Journal
doi.XX.XXXX/Journal/xxxxxx
Advance Access Publication Date: 23 March 2021
Bioinformatics of Microbes and Microbiomes

Microbiome Visualization

Microbiome Maps: Hilbert Curve Visualizations of Metagenomic Profiles

Camilo Valdes [1,2], Vitalii Stebliankin [1], Daniel Ruiz-Perez [1], Ji In Park [3], Hajeong Lee [3], and Giri Narasimhan [1,4,†]

[1]Department of Food Science and Technology, University of Nebraska-Lincoln, 1901 N 21 ST, Lincoln, NE 68588, USA.
[2]Bioinformatics Research Group (BioRG), Florida International University, 11200 SW 8th Street, Miami, FL 33199, USA.
[3]Department of Medicine. Kangwon National University School of Medicine. Chuncheon, Gangwon-do, South Korea.
[4]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, South Korea.
[5]Biomolecular Sciences Institute, Florida International University, 11200 SW 8th Street, Miami, FL 33199, USA.

[†]To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Preprint. The work reported here is ongoing. This manuscript will be updated.

Abstract

Motivation: Abundance profiles from metagenomic sequencing data synthesize information from billions of sequenced reads coming from thousands of microbial genomes. Analyzing and understanding these profiles can be a challenge since the data they represent is complex. Particularly challenging is their visualization, as existing techniques are inadequate when the taxa number in the thousands. We present a technique for succinct visualization of abundance profiles using a space-filling curve that transforms a profile into an interpretable 2D image.
Results: Jasper is a tool for visualizing profiles from metagenomic whole-genome sequencing and 16S, and orders taxa along a space-filling Hilbert curve. The result is a "Microbiome Map", where each position in the image represents the abundance of a single taxon from a reference collection. Jasper can order the taxa in one of two ways, and depending on the ordering, the *microbiome maps* can highlight "hot spots" of microbes that are either dominant in taxonomic clades or to the biological conditions under study.
We use Jasper to visualize samples from the Human Microbiome Project and from a Chronic Kidney Disease study, and discuss a variety of ways in which the microbiome maps can be an invaluable tool to visualize spatial, temporal, disease, and differential profiles. Our approach can create detailed microbiome maps involving hundreds of thousands of microbial reference genomes with the potential to unravel latent relationships (taxonomic, spatio-temporal, functional, and other) that could remain hidden using traditional visualization techniques. The maps can be converted into animated movies that bring to life the dynamicity of microbiomes.
Availability: Jasper will be available as free software from the Mac App Store and biorg.cs.fiu.edu/jasper
Contact: giri@cs.fiu.edu
Supplementary information: Supplementary materials are available at biorg.cs.fiu.edu/jasper

1 Introduction and Background

Microbiome samples are now routinely created by means of low-cost, high-throughput metagenomics DNA sequencing. The next steps in the analysis is the creation of microbial community abundance profiles [14], obtained by mapping the sequenced reads against a collection of microbial genomes from a reference genome collection like Ensembl [1] or RefSeq [31]. Tools such as Flint [36] and Kraken 2 [40] facilitate the process to create detailed microbial abundance profiles for tens of thousands of genomes for both metagenomic whole-genome DNA sequencing (mWGS) as well as 16S-amplicon sequencing (16S). The abundance profiles are the stepping stones for

© The Author 2021.

downstream analyses such as differential abundance studies [38], co-occurrence pattern discovery [20, 22, 37, 21], Bayesian analyses [19, 34], biomarker identification [35], multi-omics analyses [8, 9], and analyses of profiles from longitudinal studies [28, 33]. Low-cost sequencing has meant that an increasing number of metagenomic data sets are being generated [29].

Metagenomic profiles contain relative abundance values for the entire collection of microbial taxa present in a sample, and these profiles can be easily viewed with stacked bar charts or pie charts using data analysis software suites such as Tableau [7], or MS Excel [5]. Such tools are readily available to the public and allow for data exploration, but are designed for the analysis of generic tabular data, and do not consider domain-specific information (taxonomic, phylogenetic, etc.) that may be crucial for the interpretation of metagenomics datasets. Recent tools such as WHAM! [18] allow for explicit metagenomics-focused analyses, making it possible to dig down into the data and create useful visualizations for descriptive analyses.

We argue that complex latent properties of microbiomes embedded in community abundance profiles such as taxonomic hierarchies and other relationships are not easily described and visualized in traditional generic plotting mechanisms such as stacked bar-charts or line-plots. The problem gets more acute as the sizes of our reference genome collections continues to grow exponentially over time [30].

In this work we consider the problem of succinctly visualizing a microbiome, and specifically, visualizing metagenomic community abundance profiles, along with their latent structured labels (e.g., taxonomic description or biological property) with the use of a visualization technique called the *Hilbert Curve Visualization* (HCV). We discuss the challenges of scalable visualization of billions of measurements for hundreds of thousands of microbial genomes, and argue that alternative visualization techniques are useful when trying to combine many factors of metagenomic information in order to create interpretable images that can lead to improved understanding.

2 Approach

As mentioned earlier, we use a technique called the Hilbert curve visualization (HCV) to succinctly visualize the microbial community abundance profiles of a large number of genomes (up to 44K). These profiles contain the relative abundance measurements of thousands of genomes, and they are ordered along a space-filling curve in a 2D square using the Hilbert curve [25], thus making it possible to visualize the profile of a single metagenomic sample. In the resulting Hilbert image, each position is a genome from the reference database, and the intensity of the position's color value represents the abundance of a genome in the sample.

As discussed below, depending on the ordering of the genomes that is selected, different "microbial neighborhoods" are created, allowing for different interpretations of the clusters of bright segments, i.e., "hotspots", of abundant genomes in the images. Fixing the position of a genome results in visualizations that allow for quick comparisons of the abundance of the same genome or sets of genomes in multiple microbiome samples.

2.1 Space-Filling Curves

Space-filling curves are popular in scientific computing applications for their ability to speed-up computations, optimize complex data structures, and simplify algorithms [13]. Trees are particularly interesting structures that can be optimized with space-filling curves because it is possible to generate sequential orderings of the nodes of the tree in which parent and children nodes are neighbors in a 2D plane. The combination of trees and space-filling curves has been shown to be useful in many fields [12], and in metagenomics this can be useful because the microbial genomes in a reference database are classified using a taxonomy tree with a hierarchy of levels (Strain, Species, Genus, etc.). For data from mWGS experiments, we can presume the leaf nodes of the taxonomy tree to be microbial strains (Figure 2, panel (B)); for 16S data, the leaf nodes are usually species or genera. Clades of the taxonomy tree correspond to microbial neighborhoods in the resulting visualization.

