A reproducible effect size is more useful than an irreproducible hypothesis test to analyze high throughput sequencing datasets

Andrew D. Fernandes[1], Michael Vu[2], Lisa-Monique Edward[2], Jean M. Macklaim[2], and Gregory B. Gloor[2]

[1]**San Diego CA, 92130**

[2]**Department of Biochemistry, University of Western Ontario, London, N6A 5C1, Canada**

Corresponding author:

G. Gloor[5]

Email address: ggloor@uwo.ca

ABSTRACT

High throughput sequencing is analyzed using a combination of null hypothesis significance testing and ad-hoc cutoffs. This framework is strongly affected by sample size, and is known to be irreproducible in underpowered studies, yet no suitable non-parameteric alternative has been proposed. Here we present implementations of non-parametric standardized effect size estimates, $\mathbb{E}$, for high-throughput sequencing datasets. Case studies are shown for modelled data, transcriptome and amplicon-sequencing datasets. The $\mathbb{E}$ statistic is shown to be more reproducible and robust than p-values and requires sample sizes as small as 5 to reproducibly identify differentially abundant features. Source code and binaries freely available at: https://bioconductor.org/packages/ALDEx2.html, omicplotR, and https://github.com/ggloor/CoDaSeq. Datasets can be found at doi://10.6084/m9.figshare.8132216

INTRODUCTION

High throughput sequencing (HTS) datasets for transcriptomics, metagenomics and 16S rRNA gene sequencing are high dimensional and generally conducted at pilot-scale sample sizes. Much effort has been spent identifying the best approaches and tools to determine what is 'significantly different' between groups (Soneson and Delorenzi, 2013; Schurch *et al.*, 2016), but the answer seems to depend on the specific dataset and associated model parameters (Thorsen *et al.*, 2016; Hawinkel *et al.*, 2018; Weiss *et al.*, 2017). As commonly conducted the investigator determines what is 'significantly different' using a null hypothesis significance approach and then decides what level of difference is 'biologically meaningful' among the significantly different features. Graphically, this approach is represented by the Volcano plot (Cui and Churchill, 2003) where the magnitude of change (difference) is plotted vs the p-value.

One under-appreciated consequence of pilot scale research is that false positive features will often have very low apparent p-values (Halsey *et al.*, 2015). This explains in part why so many observations fail to replicate in larger datasets (Ioannidis, 2005). In fact, both p-values and absolute difference are poor predictors of replication likelihood if the experiment were conducted again (Cumming, 2008; Halsey *et al.*, 2015). Null-hypothesis significance based testing methods also have the property that the number of significant features identified is affected by the number of samples being compared. This leads to the practice of prioritizing statistical positive observations over biologically significant differences.

On the other hand, a standardized effect size addresses the issues of interest to the biologist:"what is reproducibly different?" or "would I identify the same true positive features as different if the experiment were repeated?" (Coe, 2002; Nakagawa, 2004; Colquhoun, 2014; Gloor *et al.*, 2016b). Standardized effect size statistics start from the assumption that there is a difference, but that the difference can be arbitrarily close to zero. Unfortunately, standardized effect size metrics are not routinely used when analyzing HTS datasets. One potential barrier is that parametric effect size statistics may not be suitable for HTS datasets because the data may not fit a Gaussian distribution.

The most widely used standardized effect size is Cohen's d, which is a parametric standardized effect size for the difference between the means of two groups. The general formulation is given in Equation 1,

$$\text{Cohen}'\text{s d} = \frac{\text{mean}(a) - \text{mean}(b)}{\sigma_{a,b}} \tag{1}$$

and is a Z score when measured in a Normal distribution. Cohen's d measures the difference between the means of the two distributions divided by the pooled standard deviation, denoted as $\sigma_{a,b}$. However, this metric depends upon the data being relatively Normal, which cannot be guaranteed for HTS data as seen in Figure 1.

The purpose of this report is to show that we can characterize the difference between distributions in a non-parametric manner without resorting to either summarizing the data prematurely or resorting to a rank-based approach, both of which discard much information. We introduce, $\mathbb{E}$, a simple-to-calculate non-parametric standardized effect size statistic calculated on distributions directly. This measure is implemented in the ALDEx2, and CoDaSeq R packages. The $\mathbb{E}$ statistic has been used in both meta-transcriptome and microbiome studies, for example see (Macklaim *et al.*, 2013; Bian *et al.*, 2017), and has been shown to give remarkably reproducible results even with extremely small sample sizes (Nelson *et al.*, 2015). The $\mathbb{E}$ metric has a near monotonic relationship with p-values, but has the advantage of being relatively stable between sample sizes. However, it is unknown how $\mathbb{E}$ compares with parametric effect size estimates, how many samples are required, and its sensitivity and specificity characteristics.

