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&], Jeffery T. Leek[2,6&]

[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

* Equal Contribution

[&] Corresponding authors: Alexis Battle - ajbattle@cs.jhu.edu; Jeffery T. Leek - jtleek@gmail.com

Gene co-expression networks can capture biological relationships between genes, and are important tools in predicting gene function and understanding disease mechanism. We show that artifacts such as batch effects in gene expression data confound commonly used network reconstruction algorithms. We then demonstrate, both theoretically and empirically, that principal component correction of gene expression measurements prior to network inference can reduce false discoveries. Using expression data from the GTEx project in multiple tissues and hundreds of individuals, this approach improves precision and recall in the networks reconstructed.

Groups of genes are function together to perform distinct cellular processes, which are often supported by coordinated expression of functionally related genes. . Based on this, gene co-expression networks seek to identify transcriptional patterns that are indicative of functional interactions and regulatory relationships between genes[1–3]. The true *in vivo* functional interactions between genes are not fully characterized for most species, tissues, and disease-relevant contexts. Therefore reconstruction of co-expression networks from high throughput measurements is of common interest. However, accurate reconstruction of such networks remains a challenging problem.

A co-expression network is an undirected graph where genes are represented as nodes, and a functional relationship between genes is represented as an edge between the nodes - such as transcription factors to their gene targets or pairs of genes within a shared pathway. Widely used network learning methods such as Weighted Gene Coexpression Network Analysis (WGCNA)[4] and graphical lasso[5] are based on pairwise associations between genes[4,5]. Typically, these methods do not account for confounding factors such as batch effects that routinely affect measurements of high-dimensional gene expression data (microarray and RNA-seq)[6]. However, both biological and technical artifacts are known to influence expression measurements of gene expression data, sometimes substantially, introducing spurious signals, including false correlations between genes not reflective of a functional relationship[7,8]. Some of the major confounders known to impact RNA-seq measurements include RNA integrity number (RIN), sequencing library size, mapping artifacts, and GC bias, among others[9–12]. During network reconstruction, correlations introduced by these confounders are often inferred as relationships between genes, leading to inaccurate network structure and erroneous conclusions in downstream analyses[7,13,14]. 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.

In this study, we leverage the framework of scale-free networks to provide a framework for data correction. It has been shown that real world networks including co-expression networks have scale-free topology, i.e. the node degree distribution of these networks follow power laws[15–17]. These networks are characterized by a small number of influential hub nodes that link to the remainder of the lower degree nodes. This makes scale-free networks more stable and more robust to random perturbations[18].[4,19–22]. Several studies have employed the assumption of scale-free topology to infer high-dimensional gene co-expression and splicing networks[4,23]. Here, 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. The number of principal components can be estimated in multiple ways, here we use a permutation based scheme[24] for estimating the number of principal components as implemented in the sva package[25]. We prove that under a scale free network that these principal components can then be removed without affecting the signal arising from the true gene network , enabling improved network reconstruction (**Supplementary Note 1**).

Using a small simulated example, we illustrate how confounders in gene expression data can impact reconstruction of co-expression networks and how this can be corrected (**Figure 1**). We used three versions of data -- (a) simulated gene expression data with no batch effects (**Figure 1a**); (b) simulated data with a batch effect added(**Figure 1d**); and (c) confounded data from (b) corrected by regressing out the top principal component (**Figure 1g**). With each version of the data, we first computed empirical correlation matrix (**Figure 1b,e,h**). Next, we reconstructed co-expression networks using graphical lasso[5,26] (**Supplementary Note 2.2.2**). Confounders in the simulated data that affected genes 2 through 6 was evident through the block pattern in the

data matrix (**Figure 1d**). Likewise a large block of high (spurious) correlation between the same genes was observed in the empirical correlation matrix (**Figure 1e**). Moreover all genes that were affected by confounders were connected to each other in the inferred network while two true dependencies *E(3,1)* and *E(4,7)* were lost (**Figure 1f**).

