Galaxy mothur Toolset (GmT): a user-friendly application for 16S rRNA gene sequencing analysis using mothur

Saskia D. Hiltemann *
Stefan A. Boers *
Peter J. van der Spek
Ruud Jansen
John P. Hays
Andrew P. Stubbs
* These authors contributed equally to this study.

Submitted for publication.

ABSTRACT

Background

The determination of microbial communities using the mothur tool suite (https://www.mothur.org) is well established. However, mothur requires bioinformatics-based proficiency in order to perform calculations via the command-line. Galaxy is a project dedicated to providing a user-friendly web interface for such command-line tools (https://usegalaxy.org).

Results

We have integrated the full set of 125+ mothur tools into Galaxy as the Galaxy mothur Toolset (GmT) and provided a set of workflows to perform 'end-to-end' 16S rRNA gene analyses and integrate with third-party visualization and reporting tools. We demonstrate the utility of GmT by analysing the mothur MiSeq standard operating procedure (SOP) data set (https://www.mothur.org/wiki/MiSeq_SOP).

Conclusions

GmT is available from the Galaxy Tool Shed, and a workflow definition file and full Galaxy training manual for the mothur SOP have been created. A Docker image with a fully configured GmT Galaxy is also available.

KEY POINTS

- GmT provides a user-friendly interface to mothur by implementing mothur software in Galaxy
- A Galaxy workflow and full training manual for the mothur SOP are provided
- GmT provides integration with third-party visualization and reporting tools

FINDINGS

Introduction

16S rRNA gene profiling analysis can be achieved using an extensive array of sophisticated software including mothur,[1] QIIME,[2] MG-RAST,[3] and many more.[4] Whilst some of these applications have a graphical user interface (GUI) to provide access to these technologies for the research scientist, their use remains complex for non-bioinformaticians. In this respect, the Galaxy project was developed in order to simplify the use of complex command-line software tools.[5] Galaxy offers extensive support for both 16S rRNA gene-based and broader metagenomic analyses, with over 100 tools in the metagenomics section of the Galaxy Tool Shed, including QIIME,[2] KRONA,[6] PyNAST,[7] PICRUSt,[8] Kraken,[9] MetaPhlAn2,[10] HUMAnN2,[11] PrinSEQ,[12] Nonpareil,[13] VEGAN,[14] and many more.

mothur is an open-source application that was designed as a single piece of software capable of analysing and comparing microbial communities from 16S rRNA gene data derived from next-generation sequencing (NGS). The creators of mothur did not only provide an extensive set of tools, but also a collection of standard operating procedures (SOPs) that detail the recommended analytical protocol for different types of input data.

The latest version of mothur consists of over 125 components, lending it great flexibility, but at the same time, great complexity. To address this challenge, we have integrated the full set of 125+ mothur components into Galaxy that are collectively called the 'Galaxy mothur Toolset (GmT)'. To simplify usage of GmT we provide the full workflow definition files, usage of which shields the end-user from the full complexities of the analysis. By simultaneously providing access to all the individual components present in mothur as separate tools, expert users and bioinformaticians retain the ability to utilize the full flexibility of mothur by creating custom workflows or by modifying or extending our workflows to fit their use-case.

GmT also leverages Galaxy's collections framework to enable easy analysis of large numbers (many thousands) of samples at once. Many mothur components support parallel computing, and the Galaxy tools will utilize the maximum amount of processing power allotted to them by the instance administrator. As part of GmT, datatypes were also contributed to the Galaxy core codebase to facilitate the handling of mothur-specific datatypes within Galaxy. Furthermore, a Galaxy data manager was also created for the automatic installation and configuration of reference datasets utilized by the mothur tool suite. And lastly, a Galaxy interactive environment (GIE)[15] for Phinch[16] was also developed.[17]

GmT includes tools to produce standard file formats, such as the BIOM format[18] to facilitate interoperability with these downstream analysis components. Where no clear file standards exist, GmT provides custom tools for conversion of mothur datatypes to other tools (e.g. the taxonomy-2-krona tool). This allows for integration with third-party

tools such as PICRUSt for prediction of functional content, or visualisation tools such as Phinch, KRONA, and certain QIIME components. The mothur tools also natively support incorporation of some third-party analysis tools, such as UCHIME and ChimeraSlayer for chimera detection or VSEARCH for clustering, which are also available in GmT.