2.2 The Hilbert Curve

The Hilbert Curve is one of the more prominent examples of space-filling curves, and its construction is based on a recursive partitioning of a square region in the 2D plane and connecting the centers of these squares in a specific order. To provide a recursive definition of the curve, Figure 1 (A) shows the curve at Level 1 when there are only four squares to connect. The curve at Level k is defined recursively by dividing the original square into four squares, each with a Level $k-1$ curve in it and then connecting these pieces using the template of the Level 1 curve after appropriate rotation of the four curves.

Many applications exploit the order that space-filling curves impose on data, and a particular application has been the visualization of high-dimensional data. The first use of the Hilbert Curve as a visualization tool was proposed by Keim in 1996 [27] to represent stock market data. Since then, it has been used for visualizing genomic data [16, 10] and DNA alignments of whole bacterial genomes [39].

In human genomics, using the HCV is relatively straightforward as the natural linear order of genomic positions on a single chromosome can be mapped on to the Hilbert curve. Mapping the entire human genome involves stitching together the chromosomes prior to mapping. Tools such as HilbertVis [10] and HilbertCurve [24] apply HCV to visualize a single human genome. To date, the HCV technique has not been applied to metagenomics datasets.

2.2.1 Metagenomics

Visualizing microbial community profiles is non-trivial as there exists no natural linear ordering for the thousands of genomes in the reference database, and we argue that in trying to order thousands of genomes the answer is not a single "true" order, but rather, multiple different orderings that provide different perspectives of the metagenomics data being visualized.

Custom orderings based on a single sample (say, ordered by decreasing abundance values in a single sample) are not that informative when using a HCV, since the main advantage of these visualizations is in the ability to locate the same genome in the same position on the images for multiple related samples. Even using the average abundance value from a cohort is not so useful because different cohorts cannot be readily visually compared.

Thus, the value of our visualizations lies in specifying useful orderings that are not dependent on a single attribute of one, or multiple samples, but rather, on some global properties and biological interpretations that are useful for the researcher.

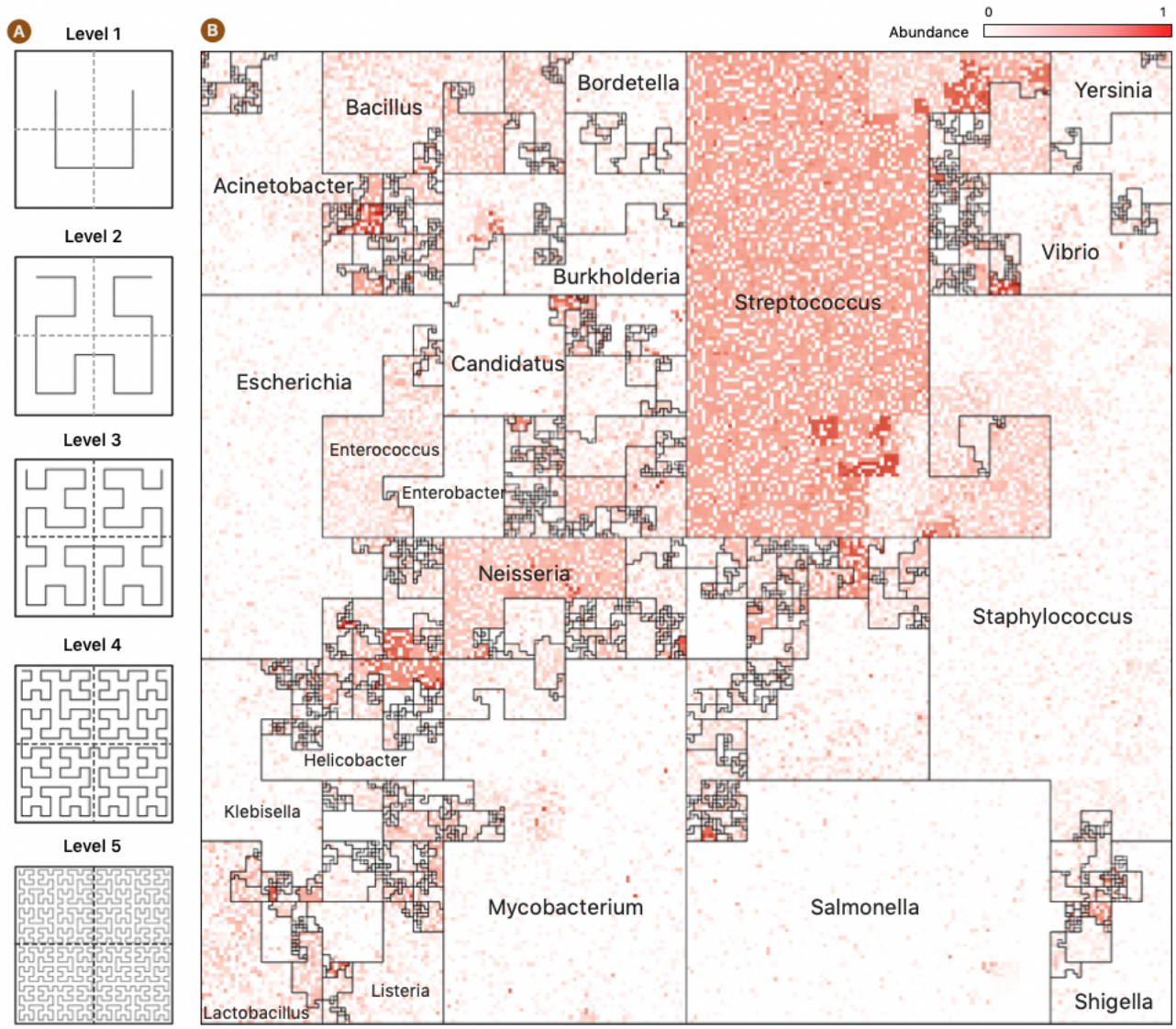

Fig. 1. Hilbert Curve Visualization of Metagenomic Samples. (A) The first five iterations of the Hilbert curve: the Level 1 curve is obtained by connecting the centers of the four initial squares as shown; the Level k curve is obtained by a recursive partitioning of each square from Level $k-1$, creating four Level $k-1$ curves and connecting them as outlined by the Level 1 curve, rotated appropriately. At level k, the original square is divided into $2^k \times 2^k$ small squares, each of whose centers is visited by the Level k Hilbert curve. **(B)** A representative image of a mWGS Buccal Mucosa sample from the Human Microbiome Project (HMP) created using a "taxonomic ordering" of 44K reference genomes from the Ensembl database. The intensity of each position in the image represents the abundance of one microbial genome. Groups of segments are labeled by the groups induced by the ordering of the taxa on the Hilbert curve. Additional Hilbert curve images for more mWGS samples, as well as other 16S samples at full resolution, are available via links from biorg.cs.fiu.edu/jasper

