RESEARCH

A pseudo-value regression approach for differential network analysis of co-expression data

Seungjun Ahn[1], Tyler Grimes[2] and Somnath Datta[1*]

*Correspondence:
somnath.datta@ufl.edu

[1] Department of Biostatistics, University of Florida, Gainesville, USA
[2] Department of Mathematics and Statistics, University of North Florida, Jacksonville, USA

Abstract

Background: The differential network (DN) analysis identifies changes in measures of association among genes under two or more experimental conditions. In this article, we introduce a pseudo-value regression approach for network analysis (PRANA). This is a novel method of differential network analysis that also adjusts for additional clinical covariates. We start from mutual information criteria, followed by pseudo-value calculations, which are then entered into a robust regression model.

Results: This article assesses the model performances of PRANA in a multivariable setting, followed by a comparison to *dnapath* and DINGO in both univariable and multivariable settings through variety of simulations. Performance in terms of precision, recall, and F1 score of differentially connected (DC) genes is assessed. By and large, PRANA outperformed *dnapath* and DINGO, neither of which is equipped to adjust for available covariates such as patient-age. Lastly, we employ PRANA in a real data application from the Gene Expression Omnibus database to identify DC genes that are associated with chronic obstructive pulmonary disease to demonstrate its utility.

Conclusion: To the best of our knowledge, this is the first attempt of utilizing a regression modeling for DN analysis by collective gene expression levels between two or more groups with the inclusion of additional clinical covariates. By and large, adjusting for available covariates improves accuracy of a DN analysis.

Keywords: Pseudo-value, Differential network analysis, Regression method, Gene regulatory network, RNA-seq data

Background

The rapid advancement of RNA-sequencing (RNA-seq) data from high-throughput sequencing technologies has provided clear advantages in gene expression studies. It has broadened our understanding of genetics and pathogenesis of human diseases [1, 2]. Compared to microarrays, RNA-seq has a wider dynamic range, the ability to detect novel transcripts, and often results in higher sensitivity and specificity of detection of differential gene expression [3, 4]. With the increase of gene expression studies, the statistical methods to analyze these gene expression data have also accordingly adapted and

progressed. The regression modelling for RNA-seq differential expression (DE) analysis has been established to compare the number of DE genes under different biological or clinical states, including linear model based *limma* [5], negative binomial model based *edgeR* [6], or Poisson log-linear model approach [7]. The analysis of DE has a major limitation in that it looks at one gene at a time, even though a set of genes are often involved in the same biological process [8, 9]. In contrast, the differential network (DN) analysis complements the DE analysis [10] by looking at genes collectively.

The DN analysis identifies changes in measures of association (i.e. network properties or topologies) of the networks across biological conditions, which makes it distinct from a single network analysis. Several groups have proposed statistical methods for DN analysis [11–14]. In particular, DINGO [11] and *dnapath* [14] have developed methods for RNA-seq data, to find differentially connected (DC) genes in subnetworks corresponding to different pathways, between two groups of patients; e.g. 'high-risk' versus 'low-risk' or 'long-term survivors' vs. 'short-term survivors.' While these methods are convenient to use and applicable, they do not consider other observed covariates that may be associated with gene expression.

For instance, a previous study has shown that the expression levels of oxidative stress-associated genes were differentially expressed with smokers with chronic obstructive pulmonary disease (COPD) through gene set enrichment analysis using microarray data [15]. Let us suppose we want to carry out DN analysis on expression data that includes oxidative stress-associated genes and smoking status, which would be used as a grouping variable. In practice, clinicians would also want to include additional covariates such as patient history of cardiac arrhythmia [16] and lung carcinoma [17] to garner more information for better prognosis. However, there is no available direct regression modeling for DN analysis regressing gene expression level between the smoking statuses with the inclusion of additional clinical covariates described above.

The pseudo-value approach was first developed from the leave-one-out jackknife subsampling procedure, applied to a marginal quantity representing some aspect of a marginal distribution of the response variable. It was originally introduced by Andersen and his colleagues [18, 19] for multi-state survival models. Several studies [20, 21] purported that the pseudo-value regression has advantages that the pseudo-values derived from an asymptotically linear and unbiased estimator are approximately independent and identically distributed with the same conditional expectation. Ahn and Logan [22] and Ahn and Mendolia [23] showed that their pseudo-value approaches controlled the type I error while maintaining high power with clustered survival data. With these benefits, we propose a regression modeling method that regresses the jackknife pseudo-values [24] derived from a measure of connectivity of genes in a network to estimate the effects of predictors. Note that the grouping variable itself could also be included in the regression model along with additional clinical covariates while regressing the pseudo-values. We loosely refer to this as a "multivariate setting", whereas in "univariate settings" only the group variable is utilized in a DN analysis.

Thus, in this paper, we introduce a Pseudo-value Regression Approach for Network Analysis (PRANA). This is a novel method of DN analysis that can adjust for additional covariates. We start from mutual information (MI) criteria, followed by pseudo-value calculations, which are then entered into a robust regression model. This article assesses

the model performances of our pseudo-value approach in a multivariable setting, followed by a comparison to *dnapath* and DINGO in the univariable setting through simulations. Lastly, we employ our method in a real data application [25] from the Gene Expression Omnibus (GEO) database [26] to identify DC genes that are associated with COPD. All statistical analyses are performed in R version 4.0.2 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Simulation study

More details of the simulation setup are available in the "Materials" section below. We select $p = 20, 50, 100$ genes to test our pseudo-value approach in smaller to larger gene networks. For each gene network, five different sample sizes $n = 40, 100, 200, 500, 1000$ are considered and in each setting, 1000 Monte Carlo replicates. We draw 1000 random samples first, then take the subsamples from this pool for a simulation with a smaller sample size to reduce computational burden. A random network is generated at each simulation replicate in which a layer of randomness is imposed to account for biological variability of the network structure. For additional details on the generation of simulated RNA-seq data, see the "Materials" section. Simulations are repeated to show the performance of our method by altering the effect size from 5%, 10%, to 20% for simulation scenarios I and II.

