




Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes

Gregory B. Gloor, Jean M. Macklaim & Andrew D. Fernandes


To cite this article: Gregory B. Gloor, Jean M. Macklaim & Andrew D. Fernandes (2016) Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes, Journal of Computational and Graphical Statistics, 25:3, 971-979, DOI: [10.1080/10618600.2015.1131161](https://doi.org/10.1080/10618600.2015.1131161)

To link to this article: <https://doi.org/10.1080/10618600.2015.1131161>

 View supplementary material [↗](#)

 Published online: 05 Aug 2016.

 Submit your article to this journal [↗](#)

 Article views: 1770

 View related articles [↗](#)

 View Crossmark data [↗](#)

 Citing articles: 35 View citing articles [↗](#)

SHORT TECHNICAL NOTE

Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes

Gregory B. GLOOR, Jean M. MACKLAIM, and Andrew D. FERNANDES

Displaying the component-wise between-group differences high-dimensional datasets is problematic because widely used plots such as Bland–Altman and Volcano plots do not show what they are colloquially *believed* to show. Thus, it is difficult for the experimentalist to grasp why the between-group difference of one component is “significant” while that of another component is not. Here, we propose a type of “Effect Plot” that displays between-group differences in relation to respective underlying variability for every component of a high-dimensional dataset. We use synthetic data to show that such a plot captures the essence of what determines “significance” for between-group differences in each component, and provide guidance in the interpretation of the plot. Supplementary online materials contain the code and data for this article and include simple R functions to produce an effect plot from suitable datasets.

Key Words: ANOVA; Bland–Altman plot; Genomics; Multivariate datasets; Transcriptomics; Volcano plot.

1. INTRODUCTION

Large datasets are ubiquitous in many domains and in some areas, such as biology, our ability to generate new datasets far outstrips our ability to analyze them. For example, a single analysis on a mid-range instrument such as an Illumina MiSeq or Ion Torrent PGM can generate millions of data points for multiple independent samples. A single instrument can generate a new dataset every day, and there are thousands of such instruments in use.

In this deluge of data, it is difficult to meaningfully examine the relationships between inter- and intragroup variation using current visualization tools. We illustrate the problem using a synthetic dataset composed of 40 random vectors containing 200 normally distributed components. Each component belongs to one of four classes, the characteristics of which are shown in [Table 1](#). Each random vector is an observation and belongs to one of two groups, “X” or “Y,” which can be thought of as “treatment” or “control” if desired.

Gregory B. Gloor (E-mail: ggloor@uwo.ca) and Jean M. Macklaim (E-mail: jean.macklaim@gmail.com), Department of Biochemistry, The University of Western Ontario, London, ON N6A 3K7, Canada. Andrew D. Fernandes, Department of Applied Mathematics, The University of Western Ontario, London, ON N6A 3K7, Canada and YouKaryote Genomics, London, ON N6K 0A1, Canada (E-mail: andrew@fernandes.org).

© 2016 *American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America*
Journal of Computational and Graphical Statistics, Volume 25, Number 3, Pages 971–979
DOI: [10.1080/10618600.2015.1131161](https://doi.org/10.1080/10618600.2015.1131161)

Table 1. Characteristics of the synthetic dataset. For each class, n represents the number of components of that class. Components in group $X \sim N(\mu \pm \delta/2, \sigma^2)$ while components in group $Y \sim N(\mu \mp \delta/2, \sigma^2)$ such that the magnitude of the expected between-group difference of the component is δ . The ratio $\theta = \delta/\sigma$ represents an effect-size along the lines of Cohen's d (Cohen 1988). Twenty observations are realized for each group

| Class | n | μ | σ | δ | δ/σ |
|-------|-----|-------|----------|----------|-----------------|
| 1 | 80 | 9 | 1 | 0 | 0 |
| 2 | 80 | 10 | 2 | 0 | 0 |
| 3 | 20 | 11 | 1 | 2 | 2 |
| 4 | 20 | 12 | 1/2 | 4 | 8 |

In the language of biology, each component would correspond to a gene, operational taxonomic unit, or genomic interval, and each random vector would correspond to a sample or observation.