2.2.2 Linear Orderings

Below, we discuss two classes of orderings that are shown to be useful for the visualization.

- **Taxonomic Ordering:** Ordering of reference genomes as per the taxonomic tree from Ensembl Genomes [2].
- **Labeled Ordering:** Oredering as per a user-supplied labeling scheme.

The resulting metagenomic Hilbert images using the orderings from Section 2.2.2 are expressive in that researchers can use them to quickly identify what taxonomic groups are abundant without losing any context imposed by the ordering (i.e., taxonomic hierarchies or biological conditions). Thus, the technique provides a framework to creatively explore the space of all possible orderings.

Metagenomic Hilbert images offer a quick visual way to identify "hotspots" in "microbial neighborhoods", i.e., groups of related microbes, that can contextualize the important features of the metagenomic samples under study. Moreover, by comparing two or much such images one can quickly identify major changes in the characteristics of these neighborhoods.

3 Methods

Visualizing a microbiome's abundance profile starts by aligning, or classifying, sequencing reads against a reference collection

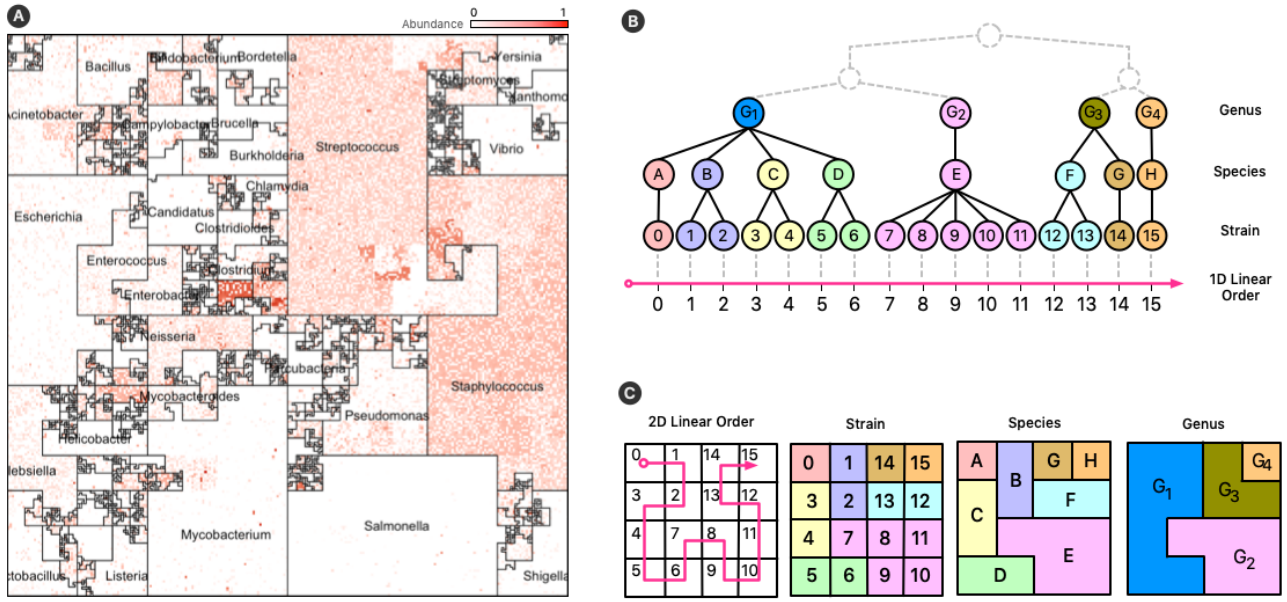

Fig. 2. Taxonomic Ordering. (A) A *Taxonomic Ordering* of 44,048 microbial reference genomes from Ensembl Bacteria. *Microbial neighborhoods* are drawn based on a taxonomic tree for microbial classification. The image depicts the distribution of Genera in the reference genome database, and the size of each neighborhood is consistent with the number of genomes that belong to it. **(B)** A taxonomic hierarchical tree with three (3) levels of rankings for the genomes in the reference database: Genus, Species, and Strain. The taxonomic tree is linearized to create a 1D linear order for the tree's leaves (Strains). **(C)** The 1D linear order is then laid out onto a 2D plane using a Hilbert curve, which creates *microbial neighborhoods* of related taxa.

of microbial genomes, and creating counts of the number of reads that align (or are classified) to each genome. The Flint [36] software facilitates the profiling of mWGS datasets, while the Kraken 2 software does it for 16S datasets. For the images shown in this paper, Flint uses a reference collection of 44,408 microbial genomes from the Ensembl Bacteria database [1], while Kraken 2 uses a "16S" reference of 5,127 genomes which contains references from Greengenes [17], SILVA [32], and RDP [15]. Once the profiles have been created, we use them as input to the Jasper tool which creates a Hilbert image for each sample.

The input to Jasper can represent a single metagenomic sample, or multiple ones if given in a labeled matrix, and its output will be metagenomic Hilbert images for each sample provided as input. Documentation on the type of profiles, along with matrix formatting guidelines, are all available in the documentation page of the GitHub repository, and detailed parameters, along with sample command calls for both Flint and Kraken 2 are available in the supplementary materials and project website.

Different linear orderings of the taxa on the Hilbert curve result in different images. Users may select from two options: a *taxonomic ordering* (Figure 2) which uses the linear order from Ensembl's taxonomic tree, or a *user-defined labeled ordering* (Figure 3) which determines the order based on a labeling from a biological grouping of samples provided by the user. Both orderings place microbial genomes along the curve in a specific order. Note that unlike other genomic HCV techniques, metagenomic Hilbert images do not depict single genomic positions, but rather, they display thousands of genomes.

