Discovery and Visualization of Latent Structure with Applications to the Microbiome

Extended Abstract

Kris Sankaran

June 28, 2018

A single question lies behind the research efforts of both the data visualization and statistical modeling communities: What are the most effective techniques for identifying and representing latent structure in data? The problem is that even moderately large collections of data are difficult to mentally process – some reduction, some more succinct representation, is necessary before the data can be used to guide reasoning. In spite of the differences in the substrates – graphical and mathematical – from which these representations are molded, the visualization and statistics communities have arrived at many similar principles for guiding this reduction. Our work blends ideas from both the modeling and visualization communities to make the representation of latent structure more accessible and automatic.

Data is never analyzed in a vacuum. Its collection and study is only valuable as far as it helps resolve important ambiguities in systems of interest. To ground our study, we focus on applications to microbiome data, seeking representations that we believe will simplify the investigation of a variety of microbiome-related questions.

The essential contribution of our work is to streamline and democratize the discovery and visualization of latent structure in the microbiome. Concretely, this involves several lines of study,

- Designing example workflows: There are many possible approaches for a microbiome analysis pipeline, from raw data to model criticism, but few references for how to choose between options and assemble a coherent workflow. One effort to provide some basic guideposts is described in section 1.
- Developing software packages: Sometimes the same conceptually complex or time-consuming representation task appears repeatedly across studies. This has motivated the creation of packages to simplify

these difficult steps, several of which are reviewed in section 2.

- Distilling relevant literature: Sometimes the barrier to effective analysis is not the implementation of a technique, but knowledge of which methods are relevant and effective. This is especially the case for more complex analysis questions, and is the underlying motivation for the studies overviewed in sections 3 and 4.

Our goal is to empower the microbiome community to make full use of ideas developed in statistics and data visualization. Indeed, it is one of the ironies data analysis that simple methods can be very labor intensive, requiring a high degree of user involvement without much guidance, while more sophisticated techniques have the potential to disappear into the background, allowing scientists to instead focus on problems of developing and evaluating theories of microbial ecology.

1 Latent variable modeling workflows

Microbiome studies attempt to characterize variation in bacterial abundance profiles across different experimental conditions [Gilbert et al., 2014]. For example, a study may attempt to describe differences in bacterial communities between diseased and healthy states or after deliberately induced perturbations [Dethlefsen and Relman, 2011, Fukuyama et al., 2017]. Such studies can be illuminating from both basic scientific and medical perspectives.

In the process, two complementary difficulties arise. First, the data are often high-dimensional, measured over several hundreds or thousands of types of bacteria. Studying patterns at the level of particular bacteria is typically uninformative. Second, it can be important to study bacterial abundances in the context of existing biological knowledge.

Viewed from this perspective, a probabilistic approach emerges as a natural candidate. However, although probabilistic latent variable models are a cornerstone of modern unsupervised learning, they are rarely applied in the context of microbiome data analysis, in spite of the evolutionary, temporal, and count structure that could be directly incorporated through such models.

The work [Sankaran and Holmes, 2018] explores the application of probabilistic latent variable models to microbiome data, with a focus on Latent Dirichlet Allocation, Nonnegative Matrix Factorization, and Dynamic Unigram models. To develop guidelines for when different methods are appropriate, we perform a detailed simulation study. We further illustrate and compare these techniques using the data of Dethlefsen and Relman [2011], a study on the effects of antibiotics on bacterial community composition.

