**iMAP: An integrative pipeline for bioinformatics analysis and exploratory visualization of microbiome data**

**Teresia M Buza1,2§, Triza Tonui3, Francesca Stomeo3,4, Christian Tiambo3, Robab Katani1,5, Megan Schilling1,6, Beatus Lyimo7, Paul Gwakisa8, Isabella Cattadori1,9, Joram Buza7, Vivek Kapur1,5,6,7**

1The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania, USA; 2Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania, USA; 3Biosciences eastern and central Africa-International Livestock Research Institute (BecA-ILRI) Hub, Nairobi, Kenya; 4The European Molecular Biology Laboratory (EMBL), Heidelberg, Germany; 5Applied Biological and Biosecurity Research Laboratory, Pennsylvania State University, University Park, Pennsylvania, USA; 6Department of Animal Science, Pennsylvania State University, University Park, PA, USA; 7Nelson Mandela African Institute of Science and Technology, Arusha, Tanzania; 8Sokoine University of Agriculture, Morogoro, Tanzania; 9Department of Biology, Pennsylvania State University, University Park, PA, USA.

§Corresponding author

Email addresses

TMB: ndelly@gmail.com

TT: [trizatonui91@gmail.com](mailto:trizatonui91@gmail.com)

FS: [stomeofra@gmail.com](mailto:stomeofra@gmail.com)

CT: [C.Tiambo@cgiar.org](mailto:C.Tiambo@cgiar.org)

RK: [rxk104@psu.edu](mailto:rxk104@psu.edu)

MS: [mas985@psu.edu](mailto:mas985@psu.edu)

BL: [beatus.lyimo@nm-aist.ac.tz](mailto:beatus.lyimo@nm-aist.ac.tz)

PG: [psgwakisa@gmail.com](mailto:psgwakisa@gmail.com)

IC:  [imc3@psu.edu](mailto:imc3@psu.edu)

JB: [joram.buza@nm-aist.ac.tz](mailto:joram.buza@nm-aist.ac.tz)

VK: [vkapur@psu.edu](mailto:vkapur@psu.edu)

**ABSTRACT**

**Background:** The major challenge facing investigators in the microbiome field is turning the large numbers of reads generated by next-generation sequencing (NGS) platforms into biological knowledge. Effective robust analytical workflows that guarantee reproducibility and result provenance are essential requirements for unleashing biological value from complex data. Several cutting-edge bioinformatics tools have been developed to help in microbiome data analysis. Few of the finest tools extensively used are Mothur and qiime. These tools follow a similar pattern where an input file typically in FASTQ format is loaded into a pipeline to produce an output file commonly known as OTU table. The OTU table is the main input for most downstream analyses. The big challenge is lack of robust review process of the intermediate output. Reviewing output progressively enables investigators to discover systematic errors, verify if the data is right for answering research questions being addressed or suitable for drawing accurate conclusions.

**Results:** We have developed a robust and reproducible bioinformatics and exploratory visualization pipeline named iMAP (Integrative Microbiome Analysis Pipeline) to address some of the challenges in microbiome data analysis. Simultaneously, we introduced a review-as-you-go (RAYG) approach which is facilitated by multiple reports generated progressively at each major step of the pipeline. For the most part, the iMAP pipeline streamlines functionalities implemented in Mothur platform and uses assorted R packages for visualization and R-markdown for generating progress reports. We have used a case study to demonstrate the robustness and reproducibility of the iMAP pipeline by analyzing 360 mouse gut samples downloaded from a public repository.

**Conclusions:** The iMAP pipeline brings a unique bioinformatics tool to the fast-growing field of the microbiome. The integrated RAYG approach provides investigators with great opportunity to discover and correct any observed problem before further downstream analyses. The intensively analyzed case study set a model for a reproducible microbiome data analysis. Adopting iMAP and RAYG approach in microbiome data analysis will enable investigators to generate more research questions that lead to precise conclusions and informed decision making. We anticipate that users will find this tool broadly useful and adaptable to their needs.

**Keywords:** iMAP, microbiome, exploratory data analysis, 16S rRNA, reproducibility, visualization, R-markdown, phylogenetic annotation.

**BACKGROUND**

Microbiome data analysis is a very complex process that requires a thorough review of the input and output generated at every step of the analysis. The core step in microbiome analysis is taxonomy classification of the representative sequences and clustering of OTUs (Operational Taxonomic Units). OTUs are pragmatic proxies for potential microbial species represented in a sample. Performing quality control of the sequences prior to taxonomy classification can highlight the basic properties of the dataset especially the poor quality reads and residual contamination. There are several public tools available for inspecting read quality and filtering the poor-quality reads as well as removing any residue contamination. For example, pre-processing tools such as Seqkit [1], FASTQC [2] and bbduk.sh command available in the BBMap package [3] are designed to help investigators review the properties and quality of reads before further downstream analyses. High quality reads coupled with stringent screening and filtering can significantly reduce the number of spurious OTUs.