3.1 Microbial Neighborhoods

Different linear orderings produce different Hilbert visualizations, with each resulting in clusters of related microbes along neighboring regions in the 2D plane. The clustering creates unique areas that resemble community neighborhoods in popular consumer mapping applications like Google Maps [3], and we term these areas "*Microbial Neighborhoods*" (Figure 2, panel (A)) as they represent microbes belonging to either the same taxonomic group, or the same biological condition — the idea being that they are clustering around a common scheme (taxonomic or biological).

One advantage of the "*Microbial Neighborhoods*", using the "*Taxonomic Ordering*" is that when the genomes are laid out in the Hilbert image without any abundance color mappings, the distribution of the reference genome database can be quickly illustrated (Figure 2, panel (A)). Larger clades with more taxa (leaves) occupy larger tracts of the image. Thus, we can see from Figure 1 (B) and 2 (A) that the Ensembl database contains large number of strains from the *Streptococcus* and *Staphylococcus* genera. Displaying a reference genome database using the Hilbert curve is a way of understanding the microbial diversity of the database.

3.1.1 Taxonomic Neighborhoods

The first option for ordering genomes along the Hilbert curve is the *taxonomic ordering* which determines a 1D linear order based on a genome's taxonomic lineage. In this ordering, pairs of taxa belonging to the same taxonomic group (say the same Genus or Species) are placed close to each other along the curve, and consequently, close to each other in the Hilbert image. This ordering scheme creates "*Taxonomic Neighborhoods*" that envelop related taxa based on their taxonomic lineage, and as

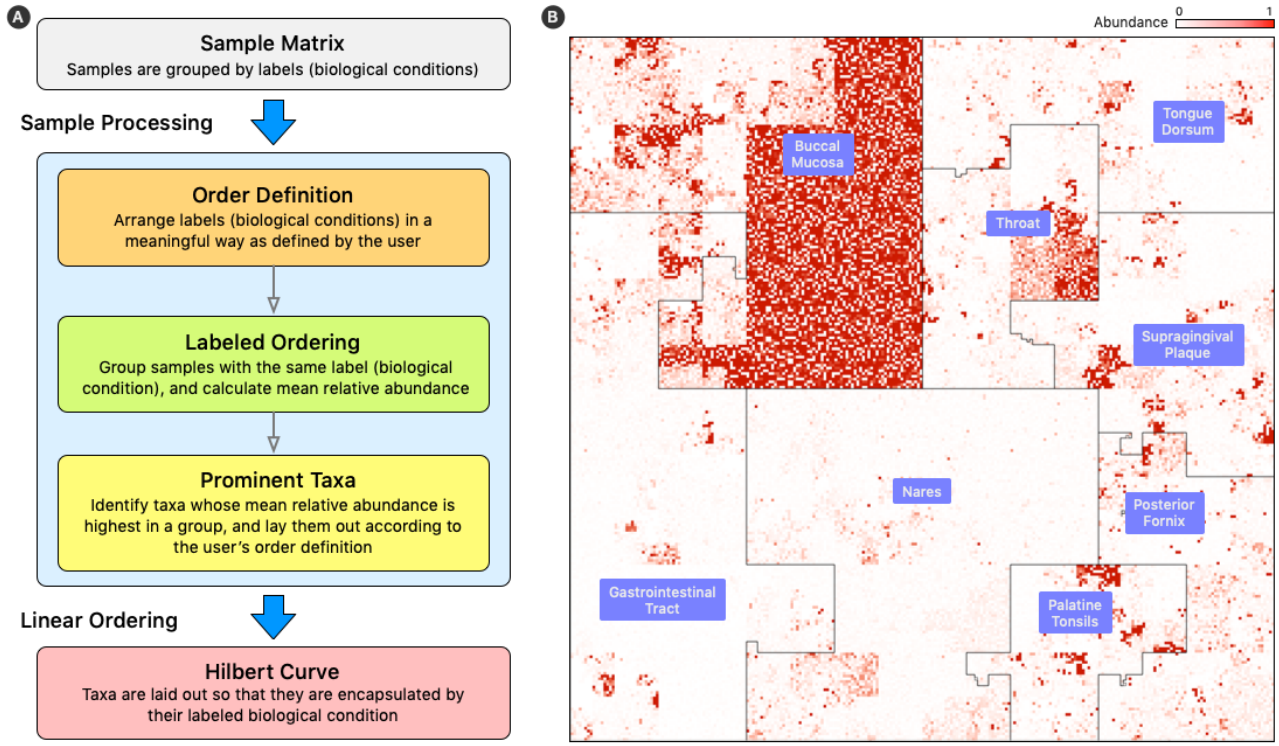

Fig. 3. Labeled Ordering. (A) Samples are processed in a M x N matrix that contains M labeled samples, and N microbial taxa. The user specifies the ordering of the biological conditions that the M samples belong to. For each labeled grouping, the taxa with the highest mean relative abundance is identified and used as an anchor for a linear order in the Hilbert curve.**(B)** Resulting Hilbert curve image for 3 samples for a buccal mucosa condition, built using an averaging ordering scheme from 8 HMP body sites.

seen in Figure 2, multiple taxonomic levels can be displayed at the same time in a single image. The ability of the Hilbert image to display multiple levels of the taxonomic tree all at once, while at the same time providing high-resolution abundance information for single genomes, is a compelling advantage over visualizing data with other means as the sheer number of data points (single genome measurements) would overwhelm any other 1D visualization.

Linearizing Trees

Illustrating a taxonomic tree as a 2D Hilbert curve starts by finding a linear order of the leaf nodes in the tree. Figure 2, panel (B), depicts a fictitious taxonomic tree with 16 leaf nodes (microbial strains in this example) that are ordered along a 1D line using a *taxonomic ordering* scheme (section 3.1.1) which groups the 16 strains according to their parent species and genus groups. Figure 2, Panel (C), illustrates how the 16 strains would be laid out on a 2D plane, and how the taxonomic hierarchies are represented as strain, species, and genus areas in the Hilbert image. Note that establishing a linear order for a tree structure is non-trivial as trees do not have a "start" or "finish", nor do they have a "right" side or a "left" side. Different orderings result by performing a permutation of the children at any given node of the tree.

