224 and 12.55% (15.59%), 70.45% (68.30%) and 29.55% (31.70%) for independent dataset GSE26887 (Fig. 2A), which were used as the p_{de}, $p_{up|de}$ and $p_{down|de}$ values for the corresponding dataset.

For glioma, using 406 normal brain tissues measured by RNA-seq and 71 normal brain tissues measured by microarray, 37 305 493 across-platform stable REOs were identified as normal background (Supplementary Table S1). Then, according to the individual-level differential expression statuses of genes in each glioma sample identified by RankComp, the f_{de}, $f_{up|de}$ and $f_{down|de}$ values were calculated for each of the glioma samples. The result showed that the distributions of the f_{de}, $f_{up|de}$ and $f_{down|de}$ values were also different across different glioma datasets (Fig. 2B). Therefore, we also evaluated the p_{de}, p_{up} and p_{down} in each glioma dataset.

3.2 Evaluation of performance in simulated data

For the heart tissue, based on the expression profiles of eight normal samples from GSE46224 measured by RNA-seq and five normal samples from GSE26887 measured by microarray, we simulated eight and five disease samples by randomly selecting 3000 genes and changing their expression values in each of the normal samples with FC levels of 1.5, 1/1.5, 2, 1/2, 2.5, 1/2.5, 3 and 1/3. The expression values of the remaining genes in disease samples were the same as those of the corresponding normal samples. The simulation experiments were repeated 100 times for GSE46224 and GSE26887, respectively. Using the same simulation method, we also produced disease samples for the brain tissues based on the 5 and 13 normal samples from TCGA and GSE50161. As shown in Figure 3, we set the medians of f_{de}, $f_{up|de}$ and $f_{down|de}$ as p_{de}, $p_{up|de}$ and $p_{down|de}$ for each dataset, PhenoComp had higher sensitivity than RankComp, with a slight decrease in specificity in GSE46224 and GSE50161, whereas no DEGs were identified by RankComp in GSE26887 and TCGA with only five disease samples. Moreover, the DEGs identified by RankComp in each simulated experiment were all included in the DEGs identified by PhenoComp in each dataset (Supplementary Table S2). This result suggested that RankComp may have insufficient power to detect DEGs for the one-phenotype data with a small number of samples.

Here, we used the heart tissues as an example to further investigate the impact of the number of samples in one-phenotype data. We selected 5, 10, 20, 30 and 40 normal samples with the minimum top 5, 10, 20, 30 and 40 GSM numbers from GSE141910 measured by RNA-seq and GSE22253 measured by microarray, respectively. Using the above simulation method, we produced 5, 10, 20, 30 and 40 disease samples based on their corresponding normal samples from both GSE141910 and GSE22253. For PhenoComp, when the number of disease samples was increased from 5 to 40 in GSE141910, the sensitivity increased from 62.82% to 76.52% at the cost of a decrease in specificity, from 77.18% to 60.28%, while RankComp failed to identify any DEG when the number of disease samples was five (Fig. 3B). In GSE22253, PhenoComp had higher sensitivity than RankComp with a slight decrease in specificity across different sample sizes (Fig. 3C). In both GSE141910 and GSE22253, DEGs identified by PhenoComp included all DEGs identified by RankComp (Supplementary Table S2).

