Developing a Density Map-based Visualization Tool for Metagenomics Analysis

Author: Daniel Munro
Faculty Mentor: Qunfeng Dong, Ph.D., Department of Biological Sciences, College of Arts & Sciences; Department of Computer Science and Engineering, College of Engineering
College and Department Affiliation: Department of Biological Sciences, College of Arts & Sciences; Honors College

Bio:

Daniel Munro graduated from the University of North Texas with a Bachelor of Science degree in May 2013, with a major in Biology and a minor in Chemistry. Through working in the bioinformatics lab of Dr. Qunfeng Dong since January 2011, Daniel has coauthored four publications in peer-reviewed journals, including PLOS ONE, The ISME Journal, and BMC Bioinformatics. His research topics include human microbiomes, animal microbiomes, and software development. In the fall of 2013 he is entering the Quantitative and Computational Biology Ph.D. program at Princeton University, and has received a National Science Foundation Graduate Research Fellowship to fund his graduate studies.

Abstract:

Microbial community composition is usually visually represented using pie charts, bar charts, and phylogenetic trees. However, such representations do not have fixed layouts, making it difficult to compare multiple figures that were created independently. A new visualization method has been developed that uses a standard two-dimensional map of bacterial diversity. Each sample is projected as a density map onto this grid, so that different figures have similar layouts and differences can be visually assessed. The reliability and validity of this method has been tested using publicly available data from the well-known Human Microbiome Project. This new visualization method can help microbial ecologists visually compare bacterial communities from different sources, which could lead to new insights in metagenomics and ecology.

Introduction

Until the last few decades, our understanding of microorganisms has been confined to a small fraction of microbial diversity because most microbes cannot be cultured in a laboratory to be studied (Riesenfeld, Schloss, and Handelsman 2004). However, the advent of DNA sequencing has allowed new microbes to be identified by extracting and sequencing their DNA from an environment. Metagenomics is the study of microbiomes, or communities of microorganisms, involving sequencing and analysis of the collective genomic DNA in a location. Such large-scale surveys of the microbial diversity in a location have only been feasible due to advances in DNA sequencing technology in the last decade. Such studies are important for understanding how these microbes affect our environment and health.

For the past two years, I have had first-hand experience in analyzing large amounts of metagenomic data, including three recently published studies that I co-authored: a study of the subgingival microbiome of periodontitis and diabetes patients (Zhou et al. 2013), a study of the microbiomes influenced by bovine mastitis (Keuhn et al. 2013), and a study of the microbiomes in blood-sucking arthropods (Hawlena et al. 2012). Such data analysis relies heavily on statistical approaches. However, due to the high dimensionality that is usually associated with microbiome data, scientists also employ visualization tools to recognize patterns in the data that can be undetected by statistical formulas. I have also gained experience in creating such tools as a co-author in the development of the Multi-Genome Synteny Viewer, which has exposed me to the challenges and the importance of visualizing bioinformatics data in a clear but informative way (Revanna et al. 2012).

Visual representation of communities is a challenge in microbial ecology because of the large amount of information that must be condensed into an image. Different methods have been devised to portray the taxonomic profiles of these samples. Some of these methods are designed to facilitate comparison within a set of samples, and may also combine samples into a single representation. However, even the popular visualization methods are not ideal for comparison involving samples from one or more existing studies. Such comparison of published data to new data is important for verification of accuracy and to reveal differences resulting from the alteration of experimental design or biological source. Given the importance of this sharing of data, it would be helpful to present microbiome sample profiles in a way that allows at least preliminary comparison to other samples without reprocessing the underlying microbial census data. How should microbial communities be visually represented to allow for comparison across studies? A new method will be presented here that meets this requirement, allowing the visual comparison of graphic representations of samples generated independently from different studies, even if the studies use different sample sizes and taxonomic resolution.

