Biological Psychiatry

Correspondence

Multiple Comparisons and Inappropriate Statistical Testing Lead to Spurious Sex Differences in Gene Expression

To the Editor:

The substantial differences in the incidence and symptoms of stress, depression, and addiction between males and females have motivated studies of sex differences in the brain's molecular responses to environmental stimuli. In a study of three mouse brain regions from the Nestler laboratory, Walker *et al.* (1) reported sex-specific transcriptional responses to cocaine and baseline sex differences. They further examined the impact of social isolation on these brain gene expression patterns. However, the transcriptome data (RNA sequencing) and analysis in this study do not support the authors' conclusions. Here, I describe two critical errors that invalidate the statistical analyses: lack of correction for multiple comparisons and drawing positive conclusions from negative (inconclusive) statistical results. I further show how a valid analysis could be performed for these data.

It is well established that transcriptome-wide analyses of gene expression must account for multiple statistical comparisons to avoid a high rate of false discoveries (2,3). Walker *et al.* defined differentially expressed (DE) genes using a nominal *p*-value threshold ($p < .05$) and fold change >1.3, with no adjustment for multiple comparisons. With these criteria, around 5% of the ∼16,000 genes that they tested would be called DE, even if no true expression differences are present. Thus, around 800 genes could be expected to contribute false positive detections. Importantly, the arbitrary fold-change threshold does not control the risk of false detections. It is standard to report DE genes with adjusted *p* value smaller than a threshold (e.g., .05 or .1), which controls the false discovery rate (<5% or 10%, respectively). Using these criteria, most (71%) of the genes reported to be DE are not significant (adjusted $p > .1$) (Figure 1A). For 39 of the 48 reported comparisons (pairs of conditions), over 95% of the reported DE genes are not significant (Figure 1B).

To illustrate the problem, Figure 2E in Walker *et al.* shows 252 genes that are enriched in males versus females in the prefrontal cortex region of group-housed mice treated with acute cocaine. However, only 12 genes survive the multiple-comparison correction, of which 11 are Y-linked genes (*Gm29650, Eif2s3y, Ddx3y, Uty, Kdm5d*) or X-linked genes involved in X-inactivation (e.g., *Xist*). These sex-linked genes are trivially differential and should be excluded from such analysis. Inclusion of sex-linked genes may account for the overall significant overlap in gene lists that the authors report using rank-rank hypergeometric overlap analysis.

If the sex differences are genuine, they should appear in both control groups: acute and chronic saline. However, <3.7% of the reported DE autosomal genes replicate in these conditions using the authors' own criteria (Figure 2). It is not plausible that sex differences in gene expression are profoundly different in animals treated with acute versus chronic saline. Instead, this is consistent with the conclusion that most of the reported genes are false positives. Given the lack of statistical significance, it is invalid to describe the genes shown in Figures 1F–H and Figures 2E–G as DE. Notably, many of the pathway analysis results (using MEGENA) in the article are also based on uncorrected *p* values.

Although correcting for multiple comparisons increases the risk of false negatives, this can be mitigated by excluding genes with a low mean expression level, or using iterative hypothesis weighting (4). These approaches can reduce the burden of multiple comparisons and improve the power for subsequent DE gene analysis while still controlling false detections.

The second critical error concerns the interpretation that transcriptional responses to cocaine are vastly different between males and females. The evidence offered for this claim is that there was very little overlap (in the DE genes) between

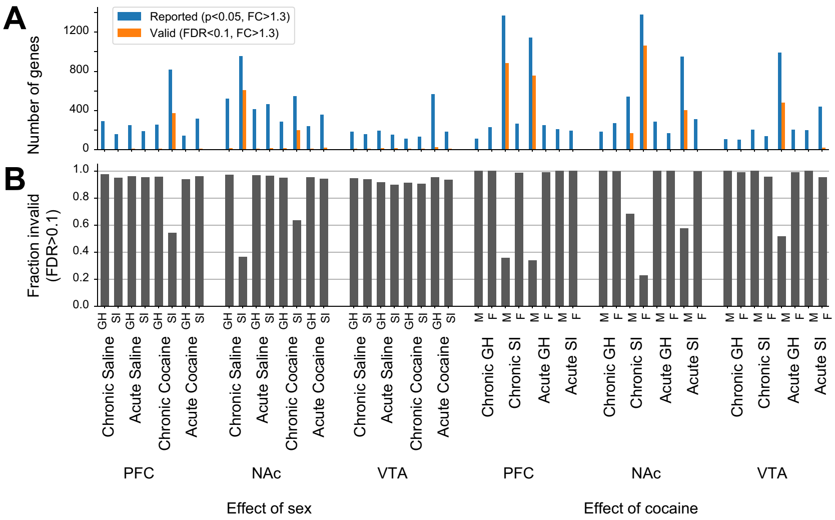

Figure 1. (A) The number of differentially expressed genes reported in each of 48 comparisons (blue) compared with the number of differentially expressed genes that survive an analysis adjusting for multiple comparisons (orange). **(B)** The fraction of invalid genes. F, female; FC, fold change; FDR, false discovery rate; GH, group housed; M, male; NAc, nucleus accumbens; PFC, prefrontal cortex; SI, social isolation; VTA, ventral tegmental area.