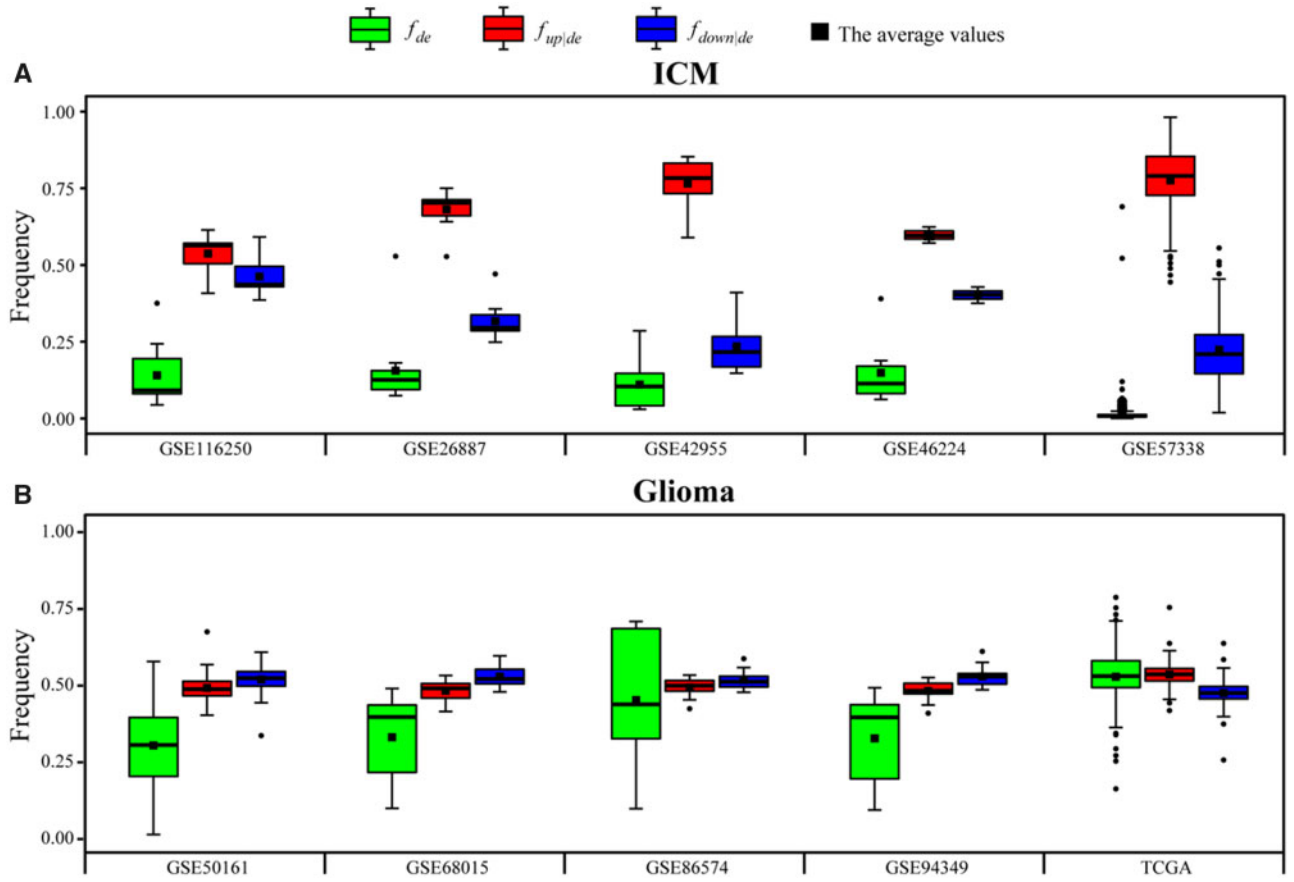

Fig. 2. Distributions of the f_{de}, $f_{up|de}$ and $f_{down|de}$ in different datasets for ICM (**A**) and glioma (**B**). The black squares on the box represent the average values of the f_{de}, $f_{up|de}$ and $f_{down|de}$ in each dataset