Results are compared with the true parent network in order to compute the performance measures described in the "Performance Measures" section. In the true network setting, a gene is considered truly DC between groups if it has at least one DC edge connected to other genes. Tables 1 and 2 summarize simulation results in the multivariable

Table 1 Scenario I simulation results of binary group variable in the multivariable robust regression model (continuous age and binary group) using pseudo-value approach with 1000 replicates

p	*n*	Effect size								
		5%			**10**%			**20**%		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
20	40	0.76	0.81	0.77	0.90	0.79	0.83	0.98	0.79	0.87
	100	0.75	0.91	0.82	0.90	0.91	0.90	0.98	0.91	0.94
	200	0.74	0.94	0.82	0.89	0.94	0.91	0.97	0.94	0.96
	500	0.73	0.97	0.83	0.88	0.97	0.92	0.97	0.97	0.97
	1000	0.73	0.98	0.83	0.88	0.98	0.92	0.97	0.98	0.97
50	40	0.95	0.65	0.77	0.98	0.65	0.78	1.00	0.64	0.77
	100	0.95	0.77	0.85	0.98	0.77	0.86	1.00	0.77	0.86
	200	0.96	0.85	0.90	0.99	0.85	0.91	1.00	0.85	0.91
	500	0.95	0.93	0.94	0.99	0.92	0.95	1.00	0.92	0.96
	1000	0.95	0.96	0.95	0.98	0.96	0.97	1.00	0.96	0.98
100	40	0.96	0.57	0.71	0.98	0.57	0.72	1.00	0.57	0.72
	100	0.96	0.67	0.79	0.99	0.68	0.80	1.00	0.67	0.80
	200	0.96	0.74	0.83	0.99	0.74	0.84	1.00	0.74	0.85
	500	0.97	0.82	0.89	0.99	0.81	0.89	1.00	0.81	0.89
	1000	0.97	0.90	0.93	0.99	0.89	0.94	1.00	0.89	0.94

Random network is generated at each simulation replicate

Table 2 Scenario II simulation results of binary group variable in the multivariable robust regression model (continuous age and binary group) using pseudo-value approach with 1000 replicates

p	*n*	Effect size								
		5%			10%			20%		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
20	40	0.75	0.59	0.64	0.90	0.69	0.70	0.98	0.61	0.73
	100	0.78	0.70	0.71	0.91	0.72	0.79	0.98	0.74	0.83
	200	0.78	0.81	0.78	0.91	0.83	0.86	0.98	0.85	0.91
	500	0.76	0.90	0.82	0.90	0.92	0.91	0.98	0.94	0.95
	1000	0.75	0.94	0.83	0.89	0.95	0.92	0.97	0.96	0.97
50	40	0.95	0.56	0.70	0.98	0.56	0.71	1.00	0.57	0.72
	100	0.95	0.65	0.77	0.98	0.66	0.78	1.00	0.68	0.80
	200	0.96	0.75	0.84	0.99	0.75	0.85	1.00	0.77	0.87
	500	0.96	0.88	0.92	0.99	0.87	0.92	1.00	0.89	0.94
	1000	0.96	0.93	0.94	0.99	0.93	0.96	1.00	0.94	0.97
100	40	0.96	0.55	0.69	0.99	0.55	0.70	0.99	0.55	0.70
	100	0.95	0.63	0.75	0.98	0.63	0.76	0.98	0.63	0.76
	200	0.96	0.68	0.79	0.99	0.68	0.81	0.99	0.68	0.80
	500	0.97	0.77	0.86	0.99	0.77	0.86	0.99	0.77	0.86
	1000	0.97	0.87	0.92	0.99	0.86	0.92	0.99	0.86	0.92

Random network is generated at each simulation replicate

setting for scenarios I and II, respectively, when the continuous variable is added as a covariate with the binary group variable in the regression. Table 2 incorporates the effect of covariate when generating random networks, whereas Table 1 does not. For both scenarios, results show that the pseudo-value regression method generally yields a high precision and recall across all specifications of network size, sample size, and effect size. The pseudo-value regression method maintains a high precision while having an acceptable recall, especially, when a smaller sample size is considered.

Tables 3 and 4 summarize simulation results for scenarios I and II, respectively, when only the binary group variable is included in the model for pseudo-value calculation. Thus, age dependent networks are simulated for Table 4 but not for Table 3. Two competing univariable methods, *dnapath* and DINGO are included for these simulations.

Overall, a similar pattern is observed in the univariable setting; i.e., PRANA consistently reaches a high precision and recall. The performance improves as *n* increases, as to be expected. It is noteworthy that PRANA outperforms *dnapath* in simulation when the sample size is relatively small regardless of the network size. Our method also shows a better recall value and F1-score than *dnapath* with small sample sizes ($n = 40, 100$). As DINGO requires substantially large computational time, it was considered for the gene network with smaller sample sizes only. To be more specific, simulations with larger sample sizes ($n = 500, 1000$) are stopped after 20 days for DINGO from the University of Florida Research Computing Linux server, HiPerGator 3.0 with 10 CPU cores and 10 GB of RAM per node. See Table S1 in Additional file 1 for more details.

Table 5 presents results of scenario III, where age acts as a confounder. That is, an observed difference in connectivity may be due to a difference in the distribution of age in the two groups. Higher precision values from the multivariable pseudo-value

Table 3 Scenario I simulation results of binary group variable in the univariable robust regression model using pseudo-value approach with 1000 replicates

p	n	Precision			Recall			F1		
		PRANA	*dnapath*	**DINGO**	**PRANA**	*dnapath*	**DINGO**	**PRANA**	*dnapath*	**DINGO**
20	40	0.90	**0.95**	0.87	**0.81**	0.64	0.78	**0.84**	0.76	0.82
	100	0.90	**0.93**	0.87	**0.90**	0.88	0.79	0.84	**0.90**	0.82
	200	0.89	**0.91**	0.87	**0.94**	**0.94**	0.76	0.91	**0.92**	0.81
	500	0.88	**0.89**	–	0.97	**0.98**	–	0.92	**0.93**	–
	1000	0.88	**0.89**	–	0.98	**0.99**	–	0.92	**0.93**	–
50	40	0.98	**1.00**	0.99	0.65	0.39	**0.70**	0.78	0.56	**0.81**
	100	0.98	**1.00**	0.98	0.77	0.61	**0.84**	0.86	0.75	**0.90**
	200	0.99	**1.00**	0.98	**0.85**	**0.85**	**0.85**	**0.91**	**0.91**	**0.91**
	500	**0.99**	0.99	–	0.92	**0.95**	–	0.95	**0.97**	–
	1000	0.98	**0.99**	–	0.96	**0.98**	–	0.97	**0.98**	–
100	40	0.98	**1.00**	0.98	0.57	0.27	**0.69**	0.72	0.42	**0.85**
	100	0.98	**1.00**	0.99	0.68	0.32	**0.75**	0.80	0.48	**0.85**
	200	0.99	**1.00**	0.98	0.74	0.62	**0.82**	0.84	0.77	**0.89**
	500	0.99	**1.00**	–	0.81	**0.88**	–	0.89	**0.93**	–
	1000	**0.99**	0.99	–	0.89	**0.94**	–	0.94	**0.97**	–