Summary of Existing Tools:

Software for processing and analyzing raw metagenomic data usually produce plots for users to interpret the results. MG-RAST, a popular online tool, produces conventional figures such as PCA plots and combined heatmap/dendrograms (Meyer et al. 2008). These combined plots convey detailed information of sample composition while still allowing quick visual comparison of samples within the data set. MEGAN is a metagenomic analysis program that can produce bar plots, pie charts, heatmaps, and phylogenetic trees for comparing the abundance profiles of samples in the data set (Huson et al. 2007). Although designed primarily for the processing and quantitative analysis of the data, these programs produce figures that allow the users to interpret their data and plan further analyses.

Other tools are designed specifically for visualization, requiring processed data as input and producing in-depth, exploratory figures. Krona is a program for displaying the taxonomic profile of metagenomic samples (Ondov, Bergman, and Phillippy 2011). It uses a circular design to show the full hierarchical taxonomy in a spatially efficient manner. It is interactive, using animated transitions to zoom into a detailed portion of the data or switch among a set of samples. The iTOL web tool also employs a circular hierarchical design, and can accommodate a variety of embedded plots to characterize the taxa (Letunic and Bork 2007). The focus of these plots, though, is the central dendrogram that can show detailed phylogenetic history. These two tools can show detailed but visually intuitive plots using the results of other sequence processing programs.

Strengths and Weaknesses:

The visualization methods described above can all effectively show the taxonomic composition of samples, though each has different strengths. Krona gives an informative view of relative abundance at all taxonomic levels (Ondov, Bergman, and Phillippy 2011). The design of iTOL seems to suggest a greater focus on the taxa themselves, with a phylogenetically accurate dendrogram in the center and additional data arranged around the taxa (Letunic and Bork 2007). Depending on the data and the goals of the analysis, one of these programs may be more effective than the other in that instance.

Also of great importance is the ability to visually compare samples to each other. The heatmaps produced by MG-RAST are perhaps the most informative in this regard (Meyer et al. 2008). Not only are samples lined up to show differences in each individual taxon’s abundance, but samples are arranged in a hierarchy according to similarity of taxonomic composition, removing some of the guesswork in this multivariate comparison. MEGAN can produce a combined bar plot and phylogenetic tree that allows comparison across samples of individual taxa at multiple levels (Huson et al. 2007). Krona uses animated transitions to visually show differences, although only one sample is visible at any given time (Ondov, Bergman, and Phillippy 2011). Comparison of samples within a dataset is vital for studies that aim to compare the microbiota in different conditions, locations, or times.

However, once they are produced for a particular sample or set of samples, these figures cannot be easily compared with other figures, even if both were produced with the same tool using the same settings. Generally, this problem occurs because the figure's layout is not static, but rather is determined by the individual sample. For example, it can be difficult to compare two separate Krona figures because the sectors that represent taxa are ordered by abundance in that sample and may differ between the separate figures. Likewise, the same taxa in different iTOL plots can be ordered differently, requiring tedious inspection to assess similarity of the samples. This can make it arduous to compare results of more than one study together.

Results

The new type of figure presented here is a two-dimensional square density plot representing a single bacterial community or a summation of multiple bacterial communities. The color intensity in different areas of the plot represents the relative abundances of the taxa assigned to points in that area. The broader the taxon, the larger and more diffuse the colored spot for that taxon, as shown in Figure 1. For example, the colored spot corresponding to a certain phylum with a relative abundance of 20% on a phylum-level plot would have a larger radius and would be more diffuse than the colored spot corresponding to a certain genus with a relative abundance of 20% on a genus-level plot. This difference in density sharpness is deliberately added depending on the specified taxonomic level of the input data. It reflects the fact that a higher-level taxon encompasses all of the lower-level taxa within it. This means that a higher-level taxon (e.g. a phylum) "occupies" more space on the square than a lower-level taxon (e.g. a genus), and thus it is given a larger and more diffuse spot. In other words, a phylum-level plot is "cloudier" than a genus-level plot because labeling taxa at only the phylum level is less precise than identifying them all the way to the genus level.