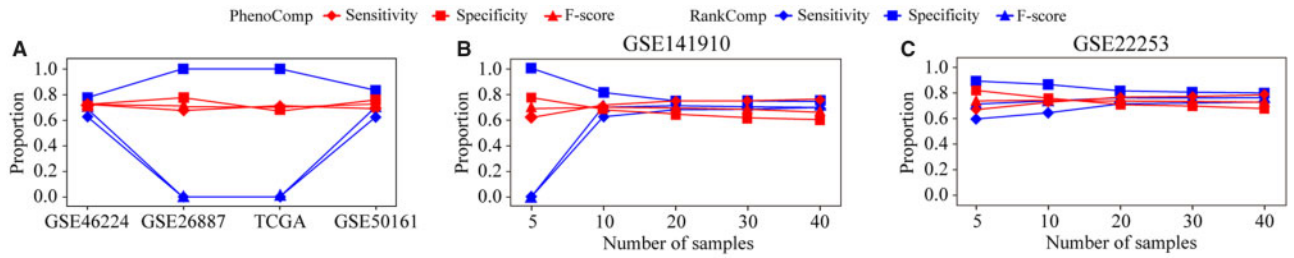

Fig. 3. Comparison of the performance of PhenoComp and RankComp in simulated data. The p_{up} and p_{down} were estimated by the mean values of the f_{de}, $f_{up|de}$ and $f_{down|de}$ for PhenoComp

We also used the mean values of f_{de}, $f_{up|de}$ and $f_{down|de}$ as p_{de}, $p_{up|de}$ and $p_{down|de}$ to infer the population-level DEGs for PhenoComp, and similar results were also observed for the simulated data (Supplementary Fig. S1 and Table S3). Overall, these results suggested that the detection power of both PhenoComp and RankComp were enhanced when more disease samples in the one-phenotype data were used to identify DEGs, and PhenoComp had more robust power than RankComp, especially in small data.

3.3 Evaluation of performance in real data

Using the median values of f_{de}, $f_{up|de}$ and $f_{down|de}$ as p_{de}, $p_{up|de}$ and $p_{down|de}$ in GSE46224 measured by RNA-seq, 1385 population-level DEGs were identified by PhenoComp for ICM (FDR < 5%). With FDR < 5%, 1677 and 1182 DEGs were identified by edgeR and limma. In comparison with edgeR and limma, although the POG_{12} and POG_{21} were about 10%, the concordance scores of DEGs reached 96%, which would be improved with a stricter FDR (Fig. 4A). Moreover, the DEGs identified by PhenoComp, edgeR and limma were separately enriched in seven, three and two pathways with FDR < 5% (Fig. 4B and Supplementary Table S4). Because the DEGs could be significantly enriched in pathways only when it contained a sufficient number of real DEGs, the results showed that the DEGs for PhenoComp were comparable to those identified by edgeR and limma. Similarly, 1963, 2519 and 4448 population-level DEGs were identified using PhenoComp, limma, and SAM, respectively, in GSE26887 measured by microarray (FDR < 5%). Compared with limma and SAM, the concordance scores were 97.38% and 92.76%, and the POG_{12} (POG_{21}) were 17.07% (13.30%) and 25.47% (11.24%) (Fig. 4A). There were 11, 10 and 23 pathways that were enriched with DEGs identified by PhenoComp, limma and SAM with FDR <5% (Fig. 4B and Supplementary Table S4).

The dysregulation directions of DEGs were essential for differential expression analysis, but RankComp did not provide a method to evaluate the dysregulation directions for the identified population-

