






Breaking Barriers with Bread: Using the Sourdough Starter Microbiome to Teach High-Throughput Sequencing Techniques

Benjamin H. Holt,^a  Alison Buchan,^b  Jennifer M. DeBruyn,^{b,c}  Heidi Goodrich-Blair,^b
Elizabeth McPherson,^b and Veronica A. Brown^{b,d}

^aDepartment of Ecology and Evolutionary Biology, University of Tennessee, Knoxville, Tennessee, USA

^bDepartment of Microbiology, University of Tennessee, Knoxville, Tennessee, USA

^cDepartment of Biosystems Engineering & Soil Science, University of Tennessee, Knoxville, Tennessee, USA

^dCenter for Environmental Biotechnology, University of Tennessee, Knoxville, Tennessee, USA

KEYWORDS high-throughput sequencing, sourdough, microbiome

INTRODUCTION

Researchers in life science disciplines are increasingly addressing biological questions using high-throughput sequencing (HTS) of nucleic acids. While technological advancements and reduced costs enable accessibility, many early career biologists lack familiarization with creating, handling, and analyzing large sequencing data sets (1). To help mitigate this barrier, academic institutions now offer classes that include HTS concepts and techniques, where students learn basic molecular biology skills, sequencing technologies, bioinformatics, and/or data analyses (2).

One of the most common applications of HTS is meta-amplicon sequencing of microbiomes—the assemblage of prokaryotes, fungi, and other microscopic eukaryotes associated with a particular environment. Meta-amplicon sequencing is ideal for HTS teaching experiences because it offers students greater hands-on opportunities than genome sequencing, it is generally more cost efficient than transcriptomics, and open source bioinformatic pipelines are available for data analysis. Here, we advocate that analysis of the microbiome of sourdough starters is effective for teaching HTS meta-amplicon sequencing, expanding student knowledge regarding contributions of microbes to everyday lives, and generating data that advance the understanding of sourdough microbiome community structure. Sourdough starters comprise flour, water or milk, and a consortium of “wild” microbes used to leaven bread via

CO₂ production (3). These consortia are relatively simple communities, containing only a few fungal and prokaryote members, offering two key advantages: (i) small data sets to facilitate analysis and (ii) easy identification of contaminants and sequencing errors. Nutrient source, storage, and geographic regions contribute to differences in microbial composition among starters (4, 5), allowing students to apply bioinformatics and statistics to analyze these differences. We have incorporated HTS meta-amplicon sequencing of sourdough microbiomes in an upper-level microbiology class at the University of Tennessee, Knoxville (UTK), where students carry out the entire HTS meta-amplicon process. The class has been held twice, with a total of 14 senior-level undergraduate and graduate students from different backgrounds, including life and agricultural sciences.

PROCEDURE

The workflow for 16S/ITS meta-amplicon sequencing follows the Illumina 16S Metagenomic Library Preparation protocol for Illumina MiSeq and contains 4 steps: library preparation, sequencing, bioinformatics, and analysis. Library preparation involves DNA extraction and a two-step polymerase chain-reaction (PCR) indexing method with two post-PCR purification steps. Following sequencing, FASTQ files are run through a bioinformatic pipeline, using students’ personal computers or campus computer labs.

SAFETY ISSUES

Students and instructors wear lab coats and nitrile gloves during laboratory procedures: gloves prevent contamination of samples; lab coats are standard protocol for microbiology labs. Ethidium bromide (EtBr) is used during gel electrophoresis; EtBr is mutagenic and can cause severe skin, eye, and lung

Editor Dave J. Westenberg, Missouri University of Science and Technology

Address correspondence to Department of Microbiology and Center for Environmental Biotechnology, University of Tennessee, Knoxville, Tennessee, USA. E-mail: vabrown@utk.edu.

The authors declare no conflict of interest.