The primary strength of this type of plot is that plots created independently from different bacterial communities can be visually compared, which, as described in the introduction, is an attribute that the major current visualization methods lack. This is possible because the layout of bacteria on the plot is fixed and identical for every plot generated. Instead of relying on whichever bacteria are present in the sample being visualized, the figure layout is predetermined by a universal file that contains the names of all possible taxa that can be included and an X and Y coordinate for each name. For each taxon present in a sample, a spot is produced at the coordinates listed in this file, as shown in Figure 2.

The taxa coordinates could be determined a number of ways, with the requirement that more closely related taxa are positioned close together and distantly related taxa are further apart. The present method used the taxonomic classifications, from domain to genus, of all entries in The All-Species Living Tree Project (LTP) version 108 (Yarza et al. 2008). Distances were then calculated between each pair of genera according to how much of their classifications differ. For example, two genera in the same family have a distance of 1, two genera in the same order but different families have a distance of 2, and so on up to the domain level, where the distance between a genus in the Bacteria domain and a genus in the Archaea domain is 6. Note that this method of distance calculation does not take evolutionary divergence into account, only nomenclature, for simplicity of demonstration. Alternative methods are described in the Discussion section.

These distances were then used for non-metric multidimensional scaling (NMDS), which arranges the elements (genera in this case) in a two-dimensional plot that visually approximates the given non-Euclidean distances (Ramette 2007). NMDS was chosen because it ranks the distances between objects, and then uses that ranking information, rather than the distances themselves, to map the objects onto the two-dimensional space. This tends to produce a more even and distributed map of the objects than would other common ordination methods, while still showing which objects (i.e. genera) are more (taxonomically) similar than others. Because the distances were based on taxonomic similarity, similar genera tend to cluster together, and the plot approximates a map of taxonomic diversity. To obtain coordinates for higher-level taxa, the coordinates for all genera within each taxon were averaged. Thus, for example, the coordinate for a phylum is at the "center of mass" of all the genera within that phylum. That way, when a diffuse spot is placed at that phylum's coordinate, it reflects the notion that those specimens could belong to any of the surrounding genera within that phylum. This generic map of all taxa coordinates was created once and saved as the coordinates file, which is then used to create all microbial community plots as described earlier.

Methods

Taxonomic information from the All-Species Living Tree Project (Yarza et al. 2008) was downloaded (available at http://www.arb-silva.de/fileadmin/silva_databases/living_tree/LTP_release_108/LTPs108_SSU.csv) and processed for further use. Taxonomic distances were calculated from this information using a Perl script. Using the "ecodist" R package (Goslee and Urban 2007), these distances were used with non-metric multidimensional scaling to produce coordinates for each genus. Coordinates for higher taxonomic levels were calculated as the arithmetic mean of the coordinates of all genera within each level.

The R programming language and the R packages "MASS" and "ggplot2" were used to produce the density plots. Data from the Human Microbiome Project was downloaded to test the new method (The Human Microbiome Project Consortium 2012). Specifically, the "Phylotype Counts" and "Phylotype Lookup" files produced from the 16S V1-V3 region data, and analyzed with the "mothur" program, were used (available at http://www.hmpdacc.org/HMMCP/). The R language was used to process this data into usable form and perform the various analyses presented in this paper. Additionally, a customized version of the filled.contour() R command was downloaded (QERM).

Discussion:

The general method of assigning bacterial taxa to fixed locations within a plot does not yet seem to have been widely used in microbial ecology. It is important that the figures created using this method accurately show similarities and differences between bacterial community compositions, since that is the primary purpose of these figures. While the plots are intended to be compared qualitatively, a quantitative measurement of difference between two plots was devised to assess the legitimacy of visual similarities and differences. The density plots are drawn from a 100 by 100 square matrix of values corresponding to the density at those coordinates. The distance between two density plots was calculated by first subtracting each element of one density matrix from the corresponding element of the other, producing a new matrix of the same size. Then, the Frobenius norm of this new matrix was calculated, which is analogous to the Euclidean distance in a two- or three-dimensional space, but instead treats each element of the matrix as a separate dimension (Golub and Van Loan 1996). The rationale for choosing this method of distance calculation was simply to try to quantify the visual difference between two density plots.