The network structure does not depend on age covariate. Random network is generated at each simulation replicate. Sample size $n = (500, 1000)$ for gene size $p = 100$ were not included for DINGO due to heavy computational time. The best results are highlighted in boldface

Table 4 Scenario II simulation results of binary group variable in the univariable robust regression model using pseudo-value approach with 1000 replicates

p	n	Precision			Recall			F1		
		PRANA	*dnapath*	**DINGO**	**PRANA**	*dnapath*	**DINGO**	**PRANA**	*dnapath*	**DINGO**
20	40	0.90	**0.97**	0.89	0.59	0.38	**0.63**	0.69	0.53	**0.72**
	100	0.91	**0.97**	0.88	0.72	0.66	**0.75**	0.79	0.77	**0.80**
	200	0.91	**0.96**	0.89	**0.83**	**0.83**	0.73	0.86	**0.88**	0.79
	500	0.90	**0.93**	–	0.93	**0.94**	–	0.91	**0.93**	–
	1000	0.89	**0.91**	–	0.95	**0.97**	–	0.92	**0.94**	–
50	40	0.98	**1.00**	0.99	0.56	0.28	**0.67**	0.71	0.43	**0.80**
	100	0.98	**1.00**	0.98	0.65	0.45	**0.71**	0.78	0.61	**0.82**
	200	0.99	**1.00**	0.98	0.75	0.68	**0.80**	0.85	0.80	**0.88**
	500	**0.99**	**0.99**	–	0.87	**0.91**	–	0.93	**0.95**	–
	1000	**0.99**	**0.99**	–	0.94	**0.97**	–	0.96	**0.98**	–
100	40	0.99	**1.00**	0.55	0.55	0.21	**0.77**	**0.70**	0.34	0.63
	100	0.98	**1.00**	0.55	0.63	0.27	**0.75**	**0.76**	0.42	0.63
	200	0.99	**1.00**	0.55	0.68	0.46	**0.77**	**0.80**	0.62	0.63
	500	0.99	**1.00**	–	0.77	**0.82**	–	0.86	**0.90**	–
	1000	0.99	**1.00**	–	0.86	**0.92**	–ara>	0.92	**0.96**	–

The network structure depends on age covariate. Random network is generated at each simulation replicate. Sample size $n = (500, 1000)$ or gene size $p = 100$ were not included for DINGO due to heavy computational time. The best results are highlighted in boldface

regression indicate that PRANA correctly identifies the DC genes, compared with *dnapath* and DINGO, neither of which accounts for the effects of age. By and large, PRANA has higher precision than DINGO and higher recall than *dnapath*.

Table 5 Scenario III simulation results of binary group variable in the multivariable and univariable robust regression model using pseudo-value approach with 1000 replicates

p	n	Precision				Recall				F1			
		PRANA (Mult)	PRANA (Univ)	dnapath	DINGO	PRANA (Mult)	PRANA (Univ)	dnapath	DINGO	PRANA (Mult)	PRANA (Univ)	dnapath	DINGO
20	40	**0.67**	0.60	0.64	0.58	0.57	0.74	0.50	**0.75**	0.59	**0.65**	0.55	0.64
	100	**0.67**	0.58	0.61	0.58	0.65	**0.85**	0.73	0.79	0.64	**0.68**	0.65	0.66
	200	**0.66**	0.58	0.59	0.58	0.76	**0.91**	0.85	0.80	0.69	**0.70**	0.69	0.67
	500	**0.64**	0.57	0.58	–	0.87	**0.95**	**0.95**	–	**0.73**	0.71	0.71	–
	1000	**0.64**	0.57	0.58	–	0.92	**0.97**	**0.97**	–	**0.75**	0.71	0.72	–
50	40	**0.57**	0.50	0.54	0.49	0.47	0.62	0.33	**0.67**	0.50	0.54	0.40	**0.55**
	100	**0.58**	0.49	0.52	0.49	0.50	0.71	0.51	**0.79**	0.52	0.58	0.51	**0.60**
	200	**0.58**	0.49	0.51	0.49	0.52	0.76	0.60	**0.83**	0.53	0.59	0.54	**0.61**
	500	**0.55**	0.48	0.48	–	0.67	**0.88**	0.83	–	0.59	**0.61**	0.60	–
	1000	**0.54**	0.48	0.48	–	0.81	**0.93**	0.91	–	**0.64**	0.62	0.62	–
100	40	**0.57**	0.49	0.53	0.48	0.44	0.57	0.22	**0.75**	0.49	0.52	0.30	**0.58**
	100	**0.58**	0.49	0.53	0.48	0.47	0.64	0.35	**0.76**	0.51	0.55	0.41	**0.58**
	200	**0.58**	0.49	0.53	0.48	0.46	0.65	0.40	**0.80**	0.50	0.55	0.45	**0.60**
	500	**0.56**	0.47	0.49	–	0.44	**0.74**	0.60	–	0.48	**0.57**	0.53	–
	1000	**0.53**	0.47	0.47	–	0.63	**0.85**	0.82	–	0.57	**0.60**	0.59	–

The network structure depends on age covariate by unequal sampling proportion depending on age categories. Random network is generated at each simulation replicate. Sample size $n =$ (500, 1000) or gene size $p = 100$ were not included for DINGO due to heavy computational costs. The best results are highlighted in boldface

Application study

Analysis of COPDGene data

23 out of 28 COPD-related genes are predicted to be DC between current and non-current smokers with PRANA while accounting for smoking pack years, age, gender, race, and FEV1. A complete list of DC genes found from the pseudo-value approach are CITED2, TESK2, AMZ1, DDX1, DMWD, MED13L, ZBTB38, EML4, HSPA4, ITGB8, TEPP, TNPO1, ARNTL, DTWD1, ADAMTSL3, THRA, SLMAP, DENND2D, STN1, SYN3, ASAP2, IER3, and MFHAS1.

We compared results of PRANA with *dnapath* [14] and DINGO [11]. With DINGO, a total of 19 out of 28 COPD-related genes were selected as DC genes between current and non-current smokers. A complete list of DC genes found in DINGO are the following: ARNTL, DDX1, HSPA4, ITGB8, SLMAP, SYN3, ASAP2, IER3, MFHAS1, VGLL4, CITED2, TESK2, CCDC69, EML4, ADAMTSL3, DENND2D, AMZ1, RASEF, and ZBTB38. Lastly, 3 genes were found DC between current smoking groups with *dnapath*, namely DTWD1, EML4, and TEPP.

