Interactive exploratory data analysis of Integrative Human Microbiome Project data using Metaviz [version 2; peer review: 3 approved]

Justin Wagner¹⁻³, Jayaram Kancherla²⁻³, Domenick Braccia²⁻³, James Matsumara⁴, Victor Felix⁴, Jonathan Crabtree⁴, Anup Mahurkar¹⁻⁴, Héctor Corrada Bravo¹⁻³

¹Department of Computer Science, University of Maryland, College Park, College Park, Maryland, 20742, USA
²Center for Bioinformatics and Computational Biology, University of Maryland, College Park, College Park, Maryland, 20742, USA
³Institute for Advanced Computer Studies, University of Maryland, College Park, College Park, Maryland, 20742, USA
⁴Institute for Genome Sciences, University of Maryland, Baltimore, Baltimore, Maryland, 21201, USA

Abstract
The rich data produced by the second phase of the Human Microbiome Project (iHMP) offers a unique opportunity to test hypotheses that interactions between microbial communities and a human host might impact an individual’s health or disease status. In this work we describe infrastructure that integrates Metaviz, an interactive microbiome data analysis and visualization tool, with the iHMP Data Coordination Center web portal and the HMP2Data R/Bioconductor package. We describe integrative statistical and visual analyses of two datasets from iHMP using Metaviz along with the metagenomeSeq R/Bioconductor package for statistical analysis of differential abundance analysis. These use cases demonstrate the utility of a combined approach to access and analyze data from this resource.

Keywords
metagenomics, visualization, R/Bioconductor, Integrative Human Microbiome Project

Open Peer Review

This article is included in the Bioconductor gateway.
Introduction

Metagenomics allows researchers to perform a microbial community census and investigate associations between host phenotype and community status. Metagenomics has been used successfully to track pathogen spread and identify intervention strategies in childhood malnutrition. Integrative analysis of samples using multiple sequencing technologies allows for comparison at various levels of granularity. The second phase of the Human Microbiome Project (iHMP) offers a unique opportunity to test hypotheses of interactions between the microbial community and the human host. To examine the iHMP data resource, we use Metaviz, an interactive microbiome exploratory data analysis and visualization tool, and metagenomeseq, an R/Bioconductor package for statistical analysis of differential abundance analysis, for combined visual and statistical analysis.

Human Microbiome Project Phase II

The second phase of the HMP, also called the Integrative Human Microbiome Project (iHMP), consisted of focused studies of three diseases – Inflammatory Bowel Disease (IBD), Type II Diabetes (T2D), and Multi-Omic Microbiome Study: Pregnancy Initiative (MOMS-PI). 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 the specific disease and consisted of separate cohorts.

Metaviz

Metaviz is a web-based interactive visualization tool for microbiome data analysis. The architecture consists of a JavaScript front-end suite of charts (based on D3.js and Canva) and a navigation component that lets users select portions of taxonomic hierarchies to visualize and analyze. Metaviz supports two backend data stores – a graph database and the metaviz 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.cbeb.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. The Earth Microbiome Project provides an interactive visualization web-application to analyze its data. EMPeror offers interactive 3D visualizations of PCA plots to show distances between microbiome samples. QiIME packages a number of tools for static plotting of Principal Coordinate Analysis and stacked bar plots. MetaPhlAn2 uses a visualization package called GraphPhlan to produce phylogenetic trees and other plots. The HMP2Data R/Bioconductor provides processed 16S sequencing data from the iHMP project in Bioconductor data structures. We implemented Metaviz using design patterns from Epiviz, an interactive epigenetics visualization tool, that visualizes data from a variety of epigenetic sequencing projects. We show how we leverage the microbiome measurement-based design of Metaviz to implement interactive exploration and hypothesis-testing of the iHMP resource.

