QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science

Evan Bolyen\(^1\),*\, Jai Ram Rideout\(^1\,\ast\), Matthew R Dillon\(^1\,\ast\), Nicholas A Bokulich\(^1\,\ast\), Christian Abnet\(^2\), Gabriel A Al-Ghalith\(^3\), Harriet Alexander\(^4\),\(^5\), Eric J Alm\(^6\),\(^7\), Manimozhiyan Arumugam\(^8\), Francesco Asnicar\(^9\), Yang Bai\(^1\,\(^10\),\(^11\),\(^12\), Jordan E Bisanz\(^13\), Kyle Bittinger\(^1\,\(^14\),\(^15\), Asker Brejnrod\(^1\), Colin J Brislawn\(^1\), C Titus Brown\(^9\), Benjamin J Callahan\(^1\,\(^16\),\(^19\), Andrés Mauricio Caraballo-Rodriguez\(^20\), John Chase\(^1\), Emily Cope\(^1\),\(^21\), Ricardo Da Silva\(^20\), Pieter C Dorrestein\(^20\), Gavin M Douglas\(^22\), Daniel M Durall\(^23\), Claire Duvall\(^6\), Christian F Edwardson\(^24\), Madeleine Ernst\(^20\), Mehrbod Estaki\(^25\), Jennifer Fouquier\(^26\),\(^27\), Julia M Gauglitz\(^20\), Deanna L Gibson\(^28\),\(^29\), Antonio Gonzalez\(^30\), Kestrel Gorlick\(^1\), Jiarong Guo\(^31\), Benjamin H Hillmann\(^32\), Susan Holmes\(^33\), Hannes Holste\(^30\),\(^34\), Curtis Huttenhower\(^35\),\(^36\), Gavin A Hutley\(^37\), Stefan Janssen\(^38\), Alan K Jarmusch\(^20\), Lingjing Jiang\(^39\), Benjamin D Kaehler\(^37\), Kyo Bin Kang\(^40\),\(^20\), Christopher R Keefe\(^1\), Paul Keim\(^1\), Scott T Kelley\(^41\), Dan Knights\(^42\),\(^32\), Irina Koester\(^43\),\(^20\), Tomasz Kosciolek\(^44\), Jorden Kreps\(^1\), Morgan GI Langille\(^45\), Joslynn Lee\(^46\), Ruth Ley\(^47\),\(^48\), Yong-Xin Liu\(^10\),\(^11\), Erika Loftfield\(^2\), Catherine Lozupone\(^49\), Massoud Maher\(^50\), Clarisse Marotz\(^30\), Bryan Martin\(^51\), Daniel McDonald\(^30\), Lauren J McIver\(^35\),\(^36\), Alexey V Melnik\(^20\), Jessica L Metcalf\(^52\), Sydney C Morgan\(^53\), Jamie T Morton\(^30\),\(^50\), Ahmad Turan Naimey\(^1\), Jose A Navas-Molina\(^50\),\(^30\),\(^54\), Louis Felix Nothias\(^20\), Stephanie B Orchanian\(^55\), Talima Pearson\(^1\), Samuel L Peoples\(^56\),\(^57\), Daniel Petras\(^20\), Mary Lai Preuss\(^58\), Elmar Pruesse\(^49\), Lasse Buur Rasmussen\(^16\), Adam Rivers\(^59\), Michael S Robeson, II\(^60\), Patrick Rosenthal\(^58\), Nicola Segata\(^9\), Michael Shaffer\(^40\),\(^61\), Arron Shiffer\(^1\), Rashmi Sinha\(^2\), Se Jin Song\(^30\), John R Spear\(^42\), Austin D Swafford\(^55\), Luke R Thompson\(^63\),\(^64\), Pedro J Torres\(^65\), Pauline Trinh\(^66\), Anupriya Tripathi\(^20\),\(^30\),\(^67\), Peter J Turnbaugh\(^58\), Sabah Ul-Hasan\(^69\), Justin JJ van der Hooft\(^70\), Fernando Vargas\(^57\), Yoshiki Vázquez-Baeza\(^30\), Emily Vogtmann\(^2\), Max von Hippel\(^71\), William Walters\(^47\), Yunhu Wan\(^2\), Mingxun Wang\(^20\), Jonathan Warren\(^72\), Kyle C Weber\(^59\),\(^73\), Chase HD Williamson\(^1\), Amy D Willis\(^74\), Zhenjiang Zech Xu\(^30\), Jesse R Zaneveld\(^75\), Yilong Zhang\(^76\), Rob Knight\(^30\),\(^77\),\(^55\), and J Gregory Caporaso\(^1\),\(^21\),\(^21\),+

* These authors contributed equally to this work.
+ Please address correspondence about this document to gregcaporaso@gmail.com.