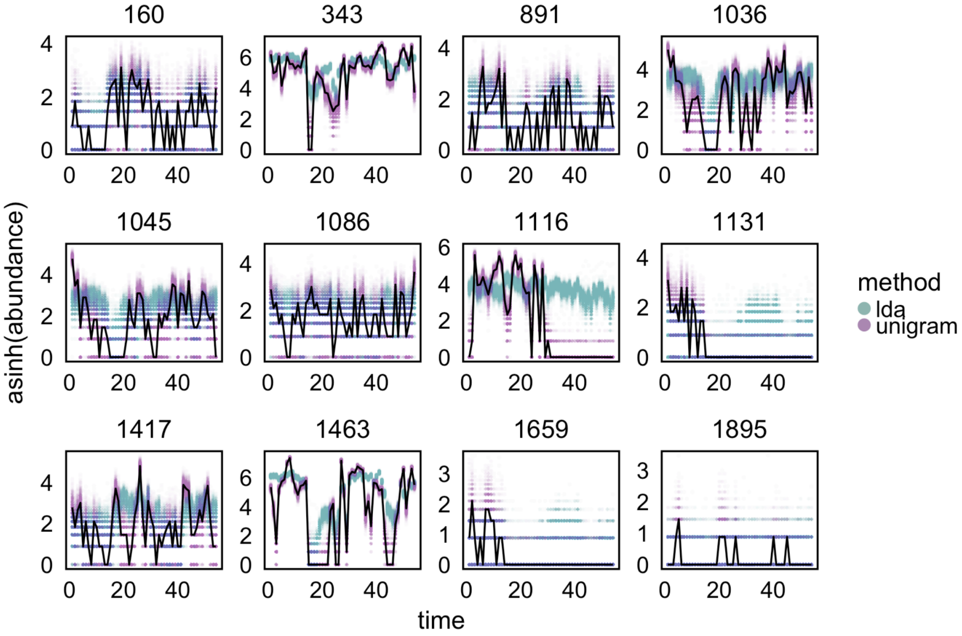

Figure 1: We can visualize the simulated time series for a subset of species and compare them with the observed ones, as a posterior check. Each panel represents one species. The black lines represent the observed asinh-transformed abundances for a subject over time. The blue and purple dots give the posterior predictive realizations for these species over time, according to LDA and the Dynamic Unigram model, respectively.

Model assessment is important for qualifying interpretations, and can guide refinements in subsequent analyses. Our work proposes novel, visual posterior predictive checks tailored to latent variable models, an example of which is available in Figure 1. Code for all algorithms, experiments, and visualizations is available at `github.com/krisrs1128/microbiome_plvm`. A docker image providing a suitable software environment for reproducing the analysis is linked from there as well.

One of the primary contributions of this study is to develop the observation that methods popular in text analysis can be adapted to the microbiome setting in a way that produces useful summaries. We develop the analogy between these text and microbiome analysis and also draw attention to points where the parallels break down. For example, topics in document modeling are analogous to communities in microbiome analysis – these are “prototypical” units which can be used as a point of reference for observed samples. In the same way that it is common to assign topics like“business” or “politics” to newspaper articles, summarizing microbiome samples by their essential bacterial signatures can be a useful mental device.

Critical reflection highlights important discrepancies, however. Among the most fundamental is that unsupervised text analysis techniques are often embedded within automatic systems, for text classification or information retrieval, say, which do not require the intervention of a scientific investigator. In contrast, in microbiome studies, researchers often have control over specific experimental design structure, and collect and analyze data on a per-study basis. In this setting, success is defined somewhat amorphously as an ability to describe the structure and function underlying a biological system of interest. The differences between these fields opens up the possibility for an interesting cross-pollination of ideas, however.

2 Interactive visualization packages

Paralleling our comparison of latent variable modeling techniques, we have evaluated a suite of visualization methods with the goal of speeding up the cycle from data preparation and modeling to interactive exploration and back.

Our approaches are encapsulated three publicly available R packages – treelapse, centroidview, and mvarVis – which encourage data analysts to work at the border between data modeling and visualization, and more generally empower a wider audience to apply less widely known, but powerful, visualization ideas.

The key contributions of these studies are,

- Proposals for visualizing hierarchically structured or high-dimensional data, based on principles from the data visualization community.
- The implementation of these proposals in a publicly available R packages.
- Illustrations of the value of interactive data visualization in scientific contexts, through diverse case studies.

Our treelapse package is motivated by problems that we call tree-structured differential abundance and differential dynamics [Sankaran and Holmes, 2017a]. In the differential abundance problem, we attempt to compare the abundances of individual bacteria across experimental conditions – for example, treatment vs. control or healthy vs. diseased. We call this analysis "tree-structured" because, in practice, researchers generate interpretations about intermediate taxonomic orders – it is more interesting to discover novel behavior taxonomic levels between high-order phyla and low-level species. In the tree-structured bacterial dynamics problem, the goal is to describe changes in bacterial abundances in an environment over time. As in the differential abundance problem, it is useful if these descriptions can be given at the highest subtree at which the pattern appears.

