Parsana et al. Genome Biology (2019) 20:94
https://doi.org/10.1186/s13059-019-1700-9

Genome Biology

SHORT REPORT Open Access

Addressing confounding artifacts in reconstruction of gene co-expression networks

Princy Parsana[1†], Claire Ruberman[2†], Andrew E. Jaffe[2,3,4,5,6], Michael C. Schatz[1,7], Alexis Battle[1,8*] and Jeffrey T. Leek[2,6*]

Abstract

Gene co-expression networks capture biological relationships between genes and are important tools in predicting gene function and understanding disease mechanisms. We show that technical and biological artifacts in gene expression data confound commonly used network reconstruction algorithms. We demonstrate theoretically, in simulation, and empirically, that principal component correction of gene expression measurements prior to network inference can reduce false discoveries. Using data from the GTEx project in multiple tissues, we show that this approach reduces false discoveries beyond correcting only for known confounders.

Background

Gene co-expression networks seek to identify transcriptional patterns indicative of functional interactions and regulatory relationships between genes [1–3]. These are not yet fully characterized for most species, tissues, and disease-relevant contexts. Therefore, reconstructing co-expression networks from high-throughput measurements is of common interest. However, accurate reconstruction of such networks remains a challenging problem.

Though some specialized methods for the reconstruction of co-expression networks do consider confounding signals within their model [4, 5], routinely used network learning methods [6, 7] do not directly account for technical and unwanted biological effects known to confound gene expression data. Despite this, many studies do not employ any form of data correction or correct only for known confounders prior to network reconstruction (Additional file 1: Table S1). These artifacts influence gene expression measurements, often introducing spurious correlations between genes [8–10]. These correlations are often inferred as relationships between genes, leading to inaccurate network structure and erroneous conclusions in downstream analyses [4, 5, 8, 11, 12]. Therefore, it is critical to correct gene expression data for unwanted biological and technical variation without eliminating signal of interest before applying standard network learning methods.

Results and discussion

In this study, we provide a framework for data correction leveraging the structure of scale-free networks. We show that for scale-free networks, principal components of a gene expression matrix can consistently identify components that reflect artifacts in the data rather than network relationships. It has been shown that real-world networks including co-expression networks often have scale-free topology, i.e., the node degree distribution of these networks follow a power law [13–15]. Several studies have employed the assumption of scale-free topology to infer high-dimensional gene co-expression and splicing networks [6, 16].

Latent factor-based data correction has been successfully employed in many applications in genomics from genome-wide association studies, cis- and trans-eQTL mapping, to differential expression analysis [9, 17–20]. In genome-wide association studies investigating the association between genotype and complex traits, it has been shown that top principal components explain the

* Correspondence: ajbattle@jhu.edu; jtleek@gmail.com
†Princy Parsana and Claire Ruberman contributed equally to this work.
[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA
[2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
Full list of author information is available at the end of the article

broad correlation between genotypes which generally reflects population structure rather than a desired functional biological signal of interest [20]. Co-expression analysis is more complicated because confounders affect sets of genes in ways that resemble co-expression. Here, we show mathematically, through simulation (Fig. 1, Additional file 1: Notes 1, 2.1, and 2.2; Additional files 2 and 3) and through real data examples that similar to genetic association studies, the broad correlation between gene expression levels in uncorrected data appears to reflect artifacts. We expect that most real co-expression networks are sparse which means that most genes are only connected to a small subset of other genes. We prove that when such networks satisfy the scale-free property, the signals from the network will not be sufficiently broad across genes to influence the latent variable estimates from PCA. Thus, principal components will primarily capture latent confounders, which can then be regressed from the expression data before network reconstruction is performed (Additional file 1: Note 1).

