Compositional transformations can reasonably introduce phenotype-associated values into sparse features

George I. Austin[1,2], Tal Korem[2,3,†]

Author affiliations
[1] Department of Biomedical Informatics, Columbia University Irving Medical , New York, NY, USA
[2] Program for Mathematical Genomics, Department of Systems Biology, Columbia University Irving Medical Center, New York, NY, USA
[3] Department of Obstetrics and Gynecology, Columbia University Irving Medical Center, New York, NY, USA
[†] Corresponding author: tal.korem@columbia.edu

Abstract

It was recently argued[1] that an analysis of tumor-associated microbiome data[2] is invalid because features that were originally very sparse (genera with mostly zero read counts) became associated with the phenotype following batch correction[1]. Here, we examine whether such an observation should necessarily indicate issues with processing or machine learning pipelines. We show counterexamples using the centered log ratio (CLR) transformation, which is often used for analysis of compositional microbiome data[3]. The CLR transformation has similarities to Voom-SNM[4,5], the batch-correction method brought into question[1,2], yet is a sample-wise operation that cannot, in itself, "leak" information or invalidate downstream analyses. We show that because the CLR transformation divides each value by the geometric mean of its sample, common imputation strategies for missing or zero values result in transformed features that are associated with the geometric mean. Through analyses of both synthetic and vaginal microbiome datasets we demonstrate that when the geometric mean is associated with a phenotype, sparse and CLR-transformed features will also become associated with it. We re-analyze features highlighted by Gihawi et al.[1] and demonstrate that the phenomenon of sparse features becoming phenotype-associated can also be observed after a CLR transformation, which serves as a counterexample to the claim that such an observation necessarily means information leakage. While we do not intend to address other concerns regarding tumor microbiome analyses[1,6], validate Poore et al.'s[2] results, or evaluate batch-correction pipelines, we conclude that because phenotype-associated features that were initially sparse can be created by a sample-wise transformation that cannot artifactually inflate machine learning performance, their detection is not independently sufficient to demonstrate information leakage in machine learning pipelines. Microbiome data is multivariate, and as such, a value of zero carries a different meaning for each sample. Many transformations, including CLR and other batch-correction methods, are likewise multivariate, and, as these issues demonstrate, each individual feature should be interpreted with caution.

Introduction

In two critiques published last year[1,6], Gihawi et al. raised several concerns regarding an analysis of tumor microbiome in The Cancer Genome Atlas (TCGA) data[2]. Among these concerns, they highlight several taxa that have mostly zero counts in raw data, but that following batch correction have values that are correlated with specific tumor types. They claim that this is sufficient to indicate information leakage[1,6]. This implies a general principle that we wished to tackle: is finding that a transformation turns a sparse feature into a feature that is associated with a phenotype sufficient to conclude that it introduced information leakage[7]?

We demonstrate that this is not the case by constructing a counterexample, which involves no information leakage but still produces the same observation. To this end, we use the centered log ratio (CLR) transformation[8], a transformation that is simple, widely used in the microbiome field, bears similarities to the transformation in question[4,5], and, most importantly, is a sample-wise transformation that does not use any metadata labels and cannot, in itself, induce information leakage.

Most microbiome datasets have an arbitrary total read count that is not reflective of sample properties, and should therefore be interpreted as relative abundances that are compositional[3,8,9]. Compositional data violates many of the assumptions underlying common data analysis strategies, and, for example, exhibit a negative correlation bias and sub-compositional incoherence[8,9]. As components of compositional data can only be understood relative to one another, compositional transformations will generally transform them with respect to a reference - in the case of CLR, with respect to the geometric mean of the sample[10]. Additionally, as most compositional transformations involve the use of a logarithm, they cannot handle zeros. Therefore, a common strategy is to introduce pseudocounts prior to the transformation. Previous studies have highlighted the difficulty in interpreting compositionally transformed features, which tend to contain information determined by the rest of the sample[9–11]. As a result, some have advocated for alternatives to normalization-based methods, such as quantifying scale uncertainties[12,13], although CLR and other normalization methods remain commonplace. In general, these intricacies have complicated inference from microbiome data, which recent work has demonstrated to be an ongoing challenge[14].

Here, we start by showing via simulations that as CLR-transformed sparse features are, by definition, negatively correlated with the geometric means of their corresponding samples, they could be associated with a phenotype that is itself associated with the geometric mean. Furthermore, as the geometric mean is related to α diversity[15], we demonstrate, through an analysis of a vaginal microbiome study[16], that CLR-transformed sparse features can be associated with a phenotype in cases where the sample α diversity is also associated with it. Finally, we reanalyze examples of sparse features highlighted by Gihawi et al.[1], and show that in cases where the α diversity was also associated with the phenotype in question, the phenomenon observed by Gihawi et al. could also be observed when performing CLR transformation. Thus, we show that the CLR transformation, which does not utilize any phenotype labels or information from other samples and therefore has no risk of

independently inducing information leakage, can reasonably transform sparse features to non-sparse features with phenotype associations. While we recommend caution with univariate analysis or other attempts to interpret these transformed sparse features or to assign them any biological significance, their existence does not, on its own, discredit a data processing pipeline or downstream machine learning models as they are the product of a multivariate transformation.

Results

CLR-transformed sparse features are strongly associated with a phenotype in a simulated dataset

We simulated a dataset with 100 samples, 50 with a positive label and 50 with a negative label (**Fig. 1a,b**). By construction, the data had one completely empty feature, and higher geometric means for the positive samples (**Methods**) with perfect separation from the negative samples (median [range] geometric means of 0.018 [0.012-0.022] and 0.00090 [0.00069-0.00099] for positive and negative samples, respectively; Mann-Whitney *U* $p < 10^{-12}$; **Fig. 1c**). As commonly done prior to CLR transformation, we added a pseudocount of 10^{-6} to all values (**Methods**). For the empty feature, dividing the pseudocounts by the geometric mean of each sample during CLR transformation created large differences between the positive and negative samples (Median [range] values -9.80 [-9.99 – -9.41] vs. -6.80 [-6.90 – -6.53], respectively; $p < 10^{-12}$; **Fig. 1d,e**). Of note, the same results hold when the added feature is non-zero ($p = 2.1\times10^{-17}$; **Fig. S1a,b**), partially zero ($p = 1.3\times10^{-5}$; **Fig. S1c,d**), as well as for a range of pseudocount values ($p < 0.01$ for pseudocounts ranging from 10^{-8} to 0.01; **Fig. S2a**). Based on this perfect separation, in most reasonable machine learning models this feature would be highly predictive of our synthetically constructed label. However, this does not discredit the validity of neither the models nor of the transformation we performed. The CLR transformation is a sample-wise transformation, which does not observe labels or values from other samples. It therefore cannot "leak" information from the labels or a test set.