One frequently desired analysis of this type of data seeks to describe the univariate component-wise differences between observational groups. When dealing with small numbers of components, it is convenient to visualize the between-group difference via a boxplot or a strip chart, as illustrated in Figure 1. For the four components depicted in this figure, it is easy to visually assess the differences between groups. Subjectively speaking, small differences can be important if the within-group variance is small. Conversely, large between-group differences may be meaningless if within-group variance is also large.

It is obviously impractical to examine plots such as Figure 1 for *all* 200 observational components. Doing so in a meaningful way becomes even more of a challenge upon the realization that actual biological datasets consisting of 10–50,000 dimensional observations are commonplace (Frazee et al. 2011). For example, a typical biological experiment may

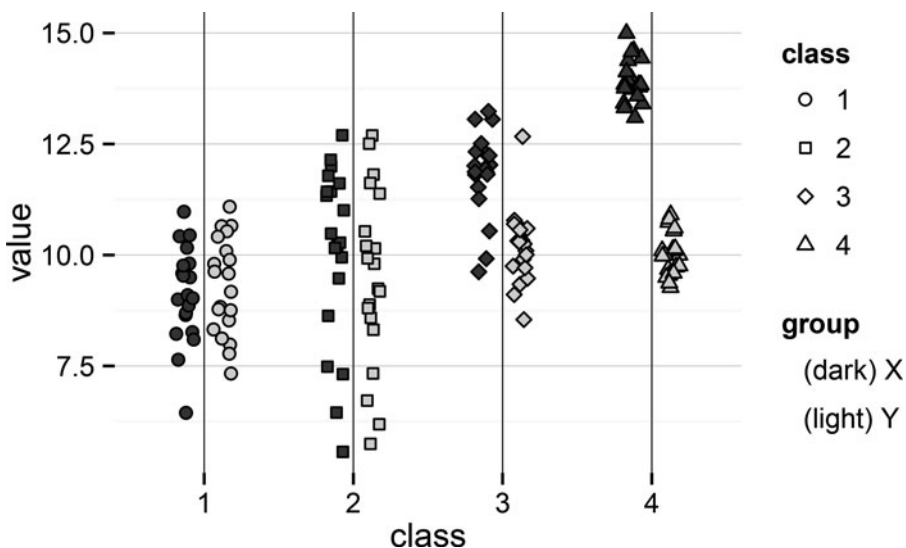


Figure 1. Representative simulated data. One observed component from each class and group is shown. Note how the between-group characteristics visually correspond to the group and class characteristics described in Table 1.

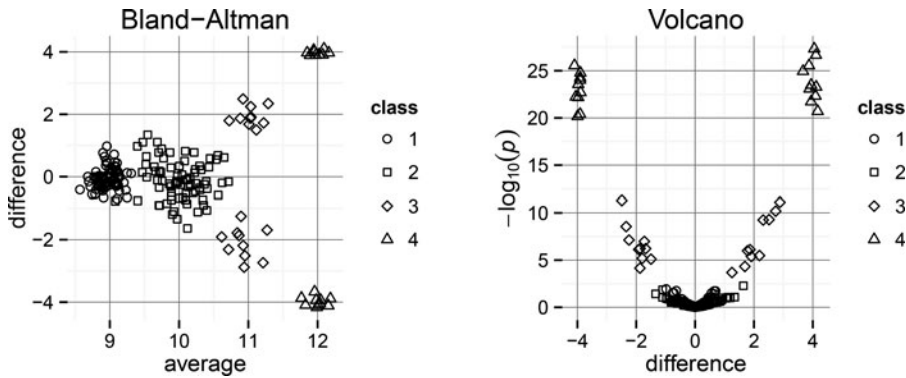


Figure 2. Bland–Altman and Volcano plots of the synthetic dataset described in Table 1. For the Volcano plot, p -values are computed via Welch’s t -test. Note that it is often unclear and somewhat controversial whether raw or corrected p -values should be depicted, where any such corrections would be to control the false-discovery rate.

encompass hundreds of operational taxonomic units (Proctor 2011), tens of thousands of genes (Bottomly et al. 2011), and hundreds of thousands of genomic intervals (Sims et al. 2014). Thus, it is also obviously impractical to summarize the data in a table, although it is common practice to summarize the “most significant” components in this way. Tables for large datasets are cumbersome, and display these “meaningful hits” out of context (Feinberg and Wainer 2011). Therefore, our goal is to display a summary of between-group differences for *all* components in an intuitive yet rigorous and quantitative manner.

