General Comments

The authors present an effect size statistic, $\mathbb{E}$, for use with high-throughput sequencing (HTS) data. They argue that use of this statistic provides more reproducible results than significance-testing methods. While reproducibility is an important consideration in development and use of statistical methods, the analysis presented in this manuscript is not, in my opinion, currently sufficient to support the claims advanced by its authors.

Basic Reporting

- The model the authors use to motivate and calculate $\mathbb{E}$ should be presented explicitly in terms of (1) a prior, (2) a likelihood, (3) data. In particular, all parameters in the model should be explicitly defined. See ch. 3 of Wakefield (2013) for examples of this sort of presentation.
- At least within the context of the model mentioned in the last point, the quantity estimated by $\mathbb{E}$ should be given explicitly. (I.e., what parameter is $\mathbb{E}$ an estimate of?)
- Some special terminology is not explicitly defined. In particular, the terms "reproducible," "parametric," and "non-parametric" have various meanings in various statistical contexts, so the authors should indicate in which sense they are using them. (In a similar vein, "linearize" in lines 80 and 82 of the manuscript should be defined or omitted.)
- The number of observations in each of the two datasets examined in the manuscript is not to my knowledge given; it should be included in the main text of the manuscript.
- Figure 1 is described as giving distributions of various quantities, which appear from context to be samples from a posterior distribution. If this is the case, how many Monte Carlo samples were taken to generate these plots, and how was smoothing performed? In addition, the caption of this plot begins "density of read counts may not follow a Normal distribution." Is this plot a descriptive summary of read counts or is it a plot of posterior distributions of transformed relative abundances? The distinction should be made clear, or if the authors believe the two coincide, this should be explained.
- Figure 2 includes p-values from t- and Wilcoxon tests. Were these tests applied to raw read counts? Transformed counts? More details are needed.
- Lines 118-120: "These were compared to a 'truth' set determined by identifying those features that were identified in all of 100 independent tests of the full dataset with outliers removed using the same tests and cutoffs."
  - I think there may be a typo in here – all 100 *tests*?
  - Is the truth set the same for each test, or are different truths used for different tests?
  - Is the truth set determined by results on 100 subsets or on the dataset from which they were taken?

Experimental Design

- The authors apply methods for estimation of central tendency and dispersion in *unknown* distributions possibly subject to contamination to summarizing a posterior distribution. However, data, prior, and likelihood together completely determine the form of the posterior distribution. In light of this, the decision to use tools developed for unknown distributions in this setting (rather than, for example, the more standard posterior median and equal-tailed credible interval) should be addressed and explained.

- The authors mention that they place independent Dirichlet$(1/2, \ldots, 1/2)$ priors on read probabilities in each sample.

  - This is described as the conjugate prior for a multivariate Poisson distribution (line 79). It is not. Is the multinomial distribution meant here rather than multivariate Poisson?
  - This implicitly induces a (sample-size-dependent, I believe) prior on effect size – this prior choice should be characterized, discussed, and justified. Why not explicitly put a prior on effect size (e.g., via a hierarchical model)?

- The authors motivate their work by asserting that commonly effect size estimates such as Cohen's d are "parametric" (unnumbered line after line 45 in submitted manuscript) and "depend upon the data being relatively normal" and hence are not expected to perform well on HTS data. H

  - Why is no such effect size estimate compared to $\mathbb{E}$ in simulations presented in this manuscript?
  - In what sense is Cohen's d parametric, and in what sense does it assume data are normally distributed? Means and standard deviations may be informative (though perhaps difficult to estimate) even when data is highly skewed. For discussion in the context of public health data, see Lumley et al. (2002).

Validity of Findings

- This manuscript is titled "A distribution-based effect size is more reproducible than hypothesis testing when analyzing high throughput sequencing datasets." The basis for this claim appears to be a comparison of a single method, edgeR, that employs significance testing to its proposed method(s). Moreover, no significance testing method is compared on 16S data (see figure 4). Significance testing is a very general paradigm; comparison to performance of a single method is not a sufficient basis for the broad claim made in the title and elsewhere in the manuscript.

- The authors argue both that p-values are poor predictors of reproducibility (lines 34-35) and that the "very good correspondence [between] features with very high effect sizes [and] low adjusted p-values" is evidence that "$\mathbb{E}$ is an appropriate metric for effect size" (lines 153-156). If p-values are unreliable, why should a strong association between $\mathbb{E}$ and p-values give us confidence in $\mathbb{E}$?

- Comparisons between the true positive rates for methods that do and do not control the false positive rate is uninformative; at the extreme, a method that reports every feature as significantly different will have a perfect TPR, but this is hardly a useful tool, and such

results are not reproducible in any particularly meaningful way. Accordingly, in order for comparisons of TPR to be informative, the comparisons presented in figure 4 should include either edgeR results with a q-value cutoff chosen so the FPR approaches the FPR of other methods, or adjusted (more stringent) cutoffs for other methods so that the FPR is no more than the nominal 0.1.

References

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. Annual review of public health, 23(1), 151-169.

Wakefield, J. (2013). Bayesian and frequentist regression methods. Springer Science & Business Media.