Using a toy and scale-free simulation, we first showed that confounding can introduce false correlations between sets of genes that can mimic co-expression and can lead to false edge discovery during reconstruction of co-expression networks with graphical lasso—sometimes at the expense of losing true connections (Fig. 1d-f, Additional file 2). We corrected the confounded simulated data using our PC-based approach and reconstructed the network using the residuals. Graphical lasso correctly estimated the network structure obtained from corrected data, which was the same as the true network structure that was obtained from the original simulated data (Fig. 1a-c,g-h, Additional file 2). We also simulated multivariate Gaussian data with 350 samples and 5000 genes from an underlying scale-free network (Additional file 3). Similar to the previous simulation, we found that confounding in data can introduce a lot more false positives in reconstructed co-expression networks. We also showed that networks reconstructed with PC corrected data in this setting were more similar to original simulated data compared to confounded data (Additional file 3). Throughout our analysis, to estimate the number of principal components to be removed, we used a permutation-based scheme [21] as implemented in the sva package [22].

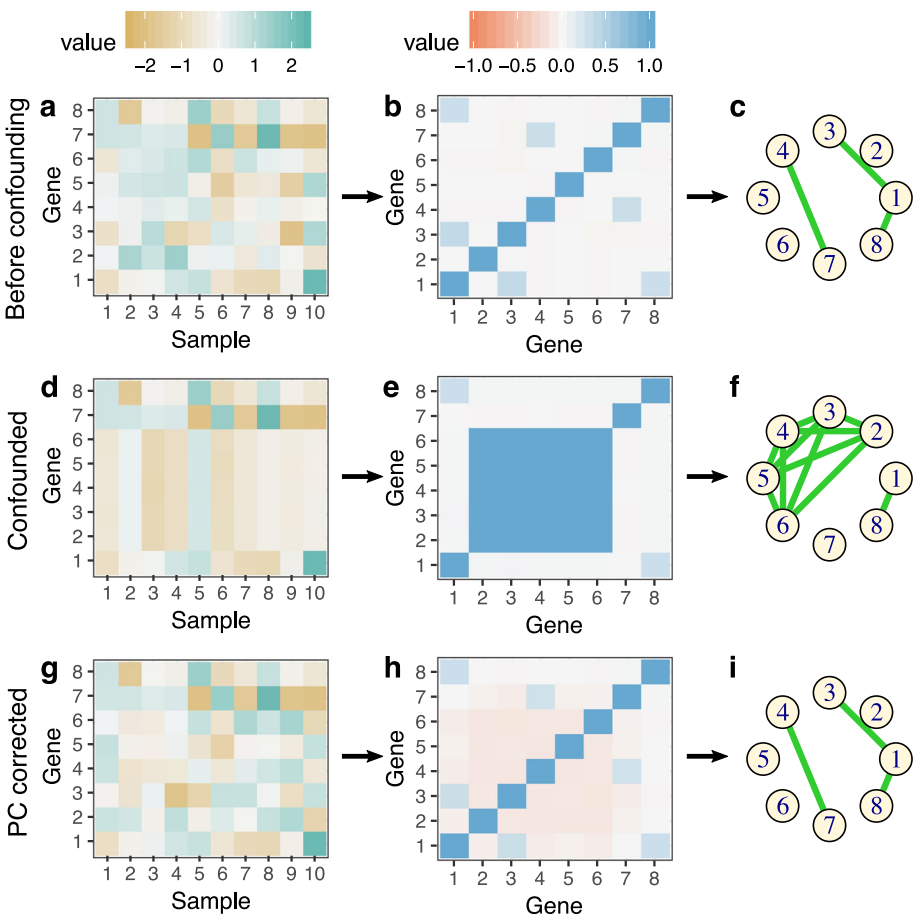

Fig. 1 Toy simulation example. (**a-f**) This toy simulation shows the reconstruction of gene co-expression networks is affected by confounders. (**g-i**) True underlying network structure can be reconstructed after principal component correction of gene expression data as described in the paper

To demonstrate the impact of latent confounders and principal component correction on the reconstruction of co-expression networks from real large-scale human gene expression measurements, we applied our method to RNA-Seq data from the Genotype-Tissue Expression (GTEx) project v6p release. We considered data from eight diverse tissues containing between 304 and 430 samples each (Additional file 1: Table S2): Subcutaneous adipose, lung, skeletal muscle, thyroid, whole blood, tibial artery, tibial nerve, and sun-exposed skin. Using the most variable 5000 genes (Additional file 1: Notes 2 and 4), we reconstructed co-expression networks for each tissue with two popular methods: (a) weighted gene co-expression network analysis [6, 23] and (b) graphical lasso [7, 24]. Since the true underlying co-expression network structure is not known, we assessed the networks using gene pairs annotated to function in the same pathways [25, 26] as ground truth edges.

