Title

PUMAA: A Platform for Accessible Microbiome Analysis in the Undergraduate Classroom

Authors

Mitchell, Keith
Ronas, Jiem
Dao, Christopher
et al.

Publication Date

2020

DOI

10.3389/fmicb.2020.584699

frontiers in Microbiology

ORIGINAL RESEARCH
published: 06 October 2020
doi: 10.3389/fmicb.2020.584699

PUMAA: A Platform for Accessible Microbiome Analysis in the Undergraduate Classroom

Keith Mitchell[1†], Jiem Ronas[1†], Christopher Dao[1], Amanda C. Freise[1], Serghei Mangul[2], Casey Shapiro[3] and Jordan Moberg Parker[1]*

[1] Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA, United States, [2] Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, United States, [3] Center for Educational Assessment, Center for the Advancement of Teaching, University of California, Los Angeles, Los Angeles, CA, United States

Improvements in high-throughput sequencing makes targeted amplicon analysis an ideal method for the study of human and environmental microbiomes by undergraduates. Multiple bioinformatics programs are available to process and interpret raw microbial diversity datasets, and the choice of programs to use in curricula is largely determined by student learning goals. Many of the most commonly used microbiome bioinformatics platforms offer end-to-end data processing and data analysis using a command line interface (CLI), but the downside for novice microbiome researchers is the steep learning curve often required. Alternatively, some sequencing providers include processing of raw data and taxonomy assignments as part of their pipelines. This, when coupled with available web-based or graphical user interface (GUI) analysis and visualization tools, eliminates the need for students or instructors to have extensive CLI experience. However, lack of universal data formats can make integration of these tools challenging. For example, tools for upstream and downstream analyses frequently use multiple different data formats which then require writing custom scripts or hours of manual work to make the files compatible. Here, we describe a microbial ecology bioinformatics curriculum that focuses on data analysis, visualization, and statistical reasoning by taking advantage of existing web-based and GUI tools. We created the Program for Unifying Microbiome Analysis Applications (PUMAA), which solves the problem of inconsistent files by formatting the output files from several raw data processing programs to seamlessly transition to a suite of GUI programs for analysis and visualization of microbiome taxonomic and inferred functional profiles. Additionally, we created a series of tutorials to accompany each of the microbiome analysis curricular modules. From pre- and post-course surveys, students in this curriculum self-reported conceptual and confidence gains in bioinformatics and data analysis skills. Students also demonstrated gains in biologically relevant statistical reasoning based on rubric-guided evaluations of open-ended survey questions and the Statistical Reasoning in Biology Concept Inventory. The PUMAA program and associated analysis tutorials enable students and researchers with no computational experience to effectively analyze real microbiome datasets to investigate real-world research questions.

Keywords: microbiome, 16S rRNA, software tool, GUI (Graphical User Interface), undergraduate education, curriculum, data visualisation, targeted amplicon sequencing

OPEN ACCESS

Edited by:
Mel Crystal Melendrez, Anoka-Ramsey Community College, United States

Reviewed by:
Tammy Tobin, Susquehanna University, United States
Henrik R. Nilsson, University of Gothenburg, Sweden

***Correspondence:**
Jordan Moberg Parker
jmobergparker@ucla.edu

[†] These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *17 July 2020*
Accepted: *14 September 2020*
Published: *06 October 2020*

Citation:
Mitchell K, Ronas J, Dao C, Freise AC, Mangul S, Shapiro C and Moberg Parker J (2020) PUMAA: A Platform for Accessible Microbiome Analysis in the Undergraduate Classroom. Front. Microbiol. 11:584699. doi: 10.3389/fmicb.2020.584699

INTRODUCTION

Engaging undergraduates in research has been consistently demonstrated to increase students' performance, attitudes, and retention in sciences (Lopatto, 2004; Russell et al., 2007; Eagan et al., 2013). In particular, course-based undergraduate research experiences (CUREs) have been touted as an inclusive and scalable model to bring these benefits to a diverse set of student populations (Harrison et al., 2011; Bangera and Brownell, 2014; Corwin et al., 2015; Shapiro et al., 2015; Hanauer et al., 2017). Microbiome research using marker gene metabarcoding is an attractive direction for CUREs, as sample collection is relatively straightforward and advances in sequencing technologies and reduced cost have made the acquisition of marker gene microbiome data easier than ever (Clooney et al., 2016; Jovel et al., 2016). The large microbiome datasets using a combination of marker genes targeting bacteria and archaea (16S), eukaryotes (18S), and fungi (ITS) give students an opportunity to ask a variety of questions ranging from the composition of their own oral microbiome to plant–microbe interactions (Rosenwald et al., 2012; Sanders and Hirsch, 2014; Wang et al., 2015; Weber et al., 2018; Parks et al., 2020; Sewall et al., 2020).

We designed a microbial ecology CURE as part of the interdepartmental Competency-Based Research Laboratory Curriculum at the University of California, Los Angeles (Shapiro et al., 2015). In this two-term (two 10-week quarters) curriculum students work in teams to conduct self-directed research projects, with a focus on developing critical thinking and quantitative skills. Under the umbrella of an instructor designated overarching research question, students in the microbial ecology CURE formulate and test hypotheses about the microbiomes of different environments. The functional profiles of microbial communities are just as important as the taxonomic composition (Langille, 2018), and the questions of "who is there?" and "what are they doing there?" are the guiding questions for the curriculum. In the first wet-lab term they use both cultivation-dependent techniques such as isolating bacteria from the soil and characterizing their functional capabilities, and cultivation-independent techniques such as extraction of environmental DNA (eDNA) for 16S rRNA (16S) sequencing. In the second computer-lab term they use a variety of phylogenetics programs and bioinformatics tools for analysis of microbiome taxonomic community profiles and Piphillin predicted functional profiles (Narayan et al., 2020).

A major challenge for the development of microbiome research for undergraduates is that marker gene amplicon microbiome data provided by sequencing providers requires a number of bioinformatic processing steps before it can be easily analyzed and visualized, a process with which not all instructors or researchers have familiarity (Carey and Papin, 2018; Garcia-Milian et al., 2018). Many of the available end-to-end data analysis packages such as Quantitative Insights Into Microbial Ecology (QIIME/QIIME 2) (Caporaso et al., 2010; Bolyen et al., 2019), mothur, and the Pipeline for Environmental DNA Metabarcoding Analysis (PEMA) (Zafeiropoulos et al., 2020) have steep learning curves, requiring at least some command line interface (CLI) programming skills, or familiarity with R (R: The R Project for Statistical Computing) in the case of phyloseq (McMurdie and Holmes, 2013, 2015) and PEMA, in order to perform data analysis and visualization. Teaching these skills may be outside the scope of the average undergraduate microbiology classroom. Fortunately, there are several microbiome data analysis and visualization tools that do not require command line, such as the Shiny web app ranacapa (Kandlikar et al., 2018) or locally installed programs with graphical user interfaces (GUIs) such as Statistical Analysis of Metagenomic Profiles (STAMP) (Parks and Beiko, 2010; Parks et al., 2014) and Cytoscape (Shannon et al., 2003). These are attractive tools for use in the undergraduate bioinformatics classroom where there is lack of time to devote to the steep learning curve necessary for installation and use of command line programs (Mangul et al., 2017).

Even with the increasing availability of GUI analysis tools, there is still the problem that the data output file formats from QIIME or custom commercial and academic pipelines such as MrDNA (mrdnalab, 2020) and Anacapa (Curd et al., 2019) do not match the data input file formats required for the GUI and web-based analysis and visualization tools. Formatting the different analysis pathway files into a single pipeline is a non-trivial task requiring either running scripts or hours of manual reformatting. To address this problem, we created PUMAA, the Program for Unifying Microbiome Analysis Applications, which takes the output files from QIIME, Anacapa, or MrDNA and reformats them directly for use in downstream GUI or web-based applications for microbiome analysis. Additionally, PUMAA both prepares files for upload to Piphillin for prediction of functional genes from the 16S taxonomy data, and queries the KEGG database to annotate the Piphillin gene predictions (Iwai et al., 2016; Narayan et al., 2020). Inferring functional profiles from 16S rRNA marker genes using programs like PiCRUSt (Langille et al., 2013; Douglas et al., 2020) or Piphillin are accessible options for researchers without the resources to perform full functional metagenomics (Laudadio et al., 2019).

Since classroom time is limited and our curriculum learning objectives focus on microbiome data analysis, visualization, and statistical reasoning rather than learning programming languages, the instructional staff runs the PUMAA program to generate the files necessary for several different GUI or web-based tools and provide them to students. The bioinformatics curriculum is scaffolded such that the students' progress in their microbiome research from phylogenies of individual bacterial isolates, to simple microbial community qualitative analyses, to quantitative diversity metrics, to statistical analysis of the microbial community profiles. We developed accompanying instructional modules, video tutorials, and a lab manual to teach students both the theory behind the analysis tools and the skills needed for visualizing and performing biostatistical methods on the data. The key tools and tutorials include inferring phylogenetic trees, analyzing community profiles and diversity metrics using Microsoft Excel pivot tables and ranacapa, statistical analysis of taxonomic and inferred functional profiles using STAMP, and using KEGG to assign functions to genes.