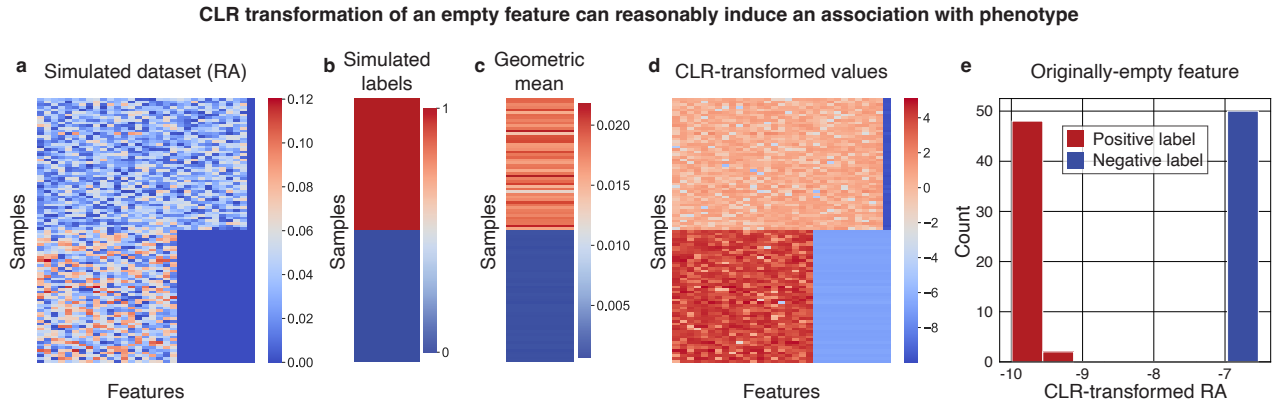

Figure 1 | CLR transformation following a pseudocount can make a sparse feature highly predictive. **a,b,** Heatmaps visualizing the relative abundances (RA; **a**) and labels (**b**) of our simulated dataset, in which there are 10 features present only in the "positive" samples, and one "empty" feature that is zero for all samples (shown in the rightmost column). **c,** The geometric mean of each sample, which is by construction highly correlated with the labels. **d,** Heatmap of the relative abundances (a) after a CLR

transformation in which they are divided by the geometric mean (c) following the addition of a pseudocount. **e,** Histogram of the CLR-transformed empty feature (rightmost column of **d**) which was originally all zero in the simulated dataset (rightmost column of a). The dataset was constructed to have different geometric means between the samples with synthetic positive and negative labels. Thus, a CLR transform creates a perfect separation between this feature's values in the positive and negative samples (Mann-Whitney *U* $p < 10^{-12}$).

CLR-transformed sparse features can be associated with clinical phenotypes

After observing that the CLR transformation of empty features can be informative for predicting an *in silico* phenotype label, we sought to demonstrate that similar conclusions can hold in real microbiome datasets. To do this, we wished to analyze a dataset that demonstrated an association between a phenotype and Shannon α diversity, which is related to the geometric mean as it is the logarithm of the inverse of the geometric mean of a sample that is weighted by the relative abundances[15]. The biological relevance of α diversity to the microbiome has been extremely well documented across a wide range of scenarios (e.g., refs. [16–24]). Specifically, we reanalyzed data from a study of the vaginal microbiome and preterm birth, and examined the first time point from each of the 40 individuals included in the study[16] (**Methods**). As previously reported[25–27], we observed an association between α diversity and subsequent preterm birth (**Fig. 2a**, Mann-Whitney *U* $p = 0.037$).

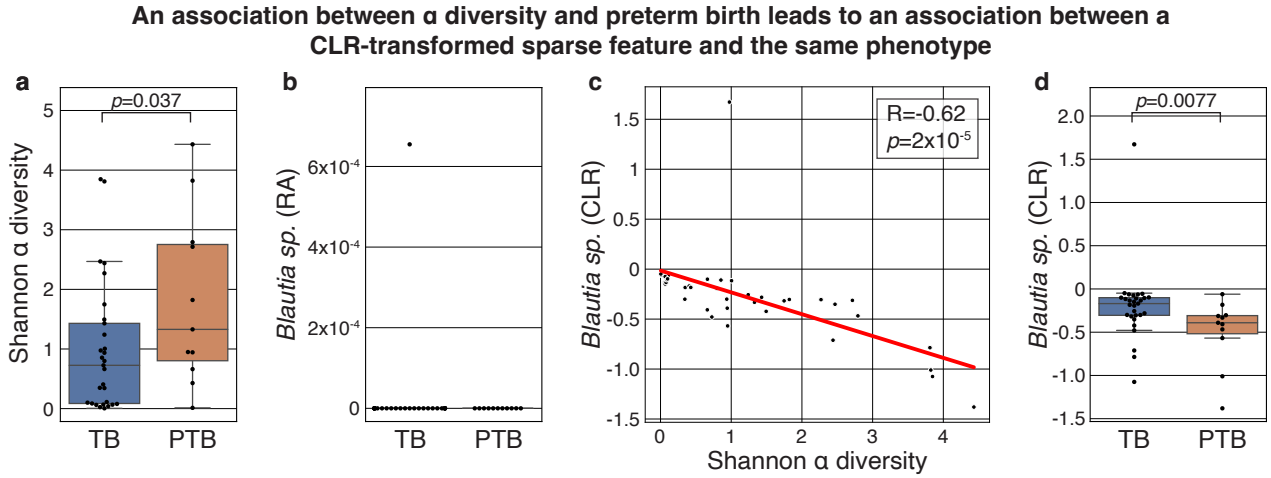

Figure 2 | CLR transformation generates associations between a sparse taxon and preterm birth. a, Box and swarm plots of the α diversity of vaginal microbiome samples collected during pregnancy, separated by subsequent term (TB) or preterm birth (PTB). As was previously noted, the α diversity of the vaginal microbiome is associated with preterm birth (Mann-Whitney *U* $p = 0.037$). **b**, Box and swarm plots showing the relative abundance (RA) of a *Blautia sp.* (OTU 4465907), a sparse feature that was only detected in a single sample. **c**, Scatterplot and fitted Ordinary Least Squares curve of the same *Blautia sp.* that was CLR transformed (y axis) and the α diversity of the same sample (x-axis). Because the α diversity is related to the geometric mean[15], we observe a strong negative correlation (Pearson's R = -0.62, $p = 2.0\times10^{-5}$). **d**, Same as **b**, showing the CLR-transformed relative abundances of the same *Blautia sp.* Since preterm birth is associated with α diversity and α diversity is negatively associated with the CLR-transformed sparse feature, the latter becomes associated with preterm birth (Mann-Whitney *U* $p = 0.0077$). Box, IQR; line, median; whiskers, nearest point to 1.5*IQR.