This "plot distance" was then compared against an "established" distance between the two communities from which the plots were generated. Bray-Curtis dissimilarity was chosen since it is widely used in microbial ecology studies (Goslee and Urban 2007). To measure the correlation between the "density plot distances" described earlier and the Bray-Curtis dissimilarities of the underlying communities, bacterial community data from the Human Microbiome Project was downloaded. This was a major project that was presented in 2012, and the data has been released to the public to aid in future studies such as the present one. One hundred samples were randomly selected from the genus-level mothur Community Profiling data set (also known as HMMCP), and density plots were produced for each using the method created in the current study. Both Bray-Curtis dissimilarity and "density plot distance" were calculated between every pairwise combination of these samples, totaling 4,950 comparisons. Figure 3 shows that there is a moderate correlation between these two measurements, indicating that the density plot method presented here does indeed portray true ecological similarities and differences between microbial communities. One possible reason for the correlation not being stronger is that Bray-Curtis dissimilarity does not take into account relatedness of different taxa, whereas the density plots do take taxonomic "relatedness" into account.

As mentioned in the Results, the genera distances used for non-metric multidimensional scaling were calculated from taxonomic differences. One advantage of this measure of distance is that genera tend to be grouped by shared domain, class, order, and then family levels. This means that when a class-level plot, for example, contains a diffuse spot at a certain location, the genera whose coordinates would be in that area in a genus-level plot tend to belong to that class. This distance measure also has the advantage over phylogenetic (i.e. evolutionary) distance in that a single genus can contain many strains that have diverged different amounts from those in another genera. It would be unclear how a single distance would be calculated between those two genera. Nevertheless, other measures of distance could theoretically be used to create these types of density plots. In addition, other methods of ordination besides NMDS could be used, depending on the desired spacing and organization of the bacteria.

In this regard, the specific method described in this paper is a proof of concept, demonstrating that this general procedure produces figures that can be created independently, yet are still useful for visual comparison. Several steps must be taken for this method to be useful to most microbial ecologists. A system for creating and updating a universal bacterial “map” must be devised so that figures produced by different researchers can be compared. Because not all of these researchers are likely to be proficient in the R programming language, a standalone program could be created that is more user-friendly. Better still, these density plots could be automatically generated by existing analysis tools such as MG-RAST (Meyer et al. 2008), further increasing their accessibility.

Finally, while this method was devised with microbial ecology in mind, it could be used for different realms of ecology, such as plants and animals, or could even cover the entire diversity of known living organisms in a single figure. On the other hand, the scope of the figure could be narrowed to a single genus or species that contains many strains. Such figures would be especially useful for a community of researchers studying a specific organism. In this case, phylogenetic distance would likely be used in place of taxonomic distance.

Conclusions

Metagenomic sample visualization is important for comparing different samples, both within and among studies. Different tools use a variety of methods to display the taxonomic content of these samples, and each method has different strengths. However, current methods are particularly weak in their ability to compare figures that were created independently. A visualization method with more defined layout has been developed for this purpose, in which a taxonomic map produced by an ordination method already in use in microbial ecology is used to produce two-dimensional density plots. Using existing published microbiome data, this new method was found to accurately portray inherent similarities and differences. This method can be implemented and modified to suit the needs of microbial research, and can be extended to be of use in other scientific areas. This method of visualization will help researchers to compare each other's data, improving the synergy of the microbial ecology community.

References

Golub, Gene H., and Charles F. Van Loan. 1996. *Matrix Computations.* Baltimore: Johns Hopkins Univ. Press.

