Supplemental Methods and Results (Holt et al.)

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

Methods: Bioinformatics

Since reads coming off the Illumina MiSeq at the University of Tennessee Genomics Core are already demultiplexed, we were able to skip a manual demultiplexing step and move directly to primer removal. We removed primers using Cutadapt before entering the DADA2 pipeline in R, beginning with the existing DADA2 pipeline tutorial, and modifying if needed (1, 2). While it generally runs slower than command line, its open-source availability for Mac and PC, ease of installation, and plethora of packages makes usage on students’ laptops easier than setting up virtual machines (3). Students visualized quality profiles of forward and reverse reads and decided to trim forward reads at 250bp and reverse reads at 225bp for 16S. ITS reads were not trimmed, as is common practice for fungal sequences with variable lengths. The filtering parameter (maxEE), which is the maximum expected error rate, was set to c(2,3) for both 16S and ITS. We encouraged students to manipulate this parameter to see how reads were lost with more stringent and more relaxed parameters for trim lengths and maxEE. Taxonomic assignments for 16S were carried out using the IdTaxa algorithm on the Silva 138 classifier, while ITS used the naïve Bayesian classifier on the Unite ITS reference database (4, 5, 6).

Following taxonomic classification, singletons and reads assigned to mitochondria and chloroplast were removed after Zymo mock community analyses were done. Contaminants were then removed using the ‘decontam’ package in R (7). We used the ‘prevalence’ method with a threshold of 0.5, meaning any taxa more prevalent in ‘blanks’ than ‘samples’ was noted as a contaminant. Since students only did one or two blanks, we did this for the instructor samples only and thus used these uncontaminated samples as an ‘idealized’ sample to identify student contaminants. For ITS samples, this resulted in the removal of six taxa, with relative abundance loss ranging from 0.001 to 0.057 from the instructor’s samples. Instructor’s 16S rRNA gen (hereafter “16S”) sequences had no contaminants identified. We then assessed how student samples compared to an idealized sample by tracking the number of reads not detected in the instructor’s samples.

Methods: Statistics

Due to our paired sample design (instructor vs student) and each student having a different sourdough starter sample, we statistically compared α diversity using a paired t-test. Since pairwise dissimilarities are rarely zero for two different samples, we assessed pairwise β diversity through a multivariate Levene’s test of group dispersion, and a distance-based redundancy analysis (dbRDA) for each of three β diversity metrics (Sørensen, Horn, and Morisita Horn dissimilarities), with significance tested using an anova, or anova.cca respectively using the R package vegan (8). Average α and β diversities used for analyses were estimated from each of 1000 rarefactions to standardize sequencing effort.

Results: ITS
For our ITS dataset, after removing contaminants and filtering as described in the main text, we obtained 3,548,011 non-chimeric reads mapped to 931 sequence variants in 20 paired samples. Total run time on a 2015 MacBook Pro (8 GB of RAM, 2.7 GHz Dual-Core) was around 10 hours for ITS, with over 7 hours dedicated to taxonomic classifying.

We rarefied our ITS data to 6695 reads and estimated diversity from each of 1000 rarefactions using Hill Numbers (9). Our results indicate a significant decline in effective species from \( q=0 \) to \( q=1 \) \( (p<0.001) \), implying uneven abundances within our samples (Fig. S1). There was no significant difference between \( q=1 \) and \( q=2 \) comparisons, indicating samples are dominated by a small number of taxa not significantly differentiated by a higher diversity order. When comparing student vs instructor, paired t-test indicated significant differences in each order of diversity comparison \( (p<0.01, p<0.05, p<0.05) \). However, in most cases, students showed strong similarities to the instructor, apart from a few outliers (Fig. S2).

To compare compositional changes, we estimated pairwise dissimilarities corresponding to the three Hill Numbers for each of 1000 rarefactions and used the average pairwise distances for statistical comparisons. We first looked at the dispersion of group variances between the instructor’s samples and student’s samples. In every comparison, there were no significant differences of group dispersion between the two groups \( (p=0.36, p=0.68, p=0.74) \). To isolate and only test for student vs instructor, we parcialled out the effects of sample source and found only the binary based dissimilarity (Sørensen) to be significant between the instructor and students, while dissimilarities incorporating relative abundances were not \( (p=0.011, p=0.201, p=0.175) \). The average pairwise paired difference between the student and the instructor was 0.405, 0.089, and 0.116 for each dissimilarity, respectively. Together, these results imply differences between instructor and student are most notable when using a binary dissimilarity, but all students effectively captured the most abundant taxa. Furthermore, variance in fungal communities were best explained by differences in sample origin \( (e.g.: \) where the sample came from).