Received: 8 November 2021, Accepted: 26 April 2022,

Published: 16 May 2022

		Learning Outcomes	Tips to Enhance Student Learning
Students	DNA extraction 1	Isolate high-quality DNA from biologically and chemically complex environmental samples.	A. Remind students to save FINAL elution from spin columns. B. Sample labeling consistency throughout process is essential.
	PCR with two-step Illumina protocol adapters 2	Amplify DNA through the biological process of polymerase chain amplification and explain the molecular steps in this process.	A. Remind students that thorough mixing of reactions is important. B. Remind students of sterile technique and positive and negative controls.
	Confirmation by 2% agarose gel electrophoresis 3	Compare and contrast different methods of visualizing DNA through gel electrophoresis.	A. Remind students to track and document sample loading order. B. Gauge amplification success and provide instructor samples if needed.
	Magnetic bead cleanup using 1.5 ml tubes 4	Compare and contrast DNA purification and recovery methods.	A. Protocol is adapted from the typical 96-well format. B. Remind students to keep tubes on magnetic racks until told to remove them.
	Index PCR with premixed forward and reverse indexes 5	Explain the two-step PCR method for attaching indexes/barcodes to PCR products with the appropriate adapters.	Remind students to track and document which indexes are added to which samples.
	Magnetic bead cleanup using 1.5 ml tubes 6	Discuss the different purposes for DNA purifications.	Remind students to keep tubes on magnetic racks until told to remove them.
	DNA quantification with Nanodrop spectrophotometer 7	Compare and contrast DNA quantification methodologies.	Other methods of quantification, such as fluorometer or Qubit, can teach the same concepts here, based on available equipment.
Experienced User/Instructor	Sample pooling 8	Calculate sample normalization concentrations and volumes. Describe the principles of barcoding and pooled sample tracking in high throughput sequencing.	Reinforce the concepts of multiplexing, demultiplexing, and indexing.
	Quantification and visualization with Agilent Bioanalyzer 9	Describe the expectations of final quality control in amplicon high-throughput sequencing.	Students should see the Bioanalyzer, either by watching it be prepared or through videos, since the instrument must be loaded by an experienced user.
	Final library preparation 10	Describe the concept of Sequencing by Synthesis. Discuss the importance of nucleotide diversity in amplicon high-throughput sequencing.	Students should be walked through the final library preparations done by an expert to minimize the Black Box feel of the steps they cannot perform themselves.
	Sample loading on Illumina MiSeq 11	Compare and contrast different high-throughput sequencing technologies.	Students should see the MiSeq, either by watching it be prepared or through videos, since the instrument must be loaded by an experienced user.

FIG 1. Outline of the lab work used in the high-throughput sequencing (HTS) class, which closely follows the Illumina 16S Metagenomic Library Preparation protocol.

irritation. To minimize contact, a staining box with diluted EtBr, rather than adding concentrated EtBr directly to gels, is used. Alternative nucleic acid stains (e.g., Midori Green) could be used to mitigate this risk.

METHODS

Forty-one sourdough starter samples were solicited from UTK representatives. Samples were stored at -20°C until use,

		Learning Outcomes	Tips to Enhance Student Learning
Demultiplexing	1	Describe the concepts of separating samples by indexes/barcodes.	Done automatically by the UT Genomics Core MiSeq, but not by all sequencing instruments, so this may need to be performed first.
Primer removal	2	Identify the segments of DNA that make up a sequencing read product, including initial PCR primers.	Discuss the expected length with students so they know if they have done this step correctly.
Quality filtering and trimming	3	Assess quality of sequencing product.	Discuss where and why quality decreases and how decisions made here will influence all downstream processes.
Merge paired-end sequences	4	Discuss the importance of sufficient overlap in forward and reverse sequences.	Discuss the expected length with students so they know if they have done this step correctly.
Remove Chimeras	5	Describe chimeras and how to remove them from the samples.	Discuss what chimeras are, how they can arise, how they are identified, and why they should be removed.
Taxonomic classification	6	Compare and contrast different databases used to assign taxonomy to OTUs/ASVs.	Silva is the current preferred database for 16S bacterial analysis and is easily accessible to students.
Post bioinformatic filtering	7	Discuss the importance of removing singletons, mitochondrial sequences and any specific additional filtering from the dataset.	Discuss how to deal with contamination and rare species.
Data analysis	8	Compare and contrast different types of statistical analyses to address project-specific questions.	Examples include α and β diversities, compositional visualizations, and hypothesis testing of research questions.
Oral presentation of results	9	Illustrate results in a short PowerPoint presentation and discussion amongst class.	Encourage students to provide each other with positive, constructive feedback.
Written presentation of results	10	Summarize results in a short paper, incorporating comments from the presentations, with heavy emphasis on the methods.	Having students turn in a rough draft of this paper as a homework assignment allows students to make edits and have more accountability.

