ABSTRACT

Title of Dissertation: SOFTWARE INFRASTRUCTURE FOR VISUAL AND INTEGRATIVE ANALYSIS OF MICROBIOME DATA

Justin Max Wagner, Doctor of Philosophy, and 2018

Dissertation directed by: Associate Professor, Hector Corrada Bravo, Department of Computer Science

Microbiome sequencing allows researchers to reconstruct bacterial community census profiles at resolutions greater than previous methodologies. As a result, increasingly large numbers of these taxonomic community profiles are now generated, analyzed, and published by researchers in the field. In this work, I present new methods and software infrastructure for visualization and sharing of microbiome data. The overall goal is to enable a researcher to complete cycles of exploratory and confirmatory analysis over metagenomic data. I describe Metaviz, an interactive statistical and visual analysis tool specifically designed for effective taxonomic hierarchy navigation and data analysis feature selection. I next detail the incorporation of Metaviz into the Human Microbiome Project Data Portal. I then show a novel method to visualize longitudinal data across multiple features built as an extension over Metaviz. Finally, previous work has shown that specific subjects in an experimental cohort can be identified using their microbiome data. I developed software using a secure multi-

party computation library to complete comparative analyses of metagenomic data across cohorts without directly revealing feature count values for individuals.

SOFTWARE INFRASTRUCTURE FOR VISUAL AND INTEGRATIVE ANALYSIS OF MICROBIOME DATA

by

Justin Max Wagner

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Hector Corrada Bravo, Chair
Professor Mihai Pop
Professor Niklas Elmqvist
Professor Max Leiserson
Professor Michael Cummings

Dedication

To Katherine, Vitas, Ruta, Paul and In Memory of George Powers Dirth III

Acknowledgements

I would first like to acknowledge the guidance and effort of my advisor, Hector Corrada Bravo. I would also like to acknowledge the members of my thesis committee: Mihai Pop, Niklas Elmqvist, Max Leiserson, and Michael Cummings. Further, the input of my proposal committee members Amitabh Varshney and Jonathan Katz was invaluable. I owe a special gratitude to the members of the Corrada Bravo lab.

This work involved a collaboration with the Institute for Genome Sciences at the University of Maryland School of Medicine. I recognize the wonderful work done by Anup Mahurkar, Victor Felix, Jonathan Crabtree, Heather Creasey, James Matsumura, and the rest of the Institute.

My scientific training began under the guidance of my father Paul Wagner and brother Vitas Wagner. My mother has always encouraged my education and her reassurance has been critical to my career. My wife Katherine provided tireless support and encouragement from the first day of my graduate education. All my extended family members encouraged and supported me during my graduate research career and I thank each of them.

My formal scientific career began in the laboratory of Dr. William Skach and under the direct supervision of Dr. Teresa Buck. I thank them for their excellent example and early guidance.

George Dirth was the first person to encourage my study of computer science and I thank him for his friendship and counsel.

My interactions in all the educational institutions I was part of helped me develop as a member of the academic community. I would like to highlight those during graduate school including the students, staff, and faculty of the UMD Computer Science Department and the Center for Bioinformatics and Computational Biology. Thank you for all the friendships, conversations, and assistance.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

1.1 Microbiome sequencing

A microbiome is the collection of microbial organisms in an environment. High throughput DNA sequencing provides a mechanism to generate a microbial community census. Current research focuses on identification of the microbiome in human body sites[1] and different ecological domains[2]. For human health, studies are designed as large observational epidemiological studies or smaller controlled experiments. Initial large observational studies focused on identifying the microbiome of healthy individuals, examining known and detecting novel pathogens in diarrheal diseases[3] and observing the relationship between the obesity and an individual's microbiome[4]. One large epidemiological study of note is the Global Enteric Multi-Center Study, which gathered stool samples from children with diarrheal disease and matched controls in four countries to identify associations between microbiome structure and disease status[5]. Another prominent study examined the microbiome of individuals with Inflammatory Bowel Disease with a focus on Crohn's disease[6]. Recent and ongoing work in the field investigates the feasibility and effectiveness of modifying the microbiome of an organism to potentially alter host health.

Researchers create microbiome community profiles for a community by first taking a sample and extracting DNA. Next, one of two high-throughput sequencing methods is employed. The first method amplifies specific variable regions of the 16S ribosomal RNA gene. After the products are sequenced, the reads are clustered and annotated against a taxonomic annotation reference database. The number of times a given taxonomic unit is observed for each sample is computed into a count table that serves as the main object of subsequent downstream analysis. The other sequencing method is whole metagenome shotgun sequencing. The reads from this sequencing approach are either aligned to reference genomes, assembled, used in k-mer based taxonomic classification[7], or compared against clade-specific gene catalogues[8] to produce taxonomic profiles. Marker gene sequencing surveys are more accessible to perform than whole metagenome shotgun sequencing and are more often used. Metagenome sequencing allows for gene-level resolution and functional profiling while marker gene surveys must rely on a functional inference estimation.

1.2 Microbiome Data Analysis

Examining collections of microbiome requires processing pipelines and robust analysis methods. Looking at historical data analysis techniques, the recommendation from John W. Tukey is for multiple rounds of exploratory data analysis and confirmatory data analysis[9]. Figure 1 shows the rounds of successive refinement of trend identification and testing if the result is more likely than random chance. With this approach, robust confirmatory and exploratory methods are needed to interrogate datasets for results.

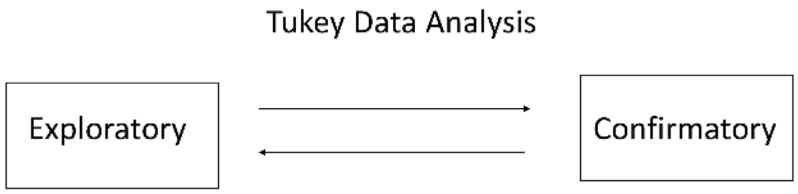


Figure 1: Tukey Data Analysis Ideal

Data visualization is an essential aspect of exploratory data analysis. Several projects provide mechanisms for visualization of microbiome data. One widely used approach is Krona that displays the taxonomic hierarchy as a Sunburst diagram with relative abundance of a given taxa represented as the arc length at that level of the taxonomy[10]. Taxonomer performs both read taxonomic assignment and visualization of results using a sunburst diagram to visualize features[11]. The R package Pavian incorporates Shiny and D3.js components to enable interactive analysis of results for metagenomic classification tools[12]. VAMPS is a web-service that provides a JavaScript and PHP-based metagenomics visualization toolkit of datasets uploaded by researchers[13]. Anvi'o is a multiomics platform that supports analysis using custom JavaScript visualizations[14]. The web-service MicrobiomeDB hosts microbiome community taxonomic profile data from open datasets and uses Shiny to visualize data[15].

The bulk of confirmatory data analysis is often carried out using statistical methods that implement a hypothesis testing procedure. Statistical methods for microbiome sequencing data include biomarker discovery and phylogenetic analysis. *metagenomeSeq* is an R/Bioconductor package which implements a method for normalization and differential abundance testing using a zero-inflated Gaussian mixture model[16]. *phyloseq* is an R/Bioconductor package for analysis of microbiome data including ordination methods and diversity analysis[17]. For visualization, *metagenomeSeq* and *phyloseq* offer static plotting utilities.

Metaviz provides a comprehensive interactive exploratory utility for microbial marker-gene sequencing and whole metagenome shotgun sequencing data with integration to confirmatory analysis utilities from the R/Bioconductor *metagenomeSeq* package. Metaviz is unique compared to the other tools listed above as it works with data from both sequencing methods, hosts datasets as a web-service, can be used as a standalone instance, generates high quality graphics, and links tightly with a statistical testing package.

1.3 Genomic Sequencing Project Data Access and Coordination

Prominent Projects

The ENCODE Data Coordination Center hosts a comprehensive data portal which includes dataset curation and archiving, reprocessing capabilities on updates of reference genomes, and sample selection utilities[18]. The portal serves data from several projects including the Epigenome Roadmap and links to UCSC Genome Browser and ENSEMBLE Browser for visualization[19]. The TCGA Data Portal was a comprehensive resource of cancer genome data and now is moving the Genomic Data Commons TCGA[20].

Data Security and Privacy

Addressing privacy and security concerns is a vital aspect of large-scale sequencing projects. There are two motivations to share genomic and physical attribute data broadly from a scientific perspective. First is the statistical power of observational tests. The power of a statistical test relies on the sampling size and method, therefore enabling many researchers to share data can lead to a stronger statistical result. The second reason for sharing genomic data is reproducibility of results. Scientific studies are designed and described so that other researchers can produce the result following the same procedure over the same material. For genomic data analysis, the functions are made public but the genomic data itself also needs to be accessible by other scientists in a way that addresses privacy concerns.

Several proposals exist for accomplishing the goal of sharing data between researchers while offering various level of privacy guarantees. These fall into three broad areas: (1) Access control in which a data custodian manages which parties can view and run analysis over research participants' data, (2) Statistical perturbation of data or output where the identity of any individual sequence is guaranteed to not be inferred by other parties running queries over participant data, (3) Secure multi-party computation in which nothing except the function output, and anything that can be logically inferred from it, is revealed during evaluation.

Access control forms the basis of the current policy for an NIH genomic research data-management system known as the Database for Genotypes and Phenotypes (dbGAP). In this setting, the security of private medical information rests on data analysts and the institutions that hold the data. Once a participant provides data to a research group, those analysts are responsible for IT security while processing data. When sharing data with other institutions, data is sent to another researcher who is then trusted to securely store and manage access to a copy of the data during the study. As recent attacks on medical records systems at various US hospitals, health insurance companies, and government employee record databases demonstrate, warehousing vast amounts of data leaves them particularly vulnerable to attack. Further, in dbGAP, patients need to provide broad consent to allow research data to be released to other parties. When additional data needs to be gathered or released to other analysts, re-consenting patients is a time-consuming and cumbersome task.

Statistical perturbation of analysis results, most widely implemented as differential privacy, is a second approach for researchers to provide privacy guarantees to participants. In this setting, a researcher maintains a data set and allows other researchers to perform queries over the data. Informally, the results of these queries are perturbed in such a manner that an adversary, with access to query results over a data set in which one specific participant has a set of values and results from another data set with that specific participant having a different set of values, will not be able to infer any information about that individual by examining the results[21]. While this approach enables provable security, it is complicated for users to reason about tradeoffs between privacy budget and usability of data.

Secure multi-party computation is a current research area and its intended use case closely matches that of researchers sharing data. This approach provides another track to navigate the intersection of data sharing utility and research participant privacy. The security guarantee with secure multi-party computation (SMC) is that nothing beyond the function output can be learned about the private input of either party. A recent contribution in the SMC space for genomic data analysis shows the promise of this technique for data[22]. This work introduces a "percent revealed" metric which is an appropriate mechanism to differentiate between the guarantees that SMC provides compared to differential privacy or access control.

1.4 Contributions

This dissertation contributes microbiome data analysis software along with data access infrastructure for integrative and comparative studies.

1. Metaviz – Interactive visualization for exploratory analysis of community taxonomic profile data. Metaviz is a web application for visualization of microbiome abundance data. The application can visualize marker-gene or whole metagenome shotgun sequencing data. Metaviz introduces a navigation utility for the taxonomic hierarchy.

2. Metaviz integration with the Human Microbiome Project (HMP) Data Infrastructure. We describe the design and implementation of linking between the HMP Data Portal and Metaviz. Also, we present an analysis of a subset of data from the HMP using Metaviz and *metagenomeSeq*.

3. Microbial community longitudinal and functional profiling visualizations in Metaviz. This work expands the visualizations available in Metaviz for longitudinal data using sparklines as the entries of a heatmap to show trends across the set of features. This work also introduces an interactive filter for community functional profile data using the navigation mechanism in Metaviz, provides a mechanism to import and export taxa of interest, and connects Metaviz to external information sources.

4. Privacy-preserving microbiome analysis using secure computation. In 2015, Franzosa *et al*. showed that it was possible to use microbiome features to identify individuals at different time points in the HMP dataset[23]. This work implements statistical analysis functions using a library for secure multi-party computation. The goal of this project is to allow researchers to compute analyses over shared microbiome abundance matrices without revealing the underlying counts directly.

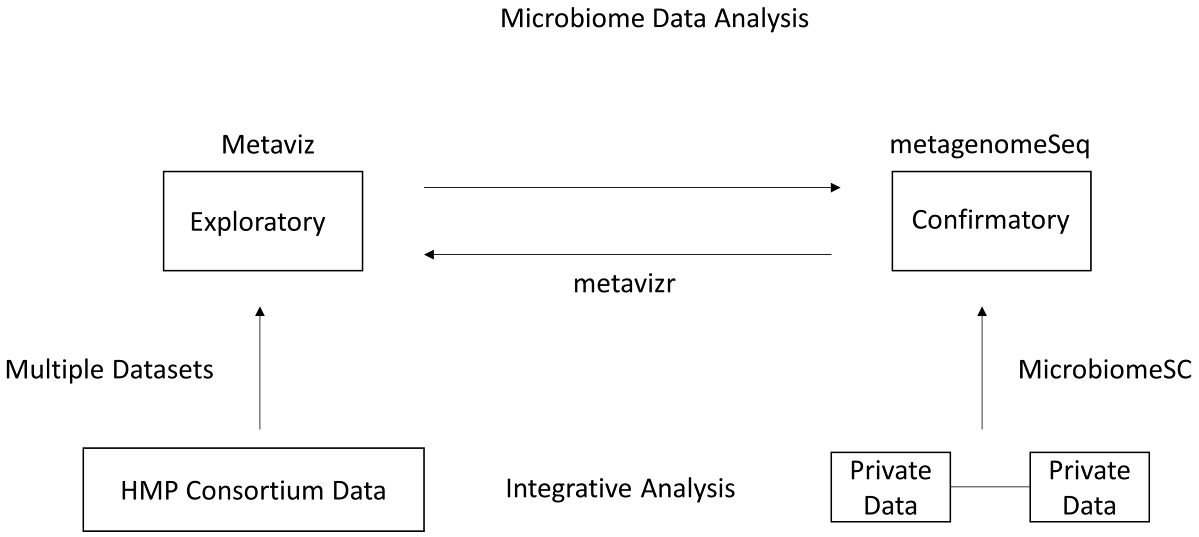


Figure 2: Microbiome Data Analysis Contributions

These contributions address specific aspects of the exploratory and confirmatory data analysis cycle. Figure 2 shows the relationships between contributions to the microbiome data analysis area. Metaviz is an interactive utility for navigating a taxonomic hierarchy and linked quantitative measurement visualizations. Linking Metaviz to the Human Microbiome Project data access center web portal leverages existing data resources to provide the community with a powerful analysis approach. MicrobiomeSC addresses privacy concerns through using a secure data-sharing protocol to ensure data is widely-accessible in the current phase of expanding microbiome research. These contributions fit within the data analysis model championed by Tukey, advance exploratory data analysis with novel visualizations, and ease the burden on researchers of sharing data for confirmatory analysis.

Chapter 2: Metaviz: interactive statistical and visual analysis of metagenomic data

A version of this work is published in Nucleic Acids Research with the following citation: Justin Wagner, Florin Chelaru*, Jayaram Kancherla*, Joseph N Paulson*, Alexander Zhang, Victor Felix, Anup Mahurkar, Niklas Elmqvist, Héctor Corrada Bravo; Metaviz: interactive statistical and visual analysis of metagenomic data, Nucleic Acids Research, Volume 46, Issue 6, 6 April 2018, Pages 2777–2787, https://doi.org/10.1093/nar/gky136.*
This work is joint with: Florin Chelaru, Jayaram Kancherla, Joseph N. Paulson, Victor Felix, Anup Mahurkar, Niklas Elmqvist, and Héctor Corrada Bravo.

2.1 Introduction

High-throughput sequencing of microbial communities provides a tool to characterize associations between the host microbiome and health status, detect pathogens, and identify the interplay of an organism's microbiome with the built environment. Recent highlights include work on the specificity of the human skin microbiome[24], diversity in the ocean microbiome[2], and cataloging the global virome[25]. Effective analysis tools and appropriate statistical models for this type of data are vital to derive and communicate significant insights from these experiments. In other high-throughput sequencing assays, including those for genome[26], transcriptome, and epigenome[27], next-generation genome browsers that integrate exploratory computational and visual analysis have proven to be effective analysis tools. Exploratory analysis tools for microbiome data are scarce however, partially stemming from the challenge that microbiome features, the units of measurement and analysis, are organized in a taxonomic hierarchy. Specifically, while the linear structures of tracks and ranges used in genome browsers provide a natural scheme for navigation in genomic visualization, a hierarchical exploration technique is not readily available. In this paper, we present the Metaviz tool for effective interactive exploration, analysis, and data visualization of hierarchically organized metagenomic features.

Motivation

As an illustrative use case for statistically-guided interactive visualization, we consider a data analysis from the Moderate to Severe Diarrheal (MSD) disease study among children in four countries of the developing world [3]. A typical analysis for this case-control study includes statistical testing to compare taxa abundance between children with and without diarrhea to find novel associations between health and disease. The *metagenomeSeq* Bioconductor package [http://bioconductor.org/packages/release/bioc/html/metagenomeSeq.html] is a popular tool to identify differentially abundant features[16]. In this tool, we target workflows after an abundance matrix has been computed. A standard workflow starts with the data analyst obtaining sequence counts indicating the abundance of annotated operational taxonomic units (OTUs) for each sample in a study with phenotypic and experimental characteristics of these samples available as metadata. The workflow proceeds by the data analyst aggregating counts to a specific level of the taxonomic hierarchy (e.g. species or genus) and obtaining differential abundance inferences by computing log fold changes and p-values for each taxon between case and control groups. She then selects features with a log fold change beyond a given threshold and p-value cutoff as differentially abundant taxa. Next, she visualizes the abundance of these filtered features across samples in a heatmap. After interpreting the plot, she may decide to change the feature selection parameters or further explore the taxonomic hierarchy which requires another iteration of computing the feature set and visualization. In this case, each refinement of statistical analysis parameters produces another visualization with no linking between results.

Our design of the Metaviz application for interactive visualization and analysis makes this workflow much more effective: for instance, once a set of differentially abundant features is selected, the data analyst can interactively visualize abundance data for those specific features. She can then explore the hierarchy of features, aggregate counts to any level of the taxonomy, and identify sub-structures that are difficult to ascertain at lower levels of the taxonomic hierarchy. Further, she may calculate differential abundance at a different level of the hierarchy then dynamically explore these inferences in the same Metaviz workspace, thus streamlining her exploration of a complex set of differential abundance results using statistical and visualization tools.

Related Work

Taxonomer performs both read taxonomic assignment and visualization of results using a sunburst diagram to visualize features[11]. Pathostat is a Shiny application that computes statistical metagenomic analyses, visualizes results, and is integrated with different Bioconductor packages [http://bioconductor.org/packages/release/bioc/html/PathoStat.html]. Pavian is an R package which incorporates Shiny and D3.js components to enable interactive analysis of results for metagenomic classification tools [https://doi.org/10.1101/084715]. Panviz is a tool for exploring annotated pan genome datasets based on D3.js libraries[28]. Krona is a web-based tool for metagenomics visualization that provides a sunburst diagram to navigate the feature space[29]. VAMPS is a web-service that provides a JavaScript and PHP-based metagenomics visualization toolkit of datasets uploaded by researchers[13]. Anvi'o is a multiomics platform that supports analysis using custom JavaScript visualizations[14]. MicrobiomeDB is a web-service that hosts microbiome community taxonomic profile data from open datasets and uses Shiny to visualize data[15].

Encompassing the features of these tools, Metaviz provides a comprehensive interactive visualization environment using JavaScript and D3.js for microbial marker-gene sequencing and whole metagenome shotgun sequencing data with integration to R/Bioconductor. In contrast to these tools, Metaviz uses FacetZoom which is more suited than sunburst diagrams for browsing the hierarchical structure of metagenomic data across many samples by enabling taxonomic feature selection spanning multiple levels of a taxonomy. Further, Metaviz can analyze data from either a database or R which makes it more efficient and scalable than Shiny based tools which are limited by in-memory processing. Metaviz implements the WebSockets protocol directly, which allows for use of data transfer types beyond those specified in Shiny to support flexible and extensible custom JavaScript visualizations.

2.2 Materials and Methods

Metaviz is a web browser-based tool for interactive exploratory microbiome data analysis. It can visualize abundance data served from an interactive R session or query data from a graph database server. Here, we present the architecture of Metaviz from the web-browser application to database storage. A web-browser based application provides flexibility for users and “run anywhere” functionality when deploying the tool. We built upon the D3.js project for an aesthetically pleasing and effective suite of plots and charts. The data back end serves an abundance matrix with taxonomic annotation for features, in our case OTUs, and the front end is a JavaScript application for data visualization. Given the structure of metagenomic data, the user navigation tools and the database storage are tailored to taxonomic hierarchies. We moved from a relational database model used in Epiviz[30], our previous interactive data analysis tool for functional genomic data such as gene expression and methylation data, to a graph database to manage the feature hierarchy and abundance counts. The fundamental operation enabled by this data backend is to efficiently aggregate abundance counts to a specific subset of nodes in the taxonomic hierarchy during interactive exploration.

Visualization layer

Implementing the visualization layer for this application presents several challenges for displaying, navigating, and manipulating data from a feature-rich hierarchy. Design considerations for metagenomic data analysis include: 1) *size of the feature space*; 2) *depth of the feature hierarchy*; and 3) *number of samples*. Given these characteristics, we focused the design of Metaviz on efficient traversal of the feature space and defining feature selections across the taxonomy. In addition, we engineered the navigation tools to be applicable across datasets and persistent between user sessions for collaboration and publication of results.

In Figure 3 we demonstrate the visualization layer of Metaviz on the MSD marker-gene survey dataset. The bottom panel is a navigation control designed to effectively explore the taxonomic feature hierarchy and aggregate count values of features to any set of taxonomic nodes. The top panel consists of a heatmap with the color intensity set as the observed count of a feature (column) in a sample (row). The rows are dynamically clustered based on Euclidean distance of the count vectors for each sample and a dendrogram shows the clustering result. The top panel also includes a PCA plot over all the features of the samples in the heatmap. The stacked bar plots in the second row render, for each sample (column), the proportion of counts for each microbial feature. The separate plots show case (left) or control (right) samples based on dysentery status and the columns are samples grouped by age range. This collection of charts provides multiple views of the same data and is dynamically updated upon user interaction with the navigation tool to achieve exploratory iterative visualization.

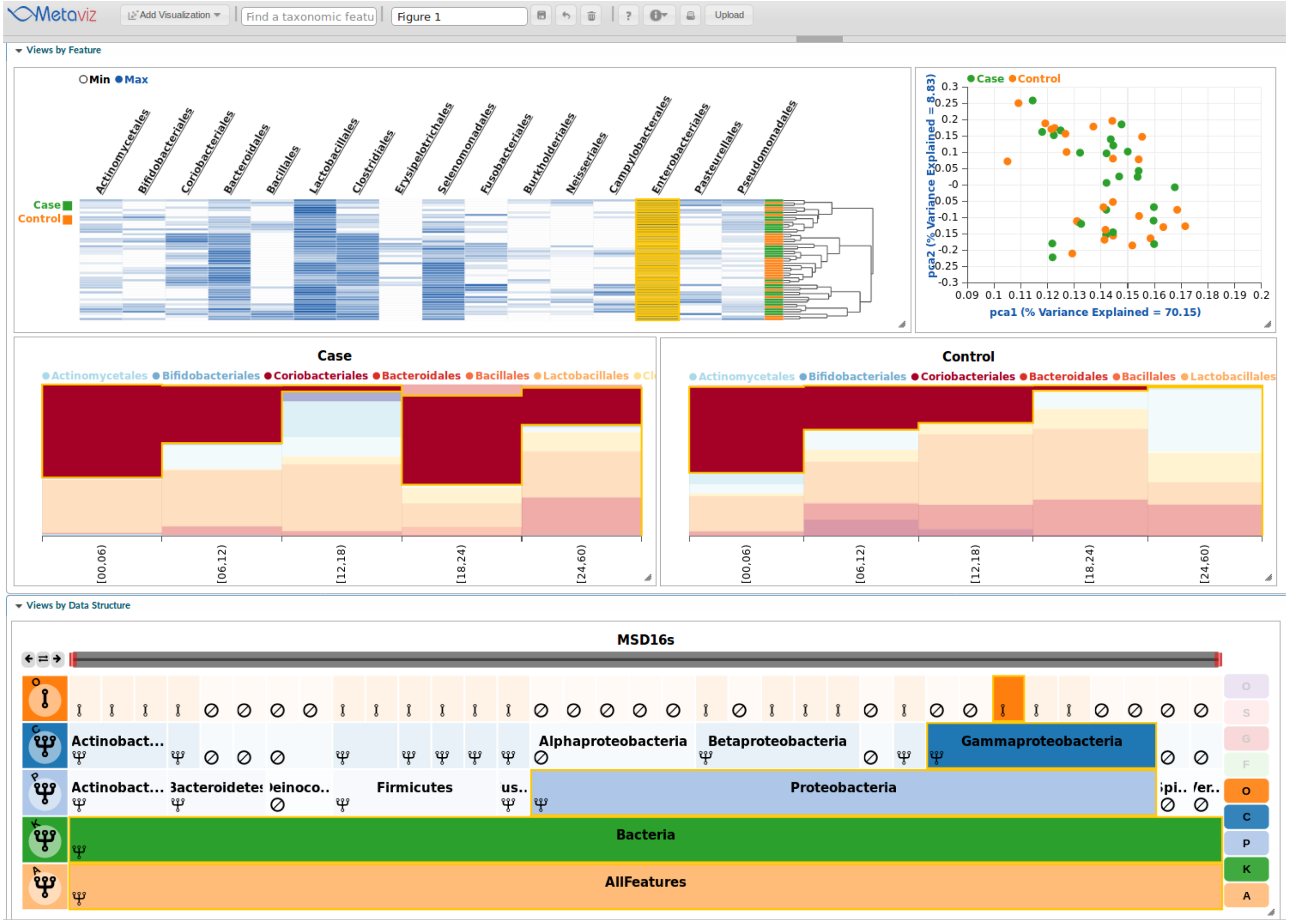


Figure 3: Metaviz interactive visualization of childhood severe diarrhea study

A subset of 50 samples (25 case and 25 control for dysentery) from the Moderate to Severe Childhood Diarrheal Disease study[3]. The FacetZoom control on the bottom panel is used for exploration of the taxonomic organization of metagenomic features. Node opacity in the FacetZoom indicates the set of taxonomic features selected across all appropriate visualizations in the Metaviz workspace. Each node can be in one of three possible states as indicated by the icon in its lower left corner: 1) aggregated, where counts of descendants of this node are aggregated and displayed in other charts, 2), expanded, where counts for all descendants of this node are visualized in other charts, or 3) removed, where this node and all its descendants are removed from all the other charts. The left column of the FacetZoom control indicates the levels of the taxonomy and the overall selection for nodes at each taxonomic level. Hovering the mouse over FacetZoom panels highlights the corresponding features in other charts through brushing. The top left chart is a heatmap showing log-transformed counts with color intensity corresponding to the abundance of that feature (column) in that sample (row). The dynamically computed and rendered row dendrogram shows Euclidean distance hierarchical clustering of samples with color indicating case/control status of each sample. The yellow highlighted column is linked between charts and FacetZoom control through brushing. The top right chart is a PCA plot over all features at the current aggregation level (order). The stacked bar plot on the left of the second row shows proportion of selected features in each case sample (columns) while the right chart shows control samples. In both, sample counts are grouped and aggregated by age range. This is available as a Metaviz workspace at http://metaviz.cbcb.umd.edu/?ws=yA4BWgUOTiq.

Navigation Mechanism - FacetZoom

We developed Metaviz to navigate the complex hierarchical structure of microbiome feature data and perform the visualization tasks of *overview*, *zoom*, and *filter*. We incorporate the FacetZoom[31] design, which visualizes a hierarchy using a tree structure showing a subset of levels at one time. We chose this approach to handle the limitations in the screen size and performance of rendering trees with tens of thousands of nodes. We extended the original FacetZoom design to perform interactive aggregation and removal of microbial lineages. We refer to our navigation tool, shown in the bottom panel of Figure 3, as a FacetZoom control for the rest of the manuscript.

The nodes of the FacetZoom control indicate how the abundance counts for taxonomic features are displayed in the other charts of the Metaviz workspace. Every node of the FacetZoom control can receive mouse-click input from the user. A click on a node sets that feature as the root of a dynamically rendered subtree. Each node can be in one of three possible states as indicated by an icon in its lower left corner: 1) *aggregated*, where counts of descendants of this node are aggregated and displayed as a single feature in other charts, 2) *expanded*, where counts for all descendants of this node are visualized as separate features in other charts, or 3) *removed*, where this node and all its descendants are removed completely from the other charts. The state of a node determines the state of its descendants. Node opacity in the FacetZoom control indicates the set of taxonomic units selected across all appropriate visualizations in the Metaviz workspace. Hovering the mouse over FacetZoom nodes highlights the corresponding features in other charts through brushing as shown in Figure 3. The bottom node of the FacetZoom visualization displays the taxonomic lineage of the corresponding feature at the root of the subtree currently in view.

The FacetZoom control includes a level-wise aggregation indicator panel on left side. Each element of the indicator panel can be used to set the aggregation state of all nodes at a given depth. The letter on each element of the panel identifies the taxonomic level with “P” denoting phylum and “O” signifying order, for instance. The panel on the right provides a persistent global view of the hierarchy to identify where in the full taxonomy the current subtree selected by the user is located. As an example, when the FacetZoom is displaying nodes from class to genus, only these elements are highlighted in the levels indicator panel.

The bar at top of the FacetZoom sets the range of features shown in the other charts in the visualization workspace. The bar is a flexible component with arrows to control movement left or right and expansion over the full range of the current subtree. Updates to the filter bar triggers queries over the count data and those results are automatically propagated to the other charts in the workspace.

As described, the FacetZoom controls which features are included in plots and charts of count data in a Metaviz workspace. We detail our implementation of heatmaps, stacked bar plots, scatter plots, and line plots in Appendix A Materials Section II.

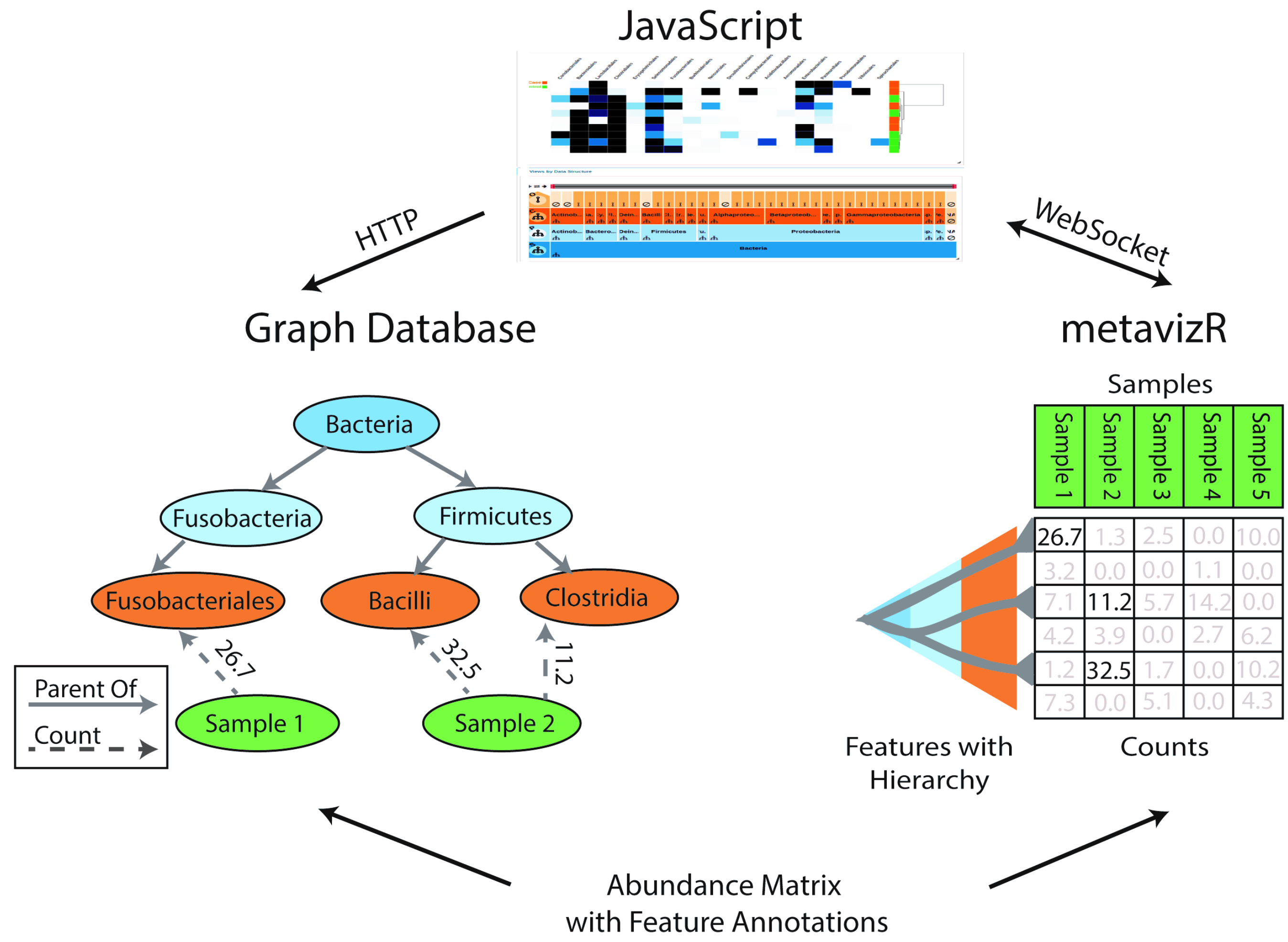
Metaviz supports text-based search for quick navigation to specific taxonomic features. A user can enter the name of a taxonomic feature of interest into a search box on the toolbar. The search provides auto-complete and lists features that contain the character string in a drop-down list. Once a user selects a feature, the navigation bar in the FacetZoom control will update to encompass that feature and all linked data visualizations update as well.

Metaviz includes a dynamic boxplot, created by clicking on column labels of a heatmap, to offer details-on-demand of taxonomic feature count distributions across samples of interest. A box and whisker glyph are created for each sample group selected based on criteria defined over sample metadata criteria. Text-search can also be used within the boxplot to select any feature in the hierarchy and display counts aggregated to that feature.

Data layer

A key difference between microbiome sequencing data and other genomic data is the hierarchical organization of its features, which drives the design of the Metaviz back end. Our data model of microbiome datasets includes the observed counts for each feature in every sample, the hierarchical taxonomic feature annotations and metadata such as phenotypic, behavioral, and environmental information for each sample. A query triggered from user interaction operates over these three data types and computes aggregations on the count data to the specified hierarchy level.

To achieve interactive visualizations with reasonable query response times, we used a graph database architecture. In a graph database, nodes and edges in a graph are objects that can be queried directly. This is a contrast to relational databases in which samples are rows and sample attributes are columns. Each table in a relational database encompasses all the required data fields for the observations in that table while keys handle relationships between tables. We use a graph database to store each taxonomic feature as a node in the graph with edges connecting nodes as specified by the taxonomic information. This system uses a natural representation of the hierarchical organization of this data while avoiding costly join operations in a relational database. We also store samples as nodes and the count value for a feature in a sample is an edge between leaf feature nodes and sample nodes. This graph database structure is shown in Figure 4.



Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
26.7	1.3	2.5	0.0	10.0
3.2	0.0	0.0	1.1	0.0
7.1	11.2	5.7	14.2	0.0
4.2	3.9	0.0	2.7	6.2
1.2	32.5	1.7	0.0	10.2
7.3	0.0	5.1	0.0	4.3

Figure 4: Metaviz query processing and Graph DB structure
Shown are two Metaviz deployment options, which can be used concurrently if desired. In one deployment option (left), the Metaviz JavaScript front end makes requests to a Python application querying a graph database using HTTP. In the other deployment option (right), abundance matrices are loaded into a *metavizr* session which uses the WebSocket protocol to communicate to the JavaScript component, allowing two-way communication between JavaScript and an interactive R session. The graph on the left shows how abundance matrices are stored in the graph database. Nodes in the graph correspond to metagenomic features or samples, edges between metagenomic features denote taxonomic relationship, edges at the leaf level of the taxonomy connect to samples and store the corresponding abundance counts. In either deployment option, aggregation queries are evaluated in response to FacetZoom control selections in the UI and require summing, for each sample, the counts for features in a selected taxonomic subtree.

Materials

We utilized several datasets during the design and testing of Metaviz. The first is the MSD dataset which was gathered from a cohort of 992 children across four countries with an age range of 0-60 months. Fecal samples were gathered from subjects with diarrhea and healthy controls. Specific details for data generation, preprocessing, and annotation are covered in Pop *et al.*[3]. To study time series, we used a longitudinal *E. coli* analysis dataset gathered from 12 participants who were challenged with *E. coli* and subsequently treated with antibiotics. Stool samples were gathered from participants each day starting 1 day pre-infection until 9 days post-infection. Experimental and sample details are available in Pop *et al.*[32]. We benchmarked our system with data from the Human Microbiome Project which is available at the Data Analysis and Coordination website [https://www.hmpdacc.org/hmp/]. We retrieved the data as a prepared *phyloseq* object [http://joey711.github.io/phyloseq-demo/HMPv35.RData] and chose the subset of samples processed at the Washington University Genome Center.

2.3 Results and Discussion

To inform the choice of database architecture, we benchmarked an implementation using a relational database against using a graph database. The relational database uses MySQL [https://www.mysql.com/] as the database management system and PHP [http://php.net/] to handle requests from the web browser client. The graph database configuration uses Neo4j [https://neo4j.com/] and the Flask web development framework [http://flask.pocoo.org/]. In the benchmarks, we deploy our backend services on an Amazon EC2 t2.small instance and used the *wrk* tool [https://github.com/wg/wrk] to send HTTP requests. The testing dataset consisted of 62 samples, 973 features, and 7 hierarchy levels. We observed that the graph database provides approximately 5x lower latency. We also modified the relational design to pre-compute a join operation between the sample, hierarchy, and count tables then store that in the database. This design decreases query response time but increases the size of the database. Compared to this implementation, our graph database implementation showed approximately 50% lower latency. We present our benchmark results in Figure 5.

