frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 19 June 2023
DOI 10.3389/fbinf.2023.1154588

OPEN ACCESS

EDITED BY
Helena Jambor,
Technical University Dresden, Germany

REVIEWED BY
Mamoon Rashid,
King Abdullah International Medical Research Center (KAIMRC), Saudi Arabia
Marco Agus,
Hamad Bin Khalifa University, Qatar
Sayaka Mizutani,
Tokyo Institute of Technology, Japan

*CORRESPONDENCE
Camilo Valdes,
valdes2@llnl.gov
Giri Narasimhan,
giri@cs.fiu.edu

†PRESENT ADDRESS
Daniel Ruiz-Perez, Meta, Menlo Park, CA, United States

RECEIVED 30 January 2023
ACCEPTED 22 May 2023
PUBLISHED 19 June 2023

CITATION
Valdes C, Stebliankin V, Ruiz-Perez D, Park JI, Lee H and Narasimhan G (2023), Microbiome maps: Hilbert curve visualizations of metagenomic profiles. *Front. Bioinform.* 3:1154588. doi: 10.3389/fbinf.2023.1154588

Microbiome maps: Hilbert curve visualizations of metagenomic profiles

Camilo Valdes[1]*, Vitalii Stebliankin[2], Daniel Ruiz-Perez[2†], Ji In Park[3], Hajeong Lee[4] and Giri Narasimhan[2,5]*

[1]Lawrence Livermore National Laboratory, Physical and Life Sciences Directorate, Livermore, CA, United States, [2]Bioinformatics Research Group (BioRG), Florida International University, Miami, FL, United States, [3]Department of Medicine, Kangwon National University Hospital, Kangwon National University School of Medicine, Chuncheon, Gangwon-do, Republic of Korea, [4]Department of Internal Medicine, Seoul National University College of Medicine, Seoul, Republic of Korea, [5]Biomolecular Sciences Institute, Florida International University, Miami, FL, United States

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 are complex. Particularly challenging is their visualization, as existing techniques are inadequate when the taxa number is in the thousands. We present a technique, and accompanying software, for the visualization of metagenomic abundance profiles using a space-filling curve that transforms a profile into an interactive 2D image. We created Jasper, an easy to use tool for the visualization and exploration of metagenomic profiles from DNA sequencing data. It orders taxa using a space-filling Hilbert curve, and creates a "*Microbiome Map*", where each position in the image represents the abundance of a single taxon from a reference collection. Jasper can order taxa in multiple ways, and the resulting *microbiome maps* can highlight "hot spots" of microbes that are dominant in taxonomic clades or biological conditions. We use Jasper to visualize samples from a variety of microbiome studies, and discuss ways in which *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 also be converted into animated movies that bring to life the dynamicity of microbiomes.

KEYWORDS

visualization, Hilbert curve, image analysis, microbiome, metagenomics, profiling, maps, DNA sequencing

1 Introduction

Microbiome samples are routinely processed by means of low-cost, high-throughput metagenomics DNA sequencing, followed by the creation of microbial community abundance profiles (Calle, 2019), where the sequenced reads are mapped against a collection of microbial reference genomes like Ensembl (EMBL-EBI, 2022a) or RefSeq (O'Leary et al., 2016). Tools such as FLINT (Valdes et al., 2019) and Kraken 2 (Wood et al., 2019) facilitate the creation of microbial abundance profiles either from metagenomic

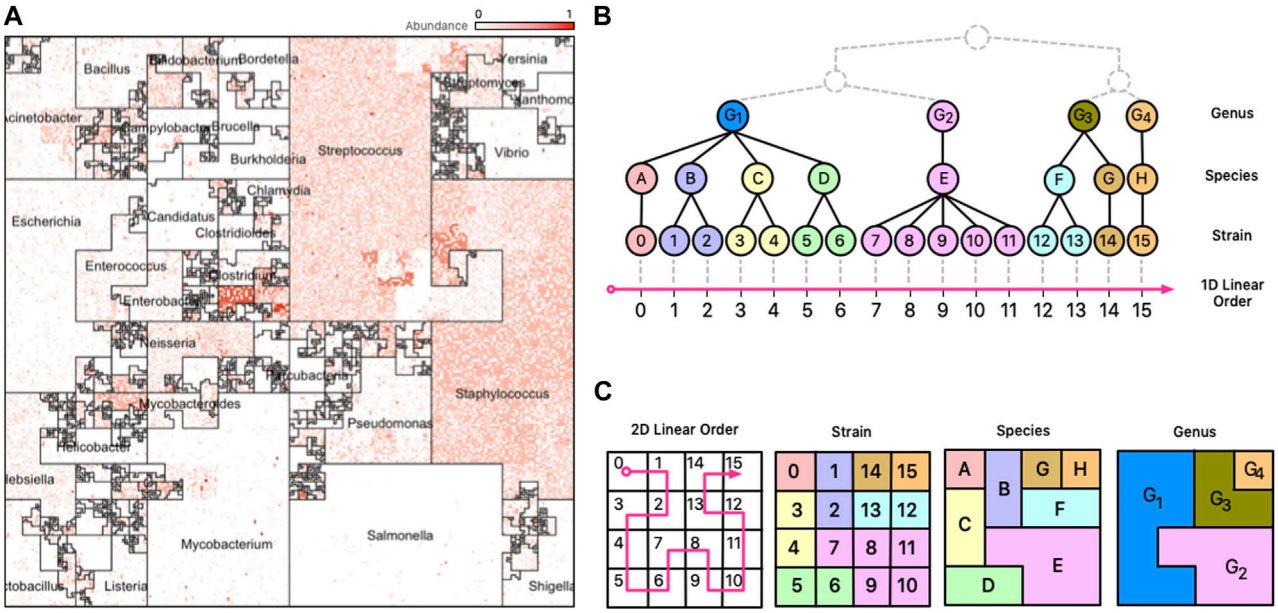

FIGURE 1
Taxonomic Ordering. **(A)** A *Taxonomic Ordering* of 44,048 reference genomes from Ensembl Bacteria. *Microbial neighborhoods* are drawn based on a taxonomic tree for microbial classification. The image depicts the distribution of Genera in the reference collection, and the size of each neighborhood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

