Cell Host & Microbe
Forum

CellPress

Bringing the Dynamic Microbiome to Life with Animations

Yoshiki Vázquez-Baeza,[1] Antonio Gonzalez,[2] Larry Smarr,[1,3] Daniel McDonald,[2] James T. Morton,[1] Jose A. Navas-Molina,[1] and Rob Knight[1,2,*]
[1]Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA 92093-0404, USA
[2]Department of Pediatrics, University of California, San Diego, La Jolla, CA 92093, USA
[3]California Institute for Telecommunications and Information Technology, University of California, San Diego, La Jolla, CA 92093-0436, USA
*Correspondence: robknight@ucsd.edu
http://dx.doi.org/10.1016/j.chom.2016.12.009

Our bodies and natural environment contain complex microbial communities, colloquially termed microbiomes. We previously created a web-based application, EMPeror, for visualizing ordinations derived from comparisons of these microbiome communities. We have now improved EMPeror to create interactive animations that connect successive samples to highlight patterns over time.

Introduction

Recent developments in high-throughput DNA sequencing, improvements in molecular methods, and the continual increase in computational power have enabled microbiome researchers to test ever-more sophisticated hypotheses and study designs. Just a decade ago, a microbiome study with a few dozen samples was considered large. However, new technologies now enable studies with thousands of samples that can explore differences at the scale of populations or environments.

One iconic result of these studies is the presentation of aggregated microbiomes from different human body sites within a single plot, where samples from equivalent body sites across many different individuals all cluster together (for an example, see Figure 1). This body site clustering was initially observed in a study of nine people (Costello et al., 2009) and then replicated across hundreds of people in the Human Microbiome Project (HMP) (HMP Consortium, 2012). As useful as these static snapshots have been in examining the variation in diversity across populations, they do not reveal temporal dynamics, which may stem from intrinsic properties of the ecosystem such as inter-species interactions or from external drivers such as drugs or diet (in humans) or environmental perturbations. As seen in macroscopic ecology, changes in the relative abundance of organisms can occur on short timescales. Given that population doubling times of some bacteria can be less than 20 min, these timescales can potentially be very short for commensal communities. A single gram of stool can contain more than one billion microbial cells, 1,000 times the density of microbes in the ocean surface. These microbes are represented by hundreds of species, thousands of strains, and an estimated 3.3 million microbial genes (Qin et al., 2010). Individual species and strains compete for resources (e.g., fiber), which can impose limitations on organism populations as the supply and demand of resources are dynamic over time. Additionally, there are time-dependent bacteriophages which target and kill specific bacteria, or in some cases, promote horizontal gene transfer (HGT), which can alter the strains (e.g., antibiotic-resistance genes are commonly transferred via HGT). Furthermore, many selective pressures act on the microbiome. For example, the human immune system acts as a regulatory force for the constituent organisms. This includes the innate and adaptive immune system, but also the variety of white blood cells and their antimicrobial glycoproteins. Additionally, microbial cells, many of which are capable of highly optimized metabolic processes, are subjected to a constantly changing set of compounds from the host's diet, as well as an intermittent exposure to antibiotics and other xenobiotics (in the case of diseased subjects, antibiotic use can lead to dysbiotic microbiomes, especially when taken frequently and at high doses).

The emergence of quantitative microbial ecology dynamics will be important for strengthening our basic scientific understanding of the microbial world. Additionally, it will also quickly become an essential tool for medicine. Currently, when antibiotics or other pharmaceuticals are prescribed, the doctor does not know the "state" of the microbiome or how it will be altered. Being able to read out and predict changes in the microbiome will enable a field of precision medicine for restoring the altered ecology to a healthy one. Similarly, there is great hope that designer pre- and probiotics will be powerful tools for "gardening" one's microbiome back to health. However, until more quantitative and large-scale trials have been performed with longitudinal tracking of the changes in the microbiome, the precise impact of these tools will remain, at best, a conjecture.

