Enabling Reproducible Microbiome Science through Decentralized Provenance Tracking in QIIME 2

Christopher R. Keefe * Ahmad Turan Naimey * Evan Bolyen ‡ Matthew Ryan Dillon ‡ J. Gregory Caporaso ‡

The Pathogen and Microbiome Institute at Northern Arizona University

* Equal Contributors ‡ Advisors

Objective and Introduction

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

QIIME 2 Key Features:

- Integrated and automatic tracking of data provenance
- Semantic type system
- Plugin system for extending microbiome analysis functionality
- Support for multiple types of user interfaces (e.g. API, command line, graphical)

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. We then demultiplex our data, passing the new artifact into `q2-demux`, alongside a column of sample metadata containing barcode sequences.

A QIIME 2 `Artifact` is an immutable collection of data and its associated metadata, including the *type*, *format*, and *provenance* (Figure 3). Using Artifacts rather than plain files helps QIIME 2 ensure actions performed are meaningful. Each 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. (Figure 1)

```
#!/bin/bash
#SBATCH --job-name="atacama-dada2"
#SBATCH --mem=320000
#SBATCH --cpus-per-task=16
#SBATCH --time=72:00:00

module load qiime2/2018.6

qiime dada2 denoise-single \
  --i-demultiplexed-seqs demux.qza \
  --p-trunc-len 0 \
  --o-table dada2-single-table.qza \
  --o-representative-sequences dada2-single-rep-seqs.qza \
  --o-denoising-stats dada2-single-denoising-stats.qza

qiime feature-table summarize
  --i-table dada2-single-results/dada2-single-table.qza  \
  --o-visualization dada2-single-results/dada2-single-table.qzv \
  --m-sample-metadata-file sample-metadata.tsv

qiime feature-table tabulate-seqs \
  --i-data dada2-single-results/dada2-single-rep-seqs.qza  \
  --o-visualization dada2-single-results/dada2-single-rep-seqs.qzv
```

Figure 1: Script used for denoising demultiplexed sequences and summarizing the results using NAU's Monsoon cluster

The integrated and decentralized handling of provenance metadata in artifacts 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 accurate provenance information.

QIIME 2 analyses are interface-agnostic

Automatic provenance tracking saves time and alleviates uncertainty. Experienced users may run downstream analysis using the Artifact API in a Jupyter Notebook [3] (Figure 2) Alternatively, analysis may proceed in QIIME 2 Studio, a GUI that provides asynchronous process handling and access to all QIIME 2 plugins in our environment (Figure 5)

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

```
In [15]: # Load a pre-trained classifier...
         greengenes_classifier = qiime2.Artifact.load(ref_dir / 'gg-13-8-99-nb-classifier.qza')

         #... and classify our sequences.
         taxonomy, = feature_classifier.classify_sklearn(sequences, greengenes_classifier)
         taxonomy.save(str(data_dir / 'taxonomy.qza'))

Out[15]: 'data-single/taxonomy.qza'

In [24]: # Compute Biplot
         relative_frequency_tbl, = feature_table.relative_frequency(filtered_table)
         pcoa_biplot, = diversity.pcoa_biplot(core_metric_results.unweighted_unifrac_pcoa_results, relative_frequency_tbl)
         taxonomy_as_md = taxonomy.view(qiime2.Metadata)
         emperor_biplot, = emperor.biplot(pcoa_biplot, sample_metadata=sample_metadata, feature_metadata=taxonomy_as_md)
         emperor_biplot.save(str(viz_dir / 'uw_unifrac_biplot.qzv'))

Out[24]: 'data-single/visualizations/uw_unifrac_biplot.qzv'
```

Figure 2: Selection from the Jupyter Notebook used for downstream analysis conducted with the Artifact API

Our Analysis and the Atacama Study

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 users interested in understanding our approach, a tutorial based on this study is available in the QIIME 2 docs.

Visualizations

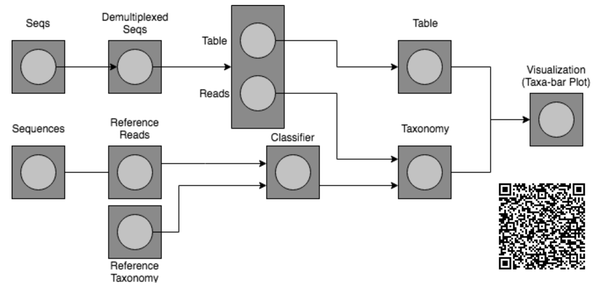

Figure 3: Provenance tree describing the production of our taxonomic bar plot visualization. QR code provides access to both provenance tree and bar plot (Figure 4).

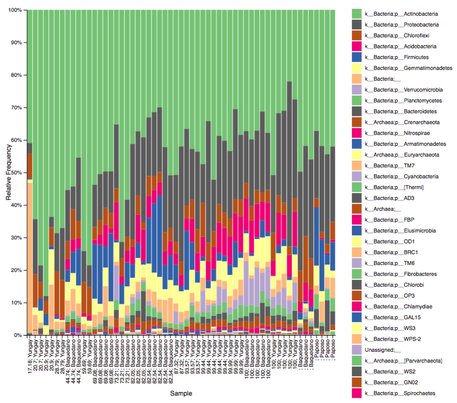

Figure 4: Taxonomic bar plot sorted by Average Soil Relative Humidity

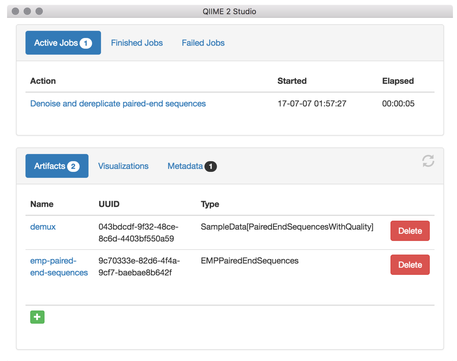

Figure 5: QIIME 2 Studio provides GUI access to all QIIME 2 plugins in the computing environment

Project Funding

ARIZONA BOARD OF Regents

ALFRED P. SLOAN FOUNDATION 1934

NATIONAL SCIENCE FOUNDATION

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. (Figure 4)
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 can use https://view.qiime2.org to confirm methods and outcomes are appropriate, without the need to download any special software. (Figure 6)

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.

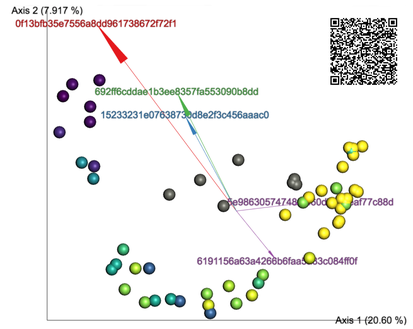

Figure 6: Color gradient represents Average Soil Relative Humidity. Vector colors represent independent clades. The magnitude and direction of a vector represents the relative contribution of a given amplicon sequence variant (ASV) to the distance between samples (points). Red: Unknown (K)Bacteria ASV, Green: (P)Actinobacteria (C)Actinobacteria, Blue: (P)Actinobacteria (C)Acidimicrobiia, Purple: (P)Proteobacteria (C)Gammaproteobacteria

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 one 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 QIIME 2 Studio might interact directly with HPC schedulers, to simplify the process of connecting with and running computationally-intensive processes remotely.

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., van Treuren, 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.

Poster Source: https://github.com/caporaso-lab/sc18-poster