# Qiita: report of progress towards an open access microbiome data analysis and visualization platform

The Qiita Development Team[‡][*]

https://www.youtube.com/watch?v=TQzXwQ9Vx08

**Abstract**—Advances in sequencing, proteomics, transcriptomics and metabolomics are giving us new insights into the microbial world and dramatically improving our ability to understand microbial community composition and function at high resolution. These new technologies are generating vast amounts of data, even from a single study or sample, leading to challenges in storage, representation, analysis, and integration of the disparate data types.

Qiita (https://github.com/biocore/qiita) aims to be the leading platform to store, analyze, and share multi-omics data. Qiita is BSD-licensed, unit-tested, and adherent to PEP8 style guidelines. New code additions are reviewed by multiple developers and tested using Travis CI. This approach opens development to the largest possible number of experts in "-omics" fields. The heterogeneous data generated by these disciplines led us to use a combination of Redis, PostgreSQL, BIOM ([Atr10]), and HDF5 for relational and hierarchical storage. The compute backend is provided by IPython's parallel framework (http://ipython.org/). In addition, the project depends on mature Python packages such as Tornado (http://www.tornadoweb.org/en/stable/), click (http://click.pocoo.org/4/), scipy (http://www.scipy.org), numpy (http://www.numpy.org), and scikit-bio (http://scikit-bio.org) among others. Most notably, the analysis pipeline is provided by QIIME (http://qiime.org), with EMPeror (http://emperor.microbio.me) serving as the visualization platform for high-dimensional ordination plots, which can be recolored interactively and manipulated using the sample metadata.

By providing the database and compute resources at http://qiita.microbio.me to the global community of microbiome researchers, Qiita alleviates the technical burdens, such as familiarity with the command line or access to compute power, that are typically limiting for researchers studying microbial ecology, while at the same time promoting an open access culture. Because Qiita is entirely open source and highly scalable, developers can inspect, customize, and extend it to suit their needs regardless of whether it is deployed as a desktop application or as a shared resource.

**Index Terms**—Microbiome, multi-omics, open science, metagenomics, metatranscriptomics, metaproteomics, metabolomics

## Introduction

In recent years, the importance of microbes, including bacteria, archaea, fungi, and unicellular eukaryotes, in ecological communities has been extensively studied ([Atr11], [Atr06]). As the costs of analytical techniques such as DNA sequencing have continued their dramatic decline and samples become relatively easy to collect, large volumes of data and new data types have allowed for the characterization of the potent effects that microbial communities can impart on both host-associated ([Atr07], [Atr03]) and environmental ([Atr09]) health. The myriad techniques that can be used to characterize each individual sample allow researchers to understand these communities in previously unattainable detail, but also pose new challenges for integrating the results from multiple levels of observational data into a coherent picture. These techniques are colloquially called "omics" techniques and allow researchers to study the entire collections of genes (genomics), gene transcripts (transcriptomics), proteins (proteomics), and metabolites (metabolomics) represented in samples.

The genome of an organism is all of its genetic material; the "metagenome" of an environmental sample is the union of all of the genomes present in the sample. Since the genome of an organism defines the organism's biological capabilities, metagenomic analysis allows researchers to approach the question of what are the organisms in a sample capable of doing, collectively? Current techniques for performing metagenomic analysis fragment the metagenome into small pieces, which are then sequenced in massively parallel fashion, and genes are identified by comparison to references containing known genes. This technique results in a highly detailed view, but is relatively expensive due to the amount of sequencing that must be performed and the computational effort required ([Atr13]). A less detailed (but much cheaper and still very useful) characterization of a microbial community can be attained by performing targeted sequencing of marker genes. Sequences from marker genes are commonly grouped by similarity into operational taxonomic units (OTUs), groupings that might correspond to species, or genera, or classes, etc. A powerful way to identify the OTUs present in a sample is to amplify and sequence genes encoding components of the ribosome (rather than all of the genes). The ribosome is a cellular component that translates transcripts into proteins that is shared across the tree of life. Because it is believed to be under neutral evolution, mutations accrue at a relatively consistent rate, allowing it to be thought of as a "molecular clock" that provides phylogenetic information about the organism it came from ([Atr15]). In bacteria and archaea, amplicons of the 16S small subunit ribosomal gene are the most commonly used, while in eukaryotes the analogous 18S small subunit ribosomal gene is used (although for fungi, often parts of the internal transcribed spacer region are included for additional phylogenetic signal).

