RESEARCH LETTER – Professional Development

A town on fire! Integrating 16S rRNA gene amplicon analyses into an undergraduate microbiology lecture class

Tammy C. Tobin[1,*,†] and Ashley Shade[2,‡]

[1]Biology Department, Susquehanna University, 514 University Avenue, Selinsgrove, PA 17870 USA and
[2]Departments of Microbiology and Molecular Genetics and Plant, Soil and Microbial Sciences, Program in Ecology, Evolutionary Biology and Behavior, and the Plant Resilience Institute, Michigan State University, 567 Wilson Road, 2215 BPS, East Lansing, MI 48824, USA

*Corresponding author: Biology Department, Susquehanna University, 514 University Avenue, Selinsgrove, PA 17870, USA. Tel: +570-372-4067; Fax: +570-372-2753; E-mail: tobinjan@susqu.edu

One sentence summary: Integration and assessment of an undergraduate bioinformatics module to study microbial ecology.

Editor: Alejandra Bravo
†Tammy C. Tobin, http://orcid.org/0000-0003-0619-1533
‡Ashley Shade, http://orcid.org/0000-0002-7189-3067

ABSTRACT

Microbiology increasingly relies upon bioinformatics to understand complex microbial interactions. Nevertheless, biology undergraduates often lack the basic quantitative and computer-based skills required for bioinformatics analyses. To address these issues, the course module 'A Town on Fire! 16S rRNA Gene Amplicon Analysis of Microbial Communities Overlying the Centralia, PA Mine Fire' was developed for an undergraduate microbiology lecture course. In this module, microbiology students used Quantitative Insights into Microbial Ecology to perform taxonomic, phylogenetic and statistical analyses on bacterial communities from three hot mine fire-impacted surface soils using 16S rRNA gene amplicon sequences. Pre- and post-module assessment data for each of 2 years were compiled, and indirect assessment indicated that students' confidence regarding their ability to perform bioinformatics analyses, as well as their ability to interpret bioinformatics data both increased, as did their enthusiasm for bioinformatics. Direct assessment demonstrated that students' understanding of topics that they actually used in the module, such as the statistical analyses that underlie bioinformatics investigations and the ability to infer phylogenetic relationships, improved during the module, but that their underlying understanding of techniques that they did not directly perform, such as sequencing and library construction, did not.

Keywords: fire; 16S rRNA; metagenomics; undergraduate; microbiology; lecture

INTRODUCTION

Typical introductory microbiology courses must cover a broad spectrum of topics, ranging from human health and disease to microbial ecology. Despite the fact that bioinformatics analyses are becoming central to all of these areas (Shade and Teal 2015), and that multiple STEM publications have identified quantitative skills as central to STEM education (National Research Council BIO2010 2003; Vision and Change 2011), the incorporation of bioinformatics into undergraduate curricula is still far from universal. Significant barriers to undergraduate implementation include the cost of sequencing materials

and the lack of training amongst older faculty (Buonaccorsi *et al.* 2011). While consortia such as the Network for Integrating Bioinformatics into Life Sciences Education (NIBSLE, https://qubeshub.org/groups/niblse) and the Genomics Consortium on Active Teaching—Sequencing Group (GCAT-SEEK, Buonaccorsi *et al.* 2011, 2014) are attempting to address these barriers, very few of the published products from these endeavors involve microbial ecology. The course module 'A Town on Fire!' was developed to teach bioinformatics to undergraduate microbiology students and to provide a resource for other undergraduate microbiology faculty.