METHODS

Calculating $\mathbb{E}$

High throughput sequencing (HTS) machines output thousands to billions of 'reads', short nucleotide sequences that are derived from a DNA or RNA molecule in the sequencing 'library'. The library is a subset of the nucleic acid molecules that have been collected from an environment and made compatible with a particular HTS platform. The HTS instruments deliver these reads as integer 'counts' per genomic feature—gene, location, etc (**?**). However, the counts are actually a single proxy for the probability of observing the particular read in a sample under a repeated sampling model. This is clear since technical replicates of the same library return different counts(**?**), and the difference between technical replicates is consistent with multivariate Poisson sampling (Fernandes *et al.*, 2013; Gloor *et al.*, 2016a). The probability estimate is delivered by the instrument as an integer representation of the probability multiplied by the number of reads (Fernandes *et al.*, 2013; Gloor *et al.*, 2016a). Thus, the data returned by HTS are a type of count compositional data, where only the relationships between the features have meaning (Aitchison,

1986; Lovell *et al.*, 2015; Fernandes *et al.*, 2014; Gloor *et al.*, 2017; Kaul *et al.*, 2017; **?**).

The ALDEx2 tool uses a combination of probabilistic modelling and compositional data analysis to determine the features that are different between groups where that difference is insensitive to random sampling. Technical replicate variance estimation and conversion of the count data to probabilities is accomplished by Monte-Carlo sampling from the Dirichlet distribution (Fernandes *et al.*, 2013; Gloor *et al.*, 2016a), which is conveniently also the conjugate prior for the multivariate Poisson process. The differences between features is linearized by applying a log-ratio transformation to the Dirichlet Monte-Carlo realizations and analyzed according to the rules of compositional data analysis (Aitchison, 1986; Fernandes *et al.*, 2013; Tsilimigras and Fodor, 2016; Gloor *et al.*, 2017).

The 'Group Distribution' panel in Figure 1 shows the distribution for a gene in a highly replicated and curated RNA-seq experiment (Schurch *et al.*, 2016) with the expression of the gene in the WT and knockout conditions shown by the two density distributions. An Anderson-Darling test indicates that a Normal distribution is a poor fit for both distributions ($p < 1e-4$). Consequently, standard effect size measures that depend on summary statistics that assume a Normal distribution are expected to perform poorly and the non-parametric method described here is to be preferred.

We will use the distributions for the gene YDR171W in Figure 1 as an example. Starting with two vectors $\vec{a}$ and $\vec{b}$ that correspond to the log-ratio transformed Dirichlet Monte-Carlo realizations of a feature in two groups, we need a method to determine the standardized effect size; that is, the difference between groups relative to an estimate of within-group dispersion. Since these posterior distributions can have heavy tails, be multimodal, and be skewed, any useful statistic should be insensitive to even extreme non-Normality and provide sensible answers even if the posterior distributions are Cauchy in one or both groups (Fernandes *et al.*, 2013). Below and in the Supplement we define the properties of the approach used.

We can define a non-parametric *difference* vector in Equation (2) as the signed difference between the two groups

$$\vec{diff} = \vec{a} - \vec{b}, \tag{2}$$

with the distribution of the vector shown in Figure 1 'Btw Group Difference'. We can further define a