Of the 23 DC genes from PRANA, 5 genes are found exclusive to PRANA (DMWD, MED13L, TNPO1, THRA, and STN1). Notably, DMWD is linked to myotonic dystrophy, a rare genetic muscular disorder [27]. Thyroid hormone receptor alpha (THRA) is related to congenital hypothyroidism [28]. These findings about additional genes will facilitate harnessing of the possible mechanisms at work in COPD exacerbation.

Heat shock protein family A (Hsp70) member 4 (HSPA4) is associated with gastric ulcer [29]. Multifunctional ROCO family signaling regulator 1 (MFHAS1) is linked to soft tissue tumor and cell cycle [30]. HSPA4 and MFHAS1 are DC genes identified in both PRANA and DINGO. Echinoderm microtubule-associated protein-like 4 (EML4) is found in all three methods. It has been studied for its association with lung cancer [30, 31]. A Venn diagram is provided to show the overlap between and among three methods (Fig. 4). In addition, a diagram is included to summarize the findings of this application study (Fig. 5).

Discussion

Simulations and real-data analysis have elucidated that PRANA is superior to existing alternatives and a practical tool, which includes covariates in the model. To the best of our knowledge, this is the first attempt to develop a regression modeling in DN analysis. Our working objective is to propose a statistical method that determines whether a gene is significantly DC between groups with the covariate included in the model. In this paper, we have shown through simulations that PRANA reaches a consistently high degree of precision and recall to identify DC genes with varying simulation parameters such as network size, sample size, and effect size. We also analyzed a COPD-related gene expression data from the GEO database. When comparing results from our method to *dnapath* and DINGO, five COPD-related genes are additionally found DC between current versus non-current smokers: DMWD, MED13L, TNPO1, THRA, and STN1.

There are a number of limitations to be highlighted in this study. We have used the absolute value of the differences between the two adjacency matrices as a proxy to determine the true DC genes. Certainly, this is a practical way to detect differences in

the number of edges for each genes in a network. The comparison of maximum values between adjacency matrices was also considered. However, we concluded that they are more useful describing the global characteristic of a network, which deviates from our objective, namely, a gene-specific characteristic of a network.

Another limitation is the inability to perturb simulated networks in a continuous way. Right now, we have discretized the effect of a covariate into three groups. Perhaps, there are other models where a truly continuous covariate could be incorporated.

Lastly, the Pearson correlation, partial correlation, and degree-weighted LASSO were also examined as alternatives to the ARACNE as a measure of association or connectedness, albeit not reported in the paper, due to relatively poor performance and heavy computational costs. It remains an interesting task for future studies to extend our work to other measures of association of a network which better assess different structural changes in the network. We conclude by foregrounding the future direction of the pseudo-value regression approach for the DN analysis, which are potentially extensible to other data types, such as the microbiome data.

Conclusion

The adjustment of covariate is an important step in differential network analysis. In this paper, we presented PRANA, a novel pseudo-value regression approach for the DN analysis, which can incorporate additional clinical covariates in the model. This is a direct regression modeling, and it is therefore computationally amenable for the most users.

Methods

Mutual information and ARACNE

Mutual information (MI) determines whether and how two genes interact. That is, it is a measure of their relatedness and calculated from their joint expression profiles. MI is zero if and only if the joint distribution between the expression level of gene j and gene k satisfies $P(g_j, g_k) = P(g_j)P(g_k)$ for $j \neq k$, or if $j = k$; in other words, they are statistically independent. Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNE) which are proposed by [32, 33] estimates MI using a computationally efficient Gaussian kernel estimator. The estimate of MI is used to quantify the connectivity of each pair of genes in a network. Given a set of two genes measurements, $\overrightarrow{u_i} \equiv (g_{ij}, g_{ik})$, $i = 1, \ldots, n$, the joint probability distribution is approximated as $f(\overrightarrow{u}) = 1/n \sum_i h^{-2} \Phi(h^{-1}(\overrightarrow{u} - \overrightarrow{u_i}))$, where Φ is the bivariate standard normal density and h is the position-dependent kernel width. Then the MI can be expressed as [32]:

$$\hat{I}_{jk} = \frac{1}{n} \sum_i \log \frac{f(g_{ij}, g_{ik})}{f(g_{ij})f(g_{ik})},$$

where $f(g_j)$ and $f(g_k)$ are the marginals of $f(\overrightarrow{u})$. The matrix containing entries $\hat{I}_{jk}$ is defined as the association matrix. The ARACNE algorithm copula-transforms the profiles for MI estimation because MI is reparameterization invariant; thus, the range of these transformed variables is between 0 and 1 [34].

Pseudo-value approach

Let $\hat{I}_{jk}$ be the MI estimate for a pair of genes $j, k \in \{1, \dots, p\}$ of an estimated network from n individuals. For each gene k, we sum the edges (MI estimates) around the gene by taking the column sum of the association matrix to obtain the total connectivity (which can be deemed as a continuous version of degree centrality) of gene k:

$$\hat{\theta}_k = \sum_{j=1}^{p} \hat{I}_{jk},$$

where $k = 1, \dots, p$.

The jackknife pseudo-values [24] for the ith individual and k^{th} gene are defined by:

$$\tilde{\theta}_{ik} = n\hat{\theta}_k - (n-1)\hat{\theta}_{k(i)}, \tag{1}$$

where $\hat{\theta}_{k(i)}$ is the column sum of a gene calculated from the re-estimated association matrix using the RNA-seq data without the ith subject. For each gene k, the re-estimation process requires n such calculations with the data size of $n-1$.

Let Z a binary group indicator. Let $\mathcal{G}_1 = \{i : Z_i = 1\}, \mathcal{G}_2 = \{i : Z_i = 2\}$, and $n_z = |\mathcal{G}_z|$ is the sample size for the two groups $z = 1, 2$ and $n = \sum n_z$. The jackknife pseudo-values are separately obtained within groups. Following the general formula above, for gene k and group z, we similarly define $\hat{\theta}_k^z$ and $\hat{\theta}_{k(i)}^z$, where $i = 1, \dots, n_z$. Then for each $i \in \mathcal{G}_z$, the kth gene jackknife pseudo-values are calculated by $\tilde{\theta}_{ik} = n_z\hat{\theta}_k^z - (n_z - 1)\hat{\theta}_{k(i)}^z$.