Author affiliations are provided following the Main Text References.

To get help with QIIME 2, visit https://forum.qiime2.org.
Abstract

We present QIIME 2, an open-source microbiome data science platform accessible to users spanning the microbiome research ecosystem, from scientists and engineers to clinicians and policy makers. QIIME 2 provides new features that will drive the next generation of microbiome research. These include interactive spatial and temporal analysis and visualization tools, support for metabolomics and shotgun metagenomics analysis, and automated data provenance tracking to ensure reproducible, transparent microbiome data science.

Main text

Rapid advances in DNA sequencing and bioinformatics technologies in the past two decades have significantly improved our understanding of the microbial world. These include our growing understanding of the vast diversity of microorganisms; how our microbiota and microbiomes impact disease\(^1\) and medical treatment\(^2\); how microorganisms impact the health of our planet\(^3\); and our nascent exploration of the medical\(^4\), forensic\(^5\), environmental\(^6\), and agricultural\(^7\) applications of microbiome biotechnology. Much of this work has been driven by marker gene surveys (e.g., bacterial/archaeal 16S rRNA genes, fungal ITS, eukaryal 18S rRNA genes), which profile microbiota with varying degrees of taxonomic specificity and phylogenetic information. The field is now transitioning to integrate other data types, such as metabolite\(^8\) or metatranscriptome\(^9\) profiles.

The QIIME 1 microbiome bioinformatics platform has supported many microbiome studies and gained a broad user and developer community. Interactions with QIIME 1 users in our online support forum, our workshops, and direct collaborations showed the potential to better serve an increasingly diverse array of microbiome researchers in academia, government, and industry. Here we present QIIME 2, a completely reengineered and rewritten system that will facilitate reproducible and modular analysis of microbiome data to enable the next generation of microbiome science.

QIIME 2 is developed based on a plugin architecture (Figure S1) that allows third-parties to contribute functionality (see https://library.qiime2.org). QIIME 2 plugins exist for latest-generation tools for sequence quality control from different sequencing platforms (DADA2\(^10\) and Deblur\(^11\)), taxonomy assignment\(^12\), and phylogenetic insertion\(^13\), that quantitatively improve results over QIIME 1 and other tools (detailed in the corresponding tool-specific publications). Plugins also support qualitatively new functionality including microbiome paired-sample and time-series analysis\(^14\), critical for studying the impact of treatment on the microbiome, and for machine learning\(^15\), including the ability to save trained models and apply them to new data and to interrogate models to identify important microbiome features. Several recently released plugins, including q2-cscs\(^16\), q2-metabolomics\(^17\), q2-shogun\(^18\), q2-metaphlan2\(^19\), and q2-picrust2\(^20\), provide initial support for analysis of metabolomics and shotgun metagenomics data. This marks
the potential of QIIME 2 to serve not only as a marker gene analysis tool, but also a multi-dimensional and powerful data science platform that can be rapidly adapted to analyze diverse microbiome features.

QIIME 2 provides many new interactive visualization tools (Figure 1), facilitating exploratory analyses and result reporting. QIIME 2 View (https://view.qiime2.org) is a unique new service (see Online Methods) that allows users to securely share and interact with results without installing QIIME 2. The QIIME 2 visualizations presented in Figure 1 are provided in Supplementary File 1 for readers to interact with using QIIME 2 View.

Reproducibility, transparency, and clarity of microbiome data science are guiding principles in the QIIME 2 design. Toward this end, it includes a decentralized data provenance tracking system: details of all analysis steps with references to intermediate data are automatically stored in the results. Users can thus retrospectively determine exactly how any result was generated (Figure 2). QIIME 2 also detects corrupted results, indicating that provenance is no longer reliable and the results no longer contain information enabling reproducibility. Provenance of the visualizations presented in Figure 1 can be interactively reviewed by loading the contents of Supplementary File 1 with QIIME 2 View, providing far more detailed information than can typically be provided in Methods text. QIIME 2 results are additionally semantically typed (Figure 2) and actions indicate acceptable input types, clarifying the data that actions should be applied to and making complex workflows less error-prone.