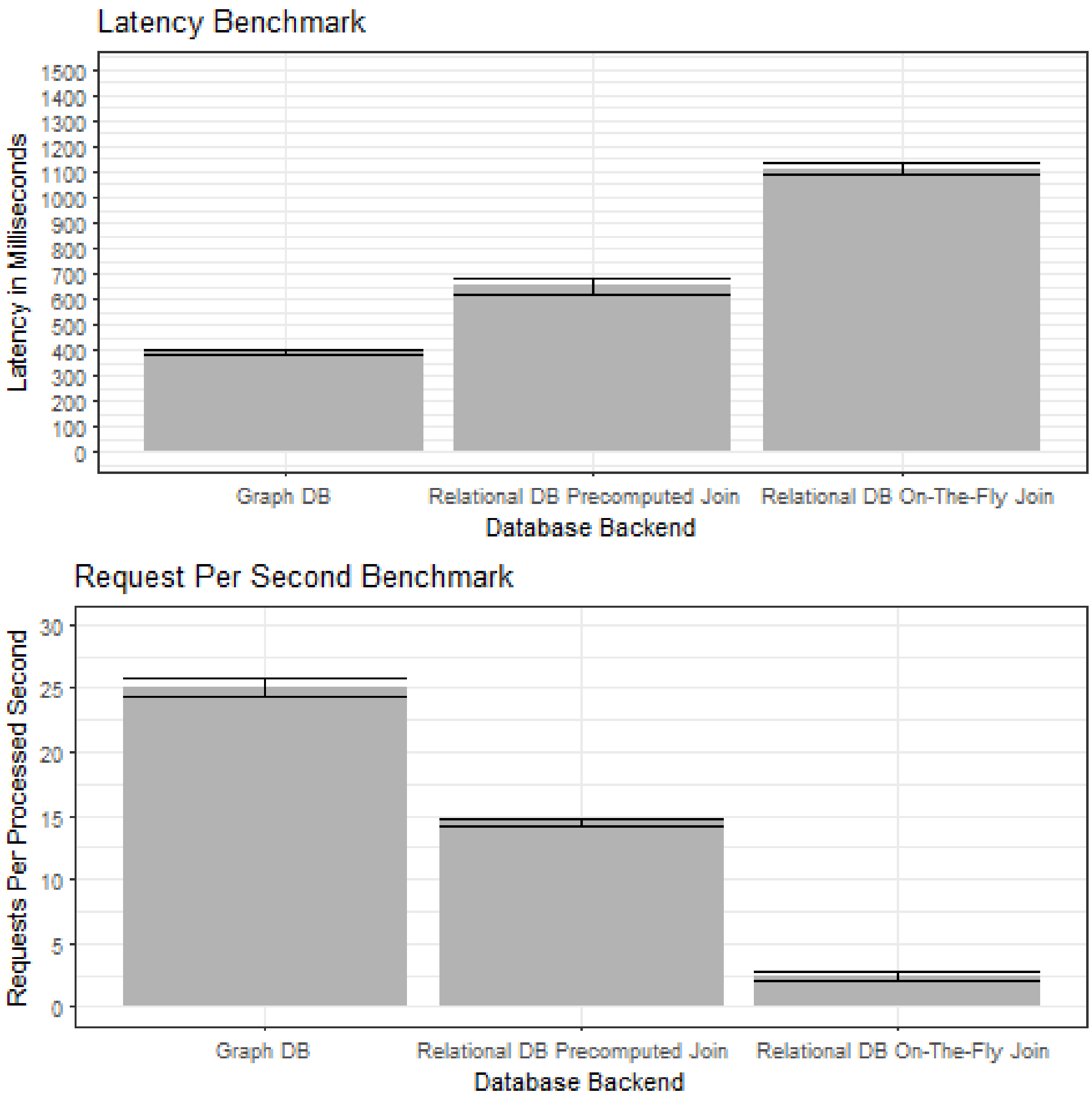


Figure 5: Metaviz database architecture benchmarks
We use the *wrk* tool to benchmark UI requests to three database architectures for storing abundance matrices and feature hierarchies (taxonomies): (1) Graph DB, using Neo4j with a Python Flask web service, (2) Relational DB Pre-computed Join, using a MySQL implementation with a JOIN of the 3 tables of features, values, and samples pre-computed and stored as a table, (3) Relational DB On-The-Fly Join, a MySQL implementation with computing a JOIN across the three tables for each query. For (2) and (3), a PHP application issues queries to the database in response to requests from the UI. We deployed each implementation on an Amazon EC2 t2.small instance and the dataset used across all instances consisted of 62 samples, 973 features, and 7 hierarchy levels. The upper panel shows query latency including standard error across 5 days of measurements. In addition to the latency of processing each request, we also measure the number of requests per second processed providing a measure of throughput in our application. In both performance measures, we see significant benefits of a Python-Neo4j deployment compared to a PHP-MySQL stack for Metaviz tasks.

Whole Metagenome Shotgun Sequencing Data

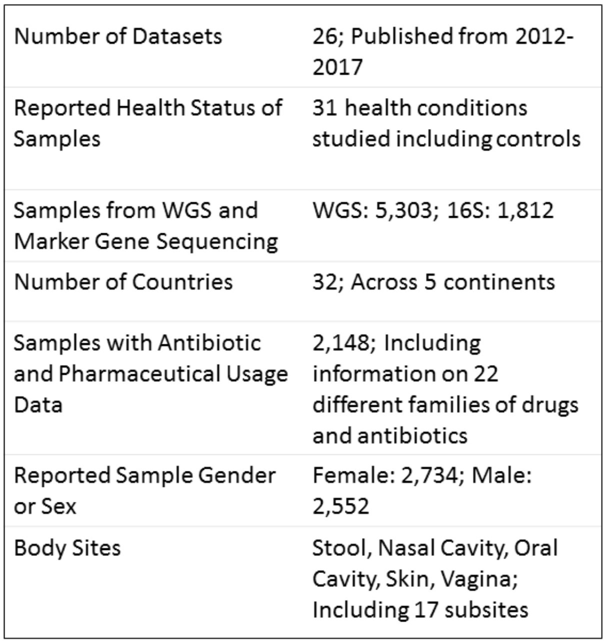
We designed Metaviz to render community taxonomic profile data derived from whole metagenome shotgun sequencing in addition to marker gene sequencing. The results of this sequencing is often reported in relative abundance, which is converted to counts through multiplying by read depth, at inner nodes of taxonomy instead of counts at leaf nodes only in marker gene data[8]. Feature selection queries, or specification of tree cuts, at various levels of the hierarchy do not compute aggregation and instead are directly returned from the inner node counts.

metavizr

Metaviz expands the analysis that can be performed from Bioconductor through the *metavizr* package [https://bioconductor.org/packages/release/bioc/html/metavizr.html]. Interactive visualization of microbiome statistical analysis results allows a user to explore the data at various levels of detail and report those findings in an accessible, aesthetically pleasing interface. *Metavizr* uses the *metagenomeSeq* Bioconductor package to load the feature, count, and sample data into a data object. *Metavizr* communicates with a Metaviz web browser application instance using a WebSocket connection. A FacetZoom control along with data charts and plots can be added to the Metaviz workspace interactively from the R session. A user can specify taxonomic features for visualization from the results of statistical testing as discussed in the *Motivation* section. Metaviz can be used with other Bioconductor packages beyond *metagenomeSeq* for analysis. As an example, we use the *vegan* CRAN package to compute alpha diversity [https://cran.r-project.org/web/packages/vegan/index.html] for microbial community-level analysis. GitHub gists can be used through *metavizr* to modify any plot or chart display setting using JavaScript in addition to customization facilities provided directly by the *metavizr* package itself. Finally, a persistent workspace identifier can be used to reproduce the visual analysis of a collaborator after *metavizr* loads the dataset. To measure the performance of *metavizr*, we benchmarked the memory usage and run-time of aggregation operations using a subset of the Human Microbiome Project dataset, which we describe in the *Methods and Materials* section. We ran the benchmark on an AWS ec2 t2.large instance to simulate the configurations of a typical laptop used for analysis using R/Bioconductor. We present the performance results in Appendix A Figure 2. We found *metavizr* to provide suitably responsive behavior for datasets up to 1,000 samples and 25,000 features and recommend switching to the graph database backend for larger datasets.

UMD Metagenome Browser

We loaded samples from a variety of marker gene and whole metagenome shotgun sequencing studies into the UMD Metagenome Browser – a Metaviz instance hosted by the University of Maryland Center for Bioinformatics and Computational Biology at http://metaviz.cbcb.umd.edu. The whole metagenome shotgun data is from the R/Bioconductor package *curatedMetagenomicData* [https://bioconductor.org/packages/release/data/experiment/html/curatedMetagenomicData.html] which provides curated data from metagenomic studies for dozens of diseases across multiple body sites[33]. A total of 7,115 samples are available from the UMD Metagenome Browser. Figure 6 lists the datasets, sample sites, and descriptions of the available metadata. With the UMD Metagenome Browser, an analyst can choose from the datasets available to complete a study and share results through a persistent Metaviz workspace.



Number of Datasets	26; Published from 2012-2017
Reported Health Status of Samples	31 health conditions studied including controls
Samples from WGS and Marker Gene Sequencing	WGS: 5,303; 16S: 1,812
Number of Countries	32; Across 5 continents
Samples with Antibiotic and Pharmaceutical Usage Data	2,148; Including information on 22 different families of drugs and antibiotics
Reported Sample Gender or Sex	Female: 2,734; Male: 2,552
Body Sites	Stool, Nasal Cavity, Oral Cavity, Skin, Vagina; Including 17 subsites

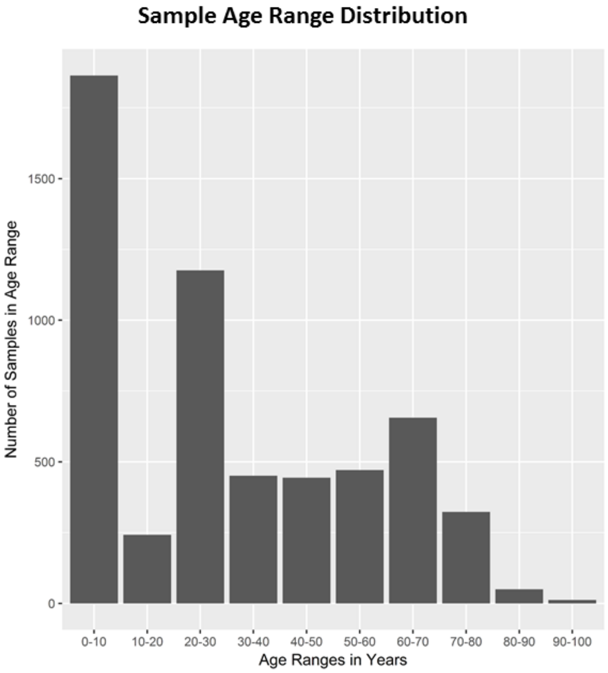


Figure 6: UMD Metagenome Browser Sample Summary
The publicly available Metaviz instance at http://metaviz.cbcb.umd.edu hosts data from several published studies which were generated using marker gene survey and whole metagenome shotgun sequencing. A total of 26 datasets with 7,115 samples across 31 health conditions and 32 countries are available. Host age ranges from 0 months to subjects over 90 years old. Among the metadata available is reported gender or sex of subject, antibiotic or pharmaceutical usage data, and time course measurements.

Deployment

We support two other deployment mechanisms of Metaviz for users to interactively visualize an abundance matrix with hierarchical feature annotations depending on analysis needs. For interactive joint exploratory statistical and visual analysis, data analysts can load the abundance matrices into a Metaviz instance through *metavizr*. Also, we provide Docker [http://www.docker.com] scripts so users can build and deploy containers of the database, load the abundance matrix to the database, and host the web-browser application as an independent Metaviz instance [https://github.com/epiviz/metaviz-docker].

Use Case 1: Exploration of MSD childhood diarrhea study in developing countries

To demonstrate the analysis utility of Metaviz we report on a new analysis of the MSD dataset. To visualize and explore samples, we examined the data from each of the four countries in the study separately and aggregated taxonomic features to the order level. In this analysis, we set case status as those with dysentery and control status as those without blood in stool, meaning that samples with diarrhea and healthy samples are in the control group for dysentery. We chose this analysis to expand upon the work from the author's original investigation, which studied healthy versus diarrhea and dysenteric versus non-dysenteric diarrhea [3]. This analysis is exemplary of case-control studies commonly employed in microbiome data investigation. For our exploration, we used three visualizations, a heatmap, a dynamic boxplot, and two stacked bar plots to identify differences in the microbial communities in case and control across age ranges by country. We created boxplots for details-on-demand of specific taxonomic features based on visual analysis of the heatmap. In the heatmap, row colors were set by dysentery status and each stacked bar plot consisted of the case and control samples for dysentery of each country. We also grouped the samples in the stacked bar plot by age range.

For visual inspection of differential abundance, we ordered each heatmap by dysentery status so that all case and control samples are grouped together. We looked at the heatmap and removed features with low abundance using the FacetZoom control. We then examined each column individually, identifying the number of samples with a feature present and the distribution of samples with high or low intensity. For features of interest, we then created a boxplot by clicking the column label in the heatmap. The boxplot shows the counts aggregated to that feature for case and control dysentery groups. Using these two visualizations of count data, we called the feature as more abundant in case samples, more abundant in control samples, or as no difference in abundance across groups. When we identified a difference in abundance for a feature, we used the FacetZoom to aggregate counts to the next level lower in the hierarchy, restrict the heatmap to show only children of that feature, and updated the boxplot to identify differences in abundance at that level of the hierarchy. We performed this systematic approach to inspect each feature from the order level to the species level. We compared the results of visual analysis by computing the log fold change using *metagenomeSeq* and report those features detected through our visualization process and list the results of statistical testing. When using *metagenomeSeq,* counts were normalized using cumulative sum scaling (with p = 0.75) and binary dysentery status as the variable of interest in the fitFeatureModel method for differential abundance. The threshold for differential abundance was an absolute log fold change of at least 1 and an adjusted p-value < .1 when comparing samples using dysentery status.

Appendix A Figures 3 and 4 show our visual analysis for Bangladesh samples. From the heatmap and boxplot analysis of these samples, the following taxa appear more abundant in the samples with dysentery than the control samples: Actinomycetales, Enterobacteriales, Lactobacillales, Pasteurellales, Pseudomonadales, Micrococcaceae, Enterobacteriaceae, Carnobacteriaceae, Streptococcaceae, Pasteurellaceae, Moraxellaceae, *Rothia*, *Escherichia*, *Shigella*, *Granulicatella*, *Streptococcus*, *Haemophilus*, *Acinetobacter*, *Escherichia coli*, *Escherichia sp. oral clone 3RH-30*, *Granulicatella adiacens*, *Streptococcus equinus*, *Streptococcus mitis*, *Streptococcus parasanguinis*, *Streptococcus salivarius*, *Haemophilus parainfluenzae*, and *Acinetobacter sp. SF6*. Correspondingly, the following taxa appear more abundant in the control samples as compared to the case samples: Coriobacteriales, Bacteroidales, Clostridiales, Coriobacteriaceae, Bacteroidaceae, Porphyromonadaceae, Clostridiaceae, Eubacteriaceae, Lachnospiraceae, Ruminococcaceae, *Collinsella*, *Bacteroides*, *Clostridium*, *Eubacterium*, *Dorea*, *Faecalibacterium*, *Ruminococcus*, *Collinsella sp. CB20*, *Bacteroides fragilis*, *Faecalibacterium prausnitzii*, *Faecalibacterium sp. DJF_VR20*, and *Ruminococcus gnavus*. Examining the stacked bar plots at the order level, Clostridiales exhibits low proportion in the case samples at 0-6 and 6-12 months, a lower level compared to control samples at 12-18 months, and then a similar proportion in both groups for 18-24 and 24-60 months. With the control samples, Bacteroidales shows a greater proportion at all intervals after 0-6 months.

Using *metagenomeSeq*, we find the following taxa to have significant difference in abundance for Bangladesh samples: Enterobacteriales (log fold change = 1.38, adjusted p-value = 1.46E-04), Pasteurellales (2.47, 4.16E-12), Coriobacteriales (-1.38, 9.88E-04), Bacteroidales (-1.19, 7.56E-04), Clostridiales (-1.09, 6.45E-04), Enterobacteriaceae (1.37, 2.26E-04), Carnobacteriaceae (1.52, 3.23E-05), Streptococcaceae (1.41, 5.00E-05), Pasteurellaceae (2.46, 1.43E-11), Coriobacteriaceae (-1.37, 1.95E-03), Bacteroidaceae (-1.09, 1.16E-02), Ruminococcaceae (-1.09, 3.17E-03), *Escherichia* (1.33, 6.50E-04), *Granulicatella* (1.51, 8.29E-05), *Streptococcus* (1.33, 2.91E-04), *Haemophilus* (2.42, 6.12E-11), *Collinsella* (-1.48, 3.89E-03), *Bacteroides* (-1.08, 2.27E-02), *Ruminococcus* (-1.18, 3.89E-03), *Escherichia coli* (1.33, 1.71E-03), *Granulicatella adiacens* (1.51, 1.92E-03), *Streptococcus mitis* (1.16, 1.50E-02), *Streptococcus parasanguinis* (1.07, 1.71E-03), *Streptococcus salivarius* (1.02, 2.11E-02), *Haemophilus parainfluenzae* (2.26, 3.04E-07), *Collinsella sp. CB20* (-1.26, 3.68E-02), and *Ruminococcus gnavus* (-1.18, 3.48E-02). We present the results for visual analysis and *metagenomeSeq* differential abundance calculation for each country in Appendix A Tables 1-4 and in Section III of Appendix A Materials.

The previously published analysis of dysenteric versus non-dysenteric diarrhea grouped samples from all countries and identified OTUs associated with dysenteric stool, including those from the following taxa: *Haemophilus*, *Streptococcus*, *Granulicatella*, *Escherichia coli*, and *Enterobacter cancerogenus* [3]. While using the heatmap, boxplot, and FacetZoom control to explore each country we observed greater abundance in case samples for *Haemophilus* in Bangladesh, The Gambia, Mali, and Kenya; *Streptococcus salivarius* in Bangladesh; *Granulicatella* in Bangladesh and The Gambia;

Escherichia coli in Bangladesh and The Gambia; and *Enterobacter cancerogenus* in Kenya. Examining results across all countries, three taxa showed greater abundance among case samples through visual inspection and were statistically significant using *metagenomeSeq*: Pasteurellales, Pasteurellaceae, and *Haemophilus*.

Features that showed statistically significant difference in abundance in more than one country but not all are Enterobacteriales and Enterobacteriaceae in Bangladesh and The Gambia. Some features with differential abundance in only one country include Coriobacteriales, Bacteroidales, Coriobacteriaceae, *Collinsella*, *Ruminococcus*, *Collinsella sp. CB20*, *Ruminococcus gnavus*, and *Streptococcus parasanguinis* in Bangladesh. A literature examination revealed that *Ruminococcus gnavus* has been identified as present in patients with Crohn's Disease that relapsed six months after surgical treatment [34]. *Streptococcus parasanguinis* has been identified as having higher relative abundance in cancers of the gastric body in patients without *Helicobacter pylori* infection [35]. In samples from The Gambia, Actinomycetales is more abundant in case than control which is notable given that *Tropheryma whipplei* is the only identified enteric pathogen in the order [36],[37] and that was not identified as differentially abundant by either visual analysis or statistical testing. It is important to note that for The Gambia, Kenya, and Mali, samples without dysentery outweighed those with dysentery.

Use Case 2: Analysis of longitudinal metagenomic studies

Another use case of Metaviz is the analysis of longitudinal metagenomic datasets. We followed the analysis using smoothing spline ANOVA as described in Paulson *et al*. [https://doi.org/10.1101/099457] for a longitudinal dataset characterizing host response to a challenge with enterotoxigenic *E. coli* [32]. The *metagenomeSeq* Bioconductor package provides the *fitMultipleTimeSeries* function for fitting a smoothing spline and performing SS-ANOVA testing. Using *fitMultipleTimeSeries*, the formula considered if diarrhea developed at any day as well when antibiotics were given to the individual. To visualize the results, we use a line plot with time points on the X-axis, log fold change on the Y-axis, and each line representing a taxonomic feature. The FacetZoom is linked to the line plot and the path through the hierarchy is highlighted when hovering over a given line. We also created a stacked line plot of counts aggregated to the species level for those species that were found to be differentially abundant for an interval of at least 2 days using the SS-ANOVA model. Figure 7 shows the Metaviz workspace for this analysis with the spline plot on the top, one sample with diarrhea on the left and one sample without diarrhea at any day on the right. We chose one pair because antibiotics were administered on different days across samples therefore averaging counts across case and control groups is not representative of response for the treatment applied. Each column in the stacked line plots represent the measurement taken at the day since infection. Antibiotics were administered at days 3 through 5 for the case sample and days 4 through 6 for the control sample. Examining the stacked plots *Bacteriodes plebeius* shows high proportion in the case sample on the day after antibiotics are administered then a decrease two days after treatment was complete to a similar level as in the control sample. This procedure can be generalized to time series analysis of microbiome data when investigating differential abundance across time points.

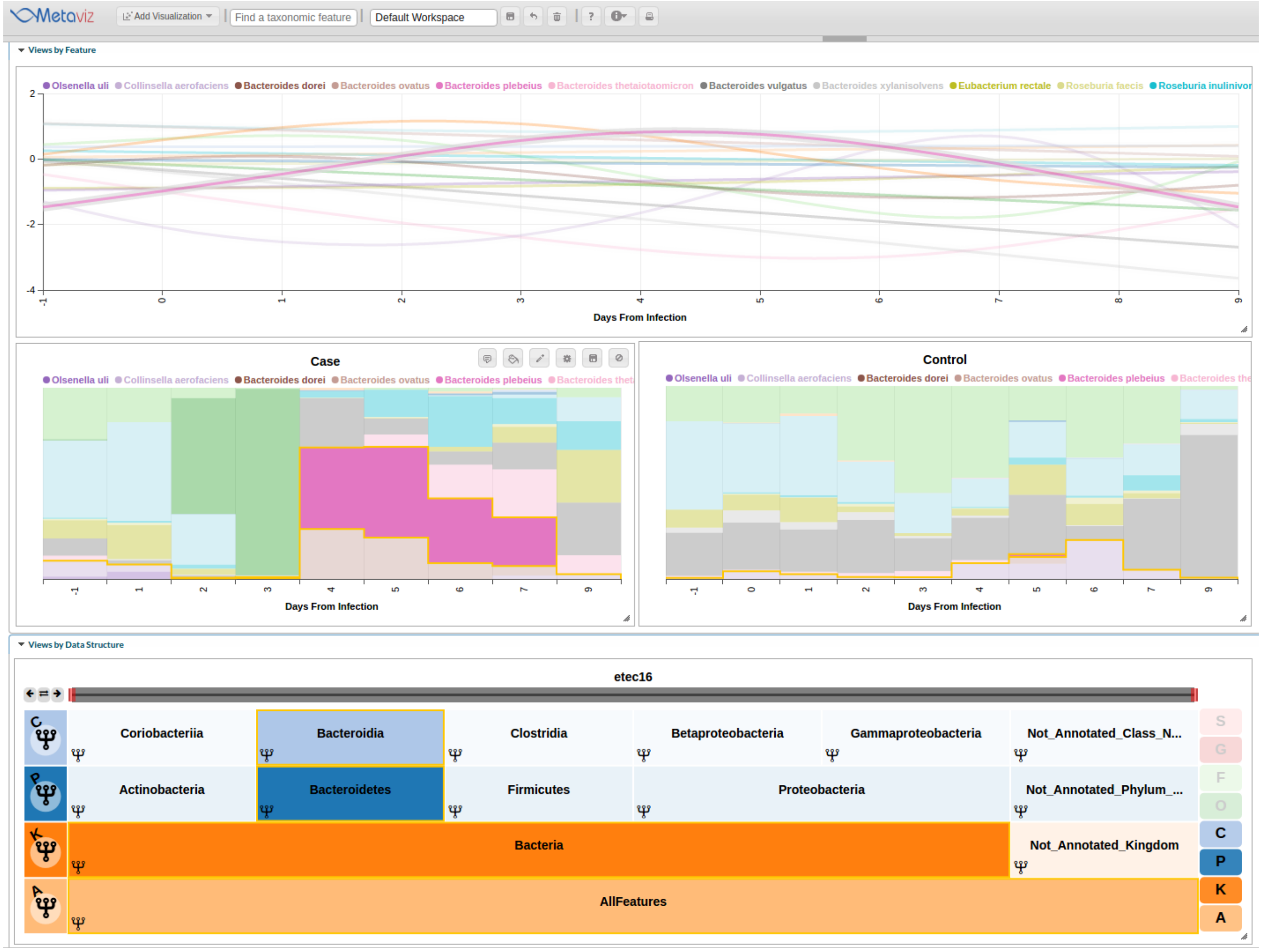


Figure 7: Interactive visualization of smoothing spline differential analysis of longitudinal study
We use Metaviz to explore a longitudinal analysis of the dataset from an enterotoxigenic *E. coli* study [32]. Count data was aggregated to the species level and a smoothing-spline ANOVA model was fit using the *fitTimeSeries* function of the *metagenomeSeq* Bioconductor package. Features with a statistically significant interval of 2 days or longer as estimated by the smoothing spline model at any time point were selected for visualization. The line plot is linked via brushing with the FacetZoom control and a stacked plot showing feature count proportions for a sample that developed diarrhea and a sample with no diarrhea.

2.4 Conclusion

In this paper, we presented the design and performance of Metaviz, a web-browser based interactive visualization and statistical analysis tool for microbiome data. We described design decisions for operating over abundance matrices with tens of thousands of features, thousands of samples, and complex feature hierarchies. We use a graph database for storing community abundance profile matrices as the features have a hierarchy derived from taxonomic databases. We also developed the *metavizr* Bioconductor package providing tight integration of the Metaviz interactive visualization tool and computational and statistical analyses using Bioconductor packages. We used Metaviz to analyze existing datasets and our results highlight the power of interactive visualization coupled with complementary statistical analysis to examine microbiome data. A major contribution of this work is the navigation utility that adapts information visualization techniques to effectively explore and manipulate the rich feature hierarchy of metagenomic datasets. Another significant contribution is the UMD Metagenome Browser web service available to host abundance matrices that allows researchers to explore and share results. We expect that Metaviz will prove useful for researchers in analyzing microbiome sequencing studies as genome browsers have for genomic data.

An avenue for continued research in this area is robust visualization of whole metagenome shotgun sequencing data. This will involve both navigation of the feature taxonomy tree as well as exploration of specific genes for each bacterial feature. This will be a useful visualization as strain level analysis of metagenomic datasets will likely be essential for research and clinical applications. Also, functional annotations could be incorporated to explore associations with host health status. These features could be examined alongside metabolome data to inspect interactions and identify the associations between microbiome community abundances and host cellular processes.

Chapter 3: Interactive Exploratory Data Analysis of Human Microbiome Project Phase II Data Using Metaviz

This work is currently in preparation for submission to the appropriate venue. This work is joint with Jayaram Kancherla, Domenick Braccia, James Matsumura, Victor Felix, Jonathan Crabtree, Anup Mahurkar, and Hector Corrada Bravo.

3.1 Introduction

Metagenomics allows researchers to perform a microbial community census and identify associations between host phenotype and community status. Metagenomics has been used successfully to track pathogen spread[38] and identify intervention strategies in childhood malnutrition[39]. Integrative analysis of samples using multiple sequencing technologies allows for comparison at various levels of granularity. The second phase of the Human Microbiome Project (HMP2) offers a unique opportunity to test hypotheses of interactions between the microbial community in humans and disease. We use Metaviz, an interactive microbiome exploratory data analysis tool, to examine this dataset.

In this work we describe infrastructure to connect Metaviz with the HMP2 Data Coordination Center web portal. We also describe analyses using both Metaviz and a statistical testing package for differential abundance analysis, *metagenomeSeq*, in illustrative use cases with the HMP2 data collection. We perform exploratory analysis with Metaviz and confirmatory analysis with *metagenomeSeq* on two datasets from HMP data portal. These studies demonstrate the usefulness of a combined approach to accessing and analyzing data from this resource. Our examples show that users can share findings and interpretations with visualizations in Metaviz and the HMP data resources.

Human Microbiome Project Phase II

The second phase of the HMP, also called the Integrative Human Microbiome Project, consisted of focused studies of three diseases – Inflammatory Bowel Disease (IBD), Type II Diabetes (T2D), and Pre-term Pregnancy (PTB)[40]. The overall goal of the project was to identify associations between human microbiome community census data and the three diseases. Each of the studies were structured for that disease and involved separate subject cohorts.

Metaviz

Metaviz is a web-based interactive visualization tool for microbiome data analysis. The architecture consists of a JavaScript and D3.js-based front-end suite of charts and a navigation component that shows a subset of taxonomic hierarchy levels at one time. Metaviz supports two backend data stores – a graph database and the *metavizr* R/Bioconductor package. Metaviz is tightly integrated with the *metagenomeSeq* statistical testing package so differential abundance testing results can be viewed directly in a Metaviz session. We host an instance of Metaviz that we call the UMD Metagenome Browser [http://metaviz.cbcb.umd.edu].

Related Work

Visualization tools for large-scale sequencing consortium projects provide a mechanism to explore and interact with data from multiple studies. These applications help users analyze individual datasets and examine trends across the entire project. MAGI is a web-application that enables a user to examine data from TCGA data[41]. The Earth Microbiome Project provides an interactive visualization web-application to analyze its data[42]. EMPeror offers interactive 3D visualizations of PCA plots to show distances between microbiome samples[43]. QIIME packages a number of tools for static plotting of Principal Coordinate Analysis and stacked bar plots[44]. Metaphlan2 uses a visualization package called GraphPhlan to produce phylogenetic trees and other plots[45].

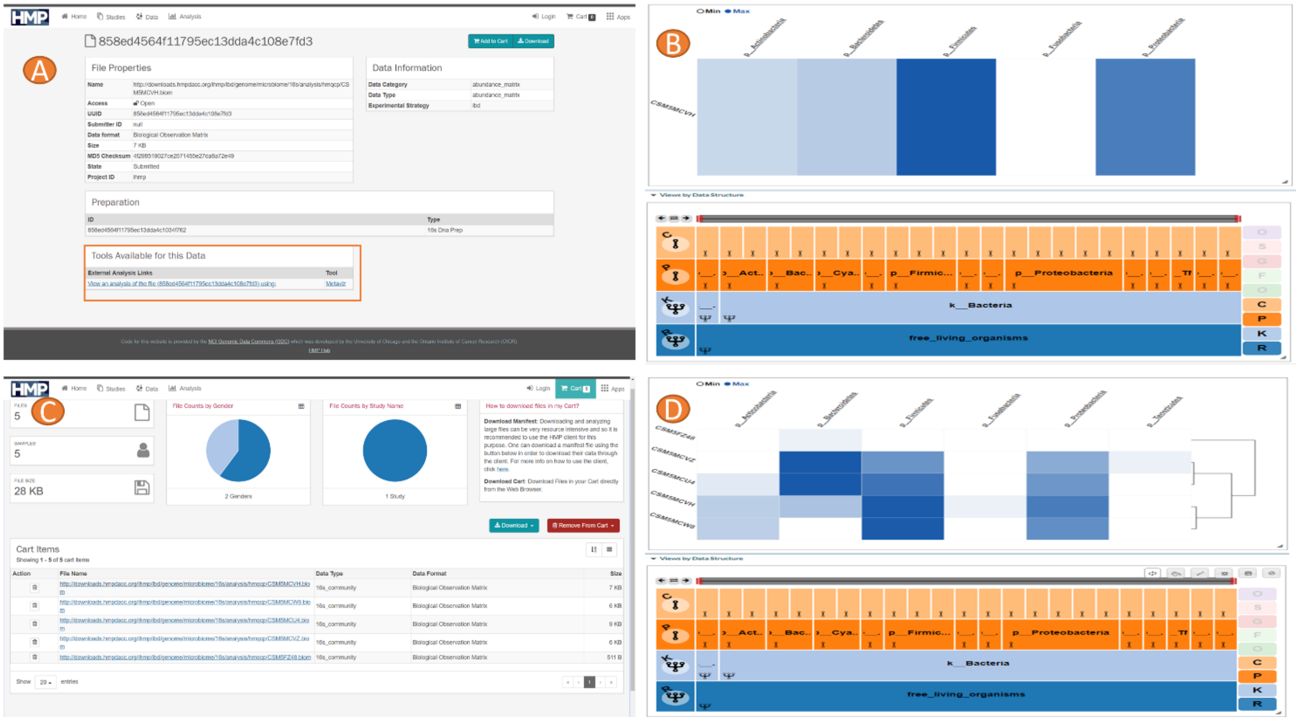
3.2 Metaviz integration with HMP infrastructure

The HMP Data Access and Coordination maintains a repository and web portal [http://ihmpdcc.org]. From this web portal, users can browse metadata for datasets, raw sequencing files, and processed files including taxonomic community profile abundance matrices. We implemented several mechanisms to interact with the HMP data resources through Metaviz.

Data loaded into UMD Metagenome Browser

We loaded the 16S community profile abundance matrices for the samples from the IBD, T2D, and PTP studies into the UMD Metagenome Browser [http://metaviz.cbcb.umd.edu]. A user can select each dataset from the application start screen. Figure 8 details the number of samples from each project currently available in the UMD Metagenome Browser.

Dataset Name	Number of Samples
IBD Stool Pilot	51
IBD Biopsy HMP2	154
PTB	6896
T2D	1900



858ed4564f11795ec13dda4c108e7fd3

File Properties

Name	http://downloads.hmpdacc.org/ihmp/ibd/genome/microbiome/16s/analysis/hmqcp/CSM5MCVH.biom
Access	Open
UUID	858ed4564f11795ec13dda4c108e7fd3
Submitter ID	null
Data format	Biological Observation Matrix
Size	7 KB
MD5 Checksum	4f288518027ce2071455e27da8a72e49
State	Submitted
Project ID	ihmp

Data Information

Data Category	abundance_matrix
Data Type	abundance_matrix
Experimental Strategy	ibd

Preparation

ID	Type
858ed4564f11795ec13dda4c1034f762	16s Dna Prep

Tools Available for this Data

External Analysis Links	Tool
View an analysis of the file (858ed4564f11795ec13dda4c108e7fd3) using:	Metaviz

Code for this website is provided by the NCI Genomic Data Commons (GDC) which was developed by the University of Chicago and the Ontario Institute of Cancer Research (OICR).
HMP Hub

How to download files in my Cart?

Download Manifest: Downloading and analyzing large files can be very resource intensive and so it is recommended to use the HMP client for this purpose. One can download a manifest file using the button below in order to download their data through the client. For more info on how to use the client, click here.

Download Cart: Download Files in your Cart directly from the Web Browser.

Cart Items
Showing 1 - 5 of 5 cart items

Action	File Name	Data Type	Data Format	Size
🗑	http://downloads.hmpdacc.org/ihmp/ibd/genome/microbiome/16s/analysis/hmqcp/CSM5MCVH.biom	16s_community	Biological Observation Matrix	7 KB
🗑	http://downloads.hmpdacc.org/ihmp/ibd/genome/microbiome/16s/analysis/hmqcp/CSM5MCW8.biom	16s_community	Biological Observation Matrix	6 KB
🗑	http://downloads.hmpdacc.org/ihmp/ibd/genome/microbiome/16s/analysis/hmqcp/CSM5MCU4.biom	16s_community	Biological Observation Matrix	9 KB
🗑	http://downloads.hmpdacc.org/ihmp/ibd/genome/microbiome/16s/analysis/hmqcp/CSM5MCVZ.biom	16s_community	Biological Observation Matrix	6 KB
🗑	http://downloads.hmpdacc.org/ihmp/ibd/genome/microbiome/16s/analysis/hmqcp/CSM5FZ48.biom	16s_community	Biological Observation Matrix	511 B

Show 20 entries

Figure 8: Metaviz and HMP 2 Data Infrastructure Integration
Top: UMD Metagenome Browser data. Middle: Single sample link from data portal to UMD Metagenome Browser. Bottom: Multiple samples manifest file upload and selection to UMD Metagenome Browser. We provide

several mechanisms to access the HMP dataset from Metaviz. First, we loaded the 3 datasets (IBD, T2D, and PTB) into the hosted instance of Metaviz directly. A user can choose any of these datasets from the data selections screen then samples can be chosen within each dataset. We also link to the HMP Data Portal for single samples as shown in the Middle panel. Finally, the HMP Data Portal provides a "cart" functionality where a user can select multiple samples and download a manifest listing those files. A user can upload a manifest file containing selections from the 16S community abundance profiles from the same dataset (IBD, T2D, or PTB) to the UMD Metagenome Browser and a new Metaviz workspace is created with those files.

HMP Data Portal linking to Metaviz

When browsing the files available from the HMP Data Portal, a user can view an individual abundance matrix in Metaviz using a link from the file description page. When the user clicks the link, a redirect occurs to the UMD Metagenome Browser with a new workspace containing a FacetZoom navigation utility and a heatmap for that file. Figure 8 shows the direct link functionality.

Metaviz import of Data Portal Manifest

In the HMP data portal, a user can select files with a shopping cart utility and download the selections as a manifest file. In the UMD Metagenome Browser, the user can upload the manifest file to create a Metaviz workspace on the fly for those samples. Currently, only files from the same project can be viewed in one workspace. Resolving taxonomic hierarchies across datasets in Metaviz is future work that could use a utility such as the *metagenomeFeatures* R/Bioconductor package[46]. Figure 8 shows the manifest file workflow.

metavizr analysis of WGS vs 16S data from same samples

In the IBD cohort of the HMP2 dataset, a subset of samples was sequenced using whole metagenome and 16S sequencing. We developed functions in *metavizr* to compare 16S and whole metagenome data for individual samples. Using the taxonomic profiles of the IBD samples, we matched the taxonomic features discovered with both sequencing methods. With this subset of features, we generated a single taxonomic hierarchy then loaded the 16S and whole metagenome abundance measurements into a *metavizr* object.

We presented this utility at a training workshop hosted by the Institute for Genome Science called the HMP Cloud Workshop. The workshop organizers developed a data analysis toolkit named Chiron [https://github.com/IGS/Chiron] for operating on microbiome in a cloud environment. We incorporated *metavizr* and Metaviz into Chiron. We created stacked plots and scatter plots that link to a single FacetZoom to compare the data from each sequencing method. Figure 9 shows an example analysis.

