AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Ecological and Evolutionary Science

A Novel Sparse Compositional Technique Reveals Microbial Perturbations

Cameron Martino,[a,b] James T. Morton,[a,c] Clarisse A. Marotz,[a] Luke R. Thompson,[d,e] Anupriya Tripathi,[a] Rob Knight,[a,c,f] Karsten Zengler[a,f,g]

[a]Department of Pediatrics, University of California San Diego, La Jolla, California, USA
[b]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, California, USA
[c]Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA
[d]Department of Biological Sciences and Northern Gulf Institute, University of Southern Mississippi, Hattiesburg, Mississippi, USA
[e]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, stationed at Southwest Fisheries Science Center, La Jolla, California, USA
[f]Center for Microbiome Innovation, University of California San Diego, La Jolla, California, USA
[g]Department of Bioengineering, University of California San Diego, La Jolla, California, USA

ABSTRACT The central aims of many host or environmental microbiome studies are to elucidate factors associated with microbial community compositions and to relate microbial features to outcomes. However, these aims are often complicated by difficulties stemming from high-dimensionality, non-normality, sparsity, and the compositional nature of microbiome data sets. A key tool in microbiome analysis is beta diversity, defined by the distances between microbial samples. Many different distance metrics have been proposed, all with varying discriminatory power on data with differing characteristics. Here, we propose a compositional beta diversity metric rooted in a centered log-ratio transformation and matrix completion called robust Aitchison PCA. We demonstrate the benefits of compositional transformations upstream of beta diversity calculations through simulations. Additionally, we demonstrate improved effect size, classification accuracy, and robustness to sequencing depth over the current methods on several decreased sample subsets of real microbiome data sets. Finally, we highlight the ability of this new beta diversity metric to retain the feature loadings linked to sample ordinations revealing salient intercommunity niche feature importance.

IMPORTANCE By accounting for the sparse compositional nature of microbiome data sets, robust Aitchison PCA can yield high discriminatory power and salient feature ranking between microbial niches. The software to perform this analysis is available under an open-source license and can be obtained at https://github.com/biocore/DEICODE; additionally, a QIIME 2 plugin is provided to perform this analysis at https://library.qiime2.org/plugins/deicode/.

KEYWORDS compositional, computational biology, matrix completion, microbiome, metagenomics

Beta diversity is an ecological concept that describes differentiation in taxonomic or phylogenetic composition between communities. Beta diversity methods are a major component of many microbiome statistical analysis pipelines. These analyses enable an overview of complex microbial communities, identifying environmental factors differentiating microbial communities. However, there are dozens of distance metrics available to microbial ecologists to analyze their data, with each distance metric tailored to capture specific data characteristics. Beta diversity plots can therefore look

Citation Martino C, Morton JT, Marotz CA, Thompson LR, Tripathi A, Knight R, Zengler K. 2019. A novel sparse compositional technique reveals microbial perturbations. mSystems 4:e00016-19. https://doi.org/10.1128/mSystems.00016-19.

Editor Josh D. Neufeld, University of Waterloo

Address correspondence to Karsten Zengler, kzengler@ucsd.edu.

A novel sparse compositional technique reveals microbial perturbations.

Received 9 January 2019
Accepted 11 January 2019
Published 12 February 2019

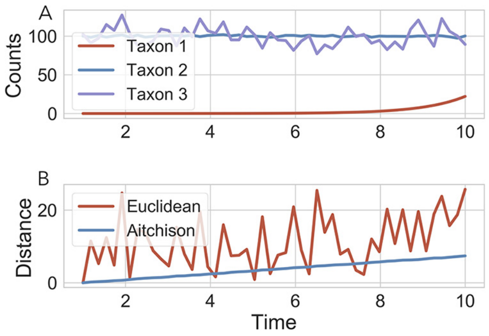

FIG 1 Benchmarking the rclr preprocessing step. Toy example with simple 3-taxon community sampled over time (A). Distance calculated between the $t = 1$ community and subsequent communities demonstrates the robustness of Aitchison distance compared to Euclidean distance (B).

dramatically different depending on the distance metric chosen, contributing to differences in interpretation of raw data (1).

One major confounding factor in beta diversity analysis is that microbiome data sets are sparse (i.e., most microorganisms are not found in most data sets), which has been shown to give rise to spike and horseshoe patterns in ordination plots (2, 3), complicating analysis. Furthermore, principal-component analysis (PCA) has common assumptions of normally distributed and linearly related variables, often violated by biological data (4–7). As a result, classical distance metrics that take into account only the presence/absence of taxa, such as the Jaccard index, or metrics that explicitly account for relative abundances, such as Bray-Curtis symmetrized distance, are commonly used. Microbial beta diversity estimation was greatly improved with the incorporation of phylogenetic information, as was shown with UniFrac (8), which can be used as either a presence/absence (unweighted) or relative abundance (weighted) metric. However, presence/absence methods often yield substantial differences between communities that are obscured by abundance-based methods. This might seem paradoxical, because abundance-based methods are integrating more information about the community. However, if the key players are rare rather than abundant species, or if abundant species display large random fluctuations, abundance information may obscure rather than clarify the result, even with phylogenetic metrics (9).

Failure to reveal associations between phenotypes and the microbiome overall may also be a symptom of methods that do not properly account for the relative changes of microbial taxon abundances. To demonstrate this principle, consider the scenario in Fig. 1A, where three taxa are simulated over time. In this scenario, taxon 1 has a much lower abundance than the other two taxa, but it is growing exponentially over time. Taxon 2 has a high abundance and is stable over time. Taxon 3 also has a high abundance but fluctuates randomly. The Euclidean distance between the first community and the other two time points is extremely variable and does not capture the change induced by the exponential growth of taxon 1. This variability in the Euclidean distance is largely driven by the random fluctuations in the high-abundance taxa.

In contrast to Euclidean distance, compositional distance metrics, such as the Aitchison distance (see equation 2), can properly account for such relative changes (10). Here, the Aitchison distance factors in only the log fold change, reflecting the fact that deviations in the high-abundance taxa are large on an absolute scale but small on a relative scale. The difference between 100 counts and 120 counts is 20 counts, which is large compared to the abundance of the first taxon but is only a 20% increase. In contrast, the first taxon increased around 2,000%, and as a result, the Aitchison distance is driven by the large relative changes, including changes in the low-abundance species.

Microbes that display large fold change across samples will be weighted more heavily in the calculation of the Aitchison distance. However, this distance metric cannot handle zeros and is thus challenging to apply to the sparse data sets that characterize microbiome studies. There are many potential processes that could give rise to zeros in microbiome data. It is possible that there was undersampling, where low-abundance microbes were not detected in the sequencing data. Another possibility is that due to the heterogeneity of the sample, the microbe was not detected, even though it is present in the environment. Furthermore, it could be possible that the microbe is not present at all in the environment. In light of all of these potential processes, it is not feasible to differentiate between these different processes from the sequencing data (11, 12). To circumvent this issue, we propose a novel, compositional distance metric that can also explicitly handle sparse data through the use of matrix completion. This is done by treating all zeros as missing values and building a model to handle this missing data using matrix completion.