We next selected a sparse taxon, which was only detected in a single sample: OTU 4465907, which was identified by the authors[16] as a *Blautia sp.* (**Fig. 2b**). When we performed a CLR transformation of this dataset (**Methods**), we noted that as expected, values were introduced to this feature, and because Shannon α diversity is related to the geometric mean, the CLR transformed *Blauria sp.* became strongly negatively correlated with the α diversity (**Fig. 2c**, Pearson's R = -0.62, $p = 2.0\text{x}10^{-5}$). As a result, we now also observe a negative association between *Blautia sp.* and subsequent preterm birth (**Fig. 2c**, Mann-Whitney U $p = 0.0077$). As before, this result holds across a wide range of pseudocounts ($p < 0.01$ for pseudocounts ranging from 10^{-8} to 0.01, **Fig. S2b**). Importantly, we note that once again, the compositional transformation of the sparse feature that we implemented is a simple sample-wise operation that is not observing any patient metadata or information from any other sample, and it is not erroneous to use this feature as a predictor for preterm birth. While it would be incorrect to interpret this association as indicating anything about the biology of *Blautia*, which is not necessarily even present in these vaginal samples, it would be accurate to interpret this as an indication of a "microbiome-wide" signature, in that the log of the inverse of geometric mean is associated with the outcome of interest. While these multivariate compositional transformations can introduce challenges with interpretation and inference, our results demonstrate one example of a valid explanation for why the transformation of an empty feature is associated with a biological phenotype on a real microbiome dataset.

The associations between tumor type and originally-sparse genera highlighted by Gihawi et al. can be reasonably explained by a CLR transformation

Recently, Gihawi et al.[1] claimed that Poore et al.[2] had major errors in data analysis of microbiome data from TCGA. Among other critiques, Gihawi et al. claimed that features that were very sparse (almost entirely zero) in the raw data were erroneously filled with an artificial tag that leaked prior information about tumor type[1], which they supported with four specific examples. Above we already showed, via counterexamples, that observing such associations is not sufficient to prove a data analysis error or leakage in machine learning analysis. Nevertheless, we wished to determine if Gihawi et al.'s observations could be explained by interpretable attributes of the samples' overall compositions, as opposed to an erroneous incorporation of sample metadata as was suggested. CLR is particularly relevant to the processing performed by Poore et al. because it bears similarity to Voom[4]. To this end, we reanalyzed the four examples Gihawi et al. provided in their Figures 2 - 5 (**Methods**). We note, however, that while this analysis evaluates this specific critique by Gihawi et al., it does not serve as a validation of the analysis by Poore et al., or of the normalization and batch-correction pipeline they used.

First, Gihawi et al. examined the values of *Hepandensovirus* in adrenocortical carcinoma (ACC). Upon reanalysis, we notice an underlying difference in α diversity between ACC and all other samples (Mann-Whitney U $p = 2.6\text{x}10^{-4}$; **Fig. 3a**). Despite the fact that *Hepandensovirus* is only detected in a single sample, and is likely not truly present in tumors[6] (**Fig. 3b**), a CLR transformation of this feature results

in values that are different in ACC compared to other samples (**Fig. 3c**, $p = 3.2\text{x}10^{-14}$) because the α diversity is related to the geometric mean. Therefore, the same association highlighted by Gihawi et al. (reproduced in **Fig. 3d**) can be observed in a compositional transformation that we know cannot cause information leakage because it does not consider values from any other samples, nor any sample metadata, but only the values of the sample itself. Of course, this does not mean that *Hepandensovirus* should be interpreted as biologically relevant to the tumor microbiome. Instead, the multivariate transformation of the data generated an association that reflects a shift in the global microbiome composition of the sample.