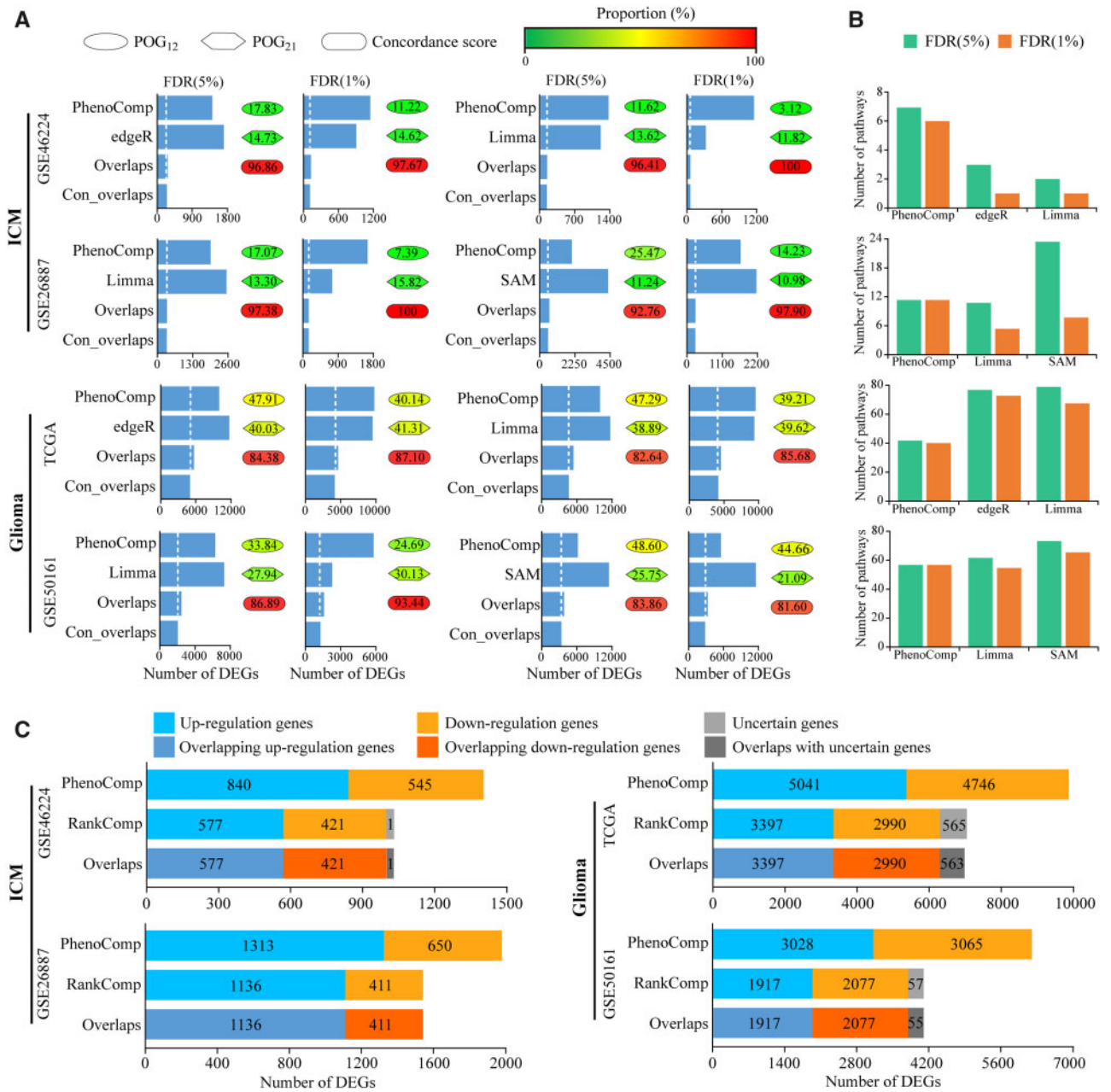

Fig. 4. Comparison of DEGs identified by different methods. (**A**) Concordance analysis of DEGs identified by two different methods. The p_{up} and p_{down} were estimated by the median values of the f_{de}, $f_{up|de}$ and $f_{down|de}$ for PhenoComp. Overlaps represent the DEGs identified by both PhenoComp and a common method (edgeR, limma or SAM). The con_overlaps denotes the overlaps that have the same dysregulated directions in two lists of DEGs. POG_{12} represents the proportion of consistent overlaps among the DEGs identified by PhenoComp. POG_{21} represents the proportion of consistent overlaps among the DEGs identified by the common method. (**B**) Number of pathways enriched with DEGs identified by different methods. The DEGs identified with FDR <5% and FDR <1% were all analyzed for pathway enrichment analysis. (**C**) Comparison of DEGs identified by PhenoComp and RankComp. Uncertain genes were those DEGs identified by RankComp with uncertain directions. Overlap with uncertain genes denotes that genes with uncertain directions identified by RankComp could be detected with clear dysregulation directions by PhenoComp