**Results: 16S**

Our 16S dataset contained 7,972,332 reads mapped to 2446 sequence variants in 30 paired samples. Total runtime for bioinformatics on all samples was just over 4 hours using the same computer as above for the ITS dataset. We then removed 256 taxa identified as chloroplast, 299 taxa identified as mitochondria, and 29 singletons, decreasing our total reads to 5,263,125 in 1863 sequence variants. Following the same process with contaminants as described above, twelve samples were removed, bringing our total to 26 pairs.

Alpha diversity was estimated over 1000 rarefactions at a rarefaction level of 12958 reads. The 16S microbiome generally lacked diversity, with only around 45 taxa and declining significantly at each order of \( q \) (Fig. S3). Paired t-test found no significant differences between students’ and instructor’s samples \( (p=0.12, p=0.16, p=0.15) \), and distribution of diversity differences showed students effectively captured an average \( \alpha \) diversity estimation similar to the instructor, with only a few instances of more or less diversity within the sample (Fig. S4).

Homogeneity of group dispersion was not significant when comparing dispersion of instructor vs student samples in any dissimilarity metric \( (p=0.525, p=0.9376, p=0.944) \). Distance-based
redundancy analysis and variance partitioning found no significant differences in communities between students and the instructor (p=0.162, p=0.544, p=0.759). The average turnover between the same sample performed by either student or instructor was 0.55 (Sørensen), 0.055 (Horn), and 0.056 (Morista-Horn). Similar to the results with ITS, our 16S results show differences between students and the instructor are most apparent at a binary based \(\alpha\) or \(\beta\) metric but incorporating relative abundance into estimations indicates students effectively capture \(\alpha\) and \(\beta\) diversity similar to the instructor.

**Results: Zymo Mock Community**

We used a commercially available reference (ZymoBiotics Microbial Community DNA Standard D6305 from Zymo Research Corp) to evaluate how PCR biases and contamination could affect downstream analyses. We analyzed the Zymo reference communities before any post bioinformatic cleanup. If uneven sequencing had occurred, we would expect relative abundance of reads in assigned student Zymo reference communities to deviate from expected ratios of the expected community. We focused on exact sequence matches to the control community and reads not assigned to an exact match were assigned as ‘Other’. Of the thirteen students who extracted and sequenced a control community, all students had over 95% of sequences that were perfect matches to a reference strain in proportions expected (**Fig. S5**).
Supplemental References

1. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. 7. *Nat Methods* 13:581–583.
2. R Core Team. 2020. R: A language and environment for statistical computing. (R Foundation for Statistical Computing). [https://www.R-project.org/](https://www.R-project.org/).
3. 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* 1:20.
4. Nilsson RH, Larsson K-H, Taylor AFS, Bengtsson-Palme J, Jeppesen TS, Schigel D, Kennedy P, Picard K, Glöckner FO, Tedersoo L, Saar I, Köljalg U, Abarenkov K. 2019. The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res* 47:D259–D264.
5. Murali A, Bhargava A, Wright ES. 2018. IDTAXA: a novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome* 6:140.
6. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Pfeiffer D, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* 41:D590–D596.
7. 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.
8. Oksanen J, Blanchet G, Friendly M, Legendre P, McGlinn PR, O’Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2019. vegan: Community Ecology Package. (R package version 2.5-6). [https://CRAN.R-project.org/package=vegan](https://CRAN.R-project.org/package=vegan).
9. Hill MO. 1973. Diversity and Evenness: A Unifying Notation and Its Consequences. *Ecology* 54:427–432.
Supplemental Figures (Holt et al.)

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

Supplemental Figure 1. Boxplot for three alpha diversity metrics for ITS samples. Point colors correspond to paired comparisons between instructor and student. Tukey HSD test shows a decrease in alpha diversity from richness to q=1, but not between q=1 and q=2.
Supplemental Figure 2. Differences in diversity between students and the instructor. Paired $t$-test found significant differences in each comparison, which can likely be attributed to outliers.
Supplemental Figure 3. Three metrics of alpha diversity for 16S samples. Colors correspond to pairs.
Supplemental Figure 4. Histogram of differences between students and the instructor for the three diversity metrics. We compared differences with a paired-t test, which was not significant at any metric.
Supplemental Figure 5. Student’s Zymo mock communities compared to expected ratios. This was on raw data, so singletons, chloroplast and mitochondria still remain. Other is described as any sequences not being a perfect match to the Zymo Community. Each sample had around 4-5% of ‘Other’.