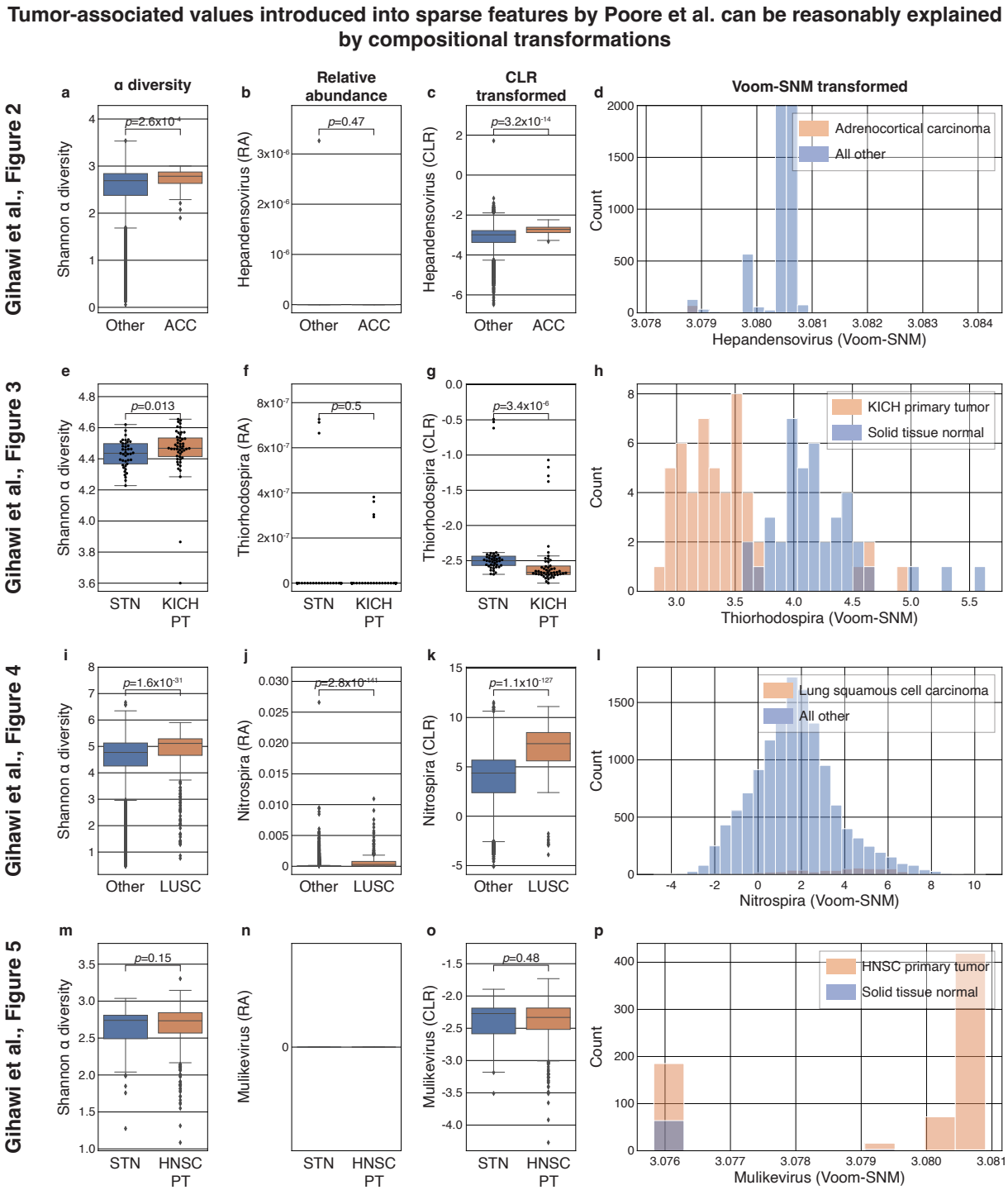

Figure 3 | CLR transformation can reproduce predictive sparse features highlighted by Gihawi et al. Each row shows the same analyses applied to a different scenario analyzed by Gihawi et al.[1] in their Figures 2-5, respectively (**Methods**). **a,e,i,m,** Boxplots of the α diversity of samples. **b,f,j,n,** Boxplots of the relative abundance (RA) of the taxon analyzed. **c,g,k,o** The CLR transformed values of the same taxa. **d,h,l,p,** Histograms replicating figures 2-5, respectively, in Gihawi et al.[1], with some differences noted in **Methods**. In the three cases in which significant differences were observed in α diversity (a,e,i), the CLR transformation produced tumor type-associated values (o,g,k) from originally sparse features (b,f,j). Box, IQR; line, median; whiskers, nearest point to 1.5*IQR. Individual dots are plotted if <= 100 samples per plot, only outliers displayed if > 100 samples.

Second, Gihawi et al. examined the values of *Thiorhodospira*, comparing whole genome sequencing (WGS) data from primary kidney chromophobe (KICH) tumor samples and normal tissue samples from the same group of patients. Upon reanalysis, we find a similar trend: an underlying difference in sample α diversity (Mann-Whitney U $p = 0.013$; **Fig. 3e**) is related to the observation that despite being originally very sparse (**Fig. 3f**) this feature becomes highly associated with KICH following CLR transformation ($p = 3.4\text{x}10^{-6}$; **Fig. 3g**). Similarly large differences were observed by Gihawi et al. (reproduced in **Fig. 3h**).

Third, Gihawi et al. examined the values of *Nitrospira*, comparing between lung squamous cell carcinoma (LUSC) samples and all other primary tumor samples. Once more, we observe a significant association of LUSC with sample α diversity (Mann-Whitney U $p = 1.6\text{x}10^{-31}$; **Fig. 3i**). Interestingly, while Gihawi et al. write that "in the raw data, there is no such shift" in *Nitrospira*[1], in this case we saw significant association with LUSC even when examining raw relative abundances ($p = 2.8\text{x}10^{-141}$; **Fig. 3j**), a difference that was present also in the raw counts. While this difference in relative abundance undermines our ability to attribute observed downstream differences solely to compositional transformations, we nevertheless continue to observe a significant association of the CLR-transformed values of *Nitrospira* with LUSC ($p = 1.1\text{x}10^{-127}$; **Fig. 3k**). As before, we conclude, using the same features highlighted in their manuscript, that the observation by Gihawi et al. (reproduced in **Fig. 3l**) cannot be attributed to data analysis error or information leakage without additional support.

Fourth, Gihawi et al. examined the values of *Mulikevirus*, comparing data from primary head and neck squamous cell carcinoma (HNSC) tumor samples and normal tissue samples from the same group of patients. In this case, we did not find the same phenomenon of associations between sample α diversity and CLR-transformed *Mulikevirus* with HNSC (**Fig. 3m-o**). However, this still is not necessarily an indication of information leakage, since we have only examined a single compositional transformation which is far less complex than the full set of transformations performed by Poore et al.

Finally, we note that while all our analyses so far used a single pseudocount that was introduced in relative abundance space (**Methods**), with the goal of isolating the impact of CLR, we observe statistically significant differences in the same three of out four taxa when varying the choice of pseudocount across most values ranging from 10^{-8} to 0.01 (**Fig. S2c-f**). We observe even greater differences when introducing a pseudocount of one to the raw counts (**Fig. S3**), because this produces differences in the relative abundances of sparse features as a result of different underlying read counts – also commonly associated with sample-wise factors such as α diversity.

Gihawi et al. did not perform an information-free analysis

Gihawi et al. conclude their analysis of sparse features by performing a classification analysis on "information-free raw data"[1]. For completeness, we note that while it may appear that this analysis includes the application of a machine learning pipeline to a matrix of zeros, this was not the analysis performed. As Gihawi et al. specify: "We then populated each cell in the empty matrix with its

corresponding value from the Voom-SNM normalized data"[1], indicating that the analysis performed was of the Voom-SNM normalized data that was subset to features that were originally zero. We note that this corresponds to an analysis of the empty feature in our simulated dataset (**Fig. 1e**), which, as we showed, contains legitimate information that can perfectly classify the label. These conclusions made by Gihawi et al. rely upon the assumption that all zeros in microbiome data should be interpreted identically. However, this assumption does not always hold for microbiome datasets, for which certain features can be zero due to undersampling, while some are structural zeros. Thus, it is common for transformations of microbiome datasets to encode additional information into features that were initially sparse, even if it is simply based on numerical summaries of the remainder of the corresponding sample, as in the CLR transform. Therefore, it is important to interpret microbiome features as multivariate. Taken together with other results presented here, we thus show that this analysis by Gihawi et al. is insufficient to demonstrate flaws in the normalization process used by Poore et al.

Discussion

Determining the validity of a transformation that induces associations with a phenotype in non-informative features bears importance for evaluation of machine learning pipelines. We present three analyses demonstrating that such a phenomenon is an expected outcome of a commonly used transformation, the CLR transformation, especially for cases in which the phenotype is associated with the geometric mean of the taxonomic composition or its α diversity. As the CLR is a sample-wise operation, which does not use an outcome label or information from other samples, it cannot independently introduce information leakage. Through these counter-examples, we therefore demonstrate that observing phenotype-associated values in previously sparse features should not be considered sufficient to conclude that an analysis suffered from an artificial inflation of predictive signals.