whole-genome DNA sequencing (mWGS) or 16S-amplicon sequencing (16S) data. These abundance profiles are the stepping stones for downstream analyses such as differential abundance studies (White et al., 2009), co-occurrence pattern discovery (Dutilh et al., 2014; Fernandez et al., 2015; Fernandez et al., 2016; Weiss et al., 2016), Bayesian analyses (Rahman Sazal et al., 2018; Adrian et al., 2019), biomarker identification (Segata et al., 2011), multi-omics analyses (IHMP Consortium, 2014; Aguiar-Pulido et al., 2016), and analyses of profiles from longitudinal studies (Jose et al., 2019; Ruiz-Perez et al., 2019). Lower sequencing costs have resulted in an increasing number of larger deep sequencing metagenomic data sets (Muir et al., 2016).

Metagenomic profiles can contain relative abundance values [either sequence abundance or taxonomic abundance (Sun et al., 2021)] for the entire collection of microbial taxa present in a sample, and these profiles can be easily visualized by many software libraries and frameworks such as Matplotlib (The Matplotlib development team, 2023) and ggplot2 (Hadley, 2023), as well as data analysis software suites such as Tableau (Inc. Salesforce, 2023), or even MS Excel (Microsoft Corp, 2022). These tools are readily available to the public and allow for data exploration, but are primarily 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 data sets. Recent tools such as WHAM! (Devlin et al., 2018)! 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 reference genome collections continues to grow exponentially over time (Nasko et al., 2018), and even novel tools such as Krona (Ondov et al., 2011) are not able to display large amounts of reference genomes.

In this work we consider the problem of visualizing a microbiome using a visualization technique called the *Hilbert Curve Visualization* (HCV). We visualize abundance profiles using the Jasper tool, a free and easy to use software application that includes both graphical and command-line versions, and discuss the challenges of visualizing billions of microbial abundance measurements for hundreds of thousands of microbial genomes. We propose an alternative visualization technique that is useful when trying to combine many factors of metagenomic information in order to create interpretable images that can lead to improved understanding.

2 Materials and methods

We use a technique called the Hilbert curve visualization (HCV) to visualize the microbial community abundance profiles of a reference collection of genomes as a "microbiome map". For our experiments, we used a reference collection of 44K genomes (EMBL-EBI, 2022a), but the approach is readily scalable to deal with considerably larger collections. These profiles contain the relative abundance measurements of thousands of genomes, and they are ordered along a space-filling

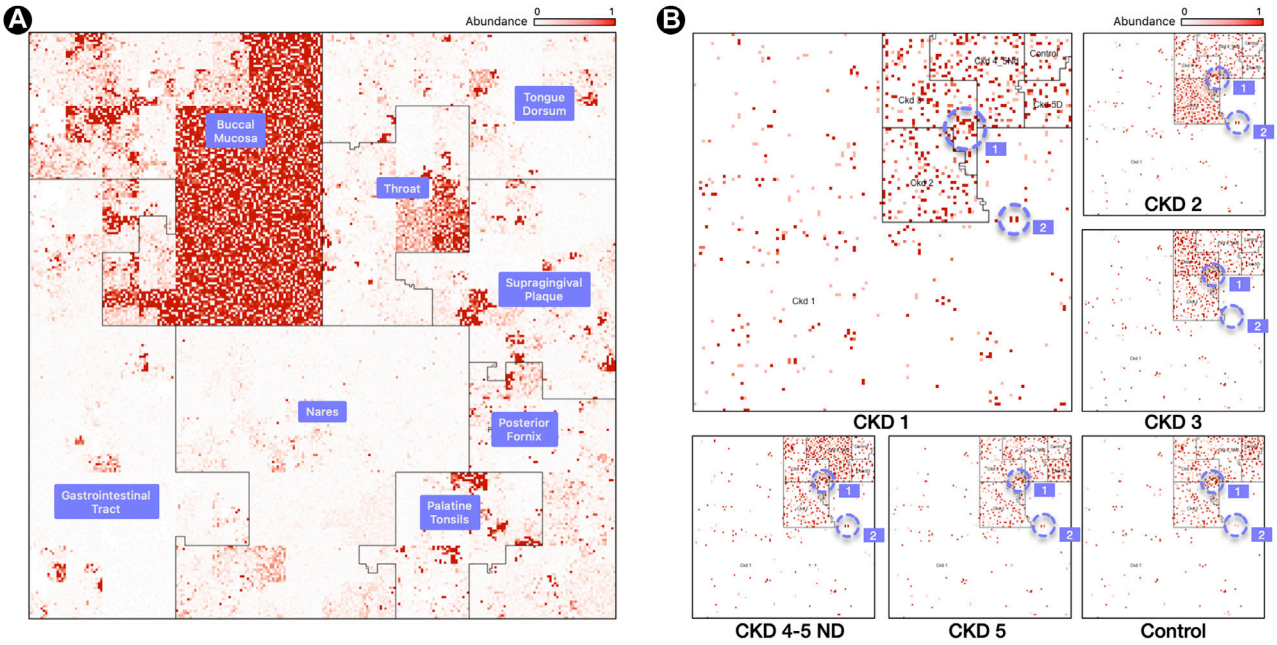

FIGURE 2
Labeled Ordering. **(A)** Samples are processed in a M x N matrix that contains *M* labeled samples, and *N* microbial taxa. The user specifies the ordering of the biological conditions that the *M* samples belong to. For each labeled grouping, the taxa with the highest mean relative abundance 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. The "1" and "2" regions within panel **(B)** highlight representative species clusters which discern the CKD stages.