With the improved understanding of human microbial dynamics, the conceptual framework of medicine will continue to experience big changes. As recently described (Costello et al., 2012), many fundamental concepts from ecological theory are directly applicable to the host-microbiome interactions. However, these topics have not yet been widely incorporated as part of general medical training. Concepts of succession and community assembly in previously unoccupied habitats (infant growth of microbial diversity), assembly after disturbance (recovery from antibiotics), and blooming of invasive species (food poisoning by pathogens) are important. Each of these topics will be elucidated in the coming decade by analyzing time series of microbial ecologies, especially in testings with defined perturbations, and integrated

CrossMark

with metabolomics, metatranscriptomics, metaproteomics, and host biomarkers.

Ordinations

Datasets that have been obtained from DNA sequencing often have very high dimensionality, up to thousands of dimensions for each sample. Given the high dimensionality of microbiome studies, dimensionality reduction techniques, otherwise known as ordinations, have been widely used in the field (some examples include principal component analysis [PCA], principal coordinates analysis [PCoA], non-metric multidimensional scaling [NMDS], and correspondence analysis [CA]). These techniques are analogous to modern photography: photographs capture 2D images of a 3D environment, and although they do not capture all of the information about the landscape, they can provide enough information to differentiate and identify objects. Similarly, ordination methods are valuable tools for capturing snapshots of these microbial populations and better understanding the major driving factors affecting these populations.

Although it is clear that the frontier of microbiome research is shifting to understanding microbiome dynamics, it remains a challenge to "see" the dynamics in these increasingly large datasets. Traditional time-series analysis techniques were developed around densely sampled and evenly collected variables, restrictions that are often impossible in clinical and ecological settings. Beyond these technical limitations with sampling, the complex dynamics of most microbial ecology datasets violate assumptions of periodicity and stationarity, used in many statistical methods for time-series analysis. For an expanded discussion and review of this topic, see Faust et al. (2015).

A few pioneering studies have been published for microbiome time-series, but these visualizations were typically performed using custom "one-off" software that is generally harder to modify, maintain, and reuse. However, these examples showed the promise of visualizing microbiome dynamics. For instance, Flores et al. (2014) described in detail the per-subject alpha and beta diversity variations, as well as the individual variability observed in the time point to time point comparisons. Another remarkable example presented by David et al. (2014) examined pre- and post-dietary intervention comparisons at the microbiome and gene expression level. Similarly, Lozupone et al. (2013) used the data from an infant's first 27 weeks of life to show that you can visualize the assembly of gut communities and contextualize it with the data of healthy human adults of the HMP. This analysis shows a step-by-step progression of the infant's communities as they transition into one of a healthy adult. Caporaso et al. (2011) showcased the variability in the three body sites of the two surveyed subjects. More importantly, it gradually painted a picture, sample by sample, of the overall stability and dynamics present in the dataset.

EMPeror

EMPeror is a response to the many requirements that are specifically relevant to high-throughput microbiome studies. The core idea behind this software is to allow users to rapidly explore a dataset, traversing hundreds of sample metadata categories and minimizing the necessity of custom plotting code. All of this can occur without requiring a complex software installation or domain-specific knowledge.

EMPeror (Vázquez-Baeza et al., 2013) is a web-enabled scatterplot viewer for microbiome analysis that was originally made available in 2013 as part of the QIIME pipeline (Caporaso et al., 2010) and was developed for the Earth Microbiome Project (http://www.earthmicrobiome.org). To use EMPeror, two files need to be provided. The first and most important is the sample metadata file, which contains covariate information about each sample; the second file is a matrix specifying coordinate positions for each sample along a set of axes. In general, for microbiome studies, the coordinate positions are derived by PCoA. Besides allowing the easy exploration of high-dimensional color ordination spaces (such as PCA or PCoA), EMPeror supports more specialized techniques, ranging from displaying jackknifed-resampled data to assess statistical confidence in clustering, to displaying procrustes analysis plots to determine the goodness of fit between two independent datatypes, and even being able to visualize biplots. As of version 0.9.5, users can now create animations inside the user interface without the necessity of any additional computation.

