iMAP: An integrative pipeline for classifying 16S rRNA gene sequences into known bacterial taxa

T. M. Buza[1], R. Katani[1], F. Stomeo[2], T. Tonui[2], J. Buza[3], I. Albert[1], V. Kapur[1]

[1]Pennsylvania State University, University Park, PA, U.S.A, [2]Biosciences Eastern and Central Africa-International Livestock Research Institute (BecA-ILRI) Hub, Nairobi, KENYA, [3]Nelson Mandela African Institute of Science and Technology, Arusha, TANZANIA

Background

Microbiome studies present great opportunities for understanding microbial communities in diverse samples and their role in environment, health and disease. Effective analytical workflows that guarantee reproducibility, repeatability and result provenance are essential requirements for analyzing highly multidimensional microbiome data. For nearly a decade, several state-of-the-art bioinformatics tools have been developed to enable investigators gain knowledge and biological value of the microbial communities living in a given sample.

Statement of problem

Most of the available bioinformatics tools are designed with multiple functions, that may require some programming skills to understand their full implementation. This is why most investigators hire bioinformaticians to help with data analysis. Microbiome data analysis is a complex process that requires thorough review of the output. However, intermediate output is less frequently reviewed and can lead to errors being propagated downstream.

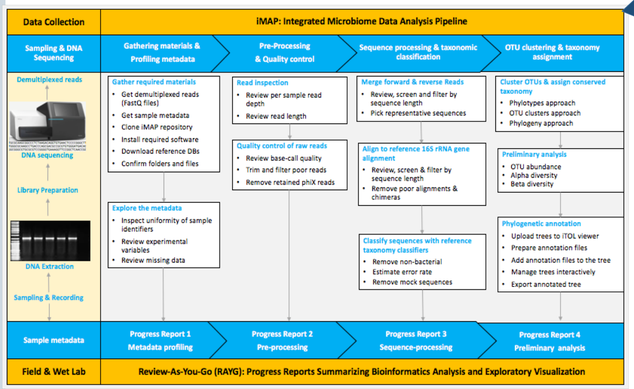
Solution

We have developed a pipeline named iMAP (integrative Microbiome Analysis Pipeline) to address some of microbiome data analysis challenges including result visualization, interpretation and reporting. 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 using 16S rRNA gene markers.

iMAP Workflow

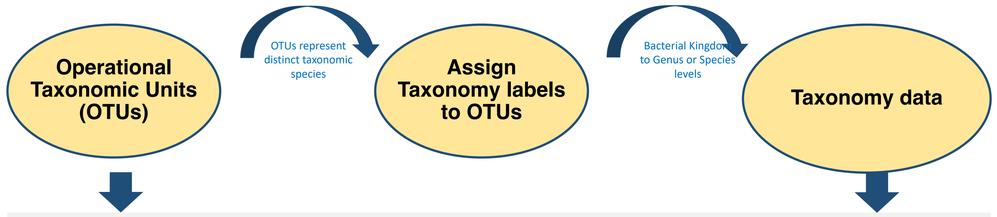
The iMAP pipeline operates on demultiplexed 16S rRNA gene reads generated using Illumina platform and performs tunable quality filtering before assigning known taxonomic nomenclatures to the representative sequences. The pipeline also generates progress reports to facilitate progressive review of the output.

Bioinformatics Analysis



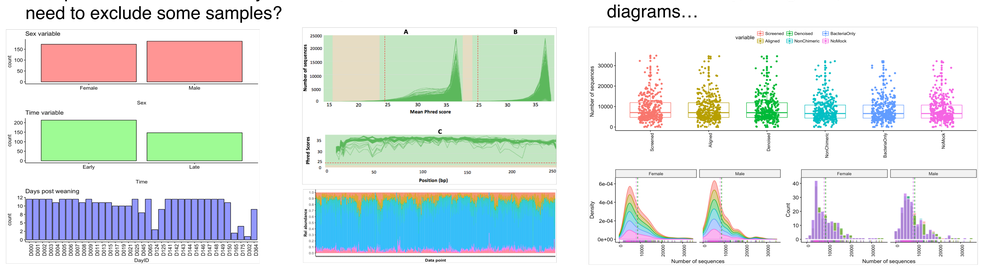
Results

The output are data tables which are then transformed into structures suitable for statistical analysis and for generating publication-quality images. Progress reports generated at each major step can be reviewed by the investigators.