Currently, there are two main methods used to depict between-group differences for the type of high-dimensional data described above.

First, the groups can be contrasted via the Bland–Altman (BA) plot (Altman and Bland 1983) (also called, for historical reasons rooted in microarray analysis, an “MA” plot) as shown in Figure 2. BA plots are widely used, especially in biology, to summarize RNA-Seq and other types of high throughput sequencing data. In its most common incarnation a BA plot assigns the between-group mean of a component to the abscissa and the between-group difference of that component to the ordinate. Such a depiction is useful to determine if the data components are symmetrical in their between-group differences. However, in the absence of other information, BA plots fail to describe how large the depicted differences are in relation to any underlying variability. In fact, the axes of BA plots contain no information regarding observational variability and thus, in the absence of additional information, yield surprisingly little information about how meaningful any observed differences are.

Partly in response to this shortcoming, the other common depiction, the Volcano plot (Cui and Churchill 2003), has become a popular adjunct to the BA plot. A Volcano plot, as shown in Figure 2, depicts the between-group difference and the p -value of a statistical test hypothesizing that the groups have identical means, most commonly a t -test of some sort. Since the p -value is implicitly dependent on underlying variability and on the sample size, the Volcano plot *appears* to display the biological significance of a given between-group difference for each component.

Unfortunately, the greatest perceived strength of the Volcano plot is also its greatest weakness because p -values are *not* effect-sizes. In fact, p -values are not even good *proxies* for effect-sizes (Halsey et al. 2015). However, belief that a p -value is “more or less indicative

of” an effect-size is a persistent and pernicious idea still endemic in the scientific community (Ziliak and McCloskey 2008).

Rather than promulgating the mistaken belief that p -values and effect-sizes have a strong meaningful relationship, we propose something surprisingly effective in its simplicity: plotting the underlying constituents of effect-size directly.

2. METHODS

There are numerous statistical notions of “effect-size,” but for describing the difference between population means the most common definition is the standardized mean-difference θ defined by

$$\theta = \frac{\mu_x - \mu_y}{\sigma} = \frac{\delta}{\sigma},$$

where μ_x and μ_y are the population means for their respective groups and σ is a dispersion parameter based on either or both population standard deviations (Hedges and Olkin 1985). Much of the literature regarding this type of effect-size has to do with *estimating* θ with as little bias as possible, especially for small sample sizes (Cohen 1988). Note that the most obvious and common choices of δ and σ define the difference δ as the “difference between means” and the dispersion σ as a standard deviation. However, *any* reasonable estimate of difference and dispersion can be used if such estimates are mutually compatible and appropriate for the hypothesized model and characteristics of the observed data.

For normally distributed data, Cohen’s d , Glass’ Δ , and Hedges’ g and g^* are common choices, with g^* being preferred due to its smaller bias for small sample sizes (Hedges and Olkin 1985). For each of these effect-size estimates, the numerator is equal to $\hat{\delta} = \bar{X} - \bar{Y}$, which is unbiased. Differences between these estimators are due to how they estimate $\hat{\sigma}$ and correct for bias, if they do so.

For nonnormal data, robust estimators of “location” and “scale” are appropriate, where the statistical term location is interpreted (somewhat confusingly) as the “*location* of the *difference* in group means” and “scale” is interpreted as the “*scale* of the random *dispersion*.” As is usual with robust estimation techniques, there is usually a trade-off between how robust the method is and how efficient it is in making the “most” use of the data (Huber and Ronchetti 2009). Robust estimates of δ can be provided by the Hodges–Lehmann difference

$$\hat{\delta} = \text{median of } \{x_i - y_j\} \text{ over all observations } i, j,$$

which has a breakdown point of 29% and a Gaussian efficiency of 86% (Lehmann and D’Abrera 2006). Note that Hodges–Lehmann estimators are known to be suitable *only* for symmetric distributions, which has historically limited them somewhat. Fortunately, the random variable defined as $X - Y$ is symmetric by its very construction.