The Galaxy Training Network (GTN) is a network of people and groups that present Galaxy and Galaxy-based training around the world. The GTN has created a central repository for Galaxy training materials.[19] In order to further facilitate the use of GmT to end-users, we have contributed training materials to the GTN that illustrate how to run mothur's MiSeq SOP within Galaxy.[20] This work has also been incorporated in a larger-scale framework to easily and quickly explore microbiota data in a reproducible and transparent environment.[21]

Purpose of this work

The work performed and described in this technical note has four objectives. First to provide end-users and bioinformaticians with easy access to all the mothur tools as the GmT. Second is to provide open-access online training material to demonstrate/complete the mothur SOP in Galaxy. Third is to deliver an 'end-to-end' workflow for the mothur SOP in Galaxy that is available for upload to any Galaxy that has the GmT installed. Fourth is to provide a summarization of results in a web report using the iReport Galaxy tool.[22] Our aim is to provide 16S rRNA gene NGS analysis tools and awareness on how to use them in a format that supports FAIR data principles.[23]

Worked example

To illustrate the utility of our toolkit, we present results on example data below. GmT is designed to take short-read 16S rRNA gene NGS data as input and to output a dynamic web report for prokaryotic taxonomical classification using the Galaxy platform. A GmT workflow follows essentially a four-step process:

(i.) *Data upload*. The Galaxy platform provides the users with standard data upload functionality for single and multi-sample datasets.

(ii.) *Collection creation*. For multi-sample and/or paired-end datasets a Galaxy collection must be created in the Galaxy interface. Here datasets can also be assigned to groups. Galaxy will make intelligent suggestions for pairings of datasets based on the file names.

(iii.) *16S rRNA gene analysis*. mothur has been wrapped as a tool suite in Galaxy. Required steps included for a full 'end-to-end' 16S rRNA gene sequencing analysis consist of read-pair merging (mothur command: make.contigs), trimming of primer sequences (trim.seqs), additional quality control (screen.seqs), alignment of sequences to a (customized) reference alignment (align.seqs, screen.seqs), removal of chimeric sequences (chimera.uchime), classifying sequences using a Bayesian

classifier in combination with a reference database such as SILVA or GreenGenes (classify.seqs), and clustering of sequences into operational taxonomic units (OTUs) at a predefined percentage – usually 97 percent – of similarity (dist.seqs, cluster, and classify.otu) (Figure 1).

(iv.) *Experimental summary and reporting*. iReport in combination with KRONA[6] is used to deliver an HTML report in Galaxy. The iReport consists of multiple tabs to group results topically (e.g. taxonomy, rarefaction, diversity, quality control) and is highly customizable and easily tailored to an end-user's specific use-case. The entire report may be downloaded from the Galaxy interface to be viewed or shared offline. To compare the output from a single experiment or across multiple experiments we utilized Phinch,[16] a dynamic web application which uses BIOM-formatted files to explore and analyse biological patterns in 16S rRNA gene NGS datasets.