curve in a 2D square using the Hilbert curve (Hilbert, 1935). Thus, in its simplest form, it is possible to visualize the profile of a single metagenomic sample. In the resulting 2D Hilbert image, each position (or pixel) corresponds to a genome from the reference collection and the position's intensity color value represents the relative abundance of a single genome in the sample.

As discussed below, depending on the ordering of the genomes that is selected in the software, different "Microbial Neighborhoods" are created, allowing for different interpretations of the "hotspots" of abundant genomes in the images. As explained later, the ordering in Figure 1 allows us to infer taxonomic clades that are most abundant in a sample, while the ordering in Figure 2 allows us to visualize site-specific or stage-specific taxa abundant in a sample.

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 (Bartholdi and Loren, 1988). 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 (Bader, 2012), 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 1, 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 into four subsquares, and then connecting the centers of these squares in a specific order. To provide a recursive definition of the curve, Figure 3A 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 (Keim, 1996) to represent stock market data. Since then, it has been used for visualizing genomic data (Deng et al., 2008; Anders, 2009) and DNA alignments of whole bacterial genomes (Wong et al., 2003).

In human genomics, the application of the HCV technique is straightforward as the natural linear order of genomic positions can be easily used by the curve, and there are tools for creating

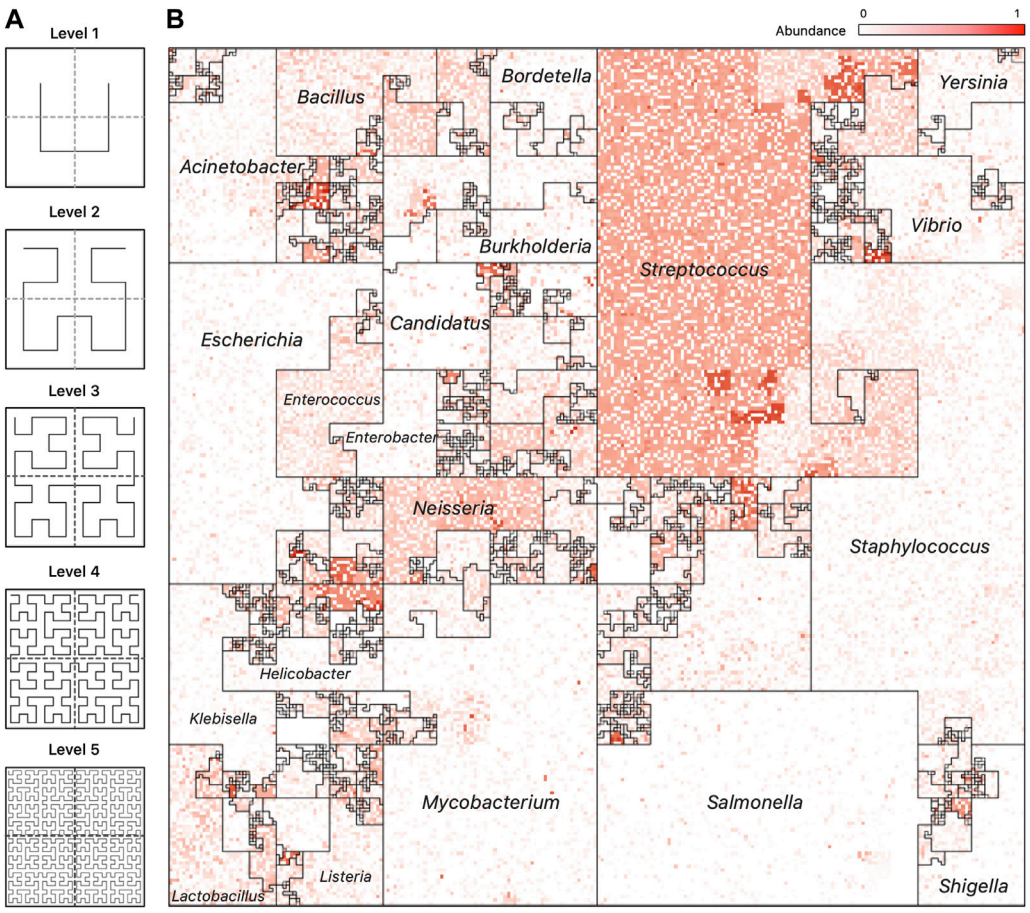

FIGURE 3
Hilbert Curve Visualization of Metagenomic Samples. **(A)** The first five iterations of the Hilbert curve: the Level 1 curve is obtained by connecting the centers of the four initial squares as shown; the Level k curve is obtained by a recursive partitioning of each square from Level $k-1$, 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 color intensity of each position in the image represents the abundance of one microbial genome. The bordered regions are "Microbial Neighborhoods" and represent groups of related microbes, and their size corresponds to the number of genomes in the reference collection.