Implementation

Metaviz integration with HMP infrastructure

The HMP Data Access and Coordination Center maintains a data repository and web portal (https://www.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 MOMS-PI studies as provided by the HMP2Data Bioconductor package into the UMD Metagenome Browser. A user can select each dataset from the application start screen. Figure 1 details the number of samples, with metadata to the extent available as of May 2020 from the HMP2Data package, from each project currently available in the UMD Metagenome Browser.

HMP Data Portal linking to Metaviz

When browsing the samples available from the HMP Data Portal, a user can view an individual abundance matrix in Metaviz using the Metaviz tool 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 sample. Figure 1A shows the direct link functionality for samples in the IBD dataset and resulting workspace in Metaviz (Figure 1B).

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...
Figure 1. Metaviz and iHMP data infrastructure integration. Top: iHMP data accessible through the UMD Metagenome Browser. Middle (A, B): Single sample link from data portal to UMD Metagenome Browser. Bottom (C, D): 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 three datasets (IBD, T2D, and MOMS-PI) 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 (A, B). Finally, the HMP Data Portal provides a "cart" functionality where a user can select multiple samples and download a manifest listing those files (C). A user can upload a manifest file containing selections from the 16S community abundance profiles from the same dataset (IBD, T2D, or MOMS-PI) to the UMD Metagenome Browser and a new Metaviz workspace is created with those files (D).

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. Figure 1C shows the manifest file workflow for samples from the IBD dataset and resulting workspace in Metaviz (Figure 1D).

Metaviz usage
For ease of use, we provide tutorials at [https://epiviz.github.io/tutorials/metaviz/](https://epiviz.github.io/tutorials/metaviz/). As a community resource, we plan to update the Metaviz database within a month of Bioconductor releases of HMP2Data. We maintain links to the HMP Data Portal through the update of HMP2Data package URLs and provide default workspaces for the HMP2 datasets as well as those in HMP16SData R/Bioconductor package (https://epiviz.github.io/metaviz-workspaces/). For generating data summaries, we recommend using the HMP2Data package with appropriate R libraries to summarize sample information and the interactive HMP Data Portal for data summaries over different samples or study attributes. We provide instructions in the metavizr vignette to handle visualizing data that would be added to an analysis session like protected variables from dbGAP.

Operation
The HMP Data Portal and Metaviz are web applications that can run in any modern browser. We recommend using Firefox (version 65 or later) or Chrome (version 65 or later) for best

<table>
<thead>
<tr>
<th>Dataset Name</th>
<th>Number of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBD stool file</td>
<td>154</td>
</tr>
<tr>
<td>IBD biopsy HMP2</td>
<td>178</td>
</tr>
<tr>
<td>T2D</td>
<td>2208</td>
</tr>
<tr>
<td>MOMS-PI</td>
<td>9107</td>
</tr>
</tbody>
</table>
performance. *Metavizr* is a Bioconductor package and general guidelines from Bioconductor for requirements and installation should be followed (https://bioconductor.org/install/).

**Use cases**

*metavizr* analysis of WGS vs 16S data from same samples

In the IBD cohort of the iHMP dataset, investigators sequenced a subset of samples 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. Figure 2 shows an example analysis with stacked plots and scatter plots that link to a single FacetZoom to compare the degree of consistency of the data across sequencing methods.

**IBD 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 iHMP. We use the taxonomic profiles for each phase available from the *HMP2Data* package and use the same taxonomic classification identifiers in the package. To upload project data on to the UMD Metagenome Browser, we extracted 16S count table and taxonomic annotation using the `otu_table()` and `tax_table()` methods of *HMP2Data* package. We then use `metagenomeseq` and *metavizr* to import the count data along with taxonomy and sample metadata into a neo4j graph database\(^5\) using the *metavizr* neo4j import functionality. We used Metaviz\(^6\) for exploratory analysis and `metagenomeseq` for confirmatory statistical testing. We examined the IBD Stool Pilot and IBD iHMP dataset separately.