In 'A Town on Fire!' students use bioinformatics analysis of 16S rRNA gene amplicon data to test hypotheses regarding the ways in which soil bacterial communities impact, and are impacted, by a very real, devastating and ongoing underground anthracite coal mine fire in Centralia, PA. This fire was ignited by burning trash in 1962, has expanded through the mine tunnels and coal seams under the town at a rate of approximately 3–7 m^2 per year since that time, and currently underlies at least 150 acres (Elick 2011). It is likely to burn for at least another 50 years (Nolter and Vice 2004). As the fire expands into new areas, it can have a dramatic impact on the surface environment. Hot combustion gases containing high levels of oxidized sulfur, nitrogen and other chemicals rise to the surface through vents, or 'anthracite smokers', and heat the surface soils, so temperatures of over 65°C are common. Additionally, as the gases rise they cool and precipitate their dissolved chemicals into the surrounding soils. Thus, extremely high levels of elemental sulfur, sulfate, ammonium and nitrate exist in the mine fire-affected soils, making Centralia an ideal habitat for thermophilic sulfur and nitrogen cycling bacteria (Tobin-Janzen *et al.* 2005; Janzen and Tobin-Janzen 2008).

While ideal for thermophiles, the high levels of noxious gases and frequent ground collapses make Centralia extremely dangerous for long-term human habitation. Thus, the entire town of over 1800 residents was relocated in the late 1980s, with only a few residents remaining today. The barren landscape that remains presents an excellent laboratory for the analysis of thermophilic bacterial communities and their environmental roles. In this module, teams of microbiology students work collaboratively as they learn about this fascinating extreme environment and the molecular biology methods used to study its resident microbes, and propose hypotheses as to which bacterial taxa (typically genera) are most likely to be present in Centralia. They test those hypotheses using bioinformatics analysis (Quantitative Insights into Microbial Ecology [QIIME] v1.9.1) of the 16S rRNA data (Caporaso *et al.* 2010b). Ultimately, the student teams write final papers that describe their research question and approach, and discuss the probable ecological impact of a single species from their chosen taxon in the Centralia soil environment.

To determine if this course module was an effective way to introduce bioinformatics concepts to undergraduates, all students enrolled in the course during each of two consecutive years completed an online pre- and post-course assessment instrument that used indirect and direct methods to measure student skills and attitudes. The results for the 2 years were compiled.

MATERIALS AND METHODS

Soil sample collection and sequencing

Three surface soil samples (without replicates) were collected from 0–10 cm boreholes in fire-affected sites as part of a small-scale pilot study during the spring of 2013 (Fig. 1). Soil temperatures were measured using an insertion probe at the bottom of each borehole. Soil chemical analyses (pH, total ammonium, nitrate, sulfur and sulfate) were performed as previously described (Tobin-Janzen *et al.* 2005; Janzen and Tobin-Janzen 2008). The samples were transported on ice to the laboratory at Susquehanna University, and stored at −80°C. Genomic DNA was extracted from all soil samples using the MoBio PowerSoil Kit was sent to the Penn State Genomics Core Sequencing facility (http://www.huck.psu.edu/content/instrumentation-facilities/genomics-core-facility) for 16S rRNA gene amplicon (V1 + V2 region) library preparation using primers 8F (Edwards *et al.* 1989) and 806R (Nikkari *et al.* 2002), and Roche 454 pyrosequencing per their standard protocols. The resulting sequence and quality files were concatenated using QIIME v1.9.0 prior to use by students in the course module. The sequence (.fas), quality (.qual) and metadata files needed for this module can be accessed via the GCAT-SEEK website (http://gcat-seek.weebly.com).

Course structure

'A Town on Fire!' was taught during four separate 1.5-h class periods interspersed throughout the 4-week microbial ecology unit of our undergraduate general microbiology class. The class enrolled an average of 24 junior and senior biology, biochemistry, ecology and neuroscience majors. While it was typical for all of the students to have had at least one biology and chemistry course, student backgrounds in more advanced biology, chemistry and ecology varied widely, and none had any previous bioinformatics experience beyond BLAST searches. During each class period, the instructor gave a brief introductory lecture, after which students worked in teams of four to complete a guided tutorial that reinforced lecture materials and led them through the bioinformatics analyses to be used that day. Short answer questions interspersed throughout the handouts were used to help guage student understanding. The student handouts for each class contain the bioinformatics workflow, and are available in the supplementary materials (Supplementary files 1–4, Supporting Information).

