JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 25, Number 6, 2018

Pp. 606–612
DOI: 10.1089/cmb.2017.0262

Simple Comparative Analyses of Differentially Expressed Gene Lists May Overestimate Gene Overlap

CHELSEA M. LAWHORN,[1,*,†] RACHEL SCHOMAKER,[2,†,‡]
JONATHAN T. ROWELL,[3] and OLAV RUEPPELL[4]

ABSTRACT

Comparing the overlap between sets of differentially expressed genes (DEGs) within or between transcriptome studies is regularly used to infer similarities between biological processes. Significant overlap between two sets of DEGs is usually determined by a simple test. The number of potentially overlapping genes is compared to the number of genes that actually occur in both lists, treating every gene as equal. However, gene expression is controlled by transcription factors that bind to a variable number of transcription factor binding sites, leading to variation among genes in general variability of their expression. Neglecting this variability could therefore lead to inflated estimates of significant overlap between DEG lists. With computer simulations, we demonstrate that such biases arise from variation in the control of gene expression. Significant overlap commonly arises between two lists of DEGs that are randomly generated, assuming that the control of gene expression is variable among genes but consistent between corresponding experiments. More overlap is observed when transcription factors are specific to their binding sites and when the number of genes is considerably higher than the number of different transcription factors. In contrast, overlap between two DEG lists is always lower than expected when the genetic architecture of expression is independent between the two experiments. Thus, the current methods for determining significant overlap between DEGs are potentially confounding biologically meaningful overlap with overlap that arises due to variability in control of expression among genes, and more sophisticated approaches are needed.

Keywords: differentially expressed genes, gene regulation, genetic architecture, transcription factor binding sites, transcriptome.

[1]Department of Mathematics, Winthrop University, Rock Hill, South Carolina.
[2]Department of Biology, Florida Southern College, Lakeland, Florida.
Departments of [3]Mathematics and Statistics and [4]Biology, University of North Carolina at Greensboro, Greensboro, North Carolina.
*Current address: Department of Statistics, North Carolina State University, Raleigh, North Carolina.
‡Current address: Department of Biological Sciences, University of South Carolina, Columbia, South Carolina.
†Shared first authorship.

1. INTRODUCTION

Through technological revolutions in molecular biology, the analysis of entire transcriptomes of cells, tissues, or organisms has become commonplace in many areas of biology. Gene expression levels are quantified as the number of messenger RNA transcripts of a particular gene, most commonly by hybridization to a microarray or high-throughput sequencing of the RNA pool of a sample (Kerr et al., 2000; Rabbani et al., 2013). Typically, genome-wide expression levels are compared between two experimental groups or targets to identify differentially expressed genes (DEGs). Due to the many parallel tests, a false discovery rate (FDR) is used to determine which of the genes can be considered significant DEGs (Benjamini and Hochberg, 1995). The rapid growth of DEG data sets allows the comparison of transcriptome changes across different experiments or studies to connect different biological processes if significant overlap in the DEG lists is found.

Overlap between two DEG lists is typically determined by comparing the number of DEGs that co-occur in both lists to the number of DEGs that is expected to co-occur in both lists by chance, given the number of DEGs and genes that are not differentially expressed in each circumstance. This simple logic and test are found extensively throughout the literature, with examples spanning human disease comparisons (Koth et al., 2011), genomic analyses of plant stress responses (Rasmussen et al., 2013), and transcriptome studies of social behavior (Toth et al., 2014; Mondet et al., 2015), and cross-species comparisons in aging (von Wyschetzki et al., 2015) and mental health (Shpigler et al., 2017).

However, not all genes have evolved equally. Genes vary in their regulatory architecture, such as the number of transcription factor binding sites (TFBS), and network analyses demonstrate that some genes are consistently at the center of transcriptome modulation, while others are at the periphery (Stuart et al., 2003). The number and nature of TFBS regulate how flexible the transcription of a particular gene responds to changes in the cellular environment (Wray, 2007). Consequently, the expression of some genes is predicted to be consistently more variable than others.