We inferred networks obtained by using (a) uncorrected expression data, the residuals after regressing out (b) RNA integrity number (RIN), (c) exonic rate—a mapping covariate that corresponds to fraction of reads mapped to exons, (d) sample-specific estimate of GC bias, all known to be common confounders in mRNA gene expression data [27–29], and (e) residuals from multiple regression model using covariates that explained at least 1% of expression variance (adjusted $R^2 \geq 0.01$, Additional file 1: Table S3–S5) [28, 30–33].

Co-expression gene modules obtained from weighted signed co-expression networks (Additional file 1: Note 2.4) were interpreted as fully connected subgraphs, as is standard. For most tissues, networks obtained from data corrected for latent confounders showed fewer false discoveries compared to those obtained from uncorrected data or from correcting for individual covariates including RIN, exonic rate (a quality metric from RNA-Seq mapping), or sample-specific GC bias (Fig. 2, Additional file 1: Figures S1, S3, and S8). Improved performance of networks obtained from PC corrected data was more evident in the whole blood, skeletal muscle, tibial artery, tibial nerve, subcutaneous adipose, and thyroid. But for some tissues such as the lung, PC correction only contributes to moderate improvement on false discovery rates in the reconstructed networks. It is possible that in these cases, the networks may violate the scale-free assumption or that true signal was already sufficiently strong in the raw data. We also observed that correcting gene expression data with multiple technical covariates (approximately 9–17 were used per tissue, Additional file 1: Table S5) sometimes improved the reconstruction of co-expression networks obtained by WGCNA (Fig. 2a–c, Additional file 1: Figure S1). Average WGCNA module size for networks with cut-height greater than 0.99 was smaller with PC-corrected data compared to uncorrected counterparts (Additional file 1: Figure S15). We also observed that the number of genes assigned to the gray (unassigned) module in WGCNA was considerably higher in PC-corrected networks (Additional file 1: Figure S15). Finally, we repeated this analysis by varying multiple settings of WGCNA and found that PC corrected showed improvement in most tissues consistently (Additional file 1: Figures S10 and S11).

In graphical lasso networks, we found that networks estimated with principal component corrected data showed fewer false discoveries compared to networks estimated with uncorrected, RIN-corrected or multiple covariates corrected data (Fig. 2d–f, Additional file 1: Figure S2). We observed that in generally improved performance on false discoveries in PC corrected networks over raw data in the whole blood, the skeletal muscle, tibial artery, and tibial nerve. Compared to raw data, jointly correcting the gene expression data for multiple technical covariates that affect expression measurements also improved reconstruction with graphical lasso in some tissues such as the whole blood, thyroid, and tibial artery, while it showed little to no improvement over uncorrected data in the lung, muscle, tibial nerve, and sun-exposed skin (Fig. 2d–f, Additional file 1: Figure S2). However, we observed that across all tissues, PC correction still shows fewer false discoveries compared to multiple technical covariate-based correction. There was no visible improvement in network reconstruction between using uncorrected data and residuals from RIN or exonic rate, thereby suggesting that RIN, exonic rate, or GC bias individually is not a sufficient alternative for the wide range of confounding variation found in gene expression data (Fig. 2, Additional file 1: Figures S2, S4, and S9). We also found that there was no improvement on false negative rates upon PC correction in networks built with WGCNA or graphical lasso (Additional file 1: Figure S14).