Matrix completion was originally developed in the context of recommender systems to predict user-item ratings (13) as a natural solution for handling sparse data. For example, the Netflix database contains a matrix detailing all customers by all movies where the entries are the movie ratings. However, each user rates only a small portion of the possible movies available on Netflix, so that only about 1% of the database contains nonzero values. As a result, when trying to recommend specific movies to specific customers, models need to be trained on the available ratings that customers have provided. Matrix completion tasks have become one of the state-of-the-art methods for performing these sorts of tasks.

Here, using simulation benchmarks and two case studies, we demonstrate the utility of preprocessing sparse microbiome data sets with matrix completion to allow compositional ordination and to preserve information about the features driving differences among samples.

RESULTS

Description of robust Aitchison PCA. Matrix completion can be interpreted as a robust dimensionality reduction technique, where PCA is performed accounting only for the observed entries (i.e., ignoring the zeros). Matrix completion relies on two major assumptions. First, it assumes that data are missing at random, meaning that the missing entries in the matrix are uniformly distributed. Second, because matrix completion is a robust form of PCA, it assumes that the data are normally distributed and centered around zero (14). To meet this assumption, a commonly applied approach is to subtract the row and column means (15, 16). However, because microbiome sequencing data are represented as counts (17), the data are strictly positive and skewed toward zero, which confounds PCA. A workaround is to first log transform the nonzero values before centering the data—we will refer to this preprocessing procedure as the robust centered log ratio (rclr) due to its links to the centered log ratio (clr) transform commonly used in compositional data analysis (10) (Fig. 2A and B). A similar procedure using interquartiles was suggested previously (18).

This procedure produces a transformed table with missing values that can be used as input for matrix completion, or robust principal-component analysis (RPCA), which provides the sample and feature loadings. These sample and feature loadings contain the ordination information directly used in beta diversity plotting and feature biclustering (Fig. 2C to E). Because PCA preserves feature information, we can use the feature loadings to determine which taxa drive the differences among sample types (Fig. 2F).

Simulations. To benchmark the effectiveness of the rclr preprocessing step, we generated simulations from a study comparing microbial communities on keyboards and human fingertips (keyboard data set) (19) (see Materials and Methods for details). Simulated data were chosen as an initial proof-of-concept benchmark due to the ease of changing data set characteristics across which to interrogate; here the primary focus was on sequencing depth.

The simulated data were generated with two clusters over various sequencing depths from 1,000 to 10,000 reads per sample. At each sequencing depth, the output of the RPCA with and without the rclr transformation was compared by Kullback-Leibler divergence (KL) (20) to the simulation ground truth between rclr preprocessed and raw

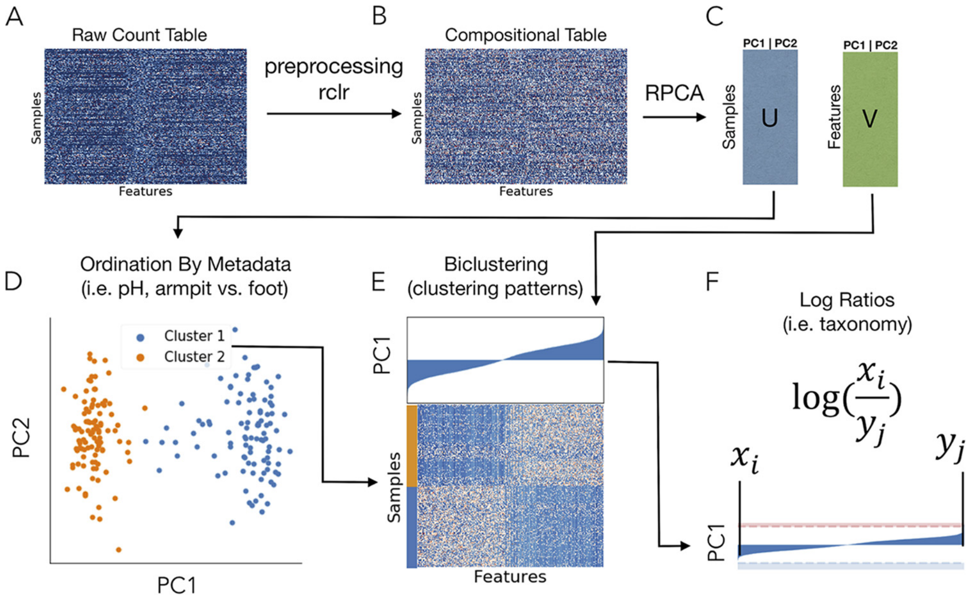

FIG 2 A general overview of the workflow. (A) A sparse, raw sequencing count table with samples on the *y* axis and features (i.e., OTUs, genes) on the *x* axis. (B) The data are preprocessed by a robust centered log ratio transform (rclr) on only the known (nonzero) values. (C) Matrix completion with a robust principal-component analysis (RPCA) that operates on only the observed values in the table resolves a loading by samples and by features. These loadings can be directly used for ordination (D), biclustering (E), and the identification of important taxa driving clustering in both the previous plots (F).

count data. Additionally, ordination output was compared by permutational multivariate analysis of variance (PERMANOVA) F-statistic and supervised k-nearest neighbor (KNN) classification cross-validation (40:60) split.

When rclr preprocessing was applied, we saw a decrease in mean KL, demonstrating a more closely matched probability distribution when using the rclr (Fig. 3A). Furthermore, when the rclr was applied, the F-statistic demonstrated a 4-fold increase (Fig. 3B) and KNN classification accuracy (Fig. 3C) increased by between 30 and 40%. All of the metrics, when applied to rclr RPCA, improved as the sequencing depth improved, following the logic that a good fit should increase performance as sequencing depth increases. A negative-control simulation with no group discrimination revealed no biclustering, RPCA clustering (Fig. 3E), low KNN classification accuracy, and PERMANOVA significance compared to a positive control (Fig. 3D) with two distinct groups (see Table S1 in the supplemental material). This demonstrates a proof of concept that rclr is less affected by outliers and is reliably reproducible at low and high sequencing depths.

Case studies. Next, we demonstrated the utility of RPCA compared to the current state of the art. To do this we used two 16S rRNA gene amplicon sequencing data sets. The first data set is a subset of the Sponge Microbiome Project (sponges) (21), where we compared sponge microbial communities classified by health status (i.e., stressed or healthy). The second data set derives from a sleep apnea study; it consists of mouse fecal samples and focuses on comparing the gut microbiome of animals exposed to intermittent hypoxia and hypercapnia (IHH; as a model of obstructive sleep apnea) to controls exposed to room air (air) (22).