Major microbiome analysis tools integrate different quality control approaches in their pipelines. Mothur [4] for example is excellently known for its intensive quality filtering of poor sequences before OTU clustering and taxonomy assignment. QIIME-2, a successor of QIIME-1 [5] (see http://qiime.org/) uses DADA2 [6] to obtain high-quality representative sequences before aligning them using MAFFT [7] software. Nevertheless, the most common sequencing error is the formation of chimeric fragments during PCR amplification process. Basically, chimeras are false recombinants formed when prematurely terminated fragments during PCR process reanneal to another template DNA. This eliminates the assumption that an amplified sequence may have originated from a single microbial organism. Detecting and removing chimeric sequences is crucial for obtaining quality sequence classification results. Both Mothur and QIIME-2 integrate special tools for chimera removal specifically UCHIME [8] and VSEARCH [9].

The sequences that pass the filtering process are typically searched against a known reference taxonomy classifier at a pre-determined threshold. Most classifiers are publicly available including the Ribosomal Database Project (RDP) [10], SILVA [11], Greengenes [12], and EzBioCloud [13] some of which are frequently updated and some are outdated. In some cases, users may opt to train their own classifiers using, for example, q2-feature-classifier protocol [14, 15] available in QIIME-2 or use any other suitable method. Over-classification of the representative sequences can result in spurious OTUs, but this can be avoided by applying stringent cut-offs. More stringent cutoffs are recommended especially when assigning conserved taxonomy to OTUs [16].

Frequently, users adopt the default settings of their preferred pipelines. For example, the 97% threshold typically expressed as 0.03 in Mothur and 70% confidence level expressed as 0.7 in QIIME-2 are customarily used in OTU clustering. The final output of most microbiome analysis pipelines is the OTU table. Basically, the OTU table is the prime data type for most downstream analyses in both Mothur and QIIME. This table is typically presented as a matrix of counts of sequences, OTUs or taxa on a per-sample basis. The quality of data in the OTU table depends primarily on the preceding analyses and their intermediate output that is normally used as input in the pipeline’s subsequent steps. Making biological conclusions from the OTU table alone without reviewing the intermediate output is a high risk that could result in inaccurate conclusions.

In this manuscript, we introduced a pipeline named iMAP (Integrative Microbiome Analysis Pipeline) that incorporates a review-as-you-go (RAYG) approach facilitated by contiguous progress reports. The robustness and reproducibility of the iMAP pipeline are features demonstrated using a case study where 360 mouse gut samples are intensively analyzed. The iMAP pipeline also extends the bioinformatics analysis to incorporate some preliminary analysis and annotation.

# METHODS

## Workflow

Code for implementing iMAP pipeline is designed with four bundles of commands wrapped individually in bash driver scripts for performing bioinformatic analysis and exploratory visualization of microbiome data (Figure 1). The output at each major step is transformed into a data structure suitable for conducting exploratory visualization of the output. A progress report that summarizes the intermediate results is generated at each major step.

## Implementation

A detailed guideline for implementing iMAP pipeline is provided (Additional file 1). Briefly, users start cloning the repository from GitHub and gather all required data files, reference databases, and software as described in the guideline. It is important that all requirements are placed in the designated folders and must remain unaltered throughout the entire analysis. A special script is included in the code to alert users when some files are either missing or found.

## Bioinformatics analysis

The iMAP pipeline is intended to be executed interactively from the command-line-interface (CLI) to optimize user interaction with generated output. However, users may want to run the code in batch mode or prepare a job scheduling script for running analyses remotely on the high-performance computing system. We recommend testing the pipeline using a small dataset included in the iMAP repository to see if the platform is set correctly. Running test analysis gives users hint of what is happening and interactively shows the expected output which enables users to familiarize with file names. If the test runs successful users must replace all data files including raw data, mapping files, and metadata before starting analyzing their dataset. Most of the analysis will run at default settings unless altered by the user. By default, the iMAP pipeline uses up-to-date SILVA seed classifiers [11] to assign taxonomy names to the representative sequences. Links to other suitable classifiers are provided in the guideline. A small modification in the sequence classification script must be made to reflect the different classifiers.