Assessment of student learning and attitudes

Student attitudes and skills were assessed using their pre- and post-course responses to the GCAT-SEEK (http://gcat-seek.weebly.com/) survey, as well as their performance on a final team paper. Unpaired *t*-tests were used to determine the significance of improvements between pre- and post-test GCAT-SEEK survey responses. The GCAT-SEEK protocol (IRB proposal 56-2013) was designated as 'Exempt, Category 2B' by the Juniata College IRB as per the Code of Federal Regulations (CFR 46.101).

RESULTS AND DISCUSSION

Class one: the history and environmental impacts of the Centralia mine fire

During the first class in this module students learned about the Centralia mine fire history, its environmental impacts, and how their 16S rRNA gene amplicon sequence data were generated. The fire started in 1962 when a trash fire ignited a surface-accessible coal seam near the Odd Fellows Cemetery in Centralia, Pennsylvania. Several unsuccessful attempts were made to extinguish the fire over the next two decades. These attempts were hampered by a bewildering array of political

Figure 1. The Centralia, Pennsylvania, mine fire has dramatically impacted the surface environment. (**A**) Aerial view of the Centralia Mine Fire in July, 2007. Bare ground indicates mine fire-affected sites. The location from which soil samples were collected is indicated by the arrow. (**B**) Land collapses have caused route 61 in Centralia to crack and buckle, while steam filled with combustion products can be seen rising from the fire below. (**C**) Elemental sulfur crystals that precipitate at the surface support sulfur-cycling bacteria.

battles that are detailed in The Day the Earth Caved In: An American Mining Tragedy (Quigley 2007). Ultimately, the fire spread underneath the town (Fig. 1A), causing land collapses and the emission of unhealthy levels of hot, oxidized combustion gases, such as carbon monoxide and sulfur dioxide, that continue to rise to the surface through cracks, or vents (Fig. 1B). As recently as 2003, the surface temperatures above the mine fire reached over 450°C. However, the fire has moved into deeper coal seams, and the current maxima are around 65°C. As the gases rise to the surface, they cool and precipitate their dissolved chemicals, particularly nitrogen and sulfur species, into the surrounding soils (Fig. 1C). Thus, the mine fire represents both a thermal and a chemical extreme environment in which thermophilic nitrogen and sulfur cycling bacteria thrive (Janzen and Tobin-Janzen 2008).

To understand the ecological roles that bacteria play in mine fire-affected soils, students must first understand the environment in which they reside. Towards that goal, concentrations of sulfur and nitrogen species that are common in other geothermal soils were analyzed in Centralia soil samples. The results are shown in Table 1. All three samples were acidic, which is typical for fire-affected soils in the area (Tobin-Janzen *et al.* 2005; Lee *et al.* 2017). Also notably, total sulfur, sulfate and nitrate levels were markedly elevated in the 52°C and 60°C sites as compared to the 37°C site, while ammonium levels were slightly elevated in those same sites. The highest levels of these combustion chemicals were not necessarily found at the hottest temperatures, with the highest sulfur values seen in the 52°C sample and the highest ammonium and nitrate values seen in the 60°C sample.

This difference is likely due to a variety of biological and physical factors that probably include the metabolic activity of resident bacteria (Tobin-Janzen *et al.* 2005).

Student teams used these temperature and chemical differences to form hypotheses regarding bacterial species that could be present and playing important biogeochemical roles in the Centralia ecosystem. For example, large numbers of ammonia oxidizing bacteria could be responsible for the high levels of nitrate in sample S2. The high levels of sulfur in S3 might be expected to support more sulfur-metabolizing bacteria, or could, conversely, remain high due to their absence. The student hypotheses frequently included members of genera such as *Geobacillus* and *Nitrospira* that are known residents in other geothermal environments (Hedlund *et al.* 2012). Thus far, even though the student teams have all used the same dataset, each team has proposed a unique hypothesis.

Class two: an introduction to LINUX and QIIME

Students began their bioinformatics analyses by performing a short, guided activity to familiarize themselves with LINUX commands (moving between directories, listing the contents of directories, etc.). Once they had successfully completed this exercise, they learned about the information present in the sequence, quality and metadata files that they would use to perform their bioinformatics analyses in QIIME and used QIIME to split and quality filter their sequence data.