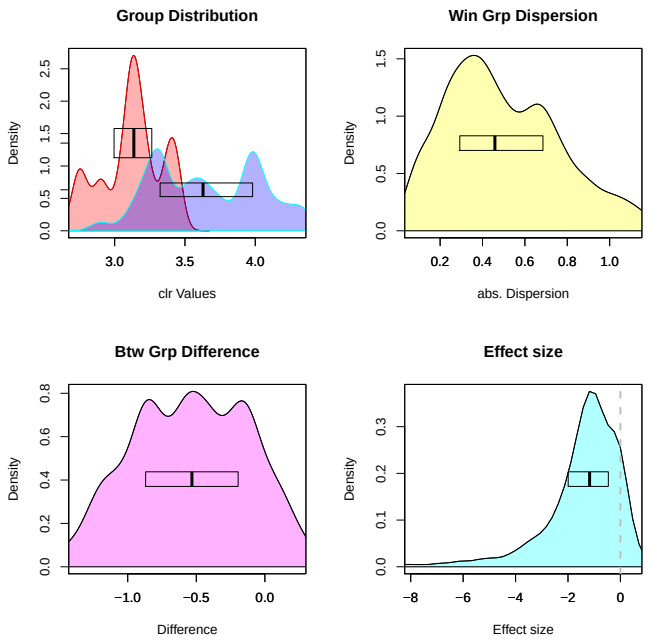

Figure 1. The density of read counts may not follow a Normal distribution. The 'Group Distribution' panel in the top left shows the density of the clr-transformed read counts in the two groups of a highly replicated RNA-seq experiment conduced in *S. cerevisiae* (Schurch *et al.*, 2016) for the gene YDR171W. We can see that the distributions are partially overlapping and are strongly multimodal. The 'Win Grp Dispersion' shows the density of the within group dispersion of the two groups calculated as outlined in equation 3. The'Btw Grp Difference' shows the density of the between group difference calculated as outlined in equation 2. The 'Effect size' shows the density of the effect size calculated as in equation 4. The dashed vertical line in this final panel shows an effect size of 0, and approximately 10% of the effect size distribution crosses this threshold for this gene. The proportion of the effect size distribution that crosses an effect of 0 is known as the 'overlap' measure. For each distribution the median and interquartile range are shown as the thick vertical line and the enclosing box. This plot was generated from the 'aldex.plotFeature()' command.

non-parametric *dispersion* vector as in Equation (3), where the notation $\rho\vec{a}$ indicates one or more random permutations of the vector

$$\vec{disp} = max\{|\vec{a} - \rho\vec{a}|, |\vec{b} - \rho\vec{b}|\}, \tag{3}$$

with the distribution shown in Figure 1 'Win Group Dispersion'. Finally, we can define an *effect* vector as

in Equation (4) that is the element-wise ratio of these two vectors

$$\vec{eff} = \frac{\vec{diff}}{\vec{disp}}, \tag{4}$$

with the distribution of the effect vector shown in Figure 1 'Effect size'.

Taking the median of $\vec{diff}, \vec{disp}$ and $\vec{eff}$ returns a robust estimate of the central tendency of these statistics ($\tilde{D}$, MMAD (median of the maximum absolute deviation), and $\mathbb{E}$), and these are the 'diff.btw', 'diff.win' and 'effect' statistics reported by ALDEx2. The median and interquartile range of these summary statistics for each distribution is shown in Figure 1. $\tilde{D}$ is the very similar to the difference between the means or the difference between medians in a Normal distribution as shown in Supplementary Figure 1. The MMAD metric is novel and the Supplement shows it has a Gaussian efficiency of 52%, a breakdown point of 20% (Supplementary Figure 2), and is 1.42 times the size of the standard deviation on a Normal distribution. The $\mathbb{E}$ statistic is a standardized effect size and is approximately 0.7 of Cohen's d when comparing the difference between two Normal distributions. Below and in Supplementary Figure 3 we show that this metric returns sensible values even with non-Normal distributions.

We used simple simulated datasets to determine baseline characteristics in a number of different distributions. Then we use the data from a highly replicated RNA-seq experiment (Schurch *et al.*, 2016) or from a large 16S rRNA gene sequencing study (Bian *et al.*, 2017) and examined 100 random subsets of the data with between 2 and 20 samples in each group. For each random subset we collected the set of features that were called as differentially abundant at thresholds of $\mathbb{E} \geq 1$, or with an expected Benjamini-Hochburg adjusted p-value of ≤ 0.1 calculated using either the parametric Welch's t-test, or the non-parametric Wilcoxon test in the ALDEx2 R package. These are output as 'we.eBH' and 'wi.eBH' by the ALDEx2 tool. 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. Note that this is simply a measure of consistency and is congruent with the approach taken in (Schurch *et al.*, 2016). We also examined subsets of these datasets where the subsets were taken from the same group. This allowed us to characterize the properties of $\mathbb{E}$ when no difference between groups was expected.

RESULTS AND DISCUSSION

The motivation for this work is to identify what features are reliably different even with the small sample sizes prevalent in high throughput sequencing experiments. Measuring differential abundance in high throughput sequencing datasets is difficult for a variety of reasons. First, almost all experiments are underpowered. Second, the true distribution of the data, and the ground truth of the data, are both unknown. Third, when sample sizes are large almost all features are identified as 'significantly different' by null hypothesis significance testing frameworks. This latter reason is why guidance is generally to ensure that the feature is below a p-value threshold (or below a q, or false discovery rate, threshold and be above a minimum difference threshold (Schurch *et al.*, 2016). Graphically, these cutoffs are represented by the volcano plot (Cui and Churchill, 2003).