Finally, QIIME 2 provides a software development kit (see https://dev.qiime2.org) that can be used to integrate it as a component of other systems (e.g., such as Qiita or Illumina BaseSpace) and to develop interfaces targeted toward users with different levels of computational sophistication (Figure S2). QIIME 2 provides the QIIME 2 Studio graphical user interface and QIIME 2 View, interfaces designed for end-user biologists, clinicians, and policy makers; the QIIME 2 application programming interface, designed for data scientists who want to automate workflows or work interactively in Jupyter Notebooks; and q2cli and q2cwl, providing a command line interface and Common Workflow Language wrappers for QIIME 2, designed for high-performance computing experts.

Advances in microbiome research promise to improve many aspects of our health and our world, and QIIME 2 will help drive those advances by enabling accessible, community-driven microbiome data science.
Figures and figure captions

Figure 1: QIIME 2 provides many interactive visualization tools. Interactive versions of these screen captures are available in Supplementary File 1 and at [https://github.com/qiime2/paper1](https://github.com/qiime2/paper1). Detailed descriptions and methods are included in Online Methods. (A) Unweighted UniFrac PCoA plot containing 37,680 samples, illustrating the scalability of QIIME 2. Colors indicate sample type as described by the Earth Microbiome Project ontology (EMPO). (B) A feature volatility plot illustrating change in Bifidobacterium abundance over time in breast-fed and formula-fed infants. Temporally interesting features can be interactively discovered with this visualization. (C) Interactive taxonomic composition bar plot illustrating phylum-level composition of microbial mat samples collected along a temperature gradient in Yellowstone National Park Hot Spring outflow channels (Steep Cone Geyser). The many interactive controls available in this plot vastly reduce the burden of exploratory analysis over QIIME 1. (D) Molecular cartography of the human skin surface. Colored spots represent the abundance of the small molecule cosmetic, sodium laureth sulfate, on the human skin. Sample data can be interactively visualized on 3D models, supporting the discovery of spatial patterns.
Figure 1

(A) Multidimensional scaling (MDS) plot of the bacterial communities in the gut microbiota across different environments. The axes represent the relative abundance of different bacterial communities. (B) Time series analysis showing the relative abundance of Bifidobacterium across different diets (breastmilk dominant vs. formula dominant). (C) Taxonomic level and color palette selection for sample sorting. (D) 3D visualization of the relative frequency of different bacterial communities across temperature gradients.
**Figure 2:** QIIME 2 iteratively records data provenance, ensuring bioinformatics reproducibility. This simplified diagram illustrates the automatically tracked information about the creation of the taxonomy barplot presented in Figure 1c. QIIME 2 results (circles) contain network diagrams illustrating the data provenance stored in the result. Actions (quadrilaterals) are applied to QIIME 2 results and generate new results. Arrows indicate flow of QIIME 2 results through actions. TaxonomicClassifier and FeatureData[Sequence] inputs contain independent provenance (red and blue, respectively) and are provided to a classify action (yellow), which taxonomically annotates sequences. The result of the classify action, a FeatureData[Taxonomy] result, integrates the provenance of both inputs with the classify action. This result is then provided to the barplot action with a FeatureTable[Frequency] input, which shares some provenance with the FeatureData[Sequence] input as they were generated from the same upstream analysis. The resulting Visualization (Figure 1c), has the complete data provenance and correctly identifies shared processing of inputs. An interactive and complete version of this provenance graph (as well as those for other Figures 1 panels) can be accessed through Supplementary File 1.
Code availability

QIIME 2 is open source and free for all use, including commercial. It is licensed under the BSD 3-clause license. Source code is available at https://github.com/qiime2.

Data availability

Data for the analyses presented in Figure 1 are available as follows: (a) Earth Microbiome Project data was obtained from ftp://ftp.microbio.me/emp/release1, and the American Gut Project (AGP) data was obtained from Qiita (http://qiita.microbio.me) study 10317. (b) Sequence data are available in Qiita under study id 10249 and EBI under accession number ERP016173. (c) Sequence data are available in Qiita under study id 925 and EBI under accession number ERP022167. (d) Data are available in the q2-ili GitHub repository (https://github.com/biocore/q2-ili). Interactive versions of the Figure 1 visualizations can be accessed at https://github.com/qiime2/paper1.