## Robustness and reproducibility

Ability to reproduce microbiome data analysis results is crucial. Challenges in robustness and reproducibility may be accelerated by lack of proper experimental design, the complexity of experiments, constant updates made to the available pipelines, lack of well-documented interactive workflows and relying on part-time developers who may leave anytime and as a result their code become inaccessible or out-of-date. We have made our iMAP pipeline robust and highly reproducible and demonstrated this using a detailed case study (next section). We extensively used R-markdown literate programming and public Git version control to document most of the analyses. We also shared the code for implementing the iMAP pipeline.

# APPLICATION

# Case study

In this case study, we used a dataset from previous microbiome study for demonstration purposes. Briefly, in 2012 Dr. Schloss (director of famous Mothur platform) and his colleagues at the University of Michigan published a paper in Gut Microbes journal entitled “Stabilization of the murine gut microbiome following weaning” [17]. In this study, 360 fecal samples were collected from 12 mice (6 female and 6 male) at 35 time points throughout the first year. Two mock community samples were added in the analysis for estimating the error rate. The mouse gut dataset was chosen because it has been successfully used in several studies for testing new protocols and workflows related to microbiome data analysis.

### Requirements

The demultiplexed paired-end 16S rRNA gene reads generated using Illumina's MiSeq platform were downloaded from http://www.mothur.org/MiSeqDevelopmentData/StabilityNoMetaG.tar. These reads were amplified from region four (V4) of the 16S rRNA gene. Sample metadata file describing the major features of the experiment and the associated variables was manually prepared. Mapping files that link paired-end sequences with the samples and design files that linked sample identifiers to individual experimental variables were manually prepared in a format compatible with Mothur (Additional file 2). Installation of software and download of required reference databases was done automatically. All required materials were placed in the designated folders exactly as described in the guideline and verified using a check file script.

## Metadata profiling

Metadata profiling was done to specifically explore the experimental variables to help in planning the downstream analysis and find out if there were any issues such as missing data. The sample identifiers were inspected and uniformly re-coded DPW numeric to three-digits (e.g. F3D1 to F3D001, F4D11 to F4D011, M4D145 to M4D145, and so on).

## Sequence pre-processing and quality control

Read pre-preprocessing included (i) general inspection using **seqkit** [1]software to provide basic descriptive information about the reads including data type, read depth and read length, (ii) assessing the base call quality using **fastqc** [2] software and (iii) trimming and filtering poor reads and removing any retained phiX control reads using **BBDuk** tool from **BBMap** [18] package. The quality of altered reads was again verified by re-running the fastqc software. The fastqc output was summarized using **multiqc** [19] software.

## Sequence processing and classification

Preprocessed paired-end reads were merged into longer sequences then screened to match the targeted V4 region of 16S rRNA gene. Representative sequences were generated and aligned to a SILVA-v132 rRNA reference alignments [11] to find the closest candidates. Post-alignment quality control involved repeating the screening and filtering the output by length and removal of poor alignments and chimeric sequences. All non-chimeric sequences were searched against SILVA-based classifiers at 80% identity using a k-nearest neighbor consensus and Wang approach exactly as described in the Mothur MiSeq SOP tutorial21. Additional quality control was done to remove any non-bacterial or unknown sequences before further analysis. The sequencing error rate was then estimated using sequences from the mock community. Finally, after error rate estimation all mock sequences where removed from further analysis.

## OTU clustering and conserved taxonomy assignment

We used a combination of phylotype, OTU-based and phylogeny methods to assign conserved taxonomy to OTUs. Briefly, in phylotype method, the sequences were binned into known phylotypes up to genus level while in OTU-based method all sequences were binned into clusters of OTUs based on their similarity at ≥97% identity. The phylogeny method was used to generate a tree that displayed consensus taxonomy for each node. The output from phylotype, OTU-based and phylogeny methods was manually reviewed, de-duplicated and integrated to form a complete OTU taxonomy output.

## Data transformation and preliminary analysis

The output was transformed into a data structure suitable for further analysis and visualization. Most analyses and visualization were conducted using R packages executed via RStudio IDE (integrated development environment). In summary, the preliminary analysis included measuring diversity in community membership using Jaccard dissimilarity coefficients based on the observed and estimated richness while diversity in community structure across groups was determined using Bray-Curtis dissimilarity coefficients. The Bray-Curtis dissimilarity coefficients were further analyzed using ordination methods to get a deeper insight into the sample-species relationships. Included in the ordination-based analysis were (i) PCA (Principal Component Analysis), (ii) Principal Coordinate Analysis (PCoA or MDS) and (iii) NMDS (Non-Metric Dimensional Scaling). Scree plot was used to find the best number of axes that explained variation seen on PCA plots while PCoA loadings and goodness function in vegan package [22] was used to generate vector of values for plotting observations into ordination space and the Shepard plot was used to compare observations from original dissimilarities, ordination distances and fitted values in NMDS.