The central dogma of biology is that genes are transcribed into messenger RNAs (mRNAs), which are then translated into specific proteins. Tight regulation at each level is required for proper

* *Corresponding author: robknight@ucsd.edu*
‡ *University of California, San Diego*

cellular function. If amplicon sequencing and metagenomics help answer the questions of who's there, and what are they capable of doing, transcriptomics help answer the question what genes are actually being expressed right now? Genes that are "on" can be recognized by the presence of mRNA transcripts that identifiably correspond to the gene. Sequencing these transcripts elucidates which genes are actually being expressed ([Atr04]).

However, even the added depth provided by transcriptomic analyses does not paint the full picture. First, because transcripts are continuously being generated and degraded by cellular processes, only a snapshot of the transcriptome can be obtained from a single sample. Second, the regulatory mechanisms that govern the translation of transcripts into proteins do not treat all transcripts uniformly. Indeed, the abundances of proteins in a cell correlate only weakly with the abundances of their respective transcripts, as reviewed in [Atr08]. Therefore, protein levels must be measured directly using proteomics techniques to answer the question of how actively are observed transcripts actually being expressed as proteins? Proteins are inherently more complex molecules than DNA and RNA, and proteomics techniques fundamentally differ from genomics and transcriptomics techniques as they do not sequence nucleic acids. Instead of using genetic sequencers, instruments called mass spectrometers are used to fragment proteins and analyze the resulting charged peptides. The spectrum of peptides produced from a fragmented protein identifies it like a fingerprint (reviewed in [Atr01]).

The last "omics" technique considered here, metabolomics, provides an even more detailed view of cellular function by observing the presence of specific metabolites or all of the metabolites in a sample. Identifying the metabolites present in an organism (or group of organisms) helps answer the question, to what extent are these organisms interacting with and affecting their environments? Similar to proteomics, mass spectrometers are used to identify compounds and gauge metabolic interactions (reviewed in [Atr05]).

More and more commonly, studies are employing two or more of these techniques in "multi-omic" analyses of samples. Integrating these analyses and gaining biological insights from the preponderance of data resulting from each applied technique is a considerable challenge. For each technique, computational tools that process and digest raw data have been developed to varying levels of maturity, but orchestrating these tools into a coherent multi-omic analysis package has not yet been accomplished.

Moreover, the extremely rich datasets generated by each multi-omic study are valuable resources that can form the basis of subsequent "meta-analyses," wherein the original data are augmented with data from other new or existing studies. Meta-analysis has already been shown to be a powerful approach (Mason et al. 2014), and the potency of the approach increases as individual studies provide more and more detailed characterizations of their samples, enabling reuse of the data. The power of this approach underscores the scientific community's need for centralized resources for standardized, open access data.

Here, we present a progress report of Qiita, a multi-omic platform for meta-analysis that stresses standardization of data formats, open access to data and results, and methods for integrating samples across studies. As we design Qiita, we intend to account for the most common use-cases that a modern microbiome researcher will face. The following list briefly describes tasks that are streamlined using Qiita:

- Perform a microbiome analysis without any required knowledge of command line tools.
- Deposition of biological sequences into a public data repository, in specific the European Bioinformatics Institute's European Nucleotide Archive (ENA).
- Searching for studies based on sample and study metadata.
- Hosting of sequence data, sample metadata and processed files like BIOM tables.
- Provide a platform to collaboratively work on a dataset.
- Combine one or more studies into a single dataset to perform further specialized analyses.
- Analyze and organize different data types (16S, 18S, WGS, etc) into a single location where the sample metadata is enforced to be consistent across representations.

The list of tasks above, while not comprehensive, exemplifies some commonly encountered scenarios where Qiita is a powerful tool. Please also note that the last point regarding integration of multiple data types is a work in progress at this point. Currently, only portions of the 16S workflow are implemented, but there are plans for adding additional workflows (see future directions). Although other platforms and individual tools exist that are capable achieving one or more of these goals independently, such ad hoc pipelines are often troublesome, time consuming, and error prone.