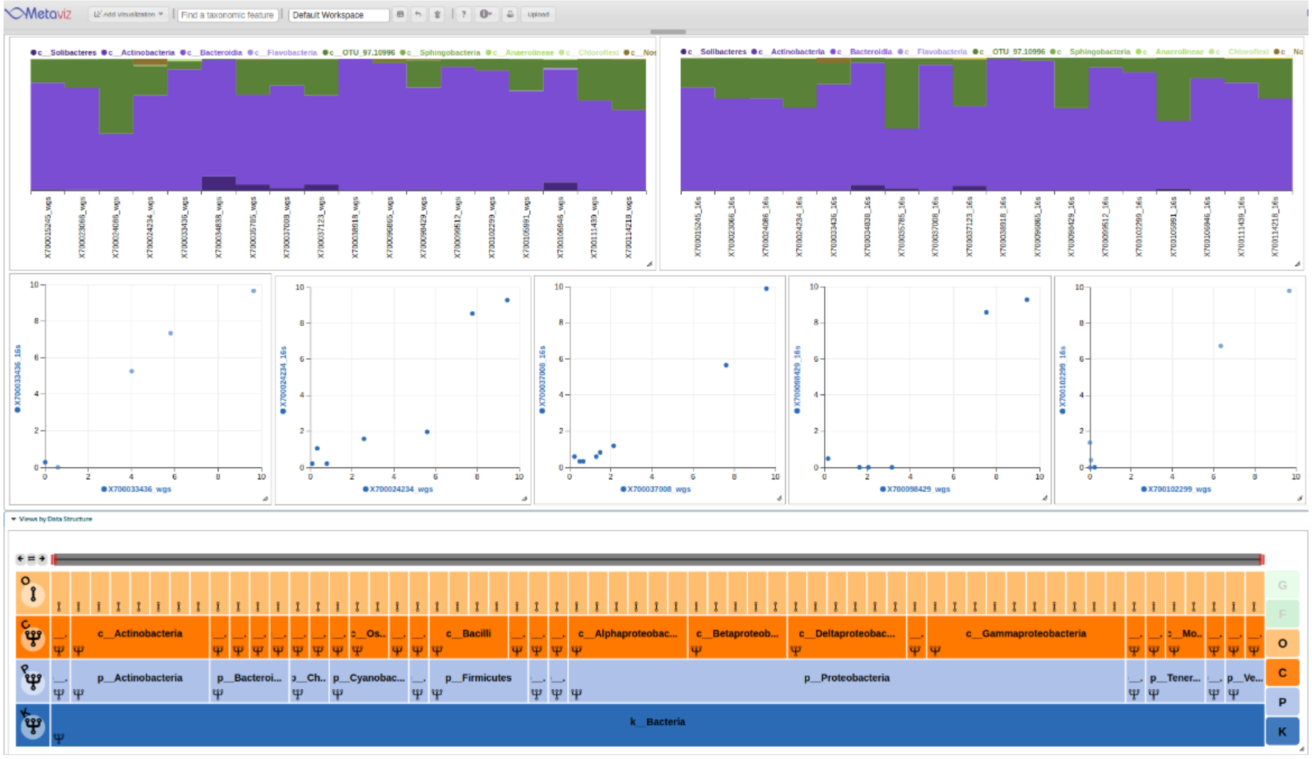


Figure 9: Comparison between 16S and WGS taxonomic profiling using metavizr
We identified taxa present in the taxonomic hierarchy for each method and created a merged dataset. A FacetZoom shows the common features, two Stacked Plots show the proportion of all features aggregated to the Order level, and a set of scatter plots for samples with WGS abundance on the X-axis and 16S abundance on the Y-axis. For WGS, the relative proportion output from MetaPhlan for taxa at the order level are multiplied by read depth. The scatter plots show the variability in taxonomic community census estimates between sequencing methods. A stacked plot visualization is shown in the main HMP consortium manuscript at the genus and species level across samples[1]. We allow a user to make specific selections of the FacetZoom to compare taxa at various levels. The scatter plot also allows resolution at the single sample.

Metaviz Usability Testing

We developed Metaviz based on input from researchers with expertise in interactive genomic visualization and microbiome association testing. The initial Metaviz prototypes identified interactive exploratory microbiome visualization needs and mapped out solutions[47]. Following design and implementation, we presented Metaviz in several public presentations. Two of these workshop presentations included extensive user interaction with the Metaviz application by audience members – one for the HMP Cloud Workshop and another at the annual conference for Bioconductor – BioC 2017. For the HMP Cloud Workshop we created a step-by-step tutorial in Chiron and instructed the 50 attendees to use Metaviz with a subset of the HMP data. We also demonstrated how to use the *metavizr* package to perform analyses as shown in Figure 9. While audience members completed the tutorial, we informally tracked user progress and asked a subset of users afterwards about overall usability. With this feedback we updated the tutorial and data selection mechanisms. We next presented a workshop tutorial at BioC 2017 with a dataset from the *curatedMetagenomicData* Bioconductor package. Through these informal user sessions, we determined the interactive data visualizations and FacetZoom navigation utility were useful for exploration of the taxonomic community profile data. We leave as future work a formal user study to identify the needs of the HMP community and areas of improvement for Metaviz visualization and navigation utilities. In these studies, we plan to measure the effort to perform a given set of tasks and identify new visualizations users want for emerging microbiome data types.

3.3 Methods and Results

Inflammatory Bowel Disease Dataset

The IBD study consisted of two phases, a pilot which we refer to in this work as the IBD Stool Pilot and a larger phase that we call IBD HMP2. We use the taxonomic profiles for each phase available from the IBD project website [www.ibdmdb.org] and use the same taxonomic classification identifiers reported in those results. We also used the ‘ExternalID’ field as a unique identifier for samples. We first loaded the taxonomic profile results for each subset into *metagenomeSeq* objects and performed filtering resulting in the IBD Stool Pilot with 51 samples and the IBD HMP2 with 154 samples. We used Metaviz for exploratory analysis and *metagenomeSeq* for confirmatory statistical testing. We examined each dataset separately and used a local copy of Metaviz with each data subset loaded.

IBD Stool Pilot

The IBD Stool Pilot dataset contains 16S and whole metagenome sequencing results of stool samples from 41 Crohn's Disease subjects and 10 Ulcerative Colitis subjects. For the analysis, we use Metaviz to visually identify taxa that showed a difference in abundance between Crohn's Disease and Ulcerative Colitis subjects. Figure 10 shows a typical visualization and Appendix B Table 1 lists the visual analysis results.

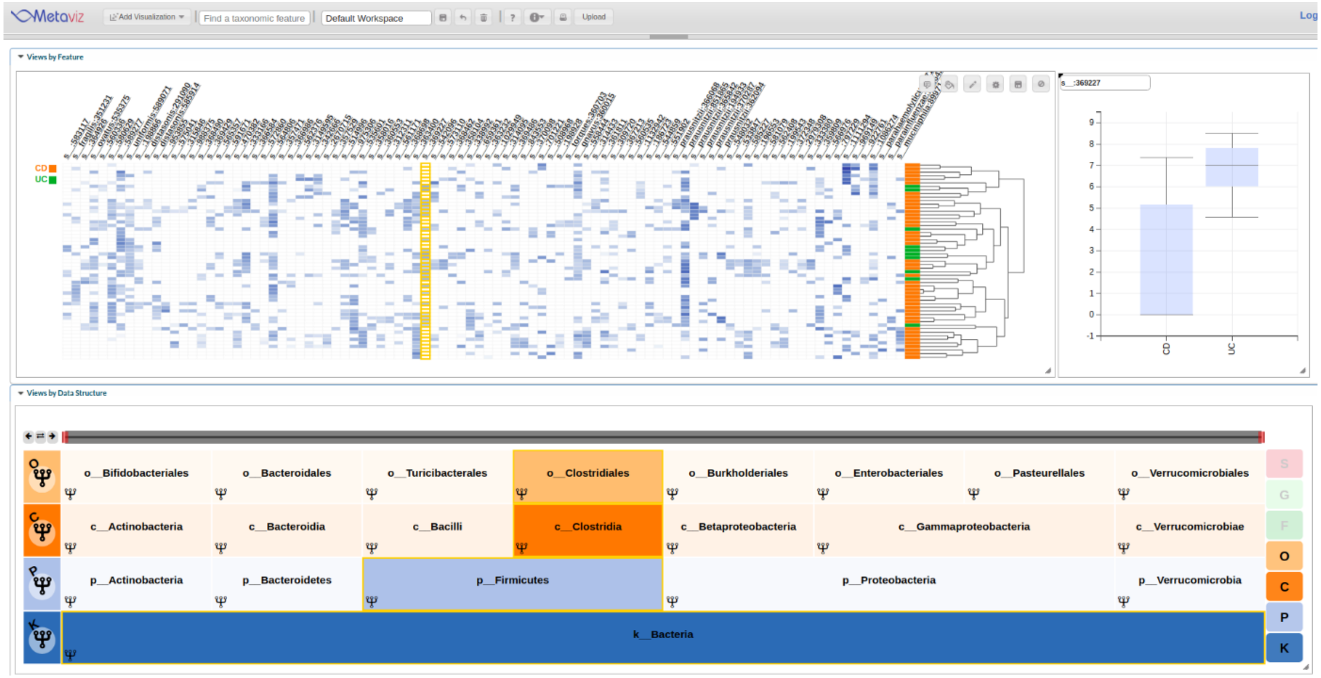


Figure 10: Metaviz Analysis of IBD Stool 16S Pilot Dataset
A Metaviz workspace with a FacetZoom taxonomic hierarchy, heatmap, and boxplot for the specific feature in this instance s__:369227. We identified features at each level of the hierarchy using this integrative view and the results for features with a potential differential abundance are listed in Appendix B Table 1.

We also used *metagenomeSeq* to test the abundance of features aggregated to each level of the taxonomy using the fitFeatureModel method. As shown in Table 1, two species had an absolute log fold-change greater than 1 and adjusted p-value less than .1. Comparing the visual analysis results in Appendix B Table 1 and the *metagenomeSeq* differential abundance testing results in Table 1 shows that the taxonomic feature s__:369227 was identified using both methods.

Table 1: metagenomeSeq analysis of IBD Stool 16S Pilot Dataset

	Log fold change	se	p-value	Adjusted p-value
s__:369227	1.864583442	0.431193725	1.53061E-05	0.000734694
s__:363232	1.193035074	0.275415013	1.47914E-05	0.000734694

We used the fitFeatureModel of *metagenomeSeq* and aggregated counts to each level of the taxonomic hierarchy. Our analysis identified s__:369227 under family *Lachnospiracea* and s__:363232 under genus *Dorea* as differentially abundant between samples from subjects diagnosed with Ulcerative Colitis and Crohn's Disease.

IBD HMP2

The IBD HMP2 dataset consists of 75 samples from Crohn's Disease (CD) subjects, 37 samples from Ulcerative colitis (UC) subjects, and 42 samples from subjects without IBD (nonIBD). For these samples, we analyzed the 16S sequencing data of an intestinal biopsy. In our analysis, we first investigated if any taxonomic features showed a difference in abundance between the three groups. Figure 11 shows an example using Metaviz for the visual inspections. We list the taxa that we found as different between groups in Appendix B Table 2. We compute an F-statistic to determine if any taxonomic feature is associated with at least one group. Currently, fitFeatureModel does not support model matrices with more than 2 columns, so we used the fitZig method and constructed contrasts for testing between groups. Appendix B Table 3 lists the results from this analysis.

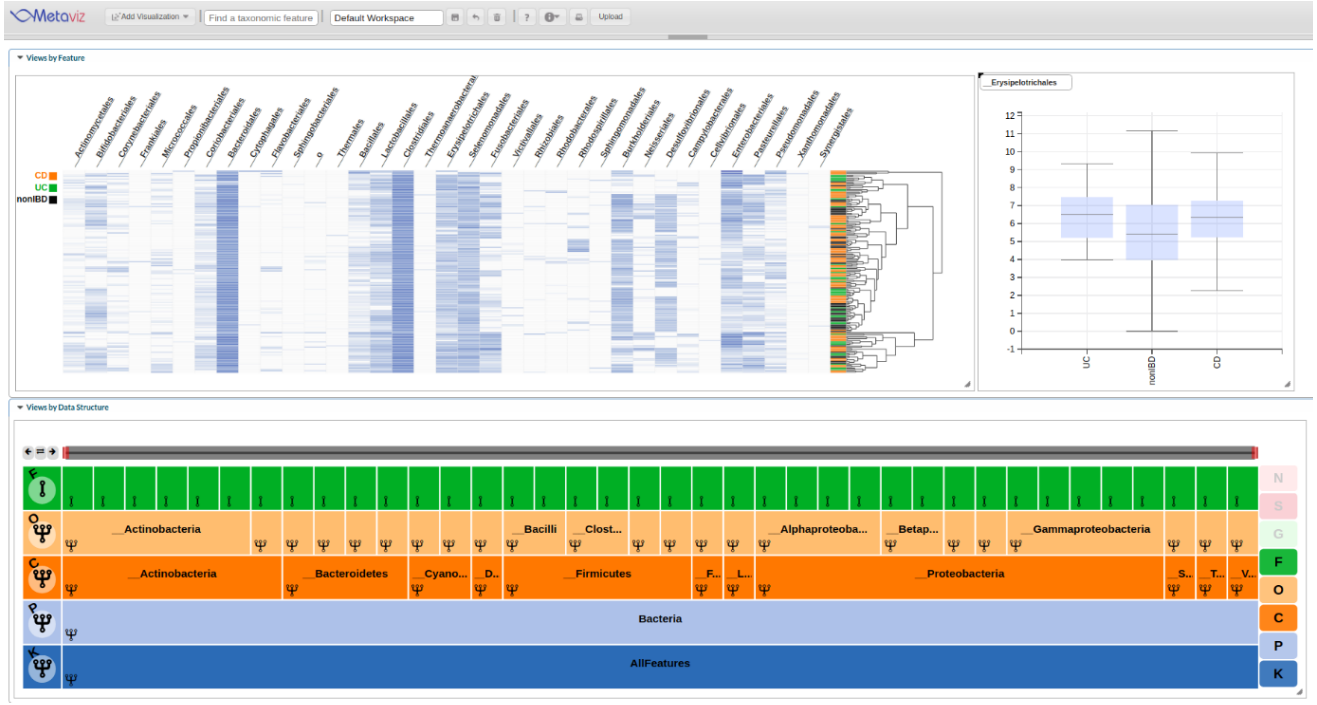
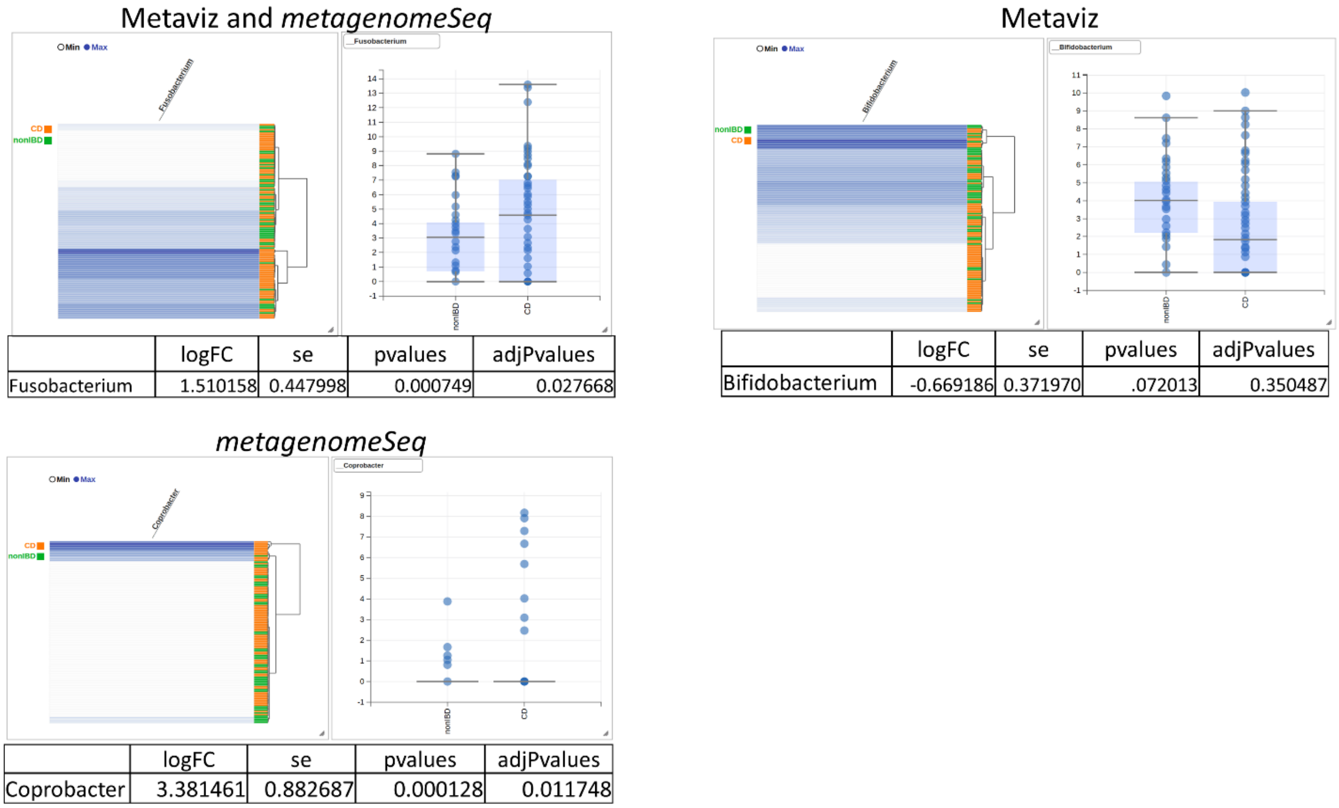


Figure 11: IBD HMP2 Multiple Groups Analysis
Using visual analysis through a heatmap and boxplots we identified taxonomic features that showed a difference in abundance between the three subject diagnosis categories: UC, CD, or nonIBD. We computed the F-statistic using the fitZig method in metagenomeSeq and list the findings in Appendix B Table 3.

For testing pair-wise comparisons between the three groups – UC, CD, and nonIBD - we used Metaviz to visually compare each group and performed statistical association testing with fitFeatureModel. Appendix B Table 4 shows the *metagenomeSeq* results for each group comparison. Appendix B Tables 5-7 list the results for Metaviz visual analysis between the groups with counts aggregated to each level of the taxonomic hierarchy.

From the pair-wise comparisons for the Crohn's Disease and subjects without IBD, we highlight the utility of tight linking between Metaviz for exploratory analysis and *metagenomeSeq* for confirmatory analysis. We focus on three taxonomic features, one identified as potentially differentially abundant with Metaviz, one identified as differentially abundant with *metagenomeSeq*, and one identified with both methods. We show the Metaviz boxplot and heatmap along with the *metagenomeSeq* log fold-change results for the three taxonomic features in Figure 12.



Metaviz and *metagenomeSeq*

	logFC	se	pvalues	adjPvalues
Fusobacterium	1.510158	0.447998	0.000749	0.027668

Metaviz

	logFC	se	pvalues	adjPvalues
Bifidobacterium	-0.669186	0.371970	.072013	0.350487

metagenomeSeq

	logFC	se	pvalues	adjPvalues
Coprobacter	3.381461	0.882687	0.000128	0.011748

Figure 12: CD vs nonIBD Metaviz and metagenomeSeq comparison
Comparing results from using Metaviz and metagenomeSeq to investigate associations between CD and nonIBD. We show one feature each from those identified using Metaviz and *metagenomeSeq,* found using Metaviz only, and found using *metagenomeSeq* only. The impact of using a mixture model is evident when considering the *metagenomeSeq* result compared to those from Metaviz. Linking exploratory analysis with confirmatory analysis helps an analyst curate results for collaborators.

3.4 Discussion

We now detail the biological significance of the results from exploratory analysis and differential abundance testing.

IBD Stool Pilot

From the *metagenomeSeq* results, the first taxonomic feature, s__:369227, is a member of the *Lachnospiraceae* family which are strictly anaerobic [49]. Members of *Lachnospiraceae* are abundant in human intestinal tracts and have been linked specifically to production of butyric acid[49]. Also, colonization with a specific strain of *Lachnospiraceae* in obese mice has been linked to development of hyperglycemia[50]. The second taxon, s__:363232, is a member of the genus *Dorea* which has recently been shown to be associated with diarrhea predominant Irritable Bowel Syndrome[51]. In the IBD Stool Pilot dataset, the number of Crohn's Disease versus Ulcerative Colitis samples is unbalanced. This is a potential cause of only one visually identified taxonomic feature being found as statistically significant. One consideration with our visual analysis approach of a heatmap and boxplot is that the effect size can be interpreted but the standard error is not as apparent.

IBD HMP2

As this dataset involved pair-wise comparison between groups, we first consider the results of comparing samples from UC and nonIBD subjects. We found *Verrucomicrobia* along with the following members of the lineage to have a statistically significant difference in abundance between groups with higher abundance in nonIBD than UC subjects: *Verrucomicrobiae*, *Verrucomicrobiales, Verrucomicrobeae,* genus *Akkermansia,* and one species. *Akkermansia* has been identified in a signaling between gut epithelial cells to control obesity related to diet[52]. The lower abundance in UC samples could indicate that the interaction between the inflamed gut epithelial cells is not functioning properly and could be a result of the lower abundance of *Akkermansia*. We also found *Megasphaera* to have statistically significant greater abundance in nonIBD compared to UC subjects. In a study of Malawian children for environmental enteric dysfunction, *Megasphaera* was shown to be more prevalent in children with the condition compared to children without[53]. In that paper, the authors also note that *Megasphaera* was identified as associated with HIV positive status[54]. Given previous findings about *Megasphaera*, the lower abundance in UC patients is notable as higher prevalence was associated with other intestinal disorders. Also, we identified *Citrobacter* as statistically significant higher abundance in

UC compared to nonIBD subjects. A recent study that involved sequencing the endoscopic equipment of subjects with UC, CD, and without IBD had a similar finding[55]. We also identified *Dielma,* a genera of the family *Erysipelotrichaceae* in the phylum Firmicutes, as higher in UC than nonIBD samples. We did not find *Dielma* to be well characterized in human health during a literature review.

From testing between subjects with CD and nonIBD, we found several bacteria with statistically significant differential abundance. Highlighting some of our findings, *Fusobacteria*, *Fusobacteriaceae*, *Fusobacteriales*, *Fusobacteriia*, and *Fusobacterium* all showed significantly greater abundance in CD compared to nonIBD subjects. *Fusobacterium* has previously been reported to have high prevalence associated with CD[56]. The taxonomic feature *Lachnospiraceae_ND3007_group* showed a significantly lower abundance in CD compared to nonIBD. This taxonomic feature is a member of the family *Lachnospiraceae,* which was observed to have lower abundance in CD subjects compared to nonIBD in a prior study[6].

Comparing the taxonomic profiles of samples from UC and CD subjects, we found *Veilloneliaceae* as significantly more abundant in CD subjects than UC. *Veilloneliaceae* was identified as associated with higher abundance in CD compared to nonIBD samples in a study of new onset IBD[6].

Exploratory and Confirmatory Analysis

The results presented in Figure 12 show the power of combining exploratory visualization and confirmatory statistical testing. The impact of using a zero-inflated model is evident when considering the *metagenomeSeq* result. During Metaviz inspection of the dataset, we did not identify *Coprobacter* as potentially differentially abundant. But we found a significant log fold-change when comparing CD to nonIBD groups using *metagenomeSeq*. This disparity between the visual inspection and statistical result helps with interpretation of the role *Coprobacter* might play in CD compared to nonIBD.

On an axis of exploratory at one end and confirmatory at the other, visualization techniques lie in the exploratory range while statistics can be used for both exploration and confirmatory analysis of a dataset. Biologists who are concerned with confirmation as opposed to exploration need to convey the results of analysis to collaborators and the scientific community. Visualizations can help curate statistical results. Interactive figures are a promising avenue for allowing researchers to curate results and make them accessible to the readers with several publishing venues incorporating visualization infrastructure for articles [57].

3.5 Conclusion

In this work we presented software infrastructure linking Metaviz to the HMP data resources. We detailed the 16S taxonomic community profile data from the HMP available in the UMD Metagenome Browser. We then described linking the UMD Metagenome Browser to the HMP Data Portal for single files and the manifest file utility for multiple file selections. We also performed visual exploratory and confirmatory differential abundance analysis of data from the IBD study. We first visualize 16S and whole metagenome sequencing abundance measurements for the same samples in *metavizr*. Then we use Metaviz and *metagenomeSeq* to analyze two datasets, IBD Stool Pilot and HMP2 IBD, to examine microbiome feature abundances in samples from subjects with Ulcerative Colitis, Crohn's disease, and without IBD. These illustrative analyses demonstrate the utility of Metaviz for integrative analysis with the HMP data resources. During this work, we identified two avenues for future research with Metaviz. First, a mechanism to filter taxonomic features for statistical testing based on visualization would be useful. Second, the metadata available from the HMP DCC is only that which has been approved for public release and methods to complete analyses over confidential data are desirable for microbiome data.

Chapter 4: Visualization of Longitudinal and Microbial Community Functional Profiling Data with Metaviz

This work is currently in preparation for submission to an appropriate venue. This work is joint with Jayaram Kancherla, Niklas Elmqvist, and Hector Corrada Bravo.

4.1 Introduction

Data visualization is a vital component in the process of data analysis. Visualization allows an analyst to gain insights into the data beyond summary statistics and to identify possible linear and non-linear trends as well as detecting outliers. In this work, we describe visualizations of longitudinal and microbial community functional profiling data in Metaviz, an interactive exploratory microbiome data analysis web-application. We also detail new utilities for microbiome data analysis available in Metaviz including a mechanism to store and lookup information on taxa of interest as well as a stack of operations to keep track of user interactions.

Related Work

Interactive and static visualization approaches are used in microbiome visualization to explore associations between disease and community profile. MEGAN is a widely used method for taxonomic analysis of microbiome sequencing data and includes a utility to create visualizations of relationships between taxonomic community members[58]. EMPeror offers interactive 3D visualizations of PCA plots to show distances between microbiome samples[43]. Phinch is a web-based interactive visualization tool that renders stacked bar plots of count data and interactive literature search to show the functions of taxonomic features[59]. BURRITO is the first interactive visualization for taxonomic and functional annotation data[60].

Metaviz

Metaviz is an interactive visualization web-application for exploratory microbiome data analysis. The application consists of a JavaScript and D3.js-based frontend suite of charts and a FacetZoom navigation component. The backend is either a graph database or the *metavizr* R/Bioconductor package. A limitation of the current Metaviz is that the FacetZoom navigation component operates with a taxonomic hierarchy but other hierarchies such as those in KEGG are used in metagenomic analysis. Also, visualizations of longitudinal data in Metaviz are mainly carried out with line plots that can be overcrowded when investigating the abundance of many features across multiple subjects. Finally, while Metaviz currently offers tight coupling with the R/Bioconductor environment for statistical testing, the user is responsible for keeping track of taxa of interest and looking up information about those features. This work introduces new methods to address these specific limitations.

4.2 Visualizations and User Interactions

We detail new utilities in Metaviz for visualizations of microbiome community functional profile and longitudinal data. We also describe a mechanism for keeping track of taxa of interest, linking to external literature resources, and a method to keep track of interactions with the FacetZoom navigation component.

Functional Profiling Data

Metaviz works well for community taxonomic abundance data but incorporating other data types is a current limitation. One specific data type is functional profile data. A visualization tool, BURRITO, provides a mechanism to investigate both community taxonomic abundance profile and functional profile data[60]. Our database architecture stores hierarchical taxonomic data and we modified it to hold functional data as well. For creating a map between the taxonomic features we use PICRUSt utilities to infer functional annotations given marker-gene data similar to BURRITO[61].

For visualizing the functional annotation hierarchy, we use a FacetZoom approach. We use the functional information primarily as a filter on the taxonomic features. Users can build a filter and apply it over the taxonomic hierarchy which is then used in heatmaps and stacked plots for taxonomic abundance measurements. The user interface for the functional annotation FacetZoom involves adding or removing a functional feature by clicking an icon on the node itself. The user then clicks a button to filter the taxonomic hierarchy FacetZoom. Figure 13 shows these a taxonomic hierarchy FacetZoom, a functional hierarchy FacetZoom, and a heatmap as well as the interaction for filtering based on functional information.

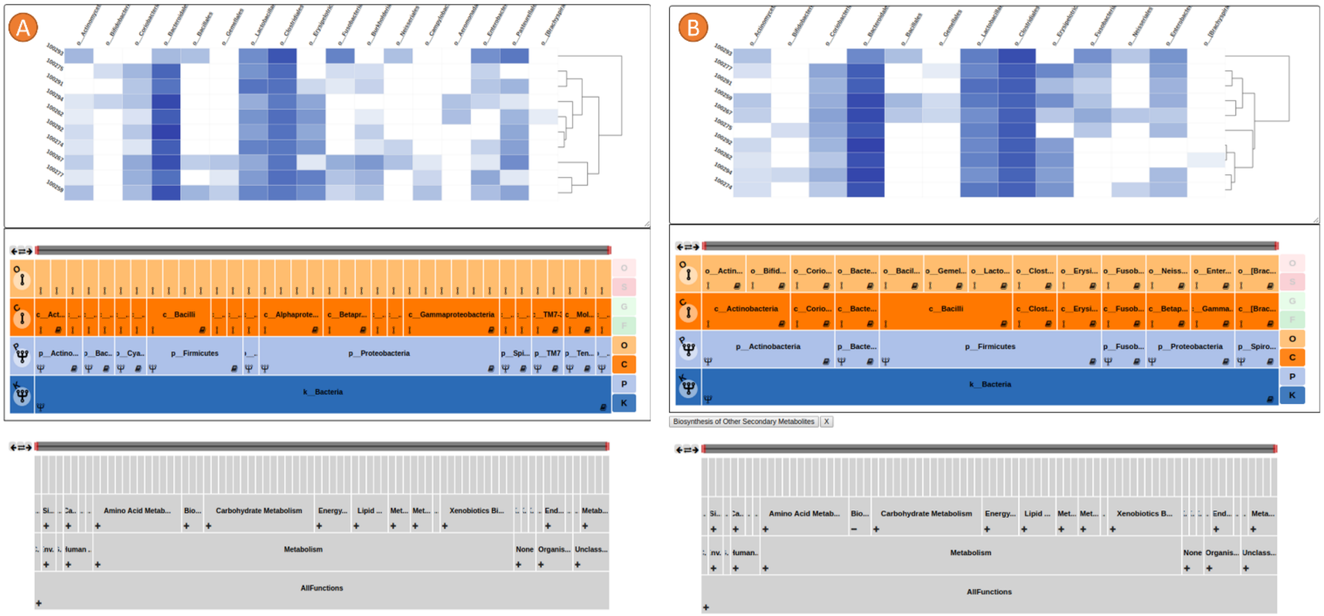


Figure 13: Functional Annotation Filter
Marker-gene sequencing provides a taxonomic community profile for a sample. Functional annotations can be inferred using this data and Metaviz includes a mechanism to filter taxonomic features based on functions. As the functional annotations have a hierarchical form, the FacetZoom can be used to show this as well. The figure shows an example with the 10 samples from the *msd16s* dataset and a subset of 176 KEGG Ortholog (KO) terms loaded into a Metaviz database. A) shows a taxonomic FacetZoom, KO FacetZoom, and a heatmap with no filter applied. B) shows the result of choosing specific KO functions to filter the taxonomic hierarchy with and the resulting heatmap of those taxonomic features.

We support what we term a “functional lens”. In the setting, the user is a microbiologist who creates a heatmap and FacetZoom to explore a dataset. The user can click on the bookmark symbol of a

FacetZoom node for any taxonomic feature of interest to keep track of it. The user can then update the functional filter to find features with similar functions. We show this utility in Appendix C Figure 1.

Longitudinal Visualization with Spark Lines

The existing Metaviz implementation offers longitudinal analysis through a line plot. Although we designed specific mechanisms for microbiome data analysis, such as interactive smoothing parameter adjustment, the number of features and samples in a line plot can lead to overcrowding. To allow a user to identify the change in a feature over time across all subjects in a dataset, we introduce a new longitudinal visualization for microbiome analysis. We adapt the heatmap that currently colors elements of a matrix according to the abundance of a given feature in a specific sample. To show the change in longitudinal measurements for a subject we use the spark line technique as the elements of a heatmap matrix. A sparkline is a small graphic that presents the trend of a dataset so a user can quickly identify changes [62]. Figure 14 shows the use of sparklines. In this dataset, which is explored in Use Case 2 of Chapter 2, study subjects were challenged with enterotoxigenic *E. coli* and then sampled for multiple timepoints. Antibiotics were administered to the subjects and this perturbation is marked in each sparkline by changing line color from blue to orange. Also, the box for each sparkline is colored to highlight series of interest. A sparkline box is highlighted if the difference between consecutive measurements for that feature in that subject is beyond one standard deviation of the mean difference across all subjects between those timepoints. For specific exploration of grouping, we developed a details-on-demand view as shown in Appendix C Figure 2. In this case we show two subject groups, those that developed diarrhea at any point during the experiment and those that did not. The user has an option to show a filled contour for each group as shown in Appendix C Figure 2 or the user can choose lines showing different summaries for each time point across all subjects in each group.

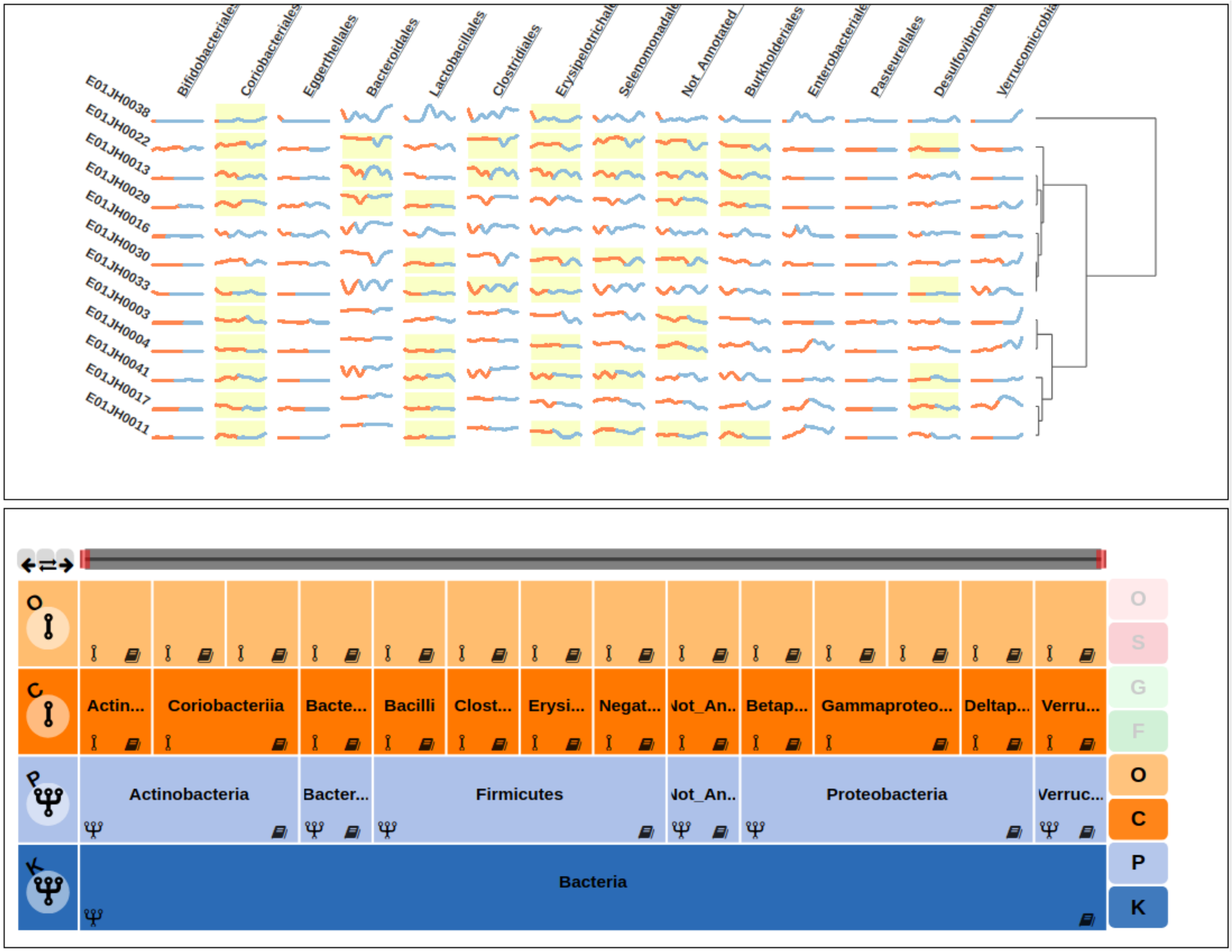


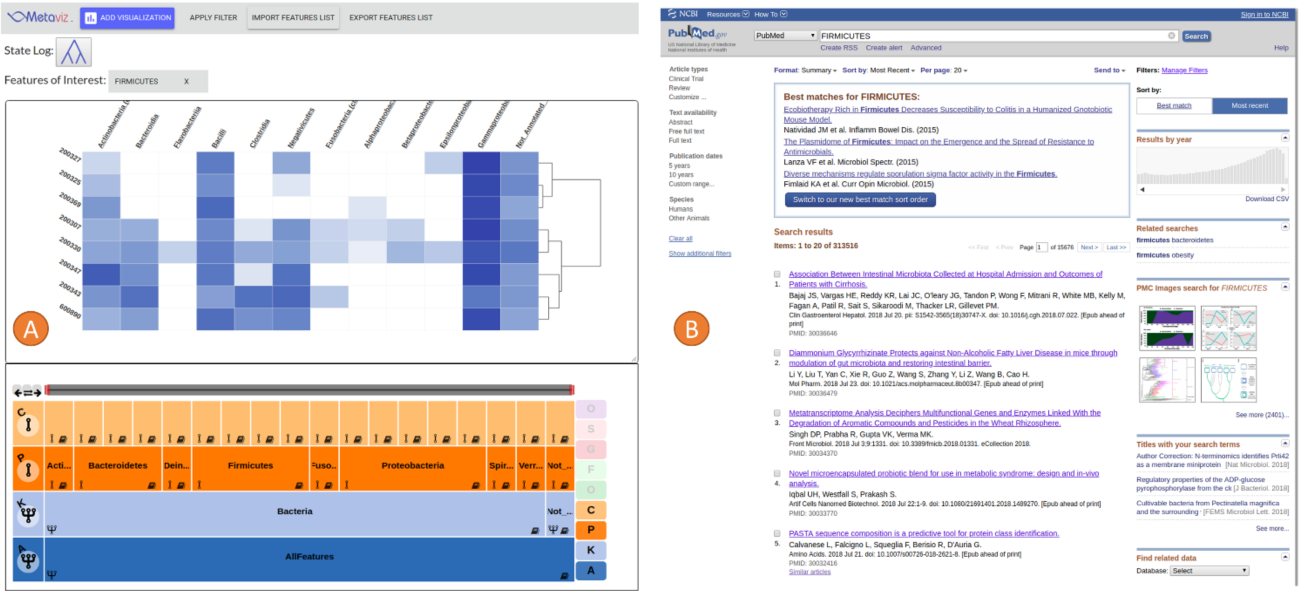
Figure 14:Heatmap with Sparklines for longitudinal data
In longitudinal microbiome experiments it is useful to get an overview of changes in feature abundance at different time points. This heatmap provides a mechanism for users to identify trends across features and subjects. Each sparkline is colored according to a perturbation which in this case was administration of antibiotics to the subjects under study. Also, each box is colored according to the difference of measurements between timepoints for that subject/feature pair being outside one standard deviation from the mean difference for all subjects between those timepoints. A user can then click on a column of the sparkline heatmap and then inspect all measurements of that feature across subjects. We provide a contour map and averages lines grouped by case/control status.

Taxa of Interest, Export and Import

We provide a utility for a user to keep track of taxonomic features of interest found through visual exploration. A user can export the list of features or import a new list. Appendix C Figure 1 shows the process of finding functions that an individual taxonomic feature is associated with then finding taxonomic features with similar functions.

External Data Source Links

We link each item in the taxa of interest space to Pubmed [http://www.pubmed.com]. This provides users a quick mechanism to review literature for a specific taxon that is of interest in a visualization. Figure 15 shows the workflow. During microbiome data analysis, understanding the role of specific bacteria and lineages is vital.