Goslee, Sarah C., and Dean L. Urban. 2007. "The Ecodist Package for Dissimilarity-based Analysis of Ecological Data." *Journal of Statistical Software* 22 (7):1-19.

Hawlena, Hadas, Evelyn Rynkiewicz, Evelyn Toh, Andrew Alfred, Lance A. Durden, Michael W. Hastriter, David E. Nelson, Ruichen Rong, Daniel Munro, Qunfeng Dong, Clay Fuqua, and Keith Clay. 2012. "The Arthropod, but Not the Vertebrate Host or its Environment, Dictates Bacterial Community Composition of Fleas and Ticks." *ISME Journal* 7:221-223. doi: 10.1038/ismej.2012.71.

The Human Microbiome Project Consortium. 2012. "Structure, Function and Diversity of the Healthy Human Microbiome." *Nature* 486:207-214. doi: 10.1038/nature11234.

Huson, Daniel H., Alexander F. Auch, Ji Qi, and Stephan C. Schuster. 2007. "MEGAN Analysis of Metagenomic Data." *Genome Research* 17:377-386. doi: 10.1101/gr.5969107.

Keuhn, Joanna S., Patrick J. Gorden, Daniel Munro, Ruichen Rong, Qunfeng Dong, Paul J. Plummer, Chong Wang, and Gregory J. Phillips. 2013. Bacterial Community Profiling of Milk Samples as a Means to Understand Culture-Negative Bovine Clinical Mastitis." *PLOS ONE* 8(4): e61959. doi: 10.1371/journal.pone.0061959.

Letunic, Ivica, and Peer Bork. 2007. "Interactive Tree Of Life (iTOL): an Online Tool for Phylogenetic Tree Display and Annotation." *Bioinformatics* 23 (1):127-128. doi: 10.1093/bioinformatics/btl529.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian et al. 2008. "The Metagenomics RAST Server – a Public Resource for the Automatic Phylogenetic and Functional Analysis of Metagenomes." *BMC Bioinformatics* 9:386. doi: 10.1186/1471-2105-9-386.

Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. "Interactive Metagenomic Visualization in a Web Browser." *BMC Bioinformatics* 12:385. doi: 10.1186/1471-2105-12-385.

Quantitative Ecology and Resource Management Program at University of Washington. "Contour Plots." Accessed February 1, 2013. http://wiki.cbr.washington.edu/qerm/index.php/R/Contour_Plots.

Ramette, Alban. 2007. "Multivariate Analyses in Microbial Ecology." *FEMS Microbiology Ecology* 62 (2):142-160. doi: 10.1111/j.1574-6941.2007.00375.x.

Revanna, Kashi V., Daniel Munro, Alvin Gao, Chi-Chen Chiu, Anil Pathak, and Qunfeng Dong. 2012. "A Web-based multi-Genome Synteny Viewer for Customized Data." *BMC Bioinformatics* 13:190. doi: 10.1186/1471-2105-13-190.

Riesenfeld, Christian S., Patrick D. Schloss, and Jo Handelsman. 2004. "METAGENOMICS: Genomic Analysis of Microbial Communities." *Annual Review of Genetics* 38:525-552. doi: 10.1146/annurev.genet.38.072902.091216.

Yarza, Pablo, Michael Richter, Jörg Peplies, Jean Euzeby, Rudolf Amann, Karl-Heinz Schleifer, Wolfgang Ludwig, Frank Oliver Glöckner, and Ramon Rosselló-Móra. 2008. "The All-Species Living Tree Project: A 16S rRNA-Based Phylogenetic Tree of all Sequenced Type Strains." *Systematic and Applied Microbiology* 31 (4): 241-250. doi: 10.1016/j.syapm.2008.07.001.