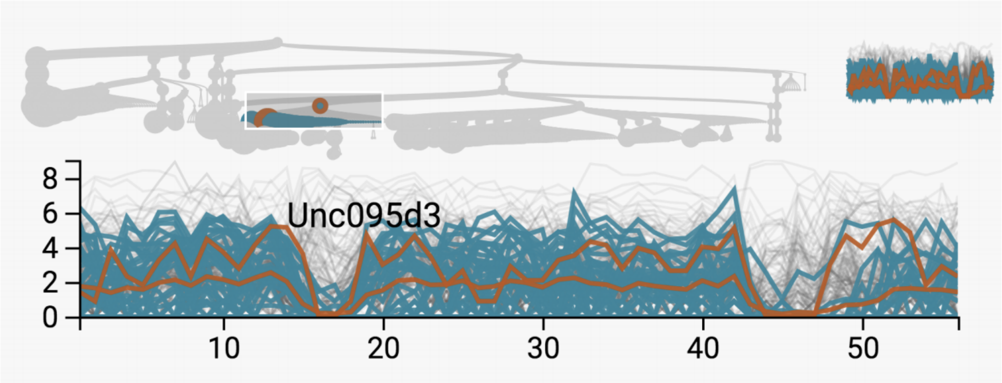

Figure 2: An example application of the treebox interactive display. By drawing a selection on the phylogenetic tree, the user has highlighted time series for species from the Ruminoccocus genus.

Our approach is most directly informed by two principles from the data visualization literature: focus-plus-context and linking [Buja et al., 1996]. From this foundation, we propose three interactive visualization methods: DOI sankeys, timebox trees, and treeboxes. DOI sankeys alow comparison of the flow of bacterial abundance across the phylogenetic tree in a way that allows rapid inspection of differential abundance, using the DOI principle to traverse large swaths of the tree. Timebox trees and treeboxes are designed to facilitate the study of differential dynamics by linking tree and time series views of bacterial abundances – visual queries on the tree can be used to highlight time series of interest, and vice versa, see Figure 2, for example. All methods are available in an R package (`http://krisrs1128.github.io/treelapse`) and a video demonstrating their usage is available at `https://youtu.be/EcmYBRMVMbI`.

The centroidview (`http://github.com/krisrs1128/centroidview`) and mvarVis (`http://github.com/krisrs1128/mvarVis`) packages adapt similar ideas for model inspection, rather than raw data exploration. Specifically, centroidview is designed to facilitate inspection of subtree centroids, a useful follow-up analysis of hierarchical clustering results, but which, in the absence of helper utilities, can be complicated to implement and difficult to visually process. See Figure **??** for an example view. mvarVis, on the other hand, is directed towards the analogous problem in the analysis of multivariate statistics output, giving an interactive alternative to printing pages of plots with slightly modified supplementary variables. Both approaches are algorithm agnostic – they can be applied to generic hierarchical clustering or multivariate analysis output.

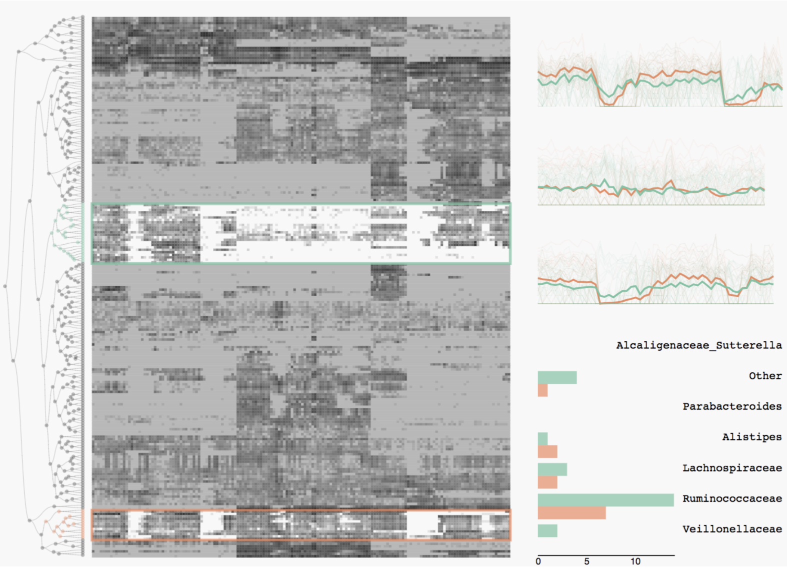

Figure 3: A centroidview display for the antibiotics data of [Dethlefsen and Relman, 2011], demonstrating the potential for interactive visualization to understand algorithmically discovered latent structure. Each row in the heatmap corresponds to one species, and each column is a sample. Samples are first grouped by person, then are sorted by time. The intensity of a cell in the heatmap reflects the abundance of that species in that sample. The hierarchical clustering tree is printed on the left. Subtree centroids are given in the panels along the top right, with one panel per subject and one line per subtree. Taxonomic breakdowns appear in the bottom right. Different colors distinguish different subtrees.