With both WGCNA and graphical lasso, networks inferred from principal component corrected data were much sparser than networks from uncorrected and RIN, exonic rate, or GC bias corrected counterparts (Fig. 2g–l). Further, PC corrected networks from graphical lasso also showed higher clustering coefficient and fewer hubs compared to others (Additional file 1: Figures S12 and S13).

Conclusion

Network reconstruction methods are vulnerable to latent confounders present in gene expression data. Co-expression networks obtained from data corrected for effects of RIN, exonic rate, or GC bias individually show little improvement on false discoveries compared to uncorrected data and are not a sufficient surrogate

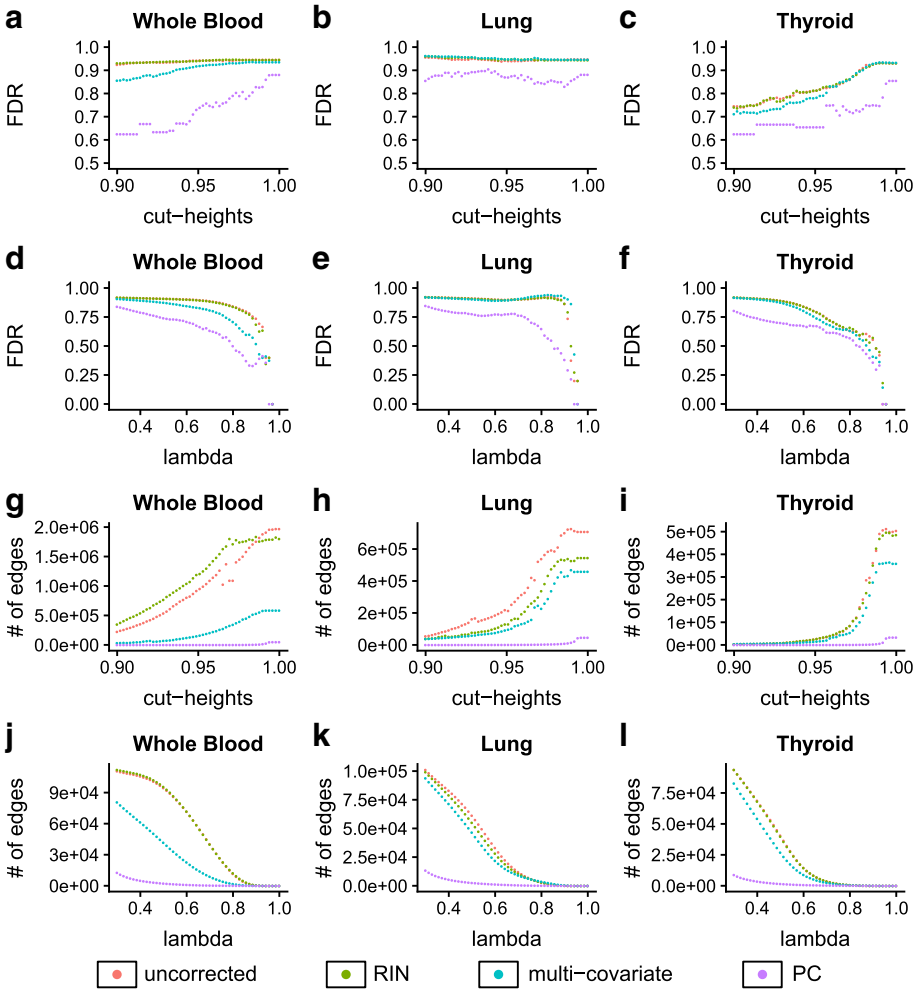

Fig. 2 False discovery rate of WGCNA modules and graphical lasso networks based on canonical pathways (**a**–**f**). The density of networks inferred from PC-corrected data is sparser (**g**–**l**). **a**–**c** FDR of WGCNA networks obtained at varying cut heights. Each point corresponds to FDR of the network obtained at a specific cut height. Each color represents networks reconstructed with a specific correction approach. **d**–f Each point in the figure corresponds to false discovery rates of networks obtained at a specific L1 penalty parameter value (lambda) in the graphical lasso. Each color represents networks reconstructed with a specific correction approach—uncorrected, multi-covariate, RIN, and PC corrected. **g**–**i** Each point corresponds to a number of edges in networks inferred by WGCNA at a cut height. **j**–**l** Each point corresponds to a number of edges inferred by graphical lasso in networks obtained at a specific L1 penalty parameter value. Networks inferred by PC-corrected data have fewer edges compared to uncorrected or RIN-corrected data