Our reanalysis of the information leakage claims raised by Gihawi et al. strongly suggests that their observations are an expected and reasonable effect of using Voom-SNM, a compositional transformation that performs linear adjustments in log-space, especially as we observed an association between the relevant phenotype and α diversity in most cases. Nevertheless, this work does not serve to validate the tumor microbiome associations reported by Poore et al.[2,28], nor their normalization or batch-correction pipeline. We therefore consider other concerns raised by Gihawi et al., such as potential contamination[6] or errors in the genome database used leading to misclassification of human reads as bacteria[1], out of scope for this analysis. Our conclusion is that Gihawi et al. did not provide robust evidence of an artificial tag, information leakage, or any synthetic inflation of predictive results in the analysis conducted by Poore et al.[2]. However, we did not demonstrate that such leakage did not exist. Furthermore, both our and Gihawi et al.'s analyses demonstrate challenges in the biological interpretation of univariate associations detected with specific taxa.

The pivotal assumption underlying Gihawi et al.'s analysis and conclusion is that features in microbiome samples are univariate and independent of each other. However, because microbiome samples are compositional, the abundance of any taxon can only be evaluated within the context of the measured abundances of other taxa within the same sample. It therefore should be considered reasonable to encode sample-wide information into transformations of features observed within multivariate and compositional data, such as microbiome data. This implies that values measured as zeros in microbiome data do not necessarily mean the same for different samples, for example due to different sequencing depths and α diversity. Additionally, some zeros may be structural (microbe is not present in the ecosystem) as opposed to sampling zeros (microbe is present in very low abundance). Such distinctions are an ongoing challenge in the field that are important for interpretation, but are independent from the issue of information leakage in this case. Therefore, evaluating if features which are initially zero are transformed into distinct values should generally be interpreted from a multivariate perspective in microbiome data, and therefore observing phenotype-associated values should not be considered a proof of a data analysis error.

There are more direct ways to check whether a machine learning pipeline is invalid or "leaks" information than examining specific features. Our recommendation, for any machine learning analysis, is to perform a permutation test in which the phenotype labels are shuffled before running any data normalization and machine learning pipelines. Under such analyses a machine learning model is expected to yield random predictions, and any observed signal is a strong indication of information leakage. We and others have also previously noted that predictive analyses are more robust if cross-validation is done across batches[7,29].

Finally, we note that there exist many reasonable variations to the data processing steps that we used in this work. We have chosen to use the CLR transformation in this work due to its simplicity, widespread use, and the fact that it is strictly sample-wise. We note, however, that it does have its downsides, as it imposes constraints on the data that make strong implicit assumptions, and are still subject to a sum constraint. Alternatives to compositional transformations have recently been proposed that instead model scale uncertainty[12,30]. We also note that while we evaluated a few variations of pseudocount methods across which our results are robust (**Figs. S2, S3**), many alternative strategies are available, including some that do not require pseudocounts[31,32]. However, this does not change our conclusions, as the processing choices used in this work do not induce any sort of information leakage, and are therefore sufficient to serve as a counter-example to the claims by Gihawi et al.[1]. Since there exists at least one reasonable transformation that introduces phenotype-associated values into sparse features, observing such a case should not be considered independently sufficient to challenge the results of a predictive pipeline.

References

1. Gihawi, A. *et al.* Major data analysis errors invalidate cancer microbiome findings. *MBio* e0160723 (2023).
2. Poore, G. D. *et al.* RETRACTED ARTICLE: Microbiome analyses of blood and tissues suggest cancer diagnostic approach. *Nature* **579**, 567–574 (2020).
3. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).
4. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).
5. Mecham, B. H., Nelson, P. S. & Storey, J. D. Supervised normalization of microarrays. *Bioinformatics* **26**, 1308–1315 (2010).
6. Gihawi, A., Cooper, C. S. & Brewer, D. S. Caution regarding the specificities of pan-cancer microbial structure. *Microb Genom* **9**, mgen001088 (2023).
7. Whalen, S., Schreiber, J., Noble, W. S. & Pollard, K. S. Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* **23**, 169–181 (2022).
8. Aitchison, J. *The Statistical Analysis of Compositional Data*. (Springer, 1986).
9. Pawlowsky-Glahn, V. & Buccianti, A. *Compositional Data Analysis: Theory and Applications*. (John Wiley & Sons, 2011).
10. Quinn, T. P. *et al.* A field guide for the compositional analysis of any-omics data. *Gigascience* **8**, (2019).
11. Greenacre, M. *Compositional Data Analysis in Practice*. (CRC Press, 2018).
12. Nixon, M. P., Gloor, G. B. & Silverman, J. D. Beyond Normalization: Incorporating Scale Uncertainty in Microbiome and Gene Expression Analysis. *bioRxiv* 2024.04.01.587602 (2024) doi:10.1101/2024.04.01.587602.
13. Nixon, M. P. *et al.* Scale Reliant Inference. (2022).
14. Wirbel, J., Essex, M., Forslund, S. K. & Zeller, G. A realistic benchmark for differential abundance testing and confounder adjustment in human microbiome studies. *Genome Biology* **25**, 1–26 (2024).
15. Tuomisto, H. A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity. *Ecography* **33**, 2–22 (2010).
16. DiGiulio, D. B. *et al.* Temporal and spatial variation of the human microbiota during pregnancy. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 11060–11065 (2015).
17. Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell* **178**, 795-806.e12 (2019).
18. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486**, 222–227 (2012).
19. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
20. Fettweis, J. M. *et al.* The vaginal microbiome and preterm birth. *Nat. Med.* **25**, 1012–1021 (2019).

21. Liao, J. *et al.* Microdiversity of the vaginal microbiome is associated with preterm birth. *Nat. Commun.* **14**, 4997 (2023).

22. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

23. Clemente, J. C. *et al.* The microbiome of uncontacted Amerindians. *Sci Adv* **1**, (2015).

24. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).

25. Sun, S. *et al.* Race, the Vaginal Microbiome, and Spontaneous Preterm Birth. *mSystems* **7**, e0001722 (2022).

26. Golob, J. L. *et al.* Microbiome preterm birth DREAM challenge: Crowdsourcing machine learning approaches to advance preterm birth research. *Cell Rep Med* **5**, 101350 (2024).

27. Kan, H. *et al.* Differential Effect of Vaginal Microbiota on Spontaneous Preterm Birth among Chinese Pregnant Women. *Biomed Res. Int.* **2022**, 3536108 (2022).

