gene-expression values[1]. The authors mixed complementary RNA from the tissues and observed similar off-diagonal effects. They concluded that the off-diagonal effects are due to technical reasons, such as nonlinear sample amplification or probe cross-hybridization, rather than statistical deconvolution.

We found that this deviation of signal reconstruction was the result of data transformation. In microarray studies, expression data are logarithm-transformed for variance stabilization or for approximation of a normal distribution[2]. However, we argue that in the context of expression-profile deconvolution, the log transformation will produce biased estimation. Deconvolution is modeled by a linear equation $\mathbf{O} = \mathbf{S} \times \mathbf{W}$, where **O** is the expression data for mixed tissue samples, *S* is the tissue-specific expression profile, and **W** is the cell-type frequency matrix. If the signal is log-transformed, the linearity will no longer be preserved. The concavity feature of the log function will induce a downward bias to the reconstructed signal (**Fig. 1a** and **Supplementary Fig. 1**). Mathematically, it can be shown that the deconvolution model used on log-transformed signals is $\log(\mathbf{O}') = \log(\mathbf{S}) \times \mathbf{W}$, where $\mathbf{O}'$ is the csSAM estimate of gene-expression profiles. As **W** is a frequency matrix and its column values sum to 1, the following is true by the properties of concave functions[3]: $\log(\mathbf{S} \times \mathbf{W}) > \log(\mathbf{S}) \times \mathbf{W}$. Taking these two equations together, we can conclude that $\log(\mathbf{O}') < \log(\mathbf{S} \times \mathbf{W}) = \log(\mathbf{O})$. Thus, we proved that when log-transformed signal is used as the input for signal reconstruction, it will always yield an underestimation of the true signal. By taking an anti-log transformation, we obtained an unbiased reconstruction of the mixed tissue samples (**Fig. 1b** and **Supplementary Fig. 2**).

The log transformation also introduced a large bias to the results of deconvolution (**Fig. 1c** and **Supplementary Fig. 3**). A substantial portion of the genes were off diagonal in the deconvolved cell type–specific gene-expression profiles. By performing the deconvolution in linear space, we achieved a considerably more accurate result (**Fig. 1d** and **Supplementary Fig. 3**).

In summary, an incorrect transformation of data can greatly bias the final results of deconvolution. In the context of gene-expression deconvolution, a linear model achieves better accuracy. Accurate deconvolution of expression profiles is important for downstream analysis, such as gene expression analysis and pathway-enrichment analysis. We urge caution in selecting data-transformation functions and any preprocessing steps in model-based statistical analysis.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS
We thank H.Y. Zoghbi, J. Botas, L.A. Chodosh, J. Alvarez, O. Lichtarge and T. Klisch for helpful insights and critical comments on this manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Yi Zhong[1,2] & Zhandong Liu[1,2]

[1]Department of Pediatrics, Jan and Dan Duncan Neurological Research Institute at Texas Children's Hospital, Baylor College of Medicine, Houston, Texas, USA. [2]Computational and Integrative Biomedical Research Center, Baylor College of Medicine, Houston, Texas, USA.
e-mail: zhandonl@bcm.edu

1. Shen-Orr, S.S. *et al. Nat. Methods* **7**, 287–289 (2010).
2. Quackenbush, J. *Nat. Genet.* **32,** 496–501 (2002).
3. Strang, G. *Calculus* 652 (Wellesley-Cambridge, 1991).

Shen-Orr *et al.* reply: We appreciate the comments made by Zhong and Liu and their hard work on the proof[1]. Indeed, removing unneeded normalization methods, including log transformation, can yield even better linearity results, optimizing the use of deconvolution methods.

Although we would expect that better deconvolution methodology will be more sensitive for detecting cell type–specific differences between groups, empirically we have found that this is not always the case. Cell type–specific significance analysis of microarrays (csSAM) compares expression between two groups[2], the gene expression data of each of which is separately deconvolved to yield cell type–specific expression. The false discovery rate for cell type–specific differences between groups is assessed via permutations, an expected side effect of which is the reduced effect of systemic biases, as those are controlled for statistically. We found that in the complex context of actual sample data, use of a log transformation on the measured 'raw' gene expression data input into the linear csSAM deconvolution model often yields improved (lower) false discovery rates between groups than when the raw data are kept as is or are log-transformed after deconvolution. Such is the case for the acute-rejection versus stable individual data we discuss in the publication[2] (Supplementary Fig. 1). A possible reason for this may be that as a function of the technology used, actual transcript abundance may be separated from what we consider as 'raw' measured gene expression by intermediate steps (for example, labeling, hybridization and scanning in the case of microarrays), which may affect linearity. Thus, we would recommend that users of csSAM try different choices of transformations, guided by the visual appearance of the results and the estimated false discovery rate.

The latest update of the csSAM R package as well as a Microsoft Excel Add-In are available at http://buttelab.stanford.edu/public:data. They include added functionality that allows effortless switching between log-transformed and anti-log-transformed gene expression values when performing either the deconvolution or comparative expression steps of csSAM.

Note: Supplementary information is available on the Nature Methods website.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Shai S Shen-Orr[1], Robert Tibshirani[2] & Atul J Butte[3]

[1]Department of Immunology, Rappaport Faculty of Medicine, Technion, Haifa, Israel. [2]Health Research and Policy and Department of Statistics, Stanford University School of Medicine, Stanford, California, USA. [3]Division of Systems Medicine, Department of Pediatrics, Stanford University School of Medicine, Stanford, California, USA.
e-mail: shenorr@technion.ac.il

1. Zhong, Y. & Liu, Z. *Nat. Methods* **9**, 8–9 (2011).
2. Shen-Orr, S.S. *et al. Nat. Methods* **7**, 287–289 (2010).