Acknowledgements

QIIME 2 development was primarily funded by NSF Award 1565100 to JGC and RK. Partial support was also provided from the following grants: NIH U54CA143925 (JGC, TP) and U54MD012388 (JGC, TP); grants from the Alfred P. Sloan Foundation (JGC, RK); ERC-STG project MetaPG (NS); Strategic Priority Research Program of the Chinese Academy of Sciences QYZDB-SSW-SMC021 (YB); from the Australian National Health and Medical Research Council APP1085372 (GAH, JGC, Von Bing Yap and RK); and from Natural Sciences and Engineering Research Council (NSERC) to DLG. Thanks to the Yellowstone Center for Resources for research permit #5664 to JRS for Yellowstone access and sample collection. We would like to thank the users of QIIME 1 and 2, whose invaluable feedback has shaped QIIME 2.

Author contributions

EB, JRR, MRD, NAB, YB, JEB, CJB, AMC, EC, RD, CFE, MEs, JMG, DLG, AKJ, KBK, STK, IK, TK, JL, YL, AVM, JLM, LFN, SBO, DP, AS, SJS, ADS, LRT, PJTo, PJTu, SU, FV, JW, RK, and JGC developed documentation, educational materials, and/or user/developer support content. EB, JRR, MRD, NAB, and JGC wrote the manuscript; all authors assisted with revision of the manuscript. EB, JRR, MRD, NAB, and JGC designed and developed the QIIME 2 framework. DMD, RL, EL, SCM, RS, JRS, WW, CHDW, and RK contributed data used in the manuscript and/or testing of the QIIME 2. CA, CTB, EC, PCD, SH, PK, EL, TP, RS, EV, YW, and RK contributed to the design of analytic methods. EB, JRR, MRD, NAB, and JGC designed and developed the QIIME 2 framework. DMD, RL, EL, SCM, RS, JRS, WW, CHDW, and RK contributed data used in the manuscript and/or testing of the QIIME 2. CA, CTB, EC, PCD, SH, PK, EL, TP, RS, EV, YW, and RK contributed to the design of analytic methods. EB, JRR, MRD, NAB, and JGC designed and developed the QIIME 2 framework. DMD, RL, EL, SCM, RS, JRS, WW, CHDW, and RK contributed data used in the manuscript and/or testing of the QIIME 2. CA, CTB, EC, PCD, SH, PK, EL, TP, RS, EV, YW, and RK contributed to the design of analytic methods.
PT, AT, JJJV, YV, MV, MW, KCW, ADW, ZZX, JRZ, YZ, and JGC contributed software to QIIME 2 plugins, interfaces, framework, and/or build and test systems.

Main text references


Author affiliations

1Pathogen and Microbiome Institute, Northern Arizona University, Flagstaff, AZ, USA. 2Metabolic Epidemiology Branch, National Cancer Institute, Rockville, MD, USA. 3Department of Computer Science and Engineering, University of Minnesota, Minneapolis, Minnesota, USA. 4Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA. 5Department of Population Health and Reproduction, University of California, Davis, CA, USA. 6Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. 7Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA. 8University of Copenhagen, Faculty of Health and Medical Sciences, Novo Nordisk Foundation Center for Basic Metabolic Research, Copenhagen, Denmark. 9Centre for Integrative Biology, University of Trento, Trento, Italy. 10State Key Laboratory of Plant Genomics, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing, China. 11Centre of Excellence for Plant and Microbial Sciences (CEPAMS), Institute of...
50Department of Computer Science & Engineering, University of California San Diego, La Jolla, CA, USA. 51Department of Statistics, University of Washington, Seattle, WA, USA. 52Department of Animal Science, Colorado State University, Fort Collins, CO, USA. 53Irving K. Barber School of Arts and Sciences, Unit 2 (Biology), University of British Columbia, Kelowna, BC, Canada. 54Mountain View, Google LLC, Mountain View, CA, USA. 55Center for Microbiome Innovation, University of California San Diego, La Jolla, CA, USA. 56School of Information Studies, Syracuse University, Syracuse, NY, USA. 57School of STEM, University of Washington Bothell, Bothell, WA, USA. 58Department of Biological Sciences, Webster University, St Louis, MO, USA. 59Agricultural Research Service, Genomics and Bioinformatics Research Unit, United States Department of Agriculture, Gainesville, FL, USA. 60College of Medicine, Department of Biomedical Informatics, University of Arkansas for Medical Sciences, Little Rock, AR, USA. 61Computational Bioscience Program, University of Colorado Denver Anschutz Medical Campus, Aurora, CO, USA. 62Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO, USA. 63Department of Biological Sciences and Northern Gulf Institute, University of Southern Mississippi, Hattiesburg, Mississippi, USA. 64Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, La Jolla, CA, USA. 65Department of Biology, San Diego State University, San Diego, CA, USA. 66Department of Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, USA. 67Division of Biological Sciences, University of California San Diego, San Diego, CA, USA. 68Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA, USA. 69Quantitative and Systems Biology Graduate Program, University of California Merced, Merced, CA, USA. 70Bioinformatics Group, Wageningen University, Wageningen, The Netherlands. 71Department of Mathematics, University of Arizona, Tucson, AZ, USA. 72National Laboratory Service, Environment Agency, Starcross, UK. 73College of Agriculture and Life Sciences, University of Florida, Gainesville, FL, USA. 74Department of Biostatistics, University of Washington, Seattle, WA, USA. 75University of Washington Bothell, School of STEM, Division of Biological Sciences, Bothell, WA, USA. 76Merck & Co. Inc., Kenilworth, NJ, USA. 77Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA.
Overview of QIIME 2