3 Analyzing variation across tables

The simultaneous study of multiple measurement types is a frequently encountered problem in practical data analysis. It is especially common in microbiome research, where several sources of data – for example, 16S, metagenomic, metabolomic, or transcriptomic data – can be collected on the same physical samples [Franzosa et al., 2015, McHardy et al., 2013]. There has been a proliferation of proposals for analyzing such multitable microbiome data, as is often the case when new data sources become more readily available, facilitating inquiry into new types of scientific questions [Fukuyama et al., 2017, Rahnavard et al.].

However, stepping back from the rush for new methods for multitable analysis in the microbiome literature, it is worthwhile to recognize the broader landscape of multitable methods, as they have been relevant in problem domains ranging from economics to robotics to genomics. The purpose of this study is not to develop new algorithms, but rather to (1) distill the relevant themes across different analysis approaches and (2) provide concrete workflows for approaching analysis, as a function of ultimate analysis goals and data characteristics (heterogeneity, dimensionality, sparsity, ...).

For more concrete motivation, we consider data from the WELL-China study, which is focused on the relationships between various indicators of wellness [Stanford Prevention Research Center]. In this study, 1969 individuals underwent clinical examinations, filled out wellness surveys (covering topics such as exercise, sleep, diet, and mental health), and provided stool samples, used for 16S sequencing and metabolomic analysis. To date, 16S sequencing data is available for 221 of these participants. To limit the scope of our case study, we focus on the question: How is the distribution of lean and fat mass across the body, measured using DEXA scans, related to patterns of microbial abundance, measured by 16S sequencing?

We provide summaries about, open-source implementations of, and practical evaluation of methods from classical ordination, multivariate analysis, probabilistic learning, and optimization. We describe approaches that usually confined to particular literature areas using shared, statistical notation and highlight certain similarities in the process – for example, PCA-IV and Bayesian multitask regression were proposed in very different contexts, but have almost the same goal. This work allows us to offer guidelines for when one model might be more appropriate than another, some of which are summarized in Table 1 in the appendix.

4 Inference of dynamic regimes

Many studies have been performed to characterize the dynamics and stability of the microbiome across a range of environmental contexts [Costello et al., 2012]. For example, it is often of interest to identify time intervals within which certain subsets of taxa have an interesting pattern of behavior. Viewed abstractly, these problems often have a flavor not just of time series modeling but also of regime detection, a problem with a rich history across a variety of applications, including speech recognition, finance, EEG analysis, and geophysics.

In [Sankaran and Holmes, 2017b], we distill the core ideas of different regime detection methods, provide example applications, and share reproducible code (`https://github.com/krisrs1128/microbiome_regime_detection`), making these techniques more accessible to microbiome researchers. Specifically, we reanalyze the data of Dethlefsen and Relman [2011] using Classification and Regression Trees (CART), Hidden Markov Models (HMMs), Bayesian nonparametric HMMs, mixtures of Gaussian Processes (GPs), switching dynamical systems, and multiple changepoint detection. Along the way, we summarize each method, their relevance to the microbiome, and the tradeoffs associated with using them. Ultimately, our goal is to describe types of temporal or regime switching structure that can be incorporated into studies of microbiome dynamics.