**Phylogenetic annotation**

Phylogenetic annotation of the groups or samples required the dissimilarity distances be converted into a Newick-formatted tree in which the samples were clustered using the UPGMA algorithm. In this demonstration, we specifically chose a phylogenetic-based annotation approach as a model for displaying multidimensional data in easily interpretable ways. To do this we uploaded the Bray-Curtis-based Newick tree into the iTOL (Integrative Tree Of Life) tree viewer[23] and annotated it interactively with species richness, diversity and relative abundances at phylum-level. Additional manual editing was done to create publication-quality figures.

# RESULTS

## Metadata profiling

Metadata profiling results were automatically summarized into a web-based progress report 1 (Additional file 3). The main variables studied were sex (female and male), time range (early and late) grouped based on days-post-weaning (DPW) (Figure 2). Reviewing the report enabled us to understand the data type and discovered incorrectly assigned missing values. We noticed that numeric variables are confused with numeric data. For example, there are 12 (3.33%) missing values in DPW variable (Table 1). As investigators, we know that the first day of sampling was recorded as day 0. However, descriptive statistics interpret zeros as missing data. To correct this, we re-coded DPW as shown in DayID column in the metadata file (Additional file 2, Excel sheet 1).

## Read pre-processing and quality control

Pre-processing results were automatically summarized into a web-based progress report 2 (Additional file 4). The whole dataset contained 7,268,922 reads including 3,634,461 reads from the forward and 3,634,461 from reverse reads. The original FastQC results showed a minimum Phred score (Q) near 10 and trimming poor quality reads at the default settings (Q = 25) and removal of phiX contaminations resulted into high quality reads (Figure 3). Distribution of changes was visualized using boxplots, density plots and histogram plots (Figure 4). The difference between the original and pre-processed reads was very small, barely visible in the distribution plots. Only 2,692 (0.07%) poor-quality reads were identified in each forward and reverse reads (Table 2) indicating that over 99.9% of the reads qualified for downstream analysis.

## Sequence processing and quality control

The sequence processing and taxonomy assignment results were automatically summarized into a web-based progress report 3 (Additional file 5). This process involved merging 3,631,769 high-quality read pairs to form much longer sequences that were then screened based on their length.

Representative sequences (non-redundant) were then searched against SILVA rRNA reference alignments[11] to find the closest 16S rRNA gene candidates for downstream analysis. The alignment report showed that forward and reverse reads overlapped very well with most sequences having approximately 250 nucleotides (Figure 5). This sequence length is perfectly in-line with the targeted V4 region of the 16S rRNA gene. Most 16S rRNA gene candidates had over 90% identity. Post-alignment quality control which involved removing poor alignments and chimeric sequences yielded 2,934,726 clean sequences for downstream analysis.

## Sequence classification

All 2,934,726 non-chimeric sequences were searched against Mothur-formatted SILVA-bacterial classifiers at 80% identity using a k-nearest neighbor consensus and Wang approach as described[20, 21]. The error rate estimated after removing any remaining non-bacterial sequences was 0.00047 (0.047%). Removal of mock community finalizes sequence processing and quality control. Tabular and graphical representation showed a slight alteration of the number of processed sequences (Table 3, Figure 6).

## OTU clustering and taxonomy assignment

OTU and taxonomy results including preliminary analysis were automatically summarized into a web-based progress report 4 (Additional file 6).Clustering of 2,920,782 clean sequences into OTUs and assigning taxonomy names was done using a combination of phylotype, OTU-based and phylogeny methods. Taxonomy assignment in OTU-based method was optimized using opticlust algorithm [24]. This algorithm yielded high-quality results with high precision and low FDR ≤ 0.002 (Table 4).

## OTU abundance and preliminary analysis

The phylotype method yielded 197 OTUs at genus level while 11,257 OTU clusters were generated by the OTU-based method at 97% identity. The phylogeny method generated 58,929 tree nodes which were taxonomically classified at 97% identity. Manual review and comparison of the taxonomy results across the three classification methods revealed high redundancy rates that inflated the number of OTUs, particularly in the OTU-based and phylogeny methods. We used the online-based Venn diagram tools[25, 26] to eliminate duplicates and reveal the actual number and the intersection of unique lineages and taxon names across the three methods (Table 5, Figure 7).