All of the sequence analyses were performed on Mac OSX laptops, which have direct access to the LINUX environment

Table 1. Thermal and chemical analysis of surface soil samples overlying the Centralia, PA, coal mine fire.

Sample	Soil temperature (°C)	Total sulfur (mg g^{-1})	Sulfate (mg g^{-1})	Ammonium (mg g^{-1})	Nitrate (mg g^{-1})	pH
S1	37	2.5	2	3.63	2.14	4.58
S2	60	125.0	50	12.63	103.10	4.12
S3	52	250.0	90	9.29	54.16	4.87

via the Mac Terminal. Students worked in teams of four and ran QIIME on Amazon Web Elastic Compute Cloud (EC2) instances (http://aws.amazon.com/ec2). Concatenated sequence (fas), quality (qual) and metadata files were uploaded to the instances using Cyberduck (https://cyberduck.io/) at the beginning of each class, and students interacted with the instances via ssh using the Mac Terminal using either their own laptops or departmental iMacs. At the end of each day, all files were downloaded to thumb drives via Cyberduck.

Class three: composition and ecological roles of Centralia bacterial communities

During the third class, the students used QIIME v1.9.1 to test their hypotheses, first by picking OTUs and assigning those OTUs to taxonomic groups using the Greengenes (v 13_5 and 13_8) database. They ultimately constructed .biom tables (McDonald *et al.* 2012) and organized their results into viewable bar charts. Typical results are shown in Fig. 2.

Student analysis of the bacterial community composition in Centralia revealed many interesting patterns that, in turn, supported student learning about microbial ecology. As can be seen in Fig. 2, Acidobacteria predominated in all three samples, but are most prevalent at 37°C. Proteobacteria (particularly the Gamma Proteobacteria *Nevskia* and *Pseudomonas*) became more prevalent at 52°C and Chloroflexi (genus *Thermogemmatispora*) and Firmicutes (genus *Geobacillus*) became more prevalent at 60°C. These latter two genera are notable as they are common members of other thermophile communities. Note also, that unassigned bacteria represented increasing proportions of soil communities as the temperatures rose, with levels of 18.6%, 32.2% and 47.3% at 37°C, 52°C and 60°C, respectively. As the Greengenes database used has not been updated since 2013, this analysis almost certainly overestimates the number of novel species. Nevertheless, Centralia's hot soils likely provide a valuable environment for novel species discovery.

Although the bar charts generated by this analysis extended to the species level, the sequencing primer used in this study
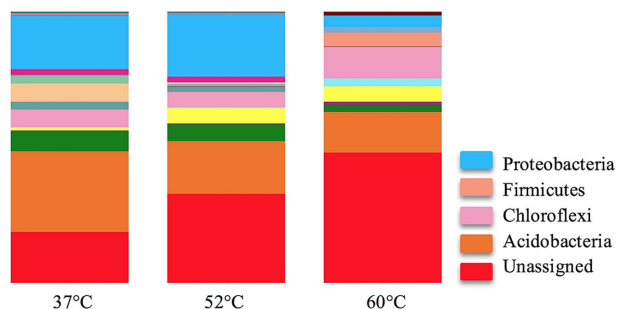
(8F) does not generally allow taxa to be discriminated beyond the genus level. This limitation turned into a valuable teaching tool, as the frustration that students felt when they realized that they could not know for sure if their species were actually present in Centralia provided a convenient opportunity to discuss the importance of experimental design. In this case, the students learned that initial primer design is critical to the ultimate phylogenetic conclusions that a 16S rRNA gene amplicon study can reach. We used primers that targeted the V1 + V2 areas of the 16S rRNA gene, which have been shown to have less phylogenetic resolution than primers targeting other variable regions (Yang, Wang and Qian 2016).