The primary contributions of this study are,

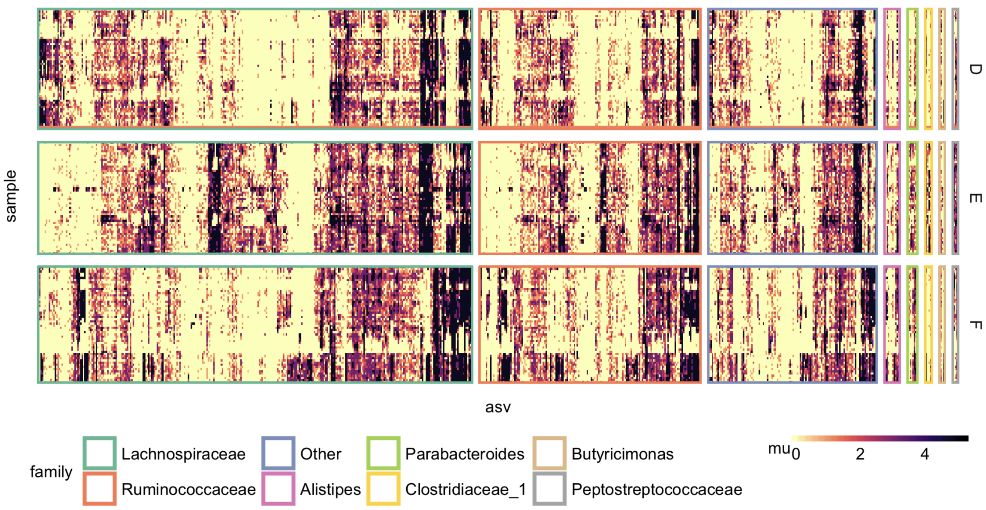

Figure 4: An example smooth from the sticky HDP-HMM, one of the regime detection methods we describe. Each column corresponds to a single species, each row is a timepoint, and panels represent different individuals. This view allows a comparison of the effects of antibiotics across different subjects and species families.

- The relation of the regime detection problem to several statistical frameworks, and a comparison of the types of interpretation facilitated by each.
- The development of experiments to evaluate the practical utility of these different formulations.
- A catalog of algorithm pseudocode and complete implementations, to serve as a reference for researchers interested in regime detection.
- The design of and code for static visualizations that can be used to evaluate the results of various methods.

We set the stage by articulating the scientific problem of interest in more detail and provide a high-level statistical formulation. To establish reference points for more complex methods, we describe approaches which are easy to implement, but that fail to incorporate temporal structure. Then, we review and apply smoothing and mixture modeling techniques relevant to this problem. An example of the type of data reduction we seek is given in Figure 4. Besides our methodological distillations, our implementations and visualizations can help researchers decide whether a certain method is appropriate for their use case, depending on the form of reduction that would be useful and the computational budget alloted.

5 Outlook

In the thesis reviewed here, we evaluated workflows, developed software, and distilled literature relevant to discovery and visualization of latent structure in the microbiome. We considered techniques from both formal modeling and exploratory data analysis, highlighting the ways in which these complementary points of view can both applied to the process of iterating towards more refined, compact representations of complex data.

This work lays the groundwork for potential projects related to visualization and workflow evaluation of biological data. More fundamentally, this work has adopted the perspective that the computational comparison of existing methods, through simulations or illustrations, is often as valuable to data analysts as the development of novel algorithms. Indeed, we hope these examples can guide the choices faced by practitioners in their day-to-day work. In this way, we emulate classical statistical theory, giving modern analogs to classical comparison of experiments. Clearly, much work remains to be done, and the proliferation of algorithms is both a blessing and a curse: while there are more options available, some possibly tailor-made to problems of interest, there are few objective guidelines available to inform the actual decision of which to apply, and there is very little sense of when one approach is optimal. We hope that the foundation laid out by our studies will be relevant to the creation and evaluation of methods related to data visualization and latent variable modeling in work to come.

A Appendix

Property	**Algorithms**	**Consequence**
Analytical solution	Concat. PCA, CCA, CoIA, MFA, PTA, Statico / Costatis	Methods with analytical solutions generally run much faster than those that require iterative updates, optimization, or Monte Carlo sampling. They tend to be restricted to more classical settings.
Require covariance estimate	Concat. PCA, CCA, CoIA, MFA, PTA, Statico / Costatis	Methods that require estimates of covariance matrices cannot be applied to data with more variables than samples, and become unstable in high-dimensional settings.
Sparsity	SPLS, Graph-Fused Lasso, Graph-Fused Lasso	Encouraging sparsity on scores or loadings can result in more interpretable, results for high-dimensional data sets. These methods provide automatic variable selection in the multitable analysis problem.
Tuning parameters	*Sparsity*: Graph-Fused Lasso, PMD, SPLS *Number of Factors*: PCA-IV, Red. Rank Regression, Mixed-Membership CCA *Prior Parameters*: Mixed-Membership CCA, Bayesian Multitask Regression *Kernel*: KCCA	Methods with many tuning parameters are often more expressive than those without any, since it makes it possible to adapt to different degrees of model complexity. However, in the absence of automatic tuning strategies, these methods are typically more difficult to use effectively.
Probabilistic	Mixed-Membership CCA, Bayesian Multitask Regression	Probabilistic techniques provide estimates of uncertainty, along with representations of cross-table covariation. This comes at the cost of more involved computation and difficulty in assessing convergence.
Not Normal or Nonlinear	KCCA, CCpNA, Mixed-Membership CCA, Bayesian Multitask Regression	When data are not normal (and are difficult to transform to normality) or there are sources of nonlinear covariation across tables, it can be beneficial to directly model this structure. KCCA allows the most general types of nonlinearity, while the probabilistic methods are suited to specific count-structure.
>2 Tables	Concat. PCA, CCA, MFA, PMD, KCCA	Methods that allow more than two tables are applicable in a wider range of multitable problems. Note these are a subset of the cross-table symmetric methods.
Cross-Table Symmetry	Concat. PCA, CCA, CoIA, Statico / Costatis, MFA, PMD, KCCA	Cross-table symmetry refers to the idea that some methods don't need a supervised or multitask setup, where one table contains response variable and the other requires predictors. The results of these methods do not change when the two tables are swapped in the method input.