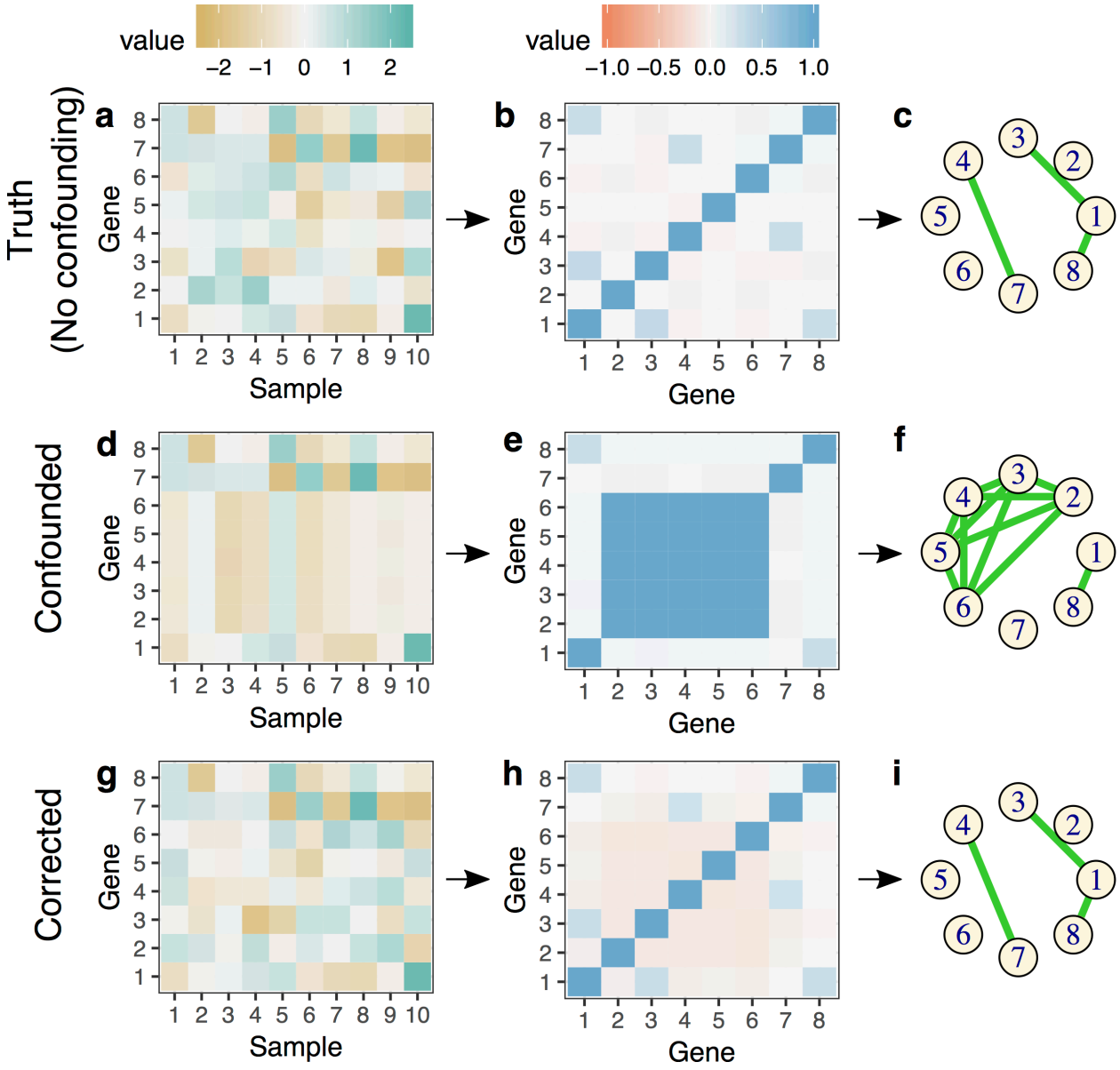

Figure 1. A simulation example. This simulation example shows reconstruction of gene co-expression networks is affected by confounders. True underlying network structure can be reconstructed after principal component correction of gene expression data as described in the paper.

Next we corrected the confounded data using a linear model with the top principal component in the confounded data as an explanatory variable. We then reconstructed the network using the residuals from this regression (see

Supplementary Note 2.1) The structure and pattern of heatmap and correlation matrices of the corrected data resembled the original simulated data with some additional negative correlation (**Figure 1a-b and g-h**). Additionally, graphical lasso correctly estimated the network structure obtained from corrected data, which was same as the true network structure that was obtained from the original simulated data (**Figure 1c and i**).