level DEGs. To compare the performance of PhenoComp and RankComp, we defined a population-level DEG in a dataset identified by RankComp as an upregulated (downregulation) gene if at least one disease sample within the dataset was upregulated (downregulated) in the gene and no disease sample within the dataset was downregulated (upregulated) in the gene. Then, the population-level DEGs, being upregulated in parts of samples and downregulated in some other parts of samples, were defined as genes with uncertain directions. Using RankComp, 999 and 1547 population-level DEGs were identified in GSE46224 and GSE26887, respectively, and almost all of the DEGs had certain dysregulation directions (Fig. 4C). Notably, all of the DEGs with certain dysregulation directions were detected by PhenoComp in GSE46224 and GSE26887 by setting the median values of the f_{de}, $f_{up|de}$ and $f_{down|de}$ as p_{de}, $p_{up|de}$ and $p_{down|de}$. Moreover, compared with RankComp, PhenoComp identified 386 and 416 new DEGs in GSE46224 and GSE26887. With $P < 5\%$, there were 17 and 4 pathways that were enriched with 386 and 416 new DEGs, respectively (Supplementary Fig. S3). For example, the adherens junction pathway (Celes *et al.*, 2007; Lee *et al.*, 2004) and cardiac muscle contraction pathway (Greiwe *et al.*, 2016) reportedly play important roles in ICM.

For glioma, we also compared the population-level DEGs identified by PhenoComp with those identified by the common methods in TCGA measured by RNA-seq and GSE50161 measured by

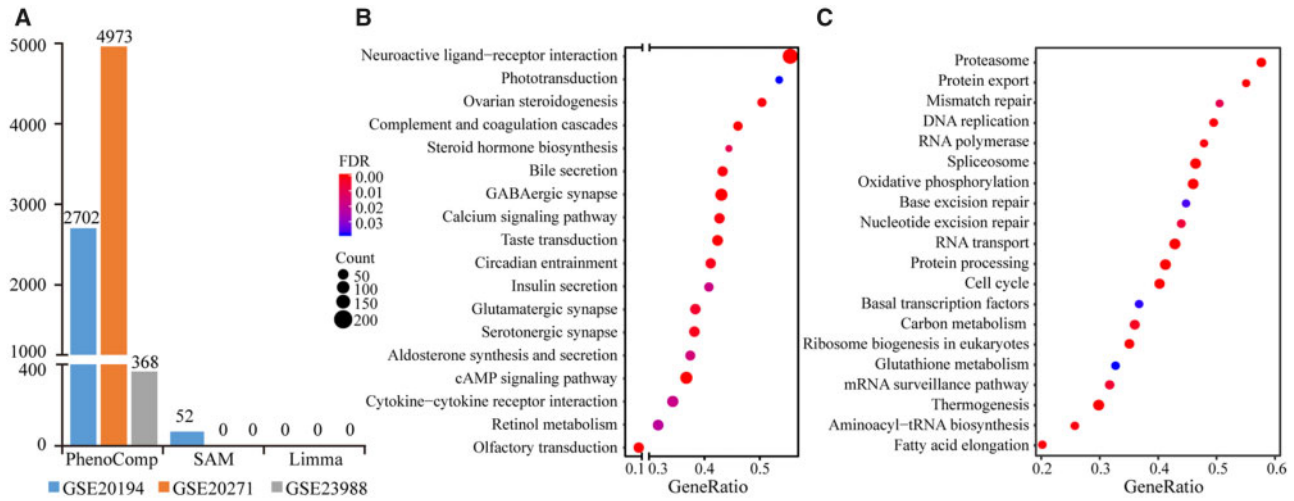

Fig. 5. Analysis of preoperative chemotherapy response data of breast cancer. (**A**) Number of DEGs identified by PhenoComp, SAM and limma in GSE20194, GSE20271 and GSE23988, respectively. The pathways enriched with upregulated genes (**B**) and downregulated genes (**C**) identified by PhenoComp. GeneRatio denoted the proportion of DEGs within a pathway among the total genes within the pathway