**IBD Stool Pilot dataset**

The IBD Stool Pilot dataset contains 16S and whole metagenome sequencing results of stool samples from 41 Crohn’s disease (CD) subjects and 10 ulcerative colitis (UC) subjects. We focused our analysis on 16S sequencing and used Metaviz to visually identify taxa that showed a difference in abundance between CD and UC subjects. Figure 3 shows a typical visualization.

We also used `metagenomeseq` to test the differential abundance of features aggregated to each level of the taxonomy using the `fitFeatureModel` method that is based on a zero-inflated

---

**Figure 2.** 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 (bottom) shows the common taxonomic features, two Stacked Plots (middle) show the proportion of all features aggregated to the Order level, and a set of scatter plots (top) 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 were transformed to counts per 1000 reads. The scatter plots show the variability in taxonomic community census estimates between sequencing methods. A static similar stacked plot visualization is shown in the main iHMP consortium manuscript at the genus and species level across samples for comparison\(^6\). Metaviz allows users to make specific selections of the FacetZoom to compare taxa at various levels. The scatter plot also allows comparison at single sample resolution. Code to create this Metaviz session is available at the following gist: https://gist.github.com/jkanche/9216d465d18a0106be7a43f5340eb38a.
log-normal linear model. As shown in Table 1, two species had an absolute log fold-change greater than 1 and adjusted (Benjamini-Hochberg) p-value less than 0.1. Visually inspecting the IBD Stool Pilot data by aggregating counts to each level of the taxonomy we found the following features appearing differentially abundant: “c__Betaproteobacteria”, “o__Burkholderiales”, “f__Ruminococcaceae”, “g__Lachnospira”, “g__[Ruminococcus]”, “g__Faecalibacterium”, “s__:589277”, “s__:333166”, “s__:564806”, “s__:369227”, “s__:358104”, “s__:369486”, “s__:gnavus:360015”, “s__:prausnitzii:851865”. These taxonomic features describe paths in the taxonomy of the Kingdom Bacteria that was derived from the SILVA database. The documentation (https://ibdmdb.org/tunnel/public/HMP2/16S/1806/products) for the abundance profiles used in this analysis denotes that this taxonomic string was generated with the sequence of an OTU derived with the UPARSE algorithm that was mapped to the SILVA database. Among these features, “c__Betaproteobacteria” refers to Class Betaproteobacteria, “o__Burkholderiales” to Order Burkholderiales, while values towards the leaves of the taxonomy refer to entries in the SILVA database that have an identifier and a sequence but have not been provided formal names in the binomial nomenclature system. Comparing the visual analysis results and the metagenomeSeq differential abundance testing results in Table 1 shows that the taxonomic feature s__:369227

Table 1. metagenomeSeq analysis of IBD Stool 16S Pilot dataset.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Log fold change</th>
<th>se</th>
<th>p-value</th>
<th>Adjusted p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>s__:369227</td>
<td>1.864583442</td>
<td>0.431193725</td>
<td>1.53061E-05</td>
<td>0.000734694</td>
</tr>
<tr>
<td>s__:363232</td>
<td>1.193035074</td>
<td>0.275415013</td>
<td>1.47914E-05</td>
<td>0.000734694</td>
</tr>
</tbody>
</table>

We used the fitFeatureModel of metagenomeSeq and aggregated counts to each level of the taxonomic hierarchy. Our analysis identified s__:369227 under family Lachnospiraceae and s__:363232 under genus Dorea as differentially abundant between samples from subjects diagnosed with Ulcerative Colitis and Crohn’s Disease.
(member of the Lachnospiraceae family which are strictly anaerobic) was identified using both methods. Members of Lachnospiraceae are abundant in human intestinal tracts and have been linked specifically to production of butyric acid. Also, colonization with a specific strain of Lachnospiraceae in obese mice has been linked to development of hyperglycemia. The second taxon, , is a member of the genus which has recently been shown to be associated with diarrhea predominant irritable bowel syndrome.