28. Sepich-Poore, G. D. *et al.* Robustness of cancer microbiome signals over a broad range of methodological variation. *Oncogene* **43**, 1127–1148 (2024).

29. Austin, G. I. & Korem, T. Planning and Analyzing a Low-Biomass Microbiome Study: A Data Analysis Perspective. *J Infect Dis* (2024) doi:10.1093/infdis/jiae378.

30. Gloor, G. B., Nixon, M. P. & Silverman, J. D. Normalizations are not what you think; Explicit Scale Simulation in ALDEx2. *bioRxiv* 2023.10.21.563431 (2023) doi:10.1101/2023.10.21.563431.

31. Austin, G. I. *et al.* Processing-bias correction with DEBIAS-M improves cross-study generalization of microbiome-based prediction models. *bioRxiv* 2024.02.09.579716 (2024) doi:10.1101/2024.02.09.579716.

32. Hu, Y., Satten, G. A. & Hu, Y.-J. LOCOM: A logistic regression model for testing differential abundance in compositional microbiome data with false discovery rate control. *Proceedings of the National Academy of Sciences* **119**, e2122788119 (2022).

33. biocore/scikit-bio: scikit-bio 0.5.9: Maintenance release. doi:10.5281/zenodo.8209901.

Methods

Synthetic simulations

We generated a simple dataset of 100 synthetic samples, of which 50 had a positive phenotype label and 50 negatives (**Fig. 1b**). Among the 50 positive samples, we simulated data by drawing 30 features from i.i.d uniform distributions in [0,1], in addition to a 31st empty feature of zero counts for all samples. To simulate different geometric means for the samples with negative phenotypes, we drew 20 features from i.i.d uniform distributions in [0,1], with the remaining 11 features being empty. By construction, this leaves the 31st feature as '0' for all samples (**Fig. 1a**). We then transformed to relative abundance space, such that all rows summed to one, added a pseudocount of 10^{-6} to all samples, corresponding to 10 to the largest power that keeps the pseudocount below the dataset's smallest

observed relative abundance value. We then applied the CLR transformation using the `skbio.stats.composition.clr` function[33] (**Fig. 1d**). Following this transformation, we examined the association of the originally-zero 31st feature with the simulated labels (**Fig. 1e**). We note that our observed differences would be larger if we added the pseudocount before transforming to relative abundance space, although doing so would induce differences already in the relative abundance table prior to the CLR transformation, which would confound our conclusions. For all analyses, we re-ran the same pipelines using pseudocounts ranging across all powers of 10 from from 10^{-8} to 0.01 (**Fig. S2**).

Analysis of the vaginal microbiome and preterm birth

We obtained publicly available microbiome read counts and metadata from Datasets S1 and S2 of a study profiling the vaginal microbiome along 40 pregnancies[16]. We filtered this dataset to consider only the first vaginal sample collected from each patient, and among those kept the 222 taxa with at least one read observed in a sample. We considered a sparse feature, OTU 4465907, which was identified by the authors as a *Blautia sp.*, and which was only observed in a single sample (**Fig. 2b**). We transformed the data to relative abundance space, added a pseudocount to all samples using the same strategy as before, which yielded a pseudocount of 10^{-4}, and then ran the `skbio.stats.composition.clr` function. We then compared the associations of α diversities, clr-transformed relative abundances of *Blautia sp.*, and preterm birth using, as applicable, Mann-Whitney *U* and Pearson's R.

Reanalysis of The Cancer Genome Atlas reanalysis

We obtained publicly available microbiome read counts and metadata from the original analysis by Poore et al.[2], in order to reanalyze the four examples highlighted in the second critique by Gihawi et al.[1] We made our best effort to match the analyses performed by Gihawi et al.[1] by replicating their histograms (**Fig. 3d,h,l,p**). We performed the following analyses :

1. *Hepandensovirus* and ACC (Fig. 2 in Gihawi et al.[1], our **Figs. 3a-d, S2c, S3a,b**). We analyzed the "Most Stringent Decontamination" file of Poore et al.[2], comparing ACC samples to all other tumor types. To match the appearance of **Fig. 3d** to Fig. 2 in Gihawi et al., we artificially limited the upper limit of the x-axis to 3.0845, which removed off one outlying point, and the upper limit of the y-axis to 2,000.
2. *Thiorhodospira* and KICH (Fig. 3 in Gihawi et al.[1], our **Figs. 3e-h, S2d, S3c,d**). We analyzed the "All Putative Contaminants Removed" file of Poore et al.[2], comparing KICH primary tumor WGS data to normal tissue samples from patients with KICH.
3. *Nitrospira* and LUSC (Fig. 4 in Gihawi et al.[1], our **Figs. 3i-l, S2e, S3e,f**). We analyzed the "All Putative Contaminants Removed" file of Poore et al.[2], comparing LUSC samples to all other tumor types.
4. *Mulikevirus* and HNSC (Fig. 5 in Gihawi et al.[1], our **Figs. 3m-p, S2f, S3g,h**). We analyzed the "Most Stringent Decontamination" file of Poore et al.[2], comparing data from HNSC primary tumor to normal tissue samples from patients with HNSC. We note that **Fig. 3p** is different from

Fig. 5 in Gihawi et al.[1] because we maintained default plotting parameters to display visually consistent spacings throughout the entire plot.

For each of the four analyses, we obtained the raw read counts for the corresponding samples, removed features marked as `contaminants`, and measured their Shannon diversities (**Fig. 3a,e,i,m**), relative abundances (**Fig. 3b,f,j,n**), and CLR of the relative abundances following either a pseudocount of 10 to the largest power that keeps the pseudocount below the dataset's smallest observed relative abundance value (**Fig. 3c,g,k,o**), or a pseudocount of one introduced before relative abundance normalization (**Fig. S3**). Mann-Whitney *U* was used for all pairwise comparisons.

Code availability

All code used in this analysis is available at www.github.com/korem-lab/compositional-empty-features

Data availability

The vaginal microbiome data used in this analysis is available in Datasets S1 and S2 of the original publication[16]. TCGA data used in this analysis is available from the original publication[2].

Acknowledgements

We thank members of the Korem group for useful discussions. We thank all authors and participants involved in the generation of data used in this study. This work was supported by the Program for Mathematical Genomics at Columbia University (T.K.) and T15LM007079 (G.I.A.).

Author contributions

G.I.A. and T.K. conceived and designed the study, designed analyses, interpreted the results, and wrote the manuscript. G.I.A. conducted all analyses.

Competing interests

The authors declare no competing interests.