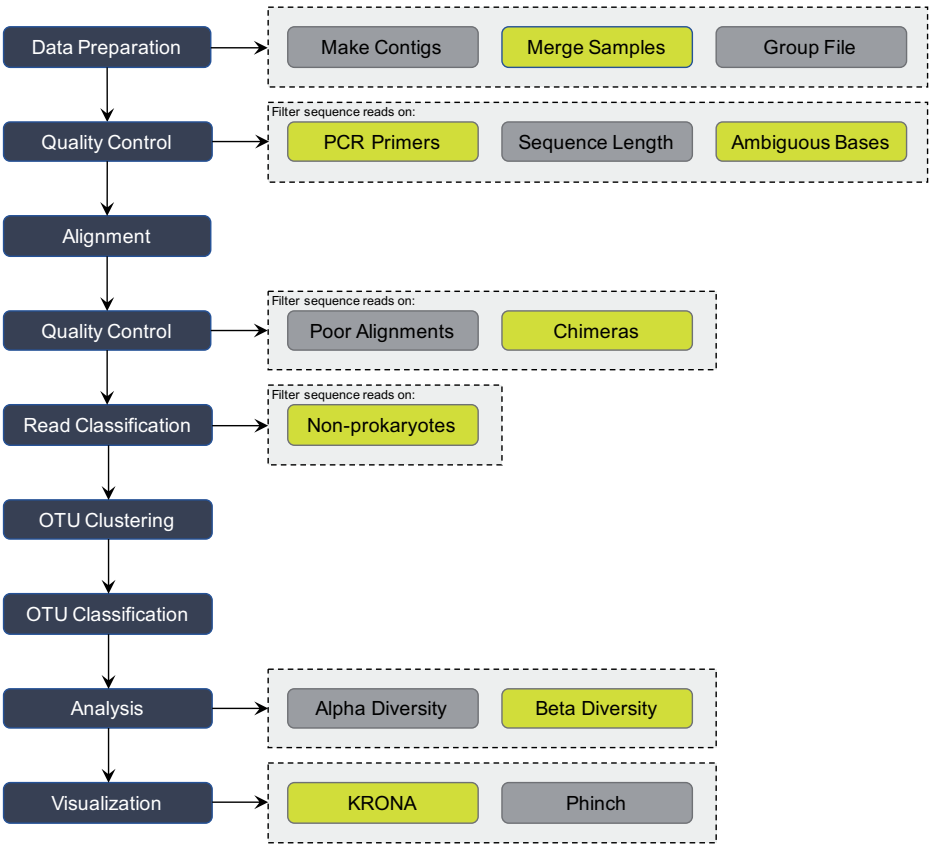

Figure 1. Conceptual view of the GmT mothur MiSeq SOP pipeline.

METHODS

Handling large datasets

Large-scale analyses have become the norm in the field, both large in disk space as in the number of files, and this can pose a challenge for analysis. For large files, Galaxy offers the option of uploading via FTP rather than web transfer. The introduction of the concept of 'collections' in Galaxy has enabled users to analyse datasets consisting of a large number of files (>100K) as easily as they would a single file.

Galaxy mothur Toolset

Many mothur components support parallelization, and our Galaxy wrappers will run these components with the maximum number of CPUs allotted to them by the Galaxy administrator. In order to diagnose potential failures, Galaxy outputs the full standard and error logs, which the users can inspect. Furthermore, we have contributed mothur datatype definitions to the Galaxy core code, meaning that the users will be protected from inputting the wrong datasets and thus reduce the number of errors they will make with the tools. All tools in GmT use only conda dependencies, making their installation in Galaxy a painless experience that requires nothing more than a single press of a button.

The mothur tool wrappers have been submitted to the Intergalactic Utilities Commission (IUC) tool repository,[24] and are available from the Galaxy Tool Shed.[25] The IUC is a group of community members dedicated to developing and upholding Galaxy tool development best practices and guidelines, thus by contributing our tools to this repo we ensure that the tools will be well-maintained. A metagenomics Galaxy flavour is available which contains all components presented here.[26] The full mothur suite has also been installed to Galaxy's main server.[27]

KRONA visualization

KRONA is a data viewer which provides the ability to interactively explore hierarchical data.[6] A Galaxy KRONA wrapper that works directly on mothur data formats was developed for this project.

Phinch visualization

Galaxy offers integration with Phinch BIOM format viewer[16] in two ways; as a Galaxy interactive environment (GIE) developed in the context of this project,[17] and more recently also as an external display application hosted by the Galaxy team.

iReport summarization

To facilitate the evaluation of 16S rRNA gene sequencing analysis results, integration with the iReport tool are also provided.[22] This tool creates a web report to present the

analysis results in an organized fashion and provides links to external resources such as BLAST searches (Figure 2).

OTUs at level 0.03

Show 10 entries Search:

OTU	DX01	DX02	Kingdom	Phylum	Class	Order	Family	Genus	BLAST
Otu0002	0	8,265	Bacteria	Proteobacteria	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Morganella	send to BLAST
Otu0003	6,565	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Anaerococcus	send to BLAST
Otu0004	0	6,607	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Alloprevotella	send to BLAST
Otu0005	0	3,868	Bacteria	Firmicutes	Bacilli	Lactobacillales	Enterococcaceae	Enterococcus	send to BLAST
Otu0006	0	502	Bacteria	Firmicutes	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	send to BLAST
Otu0007	1,219	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus	send to BLAST
Otu0008	1,108	0	Bacteria	Bacteroidetes	Bacteroidia	Bacteroidales	Prevotellaceae	Prevotella	send to BLAST
Otu0009	939	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Parvimonas	send to BLAST
Otu0010	787	0	Bacteria	Firmicutes	Clostridia	Clostridiales	Family_XI	Peptoniphilus	send to BLAST
Otu0011	16	0	Bacteria	Firmicutes	Bacilli	Lactobacillales	Streptococcaceae	Streptococcus	send to BLAST

Showing 1 to 10 of 37 entries Previous Next

Figure 2. Example iReport. This web report contains the interactive KRONA visualization, the (multi-sample) OTU table, rarefaction plots, diversity calculations, differential abundance analysis, and an extensive overview of the quality control measurements taken during the analysis. iReports are highly customizable and can be easily tailored to fit specific use-cases and end-user needs.