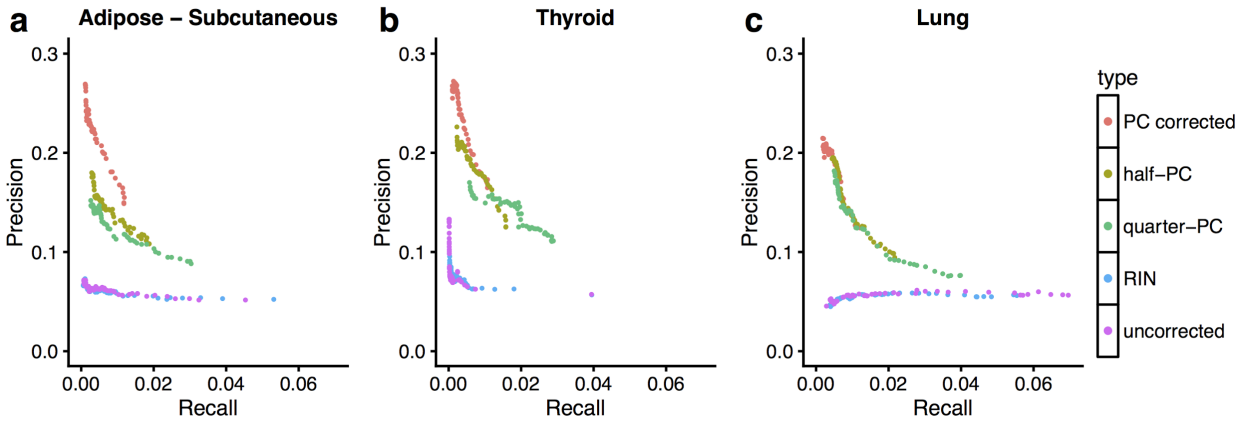

Figure 2. Precision-Recall curves of WGCNA modules based on canonical pathways. Precision and recall curves of WGCNA networks obtained at a varying cut-heights. Each point corresponds to the network obtained at a specific cut-height

To demonstrate the effect of latent confounders and principal component correction on reconstruction of co-expression networks from large-scale human gene expression measurements, we applied our method to RNA-seq data from the Genotype Tissue Expression (GTEx) project. We considered data from five diverse tissues containing between 278 and 361 samples each: Adipose Subcutaneous, Lung, Skeletal Muscle, Thyroid and Whole Blood. We used two popular methods for network reconstruction: (a) weighted gene co-expression network analysis[4,27] (WGCNA, **Supplementary Note 2.2**), and (b) graphical lasso[5,26] (**Supplementary Note 2.2**). Since the true underlying co-expression network structure is not known, we assessed the networks using genes annotated to function in the same pathways, and transcription factor with known target genes as the ground truth (**Supplementary Note 2.4.3**).

We obtained the most recent version (2016) of curated gene sets annotated as canonical pathways on MsigDB[28] available on the Enrichr library, containing information from KEGG, Reactome, Biocarta and Pathway Interaction Database. The set of transcription factors along with known target genes were obtained from ChIP Enrichment Analysis (ChEA 2016)[29–31]. Precision and recall curves (**Figure 2**) were used to evaluate and compare the performance of co-expression networks inferred from data after correcting for latent confounders to networks inferred using a) uncorrected expression data, b) the residuals after regressing out RNA integrity number (RIN), c) exonic rate - a mapping covariate that corresponds to fraction of reads mapped to exons, and d) sample specific estimate of GC bias, all shown to be common confounders in mRNA gene expression data[10,32–35].

WGCNA identifies groups of genes that form coexpressed modules of genes based on a power transform of the pearson correlation coefficient for all pairs of genes[4]. For each tissue we inferred weighted, unsigned co-expression networks using the most variable 5000 genes (**Supplementary Note 2.4**). Co-expression gene modules were identified based on fully-connected sub-graphs of the network. We observed that precision and recall curves from module assignments obtained from data corrected for latent confounders performed considerably better than those obtained from uncorrected data, or from correcting for either RIN, exonic rate (a quality metric from RNA-seq mapping), or sample-specific GC bias (**Figure 2, Supplementary Figure 1a-b, 2, 5, 7, and 9, Supplementary Note 2.3**). Note that precision and recall are both low on an absolute scale, regardless of the choice of method. It is of standard co-expression network reconstruction method to observe a high false discovery rate despite enrichment in aggregate for biologically meaningful relationships. This suggested that RIN, exonic rate, and GC bias alone were not a sufficient surrogate for the diverse sources of confounding variation in gene expression data. Since broad trends in co-expression may sometimes reflect distant regulatory relationships between genes[23] , to ensure that we are not removing true long range signals, we also reconstructed networks with data corrected for one quarter and half the number of PCs estimated by our correction method. However, we found the fully PC corrected networks reconstructed with WGCNA overall showed improved precision without compromising on recall (**Figure 2, Supplementary Figure 1a-b, 2, 5, 7, and 9, Supplementary Note 2.4**).

