Extending MetAMOS - new methods and new integrations

Marian Siwiak, Albert Bogdanowicz, Agnieszka Hajduk, Michał Krassowski, Paweł Jankowski, Mariia Savenko, Adam AP Pyzik, Paweł Szczęsny

Biodiversity analysis of metagenomic and metatranscriptomic data acquired from next-generation sequencing (NGS) requires following multiple analytic steps, often independent from each other with exception of passing output files of previous step as input for the following. If parameterization of steps following one after another is independent from one another, they may be pipelined. There are three most popular pipelines used for NGS analyses: QIIME, mothur and MetAMOS. In this work we describe our extensions to the latter. One is supplementing MetAMOS' default modes with taxonomic and metabolic biodiversity using metagenomics and metatranscriptomics data and the other provides a web-based interface to run predefined analyses that is easy to integrate with laboratory information management systems.

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1706v1 | CC-BY 4.0 Open Access | rec: 3 Feb 2016, publ: 3 Feb 2016

Extending MetAMOS - new methods and new integrations

Marian Siwiak[1,2,3#], Albert Bogdanowicz[2#], Agnieszka Hajduk[4], Michał Krassowski[4], Paweł Jankowski[4], Mariia Savenko[4], Adam Pyzik[2], Paweł Szczęsny[2,5*]

1 Oxford Data Science Centre, 154 Oxford Road, Oxford, Oxfordshire, OX4 2EB, United Kingdom
2 Institute of Biochemistry and Biophysics Polish Academy of Sciences, ul. Pawinskiego 5a, 02-106 Warsaw, Poland
3 Applied Research Institute for Prospective Technologies, Vismaliukų str. 34, Vilnius, Lithuania
4 University of Warsaw, Krakowskie Przedmiescie 26/28, Warsaw, Poland
5 Faculty of Biology, University of Warsaw, Miecznikowa, Warsaw, Poland
# both authors equally contributed to this study

* Correspondence should be addressed to:
Pawel Szczesny
Department of Bioinformatics
Institute of Biochemistry and Biophysics Polish Academy of Sciences
ul. Pawinskiego 5A
02-106 Warsaw
Poland
Email: szczesny@ibb.waw.pl

PeerJ PrePrints | https://doi.org/10.7287/peerj.preprints.1706v1 | CC-BY 4.0 Open Access | rec: 3 Feb 2016, publ: 3 Feb 2016

Abstract

Biodiversity analysis of metagenomic and metatranscriptomic data acquired from next-generation sequencing (NGS) requires following multiple analytic steps, often independent from each other with exception of passing output files of previous step as input for the following. If parameterization of steps following one after another is independent from one another, they may be pipelined. There are three most popular pipelines used for NGS analyses: QIIME, mothur and MetAMOS. In this work we describe our extensions to the last package. One extension is supplementing MetAMOS' default modes with taxonomic and metabolic analyses on metagenomics and metatranscriptomics data and the other extension provides a web-based interface to run predefined analyses that is easy to integrate with laboratory information management systems.

Keywords: metagenomics, metatranscriptomics, NGS, pipeline, web service

Introduction

There is a number of standardized protocols available for NGS data dubbed Standard Operating Procedures (SOPs). However, it is not uncommon for researchers to perform own sets of analyses, depending on specific research topic needs. The most popular pipelines for NGS analyses are QIIME (Caporaso et al. 2010), mothur (Schloss et al. 2009), and MetAMOS (Schloss et al. 2009; Treangen et al. 2013). The first two are focused on biodiversity assessment while the latter is aiming at assembly. Proper choice, deployment and pipelining of tools used in such analyses is a task non-trivial even for an experienced bioinformatician, and may pose a big problem for non-specialists.

The most easy to extended is definitely QIIME, as is it already a set of Python scripts. However its installation is so challenging, that the installation method recommended by QIIME authors is through VirtualBox or to use cloud computing instances with QIIME pre-installed. Mothur, being extremely portable due implementation in binary files, is hard to extend. In our case, requiring work with multiple instances of the software in a dynamic hardware environment, we set out to work with MetAMOS.