Like most functionality in EMPeror, animations themselves are made possible by selecting two metadata categories, a *trajectory* category and a *gradient* category. The first determines how the samples are grouped together (e.g., individual), and the second describes the order of the grouped samples. As readers might suspect, the ordering is only unambiguous when the values in the gradient category are numeric; therefore, we recommend that only these data types are used. Along with these metadata selectors, EMPeror displays a *play*, a *pause*, and a *rewind* button. These control the state of the traces being presented on screen.

With this mapping, samples that belong to the same trajectory are subsequently connected by an animated trace (sample by sample). The traces are created by linearly interpolating a series of points between two samples. The number of points present in this interpolation depends on the absolute difference between the two samples in the gradient, meaning that samples that were collected further apart along a gradient (e.g., if some samples are daily and some are weekly) will take longer to be connected in the animation than samples that were collected closer in the gradient. This allows heterogenous sampling schemes, common in clincal trials or when some samples are lost or destroyed during processing, to be accommodated in the same animation.

Importantly, these animations do not interfere with the interactivity of a plot. The rest of the controllers (used to manipulate the rotation, zooming, coloring, size, opacity, etc.) remain available and can be used interactively to create a narrative that is more relevant for the data. Lastly, EMPeror is a stand-alone application that can be easily integrated into a software workflow, for instance with Jupyter notebooks (Perez and Granger, 2007).

Example

To showcase the utility of this particular type of visualization, we will reproduce the supplementary video presented in 2015 (Weingarden et al., 2015), where four *Clostridium difficile*-infected (CDI) patients are surveyed before and after a fecal material transplant (FMT) to ameliorate the consequences of this enteric infection.