Similarly we examined the effect of confounders on networks reconstructed with graphical lasso using the same 5000 most variable expressed genes across all tissues. To test the effect of sparsity we also varied the penalty parameter in graphical lasso (λ=[0.3,1.0]). For each tissue, using the non-zero entries in the estimated precision matrix as the edges of the graph, we computed precision and recall for the inferred networks (**Supplementary Note 2.4**).

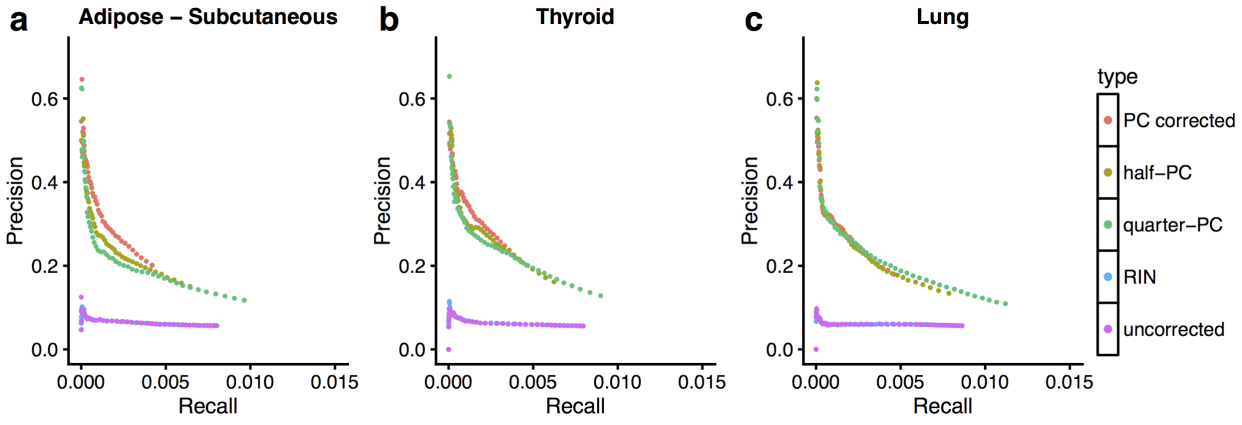

Figure 3. Precision-Recall curves of networks inferred with graphical lasso based on canonical pathways. Each point in the figure corresponds to precision and recall of networks obtained at a specific L1 penalty parameter value (penalty parameter ranges from 0.3 to 1.0, x-axis).

In adipose subcutaneous and thyroid tissues, networks estimated with principal component corrected data showed higher precision compared to the networks estimated with fewer PC corrected, uncorrected or RIN corrected data (**Figure 3a-b**). In Lung, we observed that in general PC corrected networks showed improved performance on precision and recall, however, networks obtained from one quarter, half or full PC corrected data did not show any significant differences within themselves (**Figure 3c**). Nevertheless, for all tissues, it was observed that some degree of principal component correction always improved the performance of graphical lasso. Similar to WGCNA, in all tissues, we also observed that 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 is not a sufficient alternative for the wide range of confounding variation found in gene expression data (**Figure 3, Supplementary Figure 1c-d, 3, 6, 8, and 10**). Additionally, we also found that networks inferred from principal component corrected data were much more sparse compared to uncorrected, and RIN, exonic rate or GC bias corrected counterparts (**Figure 4**).