MetAMOS focus - namely creation of automated, reproducible, assembly & analysis pipeline was in full alignment with our project. Moreover, it enabled fast and efficient implementation of extension of functionalities, developed by our team.

This work presents an example of the solution to an effective cooperation between computational and experimental biologists, merging complex analysis pipeline with user-friendly interface through combination of MetAMOS with custom pipeline and web-interface.

Custom pipeline

Synthesis of research needs and technical capabilities of the research infrastructure required development of in-house tool called bipype (Fig 1). "bipype" stands for bioinformatics-python-pipe.

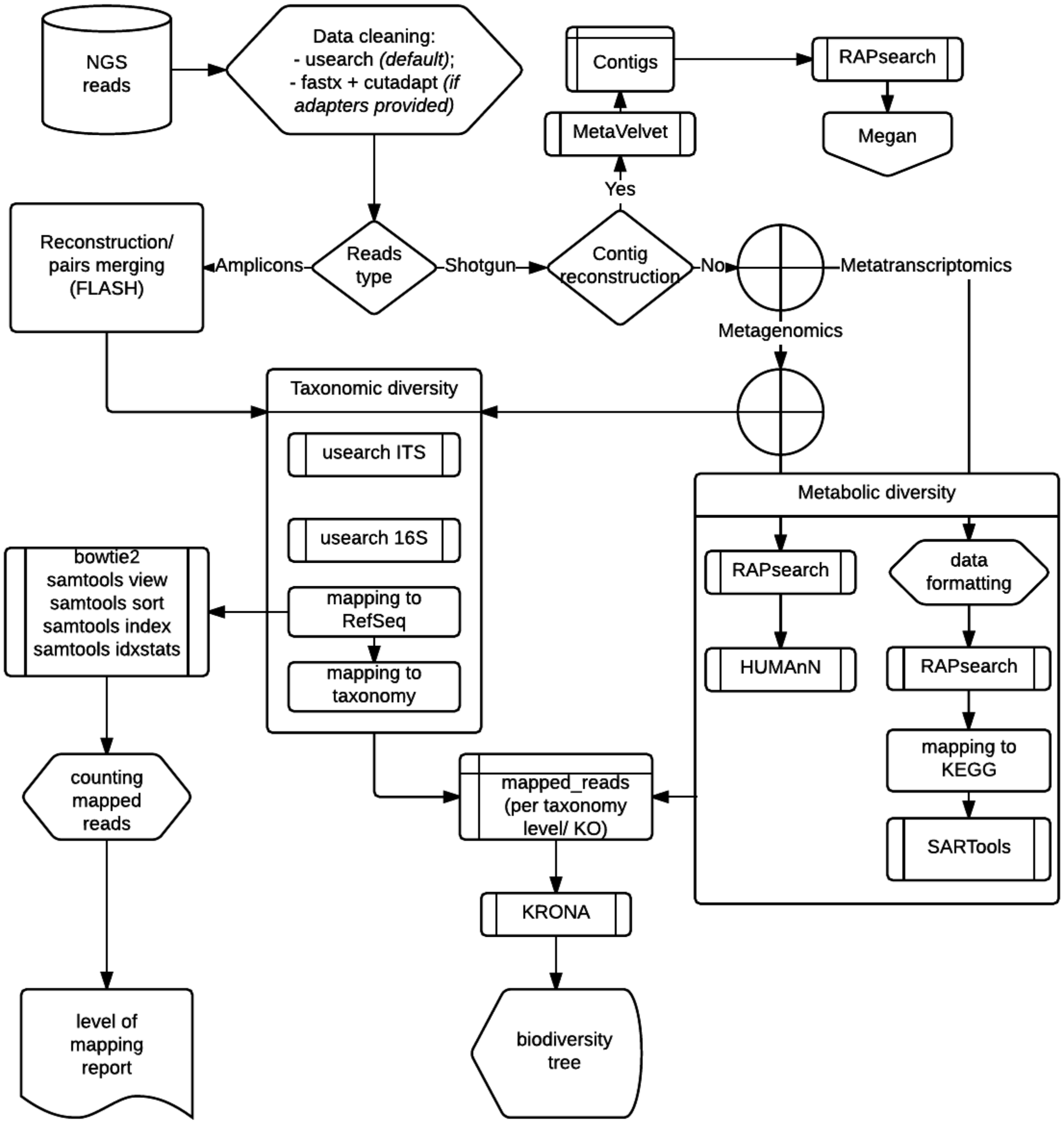


Fig. 1. Chart with bipype functionalities

Bipype accepts three types of inputs: amplicons, WGS (whole genome sequences) and metatranscriptomic data. Bipype may work with paired-end and single read sequences. For amplicon (prokaryotic or fungal) data, if needed, paired-end reads are merged and sequence reconstruction is performed. Reconstructed 16S or ITS sequences are searched in proper reference databases and hits to related taxonomic units are counted.

For WGS reads three paths are available. In one they will either be used for reconstruction of contigs, which will be further used for reference database search (outside of the pipeline). In the second they will be used directly in taxonomic diversity search, in third they will be compared to sequences related to metabolic pathways.

The biodiversity data at taxonomical and functional levels are similar in their tree structure - we provide display of results in a common form of html interactive comparative (multiple samples presented in single file) tree (Fig. 3c,). Bipype is a portable solution with a few dependencies, but to simplify analyses for the end-users, GUI presenting available analysis pathways and (optional!) parameters of each step was required. Below we present MetAMOS integration process, meeting this requirement. Bipype source code is available at: https://github.com/krassowski/bipype/

Integration with MetAMOS

As the number of customisable steps in MetAMOS is limited, we needed to overcome certain inflexibility of this system by inserting the whole bipype run in the first possible step and terminating the workflow afterwards. In order to do so, we added bipype to the list of tools used by metAMOS as a fake assembler. We added a script that created empty files required by metAMOS when it is checking if the assembly step finished successfully. Sample configuration file for such fake assembler is shown in Fig. 2. Command line options for metAMOS allow skipping most of the steps in any workflow, making it possible to finish immediately after assembly.