Following these discussions, student teams were instructed to continue with their analyses, assuming that their species was present, as long as the genus was present. Occasionally, a team did not even find genus-level evidence for their hypothesized species. These cases also provided excellent discussion points for the class, allowing us to ask if the failure to identify a genus in this unreplicated pilot dataset really falsified their hypothesis, or if a better dataset could provide more conclusive evidence. Regardless, since the final papers required students to use actual sequence data to generate phylogenetic trees, as well as in-depth discussion of their species' probable role in nitrogen- or sulfur-cycling in the Centralia ecosystem, students needed to work with a genus that was present in the soil samples. Thus, teams were required to revise their hypotheses using new species, and to retest the dataset until their hypotheses were supported.

Class four: alpha diversity and phylogenetic analysis

During the final class, students first aligned their sequences using the Greengenes core reference alignment (DeSantis *et al.* 2006; Caporaso *et al.* 2010a), and filtered the alignments to remove uninformative sequence data. They made phylogenetic assignments using the Ribosomal Database Classifier 2.2 (Wang *et al.* 2007) and constructed phylogenetic trees using FastTree (Price, Dehal and Arkin 2010), a modified neighborjoining method. The students computed alpha diversity using Heip's evenness, Faith's phylogenetic diversity and total OTUs. They collated the data and plotted it for analysis, as shown in Fig. 3.

All three measures (Fig. 3) showed decreasing values with increasing temperature. From this data, students concluded that overall community diversity decreased as temperatures increased, and that hotter sites tended to have a few predominating taxa, whose success was likely determined by their fitness in the different borehole environments. Though small, this dataset is in agreement with our recent study that used Illumina sequencing of the 16S rRNA gene with higher coverage per soil (Lee *et al.* 2017). In that study, we showed that overall prokaryotic diversity in soils that were currently affected by the fire was indeed lower than that seen in soils that were previously affected by the fire, but had since recovered, as well as in soils that had never been affected by the fire. In future years, we will
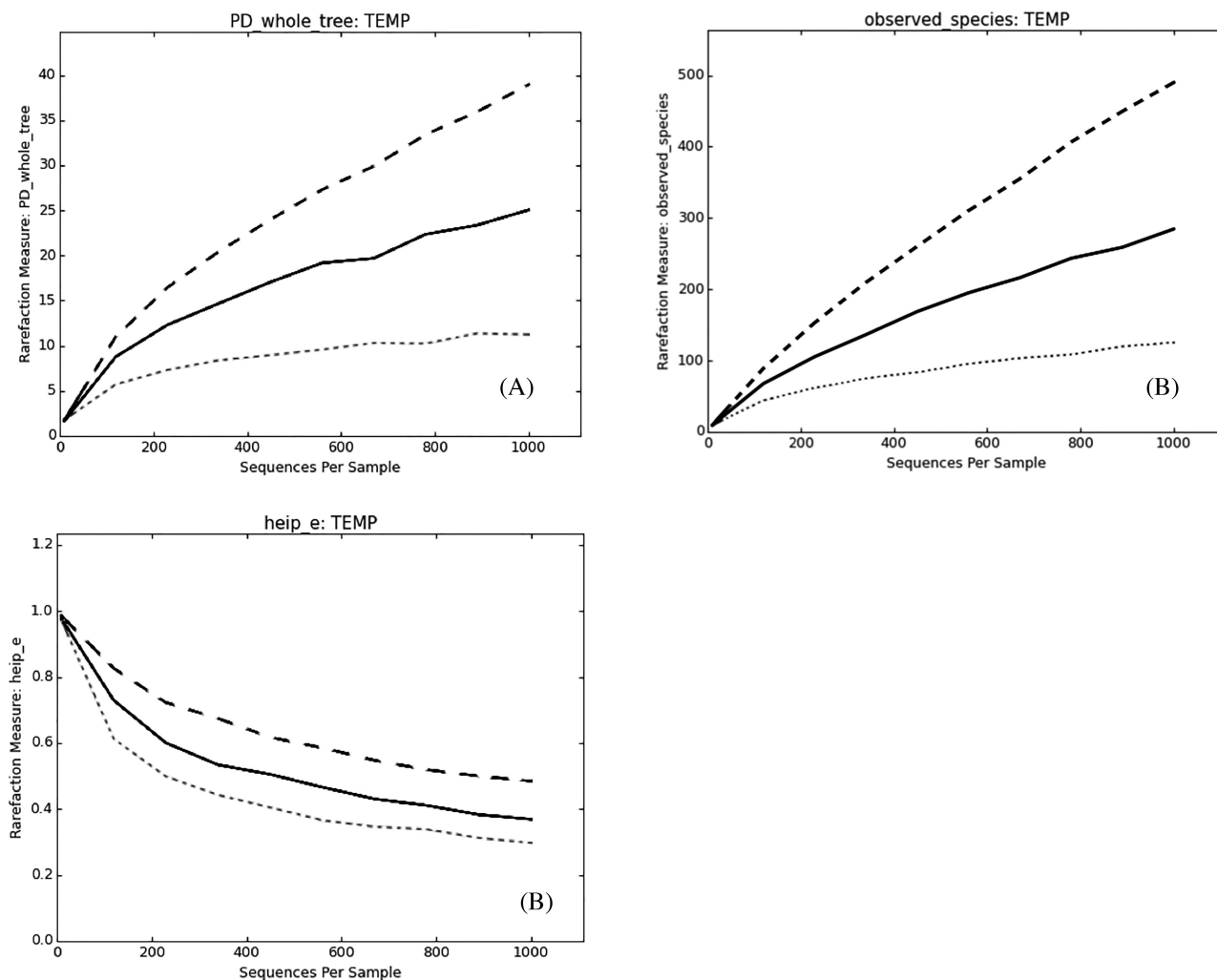