**IBD iHMP**

The IBD iHMP dataset consists of samples from subjects with CD, UC, and those without IBD (nonIBD). For these samples, we analyzed the 16S sequencing data of an ileum biopsy from the first visit for each subject, which yielded 72 samples with 32 from CD, 18 from CD, and 22 from nonIBD. We used metagenomeSeq to compute an F-statistic to determine if any taxonomic feature is associated with at least one group using the fitZig method (based on a zero-inflated Normal linear model on log-transformed counts appropriate for multi-category experiment designs). Figure 4 shows an example using Metaviz to visualize abundance profiles for phylum Fusobacteria, which was found to be differentially abundant across the three groups. Differential abundance of members of this phylum has previously been reported in studies of IBD. Analysis code and results are available as Extended data.

**Conclusion**

In this work we presented software infrastructure linking Metaviz to the iHMP data resources. We detailed the 16S taxonomic community profile data from iHMP available in the UMD Metagenome Browser. We then described linking the UMD Metagenome Browser to the iHMP 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 iHMP IBD, to examine taxonomic feature abundances in samples from UC, CD, and those without IBD. These illustrative analyses demonstrate the utility of Metaviz for integrative analysis with the HMP data resources. Visual inspection of taxonomic features coupled with statistical testing provides an effective mechanism to explore and test associations between bacterial communities and their human hosts.

*Figure 4. IBD Biopsy iHMP Multiple Groups Analysis.* Using statistical analysis we identified taxonomic features that showed a difference in abundance between the three subject diagnosis categories: UC, CD, or nonIBD in the Fusobacteria phylum. This Metaviz workspace is available at: http://metaviz.cbcb.umd.edu/?ws=wHsHT56U8Rt.
Data availability
Source data
The 16S abundance matrices for IBD, T2D and the MOMS-PI studies were downloaded from the HMP2Data Bioconductor package. These datasets are then loaded into the neo4j graph database using import methods available in the metaviz15 Bioconductor package. These import scripts are available at https://gist.github.com/jkanche/c57d8220a33b41e21c4e6769a7ae7e4.

Extended data
Figshare: Differential Abundance Analysis - IBD (Figure 4).

References
15. The Neo4j Graph Database. Reference Source

This file contains differential abundance analysis code and results. Extended data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability
Source code available from: https://github.com/epiviz/Metaviz.
License: Artistic License version 2.0.
Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 2

Reviewer Report 23 June 2021

https://doi.org/10.5256/f1000research.56081.r87967

© 2021 Smirnova E. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Ekaterina Smirnova
Department of Biostatistics, Virginia Commonwealth University, Richmond, USA

No further comments to make.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Biostatistics (working on Human Microbiome Project data)

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 20 July 2020

https://doi.org/10.5256/f1000research.26859.r64702

© 2020 Waldron L. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Levi Waldron
Graduate School of Public Health and Health Policy, Institute for Implementation Science in Population Health, City University of New York, New York, NY, USA

The authors present a web tool for exploratory analysis of iHMP data, based on their existing metaviz software. The figures from the manuscript were easy to reproduce using the links and code provided. It seems like a useful new way to view, explore, and download iHMP data. I have
only a couple minor comments:

1. The figures are low-resolution and text is a bit blurry. The lightest yellow text in Figure 2 isn't readable.

2. From http://metaviz.cbcb.umd.edu/ it's not immediately obvious how to find the iHMP (HMP2) data. IBD and T2D do appear a ways down a list of available datasets, but this list takes a very long time to load for me ("loading datasets and sample annotations..."). I didn't find MOMS-PI but I may not have waited long enough. It would be worth providing direct links in the manuscript to workspaces for each dataset, like is already done for Figure 3.