Supplementary Figures

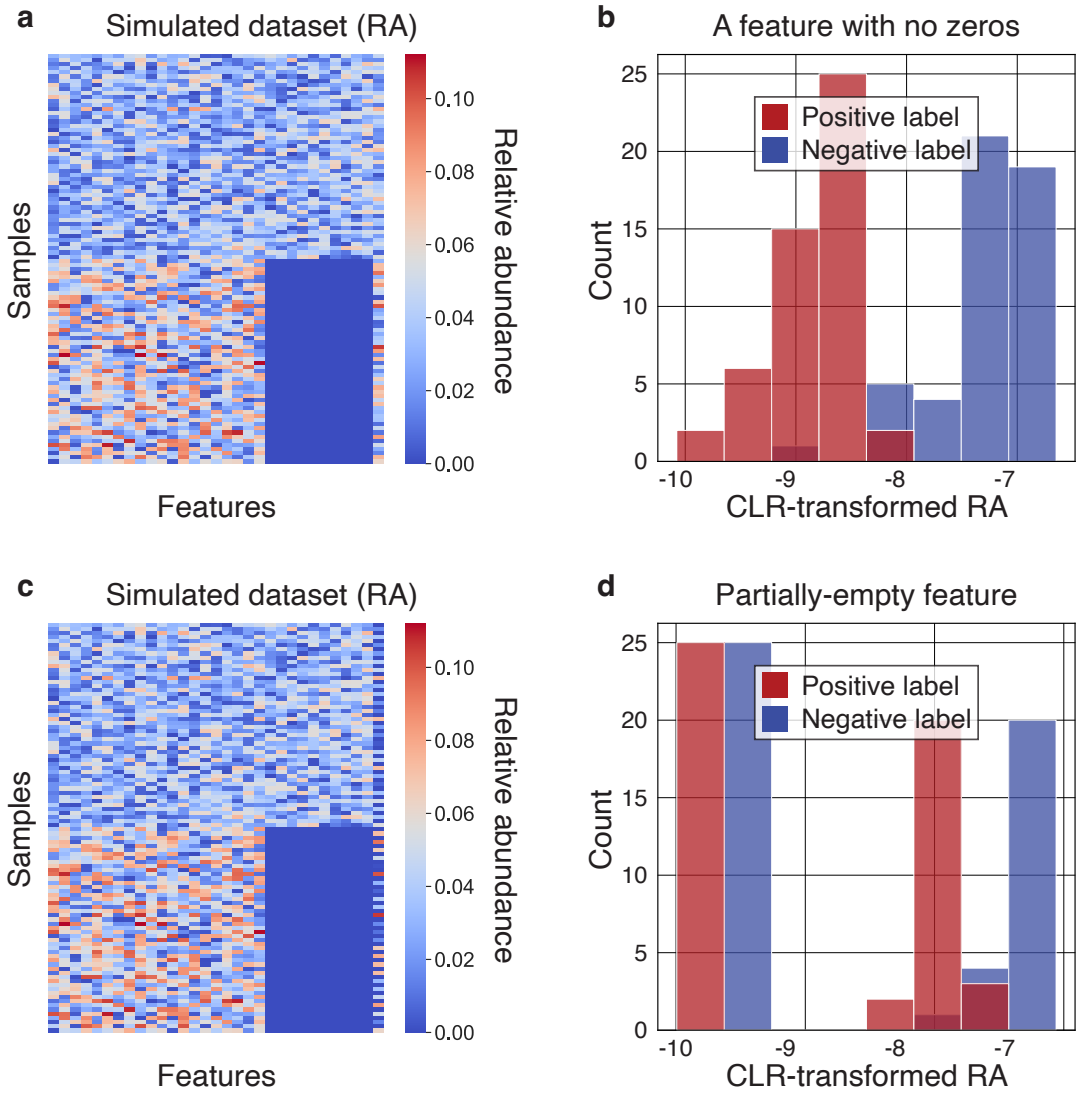

Figure S1 | Compositional transformations induce differences between groups for non-sparse and partially sparse features. a,b Same as in **Fig. 1a,d**, for a 31st feature that is drawn from the same distribution as the remaining features in "count space" before relative abundance transformation (and, in that, is equivalent to the first 20 features). **c,d** Same as in **Fig. 1a,d**, for a 31st feature that is drawn from the same distribution as the remaining features in "count space" before relative abundance transformation for 50% of samples within either group, and zeros within the remaining 50% of values.