Hilbert curve images from genomics data sets (HilbertVis (Anders, 2009) and HilbertCurve (Gu et al., 2016). Both of these tools apply HCV in the context of human genomics: a single scaffold is modeled as a single one-dimensional (1D) line in which each interval is taken to be a single genomic position. To date, the HCV technique has not been applied to metagenomics data sets.

2.2.1 Visualizing metagenomics data

Traditional 1D visualization techniques that display community abundance profiles often do not take into account latent metagenomic factors present in sequencing samples. Pie charts, line plots, etc., tend to focus on the taxa with the highest abundances, and have poor resolution for taxa with small abundances (which sometimes can be critical). Visual comparison of multiple samples is also difficult as determining the change in abundance of a single taxon between samples is not convenient.

Space-filling curves offer an intriguing scheme for visualizing metagenomics data for their ability to preserve positional data. This feature can be enhanced with a reference collection's metadata to create descriptive images that express metagenomic information succinctly. The issue of adding new genomes is discussed in detail in Section 3.2.

The Hilbert curve is not the only space-filling curve with these features, but its creation is a simple recursive partitioning that can be implemented elegantly and efficiently in software. Other curves, like the Peano curve (Peano, 1890), partition the square into 9 or more regions, which can lead to hard to interpret images when the levels of the curve are high (levels of 10 or above).

2.2.2 Linear orderings

The first challenge in visualizing abundance profiles with a space-filling curve is that there is no natural linear order for the reference collection. Below, we discuss two classes of orderings that are shown to be useful:

FIGURE 4
Jasper GUI Version The Jasper GUI software is a tool for interactively exploring *microbiome maps*. Users can click on any genome in the map and get a pop-over with detailed information, along with links to specific online resources about the genome in Ensembl, Uniprot, and Genbank. The maps can be exported and easily shared. Command-line versions of the software (non-interactive) are also available for Python and R.

- Taxonomic Ordering: Based on a taxonomic tree.
- Labeled Ordering**:** Based on a custom labeling scheme.

The area of the 2D square that bounds a *microbiome map* is proportional to the number of unique genomes in the reference collection. For a curve of size k, we have $2^k \times 2^k$ linear segments in which to place a genome in. However, the size of a reference collection will not always perfectly match the number of segments. To account for this, we merge adjacent segments along the curve's path so that their number is always equal to the number of genomes in the reference collection. Jasper has controls to understand a map's linear order: users can overlay the path of the curve to follow the map's layout (Figure 4).

3 Methods

Visualizing a microbiome's abundance profile starts by aligning DNA sequencing reads against a reference collection of genomes, and creating counts of the number of reads that align to each genome. The cloud-based tool FLINT (Valdes et al., 2019) facilitates the profiling of mWGS data sets and reports relative sequence abundances, while the Kraken 2 software does it for both mWGS and 16S data sets reporting sequence or taxonomic abundances. For the images shown in this paper, FLINT uses a reference collection of 44,408 microbial genomes from the Ensembl Bacteria database (EMBL-EBI, 2022a), while Kraken 2 uses a "16S" reference of 5,127 genomes which contains references from Greengenes (DeSantis et al., 2006), SILVA (Quast et al., 2012), and RDP (Cole et al., 2014).

Different linear orderings of the taxa on the Hilbert curve result in different images. In Jasper, users may select from two options: a *taxonomic ordering* (Figure 1) which uses the linear order from Ensembl's taxonomic tree, or a *labeled ordering* (Figure 2) which is based on a custom user-supplied label. The project website at www.microbiomemaps.org contains a manual with examples that show how to create them. Note that unlike other applications of HCV to genomics data (Anders, 2009), *microbiome maps* do not depict a single genomic object (e.g., a nucleotide), but rather, each pixel corresponds to a reference genome, and its intensity to its abundance in a metagenomic sample.

3.1 Microbial neighborhoods

Different linear orderings produce different Hilbert curve 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 (Google, 2023), and we term these areas "*Microbial Neighborhoods*" (Figure 1, 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. These neighborhoods offer a quick and visual way to readily identify abundance "hotspots" that can contextualize the important features of a metagenomic sample and identify important microbial groups.

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, remain generally 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 seen in Figure 1, multiple taxonomic levels can be displayed at the same time in a single image. The ability of a *microbiome map* to display multiple levels of a taxonomic tree at once, while at the same time providing high-resolution abundance information for single genomes, is a compelling advantage over visualizing data with 1D methods were the sheer number of data points would overwhelm the observer.

An advantage of the *Taxonomic Ordering* is that it creates a visual way of depicting the number of reference genomes in the reference collection, making it possible to compare relative sizes of clades: the *Microbial Neighborhoods* in this ordering represent large clades with many taxa, which in turn occupy large areas of the map. An example of this case isFigure 3B; 1A which shows that the Ensembl bacterial database contains large numbers of strains from the *Streptococcus* and *Staphylococcus* genera.

Another advantage of this ordering is that we can quickly understand the diversity of a reference collection by creating *microbiome maps* with no color; these colorless maps create a visual representation of the reference collection that shows us its taxonomic distribution.