We provide a high-level overview of the QIIME 2 system. Monospace font is used to indicate literal terms, such as objects defined by QIIME 2. The most up-to-date information on these topics is available in the QIIME 2 developer documentation at https://dev.qiime2.org.

There are three core components of the QIIME 2 system architecture: the framework, the interfaces, and the plugins (Figure S1). Interfaces are responsible for turning user intent into action. Plugins define all domain-specific functionality. The most important restriction of the architecture, which is evident in Figure S1, is that interfaces and plugins do not communicate directly with one another -- that communication is always mediated by the framework. In other words, the domain-specific analytic functionality (defined in plugins) is entirely decoupled from how users interface with the system (defined in interfaces). This important constraint allows multiple kinds of interfaces to be dynamically generated, and as a result QIIME 2 can adapt its user interface to the audience and the task at hand (Figure S2).

Third-party developers can create and distribute both plugins and interfaces for QIIME 2 independently of the core QIIME 2 development group, which forms the basis for our goal of decentralized QIIME 2 development (see https://library.qiime2.org and https://dev.qiime2.org). By removing our team as a bottleneck in developers delivering their new methods to users through QIIME 2, microbiome research can advance more quickly by ensuring that QIIME 2 users can have access to the latest microbiome analytic methods as quickly as bioinformatics researchers and developers can distribute them. This model makes QIIME 2 (and tools that build on it, such as Qiita) a platform for microbiome data science, not only a tool for a specific type of analysis. Since plugins conform to requirements specified by the framework, framework features such as data provenance tracking and multiple interface support are available for all plugins without the plugin developers having to be aware of these features.

In the terminology of QIIME 2, an Action creates a Result, and a Result can be either an Artifact or a Visualization. An Artifact is data generated by one or more QIIME 2 Actions which can be used as the input to other QIIME 2 Actions. A Visualization on the other hand is a terminal output of QIIME 2, which could be an interactive visualization (as in the Figure 1 examples) or any other result that is intended to be consumed by humans (not by a QIIME 2 Action). QIIME 2 assigns version 4 universally unique identifiers (UUIDs) to each execution of an Action, and to all Results.
QIIME 2 stores information about the series of Actions that led to a Result, along with information about the environment (including versions of all QIIME 2 packages and other Python dependencies) where each Action was executed, and the data itself. We refer to this process as data provenance tracking, or simply provenance tracking. We did not want to create new bioinformatics file formats to support the storage of data provenance, so QIIME 2 Results are instead stored as zip files containing a data directory that contains only the data in a relevant format (e.g., fasta or fastq for sequence data, newick for phylogenetic trees, etc), plus QIIME-2-specific metadata in other directories (such as provenance). These files use the extension .qza (for QIIME zipped artifact) or .qzv (for QIIME zipped visualization), but they are standard zip files that could be unzipped using common tools such as unzip, WinZip, or 7-Zip. Additional motivations for the storage of QIIME 2 Results in these structured zip files include the ability to submit as supplementary material to journals (the extension can simply be changed to .zip if required by the journal); “future-proofing” of QIIME 2 Results (even if QIIME 2 weren’t used anymore, Results could still be accessed by unzipping .qza or .qzv files - see Extracting data from QIIME 2 archives below); zip files contain an index, allowing them to be inspected for certain information without uncompressing them; and data are always compressed, facilitating data sharing. Because provenance is stored alongside data in .qza and .qzv files, provenance tracking is decentralized (no QIIME 2 server or database needs to be keeping track of this information) ensuring that information on how data was generated will not be lost as long as the data is intact. However, assignment of UUIDs to all QIIME 2 Results (as described above) lends itself to managing these data in a database if that is desired.