## Alpha diversity analysis

### *Species accumulation*

The number of new species added as a function of sites sampling effort was determined using four different methods described in the vegan package [22] i.e. exact, random, collector and rarefaction (Figure 8). Rarefaction method had small standard error bars making it better fit for estimating species accumulation. Example of using rarefaction and extrapolation as described in the iNEXT [27] package is provided in the progress report 4 (Additional file 6).

### *Species richness and diversity*

Estimated and observed species richness were determined using Chao and Sobs calculators, respectively (Figure 9). Three diversity indices including inverse Simpson, Shannon, and phylo-diversity indices were used to account for the abundance and evenness of species present in the samples

## Beta diversity analysis

### *Clustering and ordination projections*

The difference in microbial community composition across the groups was measured using Bray-Curtis (dis)similarity coefficients. Clustering and ordination projection methods including partition around medoids (PAM) [28], principal component analysis (PCA), principal coordinate analysis (PCoA) and non-metric multidimensional scaling (NMDS) showed similar grouping (Additional file 6). Variability of the (dis)similarity coefficients across the samples was best explained by few components in all methods. Silhouette [29] graphical representation validated the consistency within PAM clusters (Figure 10). Scree plot and data loadings were used to show components that best explained the variation in PCA and PCoA, respectively. Shepard or stress plot confirmed linearity between the original and reduced dimensions in NMDS (see details in Additional file 6).

## Phylogenetic relationship of samples

Bray-Curtis-based Newick tree was uploaded into iTOL (Integrative Tree Of Life) viewer [23] to interactively display unrooted, circular and regular cladograms or phylograms (Figure 11). Annotation of circular and regular cladograms with various datasets enabled us to see the diversity across samples.

# DISCUSSION

High-quality biological knowledge from complex and multidimensional microbiome data is dictated by the choice of bioinformatics methodology used from pre-processing of the reads through the final step of the analysis. Most of the available bioinformatics tools are complex, contain multiple functions, and may require an in-depth understanding of their implementation. This can be time-consuming and may even require more advanced programming skills which some users may be lacking. We developed iMAP, a CLI-based pipeline that streamlines diverse functionalities from published tools to collectively unleash the hidden biological knowledge from marker-based microbiome data. The development of this pipeline is guided by the need for a tool that is easily executed by a novice user to investigate bacterial communities represented in diverse samples. The iMAP pipeline is integrated with custom functions that generate reports progressively to facilitate RAYG (review-as-you-go), a new approach associated with the pipeline to encourage investigators to review intermediate output and correct any obvious errors that may lead to wrong or misleading conclusions.

The iMAP pipeline supports a wide range of crucial analyses for profiling microbial communities present in an environmental sample. Currently, the pipeline operates on demultiplexed data generated by the Illumina platform and performs metadata profiling, tunable quality filtering, sequence processing, and classification before clustering the representative sequences into OTUs and conserved taxonomy assignment. The idea of implementing RAYG approach at every major step gives the investigators an opportunity of taking care of issues that could result in spurious OTUs and misleading conclusions.

The output generated from the major analysis steps described in the workflow (Figure 1) is further transformed to simplify exploratory data analysis. We know that some of the intermediate output can be very large to explore but, in such a situation, iMAP uses custom scripts to extract important information, then transform it for applying in diverse visualization modules. Reproducibility of the iMAP pipeline is ensured by publicly sharing the code used for bioinformatics analysis and some custom R-based scripts for generating publication-ready images commonly reported in microbiome-related manuscripts. Users may want to explore the bioinformatics analysis results using methods of their choice or may modify provided scripts to best describe the different types of data being analyzed. Examples of publication-ready images are presented in each progress report (Additional Files 3, 4, 5 & 6). Choice of which type of visualization to use is entirely user-dependent and could also depend on how much detail is required.

Post-classification annotation has been emphasized in this manuscript where the transformation of some quantitative data to qualitative data may be necessary. A good example is the OTU abundance data where large numbers of output in some classification methods were due to the high redundancy rate. Oftentimes, raw OTU abundance numbers are published unfiltered. While this is not surprising it can result in misleading conclusions. The fact that there could be some species that contribute disproportionately to the community, it would be sensible to probably stratify the analyses or performing intensive data annotation. In microbiome data analysis it is very common to see species that are exceptionally abundant being analyzed in the same way as those that are extremely rare. The situation is even worse if the high abundance values are influenced by having lots of redundant values. Adopting focused annotation is the key to achieving appropriate conclusions.