PubMed ▾ FIRMICUTES Search

Create RSS Create alert Advanced Help

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Format: Summary ▾ Sort by: Most Recent ▾ Per page: 20 ▾ Send to ▾

Best matches for FIRMICUTES:

Ecobiotherapy Rich in **Firmicutes** Decreases Susceptibility to Colitis in a Humanized Gnotobiotic Mouse Model.
Natividad JM et al. Inflamm Bowel Dis. (2015)

The Plasmidome of **Firmicutes**: Impact on the Emergence and the Spread of Resistance to Antimicrobials.
Lanza VF et al. Microbiol Spectr. (2015)

Diverse mechanisms regulate sporulation sigma factor activity in the **Firmicutes**.
Fimlaid KA et al. Curr Opin Microbiol. (2015)

Switch to our new best match sort order

Search results

Items: 1 to 20 of 313516

<< First < Prev Page 1 of 15676 Next > Last >>

1. □ Association Between Intestinal Microbiota Collected at Hospital Admission and Outcomes of Patients with Cirrhosis.
Bajaj JS, Vargas HE, Reddy KR, Lai JC, O'leary JG, Tandon P, Wong F, Mitrani R, White MB, Kelly M, Fagan A, Patil R, Sait S, Sikaroodi M, Thacker LR, Gillevet PM.
Clin Gastroenterol Hepatol. 2018 Jul 20. pii: S1542-3565(18)30747-X. doi: 10.1016/j.cgh.2018.07.022. [Epub ahead of print]
PMID: 30036646

2. □ Diammonium Glycyrrhizinate Protects against Non-Alcoholic Fatty Liver Disease in mice through modulation of gut microbiota and restoring intestinal barrier.
Li Y, Liu T, Yan C, Xie R, Guo Z, Wang S, Zhang Y, Li Z, Wang B, Cao H.
Mol Pharm. 2018 Jul 23. doi: 10.1021/acs.molpharmaceut.8b00347. [Epub ahead of print]
PMID: 30036479

3. □ Metatranscriptome Analysis Deciphers Multifunctional Genes and Enzymes Linked With the Degradation of Aromatic Compounds and Pesticides in the Wheat Rhizosphere.
Singh DP, Prabha R, Gupta VK, Verma MK.
Front Microbiol. 2018 Jul 3;9:1331. doi: 10.3389/fmicb.2018.01331. eCollection 2018.
PMID: 30034370

4. □ Novel microencapsulated probiotic blend for use in metabolic syndrome: design and in-vivo analysis.
Iqbal UH, Westfall S, Prakash S.
Artif Cells Nanomed Biotechnol. 2018 Jul 22:1-9. doi: 10.1080/21691401.2018.1489270. [Epub ahead of print]
PMID: 30033770

5. □ PASTA sequence composition is a predictive tool for protein class identification.
Calvanese L, Falcigno L, Squeglia F, Berisio R, D'Auria G.
Amino Acids. 2018 Jul 21. doi: 10.1007/s00726-018-2621-8. [Epub ahead of print]
PMID: 30032416
Similar articles

Filters: Manage Filters

Sort by: Best match | Most recent

Results by year

Download CSV

Related searches
firmicutes bacteroidetes
firmicutes obesity

PMC Images search for *FIRMICUTES*

See more (2401)...

Titles with your search terms
Author Correction: N-terminomics identifies Prli42 as a membrane miniprotein [Nat Microbiol. 2018]
Regulatory properties of the ADP-glucose pyrophosphorylase from the cl [J Bacteriol. 2018]
Cultivable bacteria from Pectinatella magnifica and the surrounding [FEMS Microbiol Lett. 2018]
See more...

Find related data
Database: Select ▾

Figure 15: External Data Source Link
Metaviz links taxa of interest to Pubmed for information lookup. A user then has access to the rest of the NCBI utilities with that feature name. With microbiome analysis, the large feature space can be quickly investigated through inspecting literature for taxa of interest. A) A Metaviz workspace in which a user can click any node in the FacetZoom to link to external information. B) Pubmed search results with taxonomic feature as query.

Operation Stack

Each FacetZoom navigation operation can be stored in a stack of operations. The stack is represented by buttons with each showing a glyph of the operation that was performed. Each button can be clicked on to revert the hierarchy and the workspace to that state. Figure 16 shows the current implementation of the stack of operations in Metaviz.

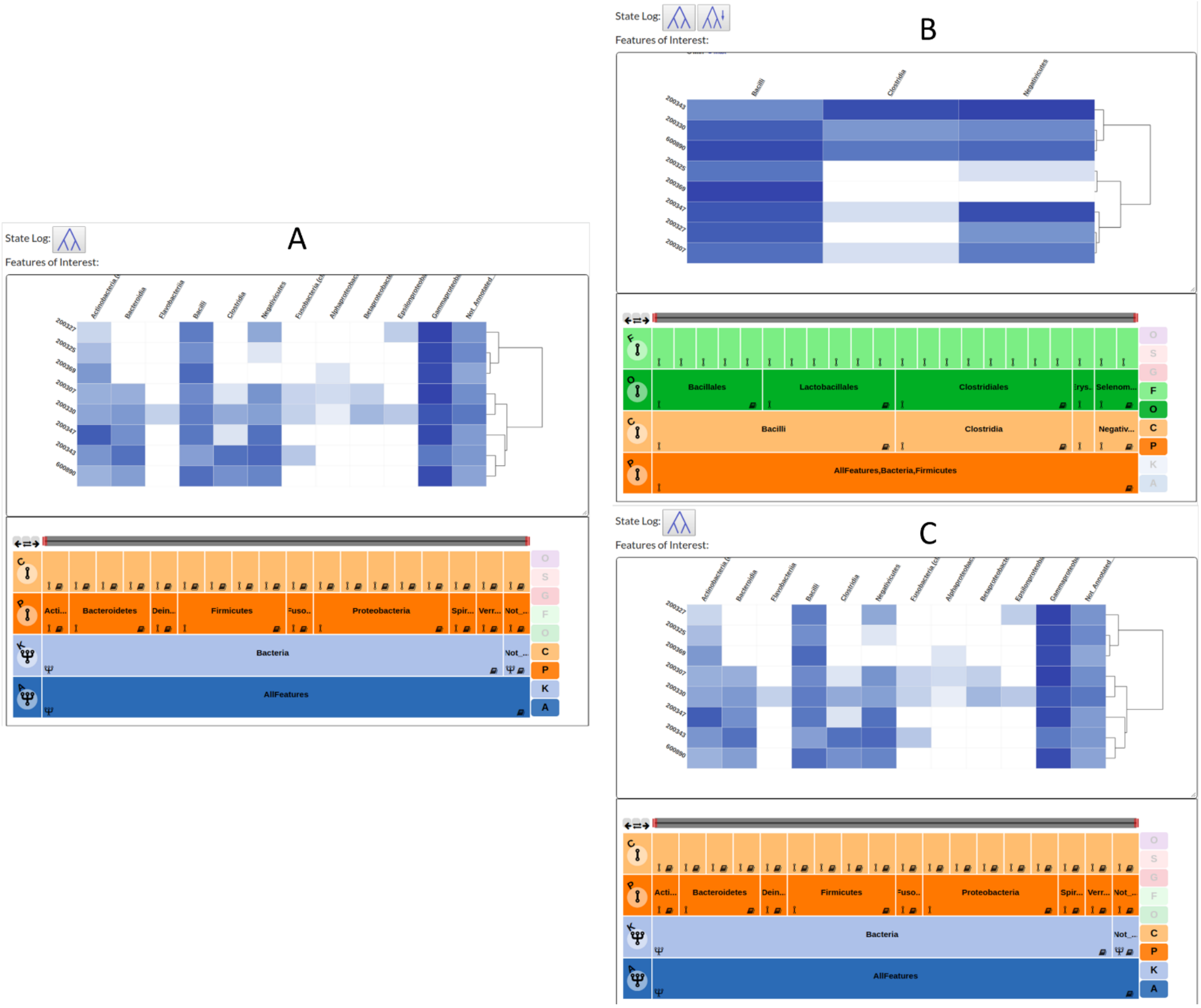


Figure 16: Stack of operations demonstration
A State Log for each user interaction with the taxonomic FacetZoom. (A) State Log is initialized with the first button showing the initial hierarchy view. (B) On descent to a lower level of the taxonomy, the State Log shows that the root of the subtree shown in the FacetZoom is now closer to leaf nodes. (C) After a click on a State Log button, the operations are popped off to recover the prior state. In this case, the subtree displayed in the FacetZoom is now at the original level.

4.3 Architecture

We designed and implemented improvements to Metaviz while keeping the existing architecture of the application.

Backend

We store community functional profile information in a graph database backend. For the taxonomic hierarchy in the Metaviz graph database, we represent taxa as nodes, taxonomic hierarchy relationships using edges, samples as nodes, and counts of taxa in specific samples as edges. Aggregation queries with sample and feature selections are then handled by the Neo4j query execution utility. With functional data, we represent the KO terms as nodes with relations between functional annotations as edges. Edges then link functional terms to taxa in the taxonomic hierarchy. A filter is generated using functional annotations and the filter is applied to taxonomic hierarchy as well as count aggregation queries.

Components library

The new Epiviz web components library provides a framework to develop extensible HTML components[63]. We plan to migrate all Metaviz utilities to this framework to ease of future development.

4.4 Discussion and Conclusions

In this work, we present new methods to use Metaviz for exploratory analysis of microbial community functional profile data and longitudinal data. We also implement several utilities to improve the Metaviz user experience including keeping a stack of user interactions, a space for storing taxa of interest identified through visual inspection and linking to external data sources. In previous work we showed that users can validate insights identified from statistical testing approaches on large microbiome sequencing datasets using interactive visualization. The mechanisms for longitudinal and community functional profile data will enable analysts to examine data from experiments targeting perturbation of microbiomes.

Chapter 5: Privacy-preserving microbiome analysis using secure computation

A version of this work is published in Bioinformatics with the following citation:
Justin Wagner, Joseph N. Paulson, Xiao Wang, Bobby Bhattacharjee, Héctor Corrada Bravo; Privacy-preserving microbiome analysis using secure computation, Bioinformatics, Volume 32, Issue 12, 15 June 2016, Pages 1873–1879, https://doi.org/10.1093/bioinformatics/btw073.
This is joint work with Joseph Paulson, Xiao Wang, Bobby Bhattacharjee, and Hector Corrada Bravo.

5.1 Introduction

Microbiome sequencing seeks to characterize and classify the composition and structure of microbial communities from metagenomic DNA samples. It is estimated that only 1 in 10 cells in and on a person's body contain that individual's DNA[4], the remainder corresponding to microbial DNA, most from organisms that cannot be cultured and studied in the laboratory.

The Human Microbiome Project (HMP)[64], the Global Enteric Multi-Center Study (MSD) [3], the Personal Genome Project[65] and the American Gut Project[66] aim to characterize the ecology of human microbiota and its impact on human health. Potentially pathogenic or probiotic bacteria can be identified by detecting significant differences in their distribution across healthy and disease populations. While the biology has led to promising results, privacy concerns of microbiome research are now being identified with no secure analysis tools available.

Recent work by Franzosa *et al.* (2015) shows that microbiome data are an unique identifier across time points in a dataset and could be used to link a sensitive attribute to an individual[23]. Earlier work by Fierer *et al.* (2010) showed that it is possible to identify an object that an individual touched by comparing microbiome samples from the object and the individual's hand[67]. We provide a thorough review of microbiome sequencing and a categorization of microbiome privacy considerations in the Appendix D. To counter these concerns, we present an implementation and evaluation of metagenomic association analyses in a secure multi-party computation (SMC) framework. For this work, we focus on garbled circuits, a cryptographic technique that evaluates a function over private inputs from two parties.

In this article, we concentrate on the case where two parties, each holding organism abundances in a set of case and control samples, are interested in performing an association analysis (e.g. determining organisms that are differentially abundant in cases) over their combined data, without revealing organism abundances in any specific sample.

We provide a detailed review of this approach in Section 3 and benchmark our secure implementation of commonly used microbiome analyses on three public datasets. We also quantify the statistical gain of analysis using combined datasets by simulation with a dataset that contains samples from four different countries.

We believe that implementing metagenomic analyses in an SMC framework will prove beneficial to researchers focused on the human microbiome as well as the secure computation community. Computational biologists will benefit from a method that allows efficient and secure function evaluation over datasets which they may be obligated to keep confidential. Security researchers can draw on the findings from our work and construct protocols that enable sharing large, sparse datasets to perform analysis.

5.2 System and methods

Our secure metagenomic analysis system is built upon garbled circuits[68], which we describe in this Section. We then detail our system including participants along with alternative approaches in the design space for privacy-preserving analysis.

Garbled circuits

Two parties, one holding input *x* and another holding input *y*, wish to compute a public function *F*(*x*, *y*) over their inputs without revealing anything besides the output. The parties could provide their inputs to a trusted third-party that computes the function and reveals the output to each party. However, modern cryptography offers a mechanism to run a protocol between only the two parties while achieving the desired functionality. The main idea behind garbled circuits is to represent the function to be computed as a Boolean circuit over the inputs from both parties and use encryption to hide the input of each party during evaluation by mapping each 0 and 1 bit of the inputs unto random strings that still compute the same result. At the end of circuit evaluation, the resulting random strings can be mapped back to appropriate 0 and 1 bit values that can then be released to each party. In this way, each party learns $F(x, y)$ without learning anything else about the input of x and y. Figure 17 illustrates the garbled circuits protocol.

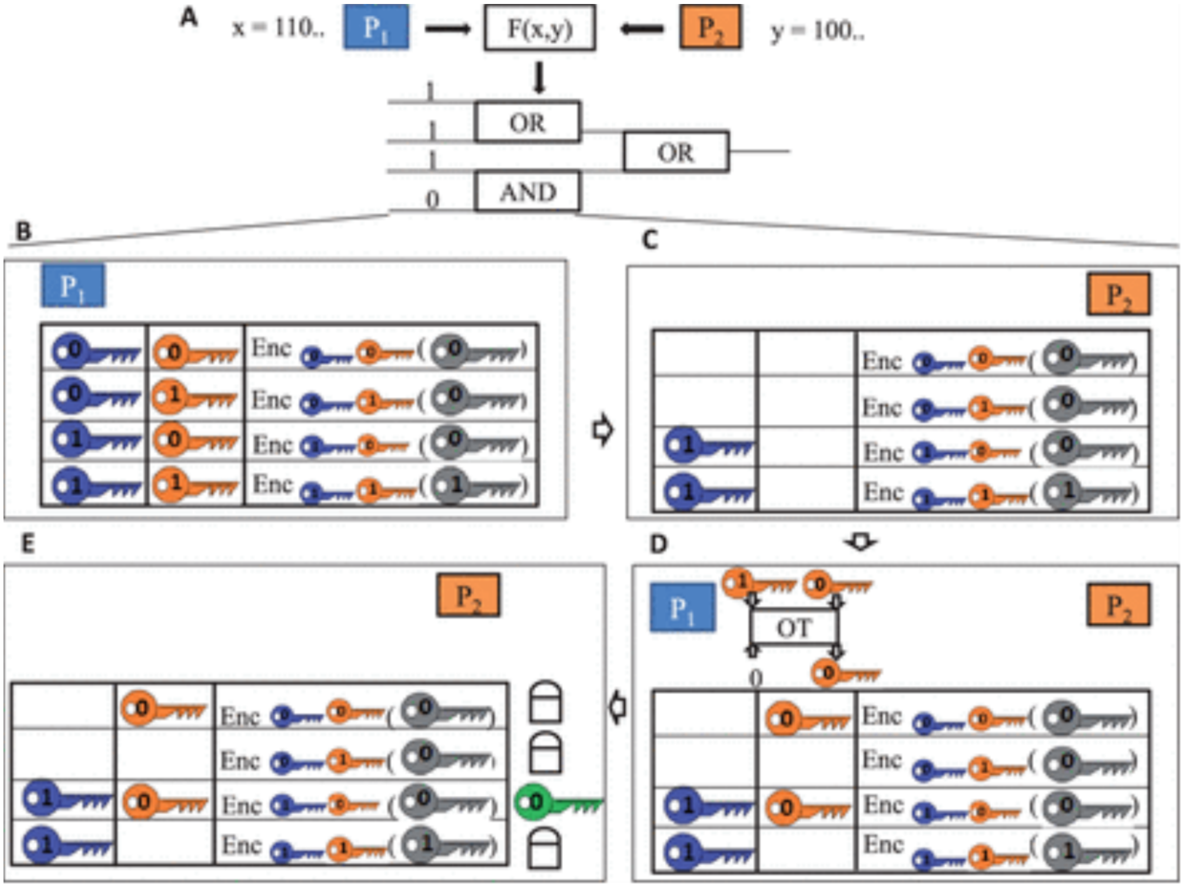


Figure 17: Schematic illustration of the garbled circuits protocol.
For analyses discussed in this paper, parties P1 and P2 are researchers performing a statistical analysis over combined data. They provide metagenomic count matrices, or locally precomputed statistics computed from count matrices, along with case/control status as input. Function F(x, y) is determined by the analysis performed, e.g. test on difference in Alpha Diversity between case and control. The ‘garbling’ in step (B) also includes randomly permuting the rows of the truth table so that the inputs are not revealed by the ordering - we omit this from the

figure for clarity. A review of the Oblivious Transfer protocol used in step (D) is provided in Appendix D Section S3.

System participants

We consider the case in which parties located in two policy-domains want to perform metagenomic analyses over shared data. Examples of policy-domains include countries with differing privacy laws or institutions (universities, companies) that stipulate different data disclosure procedures.

For $i \in 1,2$, denoting PD_i as a policy domain, R_i as a researcher in policy domain i, D_i as the data from R_i, F as the set of functions that a set of R_is would like to compute we consider the following setting:

> R_1 and R_2 would like to compute F over combined D_1 and D_2 but cannot do so by broadcasting the data as either PD_1 or PD_2 does not allow for public release or reception of individual-level microbiome data. We set $|i|=2$ but this setting could be generalized to any i.

Policy domains naturally arise due to differences in privacy laws. For example, studies currently funded by the NIH are required to release non-human genomic sequences including human microbiome data (http://gds.nih.gov/PDF/NIH_GDS_Policy.pdf). In contrast, the European General Data Protection Regulation, which is currently in draft form, lists biometric data and ‘any “data concerning health” means any personal data which relates to the physical or mental health of an individual, or to the provision of health services to the individual’ as protected information that is not to be released publicly (http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN). Therefore, researchers in the USA and EU may encounter different policies for data release but still have an interest in computing metagenomic analyses over shared data. Also, given the results published by Franzosa *et al*., some institutions may re-evaluate microbiome data release policies.

Threat model

We consider researcher R_1, who has a microbiome sample from a victim mixed with other samples, to be a semi-honest adversary, or one that follows the protocol but examines the transcript to learn more information than it should. Researcher R_2 is examining an association for a specific trait and would like to expand her study to use samples held by R_1. R_1 wants to determine if the victim is in R_2's dataset and thus learn a sensitive attribute of the victim such as disease status.

The attacks of Fierer *et al*. and Franzosa *et al*. operate over the vector of feature counts for a given sample. For the analyses studied in this article, an adversary will have no better chance of reconstructing the count vector for a specific sample than guessing the majority, or mode, of the count of any specific feature in this system. Through using a garbled circuit implementation of metagenomic analyses, R_2 will be able to keep the vector of microbiome features for any sample private, learn the outputs of functions that she would like to learn over the shared data, and prevent R_1 from completing the attack.

Solution design approaches

We consider different approaches to allow two parties to compute analyses over data which each must keep confidential.

Access control plus trusted third party

In the USA, the NIH has recognized re-identification through publicly posted genomic data as a realistic threat. Therefore, policy allows for publication of summary statistics and transfer of individual level sequencing data through access control using the Database for Genotypes and Phenotypes[69]. Once a researcher receives permission to access data, she is provided the data and is required to maintain the access control list for her research group. We look to remove the need for access control by implementing the queries that a researcher would like to run without revealing the data directly.

Differential privacy

Statistical perturbation of analysis results, most widely implemented as differential privacy, is a second approach for researchers to provide privacy guarantees to participants. In this setting, a researcher maintains a data set and allows other researchers to perform queries over the data. Informally, the results of these queries are perturbed in such a manner that an adversary, with access to query results over a data set in which one specific participant has a set of values and results from another data set with that specific participant having a different set of values, will not be able to infer any information about that individual by examining the results[21]. Although this approach provides provable privacy guarantees, the introduction of statistical noise has not gained traction in the computational biology research community. Also, recent work showed that learning warfarin dosage models on differentially private data sets introduces enough noise that the dosage recommendation could be fatal to patients[70].

Secure multiparty computation

An alternative solution which we undertake is using secure computation to perform metagenomic analyses. Other researchers have presented SMC for computing secure genome-wide association studies using secret-sharing, but that particular approach requires the use of three parties for computing tasks[71]. We address the feasibility of using garbled circuits to implement metagenomic analyses in terms of running time, network traffic, and accuracy. We believe that garbled circuits is the best approach for this scenario as it allows for direct communication between two parties and models research settings well. Further, garbled circuits can handle a variety of adversaries beyond the semi-honest one that we consider in this work.

5.3 Implementation

In this section, we describe how we implemented metagenomic analyses in garbled circuits and detail an evaluation of our system.

Metagenomics using garbled circuits

FlexSC

FlexSC, the back end of ObliVM, is a framework for secure computation including garbled circuits with a semi-honest adversary[72]. FlexSC allows users to write a function in Java for two parties to compute then compiles and evaluates the garbled circuit representation of that function. We implemented all metagenomic tests as Java packages then compiled and ran each with FlexSC. Our initial work on χ^2-test was based on a χ^2-test implementation using SNP data (https://github.com/wangxiao1254/idash_competition).

Metagenomic analysis assumptions

For this article, we perform all analyses at the species taxonomic level. As detailed in Appendix D Section 1, OTUs are generated from direct pairwise comparison of sequencing reads. This is a compute-intensive process when performed on clear text[73]. We do not attempt it in SMC for this work and assume each party performs this operation locally. We assume that each party will annotate each resulting OTU by matching to a common reference database, previously agreed upon by both parties (note that this reference database is orthogonal to sample-specific sequencing results obtained by each party). For illustration we assume that the agreed upon reference database yields annotation at the microbial species level. We also assume that parties can split data into case and control groups based on an agreed upon phenotype. Finally, we do not consider features that have all zeros in the case or control group for either party.

Design approaches

We took several approaches to implement each statistic. Since the metagenomic datasets we examined are at least 80% sparse and this trend is expected with OTU data[16], we make design choices to make computation with garbled circuits feasible. We now detail each implementation of the χ^2 -test, odds ratio, Differential Abundance and Alpha Diversity. To measure the impact of our design choices we implemented a naive algorithm for each statistic and compared results.

Precomputation

We first developed a method that finds an aggregate statistic at each party so that only those values are circuit inputs. This method is a straightforward approach to reduce the amount of operations and data in the secure computation protocol. As expected, for each statistic this approach had the best performance on all the datasets we evaluated. Appendix D Figure 2 shows the process for calculating a χ^2-test and odds ratio on precomputed contingency table counts. An issue with this approach is not all analyses that researchers are interested in computing may be able to be performed over locally generated aggregates.

Sparse matrix

We devised two methods to account for the sparsity of the feature count matrices we used for evaluation. We first followed an approach introduced by Nikolaenko *et al.* (2013) to perform sparse matrix factorization in garbled circuits[74]. We detail our work with this technique in the Appendix D Section S4. As our contribution, we took a conceptually simpler approach that input the non-zero elements for each feature to the circuit and operated over those elements directly. As shown in Figure 18 and Figure 19, this method significantly reduces the number of operations that need to be performed in the secure protocol and offers reasonable running times compared to the precomputation approach.

Presence/absence

We implemented the χ^2-test and odds ratio to perform presence/absence association testing. We provide a review of χ^2-test and odds ratio in Appendix D Section 1.

For the precomputation technique, each party splits its data into case and control groups on a characteristic determined outside of this protocol. Each party then locally computes the contingency table counts on the split data. These contingency table counts are each party's input into the circuit. Within the circuit, the counts are summed for both case and control groups then the χ^2-statistic along with the odds ratio are computed for each feature.

In the sparse matrix approach, the total number of samples and all non-zero elements for each feature are input to a garbled circuit. The circuit first adds the number of non-zero elements to compute the present contingency table counts then uses the total number of samples to find the absent counts.

Differential abundance

For calculating differential abundance, we implemented a two-sample *t*-test for testing the mean abundance between case and control groups. We assume normalization of sequencing counts can be accomplished in a preprocessing step between both parties. We make this assumption because we use normalized datasets in our evaluation. We leave implementation of normalization techniques in garbled circuits to future work.

For review of two-sample *t*-test we refer the reader to the Appendix D Section 1. We examined the process for calculating mean, variance and the t-statistic to determine what optimizations can be made for computing in a circuit. To avoid processing all samples within the computation framework, we observe transformations that reduce the total number of operations. In the Appendix D, we show how mean abundance and variance can be computed using the sum, sum of squares and total number of elements from each party. For precomputation, as each institution only needs to provide three values per feature we calculate them locally. In the circuit, a two-sample *t*-statistic to test difference between case and control groups is computed.

For the sparse matrix approach, the total sum and sum of squares are calculated in the circuit using the non-zero elements for each feature. Mean abundance along with variance can then be calculated and used compute the two-sample *t*-test. We refer the reader to Appendix D Section 4 for more detail.

Alpha diversity

We use a two-sample t-test to determine the significance of mean Alpha Diversity difference between case and control groups. Given that FlexSC does not currently compute logarithm, we measure Alpha Diversity as Simpson's index: $D = (\sum n(n-1)) \div N(N-1)$ where n is the number of OTU counts for OTU$_i$ and N is the total number of counts observed in a sample.

For precomputation, we locally compute Simpson's index for each sample. These values are input into the circuit where they are summed, mean and variance is taken, and the t-statistic is calculated. In Alpha Diversity, all samples in case and control must be processed together as opposed to Presence/Absence and Differential Abundance which can be computed per feature.

For our sparse computation design, the two values for Simpson's index, $\sum n(n-1)$ and $N(N-1)$ are generated over each sample in the circuit during one pass through the matrix. Then a pass over an array of these values using division yields Simpson's index from which the total sum and sum of squares can be used to compute the two-sample t-test between case and control groups.

Evaluation

We evaluated our implementation using two Amazon EC2 r3.2xLarge instances with 2.5 GHz processors and 61 GB RAM running Amazon Linux AMI 2015.3. We measured the size of the circuit generated, running time and network traffic between both parties for each metagenomic statistic and dataset. Circuit size serves as a useful comparison metric since it depends on the function and input sizes but is independent of hardware. Running time and network traffic are helpful in system-design decisions and benchmarking of deployments.

Datasets

We used OTU count data from the Personal Genome Project (PGP)[65], the HMP[64], and the Global Enteric MSD[3]. We retrieved the MSD data from the project website (ftp://ftp.cbcb.umd.edu/pub/data/GEMS/MSD1000.biom) as well as the PGP and HMP datasets are from the American-Gut project site (https://github.com/biocore/American-Gut/tree/master/data) [66]. We used the tongue as the case and gingiva as control for the HMP data. For PGP, we set forehead as case and left palm as control. Case and control criteria for the MSD dataset were already set by the researchers that publish the data depending on disease phenotype. After aggregating to species and removing features which hold all zeros for either the case or the control group, the PGP contains 168 samples and 277 microbiome features, the HMP has 694 samples and 97 features, and the MSD dataset consists of 992 samples and 754 features. Appendix D Table 2 summarizes the size and sparsity of each dataset.

Efficiency of secure computation

Circuit size

Figure 18 shows the circuit size per feature for each experiment. As a result of the work by Kolesnikov and Schneider (2008), XOR gates in each circuit do not require costly network traffic and computation, therefore the total number of non-XOR gates is reported for each statistic and dataset[75]. Using precomputation, the complexity of the equation in terms of arithmetic operations to calculate each statistic determines the circuit size. This explains the circuit sizes for odds ratio and χ^2 test as compared with Differential Abundance. For Alpha Diversity, all rows and columns are preprocessed with only the two-sample *t*-test computed in the circuit. With the sparse implementation, the complexity of the test along with the number of non-zero elements in the dataset directly affects circuit size.

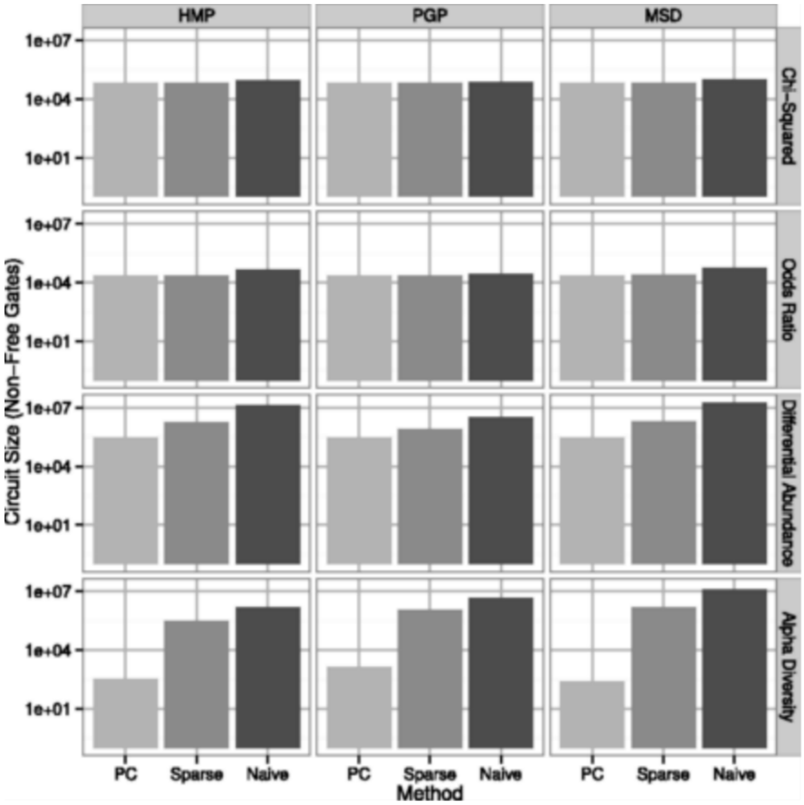


Figure 18: Circuit size per feature for each implementation and dataset.
The feature count for Alpha Diversity is the number of samples. The differences in Alpha Diversity between datasets is explained by the number of samples for PGP (168) being much lower than that of HMP (694) and MSD (992). PC, Pre-compute.

Running time

For the sparse implementation, the running time was proportional to the size and number of non-zero elements in each dataset. For precomputation, Alpha Diversity was affected by the number of samples in each dataset. The running time for the χ^2 test, odds ratio, and Differential Abundance were proportional to the number of features (rows) processed. Figure 19 summarizes the effects of input size and algorithm complexity on running time.

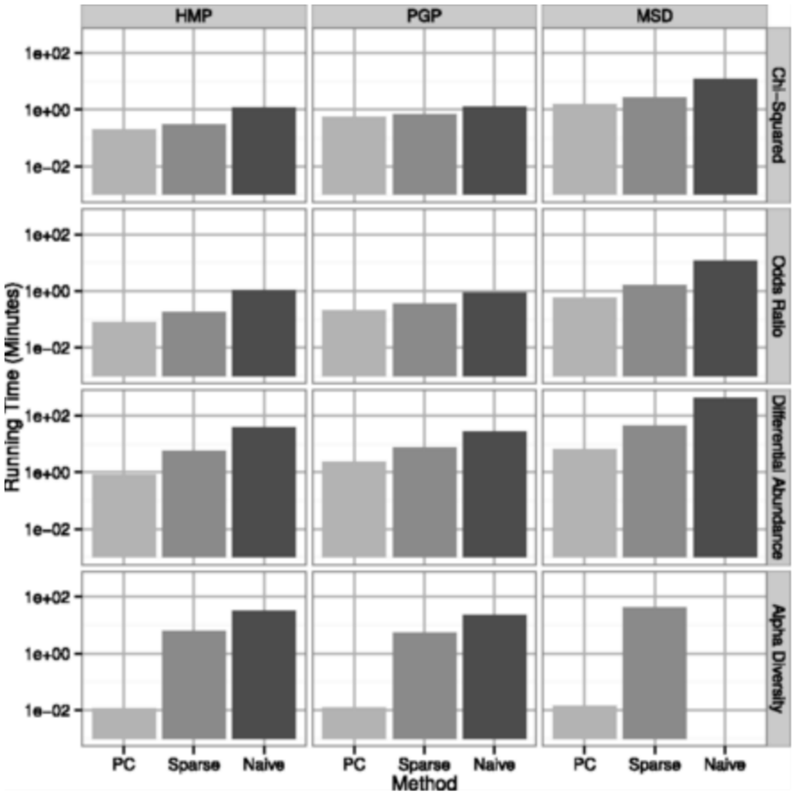


Figure 19: Running time for each statistic and each dataset in minutes.
In each statistic, the number of arithmetic operations determined the running time. The size of the dataset along with sparsity contributed to running time for the sparse implementations. Alpha Diversity MSD Naive did not run to completion on the EC2 instance size due to insufficient memory. Based on the circuit size and the number of gates processed per second for other statistics, we estimate the running time to be 378 min. PC, Pre-compute.

Network traffic

Appendix D Table 5 shows the network traffic for each experiment. The increase in network traffic between the precomputation and sparse implementations is more significant than the differences in running times of those approaches. We believe that the network traffic for the precompute implementation is quite good for the security guarantees provided with using garbled circuits while the sparse approach presents an acceptable tradeoff depending on the network resources available.

Accuracy

We compared the accuracy of our implementation results to computing the statistic using standard R libraries. Table 2 lists the accuracy of results for the χ^2 statistic, odds ratio, as well as the *t*-test results for

Differential Abundance and Alpha Diversity. The differences in our garbled circuits results compared to the *R* values appear to be the result of circuit complexity. The floating-point arithmetic operations in FlexSC are software implementations. Therefore, the operations are subject to rounding errors that are rarely observed on modern processors which have hardware level support for floating-point arithmetic.

Table 2: Accuracy

	PGP	HMP	MSD
Chi-square statistic	7.84e-07	7.48e-06	7.02e-08
Chi-square *P*-value	2.00e-07	2.14e-06	9.72e-08
odds ratio	1.60e-13	5.42e-13	2.44e-13
Differential abundance			
t-statistic	0.023	0.0017	0.0012
Differential abundance			
degrees of freedom	2.7e-4	2.5e-4	0.0028
Differential abundance			
P-value	0.0024	0.0026	0.0011
Alpha Diversity			
t-statistic	0.0038	0.017	0.0049
Alpha Diversity			
degrees of freedom	1.48e-05	9.7e-4	2.2e-4
Alpha Diversity			
P-value	0.0088	0.044	0.014

Results were generated using the R chisq.test{stats}, odds.ratio{abd}, t.test{stats}, and diversity{vegan} against our implementation in ObliVM for the χ^2-test, odds ratio, differential abundance and Alpha Diversity. We use Normalized Mean Squared Error: $\|x-y\|2/\|x\|2$ with x as the value output by R and y the value from our implementation. For comparing P-values, we use the $\log_{10}$P-value and exclude any exact matches [since $\log_{10}(0) = -\text{Inf}$ in R] while computing the mean.

We investigated if our implementation yielded any false positives and false negatives with the results from R acting as ground truth. For the *P*-values of Differential Abundance in PGP, HMP, and MSD datasets we found no false positives or false negatives for a significance level of 0.05.

Significant features discovered through data-sharing

Researchers in different policy domains may be forced to compute analyses on partial data. We measured the effect of using our implementation for data-sharing between policy domains. The MSD dataset provides a means to simulate secure computation of microbiome analyses between different countries. The data were gathered from Kenya, The Gambia, Bangladesh and Mali. We simulate each country performing secure Differential Abundance pair-wise with the other countries. We observed that sharing data resulted in a substantial increase (at minimum a 98% increase) in the number of species found to be differentially abundant between case and control groups. Table 3 summarizes the results.

Table 3: Feature Testing Across Domains

	Features found	**Total increase**
Kenya only	47	N/A
Gambia only	84	N/A
Mali only	58	N/A
Bangladesh only	75	N/A
Kenya + The Gambia	133	86
Kenya + Mali	112	65
Kenya + Bangladesh	138	91
Gambia + Bangladesh	166	82
Mali + Gambia	167	109
Mali + Bangladesh	169	111

When computing data with another policy domain, each country saw an increase in the number of features detected to be significantly different between case and control groups.

Metagenomic codes

We also evaluated our implementation on the genetic marker data that showed the greatest identification power in the metagenomic codes analysis [23]. The data are also from the HMP and consists of a total of 85 samples and 221,111 features. Due to the large number of features and sparsity of the data, we implemented a filtering garbled circuit in which we first return a vector to each party denoting if a given feature meets a presence cutoff and then have each party input those features into our existing implementations to compute the statistical test. For χ^2, the 1,729,851,751 gate circuit (circuit size of 7823 Non-Free gates per feature) is evaluated in 67.4 min, with 51,926.35 MB sent to the evaluator, and 1 642.53 MB sent to generator. For odds ratio, the 632,918,505 gate circuit is evaluated in 33.18 min, with

20,542.84 MB sent to the evaluator, and 1,642.29 MB sent to generator. This result shows that the secure comparative analyses we would like to perform are possible given the legitimate concerns raised by Franzosa *et al*.

5.4 Discussion

In this section, we describe related work and provide a context for our contribution. We also discuss a use case for our solution in building datasets and finally present conclusions we formed during our work.

Related work

As we are the first, to our knowledge, to approach secure microbiome analysis, we review related work on privacy-preserving operations over human DNA.

Secure DNA sequence matching and searching

Comparing two DNA segments is essential to genome alignment and identifying the presence of a disease-causing mutation. One approach is to use an oblivious finite state machine for privacy-preserving approximate string matching[76]. FastGC, the predecessor of the FlexSC library, was benchmarked by computing Levenstein distance and the Smith-Waterman algorithm between private strings held by two parties[77]. More recently, Wang *et al.* (2015) compute approximate edit-distance using whole genome sequences[78].

Privacy-preserving Genome-wide association studies

Prior work has shown that secure computation between two institutions on biomedical data is possible by using a three-party secret-sharing scheme[71]. The authors present an implementation of a χ^2-test over SNP data using the Sharemind framework. Other researchers have presented a modification of functional encryption that enables a person to provide her genome and phenotype to a study but only for a restricted set of functions based on a policy parameter[79].

Prior works have built systems for genomic studies using different cryptographic protocols, including systems using additive homomorphic encryption[80] and systems using fully homomorphic encryption[81]. When compared with these works, we use a garbled circuit protocol with circuits for floating-point operations. Our system has two unique advantages compared to these prior works: (1) We can benefit from a long line of work on improving the practicality of garbled circuits [77,75,82] and (2) Floating-point operations ensure us a small and bounded error even after multiple operations.

Secure genetic testing

For using sequencing results in the clinical realm, paternity determination and patient-matching is possible using private set intersection[83]. Also, it is feasible to utilize homomorphic encryption for implementing disease-risk calculation without revealing the value of any genomic variant[84].

Patient pool

A novel application of multi-party secure computation approaches to genomic analysis are patient pool designs that can benefit patient groups, specifically those suffering from rare diseases or those with insufficient data in existing repositories for association studies. The recent announcement by 23andMe to begin drug development on its genome variant datasets highlights the value of biomarker data. We imagine a scenario where individuals can use our solution to create and manage datasets in order to charge drug developers to run analysis functions over the data. The companies will have to be non-colluding as otherwise all function results could be shared among companies. The current regulatory process for drug development allows a mechanism to enforce this constraint.

The patient pool can be paid to compute a function to over its data and sign the output. Upon requesting drug trial permission in the USA, a company is required to hand over all data from research, which in this case would include the output of the patient pool analysis and signatures over those results. The FDA could verify the signatures to enforce non-collusion between companies. This provides a mechanism to create high-quality datasets that are accessible to a variety of companies and ensure patients are compensated for their efforts.