Many different metrics exist for beta diversity distance comparison. We compared RPCA to two of the most commonly employed abundance-based methods, Bray-Curtis and weighted UniFrac, over 10-fold random subsamples of the data. The distances between the highlighted metadata categories for the two data sets were compared over subsamples with PERMANOVA (Fig. 4A and C). The principal coordinate analysis (PCoA) was compared by supervised KNN classification cross-validation (40:60 split) accuracy for both data sets over subsamples (Fig. 4B and D). In all subsample compar-

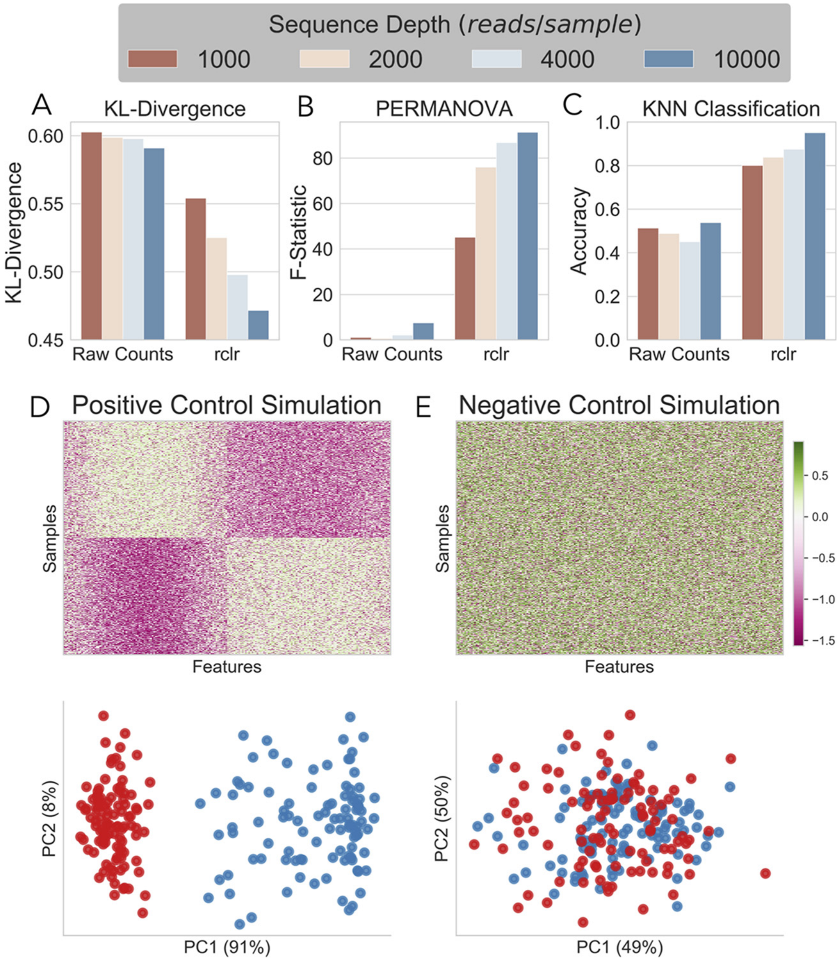

FIG 3 (A) Comparison of KL-divergence (*y* axis) between simulated base truth data between RPCA output from raw count data and rclr-preprocessed data. (B and C) Comparison between RPCA ordination by PERMANOVA F-statistic (B) and KNN classifier accuracy (C). All are at various sequencing depths from 1,000 to 10,000 reads per sample. (D and E) Comparison of positive- (D) and negative-control (E) simulation by biclustering (top) and RPCA ordination (bottom).

isons, the robust Aitchison (distance metric derived from RPCA) outperformed Bray-Curtis and weighted UniFrac. The results are qualitatively demonstrated in the PCoA clustering between metadata categories for low and high subsample depths (Fig. 4E and F).

A key benefit of RPCA over metrics, such as weighted UniFrac and Bray-Curtis, is direct access to the feature loadings. With Euclidean distance it is also possible to obtain feature loadings. However, Euclidean distance has multiple undesirable properties, such as artifacts in clustering patterns and weak discrimination in high-dimensional sparse data (2, 6, 7, 23, 24), making the values unreliable. Fortunately, feature loadings can be extracted from RPCA while utilizing the same interpretation as suggested in the work of Aitchison and Greenacre (25). These feature loadings can be largely explained by a few features (26). This ultimately allows us to rank the taxa in the data in relation to the samples and the metadata. When sorted, often referred to as biclustering, this method results in a table that reveals which taxa are driving the clustering seen in the ordinations.

In this case, we have a two-block table represented by clr-transformed heat maps for the sponges (Fig. 5A) and sleep apnea (Fig. 5B) data sets. It is evident from the heat map and ordination plots that there are some taxonomic abundance changes between the categories that are dividing the clusters. In order to compare two taxa directly, we