We began by examining the behaviour of the $\mathbb{E}$ metric and its constituent statistics. Supplementary Figure 1 shows that the difference between distributions measure is essentially as efficient and stable a measure of location as is the difference between means. When comparing measures of scale, Supplementary Figure 2 shows that the breakdown point for the MMAD is 20% and the efficiency is approximately 52% of the standard deviation in a Normal distribution. Thus, the MMAD is reasonably efficient, and much less prone to contamination than is the standard deviation. Simulation shows that the MMAD is approximately 1.418 larger than the standard deviation for a Normal distribution. Taken together, $\mathbb{E}$ is approximately 0.705 the size of Cohen's d in a Normal distribution, but $\mathbb{E}$ returns sensible estimates even for non-Normal distributions such as β and Cauchy distributions.

The remaining data and figures were generated from two real datasets. The 'yeast' dataset is derived from a highly replicated RNA-seq dataset generated by (Schurch *et al.*, 2016), and the '16S' dataset is derived from a large scale cross-sectional survey of the microbiome of healthy chinese volunteers (Bian *et al.*, 2017). Generically the genes or operational taxonomic units that compose the sequence bins will be referred to as 'features'. These two datasets are polar opposites in many ways, with the yeast dataset having very few 0-count features and low variance within groups, and the 16S dataset being very sparse and having high variance within groups. These two datasets are used to investigate the relationships between three measurable summary statistics of the distributions the difference between groups, the effect

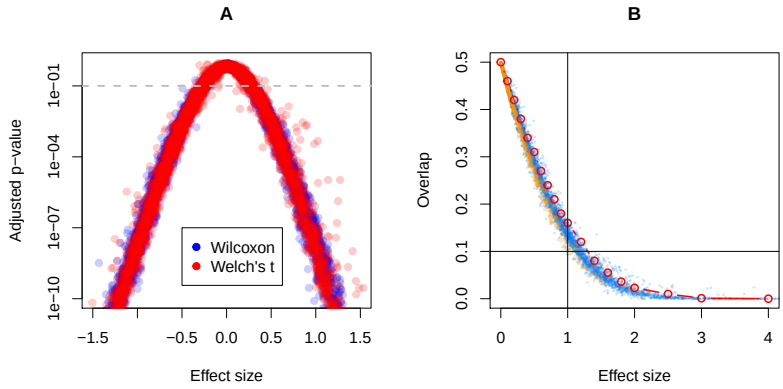

Figure 2. Distributional properties of $\mathbb{E}$. Panel A shows the relationship between $\mathbb{E}$ and Benjamini-Hochberg adjusted p-values calculated by either a non-parametric Wilcoxon test (blue points), or a parametric Welch's t-test (red points), where each point represents one of the genes in the yeast transcriptome dataset. The y axis has been truncated to highlight the p-values greater than 1e-10, and the dashed grey line shows the location of a false positive threshold of 0.1. Panel B shows the relationship between $\mathbb{E}$ and the overlap measure in the yeast dataset (blue points) and the 16S dataset (orange points). Overlaid in red is the expected relationship for a Z-score in a Normal distribution.

size and the overlap, and how these values can be used to identify reproducibly different features in high throughput sequencing datasets.

Figure 2: A shows the relationship between $\mathbb{E}$ and p-values for the 6349 genes in the yeast dataset. We can see that there is very good correspondence such that features with very high effect sizes have very low adjusted p-values. This is in line with the expected relationship between effect sizes and p-values, and provides additional confidence that $\mathbb{E}$ is an appropriate metric for effect size. However, note that the non-parametric Wilcoxon test adjusted p-values (in blue) have far fewer outliers on this plot than do the parametric Welch's t-test adjusted p-values (in red). We conclude that the majority of features likely have distributions that do not deviate to much from the Normality assumption of the parametric test, but that a significant minority of features have enough deviation to affect the calculated p-value. These outliers could contribute to both false positive and false negative identifications when using a parametric null hypothesis testing approach.