https://doi.org/10.1016/j.biopsych.2021.06.026
ISSN: 0006-3223

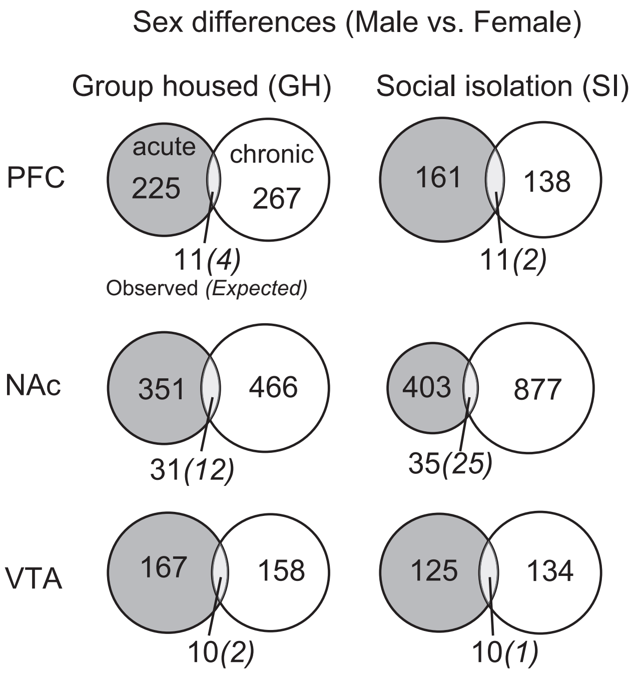

Figure 2. Using information from Table S2 in Walker *et al.* (1), the overlap was calculated between autosomal genes reported as having a sex difference in expression ($p < .05$, fold change >1.3) in mice treated with chronic vs. acute saline (gray and white circles, respectively). The observed number of genes in the intersection is shown along with the expected overlap based on a null hypothesis of completely random gene sets (shown in italics). GH, group housed; NAc, nucleus accumbens; PFC, prefrontal cortex; SI, social isolation; VTA, ventral tegmental area.

males and females across all three brain regions. However, the low degree of overlap is an expected result of the lack of control for false discoveries. Because the uncorrected test is expected to yield ~800 false positives for each group of animals, around 40 (5%) of these genes should overlap between the males and females. This is very close to the average overlap that was actually observed (39.1 genes, using the mean of the intersections shown in Figure 1F, G and Figure 2E–G). These low levels of overlap do not provide evidence for sex-specific transcriptional responses in gene expression, but rather merely reflect the fact that most of the putative DE genes are likely false positives.

More generally, it is not appropriate to draw conclusions from negative results of a statistical test (i.e., failure to reject a null hypothesis): absence of evidence is not evidence of absence. For example, a gene that has a statistically significant difference in expression between treatment groups in one sex but fails to reach significance in the other sex cannot, by virtue of those tests alone, be described as having a sex difference in transcriptional response. Instead, the hypothesis of a sex difference in response should be directly tested using a multifactor design, which is supported by popular RNA sequencing analysis software packages [e.g., edgeR (5), DE-Seq2 (6), limma (7)]. A generalized linear model can test for an interaction between sex and treatment, e.g.,

$$RNA \sim 1 + Sex + Treatment + Sex : Treatment$$

A low adjusted p value for the interaction term in this model would justify rejecting the null hypothesis and would support the authors' interpretation of a sex difference in the treatment response. Importantly, this procedure would account for the evidence from all experimental groups and provide a valid statistical measure of the significance of the interaction.

The absence of statistically sound evidence for sex differences in transcriptional responses in this study is not evidence for the absence of an effect and does not disconfirm the authors' hypotheses. Instead, the lack of statistically significant differential expression in many of the comparisons may be due to a lack of sufficient power. However, the conclusions of this and other studies that use uncorrected p values and/or negative DE test results are not supported by the data. Unfortunately, these are not isolated instances and instead reflect widespread practices in the field. Editors, reviewers, and researchers must raise the standards of statistical rigor in neurogenomics research to ensure that published findings are reproducible.

Eran A. Mukamel

Acknowledgments and Disclosures

This work was supported by the National Institutes of Health (Grant Nos. U01 MH122590 [to EAM] and RF1 MH120015 [to EAM]) and Chan Zuckerberg Initiative (Seed Networks for the Human Cell Atlas Grant No. CZF2019-002457 [to EAM]).

I am grateful to Roger Levy, Ronen Mukamel, and Francesca Telese for discussions.

EAM reports no biomedical financial interests or potential conflicts of interest.

Article Information

From the Department of Cognitive Science, University of California San Diego, La Jolla, California.

Address correspondence to Eran A. Mukamel, Ph.D., at emukamel@ucsd.edu.

See also associated correspondence: https://doi.org/10.1016/j.biopsych.2021.08.006.

Received Mar 10, 2021; revised May 13, 2021; accepted Jun 2, 2021.

References

1. Walker DM, Zhou X, Cunningham AM, Lipschultz AP, Ramakrishnan A, Cates HM, *et al.* (2021): Sex-specific transcriptional changes in response to adolescent social stress in the brain's reward circuitry [published online ahead of print Feb 24]. Biol Psychiatry.
2. Benjamini Y, Hochberg Y (1995): Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300.
3. Storey JD, Tibshirani R (2003): Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100:9440–9445.
4. Ignatiadis N, Klaus B, Zaugg JB, Huber W (2016): Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods 13:577–580.
5. Robinson MD, McCarthy DJ, Smyth GK (2010): edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140.
6. Love MI, Huber W, Anders S (2014): Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
7. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015): limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47.