Published online 2 October 2020

NAR Genomics and Bioinformatics, 2020, Vol. 2, No. 4 **1**
doi: 10.1093/nargab/lqaa079

An application of compositional data analysis to multiomic time-series data

Laura Sisk-Hackworth and Scott T. Kelley*

Department of Biology, San Diego State University, San Diego, CA 92182, USA

Received March 01, 2020; Revised August 07, 2020; Editorial Decision August 31, 2020; Accepted September 07, 2020

ABSTRACT

Compositional data analysis (CoDA) methods have increased in popularity as a new framework for analyzing next-generation sequencing (NGS) data. CoDA methods, such as the centered log-ratio (clr) transformation, adjust for the compositional nature of NGS counts, which is not addressed by traditional normalization methods. CoDA has only been sparsely applied to NGS data generated from microbial communities or to multiple 'omics' datasets. In this study, we applied CoDA methods to analyze NGS and untargeted metabolomic datasets obtained from bacterial and fungal communities. Specifically, we used clr transformation to reanalyze NGS amplicon and metabolomics data from a study investigating the effects of building material type, moisture and time on microbial and metabolomic diversity. Compared to analysis of untransformed data, analysis of clr-transformed data revealed novel relationships and stronger associations between sample conditions and microbial and metabolic community profiles.

INTRODUCTION

Technological advances in DNA-sequencing technologies, such as next-generation sequencing (NGS), combined with clever primer barcoding strategies have allowed major advances in our understanding of microbial communities across numerous environments (1). NGS allows the generation of vast numbers of DNA sequences from microbiological samples, anything from seafloor sediments (2) to hospital surfaces (3), which can be used to identify and enumerate hundreds or thousands of bacteria, fungi, archaea or virus species in a given sample (4). This data is compositional in nature: due to sequencing depth limitations, an increase in the measurement of one taxa results in a decrease in the measurement of another even if the absolute abundance of one taxa is unchanged. Furthermore, each NGS sample has a different library size and every sequence-based taxonomic abundance count represents a relative rather than an absolute abundance, meaning that taxa counts in one sample are not directly comparable to counts in other samples. Other data types used for microbiome analysis, such as data from untargeted metabolomics and transcriptomics, are also relative rather than absolute and therefore compositional (5). Typical methods used to normalize NGS-generated count data between samples, including rarefaction, spike-in normalization, normalization by library size and transcripts per million, often do not address the compositional nature of the data (6). Since NGS instruments can only sequence reads to a certain capacity, a higher number of reads of one sequence will impact the number of other reads that can be sequenced. The limitations of standard normalization methods for compositional data are described thoroughly elsewhere (6–8).

One promising approach to dealing with compositional data, such as nucleotide sequencing libraries or untargeted metabolite data from complex microbial communities, is the centered log-ratio (clr) transformation. By recasting relative count data with respect to a reference (the sample geometric mean), the clr transformation converts compositional data to scale-invariant data in real space, thereby allowing application of multivariate statistical methods (7–9). Recently, clr transformation has gained traction in the analysis of sequencing data for both RNA-seq and genomics (10–13), though fewer studies have used such transformations for microbial metagenomics and multiomics analysis (14,15). Multiomics is the application of multiple different dataset types, such as metabolomics, metagenomics and transcriptomics, to the same biological samples. While multiomics analyses can offer insight into the relationship between different levels of biology, integrating these data can be challenging. Normalization methods for one 'omics' type, such as RNA-seq, may introduce spurious results when applied to metagenomics and *vice versa* (16,17). Applying clr transformation rather than normalization to any 'omics' dataset will result in scale-invariant datasets and facilitate multiomics data integration.

Here, we explored the impact of clr transformation and other compositional data analysis (CoDA) methods on the analysis of microbiome multiomic data by reanalyzing bacterial and fungal community NGS datasets and a metabolite dataset from a prior study that had used standard normalization techniques. We reanalyzed these data using clr

*To whom correspondence should be addressed. Tel: +1 619 594 5371; Fax: +1 619 594 5676; Email: skelley@sdsu.edu

transformation and applied recently developed methods for multiomics data integration and selection of microbial balances. Our results showed that the combination of data transformation and multiomics analyses revealed novel biological patterns and interactions not identified with standard normalization approaches.

MATERIALS AND METHODS

The 'omics' portion of the original study included count data from NGS of bacterial (16S) and fungal (ITS) amplicon libraries, as well as counts of untargeted metabolomic HPLC-MS/MS analysis, obtained from microbial community samples collected longitudinally from four common building materials: gypsum, mold-free (MF) gypsum, medium-density fibreboard and plywood (18). The materials were inoculated post-sterilization by passive settling in three locations: a laboratory (control) and two residences (locations 1 and 2). Half of the samples were submerged in water after inoculation for 12 h and half were kept dry. In this study, the authors analyzed the effect of inoculation location, material type and wetting status on overall microbial and metabolite diversity and identified co-occurrences between microbes and metabolites. Samples at time point zero were taken just after materials were brought into the lab, and materials were then sampled five additional times, approximately every 5 days. The original data contained 144 samples of 6200 metabolite observations, 330 samples of 26 578 fungal operational taxonomic units (OTUs) and 338 samples of 6466 bacterial OTUs. Unequal sample sizes resulted from failures in sequencing reactions and selected use of metabolomic analyses. Data from the original paper, including OTU tables for the fungal/bacterial datasets and metabolite features, were downloaded from FigShare (https://doi.org/10.6084/m9.figshare.7865015.v2).