Next, a robust regression is applied to regress the pseudo-values on a set of covariates, including Z and $\mathbf{X}$, where Z is the group indicator and $\mathbf{X} = (X_1, \dots, X_q)$ are the potential confounders, such as age and gender. For the ith individual and k^{th} gene, we posit the model

$$E[\tilde{\theta}_{ik}|Z_i, \mathbf{X}_i] = \alpha_k + \beta_k Z_i + \gamma_{k1} X_{i1} + \cdots + \gamma_{kq} X_{iq}, \tag{2}$$

where α_k is the intercept, β_k is the regression coefficient for Z, and $\gamma_{k1}, \dots, \gamma_{kq}$ is the set of regression coefficients to be estimated for X. The main parameter of interest is β_k to test for the change in total connectivity (or degree centrality) of the kth gene between groups. Least trimmed squares (LTS), also known as least trimmed sum of squares [35], is then implemented to perform a robust regression. The LTS estimator is defined by

$$\min_{\alpha_k, \beta_k, \gamma_{k1}, \dots, \gamma_{kq}} \sum_{i=1}^{h} r_{(i)}(\alpha_k, \beta_k, \gamma_{k1}, \dots, \gamma_{kq})^2,$$

where $r_{(i)}$ is the set of ordered absolute values of the residuals (in increasing order of absolute value) and h may depend on some pre-defined trimming proportion c, for instance by means of $h = [n(1 - c)] + 1$. In general, c is chosen between 0.5 and 1 [36].

Hypothesis testing

To test whether the true difference in total connectivity of kth gene differs between groups, we test the null hypothesis of $H_0 : \beta_k = 0$ against the research hypothesis $H_1 : \beta_k \neq 0$. The t-statistic is computed by $\hat{\beta}_k / SE(\hat{\beta}_k)$, where $SE(\hat{\beta}_k)$ standard error of

$\hat{\beta}_k$, obtained using large sample theory, and which is the least-squares estimator of β_k for $k = 1, \ldots, p$ from the robust regression described in equation (2). P-values are calculated using a t-distribution as in *robustbase* R package [37, 38].

It is important to control the false discovery rate (FDR), since multiple hypothesis tests are conducted in the DN analysis. The FDR measures the proportion of false discoveries among a set of genes which are significantly DC between groups. The empirical Bayes screening (EBS) approach [39] has been applied to control the FDR, which is an extension of Westfall and Young step-down adjusted p-values [40]. The EBS procedure is robust against model mis-specification, as it utilizes nonparametric function estimation techniques for the estimation of the marginal density of the transformed p-values.

Materials

This section details step-by-step procedures how the simulation is performed. The performance of our proposed method is assessed by an extensive simulation study. Data are simulated with different number of genes p and sample size n. In this simulation, the regression model includes two covariates Z and X, where Z is the group indicator and $X \sim N(55, 10)$ is a continuous covariate such as the age of a patient. Three different simulation scenarios are considered.

Data generation

Simulate weighted networks and RNA-seq data with a dependence structure that depends on Z and/or X using the *SeqNet* R package [41]. In this setting, there are total of six networks for the combination of two groups and three age categories (younger than 50, 50-60, and older than 60). We consider three different scenarios incorporating group information only (scenario I), age and group information (scenario II), and age and group information with unequal sampling proportions with different distributions of the age in the two groups (scenario III) (see Figures 1–3 for visual demonstrations).

Scenarios I (a–b) and II (a–c)

a. Generate the first random network with p nodes for $z = 1$. The $p \times p$ adjacency matrix, where the diagonal elements are 0 and non-diagonal elements are in $\{0, 1\}$, is extracted from this first graph. It is a symmetric matrix indicating whether a pair of nodes are connected by an edge. Take the column sum of the adjacency matrix to see the total number of connected edges to the node. Record the indices of this vector with column sum for the use of effect size adjustment in later step.

b. Perturb the first network to generate the second network for $z = 2$ by removing the edges around nodes using the indices obtained in previous step. To assess the effect size of group, the top $5\%, 10\%$, and 20% of total nodes with the most number of edges in a network lose their edges (e.g. 2 nodes with the most number of edges for a network with $p = 20$ for the effect size of 10%, Fig. 1). This is the end of scenario I.