For robust estimation of σ , the interquartile range (IQR) and median absolute deviation (MAD) are popular choices, the latter especially so. Although the MAD has the highest possible breakdown point of 50%, it has a surprisingly low Gaussian efficiency of only 37%, implying that it should be used for small sample sizes only with caution. A better choice, and one recommended frequently, is Rousseeuw’s Croux’s Q_n scale estimator, corrected for bias and scaled to correspond to the unit normal distribution (Rousseeuw and Croux

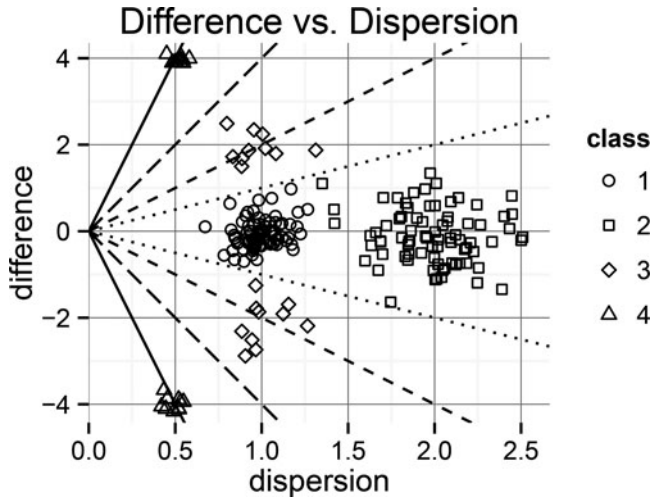


Figure 3. The effect size plot, where “difference” refers to estimates $\hat{\delta}$ of the between-group difference δ and “dispersion” refers to the estimate $\hat{\sigma}$ of the standard deviation σ . Diagonal lines are shown for zero-intercept lines with slopes of ± 1 , ± 2 , ± 4 , and ± 8 and these lines correspond to the expected location of points with the corresponding effect sizes.

1993). This estimator is computed as

$$\hat{\sigma} = Q_n = \text{first quartile of } d_n \{ |z_i - z_j|, i < j \},$$

where the set of z consists of all pairwise differences between x and y , and d_n is a constant, dependent only on the sample size, which corrects for bias and scales it to correspond to the unit normal. Q_n is easily computable via the `robustbase` package (Rousseeuw et al. 2015) of the R computational framework (R Core Team 2015). Better yet, Q_n still has the highest possible breakdown point of 50%, but also features a surprisingly high Gaussian efficiency of 82%.

2.1 GENERATION AND INTERPRETATION OF THE PLOTS

An example of the type of plot we propose is shown in Figure 3 using the same synthetic dataset as before. The “Difference versus Dispersion” plot depicts an estimate of the difference δ versus the estimated pooled standard deviation σ (Cohen 1988). This plot (herein called an “effect plot”) displays and summarizes for each component in the dataset the relationship between δ and σ in an intuitive way. The experimentalist can use this plot to explore the univariate differences between groups in the dataset efficiently. For example, the points in class 1, with $\delta = 0$ and $\sigma = 1$ can be seen to cluster around a difference of zero and a dispersion of one, and the points in class 2 with $\delta = 0$ and $\sigma = 2$ cluster around the same difference but with twice the dispersion. Thus, it is clear from the plot that the components in the X and Y groups in classes 1 and 2 differ only by intrinsic variability. The class 3 components have the same dispersion as class 1 components, but it can be seen that the class 3 components have a difference between groups of ± 2 . Class 4 components behave similarly, but are more extreme as these components have even less dispersion and a

larger difference between groups. From this plot then, it is clear that the class 4 components are the ones with the largest standardized difference (i.e., the largest effect size) between groups and the class 2 components are the ones with the smallest.