The preliminary analysis workflow included in the pipeline provides several methods for helping users in assessing diversity and statistical comparisons of the variables studied. Species accumulation and rarefaction, for example, provide the best way that allows investigators to figure out whether to continue sampling or whether the data isn’t enough for drawing a valid conclusion or for estimating a normalized sample size for statistical comparisons. Other methods such as heatmap, PAM clustering, and phylogenetic analysis are integrated into the pipeline to find out the relationship of the samples and groups while ordination projections using multivariate statistical techniques such as PCA, PCoA, and NMDS can be used to identify factors explaining differences among microbial communities. Also provided are statistical methods recommended in the Mothur platform for comparing the experimental variables. AMOVA, HOMOVA, and ANOSIM are among the methods that use P-values to determine if the observed differences are statistically significant or are by chance. The Metastats program [30] can be used for detecting differentially abundant microbial communities while the Kruskal–Wallis one-way ANOVA is commonly used to determine if there are statistically significant differences between two or more groups. The mothur-based lefse command modeled after the LEfSe program [31] is an excellent tool for biomarker discovery while the weighted and unweighted UniFrac [32] can be used to compare the samples using their phylogenetic information.

# CONCLUSIONS

The iMAP pipeline is a robust and reproducible bioinformatics tool for generating high-quality user-reviewed microbiome data analysis output. Investigators are provided with the RAYG approach to enable them to review the transitional output before further downstream analysis. Implementing the iMAP and RAYG approach simultaneously enables the investigators to discover systematic errors and generate more research questions that could lead to better conclusions. The multiple statistical methods included in the pipeline guide the investigators to make well-informed decisions backed by data and well-informed predictions in their research and also can help in generating data-driven hypotheses. No doubt that users will find this tool broadly useful and adaptable to their microbiome data analysis needs.

# LIST OF ABBREVIATIONS

AMOVA: Analysis of Molecular Variance

ANOSIM: Analysis of Similarity

CLI: Command-Line Interface

FDR: False Discovery Rate

GUI: Graphical User Interface

HOMOVA: Homogeneity of Variance

IDE: Integrated Development Environment

iMAP: integrative Microbiome Analysis Pipeline

MCC: Matthews Correlation Coefficient

NPV: Negative Predictive Value

OTU: Operational Taxonomic Unit

PAM: Partitioning Around Medoids

PCA: Principal Component Analysis

PCoA: Principal Coordinates Analysis

PPV: Positive Predictive Value

PSU-ICS: Penn State University-Institute for CyberScience

RAYG: Review-As-You-Go

# DECLARATIONS

**Acknowledgements**

Microbiome data analysis of the case study data was conducted using computational resources provided by the Huck Institutes of Life Sciences and the Institute for CyberScience Advanced Cyberinfrastructure (ICS-ACI) at the Pennsylvania State University, University Park, Pennsylvania, USA.

**Funding**

This work was funded by the Defense Threat Reduction Agency (DTRA) Cooperative Biological Engagement Program (CBEP) through Bushmeat Biosecurity Research project at the Pennsylvania State University. Project # OPP1083453 Please confirm.

**Availability of supporting data**

The data supporting the results reported in this manuscript is included within the article and its additional files. The auto generate progress reports are in HTML format and can be viewed using any preferred browser. The iMAP repository including the code and other requirements can be downloaded from https://github.com/tmbuza/iMAP.git. Guidelines for implementing iMAP are available at https://github.com/tmbuza/iMAP/blob/master/README.md. For further inquiries please contact ndelly@gmail.com.

**Disclaimer**

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Defense Threat Reduction Agency, Department of Defense, or the U.S. Government

**Authors’ contributions**

VK conceived the original idea. TMB developed the iMAP pipeline, performed the bioinformatic analysis of the case study data, and drafted the manuscript. TT, BL, RK, MS intensively tested the pipeline and provided feedback. FS, CT, PG, IC, JB and VK reviewed the manuscript and provided critical feedback and helped shape the second version of the manuscript. All authors read and approved the final manuscript.

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare that they have no competing interests.

**REFERENCES**

[1] W. Shen, S. Le, Y. Li, and F. Hu, “SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation,” *PLoS One*, vol. 11, no. 10, p. e0163962, 2016.

[2] S. Andrews, “FastQC: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.” 2018.

[3] B. Bushnell, “BBMap short-read aligner, and other bioinformatics tools.” 2016.

[4] P. D. Schloss *et al.*, “Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities,” *Appl Env. Microbiol*, vol. 75, no. 23, pp. 7537–7541, 2009.