Two different methods of zero-handling were compared: zero-replacement using the pseudo-counts method from the R package zCompositions (19) version 1.3.3 and replacement of zeroes with ones. Since we did not observe any substantial differences between zero-handling methods in the NMDS ordination plots, all subsequent analyses used data transformed after zero-handling with zCompositions pseudo-counts. Clr transformation was performed separately on all three datasets: metabolites, fungi and bacteria. To compute the clr transformation for each sample, each count value in that sample was divided by the geometric mean of the sample, then the natural log of that ratio was taken (9):

$$\mathrm{clr}\ (X_j) = \left[\ln\left(\frac{X_{1j}}{g(X_j)}\right), \ldots, \ln\left(\frac{X_{Dj}}{g(X_j)}\right) \right]$$

where X_j is a sample in a dataset, $g(X_j)$ is the geometric mean of that sample, X_{1j} is the first value in a sample and X_{Dj} is the last value in a sample of D-values.

As a guide to the reader, we have provided a diagram of the various datasets and analyses used in this study (Figure 1).

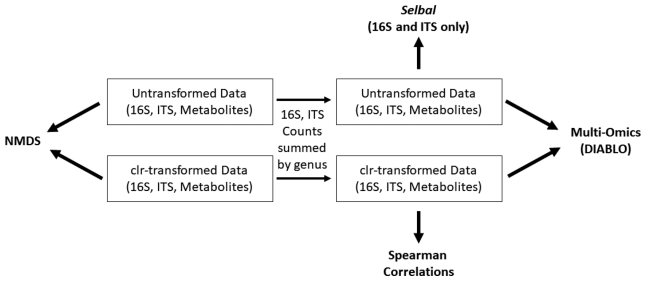

Figure 1. Schematic of the analyses performed on the clr-transformed and untransformed datasets. NMDS and DIABLO analyses were performed on both types of datasets. Only untransformed data were inputted into selbal, and only clr-transformed data were used to calculate Spearman correlations.

NMDS ordination plots

The clr-transformed values following zero replacement with the zCompositions pseudo-counts or the one-count approach, in which zeroes were replaced with ones, were used to generate NMDS ordination plots with the R vegan package (version 2.5-6) using Euclidean distances (20). NMDS plots were created in R using ggplot2 version 3.2.1c (21), and samples were colored by time point, location of initial material inoculation, material type and wetting condition, respectively (see Lax *et al.* (18) for details). We also generated NMDS plots using Bray–Curtis distance with untransformed data rarefied to 1000 for the bacterial and fungal datasets. Permutational multivariate analysis of variance (PERMANOVA) was performed for each condition (time, material, location and wetting status) in every dataset (bacteria, fungi and metabolites) using the R package vegan with 9999 permutations for all time points except time zero, with the P-values corrected for multiple comparisons using the false discovery rate (FDR) Benjamini–Hochberg method. The PERMANOVA is a multivariate test used to determine if the centroid or dispersion of a set of samples is equivalent among specified categories (e.g. time points or material types). In this case, the centroid and dispersion were estimated using the between-sample Euclidean and Bray–Curtis distances.

Spearman correlations

Genera of the fungal and bacterial taxa present in >10% of samples and the 50 most abundant metabolites were selected for correlation analysis. 16S and ITS IDs in the OTU tables were replaced with the genus name using the Greengenes (22) and Unite (23) databases, then the clr-transformed count values for each sample were summed by genus. Spearman correlations were computed using the R package corrplot (24) on wet samples only. For each material, we computed correlations between genera from three different combined multiomics datasets: (i) bacteria and fungi, (ii) metabolites and bacteria, and (iii) metabolites and fungi. P-values were adjusted with the Bonferroni correction.