FIG 2. Outline of Bioinformatic steps used in the high-throughput sequencing class. Major hinderances include the wide range of computational experience and variety of operating systems. In Step 6, OTU refers to operational taxonomic unit, while ASV refers to amplicon sequencing variant, both of which are ways of clustering sequence variants.

and metadata, such as flour type and starter age, were recorded. To compare between students and a more experienced user, the instructor replicated every sample in parallel. DNA was extracted using the DNeasy PowerSoil Kit (Qiagen). Extracted DNA was amplified using fungal (ITS

[6]) and prokaryotic (16S rRNA [7]) primers. Libraries were prepared and sequenced on the Illumina MiSeq at the UTK Genomics Core (Fig. 1). Students included extraction and PCR blanks consisting of water in the place of template. ZymoBIOMICS Microbial Community Standard (Zymo

TABLE I
Results of sourdough microbiome sequencing libraries from 14 students over two semesters

Student	Good/total ^a	% reads detected in instructor's pair ^b	# SV not in instructor's sample (potential contaminants) ^{b,c}
1	1/2	99.4	41.0
2	1/3	99.2	28.0
3	6/6	98.3	48.3
4	4/4	98.6	46.5
5	8/8	99.5	17.9
6	7/8	99.6	26.3
7	6/6	98.7	45.3
8	2/2	87.2	23.5
9	3/3	99.4	41.7
10	6/6	98.9	25.5
11	5/5	98.6	50.2
12	1/3	99.4	55.0
13	0/4	NA—sample mix-up	NA
14	0/6	NA—amplifications failed	NA

^aA sample was deemed “good” if 75% or more reads were detected in the instructor's paired sample.

^bMeans are reported if the denominator of column two is greater than one.

^cSV refers to sequence variants, and the mean number of SV present in the student sample but absent from the instructor's samples is reported. Importantly, while students often had sequences not detected in the paired instructor sample, these constituted a small proportion of total reads retained.

Research, Irvine, CA) served as a positive control for lab work and bioinformatics.

Sequencing reads are automatically demultiplexed as they are processed from the UTK MiSeq such that each student receives their individual, respective sample sequences from the pooled sequencing reaction. Students then use Cutadapt to remove primer sequences before moving to an existing DADA2 bioinformatic pipeline (Steps 3–5, Fig. 2; Text S1 in the supplemental material) in R v4.0.3 (8, 9). R has gained popularity in data science due to a large support community, highly customizable syntax, open-source availability, ease of installation, and plethora of available packages (10, 11). The RStudio IDE makes visualizations and code annotations easy for students to grasp and comprehend. After taxonomic identification of the resulting sequences is complete, students discuss the biological significance of their results and are encouraged to develop their own hypotheses for statistical analysis. Sequencing results are shared among students for diversity analyses. This portion could be easily expanded and incorporated into future classes focused on statistical analyses of HTS data. Students visualize quality profiles of sequencing reads to discuss the expected quality in HTS results.

Student experimental outcomes were evaluated by assessing the similarities between student and instructor samples (Table I). Sequence contaminants associated with the instructor's samples were removed using the “decontam” package in R to create an idealized sample (12, 13). Sequence variants in student samples absent from the instructor sample were flagged as contaminants. A student sample was deemed “good” if >75% of the reads matched reads in the instructor's replicate. In 66 pairs

of student–instructor samples, 10 samples, derived from 2 of 14 students, were dropped prior to the bioinformatic pipeline due to lab work issues (mixing up labels or lab errors). Six samples failed to meet the 75% match criterion, and 2 samples failed to provide quality sequence reads for both instructor and student. Of the 48 samples meeting the “good” criterion, the mean proportion of reads shared with the instructor was 98%, indicating most (12 of 14) students effectively captured the sourdough starter microbiome of at least one of their samples. These 48 samples were used to assess α and β diversity questions (Text S1). Since molecular work often involves failed reactions, students were not graded on accuracy of results, and troubleshooting of issues was discussed in class.