The curriculum was assessed using entry/exit surveys designed to gauge the students' confidence in integrating computational

analysis with microbiology, and the Statistical Reasoning in Biology Concept Inventory (SRBCI) (Deane et al., 2016). Analysis of entry and exit surveys saw an increase in students' self-reported conceptual understanding and confidence levels in using the analysis tools, as well as improved competencies with biostatistics as demonstrated by improvement in the SRBCI post-test. The PUMAA program and associated instructional materials provide a scaffolded learning experience for undergraduate students and make microbiome bioinformatics analyses accessible to novice researchers.

PUMAA – PROGRAM FOR UNIFYING MICROBIOME ANALYSIS APPLICATIONS OVERVIEW

Analyzing metabarcoded microbiome data is a complex multi-step process. Next-generation sequencing produces a variety of data files, which then need to be processed and quality checked before assigning taxonomic profiles (Zhang, 2016; Almeida et al., 2018). Most sequencing providers include basic bioinformatic processing in their pipelines, and provide taxonomic abundance tables and sequence FASTA files along with the raw data. These files can be then used in downstream analysis and visualization applications. However, each taxonomic assignment platform and analysis or visualization tool may have different data input and output formats that need to be reconciled, or have significant data pre-processing steps that need to occur before the various analyses can be performed.

Some sequencing providers, such as MrDNA (mrdnalab, 2020), produce taxonomy abundance tables that must be rearranged in order to be compatible with most visualization programs, but even for those that are in the right general format, many tools have specific formatting requirements. For example, the STAMP tool enforces a "strict hierarchy" requirement where no classification of taxonomy can exist at a lower level than one which was left unclassified. The following classification, from phylum to species: "Proteobacteria, Gammaproteobacteria, Enterobacteriales, unclassified, *Escherichia*, unclassified," will produce errors in STAMP because the family is unclassified even though the genus is classified. In addition, STAMP requires that all unclassified columns must be labeled so and cannot be left blank. Another tool, Cytoscape, requires that each sample identification and taxonomic identification be a unique row where the weight corresponds to the quantity of the given instance in order to create a network type visualization. Web server-based programs such as Piphillin (Iwai et al., 2016) may have file size upload limitations, necessitating sub-setting of the data. These formatting and processing steps need to be carried out independently on the taxonomy or functional data for each of the desired analysis and visualization platforms (**Figure 1A**).

PUMAA, the Program for Unifying Microbiome Analysis Applications, provides the solution to these problems by integrating all of the formatting and pre-processing steps required for the platforms and tools discussed here into a single unified protocol with an easy installation procedure (**Figure 1B**). In addition, PUMAA is easily expandable as it provides the ability to add a new analysis tool or taxonomic ID platform with one added operation. The PUMAA protocol unifies existing data analysis and visualization tools by formatting common amplicon (16S/18S/ITS) taxonomic data outputs from a variety of sources to be compatible with the input formats required for multiple basic and advanced microbiome analysis tools. Additionally, PUMAA integrates Piphillin inferred functional microbiome composition from the 16S taxonomy data. PUMAA provides both a CLI as well as a GUI to accommodate a spectrum of potential users. A CLI version is implemented to allow users with UNIX experience, or those who are interested in learning, to customize their analysis and build upon/automate the provided scripts (Mangul et al., 2017). The GUI is ideal for novice microbiome researchers with little experience on UNIX based systems, who are interested in quickly visualizing their microbiome marker gene amplicon data. Initial installation of the GUI does require running a small set of terminal installation commands, but subsequent usage is straightforward.

PUMAA Supports Input From Various Microbiome Data Pipelines

Currently PUMAA supports three microbiome raw data processing platforms and/or services: MrDNA, Anacapa, and QIIME 2 (Bolyen et al., 2019; Curd et al., 2019; mrdnalab, 2020). PUMAA formats the taxonomic abundance tables and sequence files created by these platforms for any marker gene amplicons, including 16S, 18S, ITS, and others, for downstream analysis and visualization (**Figure 2**).

MrDNA

MrDNA is a commercial full-service next generation sequencing provider that offers 16S, 18S, and ITS amplicon sequencing on a variety of platforms. Regardless of the sequencing platform, MrDNA provides free comprehensive taxonomic analysis in addition to raw data processing using their proprietary pipeline. The pipeline generates operational taxonomic unit (OTU) abundance tables with taxonomic identities and representative FASTA sequence files at each taxonomic level (kingdom, phylum, class, order, family, genus, species).

Anacapa

Anacapa is a software tool kit developed to process environmental DNA (eDNA) sequence data and assign taxonomy data for six marker genes targeting bacteria, archaea, algae, fungi, protozoa, plants, and animals (Curd et al., 2019). Anacapa creates a custom reference library for marker genes, generates amplicon sequence variants (ASV), and assigns taxonomies at each taxonomic level (domain, phylum, class, order, family, genus, species). ASVs have been proposed as a finer resolution replacement for OTU clustering based on sequence similarity (Callahan et al., 2017). Anacapa output includes a detailed taxonomy table with sequences and abundances for each ASV, as well as tables with taxonomies summarized at various percent confidence intervals.

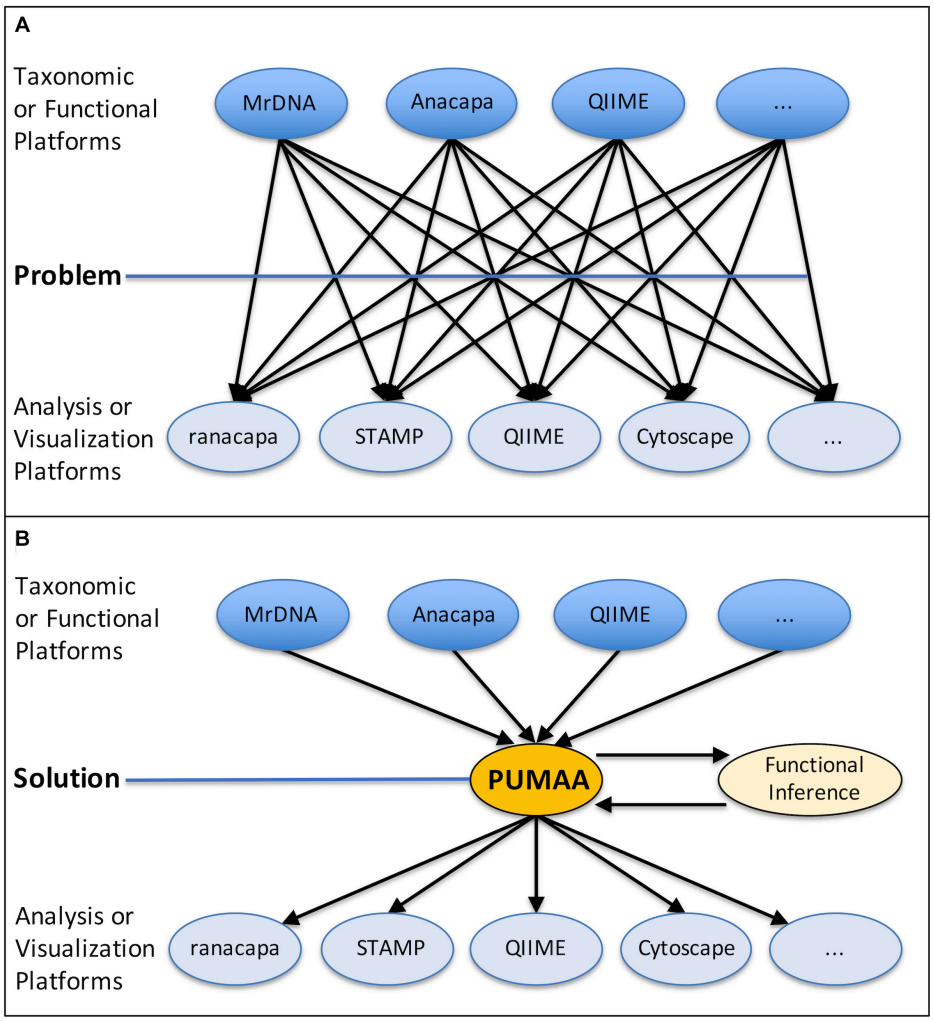

FIGURE 1 | The problem presented and the PUMAA solution. **(A)** The current problem is lack of unification of outputs from different taxonomic identification or functional inference platforms (MrDNA, Anacapa, QIIME, etc.) and the input data required by prospective analysis and visualization tools (ranacapa, STAMP, QIIME, Cytoscape, etc.). **(B)** PUMAA is a streamlined pipeline unifying the output files from multiple platforms and converting them to the input files necessary for varied analysis and visualization tools.

QIIME

QIIME is a powerful and widely adopted package for processing microbiome data, from raw sequences through taxonomy and data visualization. Tutorials and published protocols are available to walk users through standard data processing (Kuczynski et al., 2011), but the scope of QIIME may be daunting for novice users, even with the availability of the QIIME 2 Studio graphical interface (Bolyen et al., 2019). It also remains difficult to convert to other analysis/visualization platforms since QIIME provides users with OTU files and sequence files in the '.qza' format, which is unique to its platform.