5.5 Conclusions

In this article, we show that it is possible to perform metagenomic analyses in a secure computation framework. Our implementation made use of precomputation steps to minimize the number of operations performed in secure computation making the use of garbled circuits feasible. We also implemented sparse-matrix methods for each statistic. We took this step in order to prove the applicability of this solution for other analyses when the data itself acts as sufficient statistics, such as for the Wilcoxon rank-sum test. We also explored potential applications of our implementation in patient pool designs.

Although the storage and sharing of medical data is ultimately a policy matter, providing a technical solution is useful to forming good policy. We believe that given the time costs associated with re-consenting patients to release data to another researcher or creating a legal contract stipulating a data receiver's responsibility, that the running times we presented for metagenomic analyses are a reasonable tradeoff.

DNA-sequencing technologies are entering a period of unprecedented applicability in clinical and medical settings with a concomitant need for regulatory oversight over each individual's sequencing data. We believe that addressing privacy concerns through computational frameworks similar to those used in this article is paramount for patients while allowing researchers to have access to the largest and most descriptive datasets possible. We expect that secure computation and storage of DNA sequencing data, both the individual’s DNA and their metagenomic DNA, will play an increasingly significant role in the biomedical research and clinical practice landscape.

Chapter 6: Conclusion

In this dissertation we present software infrastructure and new visualization approaches for investigating microbiome data. We center the work around the data analysis ideal of successive rounds of exploratory and confirmatory analysis. Visualization is a vital component of exploratory analysis and Metaviz includes a navigation utility suited for hierarchical microbiome data. We developed Metaviz to interface with the *metagenomeSeq* Bioconductor package for microbiome differential abundance statistical testing. We support interactive, exploratory visual analysis through the *metavizr* Bioconductor package to produce visualizations based on statistical analysis results. We developed infrastructure for integrative analysis across multiple datasets from the Human Microbiome Project as well as implemented statistical tests in a secure data sharing mechanism. With microbiome sequencing projects moving towards the study of microbial community perturbation and functional community profiles, we developed a novel longitudinal visualization for multiple features along with a mechanism for inspecting functional and taxonomic hierarchies.

Specific Contributions

1. Metaviz – Interactive visualization for exploratory analysis of community taxonomic profile data. Metaviz is a web application for visualization of microbiome community abundance profile data. The application can visualize marker-gene or whole metagenome shotgun sequencing data. Metaviz introduces a navigation utility for the taxonomic hierarchy.

2. Metaviz integration with the Human Microbiome Project (HMP) Data Infrastructure. We describe the design and implementation of linking between the HMP Data Portal and Metaviz. Also, we present an analysis of a subset of data from the HMP using Metaviz and *metagenomeSeq*.

3. Microbial community longitudinal and functional profiling visualizations in Metaviz. This work expands the visualizations available in Metaviz for longitudinal data using sparklines as the entries of a

heatmap to show trends across the set of features. This work also introduces an interactive filter for community functional profile data using the navigation mechanism in Metaviz, provides a mechanism to import and export taxa of interest, and connects Metaviz to external information sources.

4. Privacy-preserving microbiome analysis using secure computation. In 2015, Franzosa *et al*. showed that it was possible to use microbiome features to identify individuals at different time points in the HMP dataset[23]. This work implements statistical analysis functions using a library for secure multi-party computation. The goal of this project is to allow researchers to compute analyses over shared microbiome abundance matrices without revealing the underlying counts directly.

Future Work

Metaviz enables exploratory analysis of microbiome feature count data and the results of confirmatory analysis. However, to fully realize the data analysis model championed by Tukey, we need a formal method to incorporate the results of interactive, visual exploratory analysis into confirmatory analysis. We envision two general approaches, one with feature selection for statistical tests based on the visualization results and another using the visualization to provide a list of recommended statistical tests or parameter selections. An issue with feature selection when applied to domains with large numbers of features such as genomics is handling the multiple testing problem. One method used in gene expression analysis filters features based on independent values to those under test such as median and variance but that method is not easily extended to a visualization-based selection as the filter would not be independent[85].

A concern with Metaviz and interactive exploratory data analysis in general is that p-hacking can arise from failing to provide an appropriate mechanism for incorporating statistical testing. An example is an analyst visually finding interesting associations and then performing confirmatory statistical tests on only those features without properly correcting for the multiple comparisons. One possible method to address this problem is to split the dataset in two randomly and perform exploratory on one part then using the hold out data for confirmatory analysis. Zgraggen *et al*. examined the issue of multiple comparison problem with interactive visualization and describe an approach using a procedure where implicit hypothesis tests from exploratory analysis are modeled along with explicit confirmatory analyses to control the false discovery rate[86]. The authors compare the results of this procedure against using a holdout validation dataset for all confirmatory analysis. As the authors highlight, keeping a holdout confirmation dataset is prohibitively expensive in many instances.

Another mechanism for operating with a holdout dataset is differential privacy. With this approach, the holdout dataset can be used repeatedly. Differential privacy provides guarantees that functions computed on a dataset are not distinguishable based on the value for a single entry. The holdout procedure operates with the function run on the available dataset then a differential privacy mechanism computes if the difference between the analysis function's result on the holdout dataset is close to the result on the training dataset within a threshold[87].

Visualizations could also be used as a test statistic directly with prior work showing the usefulness of this approach. Specifically, Wikham *et al*. develop a software package for visual statistical inference – first one for sampling from a null distribution of a dataset then presenting plots of that along with a second mechanism to sample from the given dataset[88]. The methods are named the Rorschach and Lineup techniques[89]. Majumder *et al*. apply the Lineup protocol to linear model and test with human subjects to identify if a trends in the data can be identified[90]. In related work on user ability to identify significant trends from a visualization, instructors of a massive online course tested if students inferred an association between two variables in a course work assignment[91]. These findings collectively support the idea that visualization can be suitable as a statistical test and filtering mechanism.

Another avenue of continued development of Metaviz is a recommendation system for visualizations or microbiome features of interest. We could adapt collaborative filtering techniques to identify visualizations or features that users could find interesting[92]. Collaborative filtering generally operates over measures of similarity between objects. The Voyager tool showed recommendations to be a successful approach for an interactive visualization tool[93]. In Metaviz, the similarity measures between visualizations could include summary statistics of the underlying data, distance measurements within data under examination such as the dendrogram clustering metric in the heatmap, and measures of the graphic itself including intensity of blue or total number of colored pixels.

The architecture of the Metaviz web application allows for implementing solutions to the multiple comparison problem, using visualizations as statistical tests, and a recommendation system. For instance, the data import utility could split a dataset into training and holdout subsets. A user could perform interactive visualization with the training set and then test the results of any associations using a differentially private mechanism with the holdout dataset also stored in the Metaviz backend. The web application architecture of Metaviz offers an opportunity to employ interaction logging to identify which utilities, datasets, and visualizations analysts use most frequently. Interaction logging with Metaviz could also help with addressing multiple comparison problem. The recommendation system could be refined through the interaction log data to identify similar users, visualizations for a given dataset, or similar features in other datasets.

In addition to robust exploratory and confirmatory data analysis tools, data access and data sharing are critical to advancing microbiome sequencing studies. To address concerns about data privacy and long-term data storage, security protocols could be incorporated to several Metaviz utilities. Beyond a statistical testing holdout mechanism, differential privacy protocols could be used with visualization and exploratory analysis of datasets in a could help users that need to investigate data to which they do not have direct access. Also, visualizations that rely on computed measurements or aggregations could work with secure computation protocols as the individual data counts of a matrix should not be revealed directly. Stacked bar plots that show the proportion of a feature in a given sample are one appropriate visualization. The incorporation of secure data-sharing protocols into large sequencing-based consortium projects will be vital to share data broadly. Interactive visualization for exploratory analysis coupled with robust confirmatory analysis and secure data sharing utilities is vital to advance computational biology.

Appendix

Appendix A.

Section I: Using Information Visualization Techniques for Microbiome Data

Our design for the visualization layer is motivated by results in the information visualization literature for displaying large tree structures with associated complex data. In this section, we provide a brief review of pertinent visualization techniques. To provide a basis for our design decisions, we present metagenomic visual analysis operations in relation to the Task by Data Type Taxonomy for Data Visualization[94]. In microbiome sequencing projects, sample data is multi-dimensional with study-specific attributes, e.g. age, sex, gathered in each experiment. Feature data is tree-structured with a node fan-out dependent on the bacterial hierarchy of the annotation database and the ecological community observable in each sample.

We review the tasks presented by Shneiderman for completeness. These consist of the following: 1) *Overview*: gain an overview of the entire collection; 2) *Zoom*: Zoom in on items of interest; 3) *Filter*: filter out uninteresting items; 4) *Details-on-demand*: Select an item or group and get details when needed; 5) *Relate*: View relationships among items; 6) *History*: Keep a history of actions to support undo, replay, and progressive refinement; 7) *Extract*: Allow extraction of sub-collections and of the query parameters[94]. Our task taxonomy below builds upon and generalizes the description of features presented in the Krona interactive visualization tool[29], also based on the Shneiderman interactive visualization task taxonomy.

We now discuss the specific operation and goal for each task with regards to microbiome analysis. The *overview* task consists of examining global patterns in feature abundance among samples across levels of the taxonomic hierarchy. This task is also accomplished by presenting statistics that summarize feature variance and observed ecological diversity. The *zoom* task requires navigation to the lowest levels of the feature hierarchy as well as inspection of individual sample data. The *filter* task consists of removing or expanding taxonomic features and samples. With taxonomic community profile data, several operations need to be enabled, first a level-wise filtering and then removal of features at a given depth along with aggregating to a specific point in the hierarchy. *Details-on-demand* includes showing all children of a given node, text-based search for features that contain a character string, and the utility to visualize the same data in different views. *Relate* is enabled by linking multiple data views with the feature hierarchy along with group-by and color-by operations over sample attributes. *History* requires keeping track of the current position during navigation of a feature hierarchy as well as the ability to select and remove nodes as desired. Finally, *extract* entails capturing the parameters to recreate an analysis. Specific to microbiome analysis, the *extract* task also should encompass providing a mechanism to interoperate between annotation databases and retrieving cluster center sequences from a dataset.

Section II. Data Plots and Charts

We provide several visualizations of feature count data. These allow the user to explore relationships between sample phenotype and metagenomic features. The first is a heatmap with rows as samples and columns as features[95]. The heatmap is an interactive component from which a user can select to show a dendrogram of a dynamic clustering over features or samples. If the user chooses not to employ clustering, rows can be re-ordered based on a sample metadata attribute. We also provide several utilities on the samples including color-by and modifying the displayed name of any sample attribute. Figure 1 shows a heatmap of the *msd16s* dataset with the colors for sample rows set based on dysentery status.

Metaviz heatmaps include dendrograms that are calculated with commonly used distance metrics. In addition to Euclidean distance, the following dissimilarity measures are available based on the implementations in the vegan R package [https://cran.r-project.org/web/packages/vegan/index.html]: Manhattan, Canberra, Bray-Curtis, Kulczynski, Jaccard, Gower, Morisita, Horn-Morisita, and Binomial. Another visualization in Metaviz is the stacked bar plot that shows the proportion of features in each sample. A column is a sample or group of samples, a row represents the bin counts for that feature, and each row is colored by taxa that is linked for highlighting to the FacetZoom and all other charts in the workspace. On the stacked plot, we implemented a group-by function to aggregate samples based on a sample metadata attribute. This plot is useful for comparing microbial community composition between individual samples or groups. Figure 3 shows two stacked bar plots that are split based on sample dysentery status and grouped by age range.

Metaviz supports scatter plots to visualize feature count values of selected samples in a X, Y coordinate plane. A scatter plot is useful for fast identification of distribution and spread across measurements. The scatter plot has a color-by feature to color points based on a specific sample metadata attribute. In addition, we include PCA and PCoA scatter plots for community level analysis. For instance, a PCA plot is shown in the upper right corner of Figure 1. Another scatter plot is the PCoA plot that is shown in the upper right side of Appendix A Figure 5.

Further, Metaviz includes a line plot with each line representing a feature, the height of the line denoting abundance, and the samples across the X-axis. We find the line plot useful for examining time-series data.

Metaviz allows a user to generate a boxplot of alpha diversity values for selected samples. Boxes can be generated for samples belonging to a metadata attribute for example case or control status. Appendix A Figure 6 shows a Metaviz workspace with an alpha diversity boxplot.

All data plots and charts added to the workspace are linked to the feature nodes on the FacetZoom. Hovering over a feature column in a heatmap highlights that feature in all other plots as well as the path through the hierarchy for that feature in the FacetZoom. This brushing and linking is essential to providing integrative visual analysis. Also, each plot and chart has a toolbar that can be used to modify presentation settings, the color scheme, saving the chart, and writing custom JavaScript for that chart. The toolbar is shown in the upper right-hand corner of the FacetZoom in Appendix A Figure 1. Individual charts can be saved as SVG or PDF files. Metaviz also allows users to render complete workspaces as PDF files. The process captures each SVG chart in the workspace and combines the individual charts to generate a single page. Alternatively, since Metaviz is a JavaScript application and it cannot send requests to the browser to generate a screenshot, users can capture a static image of the workspace that shows brushing or linking across charts, the user will need to use the browser screenshot function.

Section III: Exploration of MSD childhood diarrhea study in developing countries

In the main paper, we discuss results of visual and statistical analysis of Bangladesh samples in the MSD dataset. In this Section, we discuss results for the other three countries. Building the same Metaviz plots for The Gambia, we note that the number of control samples outweighs the number of case samples and no case samples from the 0-6 month age range are present. Examining the heatmap and interactive boxplots, the following taxa are more abundant in case than control samples: Actinomycetales, Lactobacillales, Campylobacterales, Enterobacteriales, Pasteurellales, Pseudomonadales, Actinomycetaceae, Micrococcaceae, Carnobacteriaceae, Streptococcaceae, Campylobacteraceae, Enterobacteriaceae, Pasteurellaceae, Moraxellaceae, Porphyromonadaceae, *Actinomyces*, *Rothia*, *Granulicatella*, *Streptococcus*, *Campylobacter*, *Citrobacter*, *Dickeya*, *Escherichia*, *Klebsiella*, *Shigella*, *Haemophilus*, *Acinetobacter*, *Parabacteroides*, *Rothia mucilaginosa*, *Granulicatella* *adiacens*, *Granulicatella sp. oral clone ASCG05*, *Streptococcus mitis*, *Streptococcus oralis*, *Streptococcus parasanguinis*, *Streptococcus sanguinis*, *Streptococcus sp. C101*, *Streptococcus sp. oral clone ASCC01*, *Streptococcus sp. oral clone ASCE09*, *Citrobacter freundii*, *Erwinia chrysanthemi*, *Escherichia coli*, *Klebsiella pneumoniae*, *Haemophilus haemolyticus*, *Haemophilus parainfluenzae*, *Haemophilus sp. oral clone BP2-46*. While the following taxa are more abundant in control samples than case: Bacteroidales, Clostridiales, Prevotellaceae, Eubacteriaceae, *Prevotella*, *Eubacterium*, *Prevotella copri*, *Prevotella histicola*, *Prevotella sp. BI-42*, *Prevotella sp. DJF_B112*, *Prevotella sp. DJF_B116*, *Prevotella sp. DJF_LS16*, *Prevotella sp. DJF_RP53*, and *Prevotella sp. oral clone BP1-28*. Examining the stacked plots, we first notice the proportion of Bacteroidales increases with age in the control samples as compared to the dysentery group. Lactobacillales decreases in proportion as age increases for both the case and control samples with a large decrease from 18-24 to 24-60 months in the case samples. In the case samples, Enterobacteriales has one of the highest proportions orders at 0-6 months, decreases for both 12-18 months and 18-24 months, but is then the highest proportion order in the 24-60 month interval. Appendix A Figure 7 shows the heatmap and stacked plot Metaviz workspace for The Gambia.

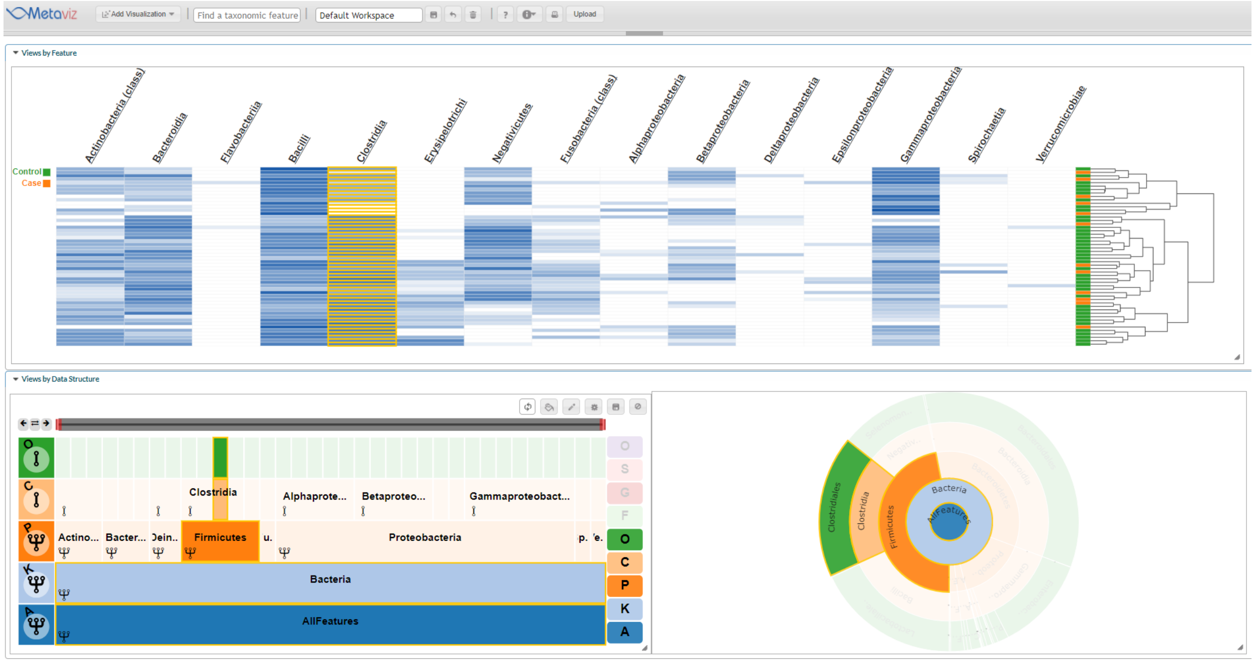
Using *metagenomeSeq*, we find the following taxa to have significant difference in abundance: Actinomycetales (1.13, 1.49E-02), Enterobacteriales (1.85, 5.20E-03), Pasteurellales (2.02, 2.00E-07), Bacteroidales (-1.36, 5.73E-03), Actinomycetaceae (1.14, 1.96E-02), Carnobacteriaceae (2.13, 2.04E-07), Enterobacteriaceae (1.83, 7.33E-03), Pasteurellaceae (2.01, 2.04E-07), *Actinomyces* (1.14, 3.00E-02), *Granulicatella* (2.13, 6.64E-07), *Escherichia* (1.88, 9.40E-03), *Haemophilus* (1.95, 9.74E-07), *Granulicatella adiacens* (1.88, 1.67E-04), *Granulicatella elegans* (1.70, 4.33E-03), *Granulicatella sp. Oral clone ASCG05* (2.64, 4.46E-07), *Streptococcus mitis* (1.64, 4.33E-03), *Streptococcus sanguinis* (1.15, 4.03E-02), *Streptococcus sp. C101* (1.24, 2.19E-02), *Streptococcus sp. Oral clone ASCC01* (2.12, 7.15E-08), *Streptococcus sp. Oral clone ASCE09* (1.49, 4.33E-03), *Escherichia coli* (1.88, 9.54E-03), *Haemophilus haemolyticus* (1.73, 3.71E-03), and *Haemophilus parainfluenzae* (2.03, 4.50E-06). We present the *metagenomeSeq* differential abundance calculations for The Gambia in Appendix A Table 2.

Inspecting the Kenya samples with Metaviz we noticed there are far fewer samples with dysentery than non-dysentery samples. From the heatmap and boxplots, we observed the following taxa as more abundant in the case samples than across the control samples: Actinomycetales, Selenomonadales, Campylobacterales, Enterobacteriales, Pasteurellales, Veillonellaceae, Campylobacteraceae, Enterobacteriaceae, Pasteurellaceae, *Megasphaera*, *Veillonella*, *Campylobacter*, *Citrobacter*, *Enterobacter*, *Escherichia*, *Klebsiella*, *Shigella*, *Haemophilus*, *Veillonella parvula*, *Veillonella sp. HF9*, *Veillonella sp. oral clone VeillC8*, *Veillonella sp. oral clone VeillD5*, *Enterobacter cancerogenus*, *Enterobacter cloacae*, *Escherichia coli*, *Escherichia sp. oral clone 3RH-30*, *Klebsiella pneumoniae*, *Haemophilus haemolyticus*, and *Haemophilus parainfluenzae*. Correspondingly, we find the following more abundant in control over case: Bacteroidales, Prevotellaceae, *Prevotella*, *Prevotella copri*, *Prevotella histicola*, *Prevotella sp. BI-42*, *Prevotella sp. DJF_B112*, *Prevotella sp. DJF_B116*, and *Prevotella sp. DJF_RP53*. As for changes across age ranges and case/control status, Campylobacterales is more prevalent in 0-6, 6-12, and 12-18 month age ranges in the case group than the control group. Appendix A Figure 8 shows the visual analysis of the Kenya samples.

Using *metagenomeSeq*, we find the following taxa to have significant difference in abundance: Pasteurellales (1.29, 7.36E-03), Pasteurellaceae (1.29, 1.25E-02), *Enterobacter* (1.05, 5.54E-02), *Haemophilus* (1.29, 1.99E-02), *Veillonella sp. Oral clone VeillD5* (1.15, 8.20E-02), *Enterobacter cancerogenus* (1.45, 4.12E-02), *Escherichia sp. Oral clone 3RH-30* (1.07, 5.76E-02), and *Haemophilus haemolyticus* (1.63, 4.12E-02).

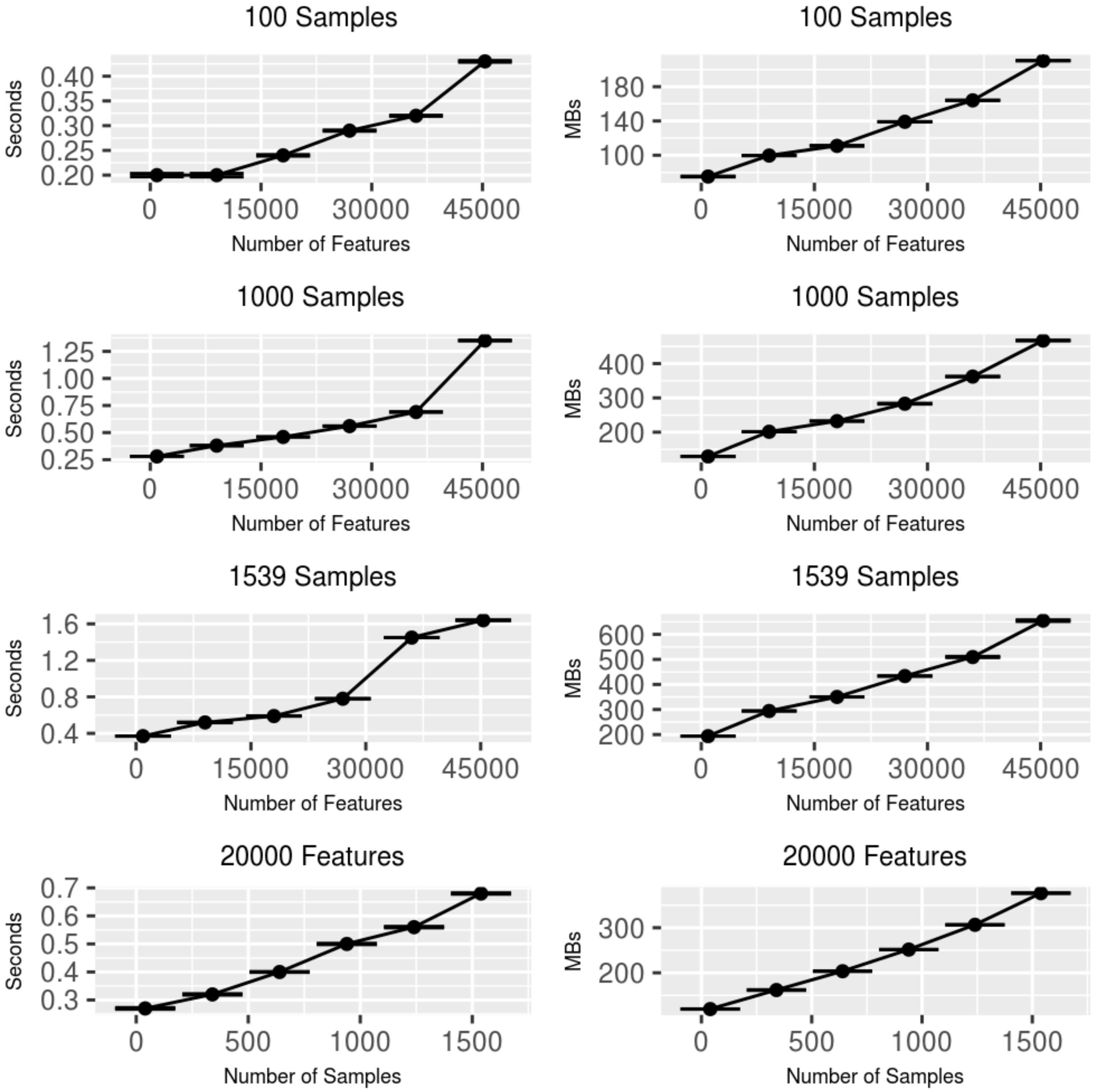
From the Metaviz plots for Mali samples, we note that the number of case samples is far smaller than the number of control samples and no case samples are from the 0-6 month age range. Examining the heatmap and boxplots, the following taxa show greater abundance in the case samples compared to control: Actinomycetales, Neisseriales, Fusobacteriales, Enterobacteriales, Pasteurellales, Pseudomonadales, Actinomycetaceae, Micrococcaceae, Neisseriaceae, Fusobacteriaceae, Enterobacteriaceae, Pasteurellaceae, Moraxellaceae, *Actinomyces*, *Rothia*, *Neisseria*, *Citrobacter*, *Dickeya*, *Enterobacter*, *Escherichia*, *Klebsiella*, *Shigella*, *Haemophilus*, *Acinetobacter*, *Rothia mucilaginosa*, *Citrobacter freundii*, *Erwinia chrysanthemi*, *Enterobacter cancerogenus*, *Enterobacter cloacae*, *Escherichia albertii*, *Escherichia coli*, *Escherichia sp. oral clone 3RH-30*, *Klebsiella pneumoniae*, *Shigella boydii*, *Shigella sonnei*, *Haemophilus parainfluenzae*, *Haemophilus sp. oral clone BP2-46*, and *Acinetobacter sp. SF6*. In contrast, these taxa exhibit higher abundance in control samples compared to the case samples: Bifidobacteriales, Bacteroidales, Clostridiales, Bifidobacteriaceae, Bacteroidaceae, Prevotellaceae, *Bifidobacterium*, *Bacteroides*, *Prevotella*, *Bifidobacterium longum*, *Bacteroides fragilis*, *Prevotella copri*, *Prevotella histicola*, *Prevotella sp. BI-42*, *Prevotella sp. DJF_B112*, and *Prevotella sp. DJF_RP53*. From the stacked plots, the proportion of Enterobacteriales among case samples in age range 6-12 and 12-18 months is much higher than that in the same age ranges for control samples. For dysentery samples, Pasteurellales shows a much higher proportion in the 18-24 month age range than for normal samples. Also, across all age ranges Bacteroidales is more prevalent in the control samples. Appendix A Figure 9 shows the visual analysis of samples from Mali.

Using *metagenomeSeq*, we find the following taxa to have significant difference in abundance: Neisseriales (1.58, 7.33E-02), Pasteurellales (2.97, 5.51E-05), Neisseriaceae (1.58, 8.33E-02), Pasteurellaceae (2.96, 9.63E-05), *Neisseria* (1.69, 9.78E-02), *Escherichia* (1.62, 9.78E-02), *Haemophilus* (2.94, 2.03E-04), *Haemophilus parainfluenzae* (2.87, 1.54E-04), *Haemophilus sp. Oral clone BP2-46* (2.39, 3.65E-03), and *Prevotella sp. DJF_RP53* (-2.91, 5.58E-02).

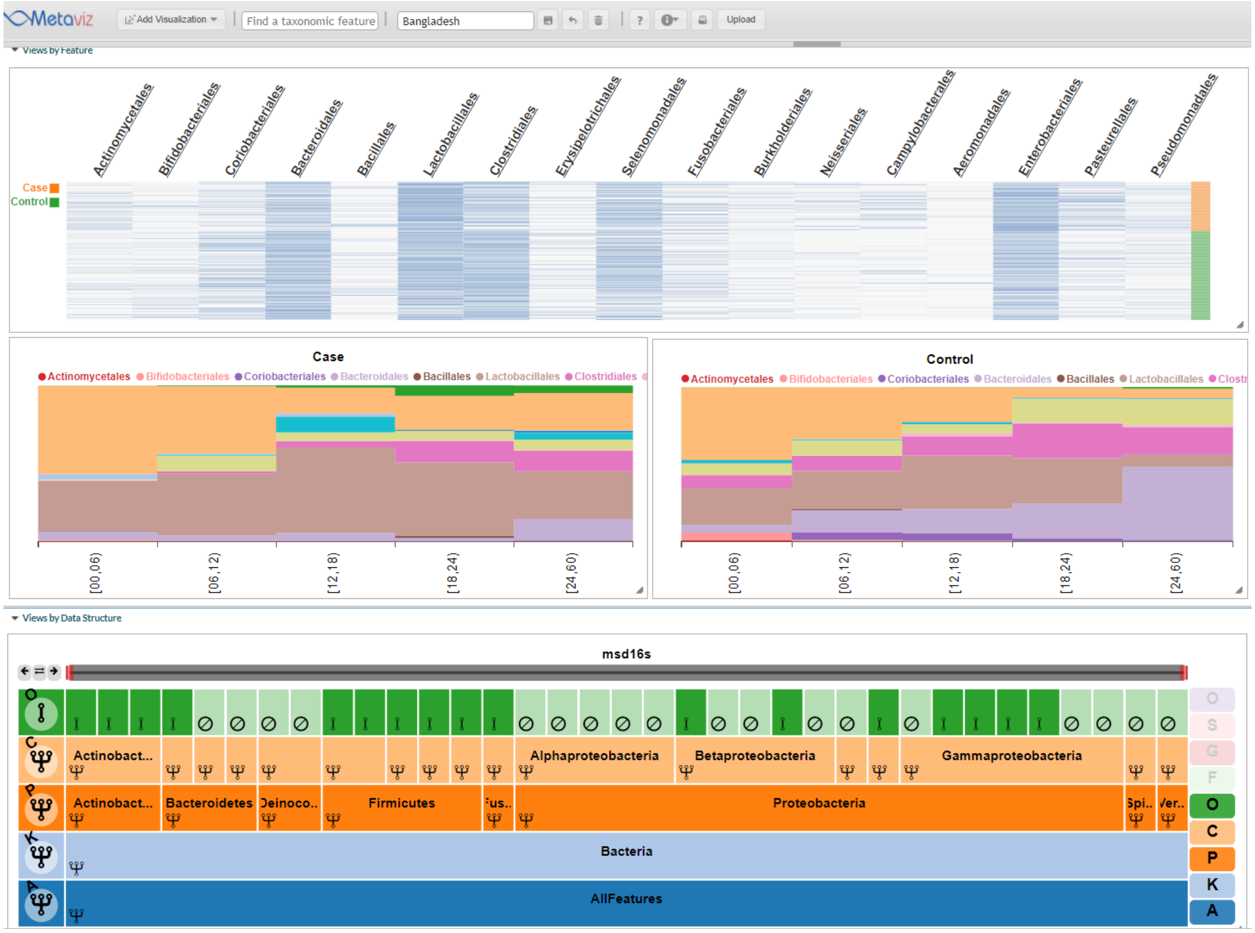


Appendix A Figure 1: Sunburst Plot

The heatmap shows 52 samples from the msd16s dataset. The sunburst diagram next to the FacetZoom is a circular taxonomy that enables viewing the lineage and hierarchy of the dataset during exploration. The sunburst is linked to all other charts in the workspace, so the lineage of a taxonomic feature is highlighted when hovering on that feature in any other chart.

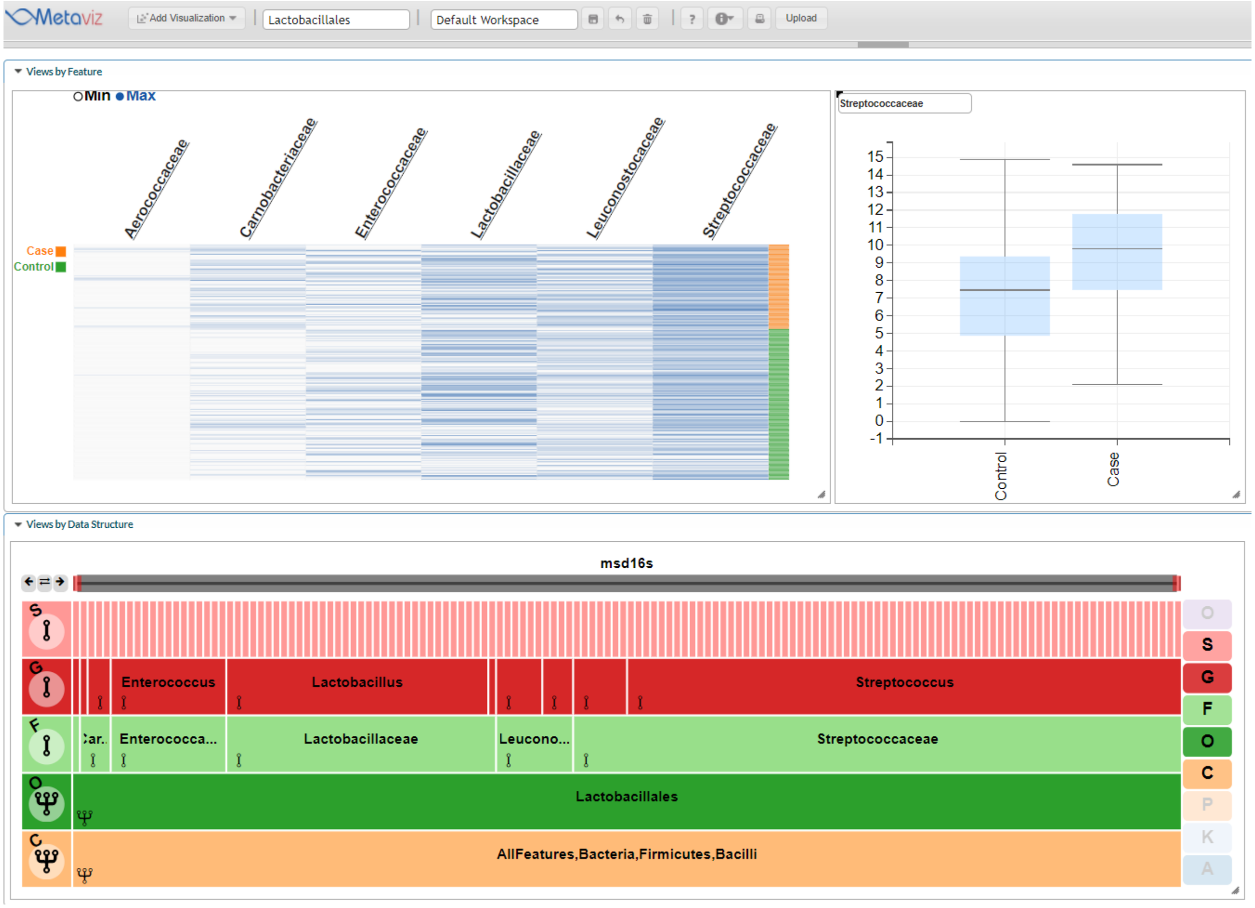


Appendix A Figure 2: metavizr benchmark
HMP dataset with 1539 samples, 45336 features, and a 7-level hierarchy. The Rprof library was used for profiling. The benchmark consisted of an aggregation query to the 3rd level of the hierarchy. The top panels show tests for keeping the number of samples at 100 and increasing the number of features over which the aggregation query is operating. The top left panel shows the aggregation query completion time in seconds and the top right panel shows the highest memory footprint in MBs during the query execution. The next two rows of show the performance on 1000 samples then all samples in the dataset, respectively. The bottom row shows keeping the number of features fixed at 20000 and increasing the number of samples. From this benchmark, datasets above 1000 samples, 25000 features, and a 7-level hierarchy are recommended to use the graph database backend for interactive query processing.



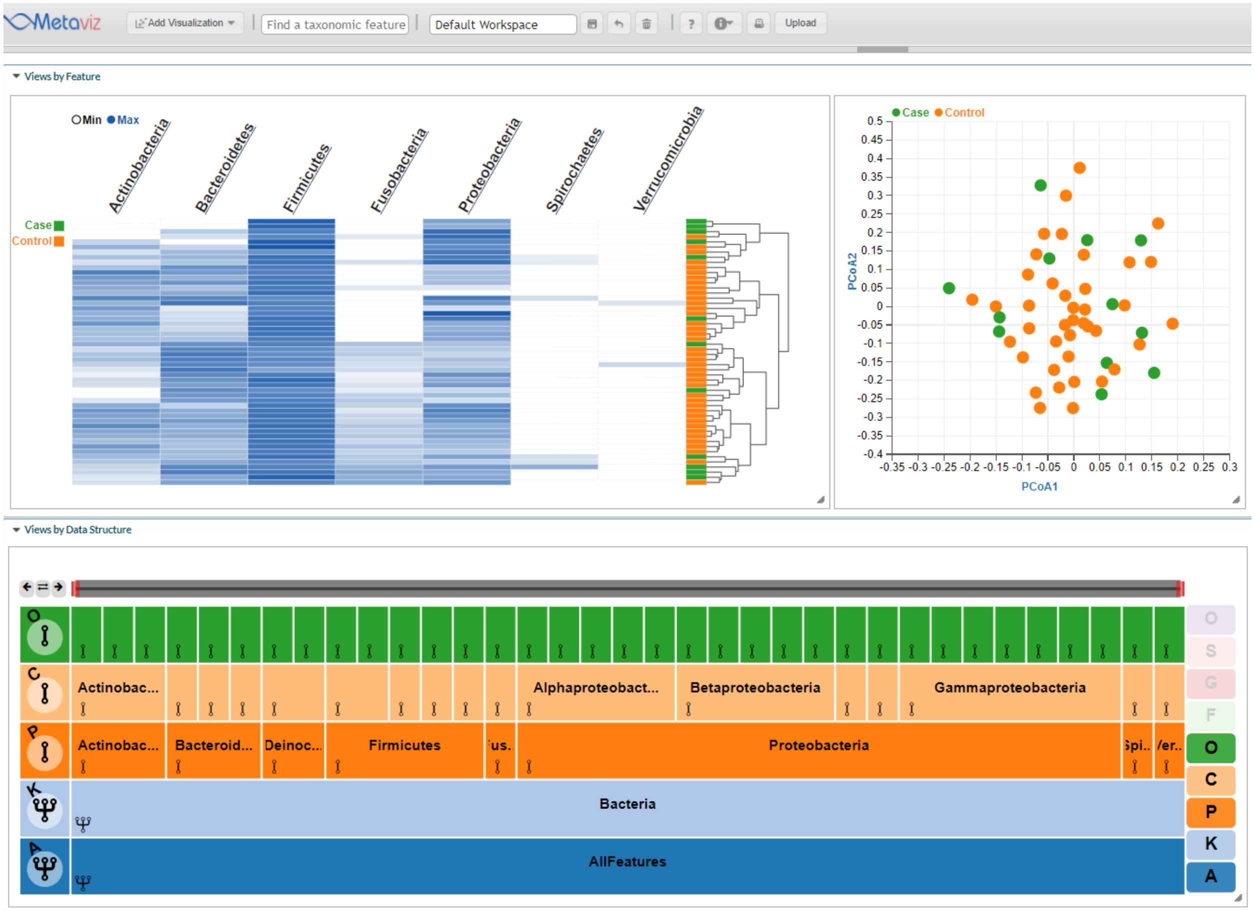
Appendix A Figure 3: Bangladesh msd16s visual analysis

From the heatmap, Actinomycetales, Enterobacteriales, Lactobacillales, Pasteurellales, and Pseudomonadales appear more abundant in the case group than the control group. Correspondingly, Coriobacteriales, Bacteroidales, and Clostridiales display higher abundance in the control samples as compared to the case samples. Using *metagenomeSeq*, Enterobacteriales (1.38, 1.46E-04), Pasteurellales (2.47, 4.16E-12), Coriobacteriales (-1.38, 9.88E-04), Bacteroidales (-1.19, 7.56E-04), and Clostridiales (-1.09, 6.45E-04) are differentially abundant while Actinomycetales (9.73E-01, 2.40E-03), Lactobacillales(1.15, 7.00E-01), and Pseudomonadales (5.36E-01, 1.05E-01) are not. Looking at the stacked bar plots, Bacteroidales shows a higher proportion in control than case samples at all intervals after 0-6 months. Finally, Clostridiales has lower proportion in case than control samples for the intervals of 0-6, 6-12, and 12-18 months then similar proportion for the last two timepoints. This workspace is available at http://metaviz.cbcb.umd.edu/?ws=iGPCfth9nQn.