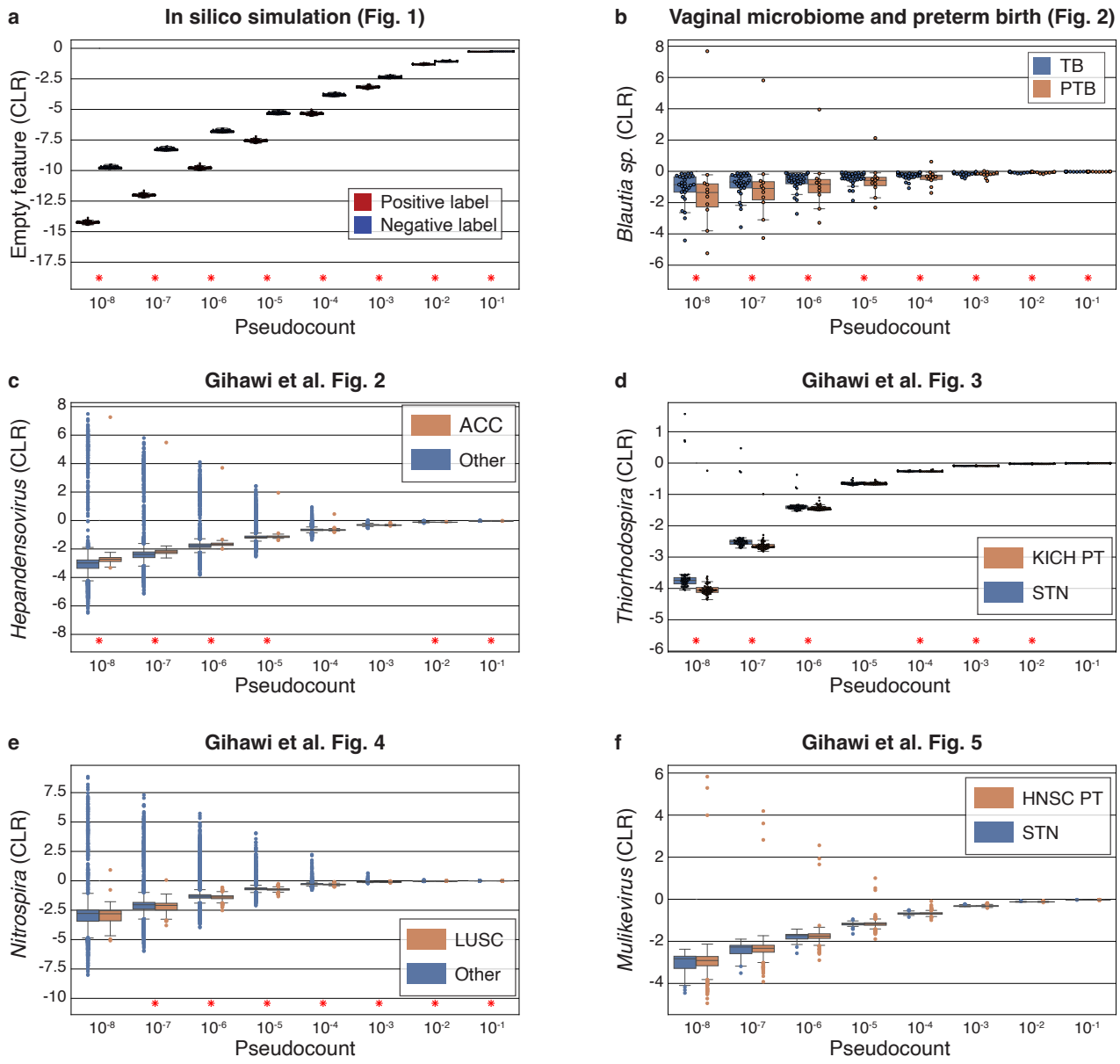

Figure S2 | Phenotype-associated values are introduced into sparse features under a range of pseudocount values. a-f, Results from the analyses performed in **Figs. 1e, 2d, and 3c,g,k,o**, but performed across a range of pseudocounts from 10^{-8} to 0.1. Box, IQR; line, median; whiskers, nearest point to 1.5*IQR. Individual dots are plotted if <= 100 samples per plot, only outliers displayed if > 100 samples. * denotes $p < 0.01$ for two-tailed Mann-Whitney *U* between groups.

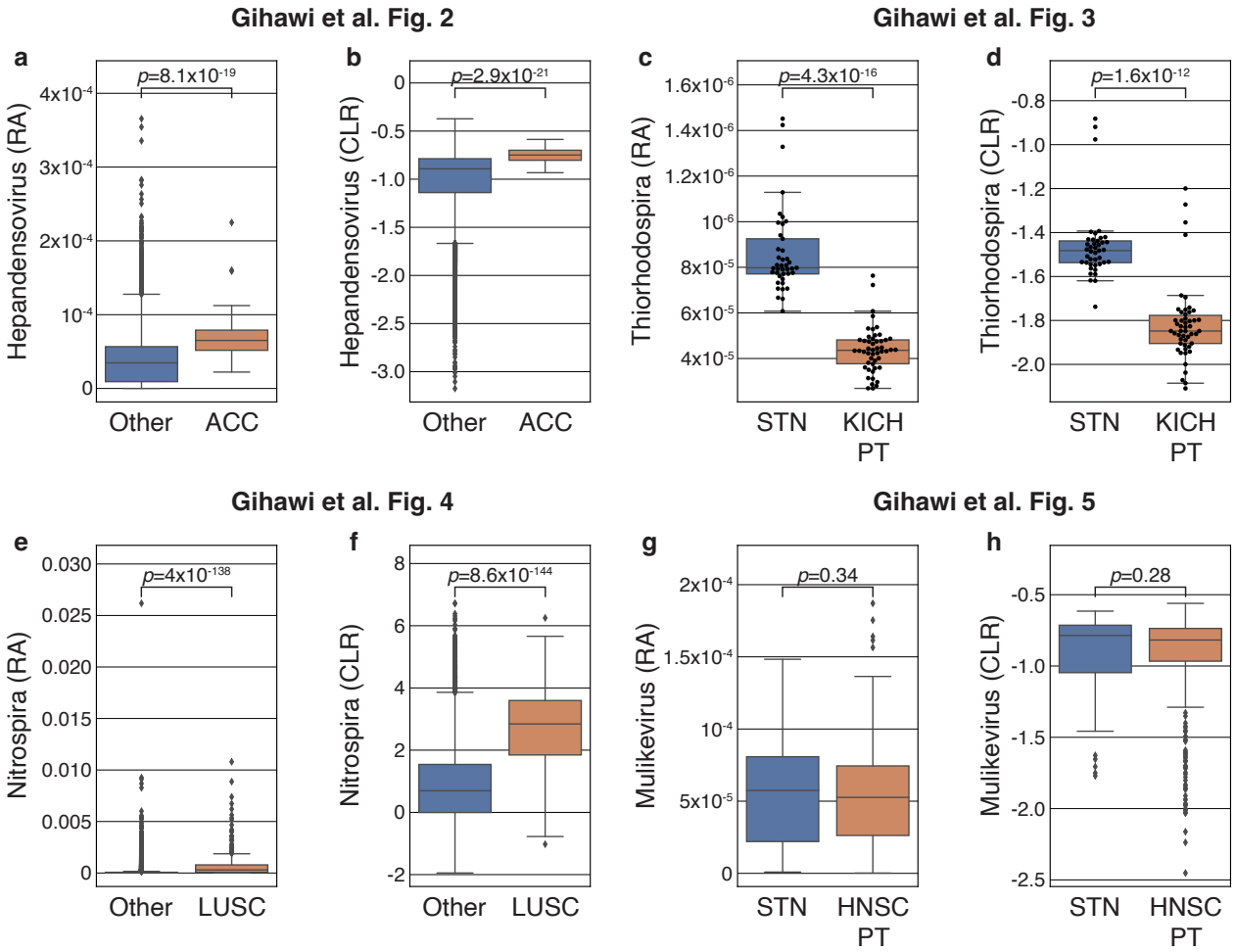

Figure S3 | An alternative pseudocount strategy increases the tumor-type associations of sparse taxa. a,c,e,g Boxplots similar to F**ig. 3b,f,j,n** respectively, but using a pseudocount of one in raw count space rather than introducing a pseudocount in relative abundance space. **b,d,f,h,** Boxplots similar to F**ig. 3c,g,k,o**, respectively, but using a pseudocount of one in raw count space rather than a pseudocount in relative abundance space (**Methods**). Introducing a pseudocount before normalization allows for the possibility of producing different relative abundances for sparse taxa due to differences in sample read counts, which we observe to further increase the separation across groups. While this approach is more consistent with standard Voom implementations, we have not used it in our main analyses so that all the variation induced into sparse features is explained by CLR, which is not the case here.