PUMAA Supports Piphillin for Inferred Functional Profile Analysis

PUMAA formats taxonomic abundance (OTU or ASV) tables and representative sequence files for prediction of metagenomic content by Piphillin, which uses nearest-neighbor matching of 16S rRNA amplicons and full genomes (Iwai et al., 2016). Piphillin has the added benefits of a web interface and the ability to use any standard abundance table and representative sequence FASTA file, rather than relying on taxonomic assignments assigned from a specific reference phylogenetic tree, as in PiCRUSt (Langille et al., 2013). PiCRUSt2 has an extended database of reference genomes and broader compatibility, but still requires use of the command line for implementation (Douglas et al., 2020). A drawback to Piphillin is the 10 MB limit placed on uploaded file sizes in the web version. PUMAA addresses this by producing subset abundance and FASTA files that comply with these limits. The subset files are uploaded to the Piphillin server[1], and reference database and percent identity cutoffs are chosen [PUMAA currently only supports

[1]https://piphillin.secondgenome.com/

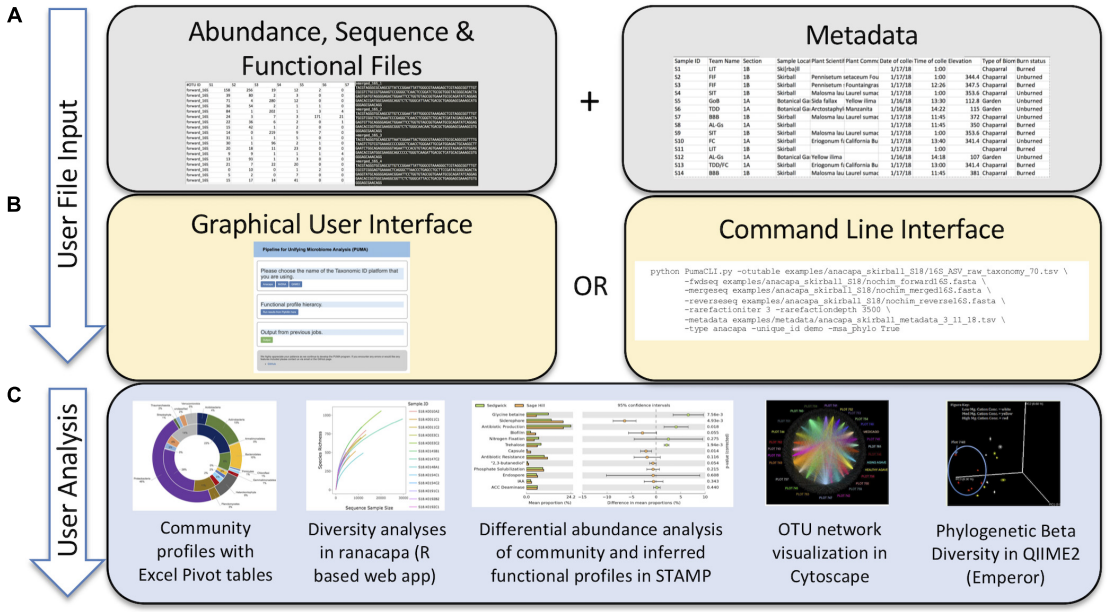

FIGURE 2 | Protocol of the PUMAA software. **(A)** The first panel as part of the "User File Input" displays the simple protocol to be performed by the user such as uploading metadata and various data formats of supported operational taxonomic unit, sequence, and functional file types. **(B)** The second panel as part of the "User File Input" displays the two forms of user interaction with PUMAA, through the GUI and CLI, which will enable community and functional profile analysis. **(C)** "User Analysis" shows the possible platforms for visualizing community/functional composition data enabled by user input such as STAMP, Excel, QIIME 2, and Cytoscape.

KEGG (Kanehisa, 2000; Kanehisa et al., 2004)], then results are emailed to the user as compressed.tar files. The other drawback to Piphillin is that it provides abundance tables for all predicted genes and pathways (identified by K and KO numbers), but not the associated annotations to assign biological information to the K/KO numbers. To address this, the PUMAA inferred function protocol also performs queries to the KEGG database in order to properly annotate the genes and pathways returned by Piphillin. Prior to PUMAA, this annotation process required command-line experience or labor-intensive manual curation.

PUMAA Supports a Variety of Analysis and Visualization Platforms

There are a wide variety of research questions that can be addressed using amplicon microbiome data, and the methods used for data analysis and visualization will vary based on the needs of the researcher. PUMAA focuses on processing and formatting user data to be compatible with a suite of readily available web-based or GUI data analysis and visualization tools. Using the PUMAA supported tools, researchers can explore data and test hypotheses by linking groups of samples or environmental parameters, otherwise known as metadata, to diversity metrics, community composition, and inferred functional profiles.

We have integrated PUMAA into a broad range of research analysis options (from simple to advanced) and visualization types (from bar charts to network analyses). In addition, PUMAA has options to complete data processing such as rarefaction subsampling to normalize for variation in sequence numbers between samples (McMurdie and Holmes, 2014; Willis, 2019), multiple sequence alignment (MSA) using MUSCLE (Edgar, 2004), and inference of phylogenetic trees using FastTree (Price et al., 2010).

Microsoft Excel

Microsoft Excel pivot tables are an easy way to begin to summarize the massive amounts of data in taxonomic abundance tables for visualizations of the overall community profile of different samples at different taxonomic levels (i.e., kingdom/domain, phylum, class, order, family, genus, species). Excel can also be easily used to make simple (non-statistical) comparisons of sample abundances at different taxonomic levels.

ranacapa

ranacapa (Kandlikar et al., 2018) is a user-friendly Shiny web application designed to explore biodiversity using environmental DNA metabarcoding data. It includes interactive visualizations and brief explanations of sequencing depth, alpha and beta diversity, and taxonomy distribution analyses such as bar plots and heatmaps. ranacapa was developed as an extension of the Anacapa toolkit (Curd et al., 2019), but can prove slightly difficult to access from other taxonomic identification platforms, like that of MrDNA.

STAMP (Statistical Analysis of Metagenomic Profiles)

STAMP (Parks et al., 2014) is a downloadable graphical interface that can quickly generate publication-quality graphics for differential abundance analysis of either taxonomy or functional

pathway data without the need to write code or use command-line interface. STAMP supports parametric and nonparametric statistical hypothesis testing for two-sample, two-group, and multiple-group comparisons. It emphasizes the use of effect size and confidence intervals in assessing biological relevance, and supports a variety of visualizations, including heatmaps, PCA plots, extended error bar plots, box plots, and bar plots.

QIIME 2 (Quantitative Insights Into Microbial Ecology)

QIIME 2 (Bolyen et al., 2019) provides numerous interactive and advanced data visualization tools and plugins for evaluation of metagenomic profiles (Caporaso et al., 2010; Kuczynski et al., 2011). Although QIIME can be used for end-to-end data analysis, some researchers may receive data processed by other platforms (e.g., MrDNA or Anacapa) and wish to feed the data back into the QIIME pipeline for analysis.

Cytoscape

Cytoscape (Kohl et al., 2011) is a unique open-source locally downloadable tool that enables the visualization of networks between community and functional profiles. Basic network analysis and visualization can be performed with the core distribution, with many additional features available as Cytoscape Apps.

Methods – PUMAA Protocol

Overview

The user executes a single script for both the GUI and CLI versions in order to execute the program. The PUMAA protocol consists of two key parts: (1) Production of all files for taxonomic community analysis, and (2) production of all files required for inferred functional analysis. PUMAA solves the problem of going from any of the taxonomic identification platforms to the multitude of visualization and analysis tools available by enforcing standardized files as part of the unification process. The user first obtains input files from one of the three supported pipelines (MrDNA, Anacapa, or QIIME2), identifies the metadata necessary for identifying and comparing samples (**Figure 2A**), and chooses to run PUMAA through either the GUI or CLI (**Figure 2B**). PUMAA verifies that the metadata sample IDs match the input data, then produces output files that can be used for a variety of analysis platforms (**Figure 2C**).

Protocol: PUMAA Installation and Requirements

PUMAA is freely available under the Apache-2.0 license at https://github.com/keithgmitchell/PUMAA and is supported by MacOSX and Linux; in addition, PUMAA works on Windows machines after installing the Linux subsystem Comprehensive installation instructions are provided on the Github page. Given software install is handled using conda, all versions of MacOSX and Linux that support the conda environment management software are viable options for usage and make for consistent and user-friendly install (Mangul et al., 2019). Issues or questions with the software can be submitted using the github issues feature: https://github.com/keithgmitchell/PUMAA/issues.

PUMAA is written in Python and the application's GUI is written using the Django web framework running locally. The example datasets all run on a laptop and use <1GB of memory when the MSA and Phylogenetic tree production is set as false. The QIIME 2 and MrDNA datasets run on a laptop and use <1GB of memory when the MSA and Phylogenetic tree production is set as true. The Anacapa dataset was unsuccessful on a laptop with 16GB RAM and was evaluated using a high-performance computing (HPC) cluster with 32GB of RAM and 3 h of runtime. Therefore, to produce a MSA and phylogenetic tree for datasets of this size, access to an HPC cluster, experience with CLI, and experience running jobs on HPC clusters may be required (**Table 1**).

Protocol: PUMAA Verifies Metadata

The user uploads their metadata describing the samples, taxonomy abundance (OTU or ASV) table and sequences from any given supported platform. The first part of the PUMAA protocol verifies the metadata and the taxonomy table to be sure the two files have consistent, alphanumeric sample identifiers which are unique compared to other forms of metadata validation (Rideout et al., 2016). This is a critical step as identifiable metadata is necessary for many downstream analysis steps, and some tools limit the types of characters accepted in the sample identifiers (e.g., underscores, but not periods, are acceptable in sample IDs in ranacapa).