The intuited effect size θ , which is the ratio of the between-group difference and the within-group dispersion, can be interpreted by examining the plot by sector, and diagonal lines are drawn on the plot to aid in this visualization. For example, the dotted lines represent the position in the plot where δ and σ are equal, and so corresponds to an effect size of one. Points contained within this sector have an effect size ≤ 1 . Consequently, any components that are found in the zone between these two lines have larger dispersion than difference, and so even if sample sizes were sufficiently large to be statistically significantly different, components in this zone may not be *meaningfully* biologically different. Proceeding outward from this sector, the subsequent zones indicate progressively larger effect sizes. The zones between the dotted and single-dashed lines represent the regions where the magnitude of the effect size is between 1 and 2; the zones between the single- and double-dashed lines represent an effect magnitude of between 2 and 4; and the zones between the double-dashed and solid lines represent an effect magnitude between 4 and 8. Note that the members of class 3, which have an expected effect size of 2, are grouped on or near the line indicating an effect size of 2, and the members of class 4 are found near the line indicating an effect size of 8. The effect-size plot thus displays the actual information that is required to differentiate and meaningfully interpret *statistical* significance, with regard to hypothesis testing, and *biological* significance, with respect to what an experimentalist expects to observe. It is telling that when the number of biological samples becomes large, common practice is to impose additional constraints, such as magnitude of change, in addition to statistical significance to identify differentially expressed genes (Love et al. 2014).

Finally, we demonstrate the utility of this plot using a historical public dataset. The “Bottomly” dataset (Bottomly et al. 2011) is a typical dataset generated through the RNA-seq technique. The table of counts was obtained from the ReCount database (Frazee et al. 2011). Here RNA corresponding to genes was sequenced on a high-throughput sequencing instrument, generating millions of sequence reads that corresponded to individual genes. In this dataset, the count of reads per gene is taken to represent the expression of that gene in the original sample and it comprises 21 samples in groups of 11 and 10 samples. The dataset was reduced to only the 13,932 genes that had at least one corresponding sequence read count. The total number of sequence reads per sample ranged from 2,717,092 to 7,339,817.

We applied a similar statistical approach to that taken in the original article whereby the read count values from the dataset were normalized to a consistent depth using the TMM method (Robinson and Oshlack 2010; Sun et al. 2013) and differential expression was determined using the exact negative binomial test instantiated in the edgeR package (Robinson et al. 2010). All p -values were adjusted for multiple hypothesis tests using the Benjamini–Hochberg approach (Benjamini and Hochberg 1995). For plotting, a pseudo-count of 0.5 was added to each actual count, and the resultant values were log transformed using base 2. The summary statistics that are plotted in Figure 4 include the following: the “average” is the mean value of each gene across the 21 samples; “difference” is the difference between the means of the two groups; the p -value was calculated using the exact negative binomial test instantiated in the edgeR package (Robinson et al. 2010); and “dispersion” is the pooled standard deviations of groups 1 and 2. Observe that the BA plots