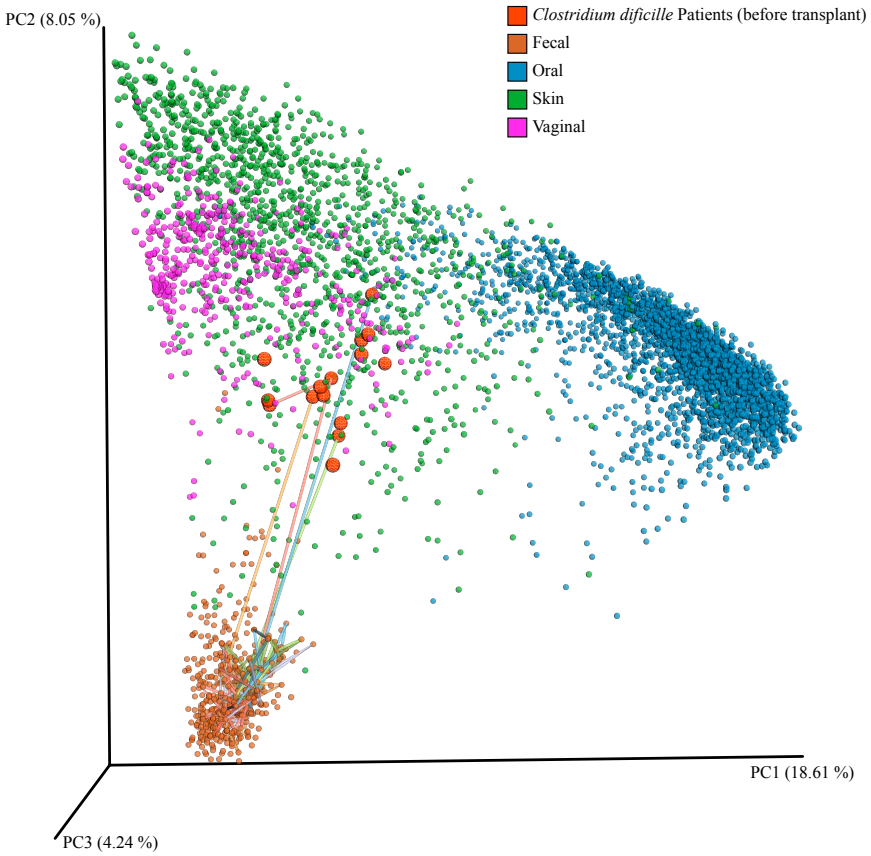


Figure 1. Case Study: Remission of *Clostridium difficile* Infection
Principal coordinates analysis plot of unweighted UniFrac distances. After receiving the fecal transplant, the patients undergo a dramatic transition in their gut microbiomes. Samples from healthy individuals are colored blue for oral, green for skin, pink for vaginal, and brown for fecal communities. Patient samples (preceding the transplant) are colored in orange, and the traces indicate the trajectory they follow over time (only four subjects receive a transplant). Access this interactive animation at http://emperor.microbio.me/animation/.

To create this visualization, we must combine this individual dataset with the 16S rRNA amplicon samples from the HMP (HMP Consortium, 2012), which includes thousands of microbiome samples collected from 242 healthy individuals. These HMP samples create a visual "map" of normal oral, skin, vagina, and stool microbiomes and provide context for the CDI samples. To combine these studies, we used Qiita (https://qiita.ucsd.edu), an open-source platform for analysis and visualization of microbiome datasets. Qiita allows users to combine samples, perform the necessary computations for PCoA (among other analyses), and visualize the result in EMPeror.

For this example, we selected the number of days since the FMT procedure as the gradient category and we selected the subject as the trajectory category. By animating these traces, we first observe that the original patients with CDI had gut microbiomes that did not resemble healthy adult fecal communities. Immediately after the FMT, we observe a rapid transition into a healthy state, which can be inferred from the movement of the sample locations by the end of the trajectory to the normal stool region defined by the HMP (Figure 1). Critically, these transitions were associated with the patient's immediate recovery from the illness.

Limitations

The objective of ordinations is to find the dimensions that best describe as much as possible the variation in a dataset, and therefore reduce the amount of information that we have to interpret. This property is most often useful for its ability to distinguish between groups of samples. However, it also makes this technique sensible to outliers. Consequently, it is important to remember that the variation explained in a statistical model does not necessarily correspond to biological importance. In fact, some methods of measuring distance between samples and ordination techniques for reducing dimensionality can explain much of the variation in a given dataset but fail to reveal biologically important patterns (Kuczynski et al., 2010). In addition, PCoA is susceptible to an artifact called the horseshoe effect, where the shared absence of taxa in the middle of the gradient causes the ends to appear similar and wrap around (other techniques such as non-metric multidimensional scaling are susceptible to a related problem called the arch effect, where the second axis is a distorted reflection of the first), and very sparse data matrices where most samples are highly dissimilar to one another can produce artifacts resembling spikes at 90 degree angles to one another. Consequently, choosing the wrong distance metric or dimensionality reduction method for a dataset can result in visual artifacts that render those visualizations unusable. These problems are less frequent, but still possible, when phylogenetic distance metrics such as UniFrac (Lozupone and Knight, 2005) are used.

Animating the transitions between samples provides a useful visual aid. However, the observed patterns of variation still need to be assessed using descriptive statistics. Differences among sets of community samples can be assessed using the analysis of similarities (ANOSIM) or a permutational multivariate analysis of variance (PERMANOVA). These statistical tests compare the groups using the distance matrix to provide a statistic that measures whether sample category labels tend to group similar samples together more often than a random shuffling of the labels. Alternatively, feature-level differences can be assessed using compositional analyses, correlation networks, or discriminant features between two groups (using supervised and unsupervised machine-learning techniques).

In our experience, the interpretability of animated ordinations is most commonly affected in two cases. The first case can be observed when an inappropriate frame of reference is used or when only the

animated traces are presented. For the example above, the HMP data provide a relevant context due to the types of samples being characterized and the size of the study. However, a reference frame that included only oral samples, or that contained only a single stool sample, would provide not nearly as relevant a data frame for display, and the resulting animations would be difficult to interpret. The second case occurs when many trajectories are animated simultaneously, making the ordination busy and increasing the difficulty of correctly identifying the patterns of variation that are most biologically relevant. This limitation can be overcome by choosing more similar sets of samples to display, such as patients with more consistently defined clinical phenotypes.