Overall our analysis shows that network reconstruction methods are vulnerable to latent confounders present in gene expression data. The simulation study demonstrates that graph estimation methods that do not account for confounders make a large number of false discoveries that may be at the expense of losing true dependencies. Similarly, in empirical analysis using GTEx data we see that the networks inferred from the expression data without any correction methods performed poorly compared to principal component corrected data. In addition, co-expression networks obtained from expression data corrected for effects of RIN,exonic rate, or GC bias show little improvement in precision and recall compared to uncorrected data, thereby suggesting that correcting gene expression data for known artifacts such as RIN,exonic rate or GC bias does not fully eliminate patterns of confounding variation from the data, even when these variables are collected. However, we do note that for particularly dense or connected sub-graphs in the underlying biological system that don't match the small-world assumption, removing principal components may remove relevant biological signal and, as with any data cleaning methodology, should be used with caution. We have implemented our PCA cleaning approach in the *sva* Bioconductor package which can be used prior to network reconstruction with a range of methods.

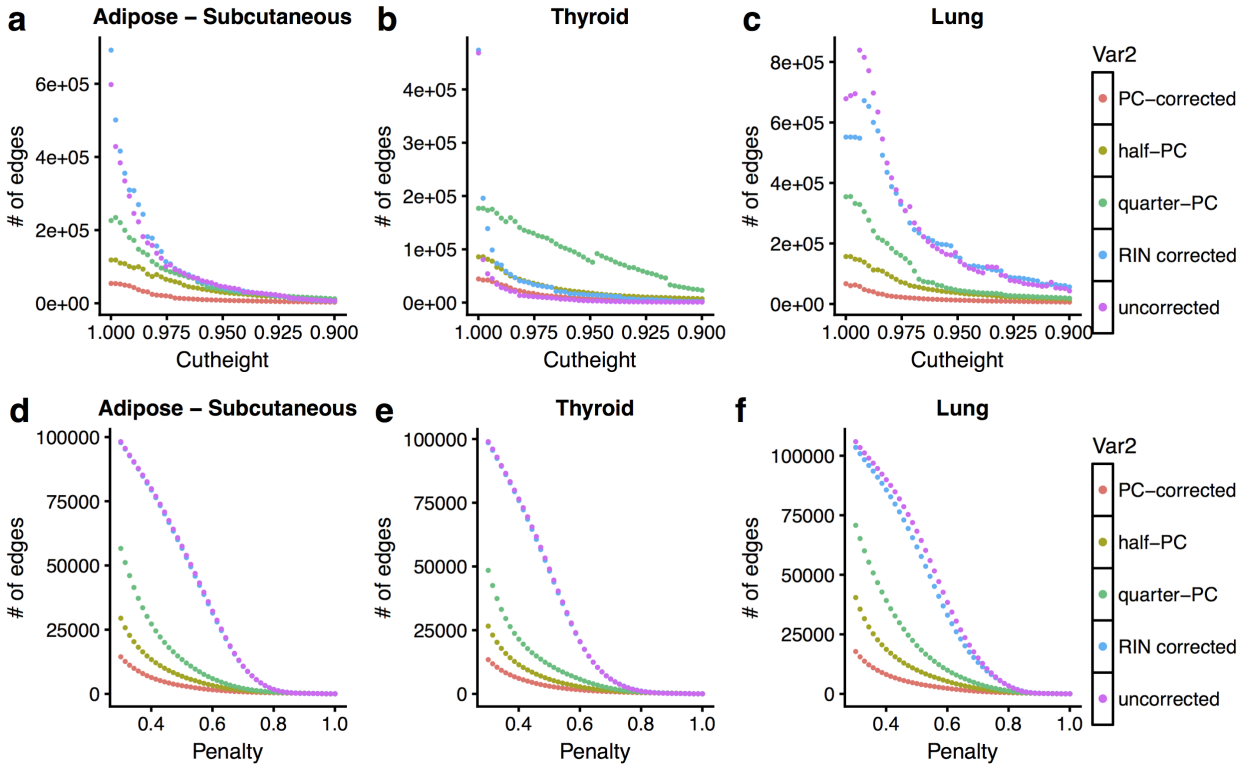

Figure 4. Density of networks inferred from PC-corrected data are sparser. a-c Each point corresponds to number of edges in networks inferred by WGCNA at a cut-height. d-e Each point corresponds to number of edges inferred by graphical lasso in networks obtained at a specific L1 penalty parameter value. In most cases, we observe that networks inferred by PC corrected data have fewer edges compared to uncorrected or RIN corrected data.