Protocol: PUMAA Produces Files for Community Profile Analysis

PUMAA performs a variety of functions on the taxonomic abundance and sequence files in order to support the suite of tools discussed above. These functions include optional sample rarefaction at a user defined depth and number of iterations (max = 10) (Weiss et al., 2017), multiple sequence alignment by MAFFT (Katoh and Standley, 2013), phylogenetic tree construction via FastTree 2 (Price et al., 2010), and file formatting and annotation for ranacapa, STAMP, QIIME 2, Piphillin, and Excel. The protocol produces files for community profile analysis in the folder 'output,' or some other specified directory as an argument in the CLI. The output folder contains time-stamped subfolders for each PUMAA run, each containing subfolders with ready-to-run files for community profile analyses in Microsoft Excel, STAMP, ranacapa, and Cytoscape. In addition, pre-processed feature table (taxonomy), metadata, and phylogenetic tree files are created that can be imported directly into the QIIME 2 pre-configured virtual machine. A variety of analyses such as alpha- and beta-diversity can be performed in QIIME 2, as well as principal component analysis based on phylogenetic diversity metrics.

Protocol: PUMAA Produces Files for Inferred Functional Profile Analysis

The PUMAA protocol consists of three steps necessary for the generation and visualization of inferred functional profiles. The first step is automatically performed at the same time as the generation of the community profile analysis files. PUMAA creates a "piphillin" subfolder in

TABLE 1 | Dataset size, runtime, and memory usage with no rarefaction performed across the three example datasets.

Dataset	Dataset size (ASV/OTU count *10,000)	Fasta file size (MB)	Runtime (minutes)	FastTree/MAFFT peak memory usage (GB)	Python memory usage (GB)
MrDNA examples	0.3229	0.868	0.0778	0.207	0.02
QIIME 2 examples	0.0759	0.115	0.00517	0.044	0.02
Anacapa examples	3.6	1.789	1.24	12	0.075

the time-stamped output subfolder. This folder contains the original data formatted as a 'phiphillinotu.csv' taxonomic abundance table and a 'phiphillinseqs.fasta' representative sequence file. If the FASTA file exceeds the file size limit of 10 MB enforced by the Piphillin server, PUMAA subsamples the data into the number of necessary file sets of '.fasta' and '.csv' files (e.g., piphillinseqs1.fasta; piphillinseqs2.fasta; piphillinotu.csv1.csv; piphillinotu.csv2.csv). Second, each of the sets of Piphillin files in the output directory are uploaded to the Piphillin functional inference web server, which returns '.tar' files to the user via email.

Finally, the '.tar' files can then be run directly in the PUMAA protocol, which produces files for functional analysis that can be visualized using many of the same tools used for community profile analysis, including STAMP, Excel, and QIIME 2. Importantly, the PUMAA protocol also performs queries to the KEGG database using the KEGG genes to pathway API in order to properly annotate the Piphillin gene estimations (Kawashima et al., 2003). The BRITE hierarchy file of the KEGG database is downloaded and used to evaluate the functional hierarchy based on Piphillin pathway estimations. This ensures that estimated gene expression levels and hierarchy levels are inferred using the actively updated information. Annotating the genes and pathway expression from Piphillin is necessary when producing data visualizations with informative identifiers, and greatly reduces the need for manual querying of KEGG.

PUMAA produces a timestamped output subfolder for the functional profile files, including a gene description and functional hierarchy file designated for use in STAMP and Excel. This file contains annotated gene names and functional pathways, as opposed to just "K number" identifiers, and vastly increases the efficiency and ease of data analysis and visualization. PUMAA also produces weighted functional network files for usage in Cytoscape, which is a platform for visualizing important gene networks between samples.

Sample Data

The sample data used here and in the tutorials was generated by UCLA students in the winter and spring quarters of 2018, where they investigated the effect on rhizosphere microbial communities following the Skirball wildfire of December 2017 (Skirball Fire, 2020). Sample collection kits and sample sequencing were provided by the California Environmental DNA (CALeDNA) program, a community science initiative monitoring California's biodiversity through eDNA (Meyer et al., 2019), and the 16S sequences were processed using the Anacapa toolkit (Curd et al., 2019). The sample data for QIIME 2 is the same as the "moving pictures" human microbiome example dataset available on the QIIME 2 website[2].

Results

PUMAA Input and Output Files

The PUMAA pipeline creates output files formatted specifically for the needed input files for each of the data analysis and visualization platforms described in **Supplementary Table 1**.

PUMAA – CURRICULUM OVERVIEW

The Microbiology, Immunology, and Molecular Genetics (MIMG) undergraduate degree program at UCLA requires the completion of a two-quarter authentic research experience. An option to fulfill this requirement is to take the MIMG 109AL/BL: Research Immersion Laboratory in Microbiology series. This laboratory series is designed to prepare its students with the proper background and training to work in microbiology research, and has been demonstrated to improve their critical thinking and research skills as part of the life science curriculum (Shapiro et al., 2015). The 109AL/BL laboratory curriculum is discovery-based and driven by student-generated hypotheses tested using both cultivation-dependent and cultivation-independent techniques. The first term emphasizes experimental design and isolation of bacteria in a wet lab environment, and the second term focuses on the analysis of 16S sequencing data from individual isolates and 16S rRNA microbial community profiles. Students work in teams to conduct an original research project within the context of an overarching research question for the microbial ecology course, focusing on the interactions between plants and soil-associated bacteria. Recent course projects have involved collaborations with researchers at UCLA and beyond studying plant–microbe interactions in California grasslands (Kandlikar et al., 2020), analysis of the soil microbial communities of a Los Angeles urban farm (St. Clair et al., 2020), and a longitudinal study on the recovery of soil microbial communities following the 2017 Skirball fire in Los Angeles, CA, United States. The Skirball fire project was conducted in conjunction with the California Environmental DNA (CALeDNA) program's efforts to catalog California's biodiversity (Meyer et al., 2019).

In order for the MIMG 109AL/BL lab series to respond to the need for more computationally minded scientists (Bialek and Botstein, 2004; Campbell et al., 2007; Brewer and Smith, 2011), it

[2]https://docs.qiime2.org/2020.2/tutorials/moving-pictures/

was necessary to introduce new modules and tutorials that would sufficiently integrate bioinformatics and statistics with biology in ways that aspiring undergraduate researchers can comprehend (Aikens and Dolan, 2014). We created a comprehensive set of step-by-step tutorials (documents, presentations, and videos) designed to provide students with the necessary theory and skills to use the GUI analysis and visualization tools described in Section 2.3 (Excel, ranacapa, and STAMP), as well as the theory behind inference of metagenomic functional profiles using Piphillin. Although not a biostatistics course, the PUMAA-associated curriculum allows these students to learn about the computational tools available to researchers and the importance of integrating their knowledge of microbiology with statistical and quantitative support.

All tutorials are publicly available at https://sites.google.com/g.ucla.edu/pumaa/home.

First Term – Sample Collection and Bacterial Isolation/Characterization

The first term of the curriculum takes place in the wet lab and closely follows the cultivation-dependent experiments described in units 1–4 of the "I, Microbiologist" (Sanders and Miller, 2010) course textbook and lab manual. In brief, students collect bulk soil and decide on enrichment strategies for isolation of bacteria related to their research questions (e.g., antibiotic production and resistance or plant growth-promoting properties). Students then perform phenotypic characterization of bacterial isolates and 16S rRNA PCR and sequencing. In addition to collecting bulk soil for cultivation-dependent experiments, students also collect separate soil samples for environmental DNA (eDNA) extraction and 16S rRNA high-throughput sequencing for bacterial community profile analysis.

Second Term – Bioinformatics Analysis of 16S rRNA Genes Using PUMAA

In the second term, students use bioinformatics to interpret, expand, or refine 16S rRNA gene datasets generated in MIMG 109AL. Students generate 16S rRNA phylogenetic trees to assign taxonomic identities to their isolates and use statistical tools to make comparisons of the microbial communities from different environments. The course is divided into five Core Concept Modules. The first module (Phylogenetic Trees) concludes the analysis of bacterial isolates, and the other four modules focus on microbiome data analysis and visualization using the PUMAA output files: Community Profiles, Diversity Metrics, Statistical Analysis of Taxonomic Profiles, and Inferring Metagenomic Functional Profiles (**Figure 3A**). Students could also elect to perform optional advanced independent analysis on their data using QIIME or Cytoscape. Each of the modules includes written and/or video tutorials and was assessed with a combination of reading assessments and reflection questions (**Figure 3B**). This bioinformatics course was assessed using pre- and post-course concept inventories and surveys. Learning objectives, activities, and tutorials for each of the Core Concept Modules are outlined in **Supplementary File 1**.

Curriculum Assessment Methods

Study Sample

The study sample consisted of six cohorts of junior and senior level students who enrolled in MIMG 109BL (Advanced Research in Microbiology) in Spring 2016, Spring 2017, Winter 2018, Spring 2018, Winter 2019, and Spring 2019. This yielded an initial population of 143 students. **Table 2** provides a summary of demographic characteristics for these students. Instructor J.M.P. taught the spring cohorts and instructor A.F. taught the winter cohorts. Prerequisites for enrollment in MIMG 109BL included MIMG 109AL (Research Immersion in Microbiology) and either Statistics 13 (Introduction to Statistical Methods for Life and Health Sciences) or Life Sciences 40 (Statistics of Biological Systems).

Assessment Data Collection and Analyses

The study utilized two sources of data: student assignments and self-report surveys. Data collected included qualitative and quantitative measures. UCLA's Institutional Review Board (IRB) gave approval to work with human subjects on all aspects of the assessment (IRB #10-000904).