Conclusions

Microbiome studies are steadily increasing in size and complexity, challenging our ability to unravel the structure of the ecological interactions occurring at the microbial scale. As researchers, the software we use greatly impacts our ability to ask questions of the data. Platforms like the Jupyter Notebook have greatly facilitated reproducibility of analyses. The success of the Notebook has been, in part, due to building on top of an open and common web infrastructure, the same that EMPeror has adopted. Interactive animations like the example presented above can be directly shared with anyone that has access to a modern web browser, whether it be in a personal computer or in a mobile device.

Sample information (also known as sample metadata) is indispensable to data reuse and necessary for the creation of complex visuals, as in Figure 1. This task was partly possible by leveraging the standardization of sample metadata enforced in Qiita. Without standardization, future generations must undertake the burden of post hoc metadata curation, a process that is deeply frustrating and impossible to automate.

Finally, although EMPeror provides a handful of ways to interrogate a dataset, we urge the community to create other open-source interactive tools that democratize robust, reproducible, and interactive analysis techniques. These tools should be usable not only by a few experts, but rather a broader audience that can benefit from these technological advances.

Availability

The source code and documentation for version 0.9.60 can be found at http://biocore.github.io/emperor/. Additionally, a tutorial (including the data) to generate Figure 1 can be found at http://biocore.github.io/emperor/build/html/tutorials/animations.html.

REFERENCES

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., et al. (2011). Moving pictures of the human microbiome. Genome Biol. *12*, R50.

Costello, E.K., Lauber, C.L., Hamady, M., Fierer, N., Gordon, J.I., and Knight, R. (2009). Bacterial community variation in human body habitats across space and time. Science *326*, 1694–1697.

Costello, E.K., Stagaman, K., Dethlefsen, L., Bohannan, B.J., and Relman, D.A. (2012). The application of ecological theory toward an understanding of the human microbiome. Science *336*, 1255–1262.

David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E., and Alm, E.J. (2014). Host lifestyle affects human microbiota on daily timescales. Genome Biol. *15*, R89.

Faust, K., Lahti, L., Gonze, D., de Vos, W.M., and Raes, J. (2015). Metagenomics meets time series analysis: unraveling microbial community dynamics. Curr. Opin. Microbiol. *25*, 56–66.

Flores, G.E., Caporaso, J.G., Henley, J.B., Rideout, J.R., Domogala, D., Chase, J., Leff, J.W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., et al. (2014). Temporal variability is a personalized feature of the human microbiome. Genome Biol. *15*, 531.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N., and Knight, R. (2010). Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. Nat. Methods *7*, 813–819.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. *71*, 8228–8235.

Lozupone, C.A., Stombaugh, J., Gonzalez, A., Ackermann, G., Wendel, D., Vázquez-Baeza, Y., Jansson, J.K., Gordon, J.I., and Knight, R. (2013). Meta-analyses of studies of the human microbiota. Genome Res. *23*, 1704–1714.

Perez, F., and Granger, B.E. (2007). IPython: A system for interactive scientific computing. Comput. Sci. Eng. *9*, 21–29.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., et al.; MetaHIT Consortium (2010). A human gut microbial gene catalogue established by metagenomic sequencing. Nature *464*, 59–65.

Vázquez-Baeza, Y., Pirrung, M., Gonzalez, A., and Knight, R. (2013). EMPeror: a tool for visualizing high-throughput microbial community data. Gigascience *2*, 16.

Weingarden, A., González, A., Vázquez-Baeza, Y., Weiss, S., Humphry, G., Berg-Lyons, D., Knights, D., Unno, T., Bobr, A., Kang, J., et al. (2015). Dynamic changes in short- and long-term bacterial composition following fecal microbiota transplantation for recurrent Clostridium difficile infection. Microbiome *3*, 10.