This heterogeneity represents a theoretical problem to the abovementioned approach to compare DEG lists by simple probability tests, which treat all genes as equal, because more flexible genes will co-occur in DEG lists more often than stable ''housekeeping'' genes (Eisenberg and Levanon, 2003). This problem needs to be recognized and overcome with entirely new approaches (Doublet et al., 2017) or at least minimized by contrasting multiple DEG list comparisons (von Wyschetzki et al., 2015; Zanni et al., 2017) to demonstrate that particular comparisons are specifically indicating biological overlap and not merely a consequence of the regulatory differences among genes.

In this study, we present results from a comprehensive simulation study that documents how variability in number and nature of TFBS of individual genes affect the degree of overlap between two random DEG lists. The results indicate that regulatory variability among genes leads to a systematic bias for concluding that DEG lists are more similar than expected by chance. Thus, we recommend caution when using this approach and recommend the development of more sophisticated procedures of comparing transcriptome profiles to study specific biological similarities.

2. METHODS

The objective is to generate two lists of DEGs with a simple computer simulation to determine their degree of random overlap, assuming variable control of individual gene expression. These results can then be compared with the conventional test for DEG list overlap used in empirical studies. Simulations are performed in MATLAB (annotated code available as Supplementary S1).

The study is designed hierarchically. Different scenarios are created to evaluate genomes with different numbers of genes (g) and maximal numbers of TFBS (n). There are s distinct types of TFBS and t types of factors. We assume that $s = t$. Two relationships between transcription factors and TFBS are considered. Either all affinities between transcription factors and TFBS are determined randomly, resulting in a probabilistic binding of transcription factors to different sites, or transcription factors bind only to one type of TFBS, representing perfect correspondence between one specific transcription factor and one specific TFBS.

For each scenario, 25 independent trials are performed. Each trial consists of 1000 replicates, each with a unique genetic architecture matrix ($T = g \times s$, where T_{ij} quantifies the number of a specific TFBS j located in the promoter of a specific gene i) and a vector Q containing the quantities of each transcription factor present. T and Q are randomly generated with T_{ij} ranging from 0 to n and Q_j ranging from 0 to $n \times g$. T and

Q are kept constant within each replicate for the two experiments that generate the two DEG lists to be compared. In each experiment, a matrix of activated transcription factors ($A = 2 \times Q$) is generated by determining a random proportion of Q_j for experimental control and treatment condition independently. The probability of occupancy is calculated for each TFBS by dividing the number of activated transcription factors that could bind to this TFBS by the total number of corresponding available TFBS. Subsequently, the summed probabilities of all TFBS for a gene determine its average level of gene expression under a particular experimental condition. The influence of the two experimental conditions on gene expression patterns is therefore statistically independent but influenced by the quantities of transcription factors and the numbers of TFBS for each gene.

For each condition in each experiment, 50 individuals are simulated and individual variation is taken into account by adding error terms that are randomly drawn from a normal distribution to the expression level of each gene. These error terms are scaled to produce a proportion of DEGs between treatment groups in each experiment that matches empirical studies. DEGs are identified by standard t-tests using an FDR = 0.05 as statistical cutoff. The probability of the degree of overlap between the DEG lists of the pairs of experiments in each replicate is determined by Fisher's exact test, and the resulting probabilities (p-values) are compared to the theoretical expectation (set to the standard probability threshold of $p = 0.05$) across all 1000 replicates for each trial.

As a negative control, the same simulations are performed without sharing the genetic architecture (T and Q) between the two paired experiments. For these control simulations, the TFBS for each gene and all other variables are generated as described above, except that the TFBS and transcription factor quantities are generated for each experiment separately.

3. RESULTS

The degree of overlap between the two simulated DEG lists varies substantially across replicates within trials for most scenarios, with corresponding probabilities ranging from many orders of magnitude below conventional significance criteria to the theoretical maximum of 1 (e.g., Fig. 1). Results indicate in the majority of scenarios that p-values below 0.05 arise by chance much more often than the expected 5% of all replicates. The overlap between DEG lists is strongest in scenarios with relatively few TFBS per gene and a large number of genes (Fig. 2).