microarray. In TCGA, 9787, 11 713 and 11 898 DEGs were identified by PhenoComp, edgeR and limma, respectively (FDR < 5%; Fig. 4A). The concordance scores of DEGs were 84.38% and 82.64% for PhenoComp compared with edgeR and limma, respectively (Fig. 4A), and the POG_{12} and POG_{21} were about 40%. Although only about 40% of the DEGs identified by PhenoComp could be reproducibly detected by the common methods, the overlapping DEGs were highly consistent. Moreover, with FDR <5%, the DEGs identified by PhenoComp, edgeR and limma were enriched in 41, 77 and 79 pathways, respectively (Fig. 4B and Supplementary Table S4). Similar results were also obtained in GSE50161.

For RankComp, 6952 and 4051 population-level DEGs were identified in TCGA and GSE50161, respectively. There were 565 DEGs in TCGA and 57 DEGs in GSE50161 with uncertain directions (Fig. 4C). Notably, in both TCGA and GSE50161, the DEGs identified by RankComp with certain dysregulation directions were all included in those identified by PhenoComp (Fig. 4C). PhenoComp also identified 2837 and 2044 new DEGs in TCGA and GSE50161, respectively. With FDR < 0.2, only one pathway, cortisol synthesis and secretion, was enriched in the 2837 new DEGs, and no pathway was enriched in the new 2044 DEGs. With a loose statistical control $P < 5\%$, the 2837 and 2044 new DEGs were enriched in 3 and 10 KEGG pathways, respectively (Supplementary Fig. S3).

Simultaneously, we also used the mean values of the f_{de}, $f_{up|de}$ and $f_{down|de}$ as p_{de}, $p_{up|de}$ and $p_{down|de}$ to infer the population-level DEGs for PhenoComp, and similar results were also observed for ICM and glioma (Supplementary Figs S2 and S4). Hence, these results suggested that both median and mean values of the f_{de}, $f_{up|de}$ and $f_{down|de}$ could be used to estimate p_{up} and p_{down} for PhenoComp.

3.4 Performance of PhenoComp in weak differential expression signal data

To further show that PhenoComp could identify DEGs even when the differential expression signals were weak, three datasets of ER-negative breast cancer patients with neoadjuvant chemotherapy response data were collected (Table 1). According to the chemotherapy response information, the ER-negative breast cancer patients were divided into two groups: pCR and RD. For limma with FDR <5%, zero DEGs were detected between RD and pCR groups in three datasets. For SAM with FDR <5%, only 52 DEGs were identified in GSE20194, and zero DEGs were identified in both GSE23988 and GSE20271 (Fig. 5A). Using the 39 347 969 stable REOs identified by the 48 pCR samples merged from GSE20194, GSE23988 and GSE20271, individual-level DEGs for each of the 79 RD samples were identified. Then, using PhenoComp by setting the medians of the f_{de}, $f_{up|de}$ and $f_{down|de}$ as p_{de}, $p_{up|de}$ and $p_{down|de}$, 2702, 4973 and 368 DEGs were identified in GSE20194, GSE23988 and GSE20271 (FDR < 5%), respectively. Moreover, the concordance scores of DEGs in any two of the three datasets were >97% (Supplementary Table S5). A total of 6250 DEGs were identified by PhenoComp in GSE20194, GSE23988 and GSE20271, including 3376 upregulated genes and 2874 downregulated genes. There were 18 and 20 biological pathways enriched with 3376 upregulated genes (Fig. 5B) and 2874 downregulated genes (Fig. 5C), respectively. Most of the enriched pathways were reportedly involved in chemotherapy resistance, such as Retinol metabolism (Chen *et al.*, 1997), DNA replication (Warner *et al.*, 2017) and cell-cycle pathways (Ingham and Schwartz, 2017). When the p_{up} and p_{down} were estimated by the mean values of the f_{de}, $f_{up|de}$ and $f_{down|de}$, the similar results were also observed (Supplementary Fig. S5 and Table S6).