c. For scenario II, further perturb remaining networks by removing edges of one additional node with the next most number of edges, coming after Step (b) above. This is to simulate networks with a covariate dependence structure on both age and group (see Fig. 2).
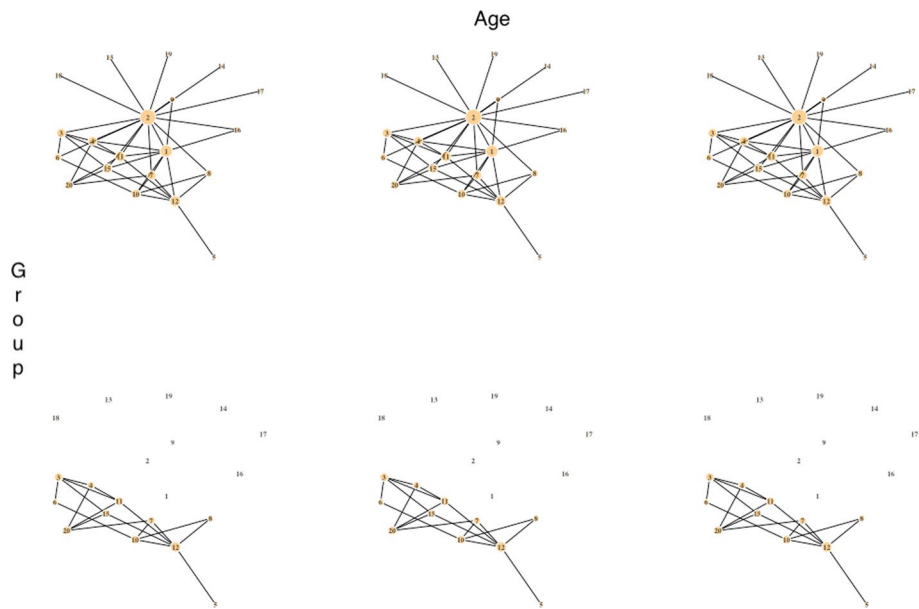
Fig. 1 Network plots visualizing the gene network ($p = 20$) without a covariate dependence structure that depends on binary group only (scenario I). The row represents group whereas the column represents age categories. The three networks in each row are identical, since there is no effect of age on the structure of network. The edges of the hub nodes are removed based on the effect size of the binary group
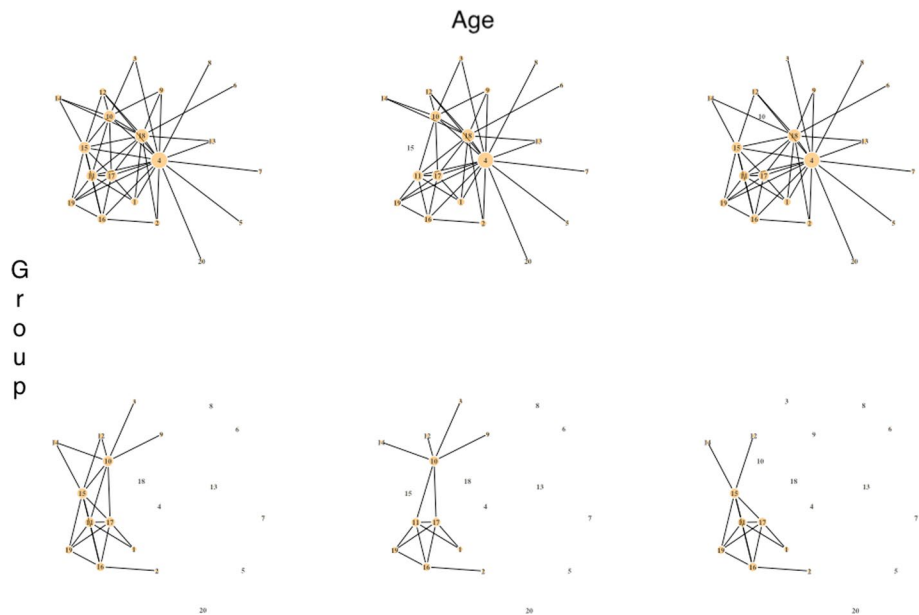
Fig. 2 Network plots visualizing the gene network ($p = 20$) with a covariate dependence structure that depends on age and group information (scenario II). The row represents group whereas the column represents age categories. All six networks have unique structure of the network. The edges of the hub nodes are firstly removed based on the effect size of the group, as shown in Fig. 1 above. For this scenario II, additional edges of nodes with greater number of connected edges are removed for each age category
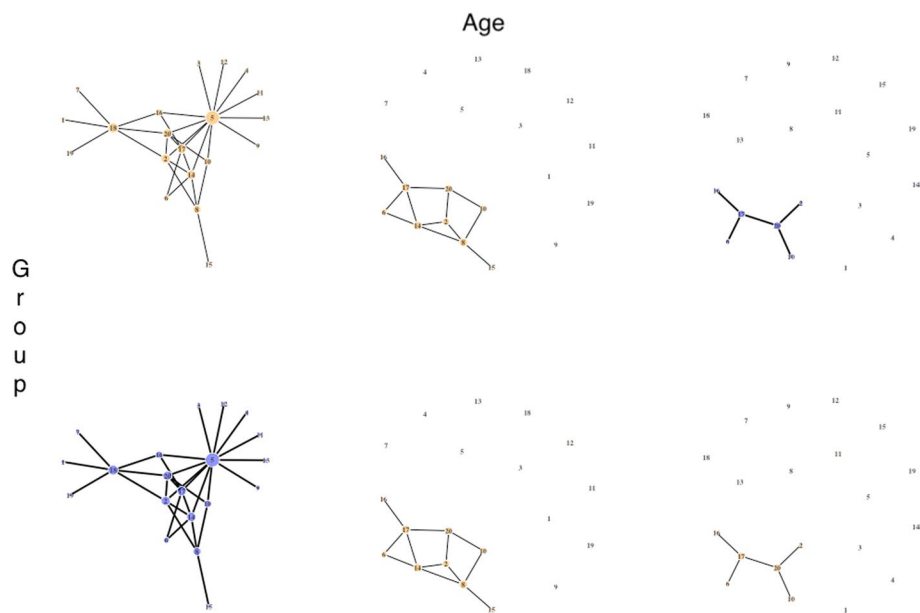
Fig. 3 Network plots visualizing the gene network ($p = 20$) with a covariate dependence structure that depends on age and group information with unequal sampling proportions with respect to different distribution of the age in the two groups (scenario III). The row represents group whereas the column represents age categories. All six networks have unique structure of the network. The edges of the two hub nodes are removed for each age category. To employ the effect of group, 10%/10%/80% of the subjects in $z = 1$ will have a network structure to each of the first, second, and third networks in the first row. In contrast, 80%/10%/10% of the subjects in $z = 2$ will have a network structure to each of the first, second, and third networks in the second row
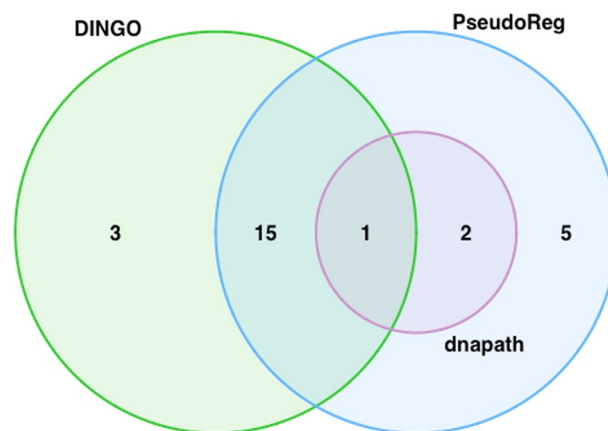
Fig. 4 A Venn diagram displaying the number of overlapping DC genes between and among univariable analysis such as DINGO and *dnapath* versus multivariable robust regression with pseudo-value approach using COPDGene study data from GEO database

Scenario III

We created a scenario where age is acting like a confounder. In other words, for a given each category that two networks are the same, but the distributions of the age of the patients are different in the two groups. Therefore, there will be an observed difference in network connectivity, which is explained through age.
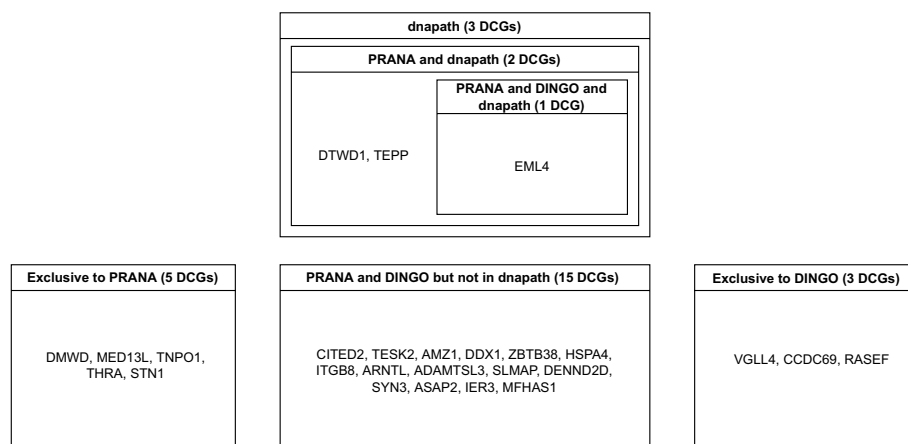
Fig. 5 A diagram summarizing results using each methods analyzing the COPDGene study data from GEO database. A full list of DCGs (differentially connected genes) are provided in each box

a. Generate the first random network with p nodes for younger than 50 category. The $p \times p$ adjacency matrix, where the diagonal elements are 0 and non-diagonal elements are either $\{0, 1\}$, is extracted from this first graph. Record the indices of connected edges for the perturbation of network in later steps below.

b. Perturb the first network to generate the second network for age 50–60 category by removing the edges of the two nodes with the most number of edges in a network lose all of their edges. In other words, we refer to the indices, recorded in the adjacency matrix from the earlier step, and remove all the connected edges around the two nodes.

c. Next, we repeat the same to perturb the second network to obtain the third network for older than 60 category (see Fig. 3).