We next focussed on the characteristics of false positive features in the two datasets and examined

the relationship between the $\mathbb{E}$ and the overlap metrics in both the yeast and 16S datasets. Recall from Figure 1 that all values are calculated from the distributions and not inferred from summary statistics. The overlap is the proportion of the a $\mathbb{E}$ distribution where the tail overlaps 0. In a Normal distribution, Cohen's d is exactly equivalent to a Z score, and we can determine the proportion of the tail distribution that corresponds to any Z score (effect size). This relationship is plotted in Figure 2:B and we can see that the non-parametric $\mathbb{E}$ and overlap metrics correspond very well to the expected relationship for a Z score and tail area in a Normal distribution. On this graph, an overlap of 0.1 corresponds to $\mathbb{E} \sim 1.2$.

Next, we examined the proportion features that were identified as being false positives as a function of per-group sample size if there was no difference between groups. For this, we generated 100 random instances from both datasets with the samples draws from only one group and the per-group sample size varying from 2 to 20. We calculated the proportion of features all features in the dataset that had a greater than 2-fold difference, or an overlap less than 0.1, or that had $\mathbb{E}$ values greater than 0.5, 1, or 2, or that had the intersect of the difference > 1, overlap < 0.1 and $\mathbb{E}$ greater than 1. This was plotted relative to the per-group sample size in Fig 3:A. A number of observations can be made. First, it is apparent that there is a strong linear relationship between the proportion of features that are identified as false positive and sample size for all the metrics. Second, a smaller number of samples was needed to ensure no false positive features were identified as the effect size increased; at the extreme an effect size of 2 would be sufficient to exclude FP features with as few as 5 samples per group. Third, difference was a poor measure by which to exclude FP features, being worse as a measure in the highly variable 16S dataset. Fourth, the effect size was highly reproducible as a measure to exclude FP features having almost the same characteristics in both datasets. The overlap measure was also highly reproducible, but this is a trivial finding since Figure 2 shows that effect size and overlap are highly correlated. Fifth, combining the three measures was able to exclude FP features better than was any single metric. However, the triple measure of effect, overlap and difference did not give the same result in the two datasets. In the low variance yeast dataset the triple measure was, if anything, even more specific than was doubling the effect size. In contrast, the triple measure was only slightly more specific in the high variance dataset than were the single measures of $\mathbb{E}$ or overlap.

We used effect plots (Gloor *et al.*, 2016b) to identify why the triple measure was more specific in

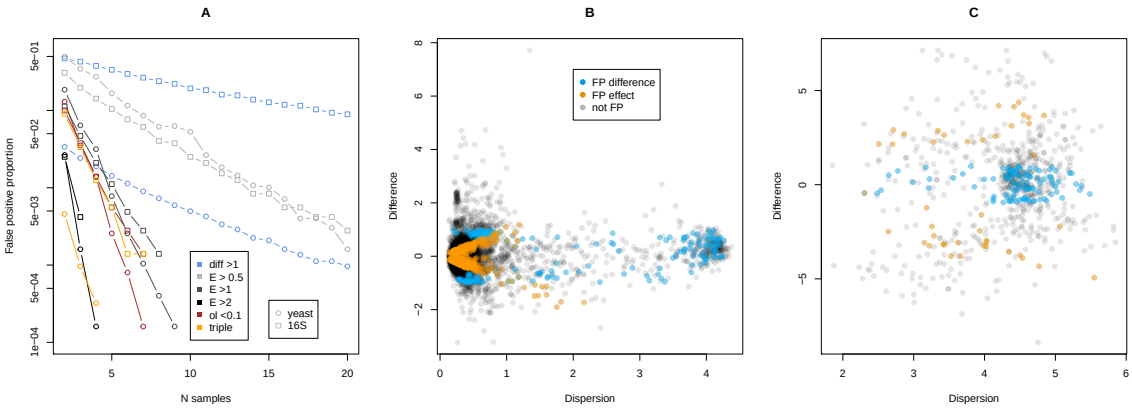

Figure 3. Characteristics of false positive features using $\mathbb{E}$. Panel A plots per-group sample size and the FP rate of difference, and the median values when applied to a samples drawn at random from the same group. Also shown is the combination of these three metrics into one metric that is the overlap between the three (triple). Panel B shows an effect plot (Gloor *et al.*, 2016b) of the whole yeast transcriptome dataset in grey points, with the false positive features identified by either difference between groups (blue) or $\mathbb{E}$ (orange) identified from a random subset of 5 samples from each group. Panel C shows the same analysis on the 16S rRNA gene sequencing dataset. The false positive features identified by each approach are restricted to features with quite separate characteristics of difference and dispersion.