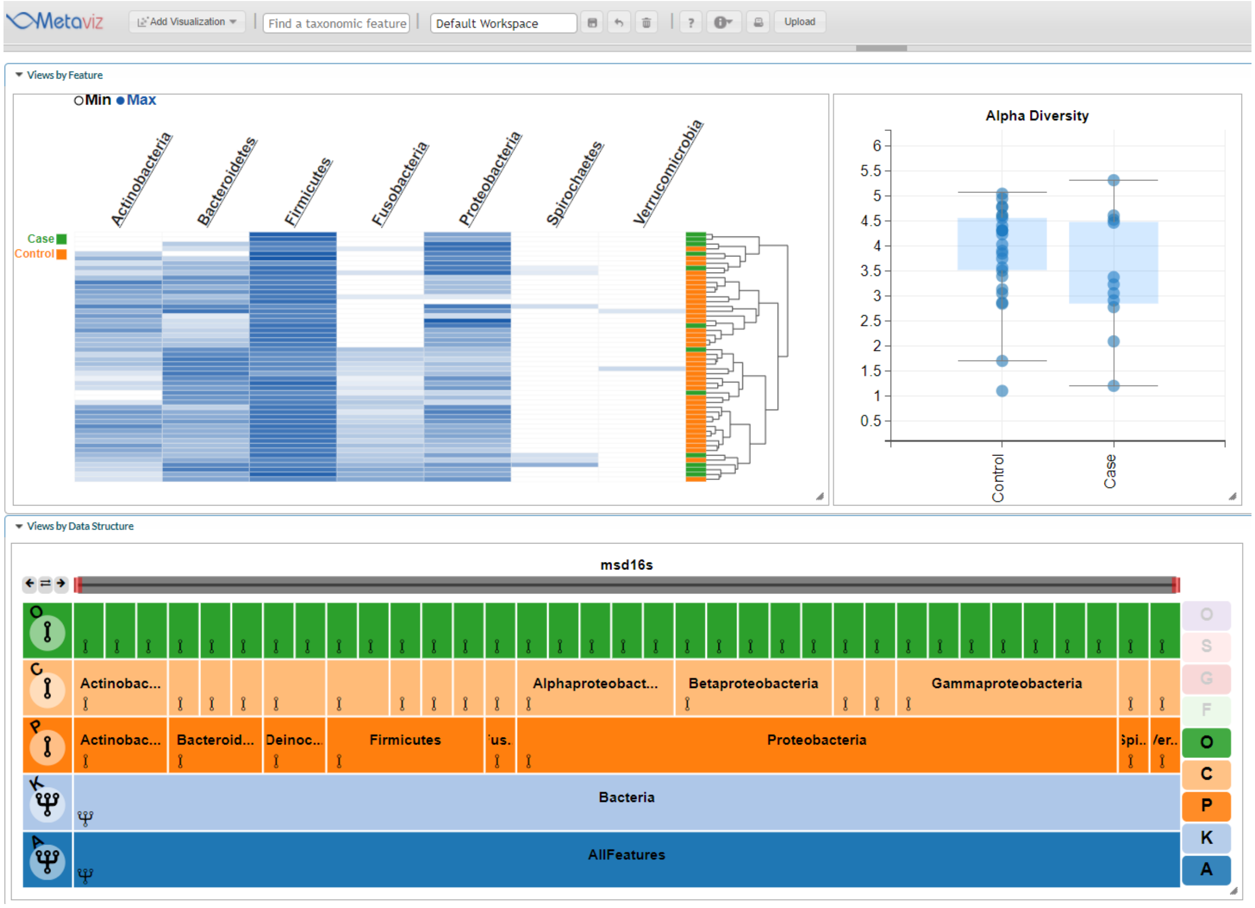
Appendix A Figure 4: Dynamic Boxplot

The order Lactobacillales in the heatmap with the family Streptococcaceae shown in the dynamic boxplot. The boxplot is generated by clicking the column name in the heatmap. The boxplot is separated into case and control samples as in the heatmap. The FacetZoom provides aggregation and filtering to a part of the taxonomy, the heatmap provides an overview of the aggregated counts for that region, and the boxplot provides details-on-demand for the specific feature of interest. Now statistical testing can be performed to find significance of the difference in abundance observed. Another round of aggregation at the genus level then inspection of each feature in species with a boxplot can be performed to inspect each level of hierarchy and observe trends in the dataset.



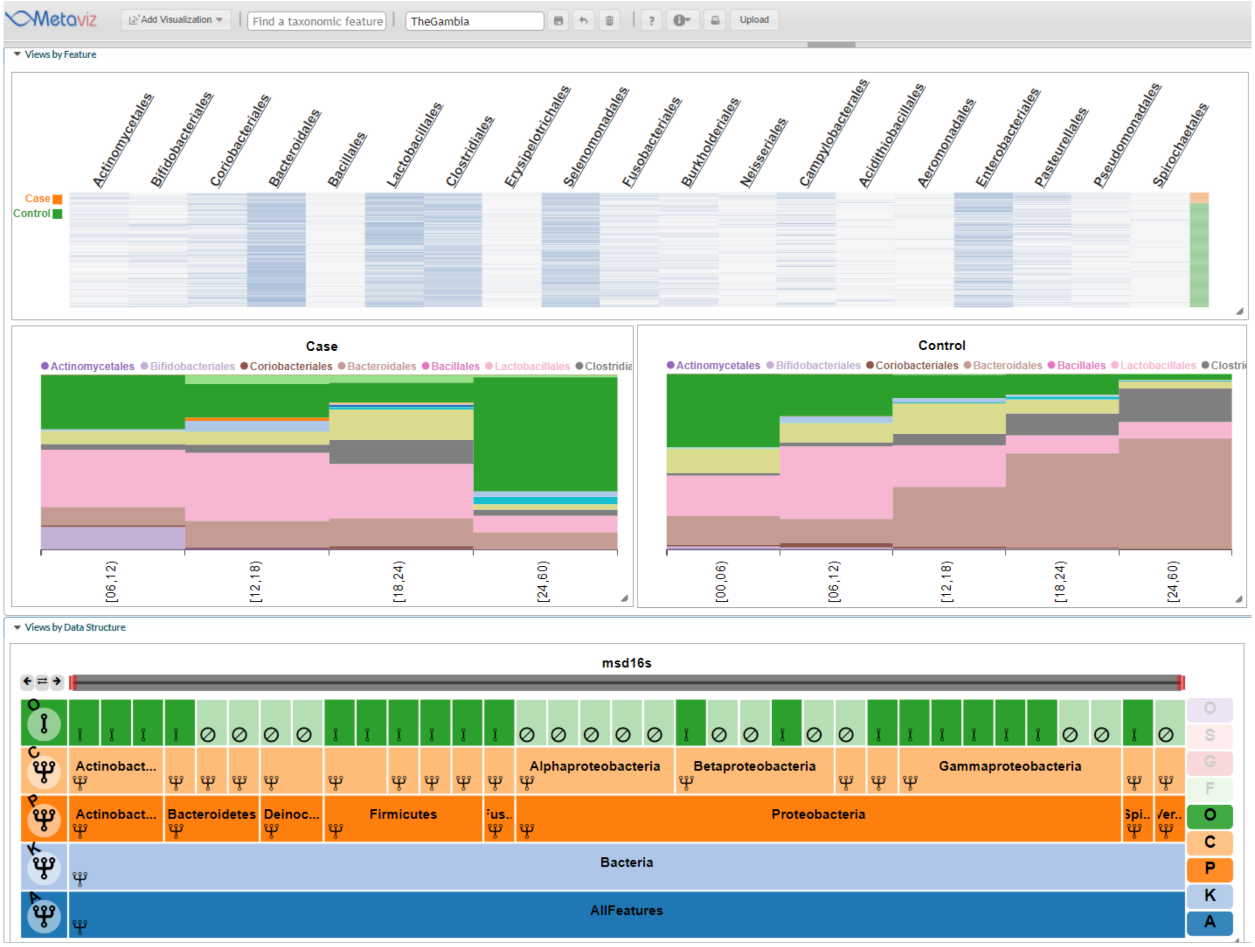
Appendix A Figure 5: PCoA Plot

The Metaviz workspace shows 52 samples from the msd16s dataset. The PCoA plot is computed over counts aggregated to the level selected in the FacetZoom control. The points are labeled based on a specified sample metadata field, in this instance dysentery case or control status.



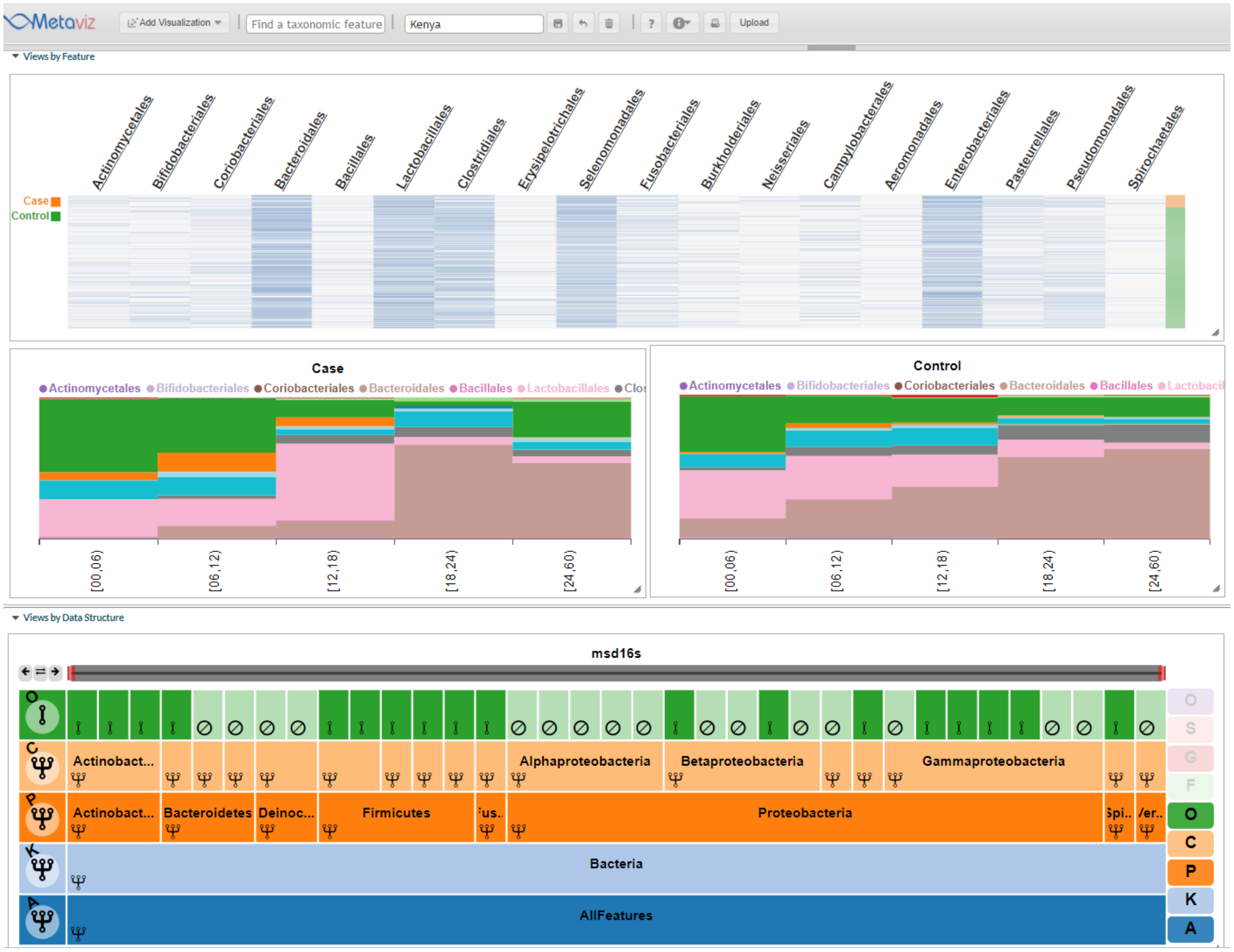
Appendix A Figure 6: Alpha Diversity Boxplot

The heatmap and boxplot displays 52 samples from the msd16s dataset. The alpha diversity boxplot is computed using Shannon Index. Samples in the boxplot can be separated on a metadata attribute with case and control dysentery status used in this example.



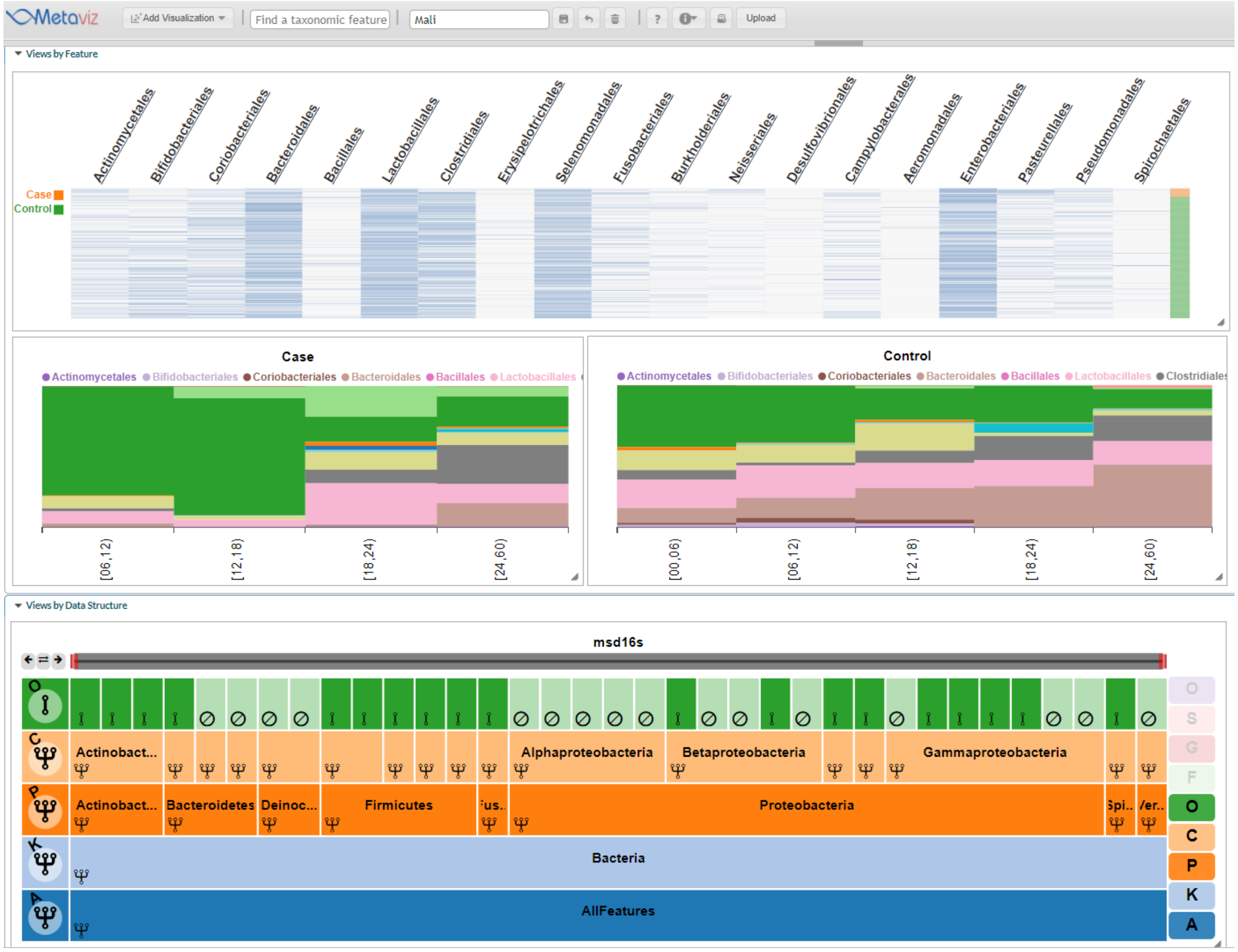
Appendix A Figure 7: The Gambia msd16s visual analysis

From the heatmap, it appears that Actinomycetales, Lactobacillales, Campylobacterales, Enterobacteriales, Pasteurellales, and Pseudomonadales are more abundant in the case samples than control samples. Bacteroidales and Clostridiales are more abundant in the control samples than case. From *metagenomeSeq*, we computed the following log-fold change and adjusted p-values: Actinomycetales (1.13, 1.49E-02), Enterobacteriales (1.85, 5.20E-03), Pasteurellales (2.02, 2.00E-07), Lactobacillales (9.21E-01, 1.42E-01), Campylobacterales (1.28E+00, 1.01E-01), Pseudomonadales (4.43E-01, 4.61E-01), Clostridiales (-4.54E-01, 4.61E-01), and Bacteroidales (-1.36, 5.73E-03). Examining the stacked bar plots, Bacteroidales shows higher proportion in control samples than case samples for 12-18, 18-24, and 24-60 month age ranges. Lactobacillales decreases in proportion as age increases for both the case and control samples, with a much large decrease from 18-24 to 24-60 months in the case samples. In the case samples, Enterobacteriales has among the highest proportion at 6-12 months, decreases in these samples at 12-18 and 18-24 months, then has the highest proportion in the 24-60 month interval. In control samples, Enterobacteriales has the highest proportion in 0-6 months and then decreases in proportion for each other age range. This workspace is available at http://metaviz.cbcb.umd.edu/?ws=Kd8O4u3zOEi.



Appendix A Figure 8: Kenya msd16s visual analysis

Examining the heatmap, Actinomycetales, Selenomonadales, Campylobacterales, Enterobacteriales, and Pasteurellales appear to be more abundant in the case samples than across the control samples. Bacteroidales appears more abundant in control over case. Using *metagenomeSeq*, Pasteurellales has a log fold-change of 1.29 and adjusted p-value of 7.36E-03 while Actinomycetales (4.29E-01, 4.04E-01), Selenomonadales (5.89E-01, 2.69E-01), Campylobacterales (1.39E+00, 2.69E-01), Bacteroidales (-8.32E-01, 2.69E-01), Enterobacteriales (8.54E-01, 2.69E-01), and are not differentially abundant. As for changes across age ranges and case/control status, it appears that Campylobacterales is more prevalent in 0-6, 6-12, and 12-18 in the case group than the control group. This is available at http://metaviz.cbcb.umd.edu/?ws=asrAc9DmK2p.



Appendix A Figure 9: Mali msd16s visual analysis

Actinomycetales, Neisseriales, Fusobacteriales, Enterobacteriales, Pasteurellales, and Pseudomonadales display increased abundance in the case samples as compared to the distribution in the control samples. Bifidobacteriales, Bacteroidales, and Clostridiales show greater abundance in control samples as compared to the case samples. With *metagenomeSeq*, we find support for these conclusions with Pasteurellales (2.97E+00, 5.51E-05) and Neisseriales (1.58E+00, 7.33E-02) but not with Actinomycetales (3.32E-01, 8.89E-01), Bifidobacteriales (-1.71E+00, 2.91E-01), Enterobacteriales (1.38E+00, 8.89E-01), Fusobacteriales (3.23E-01, 8.89E-01), Pseudomonadales (7.19E-01, 2.99E-01), Bacteroidales (-1.35E+00, 2.55E-01), or Clostridiales (-6.41E-02, 9.74E-01). From the stacked plots, the proportion of Enterobacteriales among case samples in age range 6-12 and 12-18 months is much higher than that in the similar age ranges in the control samples. In case samples, Pasteurellales shows higher proportion in the case samples as compared to the controls in the 18-24 age range. For all age ranges, Bacteroidales displays greater proportion in the control compared to case samples. This workspace is available at http://metaviz.cbcb.umd.edu/?ws=EUARocVProf.

Appendix A Table 1

	logFC	Se	pvalues	adjPvalues
Actinomycetales	0.973422	0.289342	0.000767	0.002398
Enterobacteriales	1.381503	0.321685	1.75E-05	0.000146
Lactobacillales	1.151269	1.976018	0.560149	0.700186
Pasteurellales	2.470368	0.335049	1.67E-13	4.16E-12
Pseudomonadales	0.536238	0.271541	0.048291	0.104703
Coriobacteriales	-1.38084	0.379751	0.000277	0.000988
Bacteroidales	-1.18771	0.314453	0.000159	0.000756
Clostridiales	-1.08786	0.280175	0.000103	0.000645
Micrococcaceae	0.905534	0.372999	0.015194	0.046596
Enterobacteriaceae	1.373348	0.325538	2.46E-05	0.000226
Carnobacteriaceae	1.517591	0.319974	2.11E-06	3.23E-05
Streptococcaceae	1.414031	0.307811	4.35E-06	5.00E-05
Pasteurellaceae	2.455909	0.336898	3.10E-13	1.43E-11
Moraxellaceae	0.535115	0.265456	0.043818	0.100781
Coriobacteriaceae	-1.37333	0.383064	0.000337	0.001953
Bacteroidaceae	-1.08748	0.363474	0.002772	0.011594
Porphyromonadaceae	-0.6266	0.355293	0.077797	0.155595
Clostridiaceae	-0.60554	0.291877	0.038021	0.092051
Eubacteriaceae	-0.81641	0.328938	0.013066	0.042931
Lachnospiraceae	-0.57019	0.351512	0.104782	0.200833
Ruminococcaceae	-1.08603	0.317347	0.000621	0.003175
Rothia	0.904951	0.372078	0.015009	0.057426
Escherichia	1.334016	0.32666	4.43E-05	0.00065
Shigella	0.442032	0.346428	0.201967	0.37815
Granulicatella	1.514026	0.319386	2.13E-06	8.29E-05
Streptococcus	1.326435	0.304474	1.32E-05	0.000291
Haemophilus	2.422441	0.337368	6.95E-13	6.12E-11
Acinetobacter	0.534236	0.264971	0.043778	0.118356
Collinsella	-1.47617	0.413236	0.000354	0.003894
Bacteroides	-1.08328	0.363182	0.002857	0.022669
Clostridium	-0.60009	0.289702	0.03832	0.116281
Eubacterium	-0.81469	0.328538	0.013147	0.055094
Dorea	-0.19768	0.398491	0.619836	0.70581
Faecalibacterium	-0.76861	0.333235	0.021082	0.076744
Ruminococcus	-1.18038	0.329733	0.000344	0.003894
Escherichia coli	1.334934	0.326753	4.40E-05	0.001713
Escherichia sp. oral clone 3RH-30	0.524181	0.284599	0.065501	0.278378
Granulicatella adiacens	1.511455	0.376421	5.94E-05	0.001917
Streptococcus equinus	0.803239	0.397057	0.043075	0.228084

Streptococcus mitis	1.15711	0.335784	0.000569	0.014951
Streptococcus parasanguinis	1.068448	0.262749	4.77E-05	0.001713
Streptococcus salivarius	1.016659	0.311742	0.001109	0.021077
Haemophilus parainfluenzae	2.261274	0.369539	9.41E-10	3.04E-07
Acinetobacter sp. SF6	0.469302	0.299197	0.116755	0.369725
Collinsella sp. CB20	-1.26032	0.415063	0.002394	0.036819
Bacteroides fragilis	-1.01919	0.419931	0.015223	0.119304
Faecalibacterium prausnitzii	-0.73858	0.330304	0.025347	0.174196
Faecalibacterium sp. DJF_VR20	-0.25484	0.342751	0.457174	0.695307
Ruminococcus gnavus	-1.18437	0.384095	0.002046	0.034775

Results from *metagenomeSeq* analysis of samples from Bangladesh.

Appendix A Table 2

	logFC	Se	pvalues	adjPvalues
Actinomycetales	1.127055	0.388687	0.003736	0.014943
Lactobacillales	0.921211	0.483368	0.056674	0.141684
Campylobacterales	1.282964	0.609792	0.035384	0.101097
Enterobacteriales	1.847701	0.53244	0.00052	0.0052
Pasteurellales	2.017464	0.35206	1.00E-08	2.00E-07
Pseudomonadales	0.443129	0.387984	0.253399	0.460725
Bacteroidales	-1.36212	0.41888	0.001147	0.005733
Clostridiales	-0.45426	0.384712	0.237695	0.460725
Actinomycetaceae	1.141895	0.383052	0.002873	0.019596
Micrococcaceae	0.907872	0.422592	0.031687	0.106582
Carnobacteriaceae	2.126	0.371553	1.05E-08	2.04E-07
Streptococcaceae	1.022295	0.468884	0.029237	0.106582
Campylobacteraceae	1.162876	0.64227	0.070207	0.17601
Enterobacteriaceae	1.831947	0.533449	0.000594	0.007331
Pasteurellaceae	2.005482	0.350952	1.10E-08	2.04E-07
Moraxellaceae	0.444719	0.384217	0.247081	0.481158
Porphyromonadaceae	0.130004	0.514442	0.800494	0.897523
Prevotellaceae	-1.11173	0.490627	0.023456	0.104141
Eubacteriaceae	-0.46394	0.42391	0.273762	0.506459
Actinomyces	1.142019	0.381409	0.002752	0.030019
Rothia	0.956314	0.407915	0.019058	0.120862
Granulicatella	2.126617	0.370051	9.09E-09	6.64E-07
Streptococcus	1.027824	0.468024	0.028086	0.12814
Campylobacter	1.121751	0.642578	0.080863	0.25804
Citrobacter	0.939767	0.466863	0.044121	0.18946
Dickeya	0.662075	0.48808	0.174944	0.375615
Escherichia	1.875443	0.540051	0.000515	0.009403
Klebsiella	1.111832	0.600685	0.064178	0.223096
Shigella	0.789729	0.476304	0.097311	0.284147
Haemophilus	1.94596	0.349872	2.67E-08	9.74E-07
Acinetobacter	0.458888	0.398732	0.249786	0.506511
Parabacteroides	0.10201	0.51194	0.842058	0.931368
Prevotella	-1.11195	0.48951	0.023113	0.120862
Eubacterium	-0.46404	0.422827	0.272441	0.520677
Rothia mucilaginosa	1.00853	0.423751	0.017312	0.109572
Granulicatella adiacens	1.880536	0.408159	4.08E-06	0.000167
Granulicatella elegans	1.700984	0.455292	0.000187	0.004334
Granulicatella sp. oral clone ASCG05	2.639383	0.447373	3.64E-09	4.46E-07
Streptococcus mitis	1.641897	0.440664	0.000195	0.004334

Streptococcus oralis	0.455029	0.453757	0.315955	0.629342
Streptococcus parasanguinis	0.710634	0.467002	0.128086	0.364896
Streptococcus sanguinis	1.153643	0.403807	0.004278	0.04031
Streptococcus sp. C101	1.242499	0.389737	0.001432	0.021922
Streptococcus sp. oral clone ASCC01	2.121353	0.336561	2.92E-10	7.15E-08
Streptococcus sp. oral clone ASCE09	1.488325	0.398508	0.000188	0.004334
Citrobacter freundii	0.952065	0.466131	0.041103	0.19092
Erwinia chrysanthemi	0.661991	0.484841	0.172135	0.443926
Escherichia coli	1.881366	0.541014	0.000506	0.009539
Klebsiella pneumoniae	1.024196	0.603204	0.089522	0.321673
Haemophilus haemolyticus	1.732626	0.450748	0.000121	0.003709
Haemophilus parainfluenzae	2.027288	0.376656	7.35E-08	4.50E-06
Haemophilus sp. oral clone BP2-46	2.074593	0.382352	5.77E-08	4.50E-06
Prevotella copri	-1.35683	0.457946	0.003048	0.036989
Prevotella histicola	-0.34526	0.45305	0.446016	0.753614
Prevotella sp. BI-42	-0.98653	0.468622	0.035276	0.187884
Prevotella sp. DJF_B112	-1.069	0.458181	0.01964	0.117364
Prevotella sp. DJF_B116	-0.49327	0.603958	0.414085	0.740516
Prevotella sp. DJF_LS16	-0.32501	0.695436	0.640255	0.859195
Prevotella sp. DJF_RP53	-1.54135	0.466753	0.000959	0.016783
Prevotella sp. oral clone BP1-28	-1.24241	0.474949	0.0089	0.077872

Results from *metagenomeSeq* analysis of samples from The Gambia.

Appendix A Table 3

	logFC	Se	pvalues	adjPvalues
Actinomycetales	0.429283	0.394515	0.276539	0.404172
Selenomonadales	0.589476	0.357504	0.099175	0.26919
Campylobacterales	1.387138	0.653813	0.03387	0.26919
Enterobacteriales	0.853636	0.482489	0.076855	0.26919
Pasteurellales	1.293111	0.364409	0.000387	0.007361
Bacteroidales	-0.83172	0.463252	0.072591	0.26919
Veillonellaceae	0.714247	0.382336	0.061746	0.29109
Campylobacteraceae	1.38232	0.660732	0.036429	0.234922
Enterobacteriaceae	0.853549	0.482831	0.077094	0.318014
Pasteurellaceae	1.293107	0.363709	0.000377	0.012457
Prevotellaceae	-1.05105	0.516351	0.041798	0.234922
Megasphaera	-0.85944	0.593322	0.147473	0.494034
Veillonella	1.057554	0.421157	0.012037	0.201614
Campylobacter	1.415512	0.680653	0.037559	0.27977
Citrobacter	0.433077	0.363606	0.23363	0.55881
Enterobacter	1.05059	0.333941	0.001655	0.055443
Escherichia	0.824776	0.478025	0.084459	0.404195
Klebsiella	1.133929	0.597325	0.057651	0.32778
Shigella	0.303726	0.364486	0.404676	0.595773
Haemophilus	1.288696	0.356194	0.000297	0.019895
Prevotella	-1.05483	0.518104	0.041757	0.27977
Veillonella parvula	0.954861	0.416985	0.022026	0.272035
Veillonella sp. HF9	0.378892	0.474818	0.424886	0.702802
Veillonella sp. oral clone VeillC8	1.041036	0.435346	0.01679	0.255467
Veillonella sp. oral clone VeillD5	1.152829	0.389945	0.003113	0.081982
Enterobacter cancerogenus	1.447597	0.415658	0.000496	0.041151
Enterobacter cloacae	0.763304	0.35742	0.032713	0.300472
Escherichia coli	0.82289	0.478524	0.085498	0.40776
Escherichia sp. oral clone 3RH-30	1.07018	0.329466	0.001161	0.057598
Klebsiella pneumoniae	1.096162	0.592399	0.064259	0.375394
Haemophilus haemolyticus	1.633442	0.469118	0.000498	0.041151
Haemophilus parainfluenzae	0.939895	0.37624	0.012485	0.221165
Prevotella copri	-0.61366	0.515333	0.233731	0.538174
Prevotella histicola	-0.67036	0.480563	0.163032	0.481331
Prevotella sp. BI-42	-0.90581	0.481406	0.059892	0.375394
Prevotella sp. DJF_B112	-0.86796	0.490138	0.076585	0.38661
Prevotella sp. DJF_B116	-0.56652	0.757451	0.454501	0.741555
Prevotella sp. DJF_RP53	-0.78627	0.501169	0.116677	0.444046

Results from *metagenomeSeq* analysis of samples from Kenya.

Appendix A Table 4

	logFC	Se	pvalues	adjPvalues
Actinomycetales	3.32E-01	4.64E-01	4.75E-01	8.89E-01
Neisseriales	1.58E+00	5.97E-01	8.15E-03	7.33E-02
Fusobacteriales	3.23E-01	6.47E-01	6.17E-01	8.89E-01
Enterobacteriales	1.38E+00	2.32E+00	5.51E-01	8.89E-01
Pasteurellales	2.97E+00	6.36E-01	3.06E-06	5.51E-05
Pseudomonadales	7.19E-01	4.74E-01	1.29E-01	2.99E-01
Bifidobacteriales	-1.71E+00	9.78E-01	8.09E-02	2.91E-01
Bacteroidales	-1.35E+00	6.65E-01	4.25E-02	2.55E-01
Clostridiales	-6.41E-02	6.36E-01	9.20E-01	9.74E-01
Actinomycetaceae	-0.20839	0.612033	0.733488	0.817872
Micrococcaceae	0.348041	0.456234	0.445549	0.726948
Neisseriaceae	1.577237	0.595253	0.008057	0.083251
Fusobacteriaceae	0.322954	0.646044	0.61715	0.817872
Enterobacteriaceae	1.381966	2.31667	0.55082	0.817872
Pasteurellaceae	2.961241	0.634965	3.11E-06	9.63E-05
Moraxellaceae	0.718601	0.473749	0.129307	0.317369
Bifidobacteriaceae	-1.70397	0.976725	0.081059	0.279202
Bacteroidaceae	-0.76922	0.794018	0.332659	0.572913
Prevotellaceae	-1.37776	0.762314	0.07071	0.274001
Actinomyces	-0.20772	0.608259	0.732725	0.823689
Rothia	0.380333	0.453588	0.401751	0.631323
Neisseria	1.686129	0.621543	0.006671	0.097834
Citrobacter	1.140113	0.482824	0.018209	0.100148
Dickeya	0.808741	0.402896	0.044716	0.20192
Enterobacter	0.295004	0.492357	0.549061	0.779403
Escherichia	1.620312	0.670697	0.015698	0.097834
Klebsiella	0.834308	0.705359	0.236883	0.479266
Shigella	0.990638	0.386787	0.010431	0.097834
Haemophilus	2.939186	0.635039	3.69E-06	0.000203
Acinetobacter	0.717963	0.471069	0.12748	0.304843
Bifidobacterium	-1.7011	0.976616	0.081538	0.256408
Bacteroides	-0.76787	0.794308	0.333684	0.55614
Prevotella	-1.37496	0.762848	0.071481	0.256408
Rothia mucilaginosa	0.380586	0.451422	0.399182	0.64165
Citrobacter freundii	0.782749	0.481665	0.104143	0.302511
Erwinia chrysanthemi	0.808736	0.398462	0.042393	0.209672
Enterobacter cancerogenus	-0.11498	0.518506	0.824509	0.914903
Enterobacter cloacae	0.58855	0.499883	0.239045	0.486204
Escherichia albertii	1.084531	0.651329	0.095892	0.286921

Escherichia coli	1.591776	0.662108	0.016212	0.134856
Escherichia sp. oral clone 3RH-30	0.765796	0.392548	0.051077	0.216247
Klebsiella pneumoniae	0.837909	0.715816	0.241774	0.486204
Shigella boydii	0.748402	0.48444	0.122375	0.329333
Shigella sonnei	0.92589	0.480757	0.054117	0.216247
Haemophilus parainfluenzae	2.871926	0.583115	8.43E-07	0.000154
Haemophilus sp. oral clone BP2-46	2.386879	0.583451	4.30E-05	0.003653
Acinetobacter sp. SF6	0.80256	0.470499	0.088052	0.282693
Bifidobacterium longum	-1.15348	1.280063	0.36753	0.628579
Bacteroides fragilis	-2.58961	1.238748	0.036572	0.196844
Prevotella copri	-1.64117	1.110493	0.139441	0.354414
Prevotella histicola	0.075014	0.634699	0.905918	0.950938
Prevotella sp. BI-42	-1.60745	0.882271	0.068463	0.255689
Prevotella sp. DJF_B112	-2.1359	0.862172	0.013236	0.115343
Prevotella sp. DJF_RP53	-2.91211	0.93435	0.001829	0.055776

Results from *metagenomeSeq* analysis of samples from Kenya.

Appendix B.

Appendix B Table 1: Features Identified During Visual Analysis of IBD Stool 16S Pilot Dataset

Class	Order	Family	Genus	Species
c__Betaproteobacteria	o__Burkholderiales	f__Ruminococcaceae	g__Lachnospira	s__:589277
			g__[Ruminococcus]	s__:333166
			g__Faecalibacterium	s__:564806
				s__:369227
				s__:358104
				s__:369486
				s__gnavus:360015
				s__prausnitzii:851865

Using Metaviz to aggregate counts to each level these features appeared to have a difference in mean abundance when comparing UC to CD samples. Specifically, s__:369227 was found to be statistically significant when testing for differential abundance using *metagenomeSeq*.

Appendix B Table 2: Visual Analysis of UC, CD, nonIBD

Phylum	Class	Order	Family	Genus
Proteobacteria	Erysipelotrichia	Bifidobacteriales	Rikenellaceae	Bifidobacterium
Fusobacteria	Fusobacteriia	Bacillales	Bifidobacteriaceae	Eggerthella
Bacteroidetes	Betaproteobacteria	Erysipelotrichales	Clostridiaceae_1	Alistipes
	Deltaproteobacteria	Fusobacteriales	Peptostreptococcaceae	Christensenellaceae_R_7_group
	Gammaproteobacteria	Burkholderiales	Acidaminococcaceae	Family_XIII_AD3011_group
	Actinobacteria	Desulfovibrionales	Fusobacteriaceae	Coprococcus_1
	Erysipelotrichia	Enterobacteriales	Alcaligenaceae	Fusicatenibacter
			Desulfovibrionaceae	Lachnoclostridium
			Enterobacteriaceae	
			Christensenellaceae	

Features that showed a difference in abundance between the three subjects phenotypes – Ulcerative Colitis, Crohn's Disease, and those without IBD.