## Structure and Operation

### *Modular organization*

Qiita follows a model-view-controller (MVC) architecture, with a Python module for each level (qiita_db, qiita_pet, and qiita_ware, respectively). Modularizing the platform in this way allows for flexibility in the case that different technologies are adopted as the project matures. It also permits customizability, since a user maintaining a deployment can choose to replace any of these modules with one of their own design as long as it operates using the same inter-module APIs.

### *Qiita-DB*

The qiita_db module defines a database schema in PostgreSQL (http://www.postgresql.org/) that serves to store and relate study metadata as well as system data. The schema was designed in DBSchema (http://www.dbschema.com/), which provides a convenient GUI for defining the table structure, setting constraints, and generating documentation. Although the project is under heavy development, there are active deployments of Qiita (e.g., http://qiita.microbio.me and http://qiita.ibdmdb.org). As development progresses and modifications to the database schema are required, they must be implemented and deployed in a way that preserves active deployments' data. Therefore, migrations are performed using a combination of SQL- and Python-based patches. In order to facilitate brand new deployments as well as accommodate upgrading active deployments, Qiita's GitHub repository contains the schema definition ab initio as well as all patches needed to upgrade it (modifying data of active deployments as needed) to the most up-to-date version. The database itself contains information about the currently deployed patch version so that what patches need to be applied, if any, can easily be determined. Psycopg (http://initd.org/psycopg/) provides Python bindings for interacting with PostgreSQL.

Several aspects of the data model itself bear mentioning. Users are identified by an email address and a password supplied upon

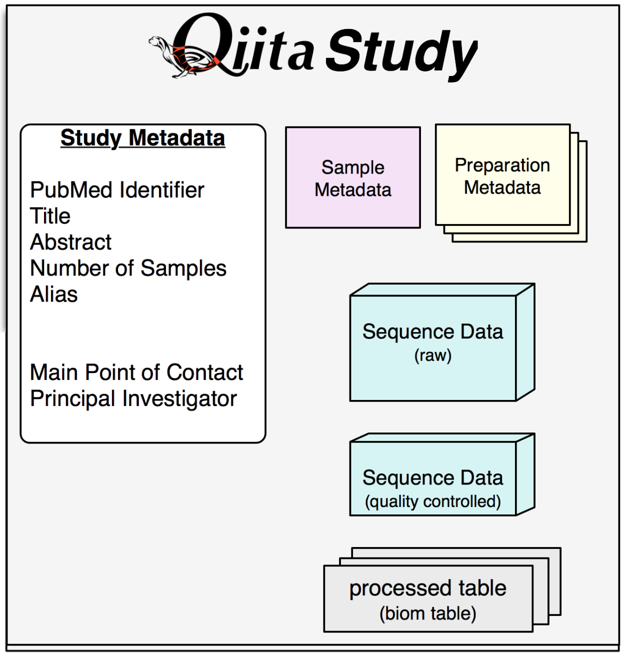

***Fig. 1:*** *Core structure of a study in Qiita. The study metadata broadly describes information about the study, the sample and preparation metadata refer to the biological specimens and their preparation method, the raw data refers to the files as generated by the instrument, quality controlled sequence data is generated for convenience and is used to create the processed tables.*

account creation. Passwords are salted and hashed with hashlib using bcrypt (https://github.com/pyca/bcrypt/). After users verify their email addresses, they are free to create "studies" by supplying some basic information such as the title of the study, an abstract, and what kind of environment is being studied, et al. Most of this information can be edited at any time after creation. Each study serves as a logical container for its associated data, metadata, and results files.

Because the system was designed with multi-omic analyses in mind, a distinction is made between metadata associated with the samples themselves (sample metadata) and metadata associated with preparations of those samples for biological processing (preparation metadata). In other words, sample metadata is invariant information about the samples themselves (e.g., the gender or age of the subject that was sampled), while preparation metadata for a 16S amplicon analysis of those samples would differ from preparation metadata for a proteomic analysis of the same samples. Note that the set of samples in two different preparation metadata might not overlap (or might overlap only partially) since not all samples are analyzed using all available techniques (see Figure 1). For example, the database currently contains a public study of about 100 samples taken from the site of the Deepwater Horizon oil spill in 2011 (study ID 1197; [Atr09]) where both 16S data and metagenomic data were collected. Some of the metadata collected (including the amounts of dissolved inorganic nitrogen, dissolved phosphate, amount of toluene, etc.) is specific to the samples themselves and will not vary with preparation; this is the sample metadata. On the other hand, some of the metadata is specific to a particular preparation of the samples for 16S analysis (including the region that was amplified, the primers that were used, the date the sequencing was performed, etc.); this is one set of preparation metadata. The subset of the full ~100 samples that were prepared for 16S analysis would be represented in this preparation metadata. For the metagenomic preparation, a smaller subset of the full ~100 samples were analyzed, so the metadata for that preparation would only contain information on those samples, and the data tracked would differ from the preparation metadata for the 16S analysis (for example, the preparation metadata for the metagenomic analysis would not contain a column for the 16S region).