the yeast dataset than in the 16S dataset and the results are shown in Figure 3 panels B and C. Here we overplotted the FP features found when an example dataset of 5 in each group was compared to the TP features found when the full dataset was examined and applying the dual cutoffs of difference and $\mathbb{E} > 1$. In the yeast dataset, the FP $\mathbb{E}$ features have very low dispersion (variance), and very low difference, while the features identified as FP when only difference was used tended to have either very high or very low dispersion. In the 16S dataset, essentially all the features have very high dispersion. Here we found that the FP identified by $\mathbb{E}$ have a large between group difference, and the FP features identified only by difference tend to have low difference between groups, and mirrors the observation seen in the yeast dataset. These observations explain why using both $\mathbb{E}$ and difference in combination are more discriminatory than either alone, as they tend to identify different sets of FP features.

With this information, we can determine the sensitivity and specificity of identifying TP and FP features in the example datasets when comparing two groups. In the yeast dataset, the edgeR tool identified over 4600 out of 6349 genes as 'significant' (Benjamini-Hochberg adjusted p-value < 0.05) when all samples were included using either the glm or exact test modes (Supplementary Table 1). Other widely used tools gave similar results (Schurch *et al.*, 2016). The null hypthesis testing framework in ALDEx2 also returned at least 4300 genes in the same dataset. Thus, identifying such a large proportion of genes as differentially abundant indicates that statistical significance is not informative for this type of experiment. Schurch et al. (and others) recommend adding a secondary threshold such as a fold-change cutoff to identify genes of interest for follow-up analyses (Cui and Churchill, 2003; Schurch *et al.*, 2016). When sample sizes are sufficiently large, we would expect that the fold-change cutoff itself would be the primary determinant of difference; however, this approach would not include either the biological variance or the uncertainty of measurement in the analysis. Furthermore, the difference metric is not sufficient to exclude FP features and this is especially relevant for features with large dispersion.

We examined the relationship between sample size and the number of features identified as significantly different using a null hypothesis testing framework in the yeast dataset. Figure 4:A shows the median True and False Positive rates. In this analysis we are testing for the ability to detect features that would have been observed as differentially abundant in the full dataset if we use a random subset of the data, and the plots show the median value of 100 trials at each sample size. The 'q+diff' example in the Proportion

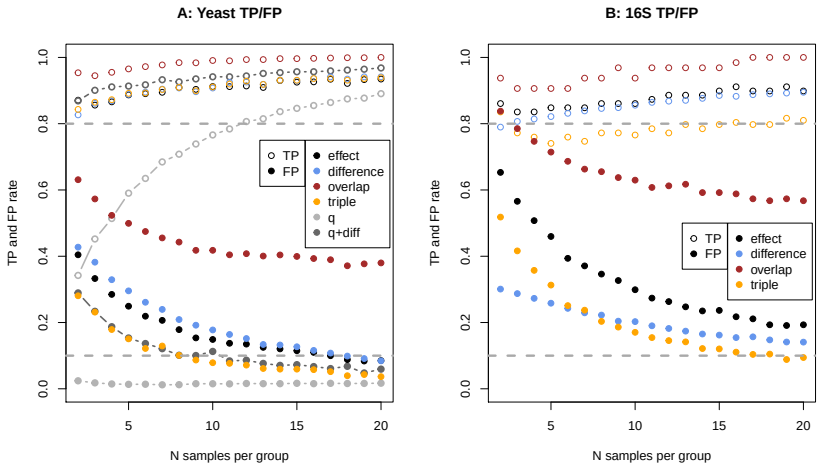

Figure 4. True Positive (TP) and False Positive (FP) identifications as a function of per group sample size in the two datases. The two panels show the results of randomly sub-sampling the yeast transcriptome and the 16S rRNA gene datasets 100 times with various numbers of features in each group. Features that were identified in the subsample and in the full dataset were counted as true positives (TP), and features that were identified in the subsets that were not identified as different in the full dataset were counted as false positives (FP). The panels show the median proportion of TP found by each approach, and the median proportion of all positives that were FP. Cutoffs used were absolute effect > 1, absolute difference > 1, overlap of < 0.1, or q score of < 0.1 (Benjamini-Hochberg adjusted p-value). The triple approach used the intersect of the effect, difference and overlap approaches.

panel shows the rate observed for the Benjamini-Hochberg corrected p-values (q-values) and a 2-fold difference observed when using the edgeR tool (REF), as advised in recent best practices (Schurch *et al.*, 2016). As expected when using p-values alone, we observe that the power of null hypothesis significance test is strongly affected by sample size, and only reaches 90% power to detect when the per-group sample size is greater than 20. However, the FP rate is effectively 0. Interestingly, applying both the significance test and the fold-change cutoff caused both the TP and FP rates to increase substantially. At small sample sizes the TP rate its greater than 80% and the FP rate ranges from $\sim 30\%$ and down. Inspection of the results indicated that this was because in this dataset the significance test was all-but irrelevant since all features with at least a 2-fold change had a q-score below the threshold of 0.1. Note that many tools have difficulty estimating the actual FDR in real datasets (Thorsen *et al.*, 2016; Hawinkel *et al.*, 2018).