While increasing gene number increases the degree of overlap under all circumstances, the effect of the maximum number of TFBS is not monotonous for all scenarios. With 1000 or 10,000 genes, more overlap results with 16 than with 4 TFBS in 3 of 4 different scenarios and the overlap decreases thereafter in scenarios with 40 or 400 different kinds of TFBS (Fig. 2).

The affinity between transcription factors and their binding sites has the most profound impact on the results. When these affinities are randomly generated for each pairing of transcription factor and TFBS,

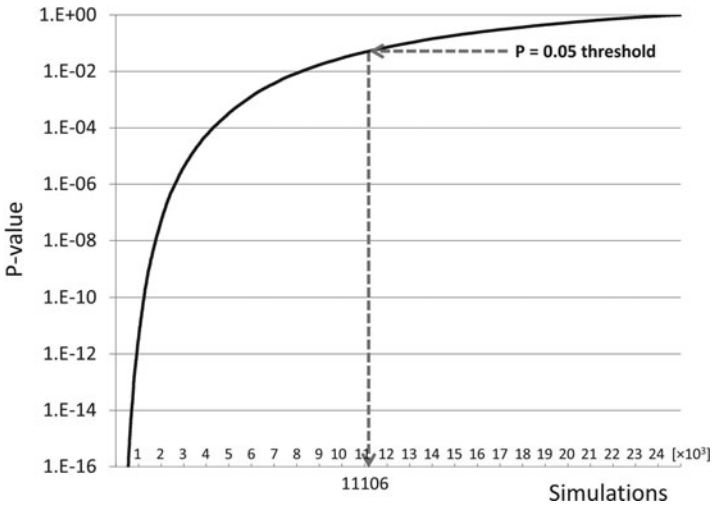

FIG. 1. Distribution of probabilities (p-values) of the overlap between two randomly generated lists of DEGs. Data from 25 trials with different gene regulatory structures that were each replicated 1000 times indicated that DEG lists can significantly overlap much more often by chance than commonly assumed. The discrepancy in this case for a maximum of 40 TFBS per gene and 10,000 genes arises because genes with many TFBS are inherently more responsive to environmental or experimental perturbation than genes with fewer TFBS. DEG, differentially expressed gene; TFBS, transcription factor binding sites.