Qiita (and the administrator(s) in a multi-user system) attempts to standardize as many fields of the metadata as possible using controlled vocabularies and ontologies when available. However, users are permitted to supply whatever sample and preparation metadata they deem relevant to their studies. Since the data that is supplied by users cannot be predicted a priori, a dynamic approach to storing the metadata must be taken. New tables are created dynamically using a consistent naming convention to keep track of each study's sample metadata and various preparation metadata, and another table keeps track of what fields are available in each metadata table and what the datatype of the field is. Like metadata fields, processing parameters are also standardized in order to minimize the impact of technical effects that would arise from heterogeneous processing. Tables for each key processing step, including demultiplexing, quality filtering, and OTU picking, keep track of these standard sets of parameters.

The qiita_db module also contains Python objects and utility functions that mediate filesystem and database interactions, similar in many respects to an object-relational mapper (ORM). Uploaded metadata files and raw data files (e.g., sequence data from a sequencing instrument) are stored in a directory structure with indirection to support horizontal scaling of file systems. Unlike the information in metadata files, the contents of raw data files are not stored in the database. Instead, the filepaths are recorded. This design facilitates processing the raw data files using external programs (e.g., programs that are implemented or wrapped in qiita_ware; see below) that need filehandles.

### *Qiita-pet*

The qiita_pet module defines components supporting a browser-based user interface. In a single-user deployment, tornado (http://www.tornadoweb.org/) handles all requests and serves all pages. In a multi-user deployment, nginx (http://nginx.org/) is required to serve downloads. While tornado is proficient at serving small or moderate files in small chunks, serving very large files can bog down the single-threaded server. Instead, tornado can be used to handle the initial request and to determine whether the file should be served (e.g., whether user has permission to access the file) before handing the request off to nginx to perform the actual file transfer. Another good use of nginx is as a load balancer sitting in front of several tornado web servers running on different ports.

Tornado templates provide a user interface that is based largely on bootstrap (http://getbootstrap.com/) and jQuery (https://jquery.com/). Other packages and extensions are used for various interface elements (for example, WTForms (https://github.com/wtforms/wtforms) is used for handling some form data, chosen (http://harvesthq.github.io/chosen/) provides improved select and multiple select form elements, and DataTables (https://www.datatables.net/) provides interactive and pleasantly formatted tabular displays). Asynchronous JavaScript and XML (AJAX) is used for the majority of asynchronous client-server communication, although websockets are employed when push notifications are useful (for example, when the server wants to notify a client that a processing job has completed).

### Qiita-ware

The qiita_ware module contains functions for manipulating input files, dispatching processing jobs, and performing operations on results files (e.g., submitting them to external data repositories like the European Bioinformatics Institute). Qiita is designed to be highly parallelizable through the use of IPython engines. Currently, the best supported workflow is for performing 16S amplicon analysis. For this workflow, scripts in the Quantitative Insights Into Microbial Ecology package (QIIME; [Atr02]) are executed from IPython engines to process users' input files and generate visualizations. Jobs are dispatched using mustached-octo-ironman (MOI; https://github.com/biocore/mustached-octo-ironman/), which serves the dual purpose of managing the submission of jobs and communicating their statuses to the browser-based interface through a websocket using pubsub calls with Redis as a message broker. Two packages are used to interface with Redis: redis-py (https://github.com/andymccurdy/redis-py) and toredis (https://github.com/mrjoes/toredis/), the latter of which provides a non-blocking mechanism for handling pubsub with Redis.