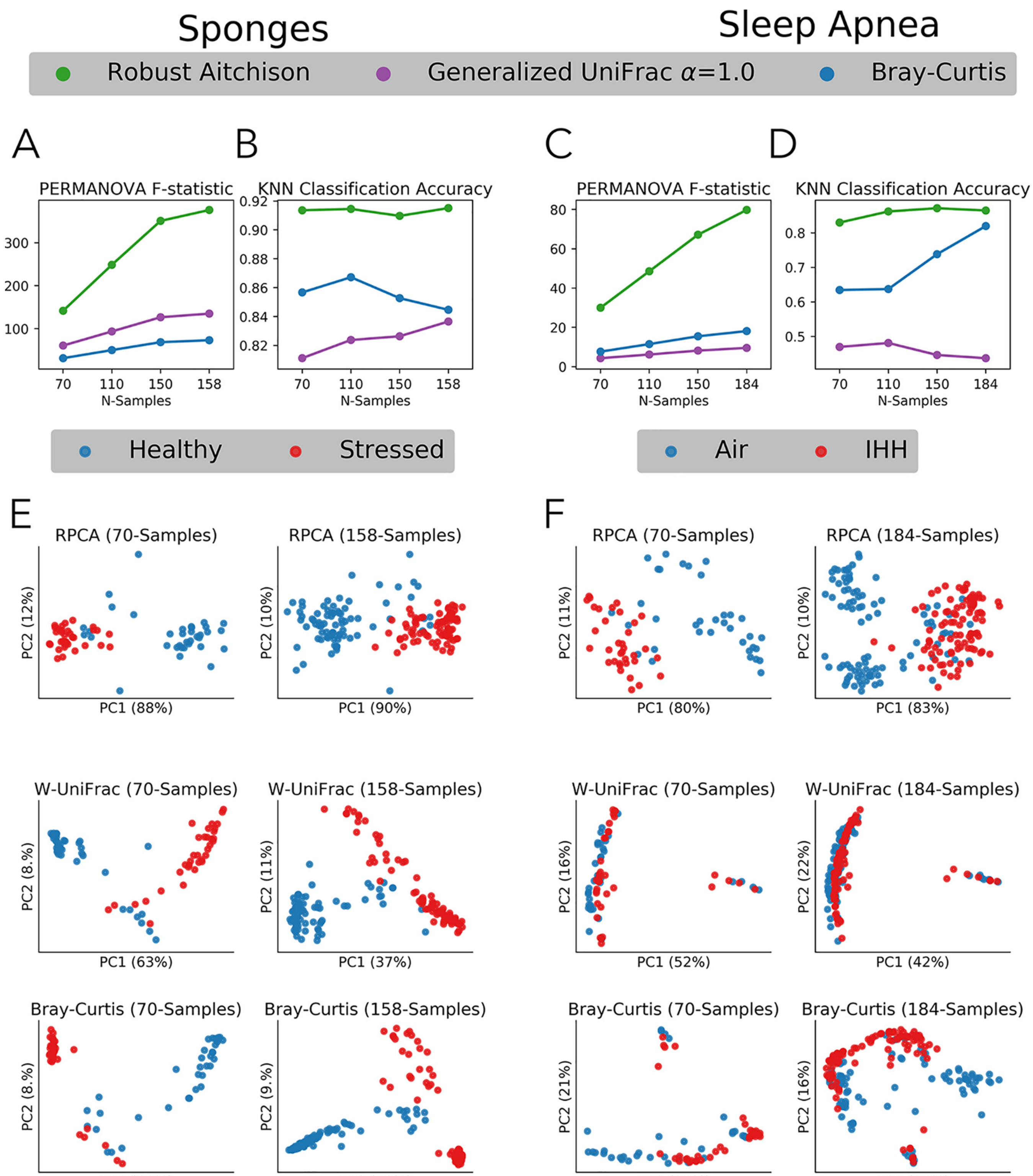

FIG 4 A case study of RPCA on real data sets; sponge (left; A, B, and E) and sleep apnea (right; C, D, and F). PERMANOVA F test statistic (*y* axis) (A and C) or KNN classifier accuracy (B and D) by subsamples of the data sets. Ordination plots between 70 samples total (left) and maximum number of samples (right) compared between RPCA (top), generalized weighted UniFrac (alpha = 1) (middle), and Bray-Curtis (bottom) (E and F). Sponge data set plotted between healthy (blue) and stressed (red) (E) along with sleep apnea data set plotted between air (blue) and IHH (red) (F).

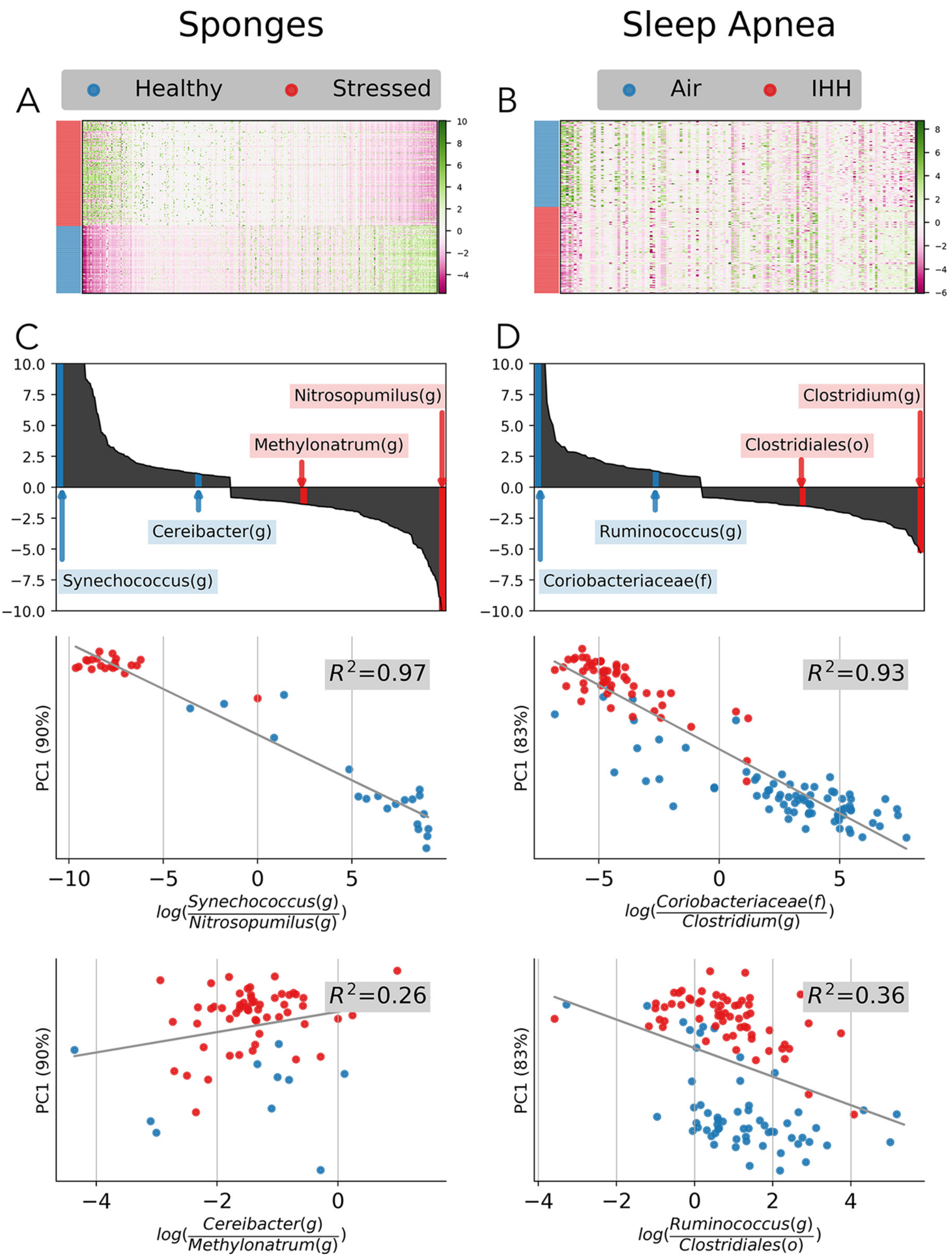

FIG 5 A case study of RPCA feature loadings on real data sets; sponge (left; A and C) and sleep apnea (right; B and D). Heat maps of clr-transformed sOTU tables with samples sorted by metadata and features sorted by RPCA feature loadings (A and B). Absolute highest (middle) and lowest (bottom) feature loading sOTUs (top) plotted as log ratios (*x* axis) by sample loading PC1 (*y* axis) (C and D).

applied log ratios on highly weighted features. The highest loaded features (most positively ranked and most negatively ranked) correspond to the most influential taxa driving the clustering. Interesting pairs of taxa were identified in the sponge data set (Fig. 5C) and the sleep apnea data set (Fig. 5D). These log ratios were shown to be correlated with the sample loadings in the PC1 axis ($R^2 = 0.97$ and 0.93). To show that not all of the taxa were significantly contributing to the variation in PC1, two pairs of insignificantly ranked log ratios were also identified ($R^2 = 0.26$ and 0.36).