AVAILABILITY OF SOURCE CODE AND REQUIREMENTS

- Project name: Galaxy mothur Toolset (GMT)
- Project home page: https://github.com/erasmusmc-bioinformatics/galaxy-mothur-toolset
- Tool shed repository: https://toolshed.g2.bx.psu.edu/view/iuc/suite_mothur/768c2e48b706
- Training manual: http://galaxyproject.github.io/training-material
- GmT Docker image: https://quay.io/shiltemann/galaxy-mothur-toolset:16.07
- Galaxy Metagenomics Docker Flavour (Docker): https://quay.io/repository/shiltemann/galaxy-metagenomics, https://github.com/shiltemann/galaxy-metagenomics
- Phinch interactive environment: https://github.com/shiltemann/phinch-galaxy-ie
- Operating system(s): Unix (Platform independent with Docker)
- License: GNU GPL v3

AVAILABILITY OF SUPPORTING DATA AND MATERIALS

The data presented here to illustrate our work is the same data used in the training manual, and is available from Zenodo.[28]

ACKNOWLEDGEMENTS

The authors would like to thank Jim Johnson and the many other contributors and reviewers of the mothur tool wrappers, including everybody who contributed to the development of these tools within the context of the Galaxy metagenomics contribution fest organized by the Galaxy community's Intergalactic Utilities Commission (IUC), a group of community members dedicated to developing and upholding Galaxy tool development best practices and guidelines.[29] We would also like to thank the Galaxy Training Network for providing the infrastructure and valuable feedback to share our training materials. This work has received funding from the European Union's Seventh Framework Programme for Health under grant agreement number 602860 (TAILORED-Treatment; www.tailored-treatment.eu).

REFERENCES

1. Schloss PD, Westcott SL, Ryabin T, et al. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 2009; **75**: 7537-7541.
2. Caporaso JG, Kuczynski J, Stombaugh J, et al. QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 2010; **7**: 335-336.
3. Glass EM, Wilkening J, Wilke A, et al. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc* 2010; **1**: pdb–prot5368.
4. Oulas A, Pavloudi C, Polymenakou P, et al. Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. *Bioinform Biol Insights* 2015; **9**: 75.
5. Afgan E, Baker D, Van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 2016; **44**: W3–W10.
6. Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomics visualization in a web browser. *BMC Bioinformatics* 2011; **12**: 385.
7. Caporaso JG, Bittinger K, Bushman FD, et al. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 2009; **26**: 266-267.
8. Langille MG, Zaneveld J, Caporaso JG, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* 2013; **31**: 814.
9. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* 2014; **15**: R46.
10. Truong DT, Franzosa EA, Tickle TL, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* 2015; **12**: 902.
11. Abubucker S, Segata N, Goll J, et al. Metabolic reconstruction for metagenomic data and its application to the human microbiome. *PLoS Comput Biol* 2012; **8**: e1002358.
12. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* 2011; **27**: 863-864.
13. Rodriguez-R LM, Konstantinidis KT. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* 2013; **30**: 629-635.
14. Dixon P. VEGAN, a package of R functions for community ecology. *J Veg Sci* 2003; **14**: 927-930.
15. Eric R, Bjorn G, John C, et al. Galaxy Interactive Environments – a new way to interact with your data. In: Galaxy Community Conference; 2015.
16. Bik HM, Pitch Interactive. Phinch: an interactive, exploratory data visualization framework for -omic datasets. *bioRxiv* 2014: 009944.
17. Phinch Galaxy Interactive Environment: https://github.com/shiltemann/phinch-galaxy-ie.
18. The Biological Observation Matrix (BIOM) format: http://biom-format.org.
19. GTN Training Materials: https://training.galaxyproject.org.
20. Training materials for using GmT to run the mothur miseq SOP: https://galaxyproject.github.io/training-material/topics/metagenomics/tutorials/mothur-miseq-sop/tutorial.html.
21. Batut B, Gravouil K, Defois C, et al. ASaiM: a Galaxy-based framework to analyze raw shotgun data from microbiota. *bioRxiv* 2017: 183970

22. Hiltemann S, Hoogstrate Y, van der Spek P, et al. iReport: a generalised Galaxy solution for integrated experimental reporting. *Gigascience* 2014; **3**: 19.
23. Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR guiding principles for scientific data management and stewardship. *Sci Data* 2016; **3**: 160018.
24. IUC tool repository: https://github.com/galaxyproject/tools-iuc.
25. Galaxy Tool Shed: https://toolshed.g2.bx.psu.edu.
26. Metagenomics Galaxy Flavour: https://github.com/shiltemann/galaxy-metagenomics.
27. Galaxy Main server: https//usegalaxy.org.
28. Mothur MiSeq SOP Galaxy Tutorial Data: https://zenodo.org/record/800651.
29. Intergalactic Utilities Commission: https://galaxyproject.org/iuc.