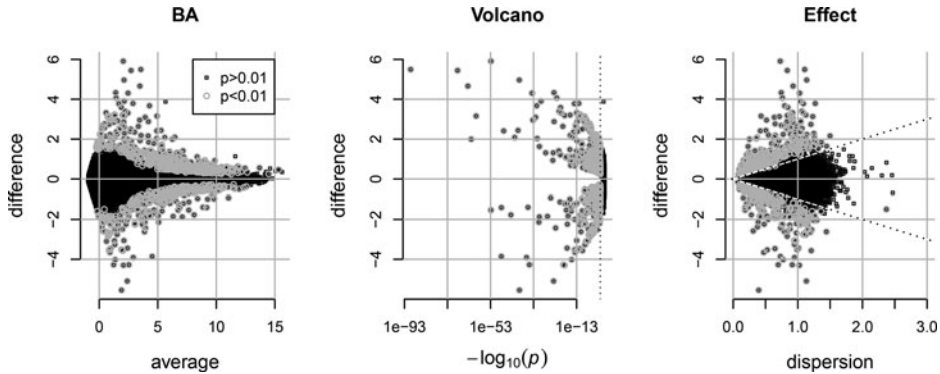


Figure 4. Three ways of viewing the Bottomly RNA-seq dataset. The dataset has 21 samples with 13,932 informative genes split into two groups. The three plots show each gene as a separate point, with small black square points indicating the bulk of the values, and large round gray points indicating a Benjamini–Hochberg corrected p -value ≤ 0.01 . The BA plot shows the difference in the means of the groups versus the mean value of all samples, the Volcano plot shows the difference versus the adjusted p -values, and the effect plot shows the difference versus the pooled standard deviation. Note that, for clarity of presentation, the effect plot omits a single point with a dispersion > 8 and a difference of ≈ 2 . The dotted line in the Volcano plot represents an adjusted p -value of 0.01, and in the Effect plot the dotted lines represent an estimated effect size of ± 1 .

is useful to determine the symmetry of the dataset, and suggests that genes with a higher mean tag count are “significant” with smaller difference. The Volcano plot shows that equivalent p -values can be obtained by genes with very different differences, but it provides no information as to the reason for the high scatter. In contrast, the effect plot contains the symmetry information found in the BA plot and, when significant genes are identified, exhibits essentially similar information regarding the relationship between p -values and difference as the Volcano plot. Finally, it is obvious from the effect plot that many genes exhibit very small differences coupled with even smaller dispersions, information that is not available from either of the other plots. In practice, all relevant information can be obtained from a combination of MA and effect plots.

3. SUMMARY

Determining differential abundance in RNA-seq datasets can be fraught with problems since analyses are exploratory, and there is rarely a standard of truth. Thus, there is a need for additional diagnostic tools, and one recent tool uses interactive graphical approach to identify structural problems with RNA-seq datasets (see, e.g., Yin et al. 2013). Here, we have described a simple, powerful, and easily interpreted graphic that plots the difference between groups versus the dispersion within the groups. As such, it plots difference versus dispersion constituents of an effect-size statistic. Using simulated data we show that the effect plot is more visually informative than other widely used plots regarding the summary statistics of interest for each component of a large dataset. The plot is simple to interpret and provides a compact way to characterize univariate differences in complex datasets, and scales to datasets containing thousands of components. Finally, the plot can be produced using statistics that are easily calculated for most datasets making it simple to implement

for many purposes, regardless of whether the underlying variables are normal or not. The ALDEx2 R package for differential abundance analysis, available through Bioconductor, has the option to generate effect plots from its output (Fernandes et al. 2014).

SUPPLEMENTARY MATERIALS

Supplemental files are contained within the archive Supplement_effect.zip. The contents include the following:

README.txt: A text file outlining all files contained in the archive.

jcgs_effect_R1.Rnw: An R noweb document containing the source for this article. It can be compiled using the knitr package.

Source and data files: Source code and data files for all figures in the directory “chunk.” In particular, two supplementary files (40_effect.r, 50_bottomly.r) contain the R code necessary to produce effect plots using simulated and real data.

build.sh: A bash script to compile the .Rnw file and R code from the command line.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the following funding sources: The Natural Sciences and Engineering Research Council of Canada for support through their Discovery Grant and CREATE programs.

[Received August 2015. Revised November 2015.]