The highly weighted log ratios in the sponge case study indicate that two sub-operational taxonomic units (sOTU) can explain a great deal of variation between healthy and thermally stressed sponges. The sOTUs most strongly associated with healthy and stressed sponges, respectively, were classified at the lowest assignment level to "*Candidatus* Synechococcus spongiarum" (species, numerator) and *Nitrosopumilus* (genus, denominator). Both of these groups are known sponge symbionts (27, 28). *Nitrosopumilus* is an ammonia-oxidizing archaeon, which nitrifies ammonia to nitrate. Nitrification by sponge-associated microbiota is thought to remove ammonia waste produced by the host sponge (27, 29). It has been proposed that ammonium, urea, and creatine leaking from host sponge tissue could promote growth of *Nitrosopumilus* (30), and this leakage may be more active in stressed hosts. "*Candidatus* Synechococcus spongiarum" has been found in numerous sponge species around the globe (28), and its photosynthetic products may contribute to host nutrition (31). From this analysis, this sOTU and several other sOTUs of "*Candidatus* Synechococcus spongiarum" (32) appear to be strongly associated with healthy sponges relative to stressed sponges.

In the sleep apnea data set, the highly weighted log ratios revealed a strong clustering of air versus IHH. These sOTUs were classified as *Coriobacteriaceae* (family) and *Clostridium* (genus). This trend was also observed by Tripathi et al. (22), where it was corroborated by the perturbations in the small molecular products attributed to members of these taxonomic classes. For example, changes in *Clostridium* were reflected in downstream changes in intestinal bile acids, as members of this genera are known to transform bile acids (33). Previous studies (34, 35) have also reported changes in these taxonomic classes in cardiometabolic comorbidities of sleep apnea, which suggests that our method potentially guides biologically relevant observations.

DISCUSSION

Here we demonstrated the ability of rclr preprocessing and RPCA to reveal salient, beta diversity ordination and factor loading. We demonstrated through simulations that rclr preprocessing dramatically improved RPCA. In two case studies (sponge and sleep apnea), RPCA presented higher PERMANOVA F-statistics and KNN classifier accuracy in small subsamples of the data. In addition, RPCA qualitatively increased the discriminative ability of clusters obtained from the ordination over beta diversity techniques widely used in the field, at both low and high levels of subsampling.

We have shown that Aitchison distance has numerous other desirable properties, such as scale invariance, negating the need to perform rarefaction. This feature is critical when one lacks access to absolute microbial abundance, because scale invariant distances ensure equivalence between distances computed from absolute and relative abundance measurements (see Materials and Methods for equation). Aitchison distance is also known to be subcompositionally coherent (36). This guarantees that distances will never decrease if additional taxa are observed (e.g., by using PCR primers with broader specificity), which has important implications for reproducibility across distance-based analyses, especially across studies that use different molecular methods.

The increased cluster separation at smaller subsamples of the data set highlights the robustness of the method. Significant partitioning of sample categories on smaller sample cohorts is particularly important in a clinical setting, due to the difficulty of large-volume sample collection. In addition, rapid resolutions of taxa driving ordination are of principal importance in translational results.

Importantly, because RPCA provides linked sample and feature information, one can directly identify which taxa are likely driving sample clustering (which are typically separate workflows in canonical amplicon analysis). However, RPCA does not currently take into account phylogenetic relationships among features. Adapting phylogenetic ilr transforms (37) could potentially improve the resulting ordinations.

It is important to note that while there have been previous methods that handle zeros in compositional data sets, such as methods available in zCompositions (38), the methodology here is fundamentally different. First, the zeros in microbiome data sets are never factored into the computation. In addition, the matrix completion approach presented here is a natural solution to high-dimensional data sets, which are not currently addressed in methods available in zCompositions.

In light of these benefits, there are still challenges that need to be considered, namely, overfitting and the low-rank approximation. Given the high-dimensional nature of microbial data sets, the number of parameters required to fit robust principal components can grow very quickly. As a result, it is still possible to overfit these methods, making them potentially sensitive to outliers and reducing their predictive power (39), although we did not notice these effects in our simulations. We therefore recommend starting fitting RPCA models with a low rank of either two or three for microbiome studies containing approximately 100 samples. The rank can be increased if there are appropriately many samples.

A low-rank constraint can possibly cause misleading results in the case of high-rank data sets. High-rank data sets may occur in microbiome data sets as a gradient between samples and features. To give intuition of what types of data may contain high-rank structure, we provide two published examples. The first example is a study of soil microbiomes representing different pH environments (40) (see Materials and Methods for detail). The second example is a case study of the gut colonization of an infant over time (41) (see Materials and Methods for detail). In both cases, a gradient forms because very few samples contain similar microbes (see Fig. S1 in the supplemental material). For example, in the infant development study very few microbes are shared between subsequent samples over time. Although the rclr transform eases the problem, it can still lead to misinterpretation in ordination (see Fig. S2 in the supplemental material). There are many possible future directions for incorporating regularization or Bayesian priors to better fit these models.

In light of the current limitations, we have shown that matrix completion resolves numerous outstanding problems in beta diversity analysis, including sparsity, compositional effects, and uneven sequencing depths, all while giving information about the taxa driving microbial perturbations. This method could possibly be adapted to or combined with other omics paradigms (e.g., metabolomics, metatranscriptomics, and metagenomics) and provides the opportunity to initiate standardization of beta diversity analyses in the microbiome field.

MATERIALS AND METHODS

Preprocessing with rclr. Prior to running matrix completion, the data need to be centered around zero and approximately normally distributed. The centered log ratio (clr) transformation is commonly applied in compositional data analysis before applying PCA. This log transforms each value and then centers them around zero. This is particularly useful when one assumes that the data are lognormally distributed as proposed in reference 42, since log-transformed lognormally distributed data are normally distributed. The clr transform is given below:

$$\mathrm{clr}(x) = \left[\log\frac{x_1}{g(x)},\ldots,\log\frac{x_D}{g(x)}\right] = \log x - \overline{\log x} \tag{1}$$

where $g(x)$ is the geometric mean of all of the taxa. The Aitchison distance can be directly calculated from the Euclidean distance of the clr-transformed data. This is given as follows:

$$d_A(x,y) = \sqrt{\sum_{i=1}^{D}\left(\mathrm{clr}[x]_i - \mathrm{clr}[y]_i\right)^2} = \sqrt{\sum_{i=1}^{D}\left(\log\frac{x_i}{x_j} - \log\frac{y_i}{y_j}\right)^2} \tag{2}$$

The Aitchison distance between the absolute abundances is equivalent to the Aitchison distance on the proportions. In order to center the samples around zero, the average clr-transformed sample needs to be calculated and then subtracted from the remaining samples. Thus, the clr-transformed results will be as follows:

$$y_{ij} = \log x_{ij} - \overline{\log x_i} - \overline{\log x_j} \tag{3}$$

This centering procedure is commonly used prior to performing PCA and eliminates the need to explicitly compute bias constants (43).