Multiomics integration

We analyzed the correlation structure of both the un-

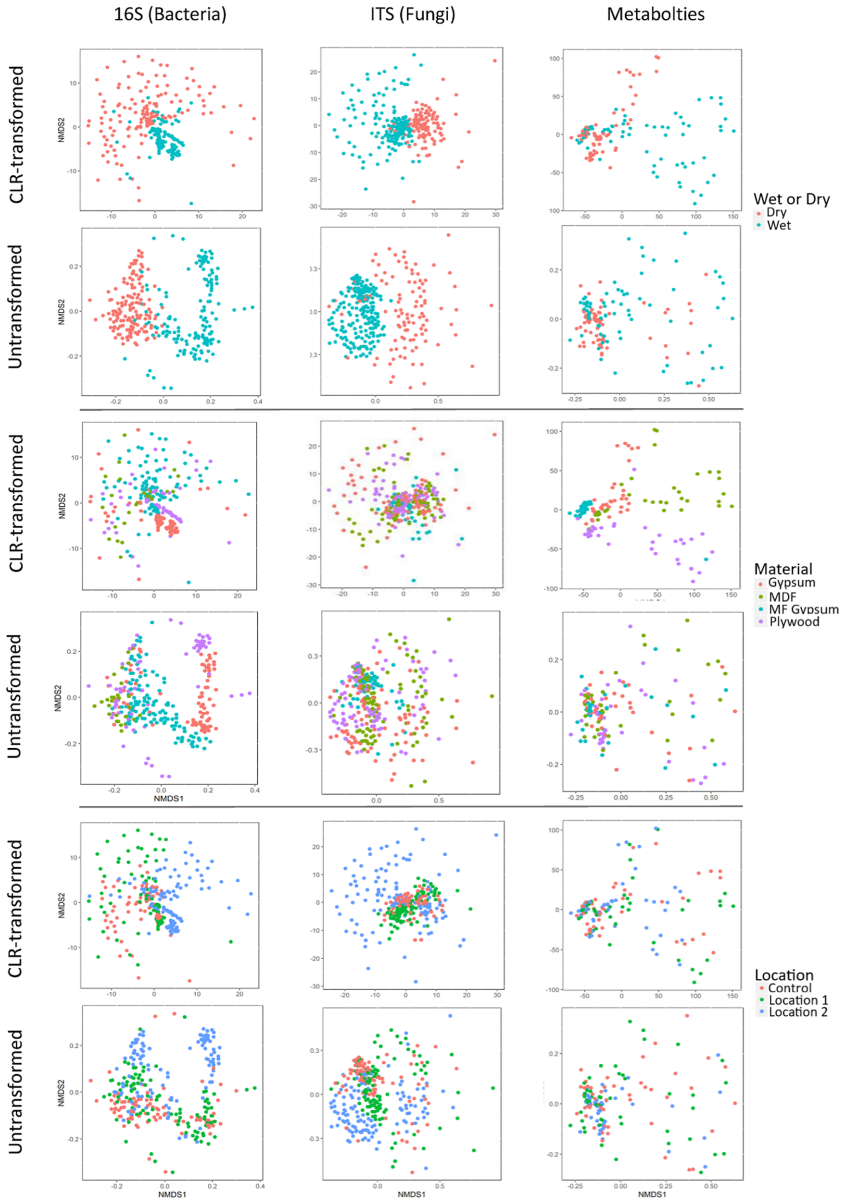

Figure 2. NMDS ordination plots showing clustering of samples. Columns correspond to dataset type; 16S, ITS and Metabolomics are columns one, two and three, respectively ($n = 294$, 319 and 144). Each row of plots is colored by wetting status, material and location. NMDS plots in rows 1, 3 and 5 were created from clr-transformed data, while NMDS plots in rows 2,4 and 6 were created from untransformed data.

transformed and the clr-transformed datasets using the mixOmics DIABLO framework, an integrative, multivariate method for multiomics classification (25). The same bacterial, fungal and metabolite datasets were integrated with the R package mixOmics version 6.10.8c (25) using the DIABLO framework to assess the correlation structure at the component level for each of the four conditions: time, material, location and wetting status.

Microbial balances

We supplemented our analysis with the R package selbal version 0.1 (26), a CoDA method that detects microbial signatures between different sample types (26). Rather than identifying large numbers of differentially abundant taxa between sample types, selbal searches for the smallest number of taxa in a microbial balance that are highly predictive of the sample conditions. selbal was used to detect fungal and bacterial signatures associated with time points in wet samples, gypsum or MF gypsum on wet samples and wet or dry condition for genera of the fungal and bacterial taxa present in >10% of samples. Input to selbal was the raw counts of fungal and bacterial taxa summed by genus as selbal has its own zero-handling and transformation method within the package. Fungal and bacterial datasets were processed in selbal separately. We did not use selbal with the metabolite dataset as the method is designed for microbial balances. Furthermore, selbal can only predict microbial balances for dichotomous and continuous response variables, so we only performed this analysis for the response

Table 1. PERMANOVA results (9999 permutations) for clr-transformed and untransformed (rarefied) data for three microbial community datasets

		clr-transformed (Euclidean)			Untransformed (Bray–Curtis)		
Dataset	Variable	R^2	p	p-adj[1]	R^2	p	p-adj[1]
Bacteria	Time	0.0352	0.0001	0.0024	0.0414	0.0001	0.0024
	Location	0.0424	0.0001	0.0024	0.0627	0.0001	0.0024
	Material	0.0907	0.0001	0.0024	0.1270	0.0001	0.0024
	Wet or dry	0.0921	0.0001	0.0024	0.1445	0.0001	0.0024
Fungi	Time	0.0297	0.0001	0.0024	0.0505	0.0001	0.0024
	Location	0.0347	0.0001	0.0024	0.0702	0.0001	0.0024
	Material	0.0270	0.0001	0.0024	0.0540	0.0001	0.0024
	Wet or dry	0.0464	0.0001	0.0024	0.0616	0.0001	0.0024
Metabolites	Time*	0.0723	0.0006	0.0144	0.0305	0.6755	1
	Location	0.0186	0.1389	1	0.0300	0.0137	0.3288
	Material*	0.2295	0.0001	0.0024	0.0262	0.1767	1
	Wet or dry	0.0965	0.0001	0.0024	0.0637	0.0001	0.0024

[1]FDR corrected for multiple comparisons.
*Significant only after clr transformation.

variables wetting status (dichotomous), wet gypsum versus wet MF gypsum (dichotomous) and time (continuous).