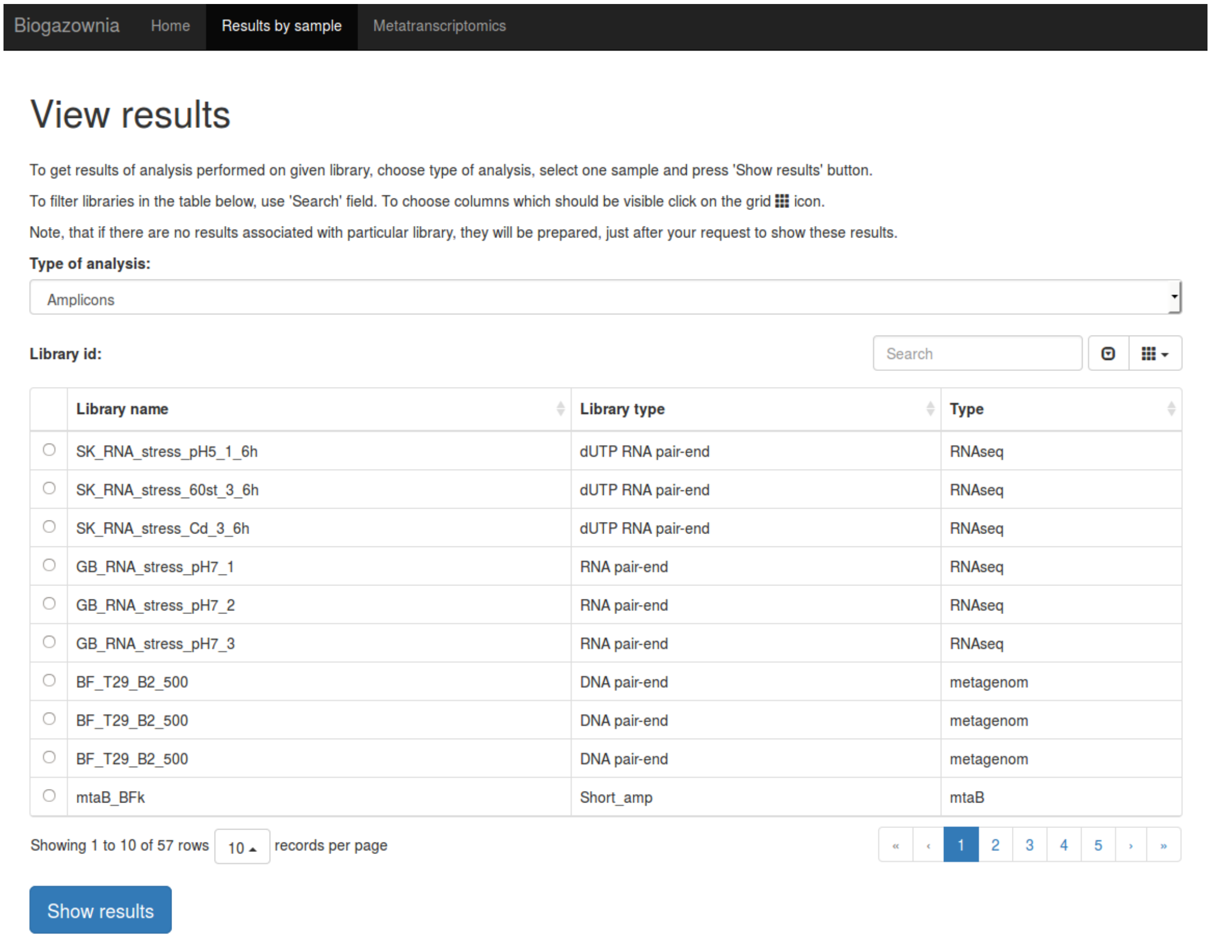
```
[CONFIG]
name bipype_amplicons
input FASTQ
scaffoldOutput [RUNDIR]/Scaffolds.fasta
location python/bipype
output 16S_ITS.krona
threads --threads
unpaired [FIRST]
paired_interleaved [FIRST] [SECOND]
paired [FIRST] [SECOND]
commands mkdir [RUNDIR] && \
        bipype --out_dir [RUNDIR] --cutadapt use_paths both --mode run -ITS -16S -ot ITS
16S --input [INPUT] && \
        bipype_cheat [RUNDIR]
```

Fig. 2. Sample configuration file

Web interface to MetAMOS

We have developed a web interface to analyses run on MetAMOS to let others not only access results but to run preconfigured analyses on their own. Fig. 3a shows the webpage where user can choose the type of analysis and library on which it would be run. Each type of analysis corresponds to one configuration file with appropriate bipype options. Status of analyses for each sample and workflow is stored in an SQL database. When user selects a workflow and a sample, web service checks if the corresponding job is finished and either shows the result (Fig. 3c), or its computation progress (Fig. 3b). If the job is not even started, the web service adds it to the database. A background service script periodically queries the database and runs queued jobs.

Additionally, we have added a simple results management, that is possibility to remove results in case re-computation is needed.



View results

To get results of analysis performed on given library, choose type of analysis, select one sample and press 'Show results' button.

To filter libraries in the table below, use 'Search' field. To choose columns which should be visible click on the grid ▦ icon.

Note, that if there are no results associated with particular library, they will be prepared, just after your request to show these results.