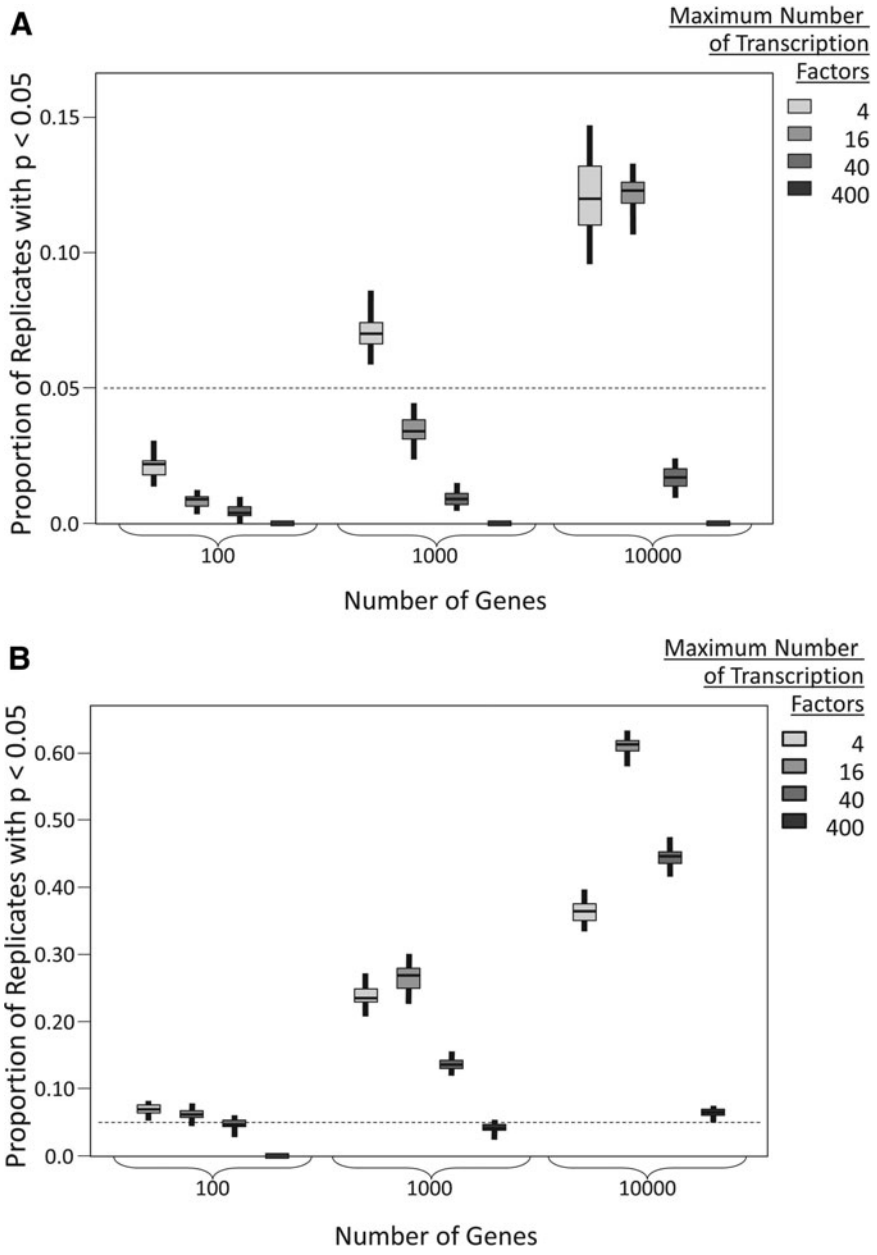

FIG. 2. All results of assessing overlap of transcriptome patterns between two experiments. Overall, the data revealed that a high proportion of simulations result in significant overlap of DEG lists between two experiments, even though gene expression patterns were randomly generated within the constraints of specific distributions of TFBS and numbers of available transcription factors. When no specific affinity of transcription factors to their binding sites is assumed, the proportion of simulations leading to random overlap is relatively low **(A)**, but when transcription factors only bind to their respective binding site **(B)**, our simulation results indicated a high chance of falsely concluding significant overlap among two lists of DEGs.

most of the simulations indicate lower proportions of simulations that resulted in significant overlap than expected, and in some cases, no single instance of significant overlap is observed (Fig. 2A). However, when transcription factors bind only to their corresponding TFBS, significant overlap between DEG lists is observed considerably more often than in 5% of the cases (Fig. 2B), with the median among trials in one scenario over 61%.

When the genetic architectures are separately generated between the two experiments, significant overlap between DEG lists of the experiments is observed in less than 0.1% of the simulations with perfect

correspondence between transcription factors and TFBS and not at all in simulations with randomly generated affinities between transcription factors and particular TFBS.

4. DISCUSSION

The dramatic increase of functional genomic studies that compare transcriptome profiles of cells, tissues, or organisms between physiological states, experimental treatments, or environmental conditions has resulted in a host of lists of DEGs. These data provide a rich community resource and are often compared among multiple experiments or studies to infer whether the biological phenomena under study are mechanistically related. Interspecific comparisons likewise use this logic to infer the evolutionary conservation of molecular pathways. Many studies determine the significance of overlap between DEGs by Fisher's exact test (Koth et al., 2011) or hypergeometric test (Mondet et al., 2015). Our study challenges this common practice of testing the overlap between DEG lists by simply comparing the degree of empirical overlap to the overlap expected by chance. The simulation results indicate that such practice will wrongly conclude nonrandom overlap between DEG lists in many cases. Our central conclusion arises because the number of DEGs that overlap by chance is commonly underestimated by assuming that all genes are identical in their degree of expression control. This assumption is simplistic because gene expression is complex and variable (De Jong, 2002).