[5] J. Kuczynski, J. Stombaugh, W. A. Walters, A. Gonzalez, J. G. Caporaso, and R. Knight, “Using QIIME to analyze 16S rRNA gene sequences from microbial communities,” *Curr Protoc Microbiol*, vol. Chapter 1, p. Unit 1E 5, 2012.

[6] B. J. Callahan, P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. Johnson, and S. P. Holmes, “DADA2: High-resolution sample inference from Illumina amplicon data,” *Nat Methods*, vol. 13, no. 7, pp. 581–583, 2016.

[7] K. Katoh and D. M. Standley, “MAFFT multiple sequence alignment software version 7: improvements in performance and usability,” *Mol Biol Evol*, vol. 30, no. 4, pp. 772–780, 2013.

[8] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, “UCHIME improves sensitivity and speed of chimera detection,” *Bioinformatics*, vol. 27, no. 16, pp. 2194–2200, 2011.

[9] T. Rognes, T. Flouri, B. Nichols, C. Quince, and F. Mahe, “VSEARCH: a versatile open source tool for metagenomics,” *PeerJ*, vol. 4, p. e2584, 2016.

[10] J. R. Cole *et al.*, “The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis,” *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D294--6, 2005.

[11] P. Yilmaz *et al.*, “The SILVA and ‘All-species Living Tree Project (LTP)’ taxonomic frameworks,” *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D643--8, 2014.

[12] T. Z. DeSantis *et al.*, “Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB,” *Appl Env. Microbiol*, vol. 72, no. 7, pp. 5069–5072, 2006.

[13] S. H. Yoon *et al.*, “Introducing EzBioCloud: A taxonomically united database of 16S rRNA gene sequences and whole-genome assemblies,” *Int. J. Syst. Evol. Microbiol.*, 2017.

[14] N. A. Bokulich *et al.*, “Optimizing taxonomic classification of marker-gene amplicon sequences with QIIME 2’s q2-feature-classifier plugin,” *Microbiome*, 2018.

[15] N. Bokulich, M. Dillon, E. Bolyen, B. D. Kaehler, G. A. Huttley, and J. G. Caporaso, “q2-sample-classifier: machine-learning tools for microbiome classification and regression,” *bioRxiv*, 2018.

[16] R. C. Edgar, “Updating the 97% identity threshold for 16S ribosomal RNA OTUs,” *Bioinformatics*, 2018.

[17] P. D. Schloss, A. M. Schubert, J. P. Zackular, K. D. Iverson, V. B. Young, and J. F. Petrosino, “Stabilization of the murine gut microbiome following weaning,” *Gut Microbes*, vol. 3, no. 4, pp. 383–393, 2012.

[18] N. Segata *et al.*, “BBMap short-read aligner, and other bioinformatics tools,” *Bioinformatics*, vol. 12, no. 1, pp. 17–23, 2016.

[19] P. Ewels, M. Magnusson, S. Lundin, and M. Kaller, “MultiQC: summarize analysis results for multiple tools and samples in a single report,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, 2016.

[20] J. J. Kozich, S. L. Westcott, N. T. Baxter, S. K. Highlander, and P. D. Schloss, “Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform,” *Appl Env. Microbiol*, vol. 79, no. 17, pp. 5112–5120, 2013.

[21] P. D. Schloss, “MiSeq SOP,” 2018. [Online]. Available: https://www.mothur.org/wiki/MiSeq\_SOP.

[22] J. Oksanen *et al.*, “Vegan: Community Ecology Package,” *R Packag.*, no. 2.5-2, 2018.

[23] I. Letunic and P. Bork, “Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees,” *Nucleic Acids Res*, vol. 44, no. W1, pp. W242--5, 2016.

[24] S. L. Westcott and P. D. Schloss, “OptiClust, an Improved Method for Assigning Amplicon-Based Sequence Data to Operational Taxonomic Units,” *mSphere*, 2017.

[25] P. Bardou, J. Mariette, F. Escudié, C. Djemiel, and C. Klopp, “Jvenn: An interactive Venn diagram viewer,” *BMC Bioinformatics*, vol. 15, no. 1, 2014.

[26] V. de Peer, “Calculate and draw custom Venn diagrams,” *Online web-based tool*, 2018. [Online]. Available: http://bioinformatics.psb.ugent.be/webtools/Venn/.

[27] T. C. Hsieh, K. H. Ma, and A. Chao, “iNEXT: An R package for interpolation and extrapolation in measuring species diversity,” *Methods Ecol. Evol.*, 2016.