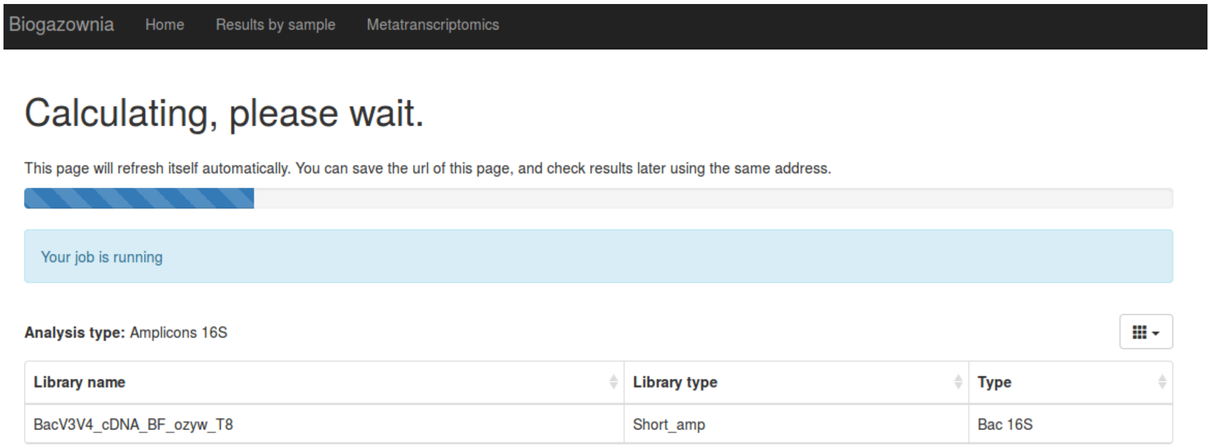
Type of analysis: Amplicons

Library id:

	Library name	Library type	Type
○	SK_RNA_stress_pH5_1_6h	dUTP RNA pair-end	RNAseq
○	SK_RNA_stress_60st_3_6h	dUTP RNA pair-end	RNAseq
○	SK_RNA_stress_Cd_3_6h	dUTP RNA pair-end	RNAseq
○	GB_RNA_stress_pH7_1	RNA pair-end	RNAseq
○	GB_RNA_stress_pH7_2	RNA pair-end	RNAseq
○	GB_RNA_stress_pH7_3	RNA pair-end	RNAseq
○	BF_T29_B2_500	DNA pair-end	metagenom
○	BF_T29_B2_500	DNA pair-end	metagenom
○	BF_T29_B2_500	DNA pair-end	metagenom
○	mtaB_BFk	Short_amp	mtaB

Showing 1 to 10 of 57 rows 10 records per page

Fig. 3a. screenshot of the interface: sample and workflow choice



Calculating, please wait.

This page will refresh itself automatically. You can save the url of this page, and check results later using the same address.