The *taxonomic ordering* linearizes a tree by using taxonomic data taken from Ensembl's Pan-taxonomic Compara [6] and Ensembl Genomes [2] databases, and we use Ensembl's taxonomic information as the basis for building a taxonomic tree structure that is linearized and used as the foundation for the Hilbert image (see supplementary materials for details). The genomes in the reference database are annotated so that we can establish a lineage up to the phylum level, and in the case when we profile mWGS data, the leafs of the tree are taken to be microbial strains; in the case when we profile 16S data, then the leafs of the tree are taken to be microbial species. For both mWGS data and 16S data, the 2D square that bounds the Hilbert image represents all the genus-level groups in our taxonomic tree; we do this because we found that drawing the curve at the genus level amounts to a good compromise between information and visual appeal.

3.1.2 Condition Neighborhoods

The second option for ordering genomes along the Hilbert curve is the *labeled ordering* (Figure 3, which creates "*Condition Neighborhoods*" by using an ordering scheme that determines the 1D linear order based on a user-supplied labeling of samples provided as a labeled $m \times n$ sample matrix M, where m are sample rows, and n are the genomes in the reference database; if we have i samples, and j reference genomes, then the cells in the matrix correspond to abundance values for genome n_j in sample m_i.

Establishing the 1D linear order for multiple biological conditions (Figure 3, panel (A)) starts by ordering the number of conditions k in some meaningful order, $C_1, C_2, ..., C_k$, provided by the user. Note that the conditions may represent different disease states (if we are comparing disease samples),

time intervals (if we are comparing a time series), or locations (if we are comparing body sites, or environmental sites); the conditions are not limited to the aforementioned list, as users can supply their own. Once we have a condition ordering established, the next task is to identify taxa whose average relative abundance is highest in C_1 and order them first, followed by taxa whose average relative abundance is highest in C_2, and so on, until we terminate the ordering by taxa that are not abundant in any of the conditions. Once we have established the ordering, we can then draw the Hilbert images for each of the samples from the input sample matrix M.

Although one can argue that every taxon must be highest in one of the k conditions, this is not meaningful unless its presence is above the threshold of noise, which we determine when we normalize the input matrix M (details available in the supplementary materials). In general, if a taxon is most abundant in multiple conditions (something that we have not seen in practice), then we assign it to the first condition as determined by the input ordering criteria. After the conditions have been organized along the curve, taxonomic information is used to order genomes within the range of the condition.

Note that in this ordering, the Hilbert image is still visualizing one sample, but it is displaying the abundance characteristics of the biological condition for which the sample is most prevalent in — "*hotspots*" will therefore appear in one of the conditions, and users can readily tell what condition the sample belongs by identifying the area, i.e., neighborhood, that represents the condition in the image. Clusters of bright positions will also appear in other neighborhoods (Figure 4), as other conditions will contain taxa with high relative abundances, but not in the same quantities as for the condition that the sample belongs to.

4 Results and Discussion

We created microbiome maps for two groups of metagenomic datasets: 24 mWGS normal samples taken from the Human Microbiome Project (HMP) [26], and 18 fecal samples (16S) from a collaboration with Kangwon National University and Seoul National University in Korea. The 24 samples from HMP represent 8 different body sites, and the 18 samples from the Korea study represent 5 stages of Chronic Kidney Disease (CKD), along with a normal control set. We analyzed the mWGS HMP samples with the FLINT software [36], and the 16S CKD samples with Kraken 2 [40] (details available in the supplementary materials). For the HMP samples, the metagenomic profiles contained relative abundance measurements for 44,048 microbial strains, and for the CKD samples, the metagenomic profiles contained relative abundance measurements for 5,127 microbial species.

Three samples were selected for our study from HMP from each of the eight following body sites: Buccal Mucosa, Gastro-Intestinal Tract, Nares, Palatine Tonsils, Posterior Fornix, Supragingival Plaque, Throat, and Tongue Dorsum.

Eighteen fecal samples were obtained from CKD patients of Kangwon and Seoul National University Hospitals. The samples were selected based on their glomerular filtration rate (see Kidney Disease Improving Global Outcomes (KDIGO) [4]), and a total of six groups were created: Control, CKD Stage 1 (CKD 1), CKD Stage 2 (CKD 2), CKD Stage 3 (CKD 3), CKD Stage 4 & 5 non-dialysis dependent (CKD 4-5ND), and CKD Stage 5 dialysis dependent (CKD 5). The CKD stages were determined based on the worsening function of the kidney patients, and three samples from each group were used.

4.1 Comparison to Other Methods

The supplementary materials contain visualizations of the HMP and CKD abundance profiles using traditional visualization techniques for tabular data (Supplementary Figures 1-4). The WHAM! and iMAP suite of tools for exploring the profiles create standard visualizations including stacked-bar plots, cluster heatmaps, etc., which are helpful for condensing the summarized information for multiple samples. However, they lack the ability to convey nuanced information on about 44,000 bacterial strains in a single sample while retaining the perspective of the latent metagenomic relationships (common and/or unique taxa). The Hilbert curve visualizations are easy to interpret: with a quick glance, one can capture the prominent taxonomic groups and even capture groups of co-occurring taxa. With a shift in perspective, the visualization can help to identify the body site (for HMP) or disease condition (CKD) that characterizes the samples.

4.2 Metagenomic Visualizations

The community abundance profiles from FLINT and Kraken 2 for the HMP and CKD datasets were converted into Hilbert curve visualizations: for both datasets we created a set of images using a *taxonomic ordering*, and another set using a *labeled ordering*. Note that the images with the *taxonomic ordering* contain taxonomic clade border lines at the Genus level, because we found this level of resolution to be a good trade-off between image interpretability and taxonomic lineages. Going to a higher level would have resulted in images with vast neighborhoods, and going a level deeper would have resulted in images that contained too many borders. For images done with the *labeled ordering*, the clade border lines are drawn so as to fence off the different labels specified by the user.

Metagenomic Hilbert images for mWGS and 16S data communicate abundance information at different levels of a genome's lineage: for mWGS samples, each position in the image displays information about microbial strains (the resolution at which abundances are reported by FLINT [36]); for 16S samples, each position in the image displays information about microbial species (the resolution at which abundances are reported by Kraken 2 [40].