Administration of Self-Report Surveys

Two self-report surveys were administered to all students in the course. Surveys included a broad collection of open- and closed-ended questions, some developed by the instructors and evaluation team. Students were given the entry survey at the start of the second term and asked to indicate how well they thought they understood key learning goals related to data analysis and their confidence in their ability to analyze data using various visualization plots. The exit survey was completed at the end of the term and had matched questions to the first survey, as well as additional survey questions asking them to assess the quality and usefulness of the tutorials and instructional materials. Both surveys also included open-ended content-related questions. The surveys were piloted in 2016 and 2017 and were given to students anonymously through the course management system as low-stakes (completion points) assessments to increase response rate and reduce response bias (Furnham, 1986). Starting in Winter 2018, these items were added to a comprehensive curricular assessment plan administered electronically by external evaluators (see Shapiro et al., 2015) for details on survey data collection). Of the 143 students who took the course between Spring 2016 and Spring 2019, 141 completed the first survey (98.6% response rate) and 132 completed the second survey (92.3% response rate). The surveys are available as **Supplementary File 2**.

Administration of SRBCI Concept Inventory

The Statistical Reasoning in Biology Concept Inventory (SRBCI) is a series of multiple-choice questions to test students on concepts including statistical significance, basic graph/trend interpretation, and assessing hypotheses based on results (Deane et al., 2016). The twelve questions on the SRBCI pre- and post-tests are designed to identify students' common misconceptions in statistical analysis and track their learning progress as a result of the pedagogical interventions. The concept inventory

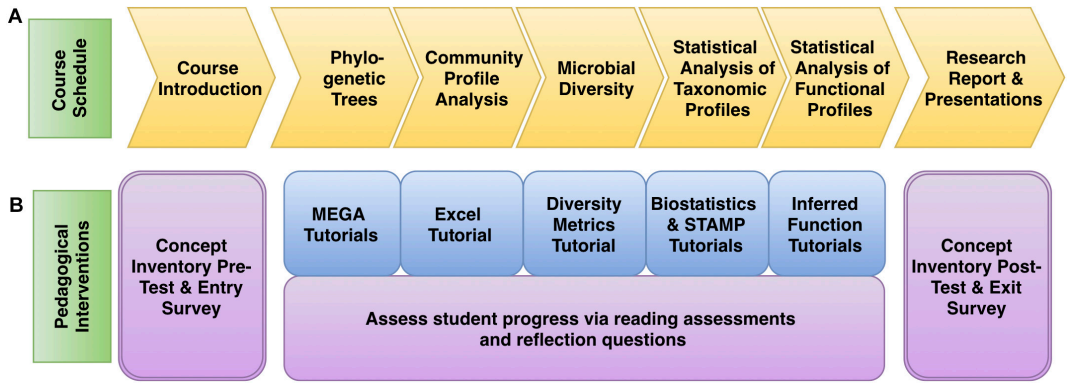

FIGURE 3 | Microbiome analysis course schedule with pedagogical interventions. **(A)** The progress of the course followed the concept goals as outlined in yellow. **(B)** The pedagogical interventions are described, with tutorials in blue and assessment materials in purple.

TABLE 2 | Study sample demographics.

	Number of students (*N*)	Percent of students (%)
Female	81	56.6%
Transfer student[a]	34	23.8%
URM[b]	34	23.8%
Pell Grant Recipient[c]	53	37.1%
Total	143	100%

Academic terms: Spring 2016, Spring 2017, Winter 2018, Spring 2018, Winter 2019, Spring 2019. [a]Transfer to UCLA, usually from a 2-year institution. [b]Under-Represented Minority (URM) students include American Indian, Native American, Black Non-Hispanic, and Hispanic students. [c]Received Pell Grant for one or more terms while enrolled at UCLA; Pell Grant Recipient is a proxy for low socioeconomic status.

was administered as an anonymous low-stakes (ungraded) in-class activity at the start and end of the second term to the first two cohorts of students in Spring 2016 and Spring 2017. The study design, intended to gauge authentic learning gains across the curriculum by reducing "math anxiety" (Ashcraft and Moore, 2009), necessarily resulted in the inability to assess individual student learning gains using this metric. The pre-test and post-test were administered to a total of 52 and 50 students, respectively. Statistical reasoning gains between the pre-test and post-test groups were assessed using descriptive and Mann–Whitney nonparametric tests to account for variations in sample size.

Analyses of Closed-Ended Quantitative Survey Data

The closed-ended survey questions quantitatively ranked the students' agreement with a statement or confidence with a certain concept using a five-point Likert scale ranging from "Not at all" to "Very well/Very confident." Scores for matched questions were averaged across all participants to compare results from the Entry and Exit Surveys. Survey items asking students about the usefulness of learning activities were rated on a five-point Likert scale where 1 = "Don't remember," 2 = "Not useful," 3 = "Somewhat useful," 4 = "Very useful," and 5 = "Essential." Descriptive analyses of matched pre/post-survey close-ended items were conducted to explore students' change in self-reported confidence and changes in their self-reported levels of understanding. To test for statistical differences between the overall means of the Entry and Exit Survey items, descriptive and Mann–Whitney nonparametric tests were performed on the combined survey data from all cohorts to account for variations in sample size. Because the responses for the Spring 2016 and Spring 2017 surveys were anonymous, we were unable to pair the data by student. Wilcoxon signed ranks (paired nonparametric) tests were conducted on just the surveys administered by the external evaluators from Winter 2018 to Spring 2019, in order to see if there were differences between the all the data and the matched data. Since both sets of tests were significant, we were confident in using the aggregated data and the Mann–Whitney nonparametric tests to report our results.

Analyses of Open-Ended Qualitative Survey Data

Open-ended questions related to course content were included in the Entry and Exit surveys, allowing students to respond in their own words. Of particular interest was a question that asked students to describe the relationship between *p*-value (statistical significance) and effect size (biological significance). A 4-point rubric assessing students' level of proficiency with statistical concepts was used to gather direct evidence of student learning gains (**Supplementary File 3**). Student responses to open-ended questions were scored on a scale of 1 point = no familiarity (i.e., students indicated that they are not familiar with the concept), and 2–4 points for novice, intermediate, and advanced proficiency, respectively. Responses left blank were unscored. All student responses (both pre and post) were randomized and pooled by the external evaluator, then provided to the raters. The rubric was developed and refined by J.R., A.F., and J.M.P. through iterative rounds of scoring a subset of sample responses followed by consensus discussion. All responses were scored independently by all three raters, and interrater reliability (IRR) as determined by Randolph's free-marginal multirater kappa, was 0.49 (61.8% overall agreement) indicating moderate agreement. To account for the IRR variations, the median score for each response was used to assess whether pre-post gains

were statistically significant between the groups using both the Mann–Whitney nonparametric test and a *t*-test.

Curriculum Assessment Results

Conceptual and Confidence Gains From Self-Reported Surveys

We wanted to assess if students would be able to formulate and statistically test hypotheses linking environmental parameters (metadata) to diversity metrics, community composition, and inferred functional profiles. Students were assessed using entry/exit surveys designed to gauge the students' comfort with integrating computational analysis with microbiology. At the beginning of the term the students reported, on average, "very little" understanding of key learning objectives such as how to use and assess the results of bioinformatics databases, and which statistical tests to use and how to interpret them (**Figure 4A**). By the end of the term students reported they understood these