Zhou, Mi, Ruichen Rong, Daniel Munro, Chunxia Zhu, Xiang Gao, Qi Zhang, and Qunfeng Dong. 2013. "Investigation of the Effect of Type 2 Diabetes Mellitus on Subgingival

Plaque Microbiota by High-Throughput 16S rDNA Pyrosequencing.” *PLOS ONE* 8(4):e61516. doi: 10.1371/journal.pone.0061516.

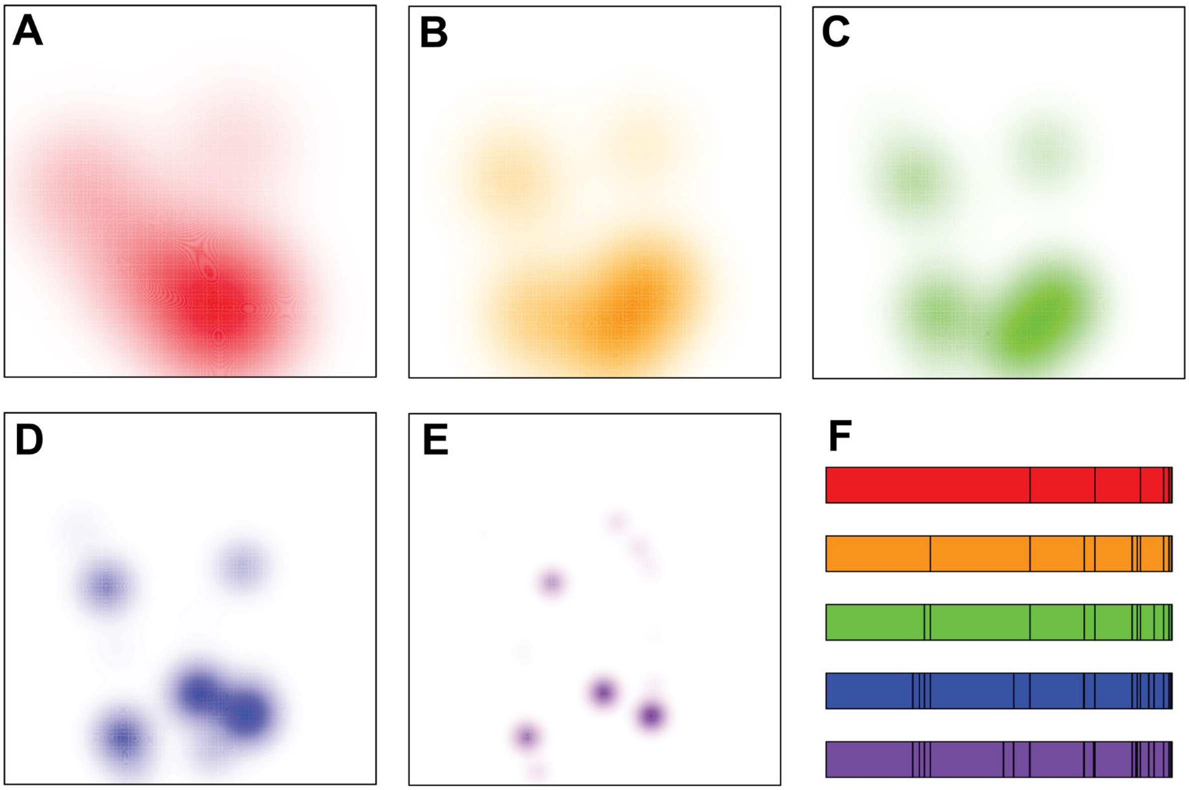
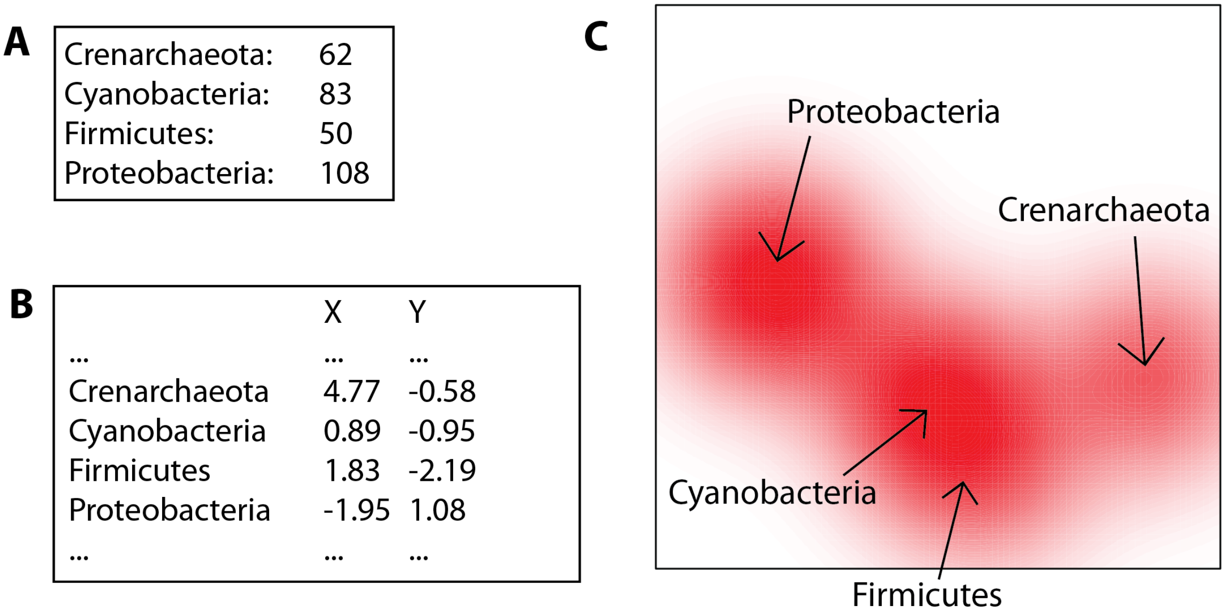