By implementing variable numbers and types of transcription factors, our simulation takes some of the complexity of the control of gene expression into account and demonstrates that this complexity can confound estimates of biological overlap between experiments or studies. This conclusion is particularly strong for the parameter space that is biologically most relevant. If transcription factors had random affinities for different TFBS, the differences in TFBS among genes would not intrinsically generate overlap between DEG lists. However, most transcription factors have quite specific binding affinities to TFBS (Wasserman and Sandelin, 2004) and in this case, the differences in genetic architecture among genes do generate the problem described above.

Moreover, our systematic variation of genes and transcription factors (and different TFBS by extension) reveals that the illusory overlap between lists of DEGs is particularly strong in scenarios with many genes. The random simulations frequently achieve significant overlap in case of the selected maximum of 10,000 genes, actually representing a lower estimate of most Metazoan genomes (Elsik et al., 2014; Ezkurdia et al., 2014). It is reasonable to assume that the higher gene numbers of many genomes would increase the overlap. The maximum number of different transcription factors used also represents a lower estimate for Metazoans (Brivanlou and Darnell, 2002; Hammonds et al., 2013). In this case, a further increase would probably decrease the overlap. It is unlikely that a simultaneous increase in both variables would invalidate our main conclusions and computational requirements limit the evaluated parameter space.

Even though our study is the first explicit demonstration, it has already been recognized that simple comparisons of DEG lists are problematic and several studies have begun to work around this problem. One feasible approach is to conduct multiple comparisons between DEG lists to contrast the overlap in question with the degree of overlap in other comparisons that share the same genetic architecture but differ in the degree of biological commonalities (Shpigler et al., 2017; Zanni et al., 2017). New methods to identify overlap are urgently needed and some methodological developments have occurred (Doublet et al., 2017; Shpigler et al., 2017). However, these new approaches are not yet sufficiently addressing the problem indicated by our findings, currently leaving investigators with generating multiple DEG list comparisons to evaluate the specificity of their findings. Therefore, current findings of gene overlap based on simple statistical tests have to be regarded with some caution.

A statistical method that takes the variable architecture of expression regulation of individual genes into account is the theoretical solution. Unfortunately, our understanding of the genetic architecture of gene control is insufficient in most biological systems to perform this more accurate testing. In some systems, the mining of combined quantitative trait locus (QTL) and transcriptome (e-QTL) data (Xia et al., 2011; Lappalainen, 2015) should prove useful in this regard. In addition, empirical measures on the variability of the expression of individual genes can be accumulated from the large number of transcriptome studies when taking individual sample variation into account instead of dismissing it as experimental noise. If these data are analyzed appropriately into gene-specific variability indices and deposited into publicly accessible databases, more biologically meaningful tests of genetic overlap between sets of DEG lists will be feasible. In addition, the search

for appropriate reference genes in targeted gene studies (Radonić et al., 2004) will also be facilitated by a database of overall gene expression variability.

ACKNOWLEDGMENTS

We thank the U.S. National Science Foundation (grant no. 1359187) and the U.S. National Institutes of Health (grant no. R15GM102753) for financially supporting this project. Furthermore, we thank all of the other Mathematical Biology REU students, Quinn Morris, and Catherine Payne for their support and guidance along the way.

AUTHORS' CONTRIBUTIONS

C.M.L. and R.S. wrote the first draft of this article and contributed to the creation of the simulation code in MATLAB. J.T.R. contributed to the revision process of the article and revised the simulation code in MATLAB. O.R. conceived the project and the study plan, served as biological advisor during the project implementation, and wrote the final version of the article.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* 57, 289–300.

Brivanlou, A.H., and Darnell, J.E. 2002. Signal transduction and the control of gene expression. *Science* 295, 813–818.