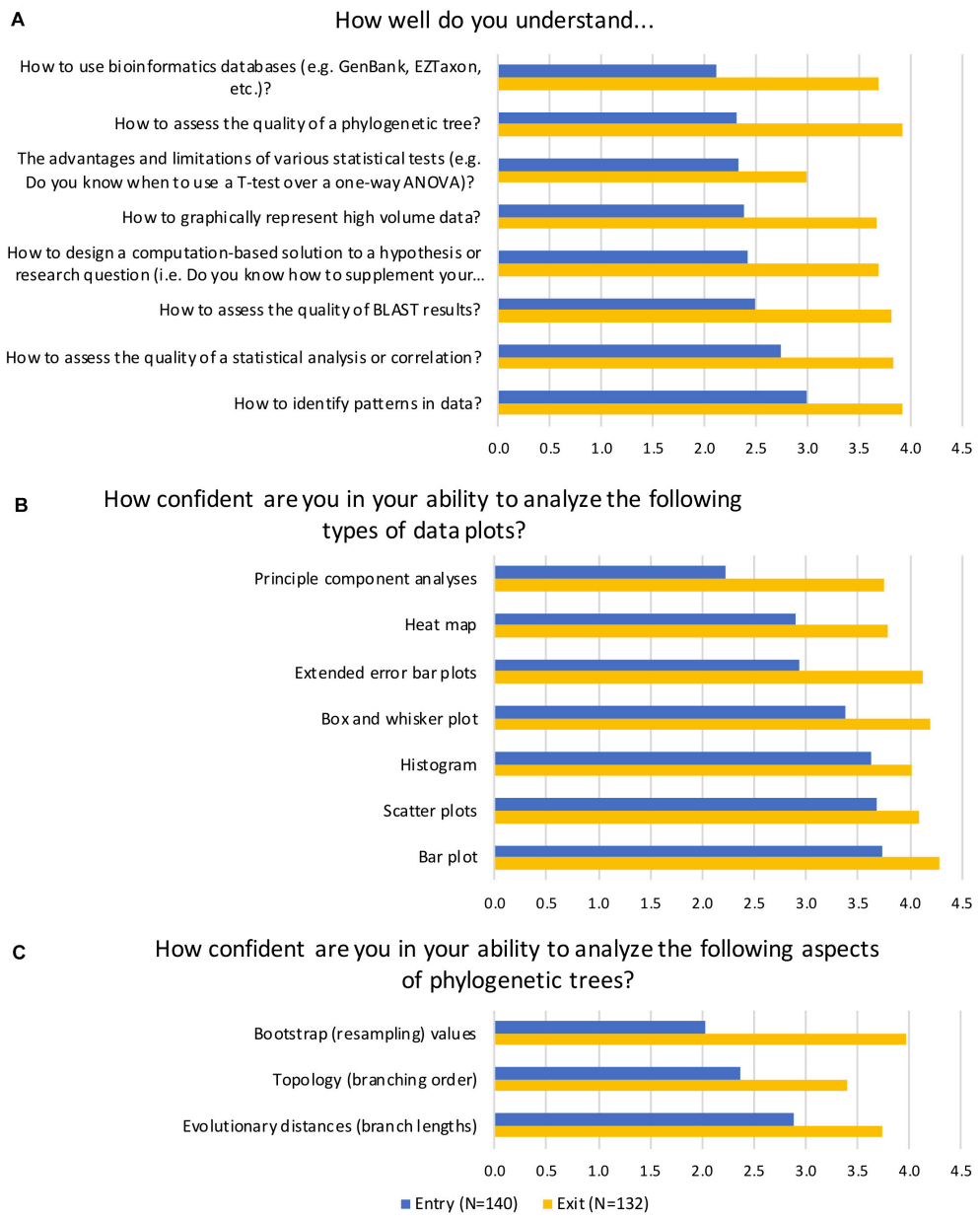

FIGURE 4 | Average ranked responses to selected entry and exit survey questions. In self-reported survey questions, students were asked to indicate **(A)** their level of understanding of key learning goals, **(B)** their confidence in their ability to analyze common data plots, and **(C)** their confidence in their ability to analyze aspects of phylogenetic trees. Average scores on a five-point Likert scale are reported for matched questions. A score of 1 = Not at all, 2 = Very little/Not very, 3 = Fairly well/confident, 4 = Quite well/confident, and 5 = Very well/confident. Students reported significant gains in their understanding and confidence in all categories ($p < 0.001$).

TABLE 3 | Ranked usefulness of STAMP learning activities.

STAMP learning activity	Average score on five-point Likert Scale (*N* = 131)
Hands-on use of the program	4.5
One-on-one discussions with instructional staff	4.3
Tutorials (documents and videos)	3.6
Reading/reading assessment of STAMP user guide or articles	3.0

concepts on average "fairly well" to "quite well," a statistically significant change based on Mann–Whitney nonparametric tests for all measures ($p < 0.001$). Of note, students were generally less confident of their understanding of "the advantages and limitations of various statistical tests (e.g., Do you know when to use a *T*-test over a one-way ANOVA)?" at the end of the term. This result was somewhat to be expected because the statistical analysis tool they used, STAMP, aims to promote best practices by suggesting a statistical hypothesis test based on the input data (Parks and Beiko, 2010). Therefore, students had limited practice with this particular skill.

In addition to performing statistical tests, STAMP generates a variety of data visualization plots, and we wanted to assess how confident students were in their ability to analyze these plots (**Figure 4B**). Mann–Whitney results indicated a statistically significant change in students' self-reported levels of confidence ($p < 0.001$). Specifically, at the start of the term students reported being "fairly" to "quite" confident in their ability to analyze common plots such as scatter plots, bar plots, and histograms. They had much less confidence, however, in their ability to interpret principal component analysis (PCA), heat maps, and extended error bar plots. By the end of the term they were "quite confident" on average in their ability to analyze most of the plots, and had dramatically improved their confidence in PCA, heat map analyses, and extended error bar plots. Another key learning objective of the course was the ability to interpret phylogenetic trees and analyze their statistical support (**Figure 4C**). At the start of the term, students reported being "not very" confident in their ability to assess bootstrap or resampling values, which are an indication of the of statistical confidence in a clade (Efron et al., 1996), and "not very" to "fairly" confident in their ability to interpret topology and evolutionary distances. By the end of the term, students had significantly increased their confidence in their ability to analyze all aspects of phylogenetic trees ($p < 0.001$).

Tutorials

STAMP was an essential component of the curriculum and was central for many of the student data analysis and visualization learning outcomes. We wanted to find out which learning activities the students found to be the most helpful in preparing them to use and interpret data in STAMP. Students reported that tutorials we created were useful, but perhaps unsurprisingly, it was actual use of the program and discussing it with the instructional staff that the students found to be essential (**Table 3**). All tutorials are publicly available at https://sites.google.com/g.ucla.edu/pumaa/home.

Statistical Reasoning and Conceptual Gains Measured by the SRBCI and Open-Ended Survey Responses

We used the SRBCI to directly assess student learning gains in core concepts related to repeatability of results, variations in data, hypotheses and predictions, and sample size. Students took the pre-test in the first week of the term and the post-test at the end of the term following the completion of all of the analysis modules. Scores for the pre-tests and post-tests were binned by number of correct responses and plotted to compare the overall distribution of scores (**Figure 5**). The distribution of the post-test scores is more skewed to the right, demonstrating overall improvement on the SRBCI for the combined cohorts. Statistical reasoning gains between the pre-test and post-test groups were assessed using a Mann–Whitney nonparametric test. There was a statistically significant increase in pre-test (Mean = 58.7%, Mean Rank = 44.3, $N = 52$) to post-test (Mean = 69.3%, Mean Rank = 59.0, $N = 50$) scores ($p = 0.01$) on the SRBCI.

A rubric-guided assessment of an open-ended survey question was used to determine whether the curricular interventions resulted in an increased understanding of the relationship between statistical significance (p-value) and biological significance (effect size). At the beginning of the term, 63.9% of students had no familiarity with the concept or held novice understanding, meaning the responses indicated they didn't know, or they had multiple or complete misconceptions (**Figure 6**). By the end of the term, 78.4% of students held intermediate to advanced levels of understanding, and were able to demonstrate conceptual understanding of the relationship to varying degrees. The rubric scores from the Exit survey (Mean = 3.14, Mean Rank = 164.4, $N = 125$) were significantly

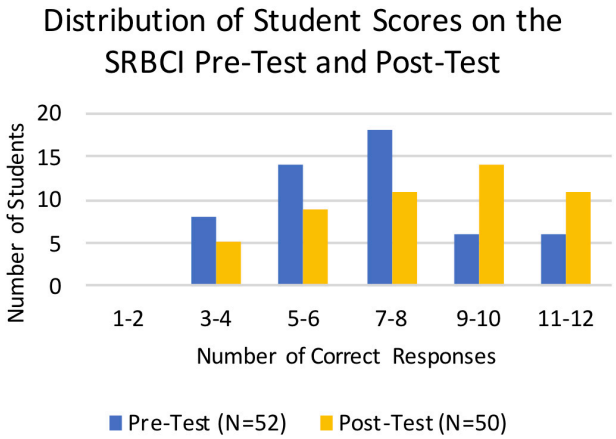

FIGURE 5 | Distribution of Student scores on the SRBCI pre-test and post-test. The number of students plotted on the vertical axis is binned by the number of correct responses shown on the horizontal axis. The blue columns represent the pre-test scores and yellow columns represent the post-test scores. There was a significant increase in the SRBCI scores from the pre-test to the post-test ($p = 0.01$).

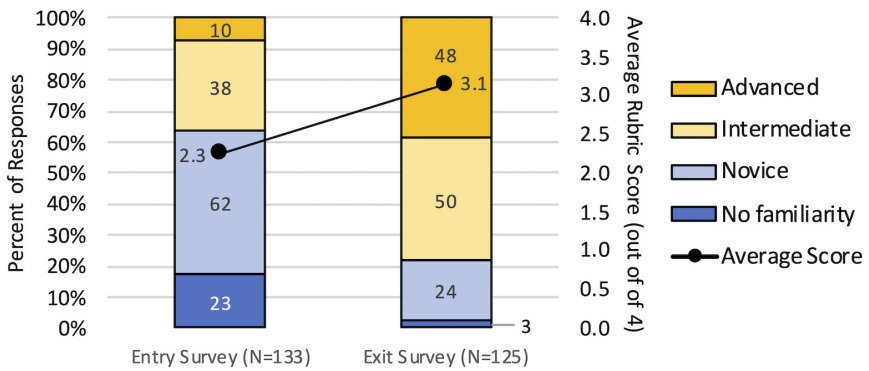

FIGURE 6 | Conceptual gains in understanding the relationship between statistical and biological significance. Student responses to the open-ended question were evaluated using a rubric to assign them a level of competency from 1 = No familiarity, 2 = Novice, 3 = Intermediate, and 4 = Advanced. The primary axis indicates the percent of student responses demonstrating each level of competency for the entry and exit surveys. Blue indicates lower competency levels and yellow indicates higher competency levels. The secondary axis indicates the average score for all responses; there was a significant increase in the average score from the entry surveys to the exit surveys ($p < 0.001$).

higher than the Entry survey (Mean = 2.26, Mean Rank = 96.7, $N = 135$) by both the Mann–Whitney and t-tests ($p < 0.001$). These results demonstrate the shift from lower levels of competency to higher levels of competency in understanding the relationship between statistical and biological significance.