[28] H. Park, J. Lee, C. Jun, M. Engineering, and S. Korea, “A K-means-like Algorithm for K-medoids Clustering and Its Performance,” *System*, 2008.

[29] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, 1987.

[30] J. R. White, N. Nagarajan, and M. Pop, “Statistical methods for detecting differentially abundant features in clinical metagenomic samples,” *PLoS Comput Biol*, vol. 5, no. 4, p. e1000352, 2009.

[31] N. Segata *et al.*, “Metagenomic biomarker discovery and explanation,” *Genome Biol*, vol. 12, no. 6, p. R60, 2011.

[32] C. Lozupone, M. E. Lladser, D. Knights, J. Stombaugh, and R. Knight, “UniFrac: An effective distance metric for microbial community comparison,” *ISME Journal*. 2011.

# TABLES

**Table 1: Descriptive statistics of the case study metadata**

Key: **q\_zeros**: quantity of zeros or missing data; **q\_na**: quantity of NA; **q\_inf**: quantity of infinite values; **type**: factor, character, integer or numeric; **unique**: quantity of unique values. The **p\_zeros, p\_na or p\_inf** are the percentages of the corresponding q values.

**Table 2: Descriptive statistics of the pre-processed reads and total count from all samples**

**Table 3: Descriptive statistics of the processed sequences and total count from all samples**

**Table 4: Statistical parameters optimized by opticlust method**

**Table 5: OTU abundance observed in three methods**

# FIGURES

**Figure 1: Schematic illustration of the iMAP pipeline**. The pipeline starts by gathering and verifying required materials including data files, software and reference databases. Then starts profiling sample metadata followed by pre-processing and quality checking of demultiplexed 16S read pairs which are then merged, aligned to reference alignments, classified then assigned conserved taxonomy names. Output from each major step is transformed, visualized and summarized into a progress report. At the end of analysis all contiguous reports are combined to form a single reproducible document that summarizes the entire analysis.

**Figure 2: Frequency of experimental variables.** Numericdays-post-weaning (DPW) were converted to character data type indicated by DayID, for example 0 DPW to D000, 1 DPW to D001 and so forth. All sample IDs maintained three digits for uniformity in axis labeling and sorting purposes.

**Figure 3: Summary of FastQC quality scores of paired-end reads from 360 samples**. Number of reads with average quality scores before (A) and after trimming at Q25 and removal of phiX contamination (B) and the mean quality values across each base position in the read (C).

**Figure 4: Distribution of pre-processed reads.** This displays the forward reads split by experimental variables (sex and time) and grouped by QC variables (see legend). The jitter boxplot (A) shows the distribution of individual observations. Mean values for the density (B) and histograms (C) are indicated by dotted lines and marginal rugs are shown at the bottom.

**Figure 6: Distribution of processed sequences.** The barplots (A) show the overall distribution without much details, the boxplots (B) clearly added more clarity showing the midpoint and outliers. The stacked density plots (C) and the histograms (D) clearly show the skewness of the sequence depth. Histograms separated the differences better than the other plots. Mean values of the density plots and histograms are indicated by dotted lines and marginal rugs are shown at the bottom. A slight shift of the mean line to the left is probably due to removal of number of sequences at denoising step.

**Figure 7: Venn diagram and taxon term representation.** Visual representation of taxon terms highlighted the most abundant taxon based on frequency of being assigned to an OTU or tree nodes. *Muribaculaceae* was the most frequently assigned family and *Muribaculaceae\_ge* was the dominating genus assigned to most sequences.

**Figure 8: Species accumulation curves.** Four methods were compared including exact (magenta), random (green), rarefaction (orange) and collector (blue). The standard deviation (except in collector curves) is indicated by the vertical lines which are highly condensed due to large dataset (360 samples).

**Figure 9: Species richness and diversity.** The observed species richness (A) is positively correlated to inverse Simpson diversity index (B), Shannon diversity index (C) and phylo-divesity index (D).

**Figure 10: Silhouette scores (Si) for a subset of 50 samples (out of 360).** Observations with a large Si (almost 1) indicates high quality clustering, a small Si (around 0) means that the observation lies between two clusters while negative value are outliers.

**Figure 11: Phylogenetic relationship and annotation of samples grouped by sex variable**. The normal (A), circular (B) and unrooted (C) cladograms displays the relationships of the 360 samples. Female (magenta) and male (light blue) are annotated with multiple dataset including number of classified and unclassified sequences (pie chart), original (blue) and final (orange) sequence depth, observed (green) and estimated (maroon) species richness and the scaled phyla abundance (heatmap). A portion of the tree (D) is enlarged to show some details.