Figure 1 - Plotting at Different Taxonomic Levels. (A-E) The organisms in a single microbial community were grouped by phylum, class, order, family, and genus, respectively, to show how the same sample visualized at different taxonomic resolutions has a similar shape but differs in the precision of the density plot. (F) These five bars from top to bottom represent the relative abundances of each taxon in plots A-E, respectively. They are aligned to show how each higher-level taxon subdivides into smaller ones, which can be faintly observed in the density plots.



A

Crenarchaeota:	62
Cyanobacteria:	83
Firmicutes:	50
Proteobacteria:	108

B

	X	Y
...	...	...
Crenarchaeota	4.77	-0.58
Cyanobacteria	0.89	-0.95
Firmicutes	1.83	-2.19
Proteobacteria	-1.95	1.08
...	...	...

Figure 2 – The Basic Plotting Method. (A) The input file for a plot lists the names and abundances of each taxon in the sample (or combination of samples). (B) A single universal file containing X and Y coordinates for all possible taxa is used to create all figures. (C) The taxa in the input file are plotted onto the figure at the coordinates found in the universal coordinates file, with an intensity corresponding to its abundance in the input file, and with a sharpness corresponding to the taxonomic level of the taxa in the input file. In this example, the spots produced for each taxon in the input file are pointed out.