4 Discussion

In this study, PhenoComp was mainly developed to identify population-level DEGs for datasets with only one phenotype, which is of special importance for heart and brain disease analysis due to the lack of normal controls. Compared with the original RankComp algorithm, the optimized PhenoComp clearly illustrated how to estimate p_{up} and p_{down}, and provided the dysregulation directions of population-level DEGs and showed better detection power in both simulated and real one-phenotype data, especially for data with small sample size. Moreover, the DEGs inferred by PhenoComp using only one-phenotype data were comparable with those identified by the common methods using two-phenotype data, and PhenoComp could capture the weak differential expression signals, which was hardly detected by the common methods. Particularly, because the REOs of genes within individual samples were insensitive to batch effects and data normalization (Geman *et al.*, 2004; Tan *et al.*, 2005), and the normal background was across-platform, the REO-based PhenoComp could be applied to analyze a particular type of disease data from different sources and by different platforms.

For each independent dataset, due to a lack of real DEGs, we used the simulated experiments to estimate the performance of PhenoComp. In a one-phenotype dataset, the detection power of both PhenoComp and RankComp will be improved when more disease samples are included in the dataset. Especially, in small data, PhenoComp showed better detection power than RankComp. On the other hand, the DEGs detected by the common methods were

also used as the 'gold standard'. Although the POG_{12} and POG_{21} scores were relatively low for both ICM and glioma, the concordance scores of DEGs between PhenoComp and one of the common methods were high in each independent dataset. The results suggested that the DEGs detected by PhenoComp were reliable, and different portions of DEGs may be detected by different methods in each dataset. For ICM, the DEGs identified by PhenoComp were significantly enriched in many biological pathways related to it, such as cytokine-cytokine receptor interaction (Cinquegrana *et al.*, 2005) and cell adhesion molecules (Ortega *et al.*, 2016). We also applied PhenoComp to analyze for glioma. The DEGs identified by PhenoComp were also significantly enriched in many biological pathways related to glioma, such as mTOR signaling pathway (Yuan *et al.*, 2017). Given that DEGs can be significantly enriched in pathways only when it contains a sufficient number of real DEGs, this result provided an additional clue to prove the reliability of the DEGs identified by PhenoComp.

In addition, PhenoComp can also be used for data lacking a sufficient number of normal controls by combining these normal controls with previously accumulated normal controls to construct a normal background. For multi-class data, PhenoComp cannot be directly used to identify population-level DEGs. A possible way to solve this problem is to transform the multi-class (e.g. five-class) samples into multiple groups (10 groups) of two-class samples and identify DEGs for each group of two-class samples by PhenoComp. Then the DEGs shared in a nonrandom high percentage of groups by using the binomial test might be defined as population-level DEGs for the multi-class samples.

In summary, PhenoComp is an efficient algorithm for analyzing specific types of data, including one-phenotype data and weak differential expression signal data, independent of the measurement platform used.

Funding

This work was supported by the National Natural Science Foundation of China [grant number 61801118]. The education research project for young and middle-aged teachers in Fujian Province [grant number JAT170214] and the Fujian Natural Science Foundation [grant number 2019J01678].

Conflict of Interest: none declared.

References

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.

Benito,M. *et al.* (2004) Adjustment of systematic microarray data biases. *Bioinformatics*, **20**, 105–114.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Cai,H. *et al.* (2018) Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. *Int. J. Biol. Sci.*, **14**, 892–900.