Another important component of QIIME 2 is its semantic type system. All Artifacts used in QIIME 2 are annotated with a semantic description of their type which conveys the meaning of the data. Semantic types differ from data types (how data is represented in memory) or file formats (how data is stored on disk), and allow QIIME 2 to constrain the composition of multiple actions to only those combinations which are semantically meaningful without needing to consider the specific file formats or data types. This also makes it possible to determine what Actions could be applied (and in what order) to generate a given Artifact from some set of input Artifacts. For example, phylogenetic trees in QIIME 2 can be either rooted or unrooted, and these two concepts are represented by the semantic types Phylogeny[Rooted] and Phylogeny[Unrooted], respectively. QIIME 2 could support loading these into multiple different data types, including a scikit-bio TreeNode object or an ete3 Tree object. Both of these types are typically stored on disk in a newick-formatted file, but this format doesn’t contain easily accessible information on whether the phylogeny is rooted or unrooted. Some QIIME 2 Actions can only generate a Phylogeny[Unrooted] (such as fasttree), and some other Actions only work on Phylogeny[Rooted] (such as beta-phylogenetic, which computes UniFrac distances). The semantic type system allows QIIME 2 to determine that the output of fasttree should not be directly provided as input to beta-phylogenetic, and to provide the user with that information prior to execution. This can help a researcher who is new to microbiome data science avoid using data incorrectly. This will also enable QIIME 2 to automatically assist users in identifying relevant workflows to generate desired data or further explore data they already have.
Due to recent advances in package management systems and bioinformatics package repositories (e.g., Anaconda, Bioconda\(^1\), and Bioconductor\(^2\)), QIIME 2 is straightforward to install.

**QIIME 2 View**

QIIME 2 View ([https://view.qiime2.org](https://view.qiime2.org)) is a unique and novel contribution to the microbiome data science ecosystem that facilitates collaborative research. A user who has generated QIIME 2 visualizations can share those visualizations with a collaborator who can explore the results interactively without having QIIME 2 installed. QIIME 2 View achieves this simplified sharing of complex interactive visualizations through a novel combination of modern web browser APIs within a single-page application. It allows a user’s browser to open and read .qza and .qzv files without the need to transfer the files over the network by utilizing a Service Worker to redirect HTTP requests directly into the archive which is retained on the user’s computer. This approach of data unpackaging and local command execution makes QIIME 2 View well suited to cases where the results are unpublished or contain private information (that information will not be stored on any remote server). It is also possible to create “smart” URLs which automatically fetch content from a CORS-enabled web-server (for example, see the links in the README.md file at [https://github.com/qiime2/paper1](https://github.com/qiime2/paper1)). This makes it very simple to share a single link with a collaborator that will be resolved into a fully interactive visualization on a user’s computer automatically. The structured nature of the archive format (Figure S3) also allows QIIME 2 View to generate a dynamic provenance visualization, summarizing the entire provenance of the archive in question.

**Extracting data from QIIME 2 archives**

QIIME 2 .qza and .qzv files are zip file containers with a defined internal directory structure. It’s very easy to get data out in the canonical formats (Figure S3). If QIIME 2 and the q2cli command line interface are installed, this can be achieved using the qiime tools export command. If QIIME 2 is not installed, this can be achieved using standard decompression utilities such as `unzip`, WinZip, or 7-zip. We illustrate how this can be achieved using `unzip` on macOS. This can similarly be achieved on Windows or Linux. We illustrate this here to further future-proof QIIME 2 Results - even if the QIIME 2 documentation were no longer accessible, users could follow these steps to access QIIME 2 Results.