For all three scenarios, we assign a partial correlation to edges to obtain weighted networks [41]. Note that adjacency matrices of these weighted networks are used for the the true connection per gene. Generate RNA-seq samples based on weighted networks with equal sampling proportions for scenarios I and II. However, specifically for scenario III, a sampling proportion differs across age categories and groups. That is, 10%/10%/80% for $z = 1$ and 80%/10%/10% for $z = 2$. The data generation involves with two major steps. Firstly, we generate gene expressions (Gaussian values) from a group-specific weighted network for each gene, denoted as $\tilde{x}_i \sim N(0, 1)$. These Gaussian values are then mapped into RNA-seq data column-wise by using the inverse CDF of empirical distribution of the reference data using expression data with accession number GSE158699 [25] from the Gene Expression Omnibus (GEO) database [26]. We will have $n_z \times p$ matrices for each group $z = 1, 2$.

Algorithm

1 Obtain an association matrix with ARACNE from the data generated in steps from the "Data Generation" section to fit an estimated network using *minet* [32, 42] for each group.

2 For each gene k, calculate the column sums of association matrix for each group z separately, denoted by $\hat{\theta}_k^z$.

3 For each gene k and individual $i \in \mathcal{G}_z$, compute $\hat{\theta}_{k(i)}^z$ from the association matrix that is re-estimated based on RNA-seq data without the ith subject of $n_z \times p$ data from the "Data Generation" section for each group z separately, where $i = 1, \ldots, n_z$.

4 Calculate $\tilde{\theta}_{ik}$ using Eq. (1) based on Step 2 and 3.

5 For each gene k, fit a multivariable robust regression with binary group variable and continuous age variable to obtain the p values of the group variable, computed from the t test. These p values are used to compute the performance measures of simulation study. More details on the performance measures are stated next.

Performance measures

To evaluate the performance of our proposed method, precision, recall, and the F1 score are calculated. Let $\Omega^z \in \mathbf{R}^{p \times p}$ be the adjacency matrix for group z, where

$$\Omega_{jk}^z = \begin{cases} 1 & \text{if } j^{\text{th}} \text{ gene and } k^{\text{th}} \text{ gene are connected} \\ 0 & \text{otherwise,} \end{cases}$$

for $z = 1, 2$. Then, for each gene k, we calculate

$$\eta_k = I\left(\sum_{j=1}^{p} |\Omega_{jk}^1 - \Omega_{jk}^2| \geq 1 \right),$$

where $I(\cdot)$ is an indicator function to determine whether gene k has differential connectivity. The gene k is truly DC if $\eta_k = 1$, and is not DC if $\eta_k = 0$ for the true gene network. Similarly, for the covariate dependence structure, the following quantities are obtained

$$\eta_k = I\left(\frac{1}{c} \sum_{c} \sum_{j=1}^{p} |\Omega_{jk}^{1,c} - \Omega_{jk}^{2,c}| \geq 1 \right),$$

where $\Omega^{z,c} \in \mathbf{R}^{p \times p}$ be the adjacency matrix for group z and age category $c = 1, 2, 3$. Denote that S is the total number of Monte Carlo simulation replicates. Let q_{ks} be adjusted p value as in following procedure [39] of kth gene at the sth simulation replicate. α represents the magnitude of error control, and 0.05 was used throughout the simulation.

- Precision is the proportion of genes that are inferred to be significantly DC from the test which have true connection between two comparing groups:

$$\text{Precision} = \frac{\sum_{k=1}^{p} \eta_k I(q_{ks} < \alpha)}{\sum_{k=1}^{p} I(q_{ks} < \alpha)}.$$

- Recall is the proportion of genes that have true connection which are correctly inferred to be significantly DC between two comparing groups:

$$\text{Recall} = \frac{\sum_{k=1}^{p} \eta_k \, I(q_{ks} < \alpha)}{\sum_{k=1}^{p} \eta_k}.$$

- F1 is calculated based on the harmonic mean of precision and recall obtained from the simulation. A higher F1 score suggests lower false negative and false positive rate:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$

COPDGene data

A recent genome-wide association study [43] identified 35 new COPD-related genes from the UK Biobank and International COPD Genetics Consortium data. Among these 35 COPD-related genes, 28 genes are available in the data from the COPDGene study for the analysis using PRANA and other DN analysis methods including *dnapath* and DINGO. The 28 COPD-related genes are the following: CITED2, TESK2, COL15A1, AMZ1, RASEF, DDX1, DMWD, MED13L, ZBTB38, CCDC69, EML4, HSPA4, ITGB8, TEPP, TNPO1, ARNTL, DTWD1, ADAMTSL3, RREB1, THRA, SLMAP, DENND2D, STN1, SYN3, ASAP2, IER3, MFHAS1, and VGLL4.

Among 2561 samples from the initial phenotype data, we have used 406 samples that were provided as the validation set in the analysis of the original study. For the analysis with PRANA, binary current smoking status variable is used as the grouping variable, and smoking pack years, age, gender, race, and FEV1 are included as additional covariates in a multivariable model. The binary current smoking status variable is used as the grouping variable for *dnapath* and DINGO.

Abbreviations

COPD	Chronic obstructive pulmonary disease
DC	Differentially connected
DE	Differential expression
DN	Differential network
GEO	Gene Expression Omnibus
LASSO	Least absolute shrinkage and selection operator
MI	Mutual information

Availability of data and materials
PRANA is available at (https://github.com/sjahnn/PRANA). The COPDGene data is available in the GEO database with accession number GSE158699 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158699). Please reach out to the author (Somnath Datta, somnath.datta@ufl.edu) if you have any further inquiries on the data or code.

Declarations

Ethical approval and consent to participate
Not applicable.

Ahn *et al. BMC Bioinformatics*   (2023) 24:8

Page 16 of 17

Consent for publication
Not applicable.