Your job is running

Analysis type: Amplicons 16S

Library name	Library type	Type
BacV3V4_cDNA_BF_ozyw_T8	Short_amp	Bac 16S

Fig 3b. screenshot of the interface: job progress

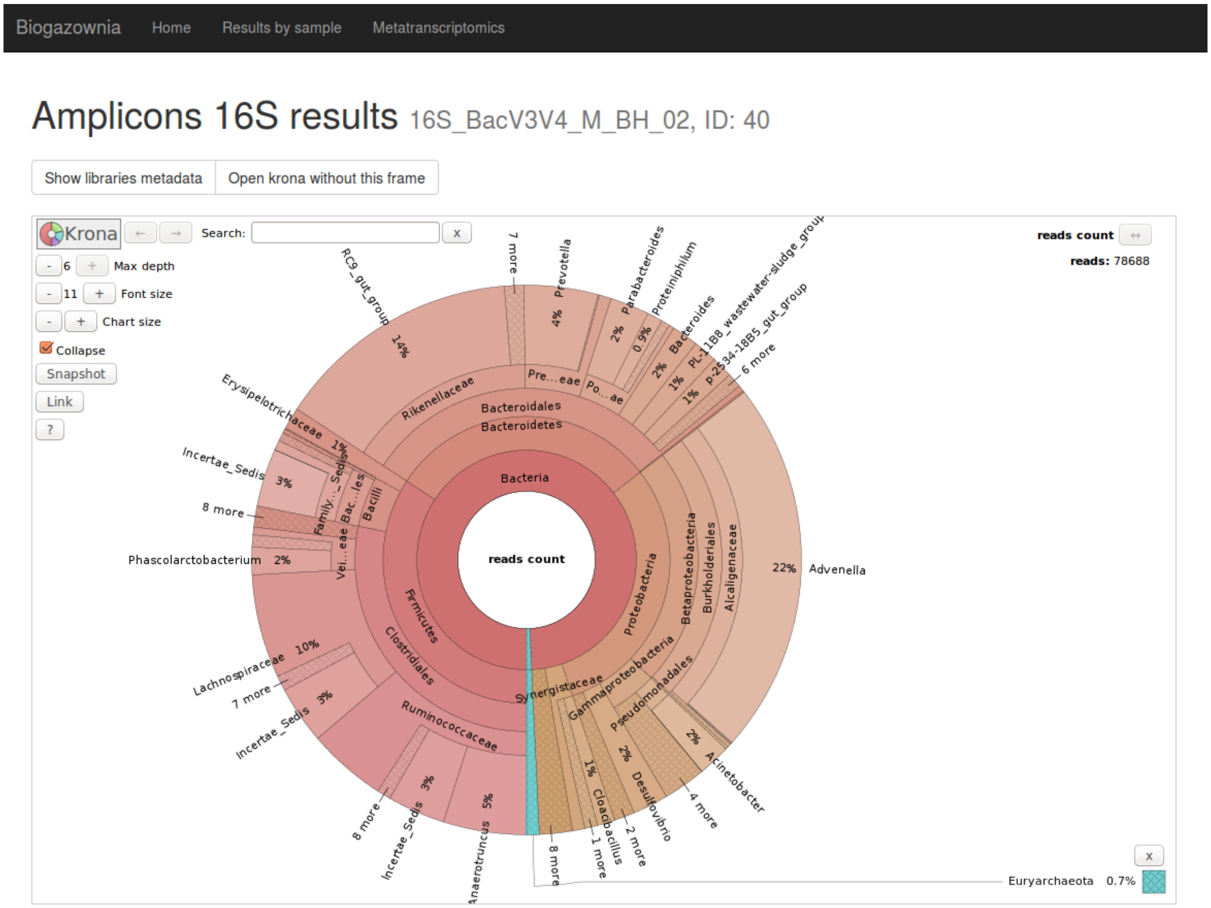


Fig 3c. screenshot of the interface: job results

Web interface source code is available at:

https://bitbucket.org/Serpens/metamos_web_interface/src

Materials and Methods

bipype is written in Python v2.7.3 and links following tools: fastx-toolkit v0.0.13 ("FASTX-Toolkit" 2016), usearch v7.0.959_i86linux32 (Edgar 2010), FLASH v1.2.7 (Magoč and Salzberg 2011), bowtie2 v2.2.4 (Langmead et al. 2009), samtools v0.1.18 (Li et al. 2009), RAPsearch 2.12_64bits (Zhao, Tang, and Ye 2012), MetaVelvet v1.2.01 (Namiki et al. 2012), HUMAnN v0.99 (Abubucker et al. 2012), SARTools v1.2.0 (Varet et al. 2015), KRONA 2.0 (Ondov, Bergman, and Phillippy 2013). Presented workflow was performed on Illumina reads with varying insert lengths, either provided in form of parameter or read from filename.

Databases used as reference include: SILVA rRNA database (Griffith, Malachi, and Griffith 2004; Quast et al. 2013), Unified system for the DNA based fungal species linked to the classification (UNITE) (Kõljalg et al. 2013), The Reference Sequence (RefSeq) Database (Griffith, Malachi, and Griffith 2004), NCBI Taxonomy Database (Wheeler 2004), Kyoto Encyclopaedia of Genes and Genomes database (Kanehisa et al. 2014),

Web interface is build using MetAMOS v.1.5rc3, Python v2.7.3, Django v1.4.5, jQuery v1.11 and Bootstrap v3.1.1. It was tested with SQLite v3.7.13 and Apache v2.2.22.

Conclusions

We present a way to develop time and resource efficient customizable pipeline serving for metagenomic and metatranscriptomic analyses, by extending MetAMOS workflow engine.

The proposed solution, besides abovementioned efficiencies presents following benefits:

- modularity allowing insertion of more custom tools and analyses (use of different assembly and display tools, search engines etc.),
- user-friendly web interface enabling easy access and steep learning curve for new team members, also enabling quick and repeated complete analyses by personnel not focused on software development,
- unified display methodology for both types of diversity data.