RESULTS AND DISCUSSION

Sample diversity by condition

We used CoDA approaches to determine the effects of sample condition (material type, inoculation location and wetting) and time on microbial and metabolite diversity using the data from Lax *et al.* (18), and compared the CoDA results to the results using non-transformed (rarefied) data using the same statistical tests. We tested two different zero-handling methods on the datasets: pseudocounts from the zCompositions package and replacement with counts of one. As the NMDS plots looked highly similar for both zero handling methods, we present only the results of the zCompositions-based analyses (Figure 2). The NMDS plots of the clr-transformed data identified several differences compared to NMDS plots of untransformed data, though the separation between wet and dry samples remained. The clr-transformed metabolite NMDS plots showed more distinct separation by wetting status and material type compared with the results of the untransformed NMDS plots, and the clustering of fungal samples for location 1 and the control location was also more evident (Figure 2). We did not observe apparent clustering of samples by time point (Supplementary Figure S1).

PERMANOVA tests indicated that location, time, material and wetting condition had significant associations with bacterial and fungal community diversity ($P = 0.0001$, p-adj $= 0.0024$ for all comparisons; Table 1). The clr transformation of metabolite data found that time, material and wetting condition, but not location, were significantly correlated with metabolite composition. This was also the case for the untransformed analyses, except for the untransformed metabolite dataset which only identified a significant difference between wetted and unwetted samples (Table 1). In the original study, the significant effect of material on the fungal community composition was not observed, meaning that our results contradicted the original study's conclusion that observed variations in fungal communities on different materials were driven only by moisture conditions. The novel associations identified in the clr-transformed metabolite datasets also demonstrated that CoDA has the potential to reveal community level associations not identified with non-CoDA approaches.

Correlations between abundant bacteria, fungi and metabolites

Applied properly, clr-transformation allows the application of standard statistical methods to analyze relationships between microbiome datasets, such as fungal and bacterial sequencing datasets and metabolomics, as we do here (9). Most correlations in all the combined datasets were within the same 'omics' datasets. For example, metabolites were mostly correlated with other metabolites rather than with bacteria or fungi. Three fungal genera were correlated significantly (p-adj < 0.05) with bacteria on wet gypsum: (i) *Neurospora* with *Acinetobacter, Agrobacterium and Erwina;* (ii) *Aureobasidium* with *Agrobacterium, Cronobacter, Enterobacter, Erwina* and *Pseudomonas;* and (iii) *Coprinopsis* with *Agrobacterium, Bacillus, Cronobacter, Erwina* and *Pseudomonas.* No significant bacteria-fungus correlations were observed on the other wetted materials. Only two genera were significantly correlated with metabolites, namely *Acinetobacter* and *Terribacillus.* Values for the Spearman correlations (p-adj < 0.05) can be found in Supplementary Table S1 and correlation matrices for correlations with p-adj < 0.05 can be found in Supplementary Figure S2. Mitochondrial sequences, which are likely an indicator of fungal abundance in general, were also identified to be correlated with *Neurospora* and some metabolites.

For all three correlation comparisons, bacteria-fungi (b-f), metabolite-bacteria (m-b), and metabolite-fungal (m-f), the number of significant correlations (p-adj < 0.05) on wet, MF gypsum (b-f $= 52$, m-b $= 23$ and m-f $= 5$) was much lower than the number of significant correlations observed for wet gypsum (b-f $= 369$, m-b $= 284$ and m-f $= 121$) and wet plywood (b-f $= 302$, m-b $= 147$ and m-f $= 159$) (Supplementary Table S1). This suggested that the decreased presence of fungal taxa on MF gypsum (MF gypsum had the lowest fungal abundance observed by quantitative polymerase chain reaction in the original study) affected the bacterial community structure in such a way as

