Enabling Reproducible Microbiome Science Through Decentralized Provenance Tracking in QIIME 2

Chris Keefe
Pathogen and Microbiome Institute
Northern Arizona University
Flagstaff, AZ
crk239@nau.edu

Ahmad Turan Naimey
Pathogen and Microbiome Institute
Northern Arizona University
Flagstaff, AZ
atn35@nau.edu

ABSTRACT

We demonstrate the ways in which automatic, integrated, decentralized provenance tracking in QIIME 2, a leading microbiome bioinformatics platform, enables reproducible microbiome science. We use sample data from a recent study of arid soil microbiomes (Significant Impacts of Increasing Aridity on the Arid Soil Microbiome; Neilson et al, 2017[2]), to illustrate specific analyses that QIIME 2 supports, and to frame our discussion of the QIIME 2 platform.

QIIME 2 actions yield as outputs Artifacts integrating the requested data or visualization with comprehensive data provenance that describes the computational history of that data or visualization, including all methods and parameters involved in its creation. This approach gives users, reviewers, and readers powerful tools for understanding, reproducing, and extending studies.

The benefits this approach provides to both the researcher and the scientific community are significant, and provide a useful model for research software developers across disciplines.

CCS CONCEPTS

• Bioinformatics • Data provenance • Computational biology

KEYWORDS

QIIME 2, Data Analysis, Bioinformatics, Reproducible Science

OBJECTIVE

To demonstrate the ways in which automatic, integrated, decentralized provenance tracking in QIIME 2 [1] enables reproducible microbiome science, we will be using a sample analysis as the framework for our discussion.

QIIME 2 Key Features:

- Integrated, automatic, and decentralized tracking of data provenance
- Semantic type system
- Plugin system for extending microbiome analysis function
- Support for multiple user interface types

Background

The analysis which frames this poster is based on work by Neilson et al in *Significant Impacts of Increasing Aridity on the Arid Soil Microbiome* [2]. This study documents correlations between aridity and changes in the soil microbiome along two arid-to-hyperarid transects in the Atacama Desert, Chile. For new QIIME 2 users interested in understanding our approach, a tutorial based on this study is available in the QIIME 2 documentation.

Creating a QIIME 2 artifact (initiating provenance tracking)

In order to work in QIIME 2, we must first import our data to create a new QIIME 2 artifact. A QIIME 2 Artifact is an immutable collection of data and its associated metadata, including the *type*, *format*, and *provenance* (Poster Figure 3). Using Artifacts rather than plain files helps QIIME 2 ensure actions performed are meaningful. We then demultiplex our data, passing the new artifact into `q2-demux`, along with a column of sample metadata containing barcode sequences. Each QIIME 2 action generates new artifacts whose provenance metadata includes a comprehensive history of the data in the artifact, including:

- the processes to which the data was subjected
- all parameters chosen when running these processes
- citation information relevant to the methods chosen
- version information for all software involved (QIIME 2 and all relevant plugins)

QIIME 2 analyses are host-agnostic

De-noising the data will be computationally intensive, so we use a high-performance cluster (Poster Figure 1). The integrated and decentralized handling of provenance metadata in ar facts makes it easy for QIIME 2 users to transfer files and run jobs through any QIIME 2 interface on any host, accumulating and retaining accurate provenance information.

QIIME 2 analyses are interface-agnostic

Automatic provenance tracking saves time and alleviates uncertainty.

More-experienced users may run downstream analysis using the Artifact API in a Jupyter Notebook (Poster Figure 2). Alternatively, analysis may proceed in QIIME 2 Studio (q2studio), a GUI that provides asynchronous process handling and access to all QIIME 2 plugins in our environment (Poster Figure 4).

Throughout this complex analysis, QIIME 2 records our methods. The provenance data in our artifacts and visualizations minimizes uncertainty about which file is correct.

Sharing QIIME 2 visualizations

QIIME 2 visualizations can be shared and viewed without a software install, and contain complete provenance information

1. Interactive visualizations are significant in understanding and communicating our data. interactive visualizations from the analysis (Poster Figure 5)
2. QIIME 2's provenance graphs allow us to review which methods and parameters were used, simplifying creation of an accurate methods section and supplementary materials for academic publications.
3. We can generate a citation list for any result in this poster with `qiime tools citations`.
4. Collaborators and reviewers of our work use https://view.qiime2.org to confirm methods and outcomes are appropriate, without the need to download any special so ware.

By packaging data with provenance, QIIME 2 reduces the risk that the methods used to produce a given outcome are accidentally misreported. QIIME 2 has integrated citation information into plugins, allowing researchers to export citations as BibTeX.

CONCLUSIONS

Decentralized provenance tracking enables study reproduction, replication & extension. QIIME 2's provenance metadata makes it easy for anyone to reproduce our computational analysis independently to confirm the quality of our results. Should we choose to expand our study, or conduct further research using similar methods, provenance graphs provide a "road map".

FUTURE WORK

- Allow QIIME 2 to consume provenance metadata: Reading provenance into QIIME 2 could allow for computer-assisted introspection into research methods, automated transcription of methods for publication, and auto-generation of "pipeline" scripts for common workflows.
- QIIME 2 Studio and HPC: In the future, we hope that q2studio might interact directly with HPC schedulers, to simplify the process of connecting with and running computationally-intensive processes remotely.

ACKNOWLEDGMENTS

Sincere thanks to Matthew Dillon, Evan Bolyen, & J. Greg Caporaso for their help and guidance.

REFERENCES

[1] Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J., and Knight, R.
Qiime allows analysis of high-throughput community sequencing data.
Nature methods 7, 5 (2010), 335.

[2] Neilson, J.W., Califf, K., Cardona, C., Copeland, A., vanTreuren, W., Josephson, K.L., Knight, R., Gilbert, J. A., Quade, J., Caporaso, J. G., and Maier, R. M.
Significant impacts of increasing aridity on the arid soil microbiome.
mSystems 2, 3 (2017).

[3] Pérez, F., and Granger, B.E. IPython: a system for interactive scientific computing. *Computing in Science and Engineering 9*, 3 (May 2007), 21--29.