The issue with applying the clr transform directly to sparse count data is that the log of zero is undefined. This motivated the construction of an approximate clr transform defined only on nonzero counts. The robust clr (rclr) transform is given as follows:

$$\mathrm{rclr}(x) = \left[\log\frac{x_1}{g_r[x]},\ldots,\log\frac{x_D}{g_r[x]}\right] \tag{4}$$

$$g_r(x) = \left(\prod_{i\epsilon\Omega_x} x_i\right)^{1/|\Omega_x|} \tag{5}$$

where x_i is the abundance of taxa i, Ω_x is the set of observed taxa in sample x, and $g_r(x)$ is the geometric mean defined only on observed taxa. The rationale behind this procedure is that due to the high dimensionality of these data sets, the robust geometric mean (the geometric mean of the log-transformed nonzero data) can serve as an approximation to the true geometric mean. We know from the Central Limit Theorem that as we collect more independent measurements, we approach the true geometric mean:

$$\frac{1}{|\Omega_x|}\sum_{i\epsilon\Omega_x} x_i \to E\left[\log\vec{x}\right] \text{ as } |\Omega_x| \to |\vec{x}| \tag{6}$$

From this we can redefine the transformed result as follows:

$$y_{ij} = \log x_{ij} - \frac{1}{\left|\Omega_{x_{i.}}\right|}\sum_{k\epsilon\Omega_{x_{i.}}} x_k - \frac{1}{\left|\Omega_{x_{.j}}\right|}\sum_{i\epsilon\Omega_{x_{.j}}} x_k \tag{7}$$

where y_{ij} is only defined when $x_{ij} > 0$. The matrix completion methods can then be directly applied to this transformed result.

Matrix completion. OptSpace is a matrix completion algorithm based on a singular value decomposition (SVD) optimized on a local manifold. It has been shown to be quite robust to noise in low-rank data sets (44). The objective function that it optimizes over is given by

$$\min_{U,V}\left|\Lambda[Y - USV^T]\right|_2^2 \tag{8}$$

where U and V are the matrices that are trying to be estimated and S is analogous to a matrix of eigenvalues. Y is the observed values, and Λ is a function such that the errors between Y and USV^T are computed only on the nonzero entries.

Simulations. Simulations were designed to replicate real data sets with low-rank clusters as a proof-of-concept test of OptSpace with and without the rclr preprocessing step. The keyboard data set was chosen as a representative data set to fit the simulation parameters due to the three distinct microbial community clusters observed in the study (M2, M3, and M9). Simulations were built by drawing blocks of n sequences with the microbial proportions given as follows (45):

$$x_{ij} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{[\mu_i - g_j]^2}{2\sigma^2}\right) \tag{9}$$

$$p_{ij} = \frac{x_{ij}}{\Sigma_k x_{kj}} \tag{10}$$

The resulting simulation was induced by multiple noise sources. There was normally distributed error that was applied to the entire matrix. There were also normally distributed errors that were randomly applied to a subset of the entries in the matrix. In addition, there were subsampling errors that were simulated from the Poisson-lognormal (PLN) distribution with an overdispersion parameter ϕ (46) where the final subsampled simulation y_{ij} is represented by:

$$\lambda_{ij} = np_{ij} \tag{11}$$

$$y_{ij} = \mathrm{PLN}\left(\lambda_{ij}, \phi\right) \tag{12}$$

The resulting optimized parameters are optimized rank (number of clusters), the intensity of noise, sequencing depth, the distribution parameters μ and σ, and overlap of features between clusters (i.e., effect size). To resolve the most realistic simulation possible, these parameters were optimized to minimize the KL-divergence between the real data and the simulation with a Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization. The resolved parameters were used to run the simulation at a rank of 2 over sequencing depths ranging from 100 to 10,000 reads/sample. At each depth, before the introduction of noise and subsampling, the sampled data were stored as a base truth to be compared to the reconstruction. Furthermore, the same noisy and subsampled simulation was run with OptSpace with or without rclr preprocessing. The resulting matrix USV^T was compared by KL-divergence to the base truth. The rclr-preprocessed data were inverse transformed by taking the exponential of USV^T before comparison to the base truth. In addition, the simulation, base truth, sample orientation U, and feature loadings V^T were saved at each iteration and compared visually.

The simulation results of improved clustering at uneven sequencing depths were also compared in the real keyboard data set (see case studies for data processing). The data were compared between two subjects at 500 and 100 reads/sample. Ordination and PERMANOVA results were compared for Jaccard, Bray-Curtis, and RPCA with rclr preprocessing. RPCA with rclr preprocessing alleviated the clustering by sequencing depth in the real data set. This was seen both qualitatively (see Fig. S3 in the supplemental material) and through the PERMANOVA F-statistic by subject ID (see Table S2 in the supplemental material).

Case studies. Case studies on real-world data sets were used to compare robust Aitchison PCA to the current state of the art in beta diversity comparison. The sponge, sleep apnea, infant, keyboard, and 88-soil data sets were acquired on 20 September 2018 from Qiita (47) with IDs of 10793, 10422, 101, 232, and 103, respectively. Each data set was run through Qiita with default trimming and Deblur (v. 1.1.0) sOTU (48) picking approach, using QIIME 2 (v. 2018.6.0) (49). The resulting BIOM (50) tables were then filtered for samples greater than 1,000 reads per sample. Phylogeny was built using the most up-to-date GreenGenes using SEPP (51), and taxonomy was assigned through scikit-learn with default QIIME 2 parameters.

The sponge data set was filtered using the metadata so that it contained only samples with either the label healthy or the label stressed. This resulted in a comparison with 218 remaining samples. Similarly, the sleep apnea study was filtered for IHH and air control samples, with a treatment duration of 6 weeks resulting in 189 remaining samples. The infant gut colonization case study was filtered for samples over 500 reads/sample and for a single sample from the mother with the title 101.Mother. The 88-soil data set was filtered for samples over 500 reads/samples. The keyboard data set was filtered for samples over 500 reads/sample and 15 reads/sOTU. Additionally, only subject IDs corresponding to M3, M2, and M9 were retained, giving 67 samples. For comparing ordinations at different numbers of samples, the data sets were filtered for having 1,000 sequences/sample and balanced to have equal numbers of each subgroup (i.e., equal Air and IHH samples). Then samples were removed randomly but equally from each subgroup; this was repeated 10 times. The first iteration was used to plot the ordinations, and the mean score of the iterations was used to plot KNN classification accuracy and PERMANOVA F-statistic.

Both data sets were then preprocessed with the robust centered log ratio (rclr) transform, and RPCA was run with a rank of 2 because there were two metadata categories of interest in each comparison. Weighted UniFrac distances were calculated using generalized UniFrac with an alpha of one (52). Bray-Curtis distances were calculated through QIIME 2 (49). Both weighted UniFrac and Bray-Curtis distances were calculated on tables rarefied to 1,000 reads per sample. PCoA and PERMANOVA analyses for the Bray-Curtis, RPCA distance matrix, and weighted UniFrac were calculated through scikit-bio. The resulting PCoA and PCA axes were plotted through matplotlib (53) with PC1 and PC2 in the *x* and *y* axes, respectively.