3.1.2 Linearizing taxonomic trees

Illustrating a taxonomic tree as a 2D Hilbert curve starts by finding a linear order of the leaf nodes in the tree. Figure 1, panel B), depicts a fictitious taxonomic tree with 16 microbial strains at the leaf nodes 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 1, 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 tree structures do not have a natural linear order—they do not have a "start/finish", or a "left/right". Different tree orderings can result by permuting the tree's children nodes at any given node of the tree, and algorithms for finding an optimal order have been proposed (Bar-Joseph et al., 2001), but the optimal order relies on the tree satisfying specific properties (e.g., a binary tree) and the existence of a good optimality measure. The *taxonomic ordering* linearizes a tree by using data from Ensembl's Pan-taxonomic Compara (EMBL-EBI, 2022c) and the Ensembl Genomes (EMBL-EBI, 2022b) databases as the foundation for the Hilbert curve. The genomes in the database are annotated so that we can establish a linear order. For mWGS data, the leaf nodes of the tree are typically at the strain level (for 16S data, the leaf nodes are at the species or genus level). As shown in Figure 1A, the 2D square that bounds the Hilbert image represents all genus-level groups. However, depending on the data and the application this could be modified to suit the needs so that it represents a good compromise between taxonomic information and visual interpretability.

3.1.3 Condition Neighborhoods

The second option for ordering genomes along the Hilbert curve is the *labeled ordering* (Figure 2) which creates "*Condition Neighborhoods*" by using an ordering scheme that determines the 1D linear order based on a user-supplied labeling of samples. This labeling is provided as a labeled $m \times n$ sample matrix M, where m are sample rows, and n are the genomes in the reference database. For sample i and reference genome j, the matrix entry $M[i, j]$ corresponds to the abundance of genome j in sample i.

Establishing the linear order for multiple conditions starts with a user-defined ordering of the set of k conditions, C_1, C_2, . . . , C_k. Conditions may represent different disease stages, time intervals, drug dosage, or sampling locations (body sites, environmental sites, etc.).

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 according to the established order.

Assigning a color to a taxon based on abundance is only meaningful if its presence is above the threshold of noise, which we determine when we normalize the input matrix M. 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 ordering criteria. After the conditions have been organized along the curve, taxonomic information is used to order genomes within the 2D region of the condition.

In this ordering, the *microbiome map* is still visualizing only one sample, but one can readily spot the relevant biological condition with which it has the most overlap. "*Hotspots*" will most likely appear in the region corresponding to one of the conditions, and users can readily tell what condition the sample belongs to by identifying the area, i.e., neighborhood, in the image with the most hotspots. Thus, it is easy to infer by visual inspection that Figure 2 panel (A) is with high probability a sample from the buccal mucosa region with some taxa that are typically abundant in the throat and other oral sites. Similar inferences are possible with the disease stage or with the environmental condition of the sample. Clusters of bright positions will also appear in other neighborhoods [Figure 2,

panel (B)], 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.

The process for defining a new custom order is simple. Detailed instructions and examples can be found at the project website at www.microbiomemaps.org.

3.2 Adding new genomes or samples

When adding new samples, preserving a genome's locality or neighborhood becomes important for drawing consistent and useful conclusions about changes. This is easy for the taxonomic ordering. For the labeled ordering, the preservation of locality for a genome is a little more delicate as the assignment of a genome to a neighborhood is done by identifying taxa whose mean relative abundance (or other user-defined metric) is highest in a group of samples. If the new sample to be added changes a taxa's mean relative abundance in the sample's group, then it could affect the map's topology. This is a disadvantage to all mapping orders that rely on precomputing a value to place taxa in a neighborhood—especially when that value is computed across a group of samples, as new samples will require their re-evaluation. Inserting new genomes to the reference collection could affect all existing plots because the positions of existing genomes may change. The addition of new taxa into an existing ordering will also change the relative abundance of all or almost all taxa, changing the color intensities of the pixels. Several suggestions to alleviate the aforementioned problems are provided in Section 4.

4 Results

We created microbiome maps for two groups of metagenomic data sets: 24 mWGS normal samples taken from the Human Microbiome Project (HMP) (Human Microbiome Project Consortium, 2012), 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 (Valdes et al., 2019), and the 16S CKD samples with Kraken 2 (Wood et al., 2019). 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) (KDIGO, 2023)], 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 deteriorating function of the patient's kidneys, and three samples from each group were used.

4.1 Microbiome maps

Microbiome maps 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 (Valdes et al., 2019)]; for 16S samples, each position in the image displays information about microbial species [abundances as reported by Kraken 2 (Wood et al., 2019)]. Both sets of profiles were then converted into *microbiome maps* using the *taxonomic*, and *labeled* orders.

Figure 1, 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 2, panel (A), contains a map from a Buccal Mucosa sample of the HMP dataset. The map contains the same 44K genomes from Figure 1, panel (A), but ordered with a *labeled ordering* scheme, based on the highest mean relative abundance of a genome in its cohort. The advantage of this scheme is that identifying the biological condition that the sample belongs to is effortless: one need only to look at the neighborhood that contains the most *hotspots* (Buccal Mucosa, in this case).

Figure 2, panel (B), shows six maps of 16S samples from the CKD analysis created using the *labeled ordering* scheme, displaying the abundances for 5,127 species each. The order is based on the mean relative abundance of the microbes in the samples belonging to each CKD stage: the most prominent taxa in Stage 1 are surrounded by the CKD 1 area, the most prominent taxa in CKD Stage 2 are surrounded by the CKD 2 area, and so on. Note that the more noticeable regions with a higher density of brighter regions are from samples of the same cohort and we can readily identify the CKD stage by looking at the density of *hotspots*.

What is more significant is the way these plots show the microbes shared by different stages of the disease. For example, the region marked "1" in Figure 1, panel (B), shows a group of microbes that appear in all stages, while the region marked "2" appears in almost all stages except CKD3 (absent) and Control (lowered abundance).

By fixing the orderings of the taxa, a *microbiome map* can be used to present groups of metagenomic samples that can be partitioned temporally (longitudinal studies), spatially (body or environmental sites), by disease (sub)type, by disease stage, and by developmental stages. Additionally, it is readily possible to create *average microbiome maps*, *aggregate* maps, and *differential* maps showing either average, aggregate, or differential abundances, respectively.