DISCUSSION

The increased availability of microbiome and other "big data" data sets has coincided with calls for life science undergraduates to have bioinformatics "minimum skill sets" or "core competencies" in order to meet the growing demand to analyze that data (Tan et al., 2009; Welch et al., 2016; Mulder et al., 2018; Sayres et al., 2018). PUMAA has been in use in the Research Immersion in Microbiology undergraduate laboratories at UCLA for a number of years, resulting in the development of a suite of instructional materials and tutorials to train students in many of the bioinformatics skills necessary to meet this demand. This curriculum focused on quantitative literacy, which is the intersection of critical thinking, math/statistics, and real-world contexts, and has been highlighted by the Association of American Colleges and Universities as an essential skill for undergraduates (Elrod, 2014). The PUMAA curriculum and associated analysis and visualization tools gave students opportunities to use multiple bioinformatic approaches to analyzing their data. Repeated practice with tools and integration of said tools into student-driven research projects increased self-reported confidence with data visualization and analysis. For example, use of STAMP enabled students to perform statistical tests on microbiome community and functional profiles, and improved their competence with statistical concepts such as statistical significance and biological significance. This was of particular interest due to the tendency of notice researchers to over interpret p-values and disregard the importance of effect sizes and confidence intervals (Nakagawa and Cuthill, 2007; Martínez-Abraín, 2008).

PUMAA presents a user-friendly, time-and-cost-effective approach to processing, analyzing, and visualizing marker gene microbiome data. It improves the accessibility and range of available microbiome investigations by providing users with a simple way to unify the output of various taxonomic identification platforms with a suite of tools for data analysis and visualization. The protocol accomplishes this by producing properly configured, formatted, and annotated files for analysis of taxonomic community profiles and inferred functional profiles. This process of data manipulation can often be performed by sequencing services for additional fees or completed by users with significant time commitment, both of which could be barriers for those with funding or time constraints. PUMAA is an open-source solution which is highly accessible to a wide spectrum of users, including undergraduates or other researchers interested in learning to conduct microbiome analyses, as it can be used as a GUI as well as a CLI. It provides an easy and flexible interface for a variety of users requiring a clear and brief interface for production of files needed for diversity analysis and data visualization for analysis of targeted amplicon sequencing studies. The demand for tools that meet this need is evidenced by the recent development of DNA metabarcoding data processing tools like the web-based SLIM (Dufresne et al., 2019) and minimal coding-required PEMA (Zafeiropoulos et al., 2020). Both of these tools produce OTU and/or ASV tables from raw metabarcode data that could be incorporated into the PUMA input pipeline for downstream data analysis and visualization.

In practice, the instructional staff runs the PUMAA program and provides students with files ready for use in Excel, ranacapa,

STAMP, and other tools. One limitation of this approach is that students do not get direct experience with command-line bioinformatics, which is one of the core competencies for undergraduate life sciences education described by several different bioinformatics curriculum committees (Tan et al., 2009; Welch et al., 2016; Mulder et al., 2018; Sayres et al., 2018). However, the International Society for Computational Biology's Curriculum Task Force has refined their core competencies and designated different user profiles requiring different *levels* of competency (Mulder et al., 2018). For example, an undergraduate in a 10-week microbial ecology course may be considered a "bioinformatics user," rather than a "bioinformatics scientist" or "bioinformatics engineer," and the steep learning curve required to gain CLI skills may not be practical with the limited time available. We focused instead on training students to perform all of the bioinformatic analyses needed for an authentic course-based undergraduate research experience in microbial ecology. PUMAA is not intended to replace comprehensive CLI tools such as QIIME or mothur, but rather serve as an entry point for novice researchers to analyze and visualize their datasets. Students that express interest in expanding their bioinformatics skills can be directed to a wealth of tutorials and resources for learning to code.

The PUMAA program and the curriculum described here have the potential to have a wide impact by making marker gene microbiome research accessible to researchers with multiple levels of experience, and with the included instructional module documents, it can be practically implemented in a classroom setting for undergraduates.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/ **Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the UCLA Institutional Review Board (UCLA IRB). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KM designed the PUMAA program and wrote the manuscript. JR created instructional materials, designed assessments, collected and analyzed assessment data, and contributed to the manuscript. CD contributed to writing the program, created instructional materials, and contributed to the manuscript. CS collected and analyzed assessment data, and contributed to the manuscript. SM consulted on the program and contributed to the manuscript. AF created instructional materials, designed assessments, and contributed to the manuscript. JMP designed the curriculum, created assessments, conceptualized the PUMAA program, and wrote the manuscript. All authors have reviewed and approved the manuscript.

FUNDING

Curricular development and assessment were supported by a grant from the UCLA Center for the Advancement of Teaching (IIP#15-02). Institutional support for the laboratory curriculum is provided by the Division of Life Sciences in the College of Letters and Science and the Department of Microbiology, Immunology and Molecular Genetics at UCLA.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.biorxiv.org/ (Mitchell et al., 2018). We would like to thank Annabel Beichman for writing early versions of the Python file conversion scripts and Joseph Olivier for contributions to analysis tutorials. We would like to thank Kris Reddi, without whom the MIMG instructional laboratories curriculum wouldn't be possible. We would like to acknowledge the Institute for Quantitative and Computational Biology (QCBio) at UCLA. We would like to acknowledge Nathan Kraft and Gaurav Kandlikar in the Ecology and Evolutionary Biology Department at UCLA, and Savanah St. Clair and Marcie Sakadjian from Pierce College, for collaborating on our research projects. We would like to thank Rachel Meyer, Emily Curd, Teia Schweizer, Miroslava Munguia Ramos, and Ana Garcia Vedrenne from the University of California Conservation Genomics Consortium CALeDNA Project for help with project design, preparation of sample collection kits, data collection and storage, eDNA extraction and sequencing library preparation, and Anacapa sequence data processing. And finally, we would like to thank and acknowledge all of the undergraduate researchers in the Research Immersion in Microbiology curriculum.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.584699/full#supplementary-material

Supplementary Table 1 | PUMMA_Input and Output files.

Supplementary File 1 | PUMAA_Curriculum details.

Supplementary File 2 | PUMAA_Surveys.

Supplementary File 3 | PUMAA_Rubric_Revised.

REFERENCES

Aikens, M. L., and Dolan, E. L. (2014). Teaching quantitative biology: goals, assessments, and resources. *Mol. Biol. Cell* 25, 3478–3481. doi: 10.1091/mbc.E14-06-1045

Almeida, A., Mitchell, A. L., Tarkowska, A., and Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience* 7:giy054. doi: 10.1093/gigascience/giy054

Ashcraft, M. H., and Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *J. Psychoeduc. Assess.* 27, 197–205. doi: 10.1177/0734282908330580

Bangera, G., and Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci. Educ.* 13, 602–606. doi: 10.1187/cbe.14-06-0099

Bialek, W., and Botstein, D. (2004). Introductory science and mathematics education for 21st-Century biologists. *Science* 303, 788–790. doi: 10.1126/science.1095480

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Brewer, C. A., and Smith, D. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washigntone, DC: Am. Assoc. Adv. Sci.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Campbell, A. M., Ledbetter, M. L. S., Hoopes, L. L. M., Eckdahl, T. T., Heyer, L. J., Rosenwald, A., et al. (2007). Genome Consortium for Active Teaching: meeting the goals of BIO2010. *CBE Life Sci. Educ.* 6, 109–118. doi: 10.1187/cbe.06-10-0196

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carey, M. A., and Papin, J. A. (2018). Ten simple rules for biologists learning to program. *PLoS Comput. Biol.* 14:e1005871. doi: 10.1371/journal.pcbi.1005871

Clooney, A. G., Fouhy, F., Sleator, R. D., O' Driscoll, A., Stanton, C., Cotter, P. D., et al. (2016). Comparing apples and oranges: next generation sequencing and its impact on microbiome analysis. *PLoS One* 11:e0148028. doi: 10.1371/journal.pone.0148028

Corwin, L. A., Graham, M. J., and Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: an agenda for future research and evaluation. *CBE Life Sci. Educ.* 14:es1. doi: 10.1187/cbe.14-10-0167

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., et al. (2019). Anacapa Toolkit: an environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods Ecol. Evol.* 10, 1469–1475. doi: 10.1111/2041-210X.13214

Deane, T., Nomme, K., Jeffery, E., Pollock, C., and Birol, G. (2016). Development of the statistical reasoning in biology concept inventory (SRBCI). *CBE Life Sci. Educ.* 15:ar5. doi: 10.1187/cbe.15-06-0131

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38, 685–688. doi: 10.1038/s41587-020-0548-6

Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J., and Cordier, T. (2019). SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics* 20:88. doi: 10.1186/s12859-019-2663-2

Eagan, M. K., Hurtado, S., Chang, M. J., Garcia, G. A., Herrera, F. A., and Garibay, J. C. (2013). Making a difference in science education the impact of undergraduate research programs. *Am. Educ. Res. J.* 50, 683–713. doi: 10.3102/0002831213482038

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13429–13429. doi: 10.1073/pnas.93.23.13429

Elrod, S. (2014). *Quantitative Reasoning: The Next "Across the Curriculum" Movement*. Available online at: https://www.aacu.org/peerreview/2014/summer/elrod (accessed June 25, 2020).