### Command line interface

In addition to the browser-based interface provided by qiita_pet, a command line interface (CLI) is also available. Qiita's scripts directory contains Python scripts that provide a command line interface to many of the system's capabilities through the click framework (http://click.pocoo.org/4/). The top-level qiita click group has subgroups (db, ware, and pet) for interfacing with each of the aforementioned modules along with a maintenance subgroup for performing administrative actions and probing the system's status. Note that all of the CLI commands assume that the user executing the commands has administrator access to Qiita.

### Data access control

Qiita can be deployed as either a single-user or multi-user system. A single-user deployment enforces virtually no data access restrictions; the sole user has ownership of all data in the system. The single-user deployment is intended for users who want a system that organizes their data and provides a graphical interface for performing analyses and meta-analyses. A multi-user deployment is more complex and depends on a group of administrators (at least one administrator is required) who moderate and curate additions and certain modifications to data in the system. Access to users' data is restricted based on the data's status, which can be one of sandboxed, private, or public.

Data that is sandboxed or private is visible only to its owner and other users with whom the owner explicitly chooses to share the data; data that is public is visible to all users of the system. Any user is free to upload, process, and explore his or her own sandboxed data using the full suite of tools provided, but the data is only minimally validated. The purpose of the sandboxed status is to allow users to get a quick look at their data -- and even rapidly compare it to other data in the system -- before expending a potentially large amount of time and effort detailing and correcting metadata-related minutiae.

Private data is assured to be maximally compatible with existing data in the system. Because computational validation can provide only a limited guarantee of compatibility, administrator approval is required to change a study from sandboxed to private status after a manual curation process. Manual curation helps ensure that new metadata uses controlled vocabulary and ontology terms where available, that applicable standards are followed (e.g., MIMARKS for marker gene sequence-related metadata), and that new user-defined metadata fields are introduced sparingly (for example, if there were already a field called "sex" in one or more existing studies, the curator would suggest amending a proposed "gender" field to avoid having multiple fields that contain the same class of information). It is possible but discouraged to revert data from private to sandboxed since another round of curation would be required to make it private again.

Once data is private, it is up to the user to decide if and when to make the data public at his or her discretion. At this stage, all users of the system are permitted to download and analyze the data, and the owner of the data can submit the data and metadata to a public repository such as the European Bioinformatics Institute (EBI; https://www.ebi.ac.uk). Reverting data from public to private has limited efficacy (since other users might have downloaded and/or performed analyses on the data) and requires administrator action.

### Configuration

By default, Qiita will look for a configuration file in a default location where an example configuration file is supplied. This behavior can be overridden by setting the QIITA_CONFIG_FP environment variable. This configuration file controls the behavior of various aspects of Qiita and its dependencies, including Postgres, IPython (http://ipython.org; [Atr12]), Redis (http://redis.io/), and MOI.

### Roadmap of future directions

Qiita is currently in alpha release and under active development. New functionality is continually being added, and these changes have the potential to affect all of the aforementioned submodules and interfaces, but any changes will maintain backwards compatibility with existing deployments. One planned enhancement will allow deployments to be "branded," so that not every Qiita deployment looks identical. In addition to supporting cosmetic changes, for example to logos or graphics, we will support the specification of multiple "portals" that coexist on one system and access a common database, but provide access to only desired subsets of the data. For example, we plan to introduce an Earth Microbiome Project ([Atr06]; EMP) portal that provides access to only EMP studies.

The most significant change currently planned will be the implementation of a plug-in system designed to support modular expansion of the system with new processing capabilities while maintaining a common user interface. We intend the plug-in system to support extensions to both the database schema and the Python framework by providing common interfaces to the main system. To demonstrate the feasibility of this approach, the current 16S analysis pipeline will be migrated to be the first plug-in. New users should note that right now, only portions of the 16S workflow are implemented. However, the data model and modularity that we have designed and built into the system will facilitate the addition of additional pipelines (including metagenomics, metaboloomics, and proteomics) through this upcoming plug-in system.