In contrast to q-scores alone, the TPR of the $\mathbb{E}$ statistic in the same random datasets is essentially independent of the number of samples for all methods and combinations. However, now the FPR is strongly affected by sample size as was observed for q-scores and difference. Note that even when only two samples are used, the $\mathbb{E}$ statistic identifies over 80% of the features as different as are identified by the same statistic in the full dataset. Thus, the simple metric outlined here can correctly identify the 'true positive' set even when the number of samples is very small. The tradeoff when using this statistic is that at very low sample sizes the False Discovery Rate (fdr) is extreme; in this dataset and with and with a cutoff of $\mathbb{E} > 1$, the fdr is 40% with two samples, but falls to less than 10% only when there are 15 or more samples. Interestingly, applying a fold-change and an overlap metric (denoted as triple in the figure) cutoff to the $\mathbb{E}$ metric reduces the false discovery rate dramatically, and it is now indistinguishable from the FP rate observed with the q-score and difference combination. A similar trend is observed with the 16S dataset, where the metrics have extremely good power to detect even at low sample sizes, but have a high false positive rate at low sample sizes. We conclude that the $\mathbb{E}$ measure, alone or in combination with the difference and overlap metrics constitute a useful and reproducible way to identify differentially abundant features in disparate datasets.

CONCLUSION

By default, we want to know both 'what is significant' and 'what is different' (Cui and Churchill, 2003). Both of these questions can be addressed with a standardized effect size statistic that scales the difference between features by their dispersion. We have found plots of difference and dispersion to be an exceeding useful tool when examining HTS datasets (Gloor *et al.*, 2016b). Furthermore, datasets analyzed by this approach have proven to be remarkably reproducible as shown by independent lab validation (Macklaim *et al.*, 2013; Nelson *et al.*, 2015).

The $\mathbb{E}$ statistic outlined here is a relatively robust statistic with the attractive property that it consistently identifies almost all the same set of true features regardless of the underlying distribution and the number of samples as shown in Figure 4. In marked contrast, even the best p-value based approaches can identify only a small proportion of the features at small samples sizes that would have been found in the full dataset (Schurch *et al.*, 2016). Thus, the simple metric outlined here can correctly identify the 'true positive' set even when the number of samples is very small. Note that fold-change thresholds as is commonly used, is not the same as an standardized effect statistic, and applying the threshold values of (Schurch *et al.*, 2016) while reducing the features that are found does not necessarily enhance reproducibility. In fact this investigation highlights the danger in relying on fold-change to identify differentially abundant features. We can see that the 16S rRNA gene sequencing datasets have substantially greater numbers of fold-change FP features than does the yeast transcritpome dataset when no difference in expected. This is because of the substantially greater dispersion observed for the features in the former dataset than in the latter.

The tradeoff when using the $\mathbb{E}$ statistic is that at very low sample sizes the False Discovery Rate can be extreme; in this dataset and with and with a cutoff of $\mathbb{E} > 1$, the FDR is 40% with two samples, but falls to less than 10% only when there are 15 or more samples regardless of dataset. Note that the FDR as measured is assessing congruence with the result in the whole dataset since the actual ground truth is not know. Given that FP are not identified if there is no difference between groups when the sample size is greater than 10 (Figure 3:A), it is likely that the FP identified when the groups are different are simply features with lower effect sized. Supplementary Figure 5 shows that this is in fact the case. Moreover,

using the combination of effect, difference and overlap enhances specificity regardless of dataset. This is because these measures are filtering out different sets of FP features, but identify substantially the same set of TP features: the $\mathbb{E}$ metric is the mid point of the effect size distribution and identifies those features with large standardized change between groups; the overlap metric corresponds to the tail of the $\vec{eff}$ distribution and identifies those features with narrow distributions; the difference between metric identifies those features with large absolute fold change.