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personal. Individ. Differ.* 7, 385–400. doi: 10.1016/0191-8869(86)90014-0

Garcia-Milian, R., Hersey, D., Vukmirovic, M., and Duprilot, F. (2018). Data challenges of biomedical researchers in the age of omics. *PeerJ* 6:e5553. doi: 10.7717/peerj.5553

Hanauer, D. I., Graham, M. J., SEA-PHAGES, Betancur, L., Bobrownicki, A., Cresawn, S. G., et al. (2017). An inclusive research education community (iREC): impact of the SEA-PHAGES program on research outcomes and student learning. *Proc. Natl. Acad. Sci. U.S.A.* 114, 13531–13536. doi: 10.1073/pnas.1718188115

Harrison, M., Dunbar, D., Ratmansky, L., Boyd, K., and Lopatto, D. (2011). Classroom-based science research at the introductory level: changes in career choices and attitude. *CBE Life Sci. Educ.* 10, 279–286. doi: 10.1187/cbe.10-12-0151

Iwai, S., Weinmaier, T., Schmidt, B. L., Albertson, D. G., Poloso, N. J., Dabbagh, K., et al. (2016). Piphillin: improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS One* 11:e0166104. doi: 10.1371/journal.pone.0166104

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* 7:459. doi: 10.3389/fmicb.2016.00459

Kandlikar, G. S., Gold, Z. J., Cowen, M. C., Meyer, R. S., Freise, A. C., Kraft, N. J. B., et al. (2018). ranacapa: an R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations. *F1000Research* 7:1734. doi: 10.12688/f1000research.16680.1

Kandlikar, G. S., Yan, X., Levine, J. M., and Kraft, N. J. B. (2020). Quantifying microbially mediated fitness differences reveals the tendency for plant-soil feedbacks to drive species exclusion among California annual plants. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.13.948679

Kanehisa, M. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kawashima, S., Katayama, T., Sato, Y., and Kanehisa, M. (2003). KEGG API: a web service using SOAP/WSDL to access the KEGG system. *Genome Inform.* 14, 673–674.

Kohl, M., Wiese, S., and Warscheid, B. (2011). "Cytoscape: software for visualization and analysis of biological networks," in *Data Mining in Proteomics: From Standards to Applications Methods in Molecular Biology*, eds M. Hamacher, M. Eisenacher, and C. Stephan (Totowa, NJ: Humana Press), 291–303. doi: 10.1007/978-1-60761-987-1_18

Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., and Knight, R. (2011). "Using QIIME to analyze 16S rRNA gene sequences from microbial communities," in *Current Protocols in Bioinformatics*, ed. A. D. Baxevanis (Hoboken, NJ: John Wiley & Sons, Inc).

Langille, M. G. I. (2018). Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *mSystems* 3:e00163-17. doi: 10.1128/mSystems.00163-17

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Laudadio, I., Fulci, V., Stronati, L., and Carissimi, C. (2019). Next-generation metagenomics: methodological challenges and opportunities. *OMICS J. Integr. Biol.* 23, 327–333. doi: 10.1089/omi.2019.0073

Lopatto, D. (2004). Survey of undergraduate research experiences (SURE): first findings. *Cell Biol. Educ.* 3, 270–277. doi: 10.1187/cbe.04-07-0045

Mangul, S., Martin, L. S., Hoffmann, A., Pellegrini, M., and Eskin, E. (2017). Addressing the digital divide in contemporary biology: lessons from teaching UNIX. *Trends Biotechnol.* 35, 901–903. doi: 10.1016/j.tibtech.2017.06.007

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., et al. (2019). Challenges and recommendations to improve the installability

and archival stability of omics computational tools. *PLoS Biol.* 17:e3000333. doi: 10.1371/journal.pbio.3000333

Martínez-Abraín, A. (2008). Statistical significance and biological relevance: a call for a more cautious interpretation of results in ecology. *Acta Oecol.* 34, 9–11. doi: 10.1016/j.actao.2008.02.004

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

McMurdie, P. J., and Holmes, S. (2015). Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking. *Bioinformatics* 31, 282–283. doi: 10.1093/bioinformatics/btu616

Meyer, R. S., Curd, E. E., Schweizer, T., Gold, Z., Ramos, D. R., Shirazi, S., et al. (2019). The California environmental DNA "CALeDNA" program. *bioRxiv* [Preprint]. doi: 10.1101/503383

Mitchell, K., Dao, C., Freise, A., Mangul, S., and Parker, J. M. (2018). PUMA: a tool for processing 16S rRNA taxonomy data for analysis and visualization. *bioRxiv* [Preprint]. doi: 10.1101/482380

mrdnalab (2020). Available online at: http://www.mrdnalab.com/16s-ribosomal-sequencing.html (Accessed June 24, 2020).

Mulder, N., Schwartz, R., Brazas, M. D., Brooksbank, C., Gaeta, B., Morgan, S. L., et al. (2018). The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Comput. Biol.* 14:e1005772. doi: 10.1371/journal.pcbi.1005772

Nakagawa, S., and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* 82, 591–605. doi: 10.1111/j.1469-185X.2007.00027.x

Narayan, N. R., Weinmaier, T., Laserna-Mendieta, E. J., Claesson, M. J., Shanahan, F., Dabbagh, K., et al. (2020). Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics* 21:56. doi: 10.1186/s12864-019-6427-1

Parks, D. H., and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26, 715–721. doi: 10.1093/bioinformatics/btq041

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494

Parks, S., Joyner, J. L., and Nusnbaum, M. (2020). Reaching a large urban undergraduate population through microbial ecology course-based research experiences. *J. Microbiol. Biol. Educ.* 21:21.1.17. doi: 10.1128/jmbe.v21i1.2047

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Rideout, J. R., Chase, J. H., Bolyen, E., Ackermann, G., González, A., Knight, R., et al. (2016). Keemei: cloud-based validation of tabular bioinformatics file formats in google sheets. *GigaScience* 5:27. doi: 10.1186/s13742-016-0133-6

Rosenwald, A. G., Arora, G. S., Madupu, R., Roecklein-Canfield, J., and Russell, J. S. (2012). The human microbiome project: an opportunity to engage undergraduates in research. *Proc. Comput. Sci.* 9, 540–549. doi: 10.1016/j.procs.2012.04.058

Russell, S. H., Hancock, M. P., and McCullough, J. (2007). Benefits of undergraduate research experiences. *Science* 316, 548–549. doi: 10.1126/science.1140384

Sanders, E. R., and Hirsch, A. M. (2014). Immersing undergraduate students into research on the metagenomics of the plant rhizosphere: a pedagogical strategy to engage civic-mindedness and retain undergraduates in STEM. *Front. Plant Sci.* 5:157. doi: 10.3389/fpls.2014.00157

Sanders, E. R., and Miller, J. H. (2010). *I, Microbiologist: A Discovery-Based Course in Microbial Ecology and Molecular Evolution*. Washington, DC: ASM Press.

Sayres, M. A. W., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS One* 13:e0196878. doi: 10.1371/journal.pone.0196878

Sewall, J. M., Oliver, A., Denaro, K., Chase, A. B., Weihe, C., Lay, M., et al. (2020). Fiber force: a fiber diet intervention in an advanced course-based undergraduate research experience (CURE) course. *J. Microbiol. Biol. Educ.* 21:21.1.40. doi: 10.1128/jmbe.v21i1.1991

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shapiro, C., Moberg-Parker, J., Toma, S., Ayon, C., Zimmerman, H., Roth-Johnson, E. A., et al. (2015). Comparing the impact of course-based and apprentice-based research experiences in a life science laboratory curriculum. *J. Microbiol. Biol. Educ.* 16, 186–197. doi: 10.1128/jmbe.v16i2.1045

Skirball Fire (2020). *Wikipedia*. Available online at: https://en.wikipedia.org/w/index.php?title =Skirball_Fire&oldid =948253595(Accessed June 27, 2020).

St. Clair, S., Saraylou, M., Melendez, D., Senn, N., Reitz, S., Kananipour, D., et al. (2020). Analysis of the soil microbiome of a Los Angeles urban farm. *Appl. Environ. Soil Sci.* 2020:e5738237. doi: 10.1155/2020/5738237

Tan, T. W., Lim, S. J., Khan, A. M., and Ranganathan, S. (2009). A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. *BMC Genomics* 10:S36. doi: 10.1186/1471-2164-10-S3-S36

Wang, J. T. H., Daly, J. N., Willner, D. L., Patil, J., Hall, R. A., Schembri, M. A., et al. (2015). Do you kiss your mother with that mouth? An authentic large-scale undergraduate research experience in mapping the human oral microbiome†. *J. Microbiol. Biol. Educ.* 16, 50–60. doi: 10.1128/jmbe.v16i1.816

Weber, K. S., Bridgewater, L. C., Jensen, J. L., Breakwell, D. P., Nielsen, B. L., and Johnson, S. M. (2018). Personal microbiome analysis improves student engagement and interest in Immunology, Molecular Biology, and Genomics undergraduate courses. *PLoS One* 13:e0193696. doi: 10.1371/journal.pone.0193696

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Welch, L., Brooksbank, C., Schwartz, R., Morgan, S. L., Gaeta, B., Kilpatrick, A. M., et al. (2016). Applying, evaluating and refining bioinformatics core competencies (an update from the curriculum task force of ISCB's education committee). *PLoS Comput. Biol.* 12:e1004943. doi: 10.1371/journal.pcbi.1004943

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Front. Microbiol.* 10:2407. doi: 10.3389/fmicb.2019.02407

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., et al. (2020). PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience* 9:giaa022. doi: 10.1093/gigascience/giaa022

Zhang, H. (2016). Overview of sequence data formats. *Methods Mol. Biol.* 1418, 3–17. doi: 10.1007/978-1-4939-3578-9_1

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