3. Perhaps explain the feature variables like c__Betaproteobacteria, o__Burkholderiales, f__Ruminococcaceae, g__Lachnospira, g__[Ruminococcus], g__Faecalibacterium, s__:589277, s__:333166, s__:564806, s__:369227, s__:358104, s__:369486, s__gnavus:360015, s__prausnitzii:851865 for readers who are familiar with scientific taxonomy names but not this format.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics, bioonductor, metagenomics, human microbiome analysis.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 21 Apr 2021

Hector Corrada Bravo, University of Maryland, College Park, College Park, USA

Reviewer Comment: The figures are low-resolution and text is a bit blurry. The lightest yellow text in Figure 2 isn't readable.
Author Response: We updated the color scheme for Figure 2. We also updated figures with a higher resolution.

Reviewer Comment: From http://metaviz.cbcb.umd.edu/ it's not immediately obvious how to find the iHMP (HMP2) data. IBD and T2D do appear a ways down a list of available datasets, but this list takes a very long time to load for me ("loading datasets and sample annotations..."). I didn't find MOMS-PI but I may not have waited long enough. It would be worth providing direct links in the manuscript to workspaces for each dataset, like is already done for Figure 3.

Author Response: We developed default workspaces for IBD, T2D, and MOMS-PI as well as put them at https://epiviz.github.io/metaviz-workspaces/

Reviewer Comment: Perhaps explain the feature variables like c__Betaproteobacteria, o__Burkholderiales, f__Ruminococcaceae, g__Lachnospira, g__[Ruminococcus], g__ Faecalibacterium, s__:589277, s__:333166, s__:564806, s__:369227, s__:358104, s__:369486, s__:gnavus:360015, s__:prausnitzii:851865 for readers who are familiar with scientific taxonomy names but not this format.

Author Response: We added the following text to the manuscript: “These taxonomic features describe paths in the taxonomy of the Kingdom Bacteria that was derived from the SILVA database. The documentation (https://ibdmdb.org/tunnel/public/HMP2/16S/1806/products) for the abundance profiles used in this analysis denotes that this taxonomic string was generated using the sequence of an OTU derived with the UPARSE algorithm that was mapped to the SILVA database. Among these features, "c__Betaproteobacteria" refers to Class Betaproteobacteria, "o__Burkholderiales" to Order Burkholderiales, while values towards the leaves of the taxonomy refer to entries in the SILVA database that have an identifier and a sequence but have not been provided formal names in the binomial nomenclature system.”

Competing Interests: No competing interests were disclosed.

Reviewer Report 16 July 2020
https://doi.org/10.5256/f1000research.26859.r64700

© 2020 Pasolli E. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Edoardo Pasolli
University of Naples Federico II, Naples, Italy

The manuscript by Justin Wagner et al. presents an infrastructure that integrates Metaviz, an
interactive microbiome data analysis and visualization tool that was previously published by the same authors, with the iHMP Data Coordination Center web portal and the HMP2Data R/Bioconductor package. The authors give an overall overview of the infrastructure in addition to a couple of use cases.

This is a nice and timely contribution that helps in using the large and complex set of available HMP data. The manuscript is already well structured, I have just few comments:

1. Please check all the links reported in the text. For example, the link to the data repository and web portal (https://ihmpdcc.org) doesn't seem to work.

2. How will the proposed infrastructure deal with likely updates (in terms of new data) of the HMP web portal?

3. Which are the options available to save/export the results generated in the proposed infrastructure? Please comment more about this in the text.

**Is the rationale for developing the new software tool clearly explained?**
Yes

**Is the description of the software tool technically sound?**
Yes

**Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?**
Partly

**Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?**
Yes

**Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?**
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Microbiome and bioinformatics

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

**Author Response 21 Apr 2021**

**Hector Corrada Bravo**, University of Maryland, College Park, College Park, USA

**Reviewer Comment:** Please check all the links reported in the text. For example, the link to the data repository and web portal (https://ihmpdcc.org) doesn't seem to work.
Author Response: We have updated the link to [http://www.ihmpdcc.org](http://www.ihmpdcc.org) and checked that it resolves.