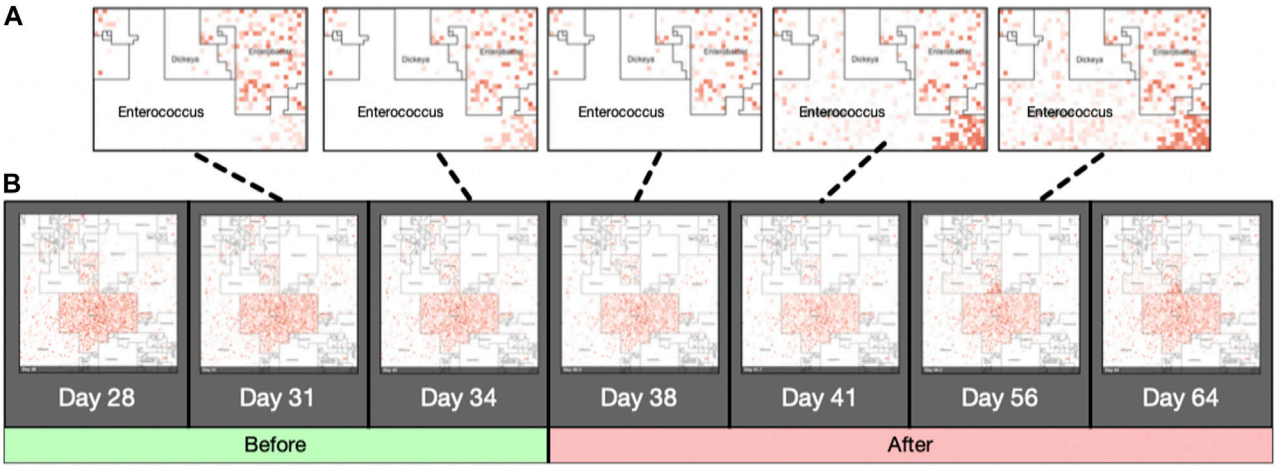

FIGURE 5
Animated Microbiome Map. Selected frames of an animated time-series visualization of 12,116 strains for a single patient from (Gibson et al., 2016). Panel **(A)**: zoomed regions of the *Enterococcus* neighborhood as it progresses through the antibiotic response. Panel **(B)**: More examples of animated microbiome maps are available at www.microbiomemaps.org.

To address the problem of adding new genomes or samples, we offer the following suggestions. One possible solution is to leave "blank" (i.e., unassigned) pixels on the map to allow for future additions. This could be implemented by inserting the gap so that the next clade always starts at the boundary of a square region of size $2^k \times 2^k$, for a predetermined value of k. Second, different parts of the ordering (e.g., a taxonomic clade or a condition associated with a label) can be assigned different colors instead of the monochromatic plots shown here. Finally, it may be useful to always provide a reference microbiome map to clarify the labeling.

4.1.1 Comparison to other methods

Using 1D visualizations are helpful for condensing information for multiple samples. However, they lack the ability to display nuanced information and to scale to deal with the exponential growth in the databases (44K bacterial strains were used in our visualizations) while retaining the perspective of the latent metagenomic relationships. The project website contains samples visualized with *microbiome maps* and compared to WHAM! and iMAP images.

5 Discussion

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 fluctuations through a time series, or accross biological conditions. Figure also 3 panel (B) 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 (Gibson et al., 2016), 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 and 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.

5.1 Jasper: visual inspection and command-line tools

The graphical user interface (GUI) version of Jasper (Figure 4) offers multiple controls for interactively inspecting *microbiome maps*, and integrates with online resources like Ensembl (EMBL-EBI, 2022b), GenBank (Benson et al., 2012), and Uniprot (The UniProt Consortium, 2014). In the GUI, users can identify any genome in the map by hovering their mouse pointer or clicking on any region. If a user does click, a pop-over is displayed with direct links to the online resources mentioned above, which provide detailed information about the genome. To make the map's layout easier to understand, the software also allows users to overlay the path of the Hilbert curve on top of the map. The Jasper GUI also includes other tools for researchers, as well as the ability to export maps as high quality vector images. The software can also be used with no profiles to create Hilbert curves which can be used for learning about space-filling curves. The GUI software is currently localized for North America and English-speaking users. We are working on adapting it to other languages and regions, and also working on accessibility features to make the software easier to use for users with disabilities. The Jasper GUI is developed with the Swift programming language (Apple Inc, 2023b), and is free to use with no restrictions. There is also a command-line version of Jasper for Python 3 (Python Software Foundation, 2023) and R (The R Foundation, 2023) that can create multiple images but is non-interactive.

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 data sets. 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 abundance of microbial taxa from a reference collection using two ordering schemes that can be used to create a *microbiome map*: the first, the *taxonomic ordering* is a default ordering that relies on taxonomic information, and can be used to create images that express abundance values in the context of taxonomic clades that the genomes belong to. The second, the *labeled ordering*, is dependent on a user-specified labeling of biological conditions, 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 release of the Jasper software, we are exploring other orderings that will be incorporated in future releases, such as orderings specific to time-series analyses, or multi-omics data sets.