Appendix B Table 3: F-statistic calculation

	UC.nonIBD	**CD.nonIBD**	**UC.CD**	**AveExpr**	**F**	**P.Value**	**adj.P.Val**
Phylum							
__Firmicutes	-3.37E-02	1.01E+01	-1.02E+01	1.27E+01	4.47E+02	1.83E-64	2.20E-63
__Bacteroidetes	4.35E-02	9.69E+00	-9.65E+00	1.20E+01	1.69E+02	2.39E-39	1.43E-38
__Proteobacteria	2.00E-01	9.75E+00	-9.55E+00	9.59E+00	1.53E+02	3.73E-37	1.49E-36
__Fusobacteria	-1.64E-01	6.68E+00	-6.84E+00	3.57E+00	2.88E+01	4.44E-11	1.33E-10
__Actinobacteria	2.79E-01	3.25E+00	-2.97E+00	4.97E+00	1.35E+01	3.99E-06	8.49E-06
__Tenericutes	-1.17E+00	-2.07E+00	9.04E-01	2.02E-01	1.46E+01	4.25E-06	8.49E-06
__Verrucomicrobia	-2.60E+00	-2.32E+00	-2.82E-01	2.41E+00	6.56E+00	2.03E-03	3.47E-03
Class							
__Clostridia	1.24E-03	9.54E+00	-9.54E+00	1.25E+01	3.21E+02	2.68E-55	6.42E-54
__Bacteroidia	4.98E-02	9.58E+00	-9.53E+00	1.20E+01	1.58E+02	9.10E-38	1.09E-36
__Gammaproteobacteria	2.47E-01	1.00E+01	-9.80E+00	7.66E+00	7.96E+01	2.25E-24	1.80E-23
__Negativicutes	-4.03E-01	5.39E+00	-5.79E+00	7.60E+00	5.32E+01	3.15E-18	1.89E-17
__Erysipelotrichia	6.95E-01	5.45E+00	-4.76E+00	6.07E+00	3.68E+01	1.00E-13	4.82E-13
__Bacilli	-1.13E-01	5.42E+00	-5.54E+00	5.64E+00	3.62E+01	1.56E-13	6.24E-13
__Fusobacteriia	-1.64E-01	6.68E+00	-6.84E+00	3.57E+00	2.87E+01	4.67E-11	1.60E-10
__Betaproteobacteria	-4.10E-01	4.26E+00	-4.67E+00	6.53E+00	1.41E+01	2.55E-06	7.66E-06
__Mollicutes	-1.17E+00	-2.07E+00	9.04E-01	2.02E-01	1.46E+01	4.15E-06	1.11E-05
__Deltaproteobacteria	-1.29E+00	2.08E+00	-3.36E+00	4.78E+00	8.57E+00	3.16E-04	7.60E-04
Order							
__Clostridiales	1.25E-03	9.54E+00	-9.54E+00	1.25E+01	3.22E+02	2.55E-55	9.93E-54
__Bacteroidales	4.98E-02	9.58E+00	-9.53E+00	1.20E+01	1.58E+02	9.17E-38	1.79E-36
__Selenomonadales	-4.03E-01	5.39E+00	-5.79E+00	7.60E+00	5.32E+01	3.13E-18	4.07E-17
__Enterobacteriales	5.27E-01	9.70E+00	-9.18E+00	6.47E+00	5.12E+01	1.14E-17	1.11E-16
__Erysipelotrichales	6.95E-01	5.45E+00	-4.76E+00	6.07E+00	3.68E+01	1.00E-13	7.84E-13
__Fusobacteriales	-1.64E-01	6.68E+00	-6.84E+00	3.57E+00	2.87E+01	4.72E-11	3.07E-10
__Lactobacillales	-2.20E-01	4.54E+00	-4.76E+00	5.27E+00	2.37E+01	1.22E-09	6.78E-09
__Bacillales	1.46E-01	3.68E+00	-3.53E+00	2.66E+00	2.34E+01	1.71E-09	8.32E-09
__Pasteurellales	4.76E-01	5.48E+00	-5.01E+00	4.46E+00	1.70E+01	2.41E-07	1.05E-06
__Mollicutes_RF9	-1.17E+00	-2.07E+00	9.04E-01	2.02E-01	1.47E+01	3.90E-06	1.52E-05
Family							
__Lachnospiraceae	1.21E-01	8.70E+00	-8.58E+00	1.14E+01	2.19E+02	1.80E-45	1.19E-43
__Bacteroidaceae	-6.23E-02	9.74E+00	-9.80E+00	1.16E+01	1.18E+02	1.26E-31	4.14E-30
__Ruminococcaceae	9.21E-02	7.30E+00	-7.21E+00	1.10E+01	9.74E+01	6.03E-28	1.33E-26
__Enterobacteriaceae	5.27E-01	9.70E+00	-9.18E+00	6.47E+00	5.12E+01	1.14E-17	1.88E-16
__Erysipelotrichaceae	6.95E-01	5.45E+00	-4.76E+00	6.07E+00	3.68E+01	1.00E-13	1.33E-12
__Veillonellaceae	-5.21E-01	6.39E+00	-6.91E+00	6.49E+00	3.31E+01	1.27E-12	1.40E-11

__Fusobacteriaceae	-9.55E-02	6.68E+00	-6.78E+00	3.52E+00	2.83E+01	6.53E-11	6.16E-10
__Streptococcaceae	-3.92E-01	4.41E+00	-4.80E+00	5.06E+00	2.39E+01	1.01E-09	8.34E-09
__Pasteurellaceae	4.76E-01	5.48E+00	-5.01E+00	4.46E+00	1.70E+01	2.41E-07	1.77E-06
__Family_XI	2.27E-01	3.67E+00	-3.44E+00	2.60E+00	1.56E+01	7.95E-07	5.25E-06
Genus							
__Bacteroides	-6.23E-02	9.74E+00	-9.80E+00	1.16E+01	1.18E+02	1.15E-31	2.11E-29
___Eubacterium_rectale_group	3.68E-01	6.83E+00	-6.46E+00	9.00E+00	6.28E+01	1.46E-20	1.35E-18
__Escherichia_Shigella	-1.35E-01	9.06E+00	-9.20E+00	6.05E+00	4.73E+01	1.20E-16	7.34E-15
___Ruminococcus_gnavus_group	2.24E-01	6.55E+00	-6.33E+00	7.45E+00	4.41E+01	9.04E-16	4.16E-14
__Lachnoclostridium	5.09E-01	6.19E+00	-5.68E+00	6.41E+00	3.12E+01	5.01E-12	1.84E-10
__Veillonella	-1.13E-01	5.62E+00	-5.74E+00	4.36E+00	2.79E+01	5.65E-11	1.65E-09
__Fusobacterium	-9.55E-02	6.68E+00	-6.78E+00	3.52E+00	2.83E+01	6.29E-11	1.65E-09
__Streptococcus	-4.31E-01	4.34E+00	-4.77E+00	5.04E+00	2.34E+01	1.55E-09	3.57E-08
__Flavonifractor	-1.79E-01	4.43E+00	-4.61E+00	4.17E+00	2.00E+01	2.30E-08	4.48E-07
__Faecalibacterium	-8.17E-02	5.61E+00	-5.69E+00	9.80E+00	1.97E+01	2.44E-08	4.48E-07
Species							
Unc054vi	-4.99E-01	8.14E+00	-8.64E+00	9.24E+00	5.65E+01	4.81E-19	1.85E-16
UncG3786	-1.35E-01	9.06E+00	-9.20E+00	6.05E+00	4.73E+01	1.20E-16	2.31E-14
UncO8895	2.24E-01	6.55E+00	-6.33E+00	7.45E+00	4.41E+01	8.98E-16	1.15E-13
Unc91005	4.57E-01	7.05E+00	-6.60E+00	7.88E+00	3.90E+01	2.62E-14	2.52E-12
UncO6361	4.74E-01	7.02E+00	-6.54E+00	3.85E+00	3.61E+01	2.85E-13	2.20E-11
Unc00a9i	-4.15E-01	5.53E+00	-5.95E+00	4.26E+00	2.99E+01	1.37E-11	8.79E-10
Unc05bd1	-6.71E-02	6.80E+00	-6.87E+00	9.88E+00	2.80E+01	4.44E-11	2.32E-09
Unc01ie9	-6.18E-01	6.42E+00	-7.04E+00	3.30E+00	2.88E+01	4.82E-11	2.32E-09
Unc64172	4.57E-02	5.77E+00	-5.72E+00	9.74E+00	2.01E+01	1.78E-08	7.60E-07
Unc054m4	-1.79E-01	4.43E+00	-4.61E+00	4.17E+00	2.00E+01	2.29E-08	8.82E-07

Calculated using fitZig function in *metagenomeSeq*. Results of F statistic comparing between Ulcerative Colitis, Crohn's Disease, and those without IBD groups. Aggregated counts to each level of the taxonomic hierarchy and used topTableF function to output 10 results from each taxonomic level.

Appendix B Table 4: Differential Abundance Testing Results for Pair-Wise comparison of CD, UC, nonIBD

	logFC	**se**	**pvalues**	**adjPvalues**
CD and UC				
__Veillonellaceae	-1.0915	0.3387	0.0013	0.0838
CD and nonIBD				
	logFC	**se**	**pvalues**	**adjPvalues**
Gt8Me241	3.4353	0.8723	0.0001	0.0147
Unc21180	2.5074	0.8225	0.0023	0.0735

GX7Fr128	1.6064	0.5351	0.0027	0.0735
UncO6361	1.5206	0.3942	0.0001	0.0147
Unc01ie9	1.4418	0.4554	0.0015	0.0659
Unc03y4v	-1.3629	0.3429	0.0001	0.0147
Unc02ruj	-1.2452	0.3615	0.0006	0.0368
Unc85953	1.0372	0.3404	0.0023	0.0735
Unc36622	-1.0354	0.3198	0.0012	0.0579
__Coprobacter	3.3815	0.8827	0.0001	0.0117
__Ruminococcus_1	-1.7493	0.5188	0.0007	0.0277
__Citrobacter	1.5917	0.5354	0.0030	0.0776
__Fusobacterium	1.5102	0.4480	0.0007	0.0277
__Lachnospiraceae_ND3007_group	-1.3585	0.3455	0.0001	0.0117
__Fusobacteriaceae	1.4750	0.4479	0.0010	0.0644
__Fusobacteriales	1.4173	0.4458	0.0015	0.0577
__Fusobacteriia	1.4237	0.4471	0.0014	0.0319
__Fusobacteria	1.4240	0.4491	0.0015	0.0167
UC_nonIBD				
	logFC	**se**	**pvalues**	**adjPvalues**
Unc04zvf	4.8441	1.4941	0.0012	0.0761
UncO1674	-4.4696	1.3109	0.0007	0.0760
GX7Fr128	3.0544	0.7973	0.0001	0.0491
Unc92642	2.5790	0.7736	0.0009	0.0760
Unc05mrd	-2.3813	0.7505	0.0015	0.0830
Unc02mpn	2.0865	0.6710	0.0019	0.0901
Unc91427	-1.8536	0.5457	0.0007	0.0760
Unc36622	-1.2034	0.3653	0.0010	0.0760
__Citrobacter	3.0057	0.7792	0.0001	0.0211
__Megasphaera	-2.5959	0.7651	0.0007	0.0525
__Dielma	2.5267	0.7923	0.0014	0.0525
__Akkermansia	-2.3501	0.7362	0.0014	0.0525
__Erysipelatoclostridium	1.6105	0.4943	0.0011	0.0525
__Verrucomicrobiales	-2.3441	0.6989	0.0008	0.0311
__Verrucomicrobiae	-2.3291	0.7176	0.0012	0.0281
__Verrucomicrobia	-2.3751	0.7534	0.0016	0.0194

Pair-wise comparison results for UC-CD, UC-nonIBD, CD-nonIBD using fitFeatureModel function of *metagenomeSeq*.

Appendix B Table 5: Visual Analysis UC vs CD

Phylum	Class	Order	Family
__Actinobacteria	__Actinobacteria	__Actinomycetales	__Actinomycetaceae
__Fusobacteria	__Fusobacteriia	__Bifidobacteriales	__Bifidobacteriaceae
__Proteobacteria	__Betaproteobacteria	__Clostridiales	__Corynebacteriaceae
	__Deltaproteobacteria	__Fusobacteriales	__Coriobacteriaceae
	__Gammaproteobacteria	__Desulfovibrionales	__Prevotellaceae
			__Carnobacteriaceae
			__Lachnospiraceae
			__Ruminococcaceae
			__Veillonellaceae
			__Alcaligenaceae
			__Desulfovibrionaceae
			__Enterobacteriaceae
			__Moraxellaceae
			__Actinomycetaceae

Results from IBD HMP2 using Metaviz to inspect each level of taxonomy.

Appendix B Table 6: Visual Analysis UC vs nonIBD

Phylum	Class	Order	Family	Genus	Species
__Actinobacteria	__Actinobacteria	__Bacteroidales	__Rikenellaceae	__Odoribacter	Unide146
__Bacteroidetes	__Bacteroidia	__Erysipelotrichales	__Christensenellaceae	__Alistipes	UncG3786
__Verrucomicrobia	__Betaproteobacteria	__Burkholderiales	__Erysipelotrichaceae	__Christensenellaceae_R_7_group	Unc01ie9
	__Erysipelotrichia	__Desulfovibrionales	__Acidaminococcaceae	__Fusicatenibacter	FNWNL294
	__Betaproteobacteria	__Desulfovibrionales	__Alcaligenaceae	__Lachnospiraceae_ND3007_group	Od8Spla3
	__Deltaproteobacteria	__Enterobacteriales	__Desulfovibrionaceae	___Eubacterium_eligens_group	Unc053aw
	__Verrucomicrobiae	__Verrucomicrobiales	__Verrucomicrobiaceae	__Ruminiclostridium_9	Unc94755
				__Ruminococcaceae_NK4A214_group	Od8Spla3
				__Ruminococcaceae_UCG_002	UncO6106
				__Ruminococcus_1	Unc01w0v
				__Erysipelatoclostridium	Unc65343
				___Clostridium_innocuum_group	Unc05768
				__Phascolarctobacterium	Unc02f9r
				__Bilophila	Unc03y4v
				__Escherichia_Shigella	Unc02q6j
				__Akkermansia	Unc057b2
					Unc36622

Results from IBD HMP2 comparing samples from UC to nonIBD subjects using Metaviz to inspect each level of taxonomy.

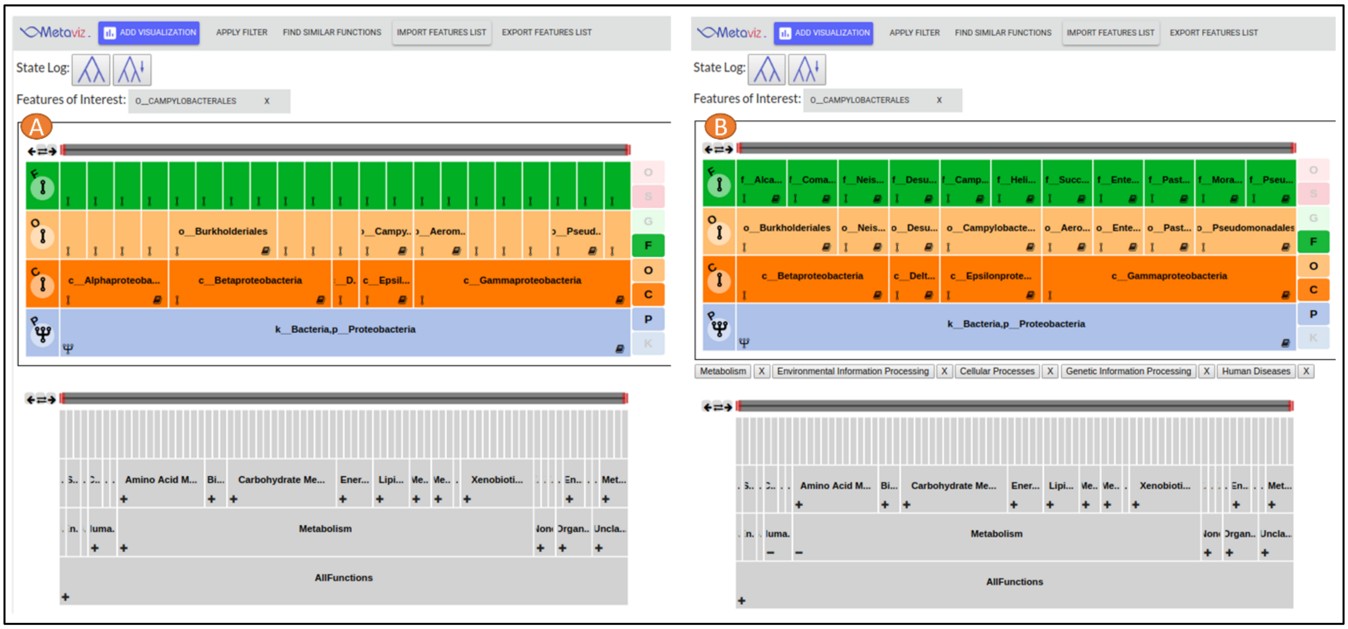
Appendix B Table 7: Visual Analysis CD vs nonIBD.

Phylum	Class	Order	Family	Genus	Species
Fusobacteria	Erysipelotrichia	Bifidobacteriales	Bifidobacteriaceae	Bifidobacterium	UncG3786
Proteobacteria	Fusobacteriia	Erysipelotrichales	Rikenellaceae	Alistipes	Unc01ie9
	Gammaproteobacteria	Fusobacteriales	Christensenellaceae	Christensenellaceae_R_7_group	FNWNL294
		Enterobacteriales	Family_XIII	Blautia	GWMAdo11
			Erysipelotrichaceae	Anaerostipes	Unc94755
			Fusobacteriaceae	Coprococcus_1	UncC1868
			Enterobacteriaceae	Lachnoclostridium	Unc02f9r
				Lachnospira	Unc94574
				Eubacterium_eligens_group	Unc02ruj
				Butyricicoccus	Unc91094
				Ruminococcaceae_UCG_002	UncO6361
				Subdoligranulum	UncO6479
				Fusobacterium	Unc02ee9
				Escherichia_Shigella	Unc03y4v
					Unc00z5u
					Unc057b2
					Unc00y95
					Unc04x9p
					Unc01iri
					Unc36622
					Unc02hhf
					Unc01t8m

					Unc65343
					Unc05o9h
					Unc01qt1
					Unc01w0v

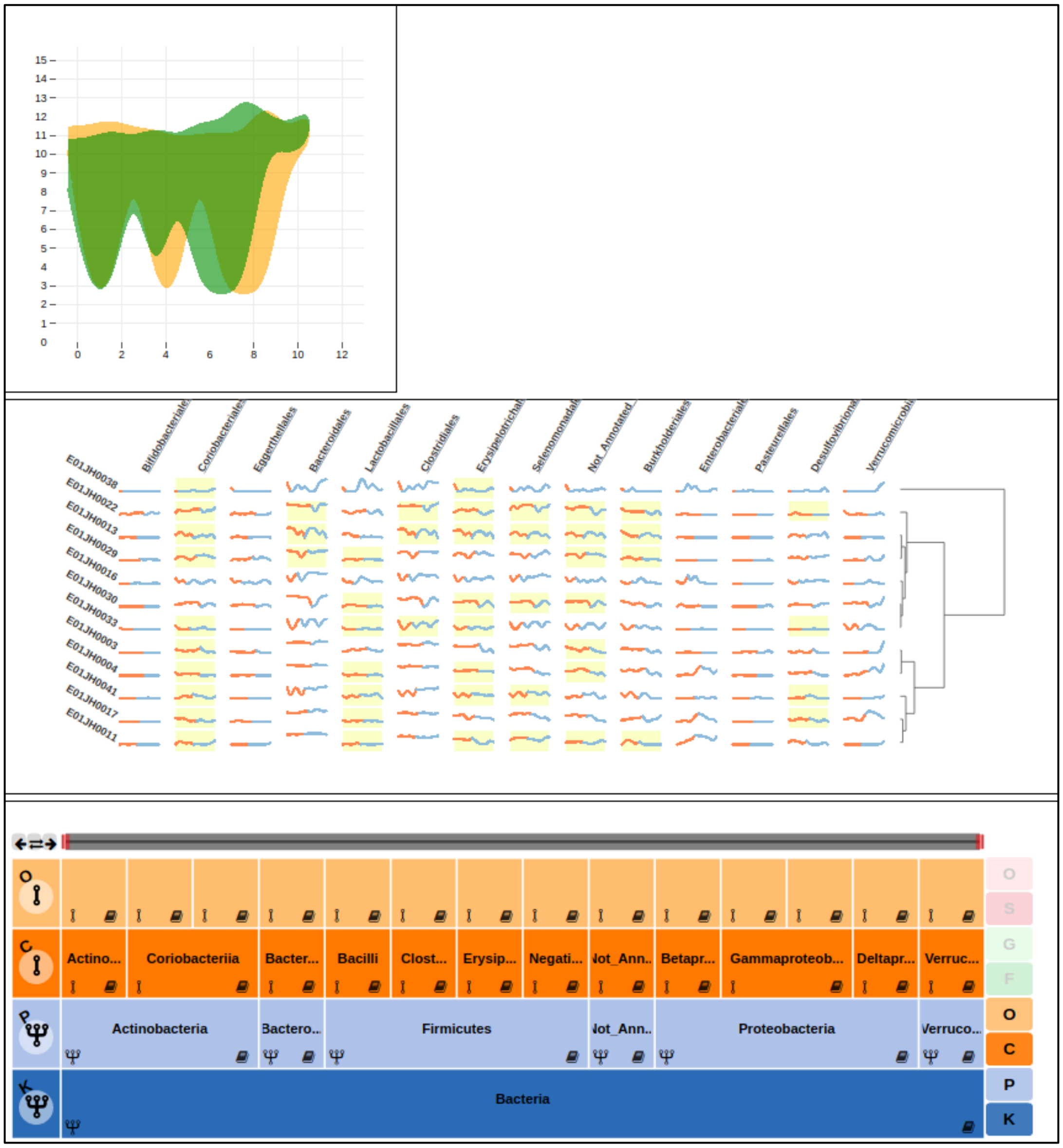
Results from IBD HMP2 using Metaviz to inspect each level of taxonomy.

Appendix C.


A
B
Metaviz
ADD VISUALIZATION
APPLY FILTER
FIND SIMILAR FUNCTIONS
IMPORT FEATURES LIST
EXPORT FEATURES LIST
State Log:
Features of Interest: O__CAMPYLOBACTERALES x
f__Alca... f__Coma... f__Neis... f__Desu... f__Camp... f__Heli... f__Succ... f__Ente... f__Past... f__Mora... f__Pseu...
o__Burkholderiales
o__Neis... o__Desu... o__Campylobacte... o__Aero... o__Ente... o__Past... o__Pseudomonadales
c__Alphaproteoba... c__Betaproteobacteria c__Delt... c__Epsil... c__Epsilonprote... c__Gammaproteobacteria
k__Bacteria,p__Proteobacteria
O S G F O C P K
Metabolism X Environmental Information Processing X Cellular Processes X Genetic Information Processing X Human Diseases X
Amino Acid M... Bi... Carbohydrate Me... Ener... Lipi... Xenobioti... Met...
Metabolism
AllFunctions

Appendix C Figure 1: Similar Function Features

A user can first select a feature of interest by clicking on a FacetZoom node. Then a user can find functions for that feature. A user can add these to the functional filter and finally apply the filter to identify all bacteria that perform functions like those of the features of interest.



Appendix C Figure 2: Sparklines details-on-demand

In this case we show two subject groups, those that developed diarrhea at any point during the experiment and those that did not. The user has an option to show a filled contour for each group as shown or can choose lines showing the minimum, maximum, and average values for each time point across all subjects in each group.

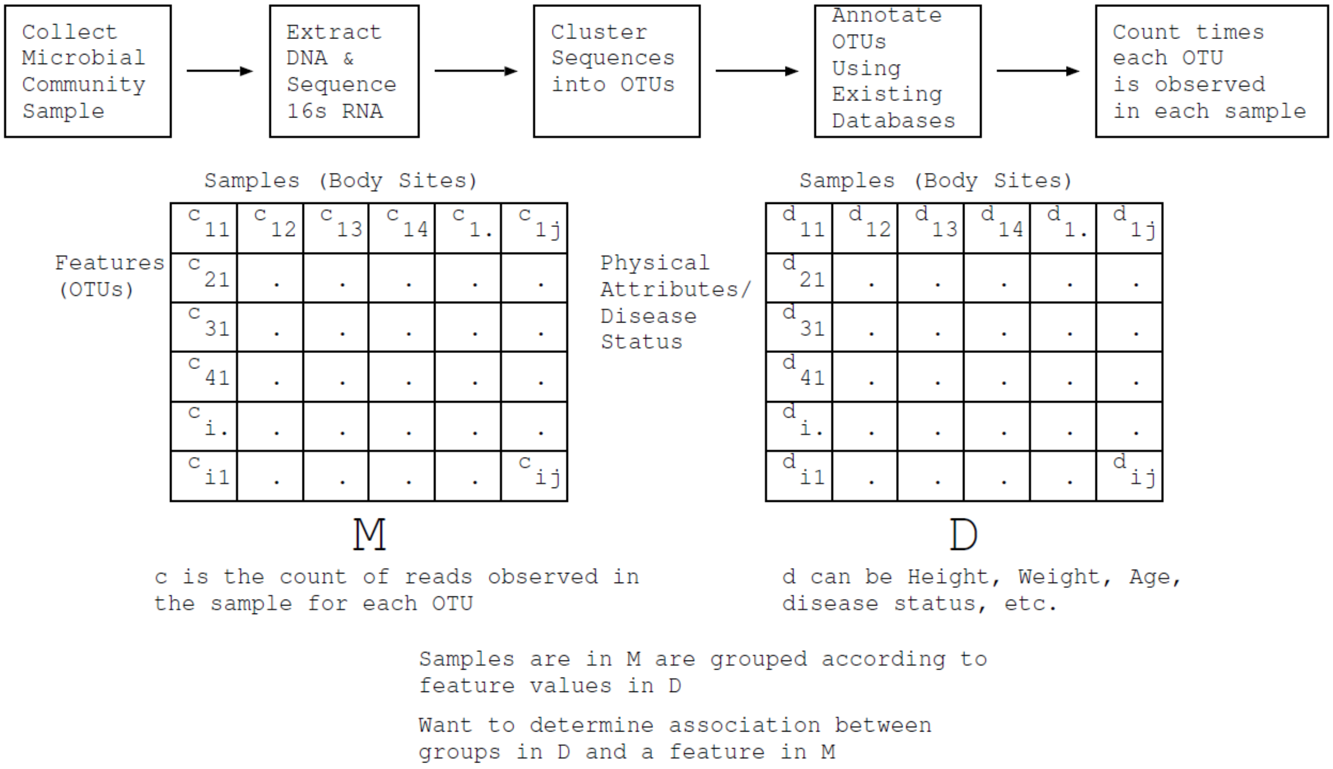
Appendix D.

Section I. Microbiome Preliminaries

In this section, we provide a background on microbiome sequencing and detail the statistics used in standard metagenomic association analyses.

Microbiome Sequencing

Human microbiome sequencing is carried out in the following steps: 1) A microbial community sample is collected from a body site such as the mouth, skin, or gut. 2) DNA is extracted from the sample. 3) The 16s rRNA gene is isolated and sequenced. All bacterial cells which are the same will contain an exact copy of the 16s RNA gene. 4) Sequences that are similar above a threshold (95, 97, or 99 percent similarity) are clustered into an Operational Taxonomic Unit (OTU) 5) OTUs are annotated through comparison to an existing microbial annotation database, 6) the number of times a given OTU is observed for each sample is computed into a count table that serves as the main object of subsequent downstream analysis. Appendix D Figure 1 shows the microbiome sequencing pipeline in more detail. The basic measurement features for metagenomics are OTUs, which are annotated corresponding to specific microbial species or strains.


Collect Microbial Community Sample
Extract DNA & Sequence 16s RNA
Cluster Sequences into OTUs
Annotate OTUs Using Existing Databases
Count times each OTU is observed in each sample
Samples (Body Sites)
Features (OTUs)
M
Samples (Body Sites)
Physical Attributes/ Disease Status
D

c_{11}	c_{12}	c_{13}	c_{14}	$c_{1.}$	c_{1j}
c_{21}	.	.	.	.	.
c_{31}	.	.	.	.	.
c_{41}	.	.	.	.	.
$c_{i.}$	.	.	.	.	.
c_{i1}	.	.	.	.	c_{ij}

d_{11}	d_{12}	d_{13}	d_{14}	$d_{1.}$	d_{1j}
d_{21}	.	.	.	.	.
d_{31}	.	.	.	.	.
d_{41}	.	.	.	.	.
$d_{i.}$	.	.	.	.	.
d_{i1}	.	.	.	.	d_{ij}

c is the count of reads observed in the sample for each OTU

d can be Height, Weight, Age, disease status, etc.

Samples are in M are grouped according to feature values in D

Want to determine association between groups in D and a feature in M

Appendix D Figure 1: Microbiome Sequencing and Metagenomic Analysis Pipeline

To determine the association between microbiota and certain phenotypes, multiple statistics are computed from these OTUs: for instance, the presence or absence of a specific OTU across samples with a given phenotype; the abundance or quantity of an OTU across samples with a given phenotype[96]; the diversity or the number of distinct OTUs in a sample; and the distribution of OTU abundances in a sample. Each of these statistics reveal a distinct view of the role microbial communities play in healthy and disease individuals. In addition, all these association statistics can be computed at any level of the OTU taxonomy. In this sense, the data used in microbiome association studies are much richer than the sets of genotypes used to describe an individual in human DNA analysis.

Metagenomic Statistics

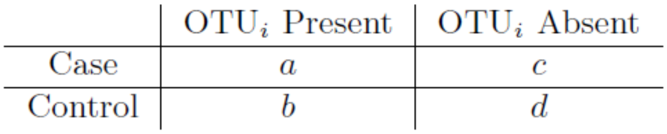
In this section, we define precisely the statistical measures mentioned in Section 3.3. These are standard statistics in the metagenomic field and we detail them here for completeness. Morgan *et al.*

provide a more thorough review of microbiome analysis procedures[97]. In this paper, we denote a metagenomic dataset **M** where $\mathbf{M}_{ij}$ contains the OTU read counts for feature i in sample j. For each sample j, an entry in a separate database **D** contains information regarding its physical characteristics and disease status. Each statistic provides a mechanism to identify associations between groups in **D** and trends in **M**.

Presence or Absence of an OTU

Identifying the role of an OTU first requires a comparison of presence or absence of that OTU in disease and non-disease groups. A χ^2 test is performed to determine the significance of an observed difference in the presence or absence of an OTU between groups. The odds ratio is another measure of association between presence or absence of an OTU and a specific phenotype.

A 2x2 contingency table is populated to compute the χ^2 test on exposure to an OTU. The contingency table counts will be calculated from $\mathbf{M}_{ij}$ by first creating a new matrix, Present, as follows: $\text{Present}_{ij} = \mathbf{M}_{ij} > 0?\ 1:\ 0$. OTU_i Present is the sum of 1s for OTU_i and OTU_i Absent is the sum of 0s.



	OTU_i Present	OTU_i Absent
Case	a	c
Control	b	d

The χ^2 statistic is calculated as:

$$\chi^2 = \frac{(a+b+c+d)(bc-ad)^2}{(a+b)(a+c)(b+d)(c+d)} \tag{1}$$

with one degree of freedom. The odds ratio describing the association between OTU exposure and case or control membership, is defined as:

$$OR = \frac{ad}{cb} \tag{2}$$

Differential Abundance

An OTU may be present in both disease and non-disease groups, but its abundance level may differ between the two groups. Computing differential abundance requires calculating the mean and variance over the counts of a given OTU for each of the two groups to compare[98].

Mean Abundance:

$$\overline{x}_i = \frac{1}{n}\sum_{1}^{j}\mathbf{M}_{ij} \tag{3}$$

Variance:

$$s_i^2 = \frac{1}{n-1}\sum_{1}^{j}(\mathbf{M}_{ij} - \overline{x}_i)^2 \tag{4}$$

A two-sample *t*-statistic is used to test difference between case and control groups.

$$t = \frac{\overline{x}_{Feature_i case} - \overline{x}_{Feature_i control}}{\sqrt{(\frac{s^2_{Feature_i case}}{n_{case}}) + (\frac{s^2_{Feature_i control}}{n_{control}})}} \tag{5}$$

Alpha Diversity

While the presence or abundance of specific OTUs may not be associated with disease, differences in microbial community structure as a whole may be associated with disease. Alpha Diversity is commonly used as a statistic to measure the evenness and richness of microbial communities. It is usually computed based on the entropy of the OTU distribution for a single sample (as Shannon's Index: $H_{\{j\}} = -\sum_i^S p_{ij} * \ln(p_{ij})$ where p_{ij} is the fraction of total OTU counts comprised by OTU i in sample j. Another Alpha Diversity measure is Simpson's index which is of the form $D = \sum_i^S p_{ij}^2$. A two-sample t-statistic is computed to test the significance of differences in statistics *H* or D between groups.

Beta Diversity

The distance of an entire microbiome community structure to that of another sample is the last metagenomic statistic that we will discuss. Beta Diversity is commonly supplied as a check of intra-individual community distance is less than that of inter-individual distance for a specific body site. It is commonly computed as Bray-Curtis dissimilarity $BC_{ij} = 1 - (2C_{ij} / (S_i - S_j))$ where C_{ij} denotes the sum of the counts of species observed at both sites i and j while S_i and S_j are the total number of species observed at sites i and j. Another metric for Beta Diversity is UniFrac which builds a phylogenetic tree across samples under study and then computes a pair-wise distance between two samples to determine if two samples are from the same source. Unweighted UniFrac uses presence/absence of an OTU while weighted UniFrac takes in account the abundance of an OTU and weights branch lengths accordingly[99].

Section II. Problem Overview

In this section we describe the privacy threats of microbiome data and annotate them according to an existing categorization of genome privacy risks. We provide a comprehensive review of microbiome sequencing and metagenomics in the Appendix D, Section 1.

Forensic Identification

One prominent study proved that a person's hand bacteria can identify objects that individual touched[67]. The authors first show the bacteria left after touching a keyboard are separate and unique between individuals. To measure the stability of the bacterial community left behind on the keyboards, the authors compared sequencing results for keyboard samples from the same person stored for 3 to 14 days at - 20 degrees C and room temperature. The community makeup for each sample was not significantly different between any sample storage method. Next the authors calculated the UniFrac distance in community membership between keyboard samples from nine people and a database of microbiome samples from 270 individual's hands. The closest match for each sample was the individual who touched the keyboard. This study was the first to show the identification power of an individual's microbiome signature.

Identification with Metagenomic Codes

A recent analysis showed that metagenomic data alone can uniquely identify individuals in the Human Microbiome Project dataset[23]. The authors build minimal hitting sets to find a collection of microbiome features that are unique to each individual compared to all others in a dataset. The minimal hitting set algorithm was built using four types of features - OTUs, species, genetic markers, and thousand base windows matching reference genomes. The authors use a greedy algorithm and prioritize features by abundance gap, the difference in abundance between a feature in one sample compared to all other samples. The authors called these sets of features "metagenomic codes" and used the codes built at the first time point in the Human Microbiome Project dataset to match individuals at a second time point. The genetic marker and base window codes were the best identifiers between the two time points. The OTU and species level codes also identified individuals but had a higher false-positive rate. As the authors note, the discovery of an identifiable microbiome fingerprint substantially changes the considerations for publicly releasing human microbiome data.

Genetic Re-identification Attacks

Through detailing attacks on genetic datasets, a recent article provided a categorization of techniques to breach participant privacy[100]. The attacks fall into several areas: *Identity Tracing* defined as determining the identity of an anonymized DNA sample using non-private attributes, *Attribute Disclosure* which uses a piece of identified DNA to discover phenotypes or activities in other protected databases, and *Completion Attacks* that use genotype imputation to uncover data that has been removed upon publication of a DNA sequence. To provide a complete overview of microbiome privacy risks, we detail each attack and then expand the categorization to include microbiome specific attacks [2].

Identity Tracing With Metadata reveals the identity of an anonymized DNA sample by using metadata such as age, pedigree information, geography, sex, ethnicity, and health condition. This attack is a concern with metagenomic comparative analysis as case and control group membership is determined by considering metadata.

Genealogical Triangulation uses genetic genealogy databases which link genealogical information, such as surname, with genetic material to allow an individual to recover ancestral information from his/her own DNA. This attack should not be a concern with microbiome data as microbiome inheritance has not been fully determined.

The microbiome presents three different methods for triangulation of a sample's identity which we term *Location Triangulation*, *Behavior Tracing*, and *Rare Disease OTU*. As evidence of the first, a recent study detailed the similarity between individuals that occupy the same dwelling[101]. Therefore, an attacker may be able to reveal the identity of an individual microbiome sample by computing similarity with a sample taken from a specific location.

Further, *Behavior Tracing* could be used to identify a microbiome. The oral microbiota of romantic partners is more similar than other individuals and it is possible to measure how long the similarity between kissing partners is maintained[102]. An attack could be mounted using the phylogenetic or feature-level distance between a known person and the sample from a suspected romantic partner.

Rare Disease Feature Tracing takes advantage of attributes of public health disease tracking and microbial disease infections. Some infections, such as antibiotic-resistant cases, are recorded by state health departments and a single microbiome feature could correspond to those infections. If an attacker is able to observe the known microbiome feature of individual in a public health database and use it to link between another dataset, this will reveal any corresponding sensitive attribute.

Identity Tracing by Phenotypic Prediction involves predicting phenotypic information from genotypic information and then using that to match to an individual. Phenotypic prediction with human DNA is quite difficult given that predictions are not currently robust for unique identifiers in the population. For identifiers such as height, weight, and age, the effectiveness of this attack is likely to be low with microbiome data.

Identity Tracing by Side Channel Leaks is possible when an identifier is apparent from the dataset entries either by data preparation techniques or data-id assignment. One example is that Personal Genome Project sequencing files which by default were named with patient first and last name included. This attack is a concern with microbiome sequencing as well given that file uploading of the Personal Genome Project is similar for microbiome results.

Attribute Disclosure With N=1 entails an attacker associating an individual's identity to a piece of DNA and that piece of DNA to a sensitive attribute, such as an element in a database of drug users. For microbiome data, the forensic identification and the metagenomic codes techniques could be used by an attacker to successfully query a dataset with a sensitive attribute.

Attribute Disclosure from Summary Statistics uses genetic information of one victim and published summary statistics from a case/control study to determine if the victim's DNA is biased towards the distribution of either the case or control group. If group membership can be determined, then the criteria to split groups (such as disease status) is revealed to the attacker. Linkage disequilibrium, or the probability that portions of DNA are more likely to be inherited together than others, provides a mechanism to increase the power of the attack. Further, genealogical information can be used to accomplish attribute disclosure.

While the authors cite Attribute Disclosure from Summary Statistics as an attack possible with all '-omic' data, linkage disequilibrium and genealogical triangulation are not applicable to microbiome sequencing. The release of summary statistics may be used to determine if a metagenomic code for an individual is present in a case/control group, but the probability of this attack needs be determined.

Completion Attacks reveal portions of a DNA sample that are not released publicly by using linkage disequilibrium to uncover the hidden SNPs. Genealogical information, such as a pedigree and the SNPs of relatives, can also be used in genotype imputation. For metagenomic data, a cohabitation mapping of individuals from the same household to distinct features could be used to mount a completion attack.