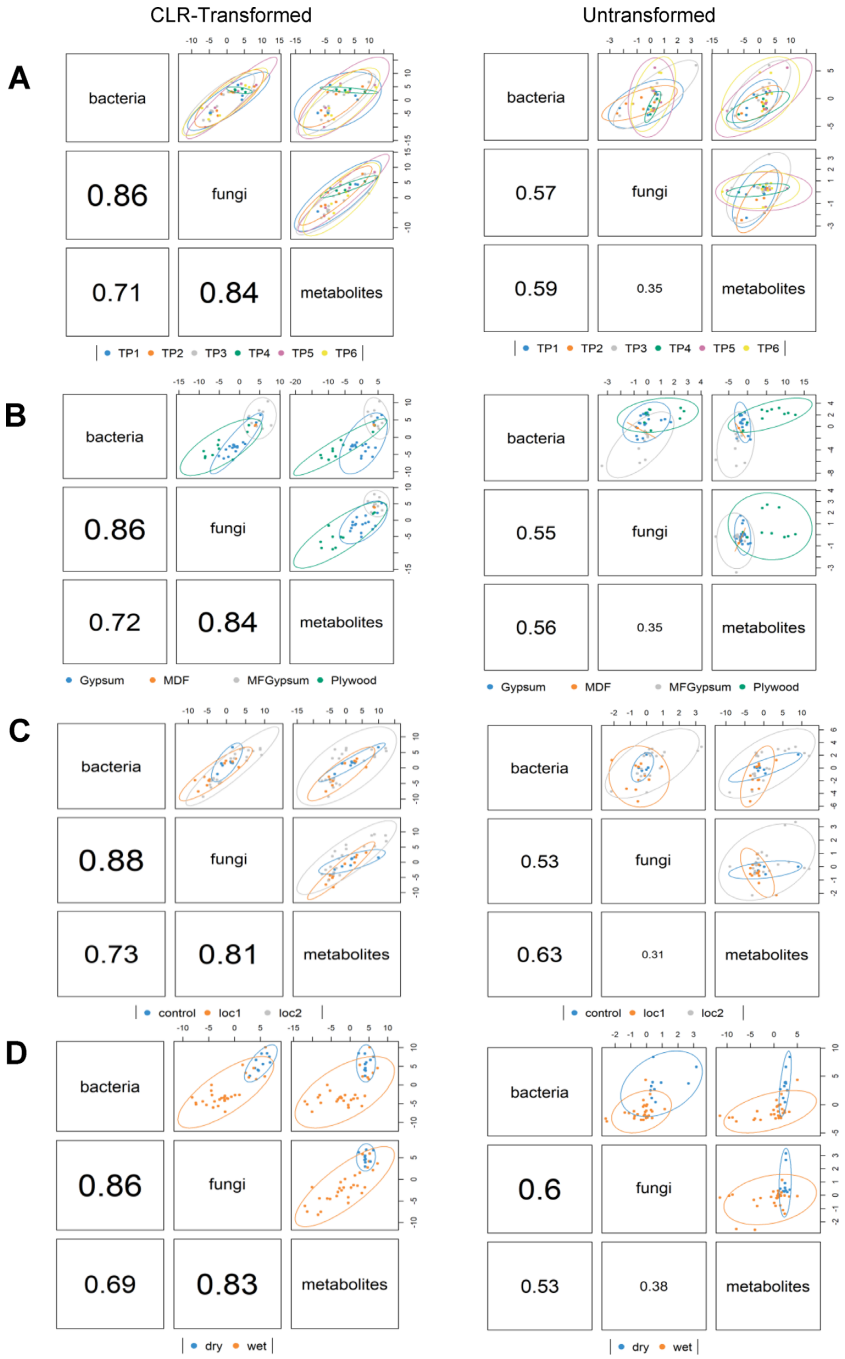

Figure 3. Correlation structure between bacterial, fungal and metabolic datasets from the mixOmics DIABLO framework, which plots the components from the framework across the datasets by sample condition: (**A**) time point, (**B**) material, (**C**) location and (**D**) wet or dry condition. Higher values indicate greater correlation structure between the compared datasets. The ellipses indicate discriminative power of the components to separate samples by colored condition. The first column of plots was created with clr-transformed data while the second column of plots was created from untransformed data.

to alter its metabolic profile. Additionally, the original study found *Bacillus* and *Pseudomonas* to be negatively correlated with one another on wet materials. However, our analysis found that found that this correlation was material specific: *Bacillus* and *Pseudomonas* were not significantly correlated on MF gypsum (corr = 0.376, p-adj = 1) or plywood (corr = −0.558, p-adj = 0.978), but were negatively correlated on gypsum (corr = −0.428, p-adj = 0.0275).

Multiomics integration

Comparisons of DIABLO mixOmics plots produced with clr-transformed and untransformed data revealed consistently greater correlation structure across all the three 'omics' datasets with the clr-transformed data (Figure 3), suggesting that clr-transformation can uncover more real correlations between multiomics data. Comparison of correlation values between clr-transformed datasets found