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

References

1. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Mol Cell. 2015;58(4):586–97.
2. Gao M, Zhong A, Patel N, Alur C, Vyas D. High throughput RNA sequencing utility for diagnosis and prognosis in colon diseases. World J Gastroenterol. 2017;23(16):2819–25.
3. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63.
4. Zhao S, Fung-Leung W, Bittner A, Ngo K, Liu X. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS ONE. 2014;9(1): e78644.
5. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
6. 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.
7. Li J, Witten DM, Johnstone IM, Tibshirani R. Normalization, testing, and false discovery rate estimation for RNA-sequencing data. Biostatistics. 2012;13(3):523–38.
8. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012;8(565):1–9.
9. Kim Y, Hao J, Gautam Y, Mersha TB, Kang M. DiffGRN: differential gene regulatory network analysis. Int J Data Min Bioinform. 2018;20(4):362–79.
10. de la Fuente A. From 'differential expression' to 'differential networking'—identification of dysfunctional regulatory networks in diseases. Trends Genet. 2010;26(7):326–33.
11. Ha MJ, Baladandayuthapani V, Do KA. DINGO: differential network analysis in genomics. Bioinformatics. 2015;31(21):3413–20.
12. McKenzie AT, Katsyv I, Song WM, Wang M, Zhang B. DGCA: a comprehensive R package for differential gene correlation analysis. BMC Syst Biol. 2016;10(106):1–25.
13. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinform. 2010;11(95):1–10.
14. Grimes T, Potter SS, Datta S. Integrating gene regulatory pathways into differential network analysis of gene expression data. Sci Rep. 2019;9(1):1–12.
15. Pierrou S, Broberg P, O'Donnell RA, Pawlowski K, Virtala R, Lindqvist E, et al. Expression of genes involved in oxidative stress responses in airway epithelial cells of smokers with chronic obstructive pulmonary disease. Am J Respir Crit Care Med. 2007;175(6):577–86.
16. Rusinowicz T, Zielonka TM, Zycinska K. Cardiac arrhythmias in patients with exacerbation of COPD. Adv Exp Med Biol. 2017;1022:53–62.
17. Durham AL, Adcock IM. The relationship between COPD and lung cancer. Lung Cancer. 2015;90(2):121–7.
18. Andersen P, Klein J, Rosthøj S. Generalised linear models for correlated pseudo-observations, with applications to multi-state models. Biometrika. 2003;90(1):15–27.
19. Andersen PK, Klein JP. Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies. Scand J Stat. 2007;34:3–16.
20. Dutta S, Datta S, Datta S. Temporal prediction of future state occupation in a multistate model from high-dimensional baseline covariates via pseudo-value regression. J Stat Comput Simul. 2017;87(7):1363–78.
21. Graw F, Gerds TA, Schumacher M. On pseudo-values for regression analysis in competing risks models. Lifetime Data Anal. 2009;15:241–55.
22. Ahn KW, Logan BR. Pseudo-value approach for conditional quantile residual lifetime analysis for clustered survival and competing risks data with applications to bone marrow transplant data. Ann Appl Stat. 2016;10(2):618–37.
23. Ahn KW, Mendolia F. Pseudo-value approach for comparing survival medians for dependent data. Stat Med. 2014;33(9):1531–8.
24. Efron B, Tibshirani RJ. An introduction to the bootstrap. Philadelphia: Chapman & Hall/CRC; 1993.
25. Wang Z, Masoomi A, Xu Z, Boueiz A, Lee S, Zhao T, et al. Improved prediction of smoking status via isoform-aware RNA-seq deep learning models. PLoS Comput Biol. 2021;17(10):e1009433.
26. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets-update. Nucleic Acids Res. 2013;41:D991–5.
27. Westerlaken JH, Van der Zee CE, Peters W, Wieringa B. The DMWD protein from the myotonic dystrophy (DM1) gene region is developmentally regulated and is present most prominently in synapse-dense brain areas. Brain Res. 2003;971(1):116–27.
28. Tylki-Szymańska A, Acuna-Hidalgo R, Krajewska-Walasek M, Lecka-Ambroziak A, Steehouwer M, Gilissen C, et al. Thyroid hormone resistance syndrome due to mutations in the thyroid hormone receptor α gene (THRA). J Med Genet. 2015;52(5):312–6.
29. Sakurai T, Kashida H, Hagiwara S, Nishida N, Watanabe T, Fujita J, et al. Heat shock protein A4 controls cell migration and gastric ulcer healing. Dig Dis Sci. 2015;60(4):850–7.
30. Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards suite: from gene data mining to disease genome sequence analyses. Curr Protoc Bioinform. 2016;54:1301–13033.

31. Adib R, Montgomery JM, Atherton J, O'Regan L, Richards MW, Straatman KR, et al. Mitotic phosphorylation by NEK6 and NEK7 reduces the microtubule affinity of EML4 to promote chromosome congression. Sci Signal. 2019;12(594):eaaw2939.
32. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla-Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinform. 2006;7 Suppl 1(Suppl 1):S7.
33. Zoppoli P, Morganella S, Ceccarelli M. TimeDelay-ARACNE: Reverse engineering of gene networks from time-course data by an information theoretic approach. BMC Bioinform. 2010;11(154):1–15.
34. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A. Reverse engineering of regulatory networks in human B cells. Nat Genet. 2005;37(4):382–90.
35. Rousseeuw PJ. Least median of squares regression. J Am Stat Assoc. 1984;79(388):871–80.
36. Pison G, Van Aelst S, Willems G. Small sample corrections for LTS and MCD. Metrika. 2002;55:111–23.
37. Todorov V, Filzmoser P. An object-oriented framework for robust multivariate analysis. J Stat Soft. 2009;32(3):1–47.
38. Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, et al. robustbase: Basic Robust Statistics; 2022. R package version 0.95-0. http://robustbase.r-forge.r-project.org/.
39. Datta S, Datta S. Empirical Bayes screening of many p-values with applications to microarray studies. Bioinformatics. 2005;21(9):1987–94.
40. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. New York: Wiley-Interscience; 1993.
41. Grimes T, Datta S. SeqNet: An R package for generating gene-gene networks and simulating RNA-seq data. J Stat Soft. 2021;98(21). https://doi.org/10.18637/jss.v098.i12.
42. Meyer PE, Lafitte F, Bontempi G. minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information. BMC Bioinform. 2008;9(461).
43. Sakornsakolpat P, Prokopenko D, Lamontagne M, et al. Genetic landscape of chronic obstructive pulmonary disease identifies heterogeneous cell-type and phenotype associations. Nat Genet. 2019;51(3):494–505.

Publisher's Note