Acknowledgements

We would like to thank Ashis Saha for providing information on the resource for transcription factors and known targets. AB is supported by NIH R01MH109905, NIH R01GM120167, and NIH R01GM121459. MCS is supported by NSF DBI-1350041 and NIH R01HG006677. JTL is supported by NIH R01GM105705 and NIH R01GM121459.

Competing financial interests

Authors declare no competing financial interests

References

1. Yang, Y. *et al.* Gene co-expression network analysis reveals common system-level properties of prognostic genes across cancer types. *Nat. Commun.* **5,** 3231 (2014).
2. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12,** 56–68 (2011).

3. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* **29,** 150–159 (2013).
4. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4,** Article17 (2005).
5. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* **9,** 432–441 (2008).
6. Freytag, S., Gagnon-Bartsch, J., Speed, T. P. & Bahlo, M. Systematic noise degrades gene co-expression signals but can be corrected. *BMC Bioinformatics* **16,** 309 (2015).
7. Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6,** e17238 (2011).
8. Leek, J. T. & Storey, J. D. Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLoS Genet.* **3,** e161 (2007).
9. Jaffe, A. E. *et al.* qSVA framework for RNA quality correction in differential expression analysis. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 7130–7135 (2017).
10. Love, M. I., Hogenesch, J. B. & Irizarry, R. A. Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation. *Nat. Biotechnol.* **34,** 1287–1291 (2016).
11. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biol.* **17,** 13 (2016).
12. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550,** 204–213 (2017).
13. Leek, J. T. *et al.* Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11,** 733–739 (2010).
14. Akey, J. M., Biswas, S., Leek, J. T. & Storey, J. D. On the design and analysis of gene expression studies in human populations. *Nat. Genet.* **39,** 807–8; author reply 808–9 (2007).
15. van Noort, V., Snel, B. & Huynen, M. A. The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep.* **5,** 280–284 (2004).
16. Carlson, M. R. J. *et al.* Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics* **7,** 40 (2006).
17. Kim, S. K. *et al.* A gene expression map for Caenorhabditis elegans. *Science* **293,** 2087–2092 (2001).
18. Hu, J. X., Thomas, C. E. & Brunak, S. Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.* **17,** 615–629 (2016).
19. Barabási, A. Emergence of Scaling in Random Networks. *Science* **286,** 509–512 (1999).
20. Cohen, R., Erez, K., ben-Avraham, D. & Havlin, S. Resilience of the internet to random breakdowns. *Phys. Rev. Lett.* **85,** 4626–4628 (2000).
21. Newman, M. E. J. The structure and function of complex networks. in *SIAM REVIEW* (2003).
22. Nguyen, K. & Tran, D. A. Fitness-Based Generative Models for Power-Law Networks. in *Handbook of Optimization in Complex Networks* (eds. Thai, M. T. & Pardalos, P. M.) **57,** 39–53 (Springer US, 2012).

23. Saha, A. *et al.* Co-expression networks reveal the tissue-specific regulation of transcription and splicing. (2016). doi:10.1101/078741
24. Buja, A. & Eyuboglu, N. Remarks on Parallel Analysis. *Multivariate Behav. Res.* **27,** 509–540 (1992).
25. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28,** 882–883 (2012).
26. Hsieh, C.-J., Sustik, M. A., Dhillon, I. S. & Ravikumar, P. QUIC: quadratic approximation for sparse inverse covariance estimation. *J. Mach. Learn. Res.* **15,** 2911–2947 (2014).
27. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
28. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27,** 1739–1740 (2011).
29. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14,** 128 (2013).
30. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44,** W90–7 (2016).
31. Lachmann, A. *et al.* ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics* **26,** 2438–2444 (2010).
32. Liebhaber, S. A. mRNA stability and the control of gene expression. *Nucleic Acids Symp. Ser.* 29–32 (1997).
33. Copois, V. *et al.* Impact of RNA degradation on gene expression profiles: assessment of different methods to reliably determine RNA quality. *J. Biotechnol.* **127,** 549–559 (2007).
34. Gallego Romero, I., Pai, A. A., Tung, J. & Gilad, Y. RNA-seq: impact of RNA degradation on transcript quantification. *BMC Biol.* **12,** 42 (2014).
35. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14,** 417–419 (2017).