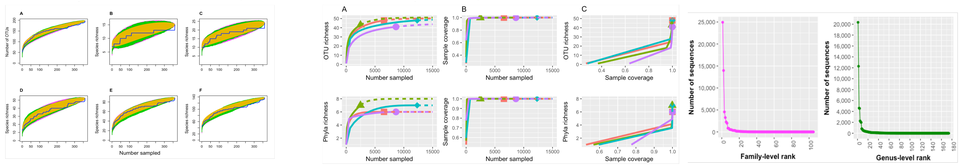


Analysis & visualization

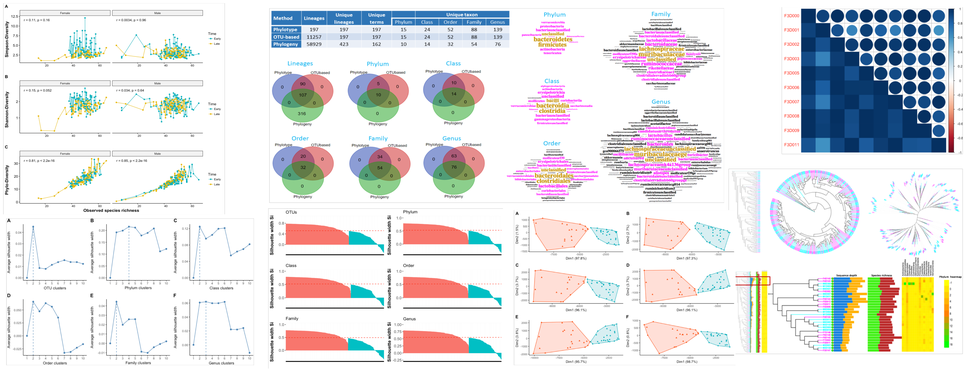
- **Metadata profiling**: Frequency of experimental variables. Any need to exclude some samples?
- **Read quality**: Default threshold is q=25.
- **General distribution:** Histograms, barplots, boxplots, density, histograms, venn diagrams…



- **Alpha diversity analysis:** Are there additional OTUs/species discovered through additional sampling? Common curves: species accumulation, rarefaction and extrapolation, rank abundance.



- **Beta diversity analysis:** Is there any similarity of data (microbial composition) between samples? Highly multidimensional OTU & taxonomic data is reduced to fewer dimensions by calculating distance matrices to specifically measure the diversity between samples. Example of visualization images: Heuristic clustering and ordination are commonly used: Scree and stress plot for pre-ordination, PCA, PAM, PCoA, NMDS for ordination, heatmaps for pairwise comparison, phylograms, cladograms…



Reporting

- Finally, the results are summarized into HTML report using Rmarkdown for easy visualization, interpretation and sharing.
- In addition to general microbiome analysis, any hypotheses tested during bioinformatics analysis will appear in the report.
- Example: In our two training datasets (i) bushmeat and (ii) dumpsite samples we profiled 13 dangerous pathogens in the category of **federal select agents.**

Abundance of select agent genera in 56 bushmeat samples

Abundance of select agent genera in 50 dumpsite samples

Conclusion

The iMAP pipeline accelerates the identification of microbial communities that are present in environmental samples. The rich visuals produced by the pipeline facilitate better understanding of the multidimensional microbiome data and the reproducible HTML reports facilitate easy visualization, interpretation and quick sharing with collaborators. We anticipate that users will find this pipeline broadly useful and adaptable to their needs.

Acknowledgement

- This work is funded by DTRA-CBEP.
- We thank Dr Kilaza S Mwaikono of Dar es Salaam Institute of Technology, Tanzania, for providing dumpsite dataset for testing the pipeline.

Disclaimer

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

References

1. Schloss PD, *et al*: **Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities**. *Appl Environ Microbiol* 2009, **75**(23):7537-7541.

2. Wood DE, Salzberg SL: **Kraken: ultrafast metagenomic sequence classification using exact alignments**. *Genome Biol* 2014, **15**(3):R46.

3. iMAP pipeline manuscript: In preparation