Section III. Oblivious Transfer

Oblivious Transfer is a subroutine that allows one party known as the sender (P1) to offer two messages and for the other party, referred to as the receiver (P2), to input a bit selecting one of the messages. Oblivious Transfer guarantees that the sender learns nothing about the receiver's selection and the receiver learns nothing about the other inputs beyond the one selected. In the semi-honest setting, one approach to Oblivious Transfer is for the receiver to generate two public-private key-pairs but with one of the public keys to not have a valid private key. The receiver then sends both keys to the sender, who encrypts its inputs with the public keys and sends them to the receiver. The receiver will only be able to properly decrypt one of the ciphertexts [103].

Section IV. Implementation

In this section we provide details on each implementation approach. In the pre-computation approach, we compute over values that are locally computed by each party. In the sparse matrix approach, we operate on the non-zero elements from each party directly.

Pre-computation approach

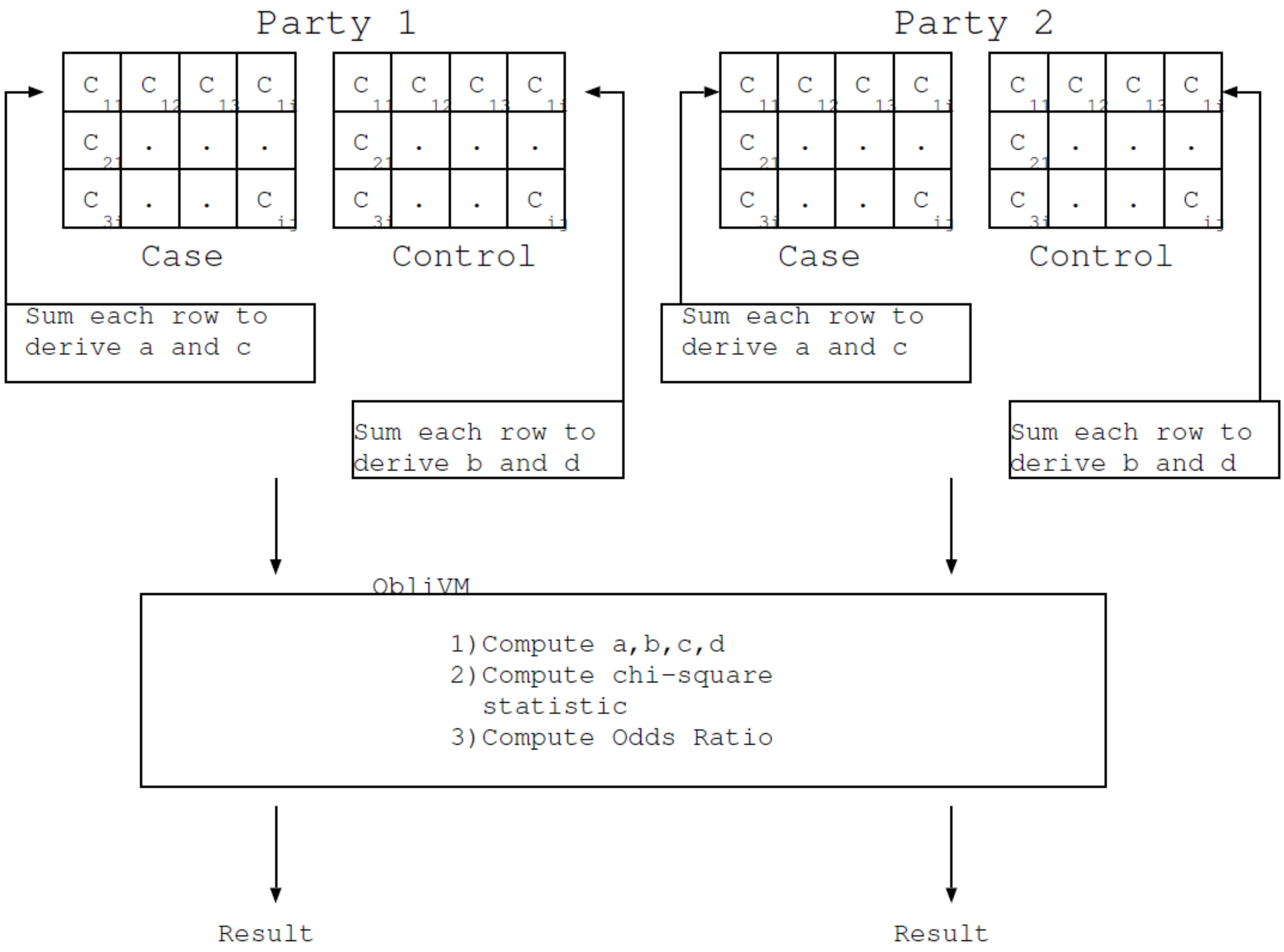
This method is a straightforward approach to reduce the amount of operations and data in the circuit. Appendix D Figure 2 shows the process for calculating a χ^2-test and odds ratio on pre-computed contingency table counts.

Sparse matrix approaches

The main idea of the technique from Nikolaenko *et al.* is to create a counting circuit using Bitonic Sort, a sorting algorithm that can be implemented as an oblivious circuit with $O(n \log^2(n))$ running time, to operate over tuples consisting of (row, column, matrix element)[74]. Appendix D Figure 3 shows the scheme in greater depth. We use the counting mechanism to implement each statistic. Surprisingly, the sorting operation for this approach outweighed the naive approach for chi-square and odds ratio. For completeness, we provide a description of this approach for implementing each statistical test.

Presence/Absence

Sparse Computation. We assume that parties first locally split their datasets on case and control criteria. For the scheme described in Appendix D Figure 3, each party will then only input the non-zero elements of the respective case and control matrices as tuples. For the χ^2 test and odds ratio, the counter can be used to find contingency table counts. The oblivious counter will be used to calculate OTU_i Present for each group. The number of samples for each party's case and control groups is shared obliviously and OTU_i Absent can be calculated. With *a, b, c, d*, χ^2 can be calculated as in Equation (1) and odds ratio as Equation (2).


Party 1
Party 2
Case
Control
Case
Control
Sum each row to derive a and c
Sum each row to derive b and d
Sum each row to derive a and c
Sum each row to derive b and d
ObliVM
1)Compute a,b,c,d
2)Compute chi-square statistic
3)Compute Odds Ratio
Result
Result

Appendix D Figure 2: Diagram of Pre-computation for Presence/Absence
The inputs to the garbled circuit are locally generated row sums from each party. Presence/Absence χ^2-test statistic and Odds Ratio are calculated in the circuit.

Differential Abundance

For calculating differential abundance, the sequencing counts need to be normalized and that can be accomplished per sample in the pre-computation phase. We examine Equations (3) and (4) to determine what optimizations can be accomplished for computing in secure computation. To avoid processing all samples within the computation framework, we observe transformations that reduce the total number of operations. With *k* for party 1 and party 2 the following can be computed:

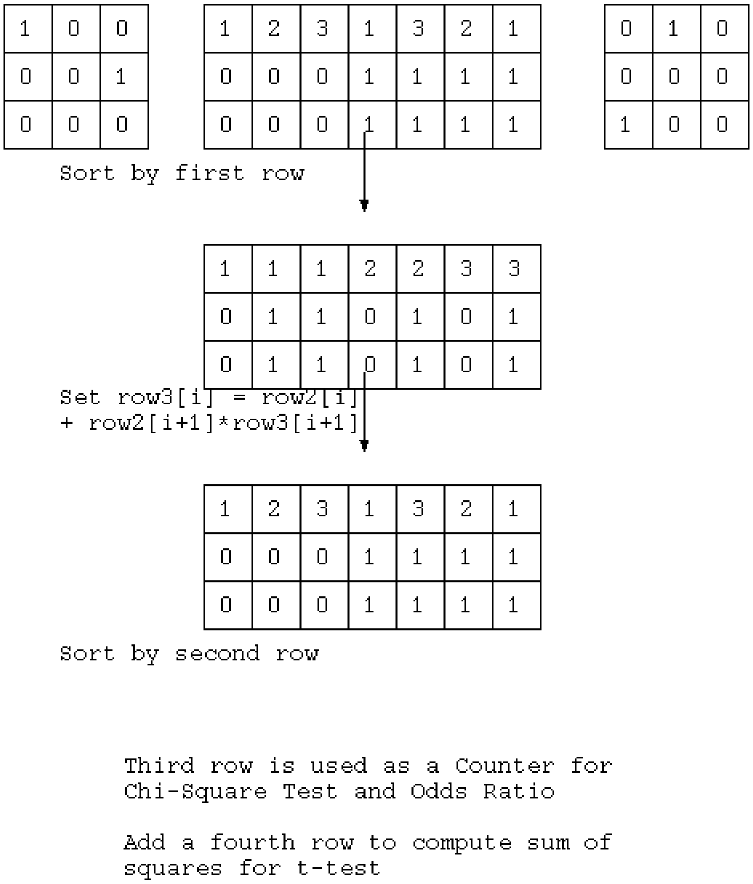
Mean Abundance

$$\overline{x}_{OTU_{i,k}} = \frac{1}{n_{k_1}+n_{k_2}}\left(\sum_{j}^{n_{k_1}} M_{ij} + \sum_{j}^{n_{k_2}} M_{ij}\right) \tag{6}$$

Variance

$$s^2_{OTU_{i,k}} = \frac{\sum_j^{n_{k_1}} M_{ij}^2 + \sum_j^{n_{k_2}} M_{ij}^2 + \frac{\left(\sum_j^{n_{k_1}} M_{ij}+\sum_j^{n_{k_2}} M_{ij}\right)^2}{n_{k_1}+n_{k_2}}}{n_{k_1}+n_{k_2}} \tag{7}$$

Sparse Computation. With Nikolaenko *et al*. sparse matrix approach, the oblivious counter is used to calculate the total sum and augmented to compute the sum of squares for each feature. Then a two-sample t-test can be performed using those values as described in Equations (6) and (7).



1	0	0
0	0	1
0	0	0

1	2	3	1	3	2	1
0	0	0	1	1	1	1
0	0	0	1	1	1	1

0	1	0
0	0	0
1	0	0

Sort by first row

1	1	1	2	2	3	3
0	1	1	0	1	0	1
0	1	1	0	1	0	1

Set row3[i] = row2[i]
+ row2[i+1]*row3[i+1]

1	2	3	1	3	2	1
0	0	0	1	1	1	1
0	0	0	1	1	1	1

Sort by second row

Third row is used as a Counter for
Chi-Square Test and Odds Ratio

Add a fourth row to compute sum of
squares for t-test

Appendix D Figure 3: Sparse Matrix Counter

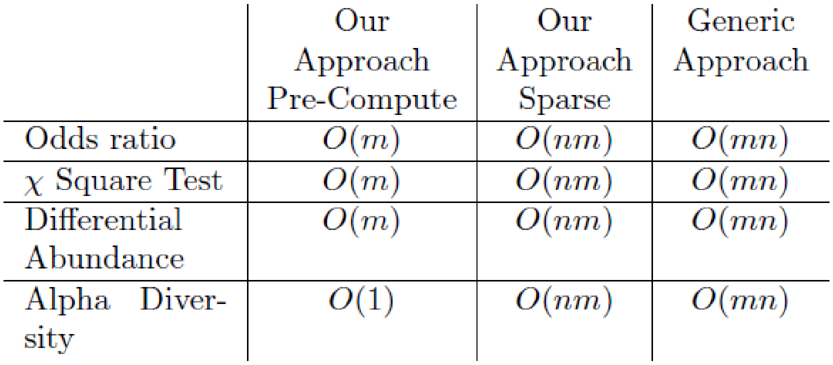
Alpha Diversity

We compute Alpha Diversity for each sample, then use a two-sample t-test to determine the significance of a difference between case and control groups. Given that ObliVM does not currently compute logarithm, we measure Alpha Diversity as Simpson's index: $D = (\sum n(n-1)) \div N(N-1)$ where n is the number of OTU counts for OTU$_i$ and N is the total number of counts observed in a sample.

Sparse Computation The two values for Simpson's index, $\sum n(n-1))$ and $N(N-1)$ are generated over each sample using the sparse matrix counter technique. Then a pass over the array using division yields Simpson's index from which the total sum and sum of squares can be computed for case and control groups.

Asymptotic Complexity

Since secure computation is orders of magnitude slower than cleartext computation, we carefully designed our protocols so that we either operate only on a sparse representation of matrix elements or can put as much computation as possible outside of the secure computation. In fact, for our pre-computation approach, as shown in Appendix D Table 1 for all test cases we evaluated, we achieved asymptotic improvement compared with a generic solution that performs all operations in secure computation directly.



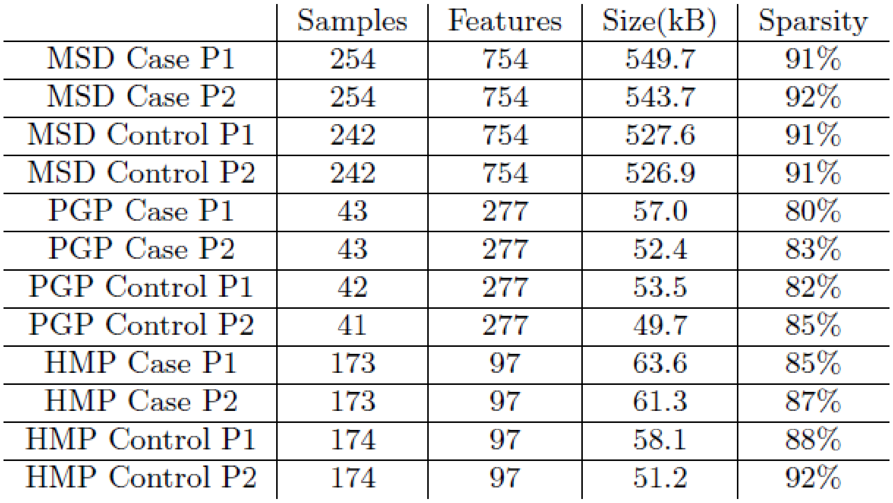
	Our Approach Pre-Compute	Our Approach Sparse	Generic Approach
Odds ratio	$O(m)$	$O(nm)$	$O(mn)$
χ Square Test	$O(m)$	$O(nm)$	$O(mn)$
Differential Abundance	$O(m)$	$O(nm)$	$O(mn)$
Alpha Diversity	$O(1)$	$O(nm)$	$O(mn)$

Appendix D Table 1: Running Time Complexity. Speedup of our approaches using local computation and sparse matrix computation compared with generic solution. For the analysis of the sparse approach, the running time is proportional to a constant k, which is the proportion of samples (n) which have a non-zero element for a given feature. For a given dataset, the total number of non-zero elements will be $(k\ n)m$. In our experiments, k took a value of $\leq .2$ for all datasets used as shown in Appendix D Table 2. While the asymptotic complexity is the same, our sparse approach ran faster than the naive approach for each dataset considered.

Section V. Evaluation

Datasets

Appendix D Table 2 summarizes the number of features, samples, file size, and sparsity. The datasets provide a good array of input sizes and sparsity to evaluate our implementations.



	Samples	Features	Size(kB)	Sparsity
MSD Case P1	254	754	549.7	91%
MSD Case P2	254	754	543.7	92%
MSD Control P1	242	754	527.6	91%
MSD Control P2	242	754	526.9	91%
PGP Case P1	43	277	57.0	80%
PGP Case P2	43	277	52.4	83%
PGP Control P1	42	277	53.5	82%
PGP Control P2	41	277	49.7	85%
HMP Case P1	173	97	63.6	85%
HMP Case P2	173	97	61.3	87%
HMP Control P1	174	97	58.1	88%
HMP Control P2	174	97	51.2	92%

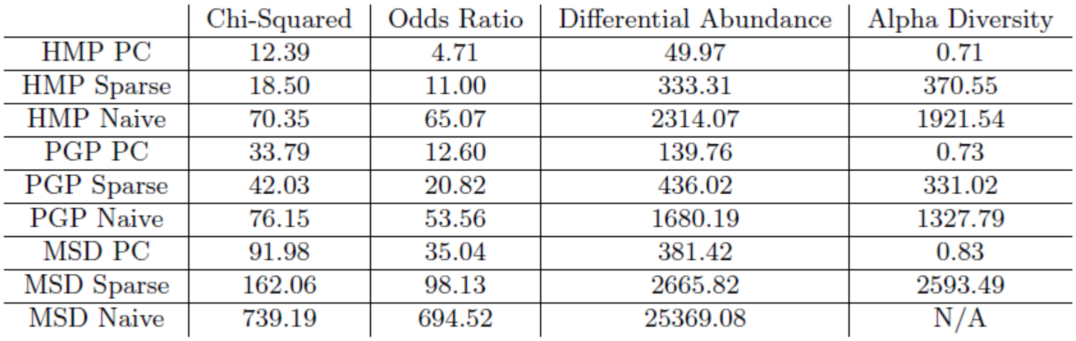
Appendix D Table 2: Dataset Sizes. Dimensions and sparsity of each dataset used for evaluation. P1 is Party 1 and P2 is Party 2 for secure computation. Sparsity is defined as 1-(Percent of Non-Zero entries).

Running Times

Appendix D Table 3 lists the running times for each statistic and dataset.

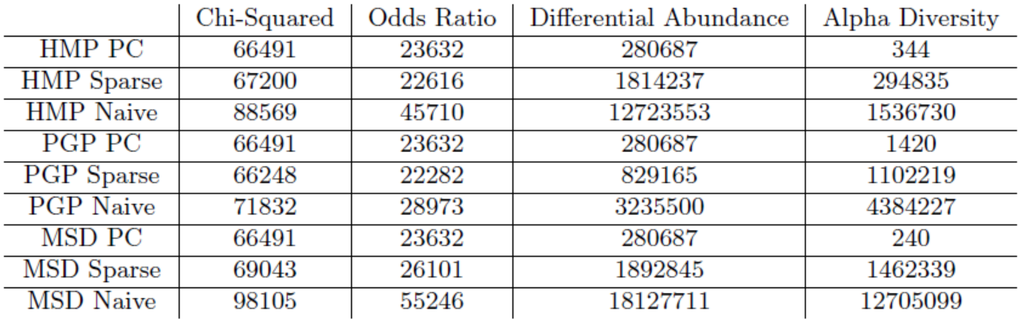
Circuit Sizes

Appendix D Table 4 lists the circuit size per feature for each statistic and dataset.



	Chi-Squared	Odds Ratio	Differential Abundance	Alpha Diversity
HMP PC	12.39	4.71	49.97	0.71
HMP Sparse	18.50	11.00	333.31	370.55
HMP Naive	70.35	65.07	2314.07	1921.54
PGP PC	33.79	12.60	139.76	0.73
PGP Sparse	42.03	20.82	436.02	331.02
PGP Naive	76.15	53.56	1680.19	1327.79
MSD PC	91.98	35.04	381.42	0.83
MSD Sparse	162.06	98.13	2665.82	2593.49
MSD Naive	739.19	694.52	25369.08	N/A

Appendix D Table 3: Running Times. Running time for each statistic and each dataset in seconds (PC stands for Pre-Compute). In each statistic, the number of arithmetic operations determined the running time. The size of the dataset along with sparsity contributed to running time for the sparse implementations. Alpha Diversity MSD Naive did not run to completion on the EC2 instance size due to insufficient memory. Based on the circuit size and the number of gates processed per second for other statistics, we estimate the running time to be 378 minutes.

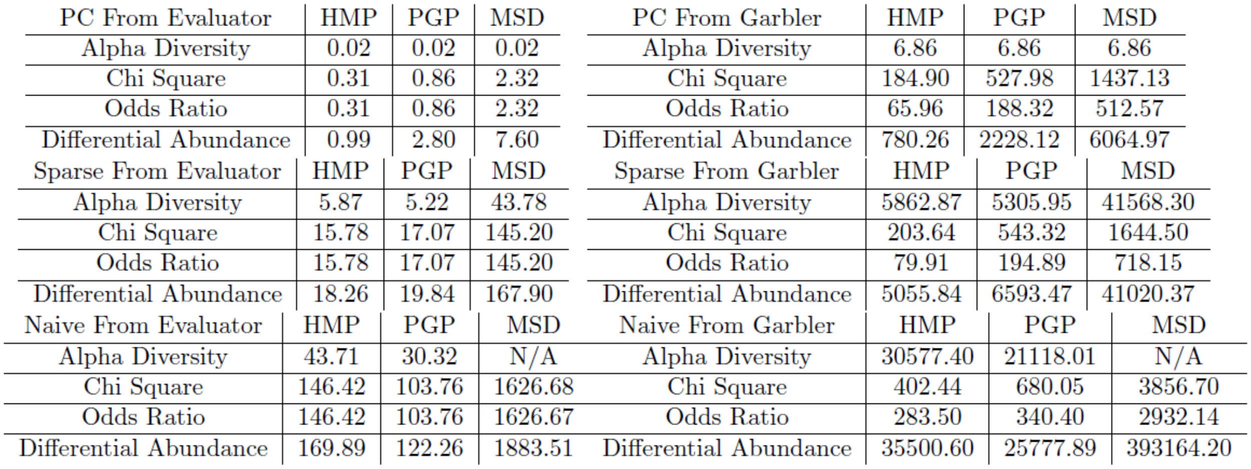


	Chi-Squared	Odds Ratio	Differential Abundance	Alpha Diversity
HMP PC	66491	23632	280687	344
HMP Sparse	67200	22616	1814237	294835
HMP Naive	88569	45710	12723553	1536730
PGP PC	66491	23632	280687	1420
PGP Sparse	66248	22282	829165	1102219
PGP Naive	71832	28973	3235500	4384227
MSD PC	66491	23632	280687	240
MSD Sparse	69043	26101	1892845	1462339
MSD Naive	98105	55246	18127711	12705099

Appendix D Table 4: Circuit Size Per Feature. Circuit size for each implementation and dataset (PC stands for Pre-Compute). The number of samples is considered as feature count for calculating Alpha Diversity circuit size. The differences in Alpha Diversity between datasets is explained by the number of samples for PGP (168) being much lower than that of HMP (694) and MSD (992).

Network Traffic

Appendix D Table 5 lists the traffic from each computation party. The pre-computation approach requires the least amount of traffic with the sparse implementation requiring several more orders of magnitude. The most costly approach is the naive approach. The increase in network traffic between the sparse and pre-computation implementations is significant as compared to the differences in running times of those approaches.



PC From Evaluator	HMP	PGP	MSD	PC From Garbler	HMP	PGP	MSD
Alpha Diversity	0.02	0.02	0.02	Alpha Diversity	6.86	6.86	6.86
Chi Square	0.31	0.86	2.32	Chi Square	184.90	527.98	1437.13
Odds Ratio	0.31	0.86	2.32	Odds Ratio	65.96	188.32	512.57
Differential Abundance	0.99	2.80	7.60	Differential Abundance	780.26	2228.12	6064.97

Sparse From Evaluator	HMP	PGP	MSD	Sparse From Garbler	HMP	PGP	MSD
Alpha Diversity	5.87	5.22	43.78	Alpha Diversity	5862.87	5305.95	41568.30
Chi Square	15.78	17.07	145.20	Chi Square	203.64	543.32	1644.50
Odds Ratio	15.78	17.07	145.20	Odds Ratio	79.91	194.89	718.15
Differential Abundance	18.26	19.84	167.90	Differential Abundance	5055.84	6593.47	41020.37

Naive From Evaluator	HMP	PGP	MSD	Naive From Garbler	HMP	PGP	MSD
Alpha Diversity	43.71	30.32	N/A	Alpha Diversity	30577.40	21118.01	N/A
Chi Square	146.42	103.76	1626.68	Chi Square	402.44	680.05	3856.70
Odds Ratio	146.42	103.76	1626.67	Odds Ratio	283.50	340.40	2932.14
Differential Abundance	169.89	122.26	1883.51	Differential Abundance	35500.60	25777.89	393164.20

Appendix D Table 5: Network Traffic. Left column details traffic in MB sent from evaluator (PC stands for Pre-compute). Right column is MB sent from garbler.

Bibliography

1. Human, T. & Project, M. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–14 (2012).

2. Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science (80-.).* **348,** 1261359 (2015).

3. Pop, M. *et al.* Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol.* **15,** 1 (2014).

4. Turnbaugh, P. J. *et al.* A core gut microbiome in obese and lean twins. *Nature* **457,** 480–484 (2009).

5. Kotloff, K. L. *et al.* Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): A prospective, case-control study. *Lancet* **382,** 209–222 (2013).

6. Gevers, D. *et al.* The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* **15,** 382–392 (2014).

7. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15,** (2014).

8. Segata, N. *et al.* Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Meth* **9,** 811–814 (2012).

9. Hoaglin, D., Mosteller, F. & Wilder Tukey, J. *Understanding robust and exploratory data analysis / edited by David C. Hoaglin, Frederick Mosteller,*

John W. Tukey. *Wiley*

10. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC Bioinformatics* **12,** (2011).

11. Flygare, S. *et al.* Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling. *Genome Biol.* **17,** 111 (2016).

12. Breitwieser, F. P. & Salzberg, S. L. Pavian: Interactive analysis of metagenomics data for microbiomics and pathogen identification. *bioRxiv* (2016). doi:10.1101/084715

13. Huse, S. M. *et al.* VAMPS: a website for visualization and analysis of microbial population structures. *BMC Bioinformatics* **15,** 1 (2014).

14. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3,** e1319 (2015).

15. Oliveira, F. S. *et al.* MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments. *Nucleic Acids Res.* gkx1027-gkx1027 (2017).

16. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat. Methods* **10,** 1200–1202 (2013).

17. McMurdie, P. J. & Holmes, S. Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8,** (2013).

18. Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal

update. *Nucleic Acids Res.* **46,** D794–D801 (2018).

19. Roadmap Epigenomics Consortium *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–329 (2015).

20. Cancer Genome Atlas Research Network *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* **45,** 1113–20 (2013).

21. Groce, A. D. New notions and mechanisms for statistical privacy. (2014).

22. Jagadeesh, K. A., Wu, D. J., Birgmeier, J. A., Boneh, D. & Bejerano, G. Deriving genomic diagnoses without revealing patient genomes. *Science (80-.).* **357,** 692–695 (2017).

23. Franzosa, E. A. *et al.* Identifying personal microbiomes using metagenomic codes. *Proc. Natl. Acad. Sci.* **112,** E2930–E2938 (2015).

24. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514,** 59–64 (2014).

25. Paez-Espino, D. *et al.* Uncovering Earth's virome. *Nature* **536,** 425–430 (2016).

26. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12,** 996–1006 (2002).

27. Chelaru, F., Smith, L., Goldstein, N. & Bravo, H. C. Epiviz: interactive visual analytics for functional genomics data. *Nat. Methods* **11,** 938–940 (2014).

28. Pedersen, T. L., Nookaew, I., Wayne Ussery, D. & Månsson, M. PanViz: interactive visualization of the structure of functionally annotated pangenomes. *Bioinformatics* **33,** 1081–1082 (2017).

29. Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic

visualization in a Web browser. *BMC Bioinformatics* **12,** 1 (2011).

30. Chelaru, F. & Bravo, H. C. Epiviz: a view inside the design of an integrated visual analysis software for genomics. *BMC Bioinformatics* **16,** 1 (2015).

31. Dachselt, R., Frisch, M. & Weiland, M. FacetZoom: A Continuous Multi-scale Widget for Navigating Hierarchical Metadata. in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 1353–1356 (ACM, 2008). doi:10.1145/1357054.1357265

32. Pop, M. *et al.* Individual-specific changes in the human gut microbiota after challenge with enterotoxigenic Escherichia coli and subsequent ciprofloxacin treatment. *BMC Genomics* **17,** 440 (2016).

33. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat. Methods* **14,** 1023–1024 (2017).

34. Mondot, S. *et al.* Structural robustness of the gut mucosal microbiota is associated with Crohn's disease remission after surgery. *Gut* **65,** 954–962 (2016).

35. Sohn, S.-H. *et al.* Analysis of Gastric Body Microbiota by Pyrosequencing: Possible Role of Bacteria Other Than Helicobacter pylori in the Gastric Carcinogenesis. *J. Cancer Prev.* **22,** 115–125 (2017).

36. Fenollar, F. *et al.* Tropheryma whipplei associated with diarrhoea in young children. *Clin. Microbiol. Infect.* **22,** 869–874 (2016).

37. Keller, P. M. *et al.* Recognition of Potentially Novel Human Disease-Associated Pathogens by Implementation of Systematic 16S rRNA Gene Sequencing in the Diagnostic Laboratory . *J. Clin. Microbiol.* **48,** 3397–3402

(2010).

38. Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M. & Tang, P. Metagenomics for pathogen detection in public health. *Genome Med.* **5,** (2013).

39. Blanton, L. V. *et al.* Gut bacteria that prevent growth impairments transmitted by microbiota from malnourished children. *Science (80-.).* **351,** (2016).

40. The integrative human microbiome project: Dynamic analysis of microbiome-host omics profiles during periods of human health and disease corresponding author. *Cell Host and Microbe* **16,** 276–289 (2014).

41. Leiserson, M. D. M. *et al.* MAGI: Visualization and collaborative annotation of genomic aberrations. *Nature Methods* **12,** 483–484 (2015).

42. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551,** 457–463 (2017).

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

44. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7,** 335–336 (2010).

45. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* **12,** 902–903 (2015).

46. Olson, N. D. *et al.* metagenomeFeatures: An R package for working with 16S rRNA reference databases and marker-gene survey feature data. *bioRxiv* (2018).

47. Chelaru, F. *Epiviz: Integrative Visual Analysis Software for Genomics*. (2015). doi:10.13016/M2ZS7X

48. Plaisant, C., Fekete, J. D. & Grinstein, G. Promoting insight-based evaluation of visualizations: From contest to benchmark repository. *IEEE Trans. Vis. Comput. Graph.* **14,** 120–134 (2008).

49. Meehan, C. J. & Beiko, R. G. A phylogenomic view of ecological specialization in the lachnospiraceae, a family of digestive tract-associated bacteria. *Genome Biol. Evol.* **6,** 703–713 (2014).

50. Kameyama, K. & Itoh, K. Intestinal Colonization by a Lachnospiraceae Bacterium Contributes to the Development of Diabetes in Obese Mice. *Microbes Environ.* **29,** 427–430 (2014).

51. Maharshak, N. *et al.* Fecal and Mucosa-Associated Intestinal Microbiota in Patients with Diarrhea-Predominant Irritable Bowel Syndrome. *Dig. Dis. Sci.* (2018). doi:10.1007/s10620-018-5086-4

52. Everard, A. *et al.* Cross-talk between Akkermansia muciniphila and intestinal epithelium controls diet-induced obesity. *Proc. Natl. Acad. Sci.* **110,** 9066–9071 (2013).

53. Isabel Ordiz, M. *et al.* Environmental enteric dysfunction and the fecal microbiota in malawian children. *Am. J. Trop. Med. Hyg.* **96,** 473–476 (2017).

54. Vázquez-Castellanos, J. F. *et al.* Altered metabolism of gut microbiota contributes to chronic immune activation in HIV-infected individuals. *Mucosal Immunol.* **8,** 760–772 (2015).

55. Nishino, K. *et al.* Analysis of endoscopic brush samples identified mucosa-

associated dysbiosis in inflammatory bowel disease. *J. Gastroenterol.* **53,** 95–106 (2018).

56. Shaw, K. A. *et al.* Dysbiosis, inflammation, and response to treatment: A longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med.* **8,** (2016).

57. Perkel, J. M. Data visualization tools drive interactivity and reproducibility in online publishing. *Nature* **554,** 133–134 (2018).

58. Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. MEGAN analysis of metagenomic data. *Genome Res.* **17,** 377–386 (2007).

59. Bik, H. M. & Inc., P. I. Phinch: An interactive, exploratory data visualization framework for -Omic datasets. *bioRxiv* 9944 (2014). doi:10.1101/009944

60. McNally, C. P., Eng, A., Noecker, C., Gagne-Maynard, W. C. & Borenstein, E. BURRITO: An interactive multi-omic tool for visualizing taxa-function relationships in microbiome data. *Front. Microbiol.* **9,** (2018).

61. Langille, M. *et al.* Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* **31,** 814–21 (2013).

62. Tufte, E. R. *Beautiful Evidence*. (Graphics Press, 2006).

63. Kancherla, J., Zhang, A., Gottfried, B. & Bravo, H. C. Epiviz Web Components: reusable and extensible component library to visualize functional genomic datasets. *F1000Research* **7,** (2018).

64. Turnbaugh, P. J. *et al.* The human microbiome project: exploring the microbial part of ourselves in a changing world. *Nature* **449,** 804–810 (2007).

65. Church, G. M. The Personal Genome Project. *Mol. Syst. Biol.* **1,** E1–E3 (2005).

66. Blaser, M., Bork, P., Fraser, C., Knight, R. & Wang, J. The microbiome explored: Recent insights and future challenges. *Nature Reviews Microbiology* **11,** 213–217 (2013).

67. Fierer, N. *et al.* Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. U. S. A.* **107,** 6477–81 (2010).

68. Malkhi, D., Nisan, N., Pinkas, B. & Sella, Y. Fairplay—a secure two-party computation system. in *SSYM'04 Proceedings of USENIX 2004* 20 (2004).

69. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics* **39,** 1181–1186 (2007).

70. Fredrikson, M. *et al.* Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. in *Proceedings of the 23rd USENIX Security Symposium* 17–32 (2014).

71. Kamm, L., Bogdanov, D., Laur, S. & Vilo, J. A new way to protect privacy in large-scale genome-wide association studies. *Bioinformatics* **29,** 886–893 (2013).

72. Liu, C., Wang, X. S., Nayak, K., Huang, Y. & Shi, E. ObliVM: A programming framework for secure computation. in *Proceedings - IEEE Symposium on Security and Privacy* **2015–July,** 359–376 (2015).

73. Ghodsi, M., Liu, B. & Pop, M. DNACLUST: Accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics* **12,** (2011).

74. Nikolaenko, V. *et al.* Privacy-preserving matrix factorization. in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13* 801–812 (2013). doi:10.1145/2508859.2516751

75. Kolesnikov, V. & Schneider, T. Improved garbled circuit: Free XOR gates and applications. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **5126 LNCS,** 486–498 (2008).

76. Troncoso-Pastoriza, J. R., Katzenbeisser, S. & Celik, M. Privacy preserving error resilient dna searching through oblivious automata. in *Proceedings of the 14th ACM conference on Computer and communications security - CCS '07* 519 (2007). doi:10.1145/1315245.1315309

77. Huang, Y., Evans, D., Katz, J. & Malka, L. Faster Secure Two-party Computation Using Garbled Circuits. in *Proceedings of the 20th USENIX Conference on Security* 35 (USENIX Association, 2011).

78. Wang, X. S. *et al.* Efficient Genome-Wide, Privacy-Preserving Similar Patient Query based on Private Edit Distance. in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15* 492–503 (2015). doi:10.1145/2810103.2813725

79. Naveed, M. *et al.* Controlled Functional Encryption. in *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS '14* 1280–1291 (2014). doi:10.1145/2660267.2660291

80. Dachman-Soled, D., Malkin, T., Raykova, M. & Yung, M. Secure efficient multiparty computing of multivariate polynomials and applications. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **6715 LNCS,** 130–146 (2011).

81. Lauter, K., López-Alt, A. & Naehrig, M. Private computation on encrypted genomic data. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **8895,** 3–27 (2015).

82. Zahur, S., Rosulek, M. & Evans, D. Two halves make a whole reducing data transfer in garbled circuits using half gates. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9057,** 220–250 (2015).

83. Baldi, P., Baronio, R., De Cristofaro, E., Gasti, P. & Tsudik, G. Countering GATTACA: Efficient and Secure Testing of Fully-sequenced Human Genomes. in *Proceedings of the 18th ACM Conference on Computer and Communications Security* 691–702 (ACM, 2011). doi:10.1145/2046707.2046785

84. Ayday, E., Raisaro, J. L., Mclaren, P. J., Fellay, J. & Hubaux, J. Privacy-Preserving Computation of Disease Risk by Using Genomic , Clinical , and Environmental Data. *Proc. USENIX Secur. Work. Heal. Inf. Technol. (HealthTech" 13)* (2013).

85. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci.* **107,** 9546–9551 (2010).

86. Zgraggen, E., Zhao, Z., Zeleznik, R. & Kraska, T. Investigating the Effect of the Multiple Comparisons Problem in Visual Analysis. in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18* 1–

12 (2018). doi:10.1145/3173574.3174053

87. Dwork, C. *et al.* The reusable holdout: Preserving validity in adaptive data analysis. *Science (80-.).* **349,** 636–638 (2015).

88. Wickham, H., Cook, D., Hofmann, H. & Buja, A. Graphical inference for infovis. *IEEE Trans. Vis. Comput. Graph.* **16,** 973–979 (2010).

89. Buja, A. *et al.* Statistical inference for exploratory data analysis and model diagnostics. *Philos. Trans. A. Math. Phys. Eng. Sci.* **367,** 4361–83 (2009).

90. Majumder, M., Hofmann, H. & Cook, D. Validation of visual statistical inference, applied to linear models. *J. Am. Stat. Assoc.* **108,** 942–956 (2013).

91. Fisher, A., Anderson, G. B., Peng, R. & Leek, J. A randomized trial in a massive online open course shows people don't know what a statistically significant relationship looks like, but they can learn. *PeerJ* **2,** e589 (2014).

92. Su, X. & Khoshgoftaar, T. M. A Survey of Collaborative Filtering Techniques. *Adv. Artif. Intell.* **2009,** 1–19 (2009).

93. Wongsuphasawat, K. *et al.* Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations. *IEEE Trans. Vis. Comput. Graph.* **22,** 649–658 (2016).

94. Shneiderman, B. The eyes have it: A task by data type taxonomy for information visualizations. in *IN IEEE SYMPOSIUM ON VISUAL LANGUAGES* 336–343 (1996).

95. Sneath, P. H. A. The application of computers to taxonomy. *Microbiology* **17,** 201–226 (1957).

96. Paulson, J. N., Colin Stine, O., Bravo, H. C. & Pop, M. Differential abundance

analysis for microbial marker-gene surveys. *Nat. Methods* **10,** 1200–1202 (2013).

97. Morgan, X. C. & Huttenhower, C. Chapter 12: Human Microbiome Analysis. *PLoS Comput. Biol.* **8,** (2012).

98. White, J. R., Nagarajan, N. & Pop, M. Statistical Methods for Detecting Differentially Abundant Features in Clinical Metagenomic Samples. *PLoS Comput. Biol.* **5,** (2009).

99. Lozupone, C. & Knight, R. UniFrac : a New Phylogenetic Method for Comparing Microbial Communities UniFrac : a New Phylogenetic Method for Comparing Microbial Communities. *Appl. Environ. Microbiol.* **71,** 8228–8235 (2005).

100. Erlich, Y. & Narayanan, A. Routes for breaching and protecting genetic privacy. *Nature Reviews Genetics* **15,** 409–421 (2014).

101. Lax, S. *et al.* Longitudinal analysis of microbial interaction between humans and the indoor environment. *Science (80-.).* **345,** 1048–1052 (2014).

102. Kort, R. *et al.* Shaping the oral microbiota through intimate kissing. *Microbiome* **2,** (2014).

103. Snyder, P. Yao's garbled circuits: Recent directions and implementations. (2014). Available at: https://www.cs.uic.edu/pub/Bits/PeterSnyder/Peter_Snyder_-_Garbled_Circuits_WCP_2_column.pdf.