5.2 Website: www.microbiomemaps.org

The website www.microbiomemaps.org is being developed into an online community resource for cataloguing maps from different data sets. Future releases of Jasper will allow users to share their maps with the website directly from the app, and also on social media. The website will eventually offer curated collections of maps which will be free to use by the community. Documentation and manuals for the Jasper software is available at www.microbiomemaps.org/manual and the GUI version can be downloaded for free from the Mac App Store (Apple Inc, 2023a), and requires macOS. A command line interface (CLI) version of Jasper (Python 3 and R) suitable for high performance computing environments is also available at www.microbiomemaps.org The CLI version is non-interactive, but can be used to create batches of images suitable for featurization tasks in machine learning workflows.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: www.microbiomemaps.org and https://biorg.cs.fiu.edu/jasper/.

Author contributions

CV conceived the idea of microbiome maps and taxonomic neighborhoods, as well as developed the GUI and CLI versions of the software, and microbiome movies. CV and GN wrote the manuscript and collaborated on the figures. GN also developed the idea of conditional and labeled neighborhoods. VS helped with the collection of the annotations for the taxonomic tree and created the abundance profiles for the microbiome movies. JP and HL provided the data and analyses for the 16S CKD data sets. DR-P created an early prototype visual inspector for inspecting the maps. All authors contributed to the article and approved the submitted version.

Funding

The work of CV was performed in part under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. The work of CV was also funded while at the University of Nebraska-Lincoln through a University of Nebraska Program of Excellence award, as well as the University of Nebraska-Lincoln (UNL) Quantitative Life Sciences Initiative. An FIU Dissertation Year Fellowship partially supported the work of CV and DR-P at FIU, and the work of DR-P, and JP were done while they were at FIU.

Acknowledgments

The authors would like to thank the members of the Bioinformatics Research Group, BioRG, at Florida International University (FIU) for their valuable feedback and comments. We also thank Dr. Jennifer Clarke at UNL, and Dr. Kalai Mathee at FIU, for their helpful comments and feedback on the project.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Adrian, D., Valdes, C., Ajdic, D., Clarke, B., and Clarke, J., (2019). Modeling association in microbial communities with clique loglinear models. *Ann. Appl. Statistics* 13 (2), 931–957. doi:10.1214/18-AOAS1229

Aguiar-Pulido, V., Huang, W., Suarez-Ulloa, V., Cickovski, T., Mathee, K., and Narasimhan, G. (2016). Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis: Supplementary issue: Bioinformatics methods and applications for big metagenomics data. *Evol. Bioinforma.* 12, EBO.S36436. doi:10.4137/ebo.s36436

Anders, S. (2009). Visualization of genomic data with the Hilbert curve. *Bioinforma. Oxf. Engl.* 25 (10), 1231–1235. doi:10.1093/bioinformatics/btp152

Apple Inc (2023). *Mac App Store*. Available at: apps.apple.com/us/genre/mac (Acessed: 05 10 2021).

Apple Inc (2023). *Swift. swift.org* (Acessed: 08 01 2023).

Bader, M. (2012). *Space-filling curves: An introduction with applications in scientific computing*. Incorporated: Springer Publishing Company.

Bar-Joseph, Z., Gifford, D. K., and Tommi, S. J. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics* 17 (1), S22–S29. doi:10.1093/bioinformatics/17.suppl_1.s22

Bartholdi, John J., and Loren, K. (1988). Heuristics based on spacefilling curves for combinatorial problems in euclidean space. *Manag. Sci.* 34 (3), 291–305. doi:10.1287/mnsc.34.3.291

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et al. (2012). GenBank. *Nucleic Acids Res.* 41 (D1), D36–D42. doi:10.1093/nar/gks1195

Calle, M. L. (2019). Statistical analysis of metagenomics data. *Genomics and Inf.* 17 (1), e6. doi:10.5808/gi.2019.17.1.e6

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., McGarrell, D. M., Sun, Y., et al. (2014). Ribosomal Database Project - data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi:10.1093/nar/gkt1244

Deng, X., Deng, X., Rayner, S., Liu, X., Zhang, Q., Yang, Y., et al. (2008). Dhpc: A new tool to express genome structural features. *Genomics* 91 (5), 476–483. doi:10.1016/j.ygeno.2008.01.003

DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., et al. (2006). Greengenes, a chimera-checked 16s rrna gene database and workbench compatible with arb. *Appl. Environ. Microbiol.* 72 (7), 5069–5072. doi:10.1128/aem.03006-05

Devlin, J. C., Battaglia, T., Blaser, M. J., and Ruggles, K. V. (2018). WHAM!: A web-based visualization suite for user-defined analysis of metagenomic shotgun sequencing data. *BMC genomics* 19 (1), 493. doi:10.1186/s12864-018-4870-z

Dutilh, B. E., Cassman, N., McNair, K., Sanchez, S. E., Silva, G. G. Z., Boling, L., et al. (2014). A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. *Nat. Commun.* 5, 4498. doi:10.1038/ncomms5498

EMBL-EBI (2022). *Ensembl Bacteria*. Available at: https://bacteria.ensembl.org (Accessed 10 15 2022).

EMBL-EBI (2022). *Ensembl genomes*. Available at: http://ensemblgenomes.org (Accessed 10 17 2022).

EMBL-EBI (2022). *Pan taxonomic Compara*. Available at: https://ensemblgenomes.org/info (Accessed 10 17 2022).

Fernandez, M., Riveros, J. D., Campos, M., Mathee, K., and Narasimhan, G. (2015). Microbial "social networks". *BMC Genomics* 16 (11), S6–S13. doi:10.1186/1471-2164-16-s11-s6

Fernandez, M., Aguiar-Pulido, V., Riveros, J., Huang, W., Segal, J., Zeng, E., et al. (2016). "Microbiome analysis: State of the art and future trends," in *Computational methods for next generation sequencing data analysis*, 401–424.