The original unprocessed (raw count) tables were sorted by feature loadings from RPCA. Features with a count sum of less than 10 across all samples were filtered out. The resulting table was then clr transformed with a pseudocount of one and plotted as a heat map. Each sOTU was given the lowest classification for the sleep apnea and sponge data sets, respectively.

The features in the PC1 axis of the feature loadings from RPCA were selected to represent a manageable number of taxa to compare between subgroups. Those selected features (sOTUs) from the feature loadings were used for log ratios. Log ratios were calculated from the table used to calculate them. The samples that contained zeros in either the numerator or denominator were removed before calculating the ratios. The correlations between the log ratio and PC1 axis were performed by Pearson correlation via SciPy (54).

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00016-19.

FIG S1, TIF file, 1.7 MB.

FIG S2, TIF file, 1.4 MB.

FIG S3, TIF file, 1.2 MB.

TABLE S1, XLSX file, 0.01 MB.

TABLE S2, XLSX file, 0.01 MB.

ACKNOWLEDGMENTS

We thank Todd P. Coleman (Department of Bioengineering, University of California at San Diego) for his insights into matrix completion. We thank Torsten Thomas for discussions about sponge-associated microbes.

This material is based upon work supported by the National Science Foundation under grant no. 1332344, the National Institutes of Health under grant no. AR071731, and the U.S. Department of Energy, Office of Science, Office of Biological & Environmental Research awards DE-SC0012658 and DE-SC0012586. Cameron Martino is supported in part by the UC San Diego Frontiers of Innovation Scholars Program (FISP). James T. Morton is supported in part by the National Science Foundation under grant no. 1144086.

REFERENCES

1. Kuczynski J, Liu Z, Lozupone C, McDonald D, Fierer N, Knight R. 2010. Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat Methods 7:813–819. https://doi.org/10.1038/nmeth.1499.
2. Hamady M, Knight R. 2009. Microbial community profiling for human microbiome projects: tools, techniques, and challenges. Genome Res 19:1141–1152. https://doi.org/10.1101/gr.085464.108.
3. Morton JT, Toran L, Edlund A, Metcalf JL, Lauber C, Knight R. 2017.

Uncovering the horseshoe effect in microbial analyses. mSystems 2:e00166-16. https://doi.org/10.1128/mSystems.00166-16.