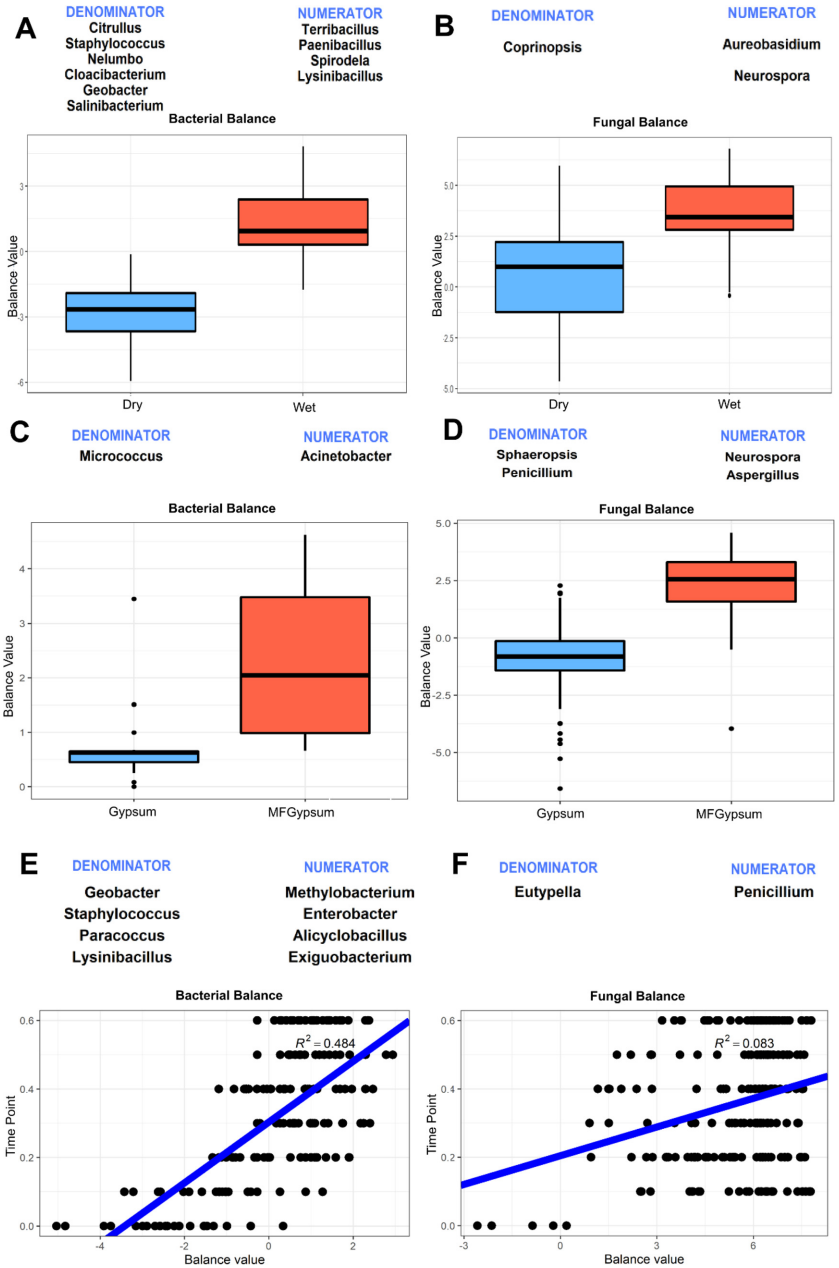

Figure 4. Microbial balances for abundant bacterial and fungal genera at different conditions computed with selbal, where numerator genera are more relatively abundant than denominator genera for higher balance values. (**A**) Microbial balance of bacteria in wet and dry samples. (**B**) Microbial balance of fungi in wet and dry samples. Microbial Balances of wet gypsum and wet MF gypsum of (**C**) bacteria and (**D**) fungi. Microbial balances for wet samples over time points for (**E**) bacteria and (**F**) fungi.

weaker correlation structure between the bacteria and metabolite datasets compared to the other pairwise dataset comparisons (Figure 3). The correlation structure of the datasets extracted by the DIABLO framework found the strongest discrimination between samples based on the material (Figure 3B) and wetting condition (Figure 3D), as seen by the greater separation of samples by those variables. This outcome agreed with the results of the NMDS ordination plots, which identified the strongest associations of the bacterial, fungal and metabolite datasets with wetting condition and material type (Figure 2).

Microbial balances between conditions

We analyzed the same abundant fungal and bacterial genera used in the correlation analyses with selbal to determine fungal and bacterial balances for wet samples over time, wet gypsum or wet MF gypsum samples and wet and dry samples (Figure 4). Interestingly, selbal was the only method that detected associations with any of the datasets and time point. selbal additionally identified many bacterial genera associated with wet or dry samples that were biologically plausible (Figure 4A). Species of one of the genera associated with wet samples, *Paenibacillus,* have been iso-

lated from wet environments, including milk, wetlands and a fresh water spring (27–29). The fungal genera *Neurospora* and *Aureobasidium* were more predictive of wet samples, while the fungal genus *Coprinopsis* was more predictive of dry samples (Figure 4B). As previously noted, these three genera were the only fungi associated with bacterial genera in wet conditions in the Spearman correlation analyses. *Neurospora* has also been identified as an important potential allergen in indoor environments (30,31). *Acinetobacter, Neurospora* and *Aspergillus* were relatively more abundant on wet MF gypsum compared to wet gypsum, whereas *Micrococcus, Sphaeropsis* and *Penicillium* were more abundant on wet MF gypsum (Figure 4C and D). Fungal and bacterial microbial signatures detected by selbal as more abundant in earlier or later time points are shown in Figure 4E and F. Of the genera more abundant in later time points, *Methylobacterium* has been shown to be associated with wet environments (32), *Alicyclobacillus* species is a known culprit in spoilage of fruit juices (33,34) and *Exiguobacterium* has been isolated from marine water (35). The identification of taxa not identified as interesting in the original study as potentially biologically relevant, indicates the potential utility of sparse models of microbial balances for investigating microbial community diversity.