Figure 2, panel (A), contains a representative image from one sample of the HMP dataset (Nares) ordered using the *taxonomic ordering* scheme. In this image we can clearly see that the *Streptococcus* and *Staphylococcus* groups are abundant in the Nares sample. While the dominant group would have been obvious even in a traditional 1D plot, the Hilbert curve visualization ensures that the smaller taxonomic groups are not overshadowed by the more abundant groups. Identifying the most abundant taxonomic clades in a sample only takes a quick glance at the image.

Figure 3, panel (B), contains another representative image from a Buccal Mucosa sample from the set of 18 HMP samples. This image also contains the same 44K reference genomes from 2, panel (A), but they have been ordered using the *labeled ordering* scheme. Using this scheme, we have ordered the reference genomes based on the highest mean relative abundance of a genome in its labeled cohort, and the resulting

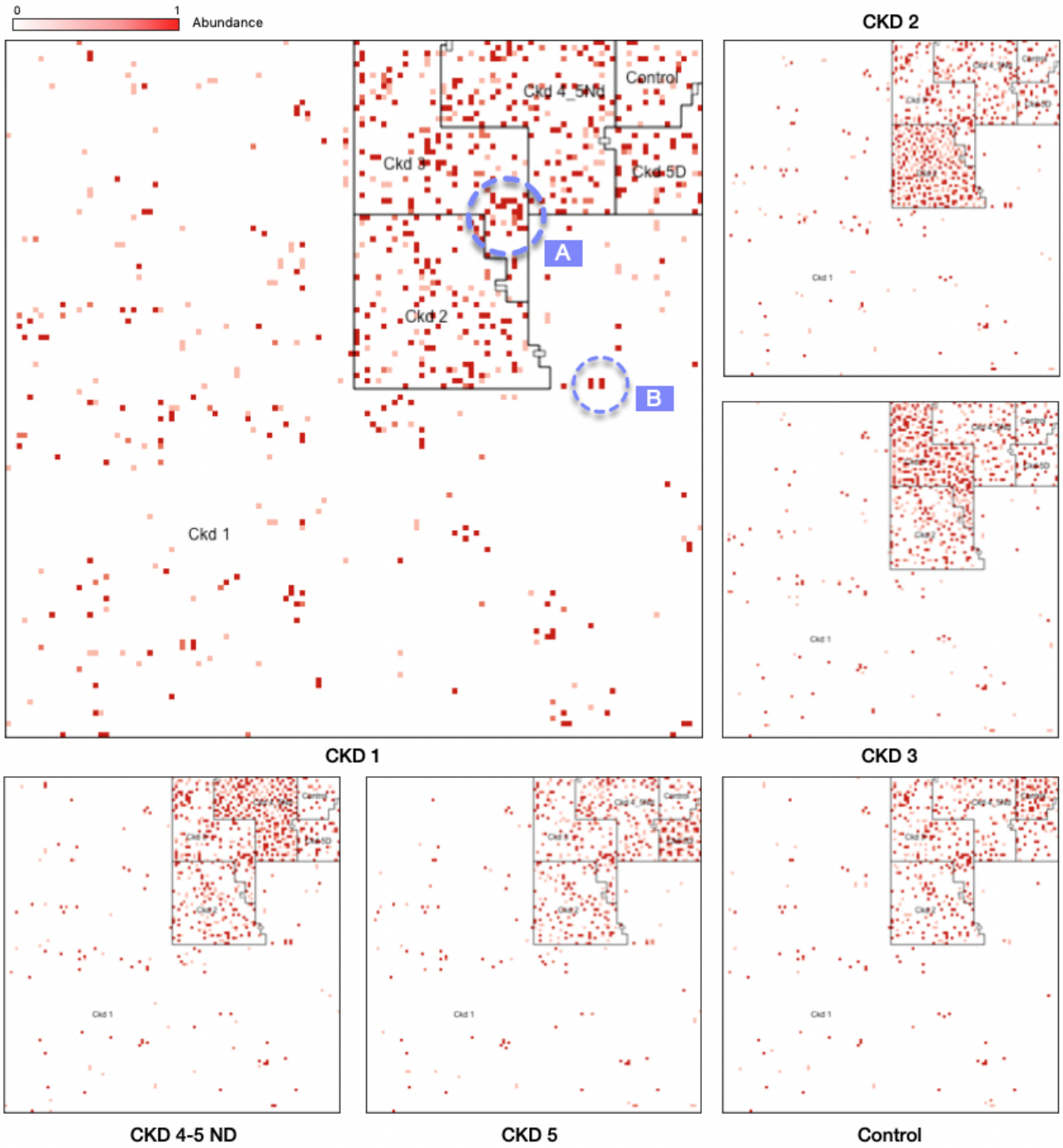

Fig. 4. Chronic Kidney Disease Stages. Hilbert curve visualizations for 16S samples of five stages of chronic kidney disease, along with a control sample. Each panel represents the mean relative abundance of 3 samples for each stage, and displays 5,127 species. Animated figures showing the progression of the stages can be found at the project's repository via links from biorg.cs.fiu.edu/jasper

image can then display the profile of any sample (or the cohort's average) using the computed order. The advantage of this ordering scheme is that identifying the biological condition that the sample belongs to is effortless: one need only look at the neighborhood that contains the most *hotspots* (Buccal Mucosa in this case). Additionally, one can readily see that the buccal mucosa site shares some microbes with those typically abundant in the throat, palatine tonsils, posterior fornix, supragingival plaque, tongue dorsum, and the GI tract.

Figure 4 comprises 16S samples from the CKD analysis arranged using the *labeled ordering* scheme. Here, we observe 5 stages of CKD, and the control, and are displaying the abundances for 5,127 species. Just as we did with the ordering in Figure 3, we ordered each of the species in the profiles based on mean relative abundance of each CKD stage: the

most prominent taxa in CKD Stage 1 are surrounded by the labeled border of the CKD 1 area, the most prominent taxa in CKD Stage 2 are surrounded by the border of the CKD 2 area, and so on. Each image on the panel corresponds to the highest mean relative abundance in the samples from each of the CKD conditions. It is not surprising that the respective regions are more lit up when the samples were from the appropriate cohort. Thus, we can readily identify the stage of each sample by looking at the density of *hotspots* in each labeled area.