4. Greig-Smith P. 1980. The development of numerical classification and ordination. Vegetatio 42:1–9. https://doi.org/10.1007/BF00048864.
5. Potvin C, Roff DA. 1993. Distribution-free and robust statistical methods: viable alternatives to parametric statistics. Ecology 74:1617–1628. https://doi.org/10.2307/1939920.
6. Tabachnick BG, Fidell LS. 2013. Using multivariate statistics: Pearson new international edition. Pearson Education Limited, London, United Kingdom.
7. Ginter JL, Thorndike RM. 1979. Correlational procedures for research. J Mark Res 16:600. https://doi.org/10.2307/3150840.
8. Lozupone C, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71: 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005.
9. Lozupone CA, Hamady M, Kelley ST, Knight R. 2007. Quantitative and qualitative beta diversity measures lead to different insights into factors that structure microbial communities. Appl Environ Microbiol 73: 1576–1585. https://doi.org/10.1128/AEM.01996-06.
10. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. 2015. Modeling and analysis of compositional data. John Wiley & Sons, New York, NY.
11. Martín-Fernández JA, Palarea-Albaladejo J, Olea RA. 2011. Dealing with zeros. Monogr Stat Appl Probab 44:43–58.
12. Silverman JD, Roche K, Mukherjee S, David LA. 2018. Naught all zeros in sequence count data are the same. bioRxiv https://www.biorxiv.org/content/early/2018/11/26/477794.
13. Candès EJ, Li X, Ma Y, Wright J. 2011. Robust principal component analysis? J ACM 58:1–37. https://doi.org/10.1145/1970392.1970395.
14. Tipping ME, Bishop CM. 1999. Probabilistic principal component analysis. J R Stat Soc B 61:611–622. https://doi.org/10.1111/1467-9868.00196.
15. Jiang B, Ma S, Causey J, Qiao L, Hardin MP, Bitts I, Johnson D, Zhang S, Huang X. 2016. Corrigendum: SparRec: an effective matrix completion framework of missing data imputation for GWAS. Sci Rep 6:37365. https://doi.org/10.1038/srep37365.
16. Cai T, Tony Cai T, Zhang A. 2016. Structured matrix completion with applications to genomic data integration. J Am Stat Assoc 111:621–633. https://doi.org/10.1080/01621459.2015.1021005.
17. McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 10:e1003531. https://doi.org/10.1371/journal.pcbi.1003531.
18. Wu JR, Macklaim JM, Genge BL, Gloor GB. 2017. Finding the centre: corrections for asymmetry in high-throughput sequencing datasets. arXiv 1704.01841 [q-bio.QM]. https://arxiv.org/abs/1704.01841.
19. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. 2010. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A 107:6477–6481. https://doi.org/10.1073/pnas.1000162107.
20. Kullback S, Leibler RA. 1951. On information and sufficiency. Ann Math Statist 22:79–86. https://doi.org/10.1214/aoms/1177729694.
21. Moitinho-Silva L, Nielsen S, Amir A, Gonzalez A, Ackermann GL, Cerrano C, Astudillo-Garcia C, Easson C, Sipkema D, Liu F, Steinert G, Kotoulas G, McCormack GP, Feng G, Bell JJ, Vicente J, Björk JR, Montoya JM, Olson JB, Reveillaud J, Steindler L, Pineda MC, Marra MV, Ilan M, Taylor MW, Polymenakou P, Erwin PM, Schupp PJ, Simister RL, Knight R, Thacker RW, Costa R, Hill RT, Lopez-Legentil S, Dailianis T, Ravasi T, Hentschel U, Li Z, Webster NS, Thomas T. 2017. The sponge microbiome project. Gigascience 6:1–7. https://doi.org/10.1093/gigascience/gix077.
22. Tripathi A, Melnik AV, Xue J, Poulsen O, Meehan MJ, Humphrey G, Jiang L, Ackermann G, McDonald D, Zhou D, Knight R, Dorrestein PC, Haddad GG. 2018. Intermittent hypoxia and hypercapnia, a hallmark of obstructive sleep apnea, alters the gut microbiome and metabolome. mSystems 3:e00020-18. https://doi.org/10.1128/mSystems.00020-18.
23. Dollhopf SL, Hashsham SA, Tiedje JM. 2001. Interpreting 16S rDNA T-RFLP data: application of self-organizing maps and principal component analysis to describe community dynamics and convergence. Microb Ecol 42:495–505. https://doi.org/10.1007/s00248-001-0027-7.
24. Aggarwal CC, Hinneburg A, Keim DA. 2001. On the surprising behavior of distance metrics in high dimensional space, p 420–434. *In* Database theory—ICDT 2001. Springer, Berlin, Germany.
25. Aitchison J, Greenacre M. 2002. Biplots of compositional data. J R Stat Soc C 51:375–392. https://doi.org/10.1111/1467-9876.00275.
26. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, Paredes R, Noguera-Julian M, Calle ML. 2018. Balances: a new perspective for microbiome analysis. mSystems 3:e00053-18. https://doi.org/10.1128/mSystems.00053-18.
27. Feng G, Sun W, Zhang F, Karthik L, Li Z. 2016. Inhabitancy of active Nitrosopumilus-like ammonia-oxidizing archaea and Nitrospira nitrite-oxidizing bacteria in the sponge Theonella swinhoei. Sci Rep 6:24966. https://doi.org/10.1038/srep24966.
28. Usher KM. 2008. The ecology and phylogeny of cyanobacterial symbionts in sponges. Mar Ecol 29:178–192. https://doi.org/10.1111/j.1439-0485.2008.00245.x.
29. Diaz MC, Ward BB. 1997. Sponge-mediated nitrification in tropical benthic communities. Mar Ecol Prog Ser 156:97–107. https://doi.org/10.3354/meps156097.
30. Moitinho-Silva L, Díez-Vives C, Batani G, Esteves AIS, Jahn MT, Thomas T. 2017. Integrated metabolism in sponge-microbe symbiosis revealed by genome-centered metatranscriptomics. ISME J 11:1651. https://doi.org/10.1038/ismej.2017.25.
31. Waterbury BJ. 1986. Biological and ecological characterization of the marine unicellular cyanobacterium Synechococcus. Can Bull Fish Aquat Sci 214:71–120.
32. Erwin PM, Thacker RW. 2008. Cryptic diversity of the symbiotic cyanobacterium Synechococcus spongiarum among sponge hosts. Mol Ecol 17:2937–2947. https://doi.org/10.1111/j.1365-294X.2008.03808.x.
33. Studer N, Desharnais L, Beutler M, Brugiroux S, Terrazos MA, Menin L, Schürch CM, McCoy KD, Kuehne SA, Minton NP, Stecher B, Bernier-Latmani R, Hapfelmeier S. 2016. Functional intestinal bile acid 7α-dehydroxylation by Clostridium scindens associated with protection from Clostridium difficile infection in a gnotobiotic mouse model. Front Cell Infect Microbiol 6:191. https://doi.org/10.3389/fcimb.2016.00191.
34. Jie Z, Xia H, Zhong SL, Feng Q, Li S, Liang S, Zhong H, Liu Z, Gao Y, Zhao H, Zhang D, Su Z, Fang Z, Lan Z, Li J, Xiao L, Li J, Li R, Li X, Li F, Ren H, Huang Y, Peng Y, Li G, Wen B, Dong B, Chen JY, Geng QS, Zhang ZW, Yang H, Wang J, Wang J, Zhang X, Madsen L, Brix S, Ning G, Xu X, Liu X, Hou Y, Jia H, He K, Kristiansen K. 2017. The gut microbiome in atherosclerotic cardiovascular disease. Nat Commun 8:845. https://doi.org/10.1038/s41467-017-00900-1.
35. Kameyama K, Itoh K. 2014. Intestinal colonization by a Lachnospiraceae bacterium contributes to the development of diabetes in obese mice. Microb Environ 29:427–430. https://doi.org/10.1264/jsme2.ME14054.
36. Greenacre M, Lewi P. 2009. Distributional equivalence and subcompositional coherence in the analysis of contingency tables, ratio-scale measurements and compositional data. J Classif 26:29–54.
37. Silverman JD, Washburne AD, Mukherjee S, David LA. 2017. A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6:e21887. https://doi.org/10.7554/eLife.21887.
38. Palarea-Albaladejo J, Martín-Fernández JA. 2015. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics Intell Lab Syst 143:85–96. https://doi.org/10.1016/j.chemolab.2015.02.019.
39. Keshavan RH, Montanari A. 2010. Regularization for matrix completion, p 1503–1507. *In* 2010 IEEE International Symposium on Information Theory. https://ieeexplore.ieee.org/document/5513563.
40. Lauber CL, Hamady M, Knight R, Fierer N. 2009. Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. Appl Environ Microbiol 75:5111–5120. https://doi.org/10.1128/AEM.00335-09.
41. Koenig JE, Spor A, Scalfone N, Fricker AD, Stombaugh J, Knight R, Angenent LT, Ley RE. 2011. Succession of microbial consortia in the developing infant gut microbiome. Proc Natl Acad Sci U S A 108(Suppl 1):4578–4585. https://doi.org/10.1073/pnas.1000081107.
42. Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 10:1200–1202. https://doi.org/10.1038/nmeth.2658.
43. Abdi H, Williams LJ, Valentin D. 2013. Multiple factor analysis: principal component analysis for multitable and multiblock data sets. WIREs Comput Stat 5:149–179. https://doi.org/10.1002/wics.1246.
44. Keshavan RH, Oh S, Montanari A. 2009. Matrix completion from a few entries. *In* 2009 IEEE International Symposium on Information Theory. https://ieeexplore.ieee.org/document/5466511.
45. Aitchison J, Shen SM. 1980. Logistic-normal distributions: some properties and uses. Biometrika 67:261–272. https://doi.org/10.2307/2335470.
46. Aitchison J, Ho CH. 1989. The multivariate Poisson-log normal distribution. Biometrika 76:643–653. https://doi.org/10.2307/2336624.
47. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC,

Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. Nat Methods 15:796–798. https://doi.org/10.1038/s41592-018-0141-9.

48. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2:e00191-16. https://doi.org/10.1128/mSystems.00191-16.

49. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Titus Brown C, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu YX, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, II, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Knight R, Gregory Caporaso J. 2018. QIIME 2: reproducible, interactive, scalable, and extensible microbiome data science. PeerJ Preprints 6:e27295v2. https://doi.org/10.7287/peerj.preprints.27295v2.

50. McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG. 2012. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. Gigascience 1:7. https://doi.org/10.1186/2047-217X-1-7.

51. Janssen S, McDonald D, Gonzalez A, Navas-Molina JA, Jiang L, Xu ZZ, Winker K, Kado DM, Orwoll E, Manary M, Mirarab S, Knight R. 2018. Phylogenetic placement of exact amplicon sequences improves associations with clinical information. mSystems 3:e00021-18. https://doi.org/10.1128/mSystems.00021-18.

52. Chen J, Bittinger K, Charlson ES, Hoffmann C, Lewis J, Wu GD, Collman RG, Bushman FD, Li H. 2012. Associating microbiome composition with environmental covariates using generalized UniFrac distances. Bioinformatics 28:2106–2113. https://doi.org/10.1093/bioinformatics/bts342.

53. Hunter JD. 2007. Matplotlib: a 2D graphics environment. Comput Sci Eng 9:90–95. https://doi.org/10.1109/MCSE.2007.55.

54. Jones E, Oliphant T, Peterson P. 2001. SciPy: open source scientific tools for Python. http://www.scipy.org/.