Reviewer Comment: How will the proposed infrastructure deal with likely updates (in terms of new data) of the HMP web portal?

Author Response: We updated the manuscript with the following: “As a community resource, we plan to update the Metaviz database within a month of Bioconductor releases of HMP2Data. We maintain links to the HMP Data Portal through the update of HMP2Data package URLs and provide default workspaces for the HMP2 datasets as well as those in HMP16SData R/Bioconductor package ([https://epiviz.github.io/metaviz-workspaces/](https://epiviz.github.io/metaviz-workspaces/))”

Reviewer Comment: Which are the options available to save/export the results generated in the proposed infrastructure? Please comment more about this in the text.

Author Response: We added the following tutorial to [https://epiviz.github.io/tutorials/metaviz/save_plots/](https://epiviz.github.io/tutorials/metaviz/save_plots/)

Competing Interests: No competing interests were disclosed.
samples by body site, etc.

3. MOMS-PI and T2D studies have protected meta data available through dbGap. If users apply for dbGap and get access, is there an option to merge meta data by sample ids and visualize using this tool?

4. It would be helpful to provide instructions how to create diversity boxplots by disease status (as in Figure 3). I tried creating these but could not produce them on the same plot.

5. I recommend making alpha diversity plots by disease status a default plot when the data is selected using visualization tool.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Biostatistics (working on Human Microbiome Project data)

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

**Author Response 21 Apr 2021**

**Hector Corrada Bravo,** University of Maryland, College Park, College Park, USA

**Reviewer Comment:** How does the tool handle HMP DAC and HMP2Data package data updates? There is discussion that user can upload additional data manifest from HMP DAC portal, but what if the data already available thought Metaviz tool is updated on the DAC or Bioconductor?

**Author Response:** We updated the manuscript with the following: “As a community
resource, we plan to update the Metaviz database within a month of Bioconductor releases of HMP2Data. We maintain links to the HMP Data Portal through the update of HMP2Data package URLs and provide default workspaces for the HMP2 datasets as well as those in HMP16SDATA R/Bioconductor package (https://epiviz.github.io/metaviz-workspaces/)

**Reviewer Comment:** Is there an option to download data summaries? Such as demographics table, number of samples by body site, etc.

**Author Response:** We updated the manuscript with the following: “For generating data summaries, we recommend using the HMP2Data package with appropriate R libraries to summarize sample information and the interactive HMP Data Portal for data summaries over different samples or study attributes. We provide instructions in the metavizr vignette to handle visualizing data that would be added to an analysis session like protected variables from dbGAP”

**Reviewer Comment:** MOMS-PI and T2D studies have protected meta data available through dbGap. If users apply for dbGap and get access, id there an option to merge meta data by sample ids and visualize using this tool?

**Author Response:** We added a vignette to the metavizR Bioconductor package (https://github.com/epiviz/metavizr/pull/2/commits/756cc049a2c9264355e15df42cfa64d71ee141ae#diff-46f4a8e59ff35c66f1220d9c6abf687d92771b424f4a6dfed1434c998cd500) for added a protected variable and then generate new plots with it.

**Reviewer Comment:** It would be helpful to provide instructions how to create diversity boxplots by disease status (as in Figure 3). I tried creating these but could not produce them on the same plot.

**Author Response:** We added a tutorial at metaviz.org for including a diversity plot (https://epiviz.github.io/tutorials/metaviz/diversity_plot/).

**Reviewer Comment:** I recommend making alpha diversity plots by disease status a default plot when the data is selected using visualization tool.

**Author Response:** We created default workspaces at https://epiviz.github.io/metaviz-workspaces/ with alpha diversity plots on either the disease or an interesting attribute.

**Competing Interests:** No competing interests were disclosed.
The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com