CONCLUSIONS

Sourdough starter microbiomes are an effective model system for teaching meta-amplicon sequencing and analysis of derived data. As these systems represent relatively low diversity microbiomes, contaminants are readily identifiable, allowing for easy assessment of student technical success. Furthermore, the bioinformatic pipeline and data analyses can be completed relatively quickly on personal laptops. Moreover, genetic variation is evident in these populations, providing opportunities for students to consider how seemingly subtle differences in nucleotide composition results in significant differences in community composition but not overall function. Students can easily relate to the sourdough starter as a part of everyday life, and this

connection facilitates hypothesis development and outcomes interpretation of their microbiome sequencing projects.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 1.1 MB.

ACKNOWLEDGMENTS

This work was, in part, supported by funds provided by the University of Tennessee, Knoxville College of Arts and Sciences. A.B. is partially supported by an NSF award (OCE-1357242).

None of the authors have any financial or personal relationships that could inappropriately bias or compromise our actions. Institutional funding supported the work, but the supporting offices/individuals had no involvement in the design, collection, analysis, or interpretation of the data; nor did they write any of the manuscript; nor were they involved in the decision to submit this report for publication.

REFERENCES

1. Marx V. 2013. The big challenges of big data. *Nature* 498:255–260. <https://doi.org/10.1038/498255a>.
2. Edwards RA, Haggerty JM, Cassman N, Busch JC, Aguinaldo K, Chinta S, Vaughn MH, Morey R, Harkins TT, Teiling C, Fredrikson K, Dinsdale EA. 2013. Microbes, metagenomes and marine mammals: enabling the next generation of scientist to enter the genomic era. *BMC Genomics* 14:600. <https://doi.org/10.1186/1471-2164-14-600>.
3. Brandt MJ. 2007. Sourdough products for convenient use in baking. *Food Microbiol* 24:161–164. <https://doi.org/10.1016/j.fm.2006.07.010>.
4. Liu X, Zhou M, Jiaxin C, Luo Y, Ye F, Jiao S, Hu X, Zhang J, Lü X. 2018. Bacterial diversity in traditional sourdough from different regions in China. *LWT Food Sci Technol* 96:251–259. <https://doi.org/10.1016/j.lwt.2018.05.023>.
5. Landis EA, Oliverio AM, McKenney EA, Nichols LM, Kfoury N, Biango-Daniels M, Shell LK, Madden AA, Shapiro L, Sakunala S, Drake K, Robbat A, Booker M, Dunn RR, Fierer N, Wolfe BE. 2021. The diversity and function of sourdough starter microbiomes. *Elife* 10:e61644. <https://doi.org/10.7554/eLife.61644>.
6. Cregger MA, Veach AM, Yang ZK, Crouch MJ, Vilgalys R, Tuskan GA, Schadt CW. 2018. The *Populus* holobiont: dissecting the effects of plant niches and genotype on the microbiome. *Microbiome* 6:31. <https://doi.org/10.1186/s40168-018-0413-8>.
7. Klindworth A, Priesse E, Schweer T, Peplies J, Quast C, Horn M, Glöckner FO. 2013. Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* 41:e1. <https://doi.org/10.1093/nar/gks808>.
8. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 13:581–583. <https://doi.org/10.1038/nmeth.3869>.
9. R Core Team. 2020. R: a language and environment for statistical computing. R Foundation for Statistical Computing. <https://www.R-project.org/>.
10. Gentleman R. 2009. R programming for bioinformatics. CRC Press, Boca Raton, FL.
11. Brittain J, Cendon M, Nizzi J, Pleis J. 2018. Data Scientist's Analysis Toolbox: Comparison of Python, R, and SAS Performance. *SMU Data Sci Rev* 1:7.
12. Davis NM, Proctor DM, Holmes SP, Relman DA, Callahan BJ. 2018. Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* 6:226. <https://doi.org/10.1186/s40168-018-0605-2>.
13. McMurdie PJ, Holmes SP. 2013. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. <https://doi.org/10.1371/journal.pone.0061217>.