Another important change will affect data processing. Right now, in order to ensure consistent processing workflows, users can upload only raw data for processing on the system using standardized methods. However, the ability to enter the data processing

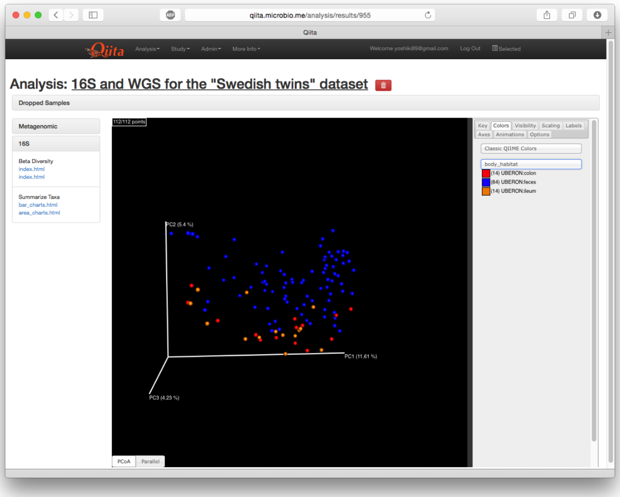

***Fig. 2:*** *Embeded beta diversity plot displayed using EMPeror showcasing an example dataset where samples are colored by the body habitat from where they were collected.*

workflows at downstream steps is a frequently requested feature that we plan to support. For the 16S analysis pipeline, users will be able to upload sequence files that have already been demultiplexed and/or quality filtered (e.g., by the sequencing center) or even BIOM tables of OTU picking results. The downside to these alternative pipeline entry points is that the standardized processing that is applied to other studies in the system cannot be guaranteed. For this reason, processing results that do not originate from raw data cannot be made available for public use like other results.

Due to the size and complexity of this nascent project, Qiita's documentation for users and developers is continuously evolving. For developers, the Numpydoc-formatted docstrings (https://github.com/numpy/numpydoc) that have already been added, which describe the system's Python objects and functions, will be rendered using sphinx (http://sphinx-doc.org/) and supplemented by markdown documents that provide additional details or instructions. For users, separate documentation will be made available covering key design concepts and how to interact with the system through the web interface.

### Interactive Visualizations

Allowing users to share, process, and combine their datasets easily does not ensure that interesting conclusions or insights will be generated. Only by carefully cross-examining results with sample metadata can correlations be observed and hypotheses developed. When working with large datasets (or combinations of datasets), effective visualizations are indispensable for presenting information in an intuitive manner and accelerating hypothesis generation. Collaborative efforts benefit greatly from visualizations that are portable and lightweight, qualities that allow researchers to communicate results and ideas to one another seamlessly.

One application that has proven useful to a large number of microbiome researchers is EMPeror ([Atr14]). While many existing tools are capable of displaying scatter plots, none of them actually integrates the sample metadata into the visualization on the fly while providing publication quality graphics. EMPeror accomplishes this integration, for example Figure 2 shows EMPeror executing within Qiita, meaning that users can interactively recolor points in space based on a metadata field using an intuitive browser-based interface. Other graphical manipulations of the points are also available, such as resizing or changing the opacity of arbitrary subsets of points. These capabilities shorten the gap between running a purely exploratory analysis and producing publication-quality figures.

As the development of EMPeror matures, other enhancements are being added, including the ability to view and interact with EMPeror plots from within an IPython notebook, supplementing textual descriptions with interactive plots. This feature is still in active development and will be available in a future release.

Since 2010, QIIME has provided the tools that utilize a sample's metadata to visualize taxonomic summaries, rarefaction curves, ordination plots, and even histograms of beta diversity distances. However these tools are usually limited, either because they are not extensible, lacking an interface that other web applications might use, or because they do not effectively provide both interactive and publication-quality static plots. The need for interactive, lightweight, and extensible browser-based visualization tools like EMPeror grows with the popularity of web-based scientific analysis platforms like BaseSpace (https://basespace.illumina.com/), Galaxy (https://galaxyproject.org/), iPlant (http://www.iplantcollaborative.org/), and KBase (https://kbase.us/), among others.

### Conclusions

Qiita provides a centralized resource where researchers can add their multi-omic datasets and process them in a standardized manner that maximizes their utility in meta-analyses. Organizing data and results, managing computational work, and interacting with all of the available tools poses a significant technical burden for researchers to surmount. Single-user deployments of Qiita help ameliorate this burden for individuals. Meanwhile, multi-user deployments serve as hubs that coordinate research efforts by facilitating the sharing of data and communication between users. Furthermore, a large, centralized, multi-user deployment that is maintained by the Qiita developers and staff at the University of California, San Diego, is available at http://qiita.microbio.me, where free data storage and compute clusters are provided to users. Regardless of the mode of deployment, a growing set of interactive results visualizations are provided by browser-based tools like EMPeror to accelerate the generation and exploration of new hypotheses.