What is more significant is the way these plots show the microbes shared by different stages of the disease. For example, region marked A in the figure shows a group of microbes that appear in all stages, while the region marked B appears in almost all stages except CKD3 (absent) and Control (lowered abundance).

A significant point to note is that by fixing the orderings of the taxa, Jasper's visualizations can be used to effectively present groups of metagenomic samples that can be partitioned temporally (longitudinal studies), spatially (body sites or environmental sites), by disease types or subtypes (e.g., ulcerative colitis vs Crohn's disease), by disease stages (as in CKD), and by developmental stages (infant gut at different stages of development). Additionally, it is readily possible to create *average* microbiome maps, *aggregate* maps, and *differential* maps showing either average, aggregate, or differential abundances, respectively. Finally, as suggested below, animations can help enhance the visual appeal for some of these groups of samples.

Animated Movies

Jasper produces a single image for each sample it is given as input, using either a *taxonomic ordering* or a *labeled ordering*. Images can also be used as a single frame of animations that show abundance "*hotspots*", and their evolution across samples or biological conditions. Figure 4 displays an example of how the microbiome "moves" throughout the CKD conditions as the disease stage progresses. Figure 5 contains image frames from a study by [23], and shows how the microbiome of a single patient behaves over the course of 2 days. One can see how microbes are affected by the antibiotic that was administered (Vancomycin & Ticarcillin-Clavulanate) on day 35, and how clades of microbes recuperate later (days 38 - 64). Full resolution images and movies are available on the project's website.

A Visual Inspection Tool

Jasper also offers the *Visual Inspector*, a tool for interactively inspecting and exploring the microbiome maps. The inspector can load an image along with a set of annotations, and users can mouse over and click on a location of the loaded microbiome map to get information about the microbial genome represented at that location. Details on the inspector can be found at the project's website and supplementary materials. An obvious enhancement to the visual inspector tool is the ability to zoom into specific regions of the microbiome map, and/or to select a region of the image for inspection and further exploration.

5 Conclusion

In this work we have shown how the Hilbert curve visualization technique can be used to visualize metagenomic community abundance profiles from both mWGS and 16S DNA sequencing datasets. The resulting microbiome maps display the relative abundance of microbial genomes in an interpretable manner, and can convey information about multiple latent factors of the reference genomes in the samples under study.

The Hilbert curve is used to lay out the microbes from the reference database in two different ordering schemes that can be used to draw a metagenomic Hilbert image: the first, the *taxonomic ordering* relies on taxonomy information from the Ensembl Genomes database, and can be used to create images that express abundance values in the context of the taxonomic clades that the microbial genomes belong to. The second, the *labeled ordering* is dependent on a user-specified labeling of biological conditions for each sample, and can express the abundance values of the profile in the context of a biological interpretation for a set of samples. Although the aforementioned two orders are the first ones to be available in the first Jasper release, we are exploring other classes of orderings that will be incorporated in future releases such as orderings specific to time-series analyses, or multi-omics datasets.

Jasper will be available for free from the Mac App Store [11]. The software requires macOS 10.15, "Catalina", (or greater) and is developed using the Swift programing language and Cocoa APIs.

Acknowledgments

The authors would like to thank the members of the Bioinformatics Research Group, BioRG, at Florida International University for their valuable feedback and comments on the project and the manuscript.

Funding

This work was supported in part by Amazon's "AWS Cloud Credits for Research" program awarded to CV, as well as Florida International University Dissertation Year Fellowships to CV and DR-P. CV is also funded by a University of Nebraska Program of Excellence award, and the University of Nebraska-Lincoln Quantitative Life Sciences Initiative. The work of JIP was done when visiting FIU.

References

[1]Ensembl Bacteria. https://bacteria.ensembl.org/index.html. Accessed: 2018-10-15.
[2]Ensembl Genomes. http://ensemblgenomes.org. Accessed: 2018-10-17.
[3]Google maps. https://www.google.com/maps. Accessed: 2018-11-17.
[4]Kidney disease improving global outcomes guidelines. https://kdigo.org/guidelines/. Accessed: 2018-11-17.
[5]Microsoft Excel. https://products.office.com/en-us/excel. Accessed: 2020-01-14.
[6]Pan Taxonomic Compara. http://ensemblgenomes.org/info. Accessed: 2018-10-17.
[7]Tableau. https://www.tableau.com. Accessed: 2020-01-14.
[8]The integrative human microbiome project. *Cell host & microbe*, 16(3):276, 2014.
[9]Vanessa Aguiar-Pulido, Wenrui Huang, Victoria Suarez-Ulloa, Trevor Cickovski, Kalai Mathee, and Giri Narasimhan. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: supplementary issue: bioinformatics methods and applications for big metagenomics data. *Evolutionary Bioinformatics*, 12:EBO–S36436, 2016.

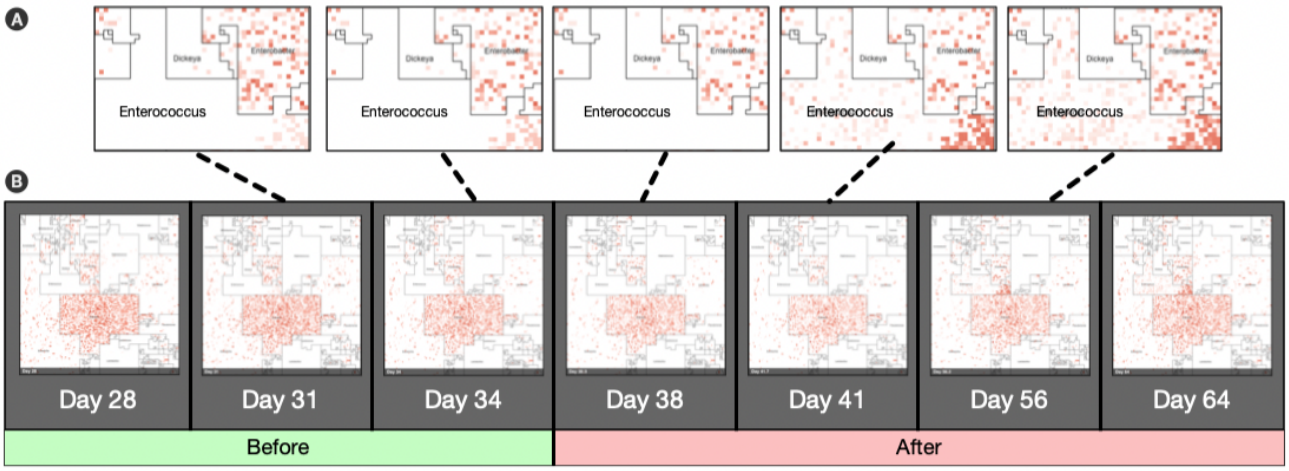

Fig. 5. Animated Microbiome Map. Selected frames of an animated time-series visualization of 12,116 strains for a single patient from [23].**Panel (A)**: zoomed regions of the *Enterococcus* neighborhood as it progresses through the antibiotic response. **Panel (B)**: Full resolution animated microbiome map is available at the project's repository via links from biorg.cs.fiu.edu/jasper