First, obtain a .qza file. Here we use the FeatureData[Sequence] artifact generated during the QIIME 2 Moving Pictures tutorial.

```bash
$ wget https://docs.qiime2.org/2018.8/data/tutorials/moving-pictures/rep-seqs.qza
```

Next, unzip that file with the macOS (or Linux) `unzip` program. This will create a new directory. The name of that directory will be the UUID of the artifact being unzipped, in this case

```
8dc793b8-7284-462a-8578-6370ffceebdc.
```
$ unzip rep-seqs.qza
 Archive: rep-seqs.qza
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/metadata.yaml
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/VERSION
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/metadata.yaml
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/citations.bib
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/VERSION
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/artifacts/bdaa3214-f883-4c8b-8db3-f6ea4910d724/metadata.yaml
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/artifacts/bdaa3214-f883-4c8b-8db3-f6ea4910d724/citations.bib
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/artifacts/bdaa3214-f883-4c8b-8db3-f6ea4910d724/VERSION
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/artifacts/bdaa3214-f883-4c8b-8db3-f6ea4910d724/action/action.yaml
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/provenance/artifacts/bdaa3214-f883-4c8b-8db3-f6ea4910d724/action/barcodes.tsv
 inflating: 8dc793b8-7284-462a-8578-6370ffceebdc/data/dna-sequences.fasta

The last entry that is unzipped in this example is data/dna-sequences.fasta. All other directories and files are QIIME 2 specific metadata (such as information about the semantic type of the artifact and the data provenance). If you’re only interested in the sequence data, you can safely ignore all of that information. The data/dna-sequences.fasta file is a typical fasta file containing sequence identifiers and sequences. The first four lines of this file can be viewed as follows:

$ head -4 8dc793b8-7284-462a-8578-6370ffceebdc/data/dna-sequences.fasta
>3f52cf1f8eefc483511c2270aabd0ae6
TACGTAGGGTGCGAGCGTTAATCCGAATTAATCGGGCGTAAACGCGAGGGTTTTGTAAGACAGAGGTGAAATCCCCGGGCT
>82e7255267397b777a1af4de4e22755
CAACGGTGAATCCGGCTTATCTCGCGGTAACGTGCGTGAGGCCTGTTCTTGTGAGAGGTGAAATCCCCGGGCT
CAACGGTGAATCCGGCTTATCTCGCGGTAACGTGCGTGAGGCCTGTTCTTGTGAGAGGTGAAATCCCCGGGCT
CAACGGTGAATCCGGCTTATCTCGCGGTAACGTGCGTGAGGCCTGTTCTTGTGAGAGGTGAAATCCCCGGGCT
CAACGGTGAATCCGGCTTATCTCGCGGTAACGTGCGTGAGGCCTGTTCTTGTGAGAGGTGAAATCCCCGGGCT
QIIME 2 user and developer community

QIIME 2 officially succeeded QIIME 1 (http://www.qiime.org)\(^3\) in January of 2018, and has developed an engaged user base and community. As of this writing there are over 1980 active users (users who have performed an action, such as creating or liking a post) on the QIIME 2 Forum; over 3000 monthly downloads of QIIME 2 from Anaconda; over 8000 unique visitors to the QIIME 2 Forum according to Google Analytics; and our multi-day workshops are frequently filled to capacity (https://workshops.qiime2.org). QIIME 2 is also being adopted by third-party bioinformatics developers who are choosing to make their software accessible through plugins, and who are motivated to develop for QIIME 2 by access to its integrated provenance tracking, multiple interfaces, standardization of data types provided by the semantic type system, large user community, and supportive developer community.

A core goal of QIIME 2 is to cultivate a diverse and inclusive community of scientists, software engineers, statisticians, educators, students, and other microbiome stakeholders who are openly sharing methods, data, and knowledge to advance microbiome research.
Supplementary figures, files, and captions