In summary, by using both standard and more recently developed statistical methods on clr-transformed sequencing data we discovered relationships and discriminatory factors not identified in the original study. Additionally, the greater separation of the clr-transformed data by variables in some of the NMDS plots and a greater correlation structure between datasets using clr-transformed data shown in the DIABLO analyses indicated that the clr transformation may model the structure of the data better than approaches like rarefaction. Studies drawing conclusions from sequencing data normalized only by traditional methods, rather than CoDA methods, may miss out on novel and interesting biological insights due to sub-optimal data standardization. There are some methods, however, to which the clr transformation of compositional counts may not be the most appropriate step. For example, some diversity metrics depend on 'absence' or 'presence' counts (zeroes or non-zeroes, respectively). Since clr transformation requires the replacement of zeroes, there are no 'absent' species; samples would have the same number of species 'present', rendering this type of metric useless. This is also true for more complicated diversity metrics like Faith's Phylogenetic Diversity (PD), which sums branch lengths for the phylogenetic tree of each sample for present species (36). If all species are 'present', Faith's PD will be the same for each sample. Even measures like evenness, which take taxa abundance into account, are sensitive to bias introduced in sample preparation, sequencing and analysis methods (37).

DATA AVAILABILITY

The data tables from the original article we reanalyzed may be accessed at: https://doi.org/10.6084/m9.figshare.7865015.v2.

SUPPLEMENTARY DATA

Supplementary Data are available at NARGAB Online.

ACKNOWLEDGEMENTS

We would like to thank Bryan Ho for assistance and advice regarding software installation.

FUNDING

We declare no funding sources for this work.
Conflict of interest statement. None declared.

REFERENCES

1. Tringe,S.G. and Hugenholtz,P. (2008) A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.*, **11**, 442–446.
2. Varliero,G., Bienhold,C., Schmid,F., Boetius,A. and Molari,M. (2019) Microbial diversity and connectivity in deep-sea sediments of the South Atlantic polar front. *Front. Microbiol.*, **10**, 1–18.
3. Hewitt,K.M., Mannino,F.L., Gonzalez,A., Chase,J.H., Caporaso,J.G., Knight,R. and Kelley,S.T. (2013) Bacterial diversity in two neonatal intensive care units (NICUs). *PLoS One*, **8**, e54703.
4. Coutinho,F.H., Gregoracci,G.B., Walter,J.M., Thompson,C.C. and Thompson,F.L. (2018) Metagenomics sheds light on the ecology of marine microbes and their viruses. *Trends Microbiol.*, **26**, 955–965.
5. Milac,T.I., Randolph,T.W. and Wang,P. (2012) Analyzing LC-MS/MS data by spectral count and ion abundance: two case studies. *Stat. Interface*, **5**, 75–87.
6. Quinn,T.P., Erb,I., Gloor,G., Notredame,C., Richardson,M.F. and Crowley,T.M. (2019) A field guide for the compositional analysis of any-omics data. *GigaScience*, **8**,1–14.
7. 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**,1–6.
8. Quinn,T.P., Erb,I., Richardson,M.F. and Crowley,T.M. (2018) Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, **34**, 2870–2878.
9. Aitchison,J. (1982) The statistical analysis of compositional data. *J. R. Stat. Soc. B*, **44**, 139–160.
10. Selechnik,D., Richardson,M.F., Shine,R., Brown,G.P. and Rollins,L.A. (2019) Immune and environment-driven gene expression during invasion: an eco-immunological application of RNA-seq. *Ecol. Evol.*, **9**, 6708–6721.
11. Jiang,P., Green,S.J., Chlipala,G.E., Turek,F.W. and Vitaterna,M.H. (2019) Reproducible changes in the gut microbiome suggest a shift in microbial and host metabolism during spaceflight. *Microbiome*, **7**, 1–18.
12. Leong,C., Haszard,J.J., Heath,A.-L.M., Tannock,G.W., Lawley,B., Cameron,S.L., Szymlek-Gay,E.A., Gray,A.R., Taylor,B.J., Galland,B.C. *et al.* (2019) Using compositional principal component analysis to describe children's gut microbiota in relation to diet and body composition. *Am. J. Clin. Nutr.*, **111**, 70–78.
13. Gao,C., Montoya,L., Xu,L., Madera,M., Hollingsworth,J., Purdom,E., Singan,V., Vogel,J., Hutmacher,R.B., Dahlberg,J.A. *et al.* (2020) Fungal community assembly in drought-stressed sorghum shows stochasticity, selection, and universal ecological dynamics. *Nat. Commun.*, **11**, 1–14.
14. Fernandes,A.D., Reid,J.N.S., 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**, 1–13.
15. Peters,B.A., Wu,J., Pei,Z., Yang,L., Purdue,M.P., Freedman,N.D., Jacobs,E.J., Gapstur,S.M., Hayes,R.B. and Ahn,J. (2017) Oral microbiome composition reflects prospective risk for esophageal cancers. *Cancer Res.*, **77**, 6777–6787.
16. 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**, 1–14.
17. Pereira,M.B., Wallroth,M., Jonsson,V. and Kristiansson,E. (2018) Comparison of normalization methods for the analysis of metagenomic gene abundance data. *BMC Genomics*, **19**, 1–17.