for the diverse sources of confounding variation in gene expression data. With empirical analysis supported by theoretical proof, we show that PC correction is a simple, yet effective approach to address confounding variation for the reconstruction of gene co-expression networks. We do note for particularly dense or connected sub-graphs in the underlying biological system that may not match the scale-free assumption, or when large differences in expression changes are expected (e.g., cancer vs normal), removing principal components may remove biological signal of interest and, as with any data cleaning methodology, should be used with caution. We have implemented our PC correction approach as a function—"sva_network" in *sva* Bioconductor package which can be used prior to network reconstruction with a range of methods (Vignette: Additional file 4).

Methods

Principal component-based correction of gene expression

Using a permutation-based approach as described in [21], we first determined the number of principal components "p" to correct the data for with the "num.sv" function in the Bioconductor package *sva* (Additional file 1: Table S4). Next, we compute the principal component loadings L of

the standardized expression matrix with singular value decomposition (SVD). Using a linear model, we regressed the top "p" principal components (p as determined by "num.sv") on each gene E_i from the expression data and computed the residuals $\widehat{E_i}$.

$$E_i = \mu_i + \beta_i \times L_{1:p}$$

$$\hat{E}_i = E_i - \left[\mu_i + \left(\beta_i \times L_{1:p}\right)\right]$$

Evaluation of co-expression networks

To evaluate our correction method and its effect on the reconstruction of co-expression networks, we used two methods to infer the structure of gene co-expression networks: (a) weighted gene co-expression networks (WGCNA) [10] and (b) graphical lasso [11] (Additional file 1: Note 2).

Since the underlying network structure is generally unknown, we used genes known to be functional in the same pathways as ground truth to assess these networks.

Any pair of genes that have at least one pathway in common were assumed as a true functional relationship. An edge that was observed between a pair of genes in the inferred network (from WGCNA or graphical lasso) and was also present in the list of real connections was called as a true positive (TP). We defined false positive (FP) to be an edge that was observed between a pair of genes in the inferred network, however was absent in the list of real connections.

- Shared true positives: We obtained a refined list of real connections described above by restricting to pairs of genes that were present in at least two pathway databases.

All TP, FP, and FN were computed with genes restricted to the most variable 5000 genes that were used for reconstructing co-expression networks. We compute the false discovery rate as given below:

$$\mathrm{FDR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}$$

Additional files

Additional file 1: Supplemental methods and results. This file contains theoretical proofs, supplemental methods, results, figures, and tables. (PDF 963 kb)

Additional file 2: Scale-free simulation (R notebook) (HTML 772 kb)

Additional file 3: Scale-free simulation with sample and gene numbers matched to GTEx (R notebook). (HTML 763 kb)

Additional file 4: Tutorial vignette to apply PC correction prior to network reconstruction in an example dataset. (HTML 726 kb)

Abbreviations

WGCNA: Weighted gene co-expression networks

Funding

AB is supported by the NIH R01MH109905, NIH R01GM120167, and NIH R01GM121459. MCS is supported by the NSF DBI-1350041 and NIH R01HG006677. JTL is supported by the NIH R01GM105705 and NIH R01GM121459.

Availability of data and materials

All analyses were performed using R and scripts which are available on GitHub at https://github.com/leekgroup/networks_correction [34].

Authors' contributions

CR, PP, AB, and JL conceived the study. PP, CR, AJ, MS, AB, and JL designed the experiments. CR performed the theoretical analysis. PP and CR performed the simulation experiments. PP performed empirical analyses. PP, CR, AB, and JL wrote the manuscript with inputs from all co-authors. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [2]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [3]Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, USA. [4]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [5]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [6]Center for Computational Biology, Johns Hopkins University, Baltimore, MD, USA. [7]Department of Biology, Johns Hopkins University, Baltimore, MD, USA. [8]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA.