Celes,M.R. *et al.* (2007) Reduction of gap and adherens junction proteins and intercalated disc structural remodeling in the hearts of mice submitted to severe cecal ligation and puncture sepsis. *Crit. Care Med.*, **35**, 2176–2185.

Chen,A.C. *et al.* (1997) Human breast cancer cells and normal mammary epithelial cells: retinol metabolism and growth inhibition by the retinol metabolite 4-oxoretinol. *Cancer Res.*, **57**, 4642–4651.

Cinquegrana,G. *et al.* (2005) Effects of different degrees of sympathetic antagonism on cytokine network in patients with ischemic dilated cardiomyopathy. *J. Card. Fail.*, **11**, 213–219.

Ein-Dor,L. *et al.* (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA*, **103**, 5923–5928.

Geman,D. *et al.* (2004) Classifying gene expression profiles from pairwise mRNA comparisons. *Stat. Appl. Genet. Mol. Biol.*, **3**, Article 19.

Greiwe,L. *et al.* (2016) The muscle contraction mode determines lymphangiogenesis differentially in rat skeletal and cardiac muscles by modifying local lymphatic extracellular matrix microenvironments. *Acta Physiol. (Oxf.)*, **217**, 61–79.

GTEx Consortium *et al.* (2017) Genetic effects on gene expression across human tissues. *Nature*, **550**, 204–213.

Ingham,M. and Schwartz,G.K. (2017) Cell-cycle therapeutics come of age. *J. Clin. Oncol.*, **35**, 2949–2959.

International Cancer Genome Consortium *et al.* (2010) International network of cancer genome projects. *Nature*, **464**, 993–998.

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

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Jung,H. *et al.* (2019) DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. *Nat. Commun.*, **10**, 4278.

Kanehisa,M. *et al.* (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.

Kim,D. *et al.* (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.

Law,C.W. *et al.* (2014) voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.*, **15**, R29.

Lazar,C. *et al.* (2013) Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform.*, **14**, 469–490.

Lee,H.Z. *et al.* (2004) The effect of elevated extracellular glucose on adherens junction proteins in cultured rat heart endothelial cells. *Life Sci.*, **74**, 2085–2096.

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

Leek,J.T. *et al.* (2012) The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, **28**, 882–883.

Leek,J.T. (2014) svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res.*, 42.

Loven,J. *et al.* (2012) Revisiting global gene expression analysis. *Cell*, **151**, 476–482.

Molina-Navarro,M.M. *et al.* (2013) Differential gene expression of cardiac ion channels in human dilated cardiomyopathy. *PLoS One*, **8**, e79792.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Nygaard,V. *et al.* (2016) Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, **17**, 29–39.

Ortega,A. *et al.* (2016) New cell adhesion molecules in human ischemic cardiomyopathy. PCDHGA3 implications in decreased stroke volume and ventricular dysfunction. *PLoS One*, **11**, e0160168.

Peixoto,L. *et al.* (2015) How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Res.*, **43**, 7664–7674.

Pertea,M. *et al.* (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.*, **33**, 290–295.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Robinson,M.D. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

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

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

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

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

Wagner,G.P. *et al.* (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, **131**, 281–285.

Wang,D. *et al.* (2012) Extensive up-regulation of gene expression in cancer: the normalised use of microarray data. *Mol. Biosyst.*, **8**, 818–827.

Wang,H. *et al.* (2015) Individual-level analysis of differential expression of genes and pathways for personalized medicine. *Bioinformatics*, **31**, 62–68.

Warner,D.F. *et al.* (2017) DNA replication fidelity in the mycobacterium tuberculosis complex. *Adv. Exp. Med. Biol.*, **1019**, 247–262.

Yuan,Y. *et al.* (2017) Activation of the mTOR signaling pathway in peritumoral tissues can cause glioma-associated seizures. *Neurol. Sci.*, **38**, 61–66.

Zhang,M. *et al.* (2008) Apparently low reproducibility of true differential expression discoveries in microarray studies. *Bioinformatics*, **24**, 2057–2063.