18. Lax,S., Cardona,C., Zhao,D., Winton,V.J., Goodney,G., Gao,P., Gottel,N., Hartmann,E.M., Henry,C., Thomas,P.M. *et al.* (2019) Microbial and metabolic succession on common building materials under high humidity conditions. *Nat. Commun.*, **10**, 1–12.
19. Palarea-Albaladejo,J. and Martín-Fernández,J.A. (2015) zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemom. Intell. Lab. Syst.*, **143**, 85–96.
20. Oksanen,J., Blanchet,F.G., Friendly,M., Kindt,R., Legendre,P., McGlinn,D., Minchin,P.R., O'Hara,R.B., Simpson,G.L., Solymos,P. *et al.* (2019) vegan: community ecology package. R package version 2.5-6.
21. Wickham,H. (2016) *ggplot2: Elegant Graphics for Data Analysis*. Springer, NY.
22. DeSantis,T.Z., Hugenholtz,P., Larsen,N., Rojas,M., Brodie,E.L., Keller,K., Huber,T., Dalevi,D., Hu,P. and Andersen,G.L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.*, **72**, 5069–5072.
23. Nilsson,R.H., Larsson,K.-H., Taylor,A.F S., Bengtsson-Palme,J., Jeppesen,T.S., Schigel,D., Kennedy,P., Picard,K., Glöckner,F.O., Tedersoo,L. *et al.* (2018) The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.*, **47**, D259–D264.
24. Wei,T. and Simko,V. (2017) R package "corrplot": visualization of a correlation matrix (Version 0.84).
25. Rohart,F., Gautier,B., Singh,A. and Lê Cao,K.-A. (2017) mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.*, **13**, e1005752.
26. Rivera-Pinto,J., Egozcue,J.J., Pawlowsky-Glahn,V., Paredes,R., Noguera-Julian,M. and Calle,M.L. (2018) Balances: a new perspective for microbiome analysis. *mSystems*, **3**, e00053-00018.
27. Baik,K.S., Choe,H.N., Park,S.C., Kim,E.M. and Seong,C.N. (2011) Paenibacillus wooponensis sp. nov., isolated from wetland freshwater. *Int. J. Syst. Evol. Microbiol.*, **61**, 2763–2768.
28. Scheldeman,P., Goossens,K., Rodriguez-Diaz,M., Pil,A., Goris,J., Herman,L., De Vos,P., Logan,N.A. and Heyndrickx,M. (2004) Paenibacillus lactis sp. nov., isolated from raw and heat-treated milk. *Int. J. Syst. Evol. Microbiol.*, **54**, 885–891.
29. Saha,P., Mondal,A.K., Mayilraj,S., Krishnamurthi,S., Bhattacharya,A. and Chakrabarti,T. (2005) *Paenibacillusassamensis* sp. nov., a novel bacterium isolated from a warm spring in Assam, India. *Int. J. Syst. Evol. Microbiol.*, **55**, 2577–2581.
30. Côté,J., Chan,H., Brochu,G. and Chan-Yeung,M. (1991) Occupational asthma caused by exposure to neurospora in a plywood factory worker. *Br. J. Ind. Med.*, **48**, 279–282.
31. Singh,A.B. and Kumar,P. (2002) Common environmental allergens causing respiratory allergy in India. *Ind. J. Pediatr.*, **69**, 245–250.
32. Kelley,S.T., Theisen,U., Angenent,L.T., St. Amand,A. and Pace,N.R. (2004) Molecular analysis of shower curtain biofilm microbes. *Appl. Environ. Microbiol.*, **70**, 4187–4192.
33. Komitopoulou,E., Boziaris,I.S., Davies,E.A., Delves-Broughton,J. and Adams,M.R. (1999) *Alicyclobacillusacidoterrestris* in fruit juices and its control by nisin. *Int. J. Food Sci. Technol.*, **34**, 81–85.
34. Cerny,G., Hennlich,W. and Poralla,K. (1984) Spoilage of fruit juice by bacilli: isolation and characterization of the spoiling microorganisms. *Z Lebensm Unters Forsch.*, **179**, 224–227.
35. Anil Kumar,P.K. and Suresh,P.V. (2014) Biodegradation of shrimp biowaste by marine *Exiguobacterium* sp. CFR26M and concomitant production of extracellular protease and antioxidant materials: production and process optimization by response surface methodology. *Mar. Biotechnol.*, **16**, 202–218.
36. Faith,D.P. (1992) Conservation evaluation and phylogenetic diversity. *Biol. Conserv.*, **61**, 1–10.
37. McLaren,M.R., Willis,A.D. and Callahan,B.J. (2019) Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, **8**, e46923.