De Jong, H. 2002. Modeling and simulation of genetic regulatory systems: A literature review. *J. Comp. Biol.* 9, 67–103.

Doublet, V., Poeschl, Y., Gogol-Döring, A., et al. 2017. Unity in defence: Honeybee workers exhibit conserved molecular responses to diverse pathogens. *BMC Genomics* 18, 207.

Eisenberg, E., and Levanon, E.Y. 2003. Human housekeeping genes are compact. *Trends Genet.* 19, 362–365.

Elsik, C., Worley, K., Bennett, A., et al. 2014. Finding the missing honey bee genes: Lessons learned from a genome upgrade. *BMC Genomics* 15, 86.

Ezkurdia, I., Juan, D., Rodriguez, J.M., et al. 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol. Genet.* 23, 5866–5878.

Hammonds, A.S., Bristow, C.A., Fisher, W.W., et al. 2013. Spatial expression of transcription factors in *Drosophila* embryonic organ development. *Genome Biol.* 14, R140.

Kerr, M.K., Martin, M., and Churchill, G.A. 2000. Analysis of variance for gene expression microarray data. *J. Comp. Biol.* 7:819–837.

Koth, L.L., Solberg, O.D., Peng, J.C., et al. 2011. Sarcoidosis blood transcriptome reflects lung inflammation and overlaps with tuberculosis. *Am. J. Respir. Crit. Care Med.* 184, 1153–1163.

Lappalainen, T. 2015. Functional genomics bridges the gap between quantitative genetics and molecular biology. *Genome Res.* 25, 1427–1431.

Mondet, F., Alaux, C., Severac, D., et al. 2015. Antennae hold a key to *Varroa*-sensitive hygiene behaviour in honey bees. *Sci. Rep.* 5, 10454.

Rabbani, B., Tekin, M., and Mahdieh, N. 2013. The promise of whole-exome sequencing in medical genetics. *J. Hum. Genet*. 59, 5–15.

Radonić, A., Thulke, S., Mackay, I.M., et al. 2004. Guideline to reference gene selection for quantitative real-time PCR. *Biochem. Biophys. Res. Commun.* 313, 856–862.

Rasmussen, S., Barah, P., Suarez-Rodriguez, M.C., et al. 2013. Transcriptome responses to combinations of stresses in *Arabidopsis*. *Plant Physiol*. 161, 1783–1794.

Shpigler, H.Y., Saul, M.C., Corona, F., et al. 2017. Deep evolutionary conservation of autism-related genes. *Proc. Nat. Acad. Sci. U. S. A.* 114, 9653–9658.

Stuart, J.M., Segal, E., Koller, D., and Kim, S.K. 2003. A gene-coexpression network for global discovery of conserved genetic modules. *Science* 302, 249–255.

Toth, A.L., Tooker, J.F., Radhakrishnan, S., et al. 2014. Shared genes related to aggression, rather than chemical communication, are associated with reproductive dominance in paper wasps (*Polistes metricus*). *BMC Genomics* 15, 75.

von Wyschetzki, K., Rueppell, O., Oettler, J., and Heinze, J. 2015. Transcriptomic signatures mirror the lack of the fecundity/longevity trade-off in ant queens. *Mol. Biol. Evol.* 32, 3173–3185.

Wasserman, W.W., and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet*. 5, 276–287.

Wray, G.A. 2007. The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet*. 8, 206–216.

Xia, K., Shabalin, A.A., Huang, S., et al. 2011. seeQTL: A searchable database for human eQTLs. *Bioinformatics* 28, 451–452.

Zanni, V., Galbraith, D.A., Annoscia, D., et al. 2017. Transcriptional signatures of parasitization and markers of colony decline in *Varroa*-infested honey bees (*Apis mellifera*). *Ins. Biochem. Mol. Biol*. 87, 1–13.

Address correspondence to:
Dr. Olav Rueppell
Department of Biology
University of North Carolina at Greensboro
321 McIver Street
Greensboro, NC 27403

E-mail: olav_rueppell@uncg.edu