[10]Simon Anders. Visualization of genomic data with the Hilbert curve. *Bioinformatics (Oxford, England)*, 25(10):1231–1235, May 2009.

[11]Apple. Mac App Store. `https://apps.apple.com/us/genre/mac`. Acessed: 2021-03-05-2021.

[12]Michael Bader. *Space-Filling Curves: An Introduction with Applications in Scientific Computing*. Springer Publishing Company, Incorporated, 2012.

[13]John J Bartholdi and Loren K. Heuristics Based on Spacefilling Curves for Combinatorial Problems in Euclidean Space. *Management Science*, 34(3):291–305, March 1988.

[14]Calle, M Luz. Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, 17(1), March 2019.

[15]James R Cole et al. Ribosomal Database Project - data and tools for high throughput rRNA analysis. *Nucleic Acids Research*, 2014.

[16]Xuegong Deng, Xuemei Deng, Simon Rayner, Xiangdong Liu, Qingling Zhang, Yupu Yang, and Ning Li. Dhpc: A new tool to express genome structural features. *Genomics*, 91(5):476 – 483, 2008.

[17]T. Z. DeSantis et al. Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Applied and Environmental Microbiology*, 72(7):5069–5072, 2006.

[18]Joseph C Devlin, Thomas Battaglia, Martin J Blaser, and Kelly V Ruggles. WHAM!: a web-based visualization suite for user-defined analysis of metagenomic shotgun sequencing data. *BMC genomics*, 19(1):493, June 2018.

[19]Adrian Dobra, Camilo Valdes, Dragana Ajdic, Bertrand Clarke, Jennifer Clarke, et al. Modeling association in microbial communities with clique loglinear models. *The Annals of Applied Statistics*, 13(2):931–957, 2019.

[20]Dutilh et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nature communications*, 5:4498, 2014.

[21]Mitch Fernandez, Vanessa Aguiar-Pulido, Juan Riveros, Wenrui Huang, Jonathan Segal, Erliang Zeng, Michael Campos, Kalai Mathee, and Giri Narasimhan. Microbiome analysis: State of the art and future trends. *Computational Methods for Next Generation Sequencing Data Analysis*, pages 401–424, 2016.

[22]Mitch Fernandez, Juan D Riveros, Michael Campos, Kalai Mathee, and Giri Narasimhan. Microbial "social networks". *BMC Genomics*, 16(11):1–13, December 2015.

[23]Molly K. Gibson et al. Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nature Microbiology*, 1(4):16024, March 2016.

[24]Zuguang Gu, Roland Eils, and Matthias Schlesner. HilbertCurve - an R/Bioconductor package for high-resolution visualization of genomic data. *Bioinformatics*, 2016.

[25]David Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. In *Dritter Band: Analysis · Grundlagen der Mathematik · Physik Verschiedenes*, pages 1–2. Springer, Berlin, Heidelberg, 1935.

[26]Human Microbiome Project Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, June 2012.

[27]Daniel A Keim. Pixel-Oriented Visualization Techniques for Exploring Very Large Data Bases. *Journal of Computational and Graphical Statistics*, February 1996.

[28]Jose Lugo-Martinez, Daniel Ruiz-Perez, Giri Narasimhan, and Ziv Bar-Joseph. Dynamic interaction network inference from longitudinal microbiome data. *Microbiome*, 7(1):54, 2019.

[29]Paul Muir et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17(1):53, 2016.

[30]Daniel J Nasko, Sergey Koren, Adam M Phillippy, and Todd J Treangen. RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome biology*, 19(1):165, October 2018.

[31]Nuala A O'Leary, Mathew W Wright, Terence D Brister, and Kim D Pruitt. Reference sequence (RefSeq) database at NCBI - current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–45, 2016.

[32]Christian Quast et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(D1):D590–D596, 11 2012.

[33]Daniel Ruiz-Perez, Jose Lugo-Martinez, Natalia Bourguignon, Kalai Mathee, Betiana Lerner, Ziv Bar-Joseph, and Giri Narasimhan. Dynamic bayesian networks for integrating multi-omics time-series microbiome data. *BioRxiv*, page 835124, 2019.

[34]Musfiqur Rahman Sazal, Daniel Ruiz-Perez, Trevor Cickovski, and Giri Narasimhan. Inferring relationships in microbiomes from signed bayesian networks. In *2018 IEEE 8th ICCABS Conference*, pages 1–1. IEEE, 2018.

[35]Nicola Segata et al. Metagenomic biomarker discovery and explanation. *Genome biology*, 12(6):R60, 2011.

[36]Valdes, Camilo, Stebliankin, Vitalii, and Narasimhan, Giri. Large scale microbiome profiling in the cloud. *Bioinformatics (Oxford, England)*, 35(14):i13–i22, July 2019.

[37]Sophie Weiss, Will Van Treuren, Catherine Lozupone, Karoline Faust, Jonathan Friedman, Ye Deng, Li Charlie Xia, Zhenjiang Zech Xu, Luke Ursell, Eric J Alm, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *The ISME journal*, 10(1):1669–1681,

2016.

[38]James Robert White, Niranjan Nagarajan, and Mihai Pop. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS computational biology*, 5(4), 2009.

[39]Pak Chung Wong, Kwong Kwok Wong, H. Foote, and J. Thomas. Global visualization and alignments of whole bacterial genomes. *IEEE Transactions on Visualization and Computer Graphics*, 9(3):361–377, July 2003.

[40]Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with Kraken 2. *Genome biology*, 20(1):1–13, December 2019.