Further tempering this, is the observation that no false positives are identified when no difference is expected in two different datasets when there are 10 or more samples per group. Taken together, we suggest that a fold change of at least two, and both $\mathbb{E} > 1$ and overlap < 0.1 are robust and reproducible measures that provide an acceptable mix of power and specificity when the sample size is greater than 10 per group in diverse datasets.

This work describes the $\mathbb{E}$ statistic that measures a standardized effect size directly from distributions and not from summary statistics. We show that it is useful when examining high throughput sequencing datasets. The statistic is relatively robust and efficient, and answers the question most desired by the biologist, namely 'what is reproducibly different'. The $\mathbb{E}$ metric is computed in the ALDEx2 R package as the 'effect' output where it is the median of the inferred technical and biological data, and in the CodaSeq R package where it acts only the point estimates of the data. Interactive exploration of effect sizes can be done in the omicplotR Bioconductor package (Giguere *et al.*, 2019).

ACKNOWLEDGEMENTS

To be determined

FUNDING

This work has been supported by NSERC, CIHR.

REFERENCES

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Bian, G., Gloor, G. B., Gong, A., Jia, C., Zhang, W., Hu, J., Zhang, H., Zhang, Y., Zhou, Z., Zhang, J.,

Burton, J. P., Reid, G., Xiao, Y., Zeng, Q., Yang, K., and Li, J. (2017). The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere*, **2**(5), e00327–17.

Coe, R. (2002). It's the effect size, stupid: What effect size is and why it is important.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *R Soc Open Sci*, **1**(3), 140216.

Cui, X. and Churchill, G. A. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, **4**(4), 210.1 – 210.10.

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci*, **3**(4), 286–300.

Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., and Gloor, G. B. (2013). Anova-like differential expression (aldex) analysis for mixed population rna-seq. *PLoS One*, **8**(7), e67019.

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.

Giguere, D., Macklaim, J., and Gloor, G. (2019). omicplotr: Visual exploration of omic datasets using a shiny app. Bioconductor v1.4.0.

Gloor, G. B., Macklaim, J. M., Vu, M., and Fernandes, A. D. (2016a). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics*, **45**, 73–87.

Gloor, G. B., Macklaim, J. M., and Fernandes, A. D. (2016b). Displaying variation in large datasets: Plotting a visual summary of effect sizes. *Journal of Computational and Graphical Statistics*, **25**(3C), 971–979.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., and Egozcue, J. J. (2017). Microbiome datasets are compositional: And this is not optional. *Front Microbiol*, **8**, 2224.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nat Methods*, **12**(3), 179–85.

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2018). A broken promise : microbiome differential abundance methods do not control the false discovery rate. *BRIEFINGS IN BIOINFORMATICS*.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, **2**(8), e124.

Kaul, A., Mandal, S., Davidov, O., and Peddada, S. D. (2017). Analysis of microbiome data in the presence of excess zeros. *Front Microbiol*, **8**, 2114.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., and Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol*, **11**(3), e1004075.

Macklaim, J. M., Fernandes, A. D., Di Bella, J. M., Hammond, J.-A., Reid, G., and Gloor, G. B. (2013). Comparative meta-rna-seq of the vaginal microbiota and differential expression by lactobacillus iners in health and dysbiosis. *Microbiome*, **1**(1), 12.

Nakagawa, S. (2004). A farewell to Bonferroni: the problems of low statistical power and publication bias. *Behavioral Ecology*, **15**(6), 1044–1045.

Nelson, T. M., Borgogna, J.-L. C., Brotman, R. M., Ravel, J., Walk, S. T., and Yeoman, C. J. (2015). Vaginal biogenic amines: biomarkers of bacterial vaginosis or precursors to vaginal dysbiosis? *Frontiers in physiology*, **6**.

Schurch, N. J., Schofield, P., Gierliński, M., Cole, C., Sherstnev, A., Singh, V., Wrobel, N., Gharbi, K., Simpson, G. G., Owen-Hughes, T., Blaxter, M., and Barton, G. J. (2016). How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA*, **22**(6), 839–51.

Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, **14**, 91.

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., Sørensen, S., Bisgaard, H., and Waage, J. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome*, **4**(1), 62.

Tsilimigras, M. C. B. and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann Epidemiol*, **26**(5), 330–5.

𝔼 Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., Lozupone, C., Zaneveld, J. R., Vázquez-Baeza, Y., Birmingham, A., Hyde, E. R., and Knight, R. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, **5**(1), 27.