Figure S1. Schematic diagram of the QIIME 2 system. Interfaces define how users interact with the system; plugins define all domain-specific functionality; and the framework mediates communication between plugins and interfaces, and performs core functionality such as provenance tracking. Arrows indicate dependencies. Interfaces interact only with the qiime2.sdk submodule, while plugins interact only with the qiime2.plugin submodule. This design has led to a system that is readily extended by third-party plugin and interface developers.
**Figure S2. QIIME 2 is interface agnostic.** The full suite of QIIME 2 functionality is useful to and usable by researchers ranging widely in their computational sophistication, a major advantage over technologies such as QIIME 1 that provide a single interface. (a) Users wanting to view QIIME 2 results or data provenance can use QIIME 2 View without installing QIIME 2, which is convenient for lead investigators, clinicians, or policy makers who may want to explore interactive visualizations generated by others. (b) Researchers who prefer graphical interfaces can use QIIME 2 Studio, our prototype graphical interface. This is convenient for users without command line or programming skills. (c) Power users (e.g., who are comfortable with the Linux command line and/or regularly work on institutional computer clusters), can use QIIME 2 through the command line interface, q2cli. (d) “Data scientists” (e.g., users who are programmers, who work in Jupyter Notebooks, or who are interested in automating QIIME 2 workflows), can use QIIME 2 through the Python 3 “artifact API”. 
Figure S3. Anatomy of a QIIME 2 Archive (i.e., .qza or .qzv file). QIIME 2 stores data in a directory structure called an Archive. These archives are zipped to make moving data convenient. The directory structure has a single root directory named with a UUID which serves as the identity of the archive.
**Supplementary File 1** contains the QIIME 2 .qzv files corresponding to **Figure 1a-d.** These are also accessible at [https://github.com/qiime2/paper1](https://github.com/qiime2/paper1) and can be viewed using QIIME 2 View ([https://view.qiime2.org](https://view.qiime2.org)) where readers can interact with the results, and explore the methods used to generate them (see the Provenance tab after loading a .qzv file with QIIME 2 View).

We describe the methods used to generate each of these visualizations here, and this information can be compared to the data provenance which contains far more detail that is possible or desirable to include in supplementary methods text.

**a-pcoa.qzv:** Emperor PCoA plot presenting a meta-analysis of the first release of the Earth Microbiome Project (EMP)⁴ and the first release of the American Gut Project (AGP)⁵. The EMP data was obtained from [ftp://ftp.microbio.me/emp/release1](ftp://ftp.microbio.me/emp/release1), and the AGP data was obtained from Qiita study 10317 for the set of samples used in its publication (samples described in the AGP supplemental data accession table). Both projects were downloaded and imported into QIIME 2 as BIOM tables⁶. The contingency matrices were combined, filtered for blooms⁷, rarefied at an even depth (1000 sequences per sample), and compared using the unweighted UniFrac⁸ metric. Lastly the samples were projected into a small dimensional space using principal coordinates analysis and visualized using Emperor⁹. The samples were colored according to the Earth Microbiome Project Ontology⁴.

**b-feature-volatility.qzv:** Data were generated on five sequencing runs of V4 16S rRNA gene amplicons from the ECAM study¹⁰. Forward reads were imported separately in EMPSingleEndDirFmt format, demultiplexed with q2-demux’s `emp_single` method, and denoised using q2-dada2’s `denoise_single` method (`trunc_len=150`, other parameters used default values)¹¹. Denoised feature tables and sequences were merged using q2-feature-table’s `merge` and `merge-seqs` methods, respectively. q2-feature-table’s `filter-samples` method was used to remove samples with fewer than 2000 sequences, and to perform metadata-based filtering to retain only children’s samples. A naive Bayes taxonomy classifier was trained on the Greengenes¹² reference sequences (clustered at 99% similarity) using q2-feature-classifier’s `fit-classifier-naive-bayes` method¹³. This classifier was used to taxonomically classify the ECAM ASVs using q2-feature-classifier’s `classify-sklearn` method¹³. ASVs were collapsed based on genus-level taxonomy using q2-taxa’s `collapse` method. Temporally predictive features were identified using q2-longitudinal’s `feature-volatility` pipeline¹⁴ using default parameters. Data contained in this artifact have been described in a previous publication¹⁴.

**c-taxa-barplot.qzv:** Data were imported into QIIME 2 as multiplexed 2x150 MiSeq reads and demultiplexed. DADA2¹¹ was applied to single-end reads (as approximately 30% of reads failed to join due to the relatively short sequence length) with no trimming of reads. Taxonomy was assigned to the resulting amplicon sequence variants (ASVs) against the Silva version 132 99% OTUs (trimmed to the 515F/806R region of the 16S) using q2-feature-classifier’s `classify-sklearn` method¹³.

**d-ili-plot.qzv:** The primary files for this visualization are a stereolithography file (STL) and a sample metadata file with a mapping between samples and the spatial coordinates (x, y and z). Both files were obtained from 'ili's GitHub page¹⁵,¹⁶. The comma-separated file was converted into a to a tab-separated format (to make it compatible with QIIME 2).
Online methods references