### References

[Atr01] Aebersold R, Mann M, "Mass spectrometry-based proteomics," Nature 2003 Mar 13;422(6928):198-207.

[Atr02] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pea AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R, "QIIME allows analysis of high-throughput community sequencing data," Nature Methods 2010 May 7;7(5):335-6.

[Atr03] Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R, "Bacterial community variation in human body habitats across space and time," Science. 2009 Dec 18;326(5960):1694-7. doi: 10.1126/science.1177486.

[Atr04] Creecy JP and Conway T, "Quantitative bacterial transcriptomics with RNA-seq," Curr Opin Microbiol. 2015 Feb;23:133-40. doi: 10.1016/j.mib.2014.11.011. Epub 2014 Dec 5.

[Atr05] Dettmer K, Aronov PA, Hammock BD, "Mass spectrometry-based metabolomics," Mass Spectrom Rev. 2007 Jan-Feb;26(1):51-78.

[Atr06] Gilbert JA, Jansson JK, Knight R, “The Earth Microbiome project: successes and aspirations,” BMC Biology 2014, 12:69 doi:10.1186/s12915-014-0069-1.

[Atr07] Goodrich JK, Di Rienzi SC, Poole AC, Koren O, Walters WA, Caporaso JG, Knight R, Ley RE, “Conducting a microbiome study,” Cell 2014,158(2):250-62. doi:10.1016/j.cell.2014.06.037.

[Atr08] Maier T, Güell M, Serrano L, “Correlation of mRNA and protein in complex biological samples,” FEBS Lett. 2009 Dec 17;583(24):3966-73. doi: 10.1016/j.febslet.2009.10.036.

[Atr09] Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Bælum J, Kimbrel J, Bouskill NJ, Prestat E, Borglin S, Joyner DC, Fortney JL, Jurelevicius D, Stringfellow WT, Alvarez-Cohen L, Hazen TC, Knight R, Gilbert JA, Jansson JK, “Metagenomics reveals sediment microbial community response to Deepwater Horizon oil spill,” ISME J. 2014 Jul;8(7):1464-75. doi: 10.1038/ismej.2013.254.

[Atr10] McDonald D, Clemente JC, Kuczynski J, Rideout JR, Stombaugh J, Wendel D, Wilke A, Huse S, Hufnagle J, Meyer F, Knight R, Caporaso JG, “The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome,” Gigascience 2012 Jul 12;1(1):7. doi: 10.1186/2047-217X-1-7.

[Atr11] NIH HMP Working Group, Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA, Bonazzi V, McEwen JE, Wetterstrand KA, Deal C, Baker CC, Di Francesco V, Howcroft TK, Karp RW, Lunsford RD, Wellington CR, Belachew T, Wright M, Giblin C, David H, Mills M, Salomon R, Mullins C, Akolkar B, Begg L, Davis C, Grandison L, Humble M, Khalsa J, Little AR, Peavy H, Pontzer C, Portnoy M, Sayre MH, Starke-Reed P, Zakhari S, Read J, Watson B, Guyer M, “The NIH Human Microbiome Project,” Genome Res. 2009 Dec;19(12):2317-23. doi: 10.1101/gr.096651.109.

[Atr12] Pérez F, Granger B, “IPython: A System for Interactive Scientific Computing,” Computing in Science and Engineering, vol. 9, no. 3, pp. 21-29, May/June 2007, doi:10.1109/MCSE.2007.53. URL: http://ipython.org

[Atr13] Scholz MB, Lo CC, Chain PS, “Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis,” Curr Opin Biotechnol. 2012 Feb;23(1):9-15. doi: 10.1016/j.copbio.2011.11.013.

[Atr14] Vázquez-Baeza Y, Pirrung M, Gonzalez A, Knight R, “EMPeror: a tool for visualizing high-throughput microbial community data,” Gigascience 2013 Nov 26;2(1):16. doi: 10.1186/2047-217X-2-16.

[Atr15] Woese CR, “Bacterial evolution,” Microbiol Rev. 1987 Jun; 51(2): 221–271.