Our extensions are obviously available as open source under GNU GPLv2 license, allowing other researchers to build upon our work.

Acknowledgements

This work was supported by the National Centre for Research and Development (Poland) grant no. 177481 (PBS1/A8/3/2012).

References

Abubucker, Sahar, Nicola Segata, Johannes Goll, Alyxandria M. Schubert, Jacques Izard, Brandi L. Cantarel, Beltran Rodriguez-Mueller, et al. 2012. “Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome.” *PLoS Computational Biology* 8 (6): e1002358.

Caporaso, J. Gregory, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D. Bushman, Elizabeth K. Costello, Noah Fierer, et al. 2010. “QIIME Allows Analysis of High-Throughput Community Sequencing Data.” *Nature Methods* 7 (5): 335–36.

Edgar, Robert C. 2010. “Search and Clustering Orders of Magnitude Faster than BLAST.” *Bioinformatics* 26 (19): 2460–61.

“FASTX-Toolkit.” 2016. Accessed January 27. http://hannonlab.cshl.edu/fastx_toolkit/.

Griffith, Malachi, Griffith Malachi, and Obi L. Griffith. 2004. “RefSeq (the Reference Sequence Database).” In *Dictionary of Bioinformatics and Computational Biology*.

Kanehisa, Minoru, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. 2014. “Data, Information, Knowledge and Principle: Back to Metabolism in KEGG.” *Nucleic Acids Research* 42 (Database issue): D199–205.

Kõljalg, Urmas, R. Henrik Nilsson, Kessy Abarenkov, Leho Tedersoo, Andy F. S. Taylor, Mohammad Bahram, Scott T. Bates, et al. 2013. “Towards a Unified Paradigm for Sequence-Based Identification of Fungi.” *Molecular Ecology* 22 (21): 5271–77.

Langmead, Ben, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. 2009. “Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome.” *Genome Biology* 10 (3): R25.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics* 25 (16): 2078–79.

Magoč, Tanja, and Steven L. Salzberg. 2011. “FLASH: Fast Length Adjustment of Short Reads to Improve Genome Assemblies.” *Bioinformatics* 27 (21): 2957–63.

Namiki, Toshiaki, Tsuyoshi Hachiya, Hideaki Tanaka, and Yasubumi Sakakibara. 2012. “MetaVelvet: An Extension of Velvet Assembler to de Novo Metagenome Assembly from Short Sequence Reads.” *Nucleic Acids Research* 40 (20): e155.

Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2013. “Krona: Interactive Metagenomic Visualization in a Web Browser.” In *Encyclopedia of Metagenomics*, 1–8.

Quast, Christian, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. 2013. “The SILVA Ribosomal RNA Gene Database Project: Improved Data Processing and Web-Based Tools.” *Nucleic Acids Research* 41 (Database issue): D590–96.

Schloss, Patrick D., Sarah L. Westcott, Thomas Ryabin, Justine R. Hall, Martin Hartmann, Emily B. Hollister, Ryan A. Lesniewski, et al. 2009. “Introducing Mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities.” *Applied and Environmental Microbiology* 75 (23): 7537–41.

Treangen, Todd J., Sergey Koren, Daniel D. Sommer, Bo Liu, Irina Astrovskaya, Brian Ondov, Aaron E. Darling, Adam M. Phillippy, and Mihai Pop. 2013. “MetAMOS: A Modular and Open Source Metagenomic Assembly and Analysis Pipeline.” *Genome Biology* 14 (1): R2.

Varet, Hugo, Varet Hugo, Coppée Jean-Yves, and Dillies Marie-Agnès. 2015. “SARTools: A DESeq2- and edgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data.” doi:10.1101/021741.

Wheeler, D. L. 2004. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research* 33 (Database issue): D39–45.

Zhao, Yongan, Haixu Tang, and Yuzhen Ye. 2012. “RAPSearch2: A Fast and Memory-Efficient Protein Similarity Search Tool for next-Generation Sequencing Data.” *Bioinformatics* 28 (1): 125–26.