Received: 5 March 2019 Accepted: 24 April 2019
Published online: 16 May 2019

References

1. Yang Y, Han L, Yuan Y, Li J, Hei N, Liang H. Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. Nat Commun. 2014;5:3231.
2. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
3. Furlong LI. Human diseases through the lens of network biology. Trends Genet. 2013;29:150–9.
4. Stegle O, Lippert C, Mooij JM, Lawrence ND, Borgwardt K. Efficient inference in matrix-variate gaussian models with\iid observation noise. Adv Neural Inf Proces Syst. 2011;630–638.
5. Gao C, McDowell IC, Zhao S, Brown CD, Engelhardt BE. Context specific and differential gene co-expression networks via Bayesian biclustering. PLoS Comput Biol. 2016;12:e1004791.
6. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:Article17.
7. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–41.
8. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One. 2011;6:e17238.

9. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3:e161.
10. Freytag S, Gagnon-Bartsch J, Speed TP, Bahlo M. Systematic noise degrades gene co-expression signals but can be corrected. BMC Bioinformatics. 2015;16:309.
11. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010;11:733–9.
12. Akey JM, Biswas S, Leek JT, Storey JD. On the design and analysis of gene expression studies in human populations. Nat Genet. 2007;39:807–8 author reply 808–9.
13. van Noort V, Snel B, Huynen MA. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. EMBO Rep. 2004;5:280–4.
14. Carlson MRJ, Zhang B, Fang Z, Mischel PS, Horvath S, Nelson SF. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 2006;7:40.
15. Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, et al. A gene expression map for *Caenorhabditis elegans*. Science. 2001;293:2087–92.
16. Saha A, Kim Y, Gewirtz ADH, Jo B, Gao C, McDowell IC, et al. Co-expression networks reveal the tissue-specific regulation of transcription and splicing. Genome Res. 2017;27:1843–58.
17. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464:768–72.
18. Flutre T, Wen X, Pritchard J, Stephens M. A statistical framework for joint eQTL analysis in multiple tissues [Internet. PLoS Genetics. 2013:e1003486 Available from: https://doi.org/10.1371/journal.pgen.1003486.
19. Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6:e1000770.
20. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38:904–9.
21. Buja A, Eyuboglu N. Remarks on parallel analysis. Multivariate Behav Res. 1992;27:509–40.
22. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–3.
23. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
24. Hsieh C-J, Sustik MA, Dhillon IS, Ravikumar P. QUIC: quadratic approximation for sparse inverse covariance estimation. J Mach Learn Res JMLR org. 2014; 15:2911–47.
25. Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. 2016;44:W90–7.
26. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.
27. Jaffe AE, Tao R, Norris AL, Kealhofer M, Nellore A, Shin JH, et al. qSVA framework for RNA quality correction in differential expression analysis. Proc Natl Acad Sci U S A. 2017;114:7130–5.
28. Love MI, Hogenesch JB, Irizarry RA. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. Nat Biotechnol. 2016;34:1287–91.
29. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13.
30. Liebhaber SA. mRNA stability and the control of gene expression. Nucleic Acids Symp Ser. 1997:29–32. https://doi.org/10.1038/npg.els.0005972
31. Copois V, Bibeau F, Bascoul-Mollevi C, Salvetat N, Chalbos P, Bareil C, et al. Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. J Biotechnol. 2007;127:549–59.
32. Gallego Romero I, Pai AA, Tung J, Gilad Y. RNA-seq: impact of RNA degradation on transcript quantification. BMC Biol. 2014;12:42.
33. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods. 2017;14:417–9.
34. Parsana P, Ruberman C, Jaffe AE, Schatz MC, Battle A, Leek JT. Addressing confounding artifacts in reconstruction of gene co-expression networks: Zenodo; 2019. Available from: https://doi.org/10.5281/ZENODO.2648667