REFERENCES

- Altman, D. G., and Bland, J. M. (1983), “Measurement in Medicine: The Analysis of Method Comparison Studies,” *Journal of the Royal Statistical Society*, Series D, 32, 307–317. Available at <http://www.jstor.org/stable/2987937>. [973]
- Benjamini, Y., and Hochberg, Y. (1995), “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing,” *Journal of the Royal Statistical Society*, Series B, 57, 289–300. [976]
- Bottomly, D., Walter, N. A. R., Hunter, J. E., Darakjian, P., Kawane, S., Buck, K. J., Searles, R. P., Mooney, M., McWeeney, S. K., and Hitzemann, R. (2011), “Evaluating Gene Expression in C57BL/6J and DBA/2J Mouse Striatum Using RNA-seq and Microarrays,” *PLoS One*, 6, e17820. [973,976]
- Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates. Available at <http://www.loc.gov/catdir/enhancements/fy0731/88012110-d.html>. [972,974,975]
- Cui, X., and Churchill, G. A. (2003), “Statistical Tests for Differential Expression in cDNA Microarray Experiments,” *Genome Biology*, 4, 210.1–210.10. [973]
- Feinberg, R. A., and Wainer, H. (2011), “Extracting Sunbeams From Cucumbers,” *Journal of Computational and Graphical Statistics*, 20, 793–810. Available at <http://dx.doi.org/10.1198/jcgs.2011.204a>. [973]
- Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (2014), “Unifying the Analysis of High-throughput Sequencing Datasets: Characterizing RNA-seq, 16s rRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis,” *Microbiome*, 2, 15.1–15.13. [978]

- Frazee, A. C., Langmead, B., and Leek, J. T. (2011), "Recount: A Multi-Experiment Resource of Analysis-Ready RNA-seq Gene Count Datasets," *BMC Bioinformatics*, 12, 449.1–449.5. [972,976]
- Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015), "The Fickle p Value Generates Irreproducible Results," *Nature Methods*, 12, 179–185. [973]
- Hedges, L. V., and Olkin, I. (1985), *Statistical Methods for Meta-Analysis*, Orlando, FL: Academic Press. Available at <http://www.loc.gov/catdir/description/els032/84012469.html>. [974]
- Huber, P. J., and Ronchetti, E. (2009), *Robust Statistics*, Wiley Series in Probability and Statistics (2nd ed.), Hoboken, NJ: Wiley. Available at <http://www.loc.gov/catdir/toc/ecip0824/2008033283.html>. [974]
- Lehmann, E. L., and D'Abrera, H. J. M. (2006), *Nonparametrics: Statistical Methods Based on Ranks* (1st rev. ed.), New York: Springer. Available at <http://www.loc.gov/catdir/toc/fy0704/2006927419.html>. [974]
- Love, M. I., Huber, W., and Anders, S. (2014), "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data With DESeq2," *Genome Biology*, 15, 550.1–550.21. [976]
- Proctor, L. M. (2011), "The Human Microbiome Project in 2011 and Beyond," *Cell Host Microbe*, 10, 287–291. [973]
- R Core Team (2015), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at <https://www.R-project.org/>. [975]
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010), "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data," *Bioinformatics*, 26, 139–140. [976]
- Robinson, M. D., and Oshlack, A. (2010), "A Scaling Normalization Method for Differential Expression Analysis of RNA-seq Data," *Genome Biology*, 11, R25.1–R25.9. [976]
- Rousseeuw, P., Croux, C., Todorov, V., Ruckstuhl, A., Salibian-Barrera, M., Verbeke, T., Koller, M., and Maechler, M. (2015), *robustbase: Basic Robust Statistics, R package version 0.92-5*. Available at <http://CRAN.R-project.org/package=robustbase>. [975]
- Rousseeuw, P. J., and Croux, C. (1993), "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, 88, 1273–1283. Available at <http://www.jstor.org/stable/2291267>. [975]
- Sims, D., Sudbery, I., Ilott, N. E., Heger, A., and Ponting, C. P. (2014), "Sequencing Depth and Coverage: Key Considerations in Genomic Analyses," *Nature Reviews Genetics*, 15, 121–132. [973]
- Sun, J., Nishiyama, T., Shimizu, K., and Kadota, K. (2013), "TCC: An R Package for Comparing Tag Count Data With Robust Normalization Strategies," *BMC Bioinformatics*, 14, 219.1–219.13. [976]
- Yin, T., Majumder, M., Chowdhury, N. R., Cook, D., Shoemaker, R., and Graham, M. (2013), "Visual Mining Methods for RNA-Seq Data: Data Structure, Dispersion Estimation and Significance Testing," *Journal of Data Mining in Genomics & Proteomics*, 4, 2153–0602. [977]
- Ziliak, S. T., and McCloskey, D. N. (2008), *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, Ann Arbor, MI: University of Michigan Press. [974]