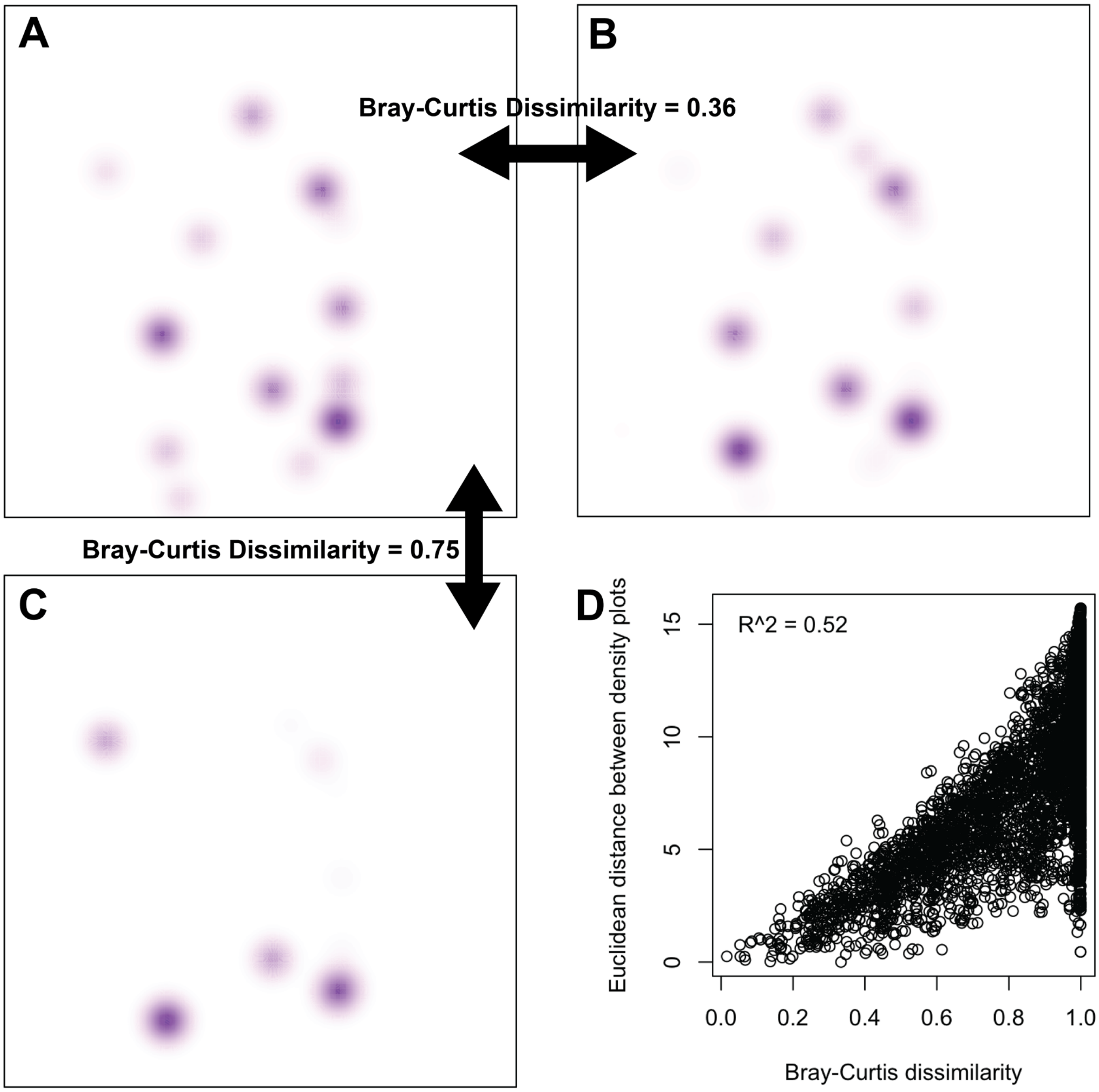


Figure 3 – Validating Visual Differences. (A) An example plot is shown for a genus-level sample. (B) Another sample, which was relatively similar according to the Bray-Curtis dissimilarity of the two samples, also produced a similar-looking density plot. (C) A third sample, which was relatively different from the first sample according to Bray-Curtis dissimilarity, has a different-looking density plot. (D) The quantitative difference between two density plots was compared to the Bray-Curtis dissimilarity of the two underlying samples using all 4,950 pairwise comparisons among 100 random samples. A coefficient of determination of 0.52 indicates a moderate correlation.