Table 1: A high-level comparison of the multitable analysis methods discussed in this review. The purpose of this table is to give rules-of-thumb that can guide practical application, where choices invariabily depend on the scale and structure of the data, the goals of the analysis, the expected number of future workflow applications, and availability of programming computation time.

References

Andreas Buja, Dianne Cook, and Deborah F Swayne. Interactive high-dimensional data visualization. *Journal of computational and graphical statistics*, 5(1):78–99, 1996.

Elizabeth K Costello, Keaton Stagaman, Les Dethlefsen, Brendan JM Bohannan, and David A Relman. The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262, 2012.

Les Dethlefsen and David A Relman. Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation. *Proceedings of the National Academy of Sciences*, 108(Supplement 1):4554–4561, 2011.

Eric A. Franzosa, Tiffany Hsu, Alexandra Sirota-Madi, Afrah Shafquat, Galeb Abu-Ali, Xochitl C. Morgan, and Curtis Huttenhower. Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. *Nature Reviews Microbiology*, 13(6):360–372, apr 2015. doi: 10.1038/nrmicro3451. URL `https://doi.org/10.1038/nrmicro3451`.

Julia Fukuyama, Laurie Rumker, Kris Sankaran, Pratheepa Jeganathan, Les Dethlefsen, David A Relman, and Susan P Holmes. Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment. *PLoS Computational Biology*, 13(8):e1005706, 2017.

Jack A Gilbert, Janet K Jansson, and Rob Knight. The earth microbiome project: successes and aspirations. *BMC biology*, 12(1):69, 2014.

Ian H McHardy, Maryam Goudarzi, Maomeng Tong, Paul M Ruegger, Emma Schwager, John R Weger, Thomas G Graeber, Justin L Sonnenburg, Steve Horvath, Curtis Huttenhower, Dermot PB McGovern, Albert J Fornace, James Borneman, and Jonathan Braun. Integrative analysis of the microbiome and metabolome of the human intestinal mucosal surface reveals exquisite inter-relationships. *Microbiome*, 1(1):17, 2013. doi: 10.1186/2049-2618-1-17. URL `https://doi.org/10.1186/2049-2618-1-17`.

Gholamali Rahnavard, Eric A. Franzosa, Lauren J. McIver, Emma Schwager, George Weingart, Yo Sup Moon, Xochitl C. Morgan, Levi Waldron, and Curtis Huttenhower. High-sensitivity pattern discovery in large multi'omic datasets. URL `https://huttenhower.sph.harvard.edu/halla`.

Kris Sankaran and Susan Holmes. Interactive visualization of hierarchically structured data. *Journal of Computational and Graphical Statistics*, pages 0–0, oct 2017a. doi: 10.1080/10618600.2017.1392866. URL `https://doi.org/10.1080/10618600.2017.1392866`.

Kris Sankaran and Susan P Holmes. Inference of dynamic regimes in the microbiome. *arXiv preprint arXiv:1712.00067*, 2017b.

Kris Sankaran and Susan P Holmes. Latent variable modeling for the microbiome. *Biostatistics*, jun 2018. doi: 10.1093/biostatistics/kxy018. URL `https://doi.org/10.1093/biostatistics/kxy018`.

Stanford Prevention Research Center. WELL-China: New wellness solutions. URL `https://prevention.stanford.edu/content/dam/sm/prevention/documents/about/WELL-CHINA.pdf`.