Gibson, M. K., Wang, B., Ahmadi, S., Burnham, C. A. D., Tarr, P. I., Warner, B. B., et al. (2016). Developmental dynamics of the preterm infant gut microbiota and antibiotic resistome. *Nat. Microbiol.* 1 (4), 16024. doi:10.1038/nmicrobiol.2016.24

Google (2023). *Google maps*. Available at: https://www.google.com/maps (Accessed 11 17, 2022).

Gu, Z., Eils, R., and Schlesner, M. (2016). HilbertCurve - an R/Bioconductor package for high-resolution visualization of genomic data. *Bioinformatics* 32, 2372–2374. doi:10.1093/bioinformatics/btw161

Hadley, W. (2023). *ggplot2*. Available at: https://ggplot2.tidyverse.org (Acessed: 01 20 2023).

Hilbert, D. (1935). "Uber die stetige abbildung einer linie auf ein flachenstuck," in *Dritter band: Analysis · grundlagen der Mathematik · physik verschiedenes* (Berlin, Heidelberg: Springer), 1–2.

Human Microbiome Project Consortium (2012). A framework for human microbiome research. *Nature* 486 (7402), 215–221. doi:10.1038/nature11209

IHMP Consortium (2014). The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell. host microbe* 16 (3), 276–289. doi:10.1016/j.chom.2014.08.014

Inc. Salesforce (2023). *Tableau*. Available at: https://www.tableau.com (Accessed 11 14 2022).

Jose, L-M., Ruiz-Perez, D., Narasimhan, G., and Bar-Joseph, Z. (2019). Dynamic interaction network inference from longitudinal microbiome data. *Microbiome* 7 (1), 54. doi:10.1186/s40168-019-0660-3

KDIGO (2023). *Kidney disease improving global outcomes guidelines*. Available at: https://kdigo.org/guidelines/ (Accessed 11 17 2023).

Keim, D. A. (1996). Pixel-oriented visualization techniques for exploring very large data bases. *J. Comput. Graph. Statistics* 5, 58. doi:10.2307/1390753

Microsoft Corp (2022). *Microsoft Excel*. Available at: https://products.office.com/en-us/excel (Accessed 11 14. 2022).

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., et al. (2016). The real cost of sequencing: Scaling computation to keep pace with data generation. *Genome Biol.* 17 (1), 53. doi:10.1186/s13059-016-0917-0

Nasko, D. J., Koren, S., Phillippy, A. M., and Treangen, T. J. (2018). RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.* 19 (1), 165. doi:10.1186/s13059-018-1554-6

O'Leary, N. A., Wright, M. W., Brister, T. D., Pruitt, K. D., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI - current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* 44 (D1), D733–D745. doi:10.1093/nar/gkv1189

Ondov, B., Bergman, N., and Adam, P. (2011). Interactive metagenomic visualization in a web browser. *BMC Bioinforma.* 1 (1), 385. doi:10.1186/1471-2105-12-385

Peano, G. (1890). Sur une courbe, qui remplit toute une aire plane. *Math. Ann.* 36 (1), 157–160. doi:10.1007/bf01199438

Python Software Foundation (2023). *Python*. Available at: https://www.python.org/ (Acessed: 08 01 2023).

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* 41 (D1), D590–D596. doi:10.1093/nar/gks1219

Rahman Sazal, M., Ruiz-Perez, D., Cickovski, T., and Narasimhan, G. (2018). Inferring relationships in microbiomes from signed bayesian networks. Proceeding of the 2018 IEEE 8th ICCABS Conference. October 2018, Las Vegas, NV, USA. IEEE, 1.

Ruiz-Perez, D., Lugo-Martinez, J., Bourguignon, N., Mathee, K., Lerner, B., Bar-Joseph, Z., et al. (2019). Dynamic bayesian networks for integrating multi-omics time-series microbiome data. *bioRxiv*. [Preprint]. doi:10.1101/835124

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W. S., et al. (2011). Metagenomic biomarker discovery and explanation. *Genome Biol.* 12 (6), R60. doi:10.1186/gb-2011-12-6-r60

Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A. P., et al. (2021). Challenges in benchmarking metagenomic profilers. *Nat. methods* 18 (6), 618–626. doi:10.1038/s41592-021-01141-3

The Matplotlib development team (2023). *Matplotlib*. Available at: https://matplotlib.org (Acessed: 01 21 2023).

The R Foundation (2023). *The R project for statistical computing*. Available at: https://www.r-project.org/ (Acessed: 01 08 2023).

The UniProt Consortium (2014). UniProt: A hub for protein information. *Nucleic Acids Res.* 43 (D1), D204–D212. doi:10.1093/nar/gku989

Valdes, C., Stebliankin, V., and Narasimhan, G. (2019). Large scale microbiome profiling in the cloud. *Bioinforma. Oxf. Engl.* 35 (14), i13–i22. doi:10.1093/bioinformatics/btz356

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10 (1), 1669–1681. doi:10.1038/ismej.2015.235

White, J. R., Nagarajan, N., and Pop, M. (2009). Statistical methods for detecting differentially abundant features in clinical metagenomic samples. *PLoS Comput. Biol.* 5 (4), e1000352. doi:10.1371/journal.pcbi.1000352

Wong, P. C., Wong, K. K., Foote, H., and Thomas, J. (2003). Global visualization and alignments of whole bacterial genomes. *IEEE Trans. Vis. Comput. Graph.* 9 (3), 361–377. doi:10.1109/TVCG.2003.1207444

Wood, D. E., Lu, J., and Ben, L. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20, 1–13.