Figure 2. Bacterial community diversity in mine fire-affected soils. The dominant classes found in each sample site are indicated in their respective bar charts. Proteobacteria, Acidobacteria and unassigned taxa were common at all three temperatures, while Firmicutes and Chloroflexi became more prevalent at 60°C.

Figure 3. Alpha diversity of Centralia soil samples decreases with temperature. Faith's phylogenetic diversity (**A**), total OTUs (**B**) and Heip's Evenness (**C**) were calculated using QIIME and are shown for the 37°C (large dash), 50°C (solid) and 60°C (small dash) samples.

use this paper as a reading assignment at the end of the module to demonstrate how a larger and more complex dataset can translate into more robust statistical analyses, and, ultimately, a publication.

These graphs also clearly demonstrated that none of the soil samples had been sequenced deeply enough to assure that sufficient phylogenetic diversity was captured. Although there was access to sequence datasets with higher coverage and replicates (e.g. from the Lee and Sorensen study), this one was chosen for the module because it allowed students to engage, once again, in robust discussions about experimental design. They were able to ask questions such as 'Would we have found our species if we had sequenced deeply enough?' 'Were some species artificially prevalent or absent due to insufficient sampling?' 'What would we need to do to improve this dataset?' These questions followed up nicely on the primer design discussions during the analysis of community composition. It turns out that the reason for the lack of sequencing depth, in this case, was poor primer quality. There was a technical problem during primer synthesis, and the resulting primers contained errors that caused too many sequences to fail the initial quality trimming.

Ultimately, students imported their sequences and metadata into FigTree (Rambout 2012) This program not only al-

lowed students to generate phylogenetic trees, but also to color tree branches to match metadata categories. Thus, the students could visually determine if their chosen genus was associated with a particular temperature/chemical characteristic, or if it was broadly represented in all Centralia soil types. In general, students found that their chosen genus was located where they would expect it to be, both with regard to soil temperature and chemistry, although there were some surprises. For example, *Nevskia* has not previously been described in thermophilic environments (Kim *et al.* 2011; Leandro *et al.* 2012) so its prevalence in 52°C soils in Centralia is quite interesting.

This unusual observation will provide a final opportunity to discuss the limitations of 16S rRNA gene amplicon analyses in future classes. In the Illumina-sequenced Centralia dataset (Lee *et al.* 2017), *Nevskia* was detected, only as part of the rare biosphere, which is a striking contrast to its prevalence in this dataset. This difference could reflect real variation in microbial community composition due to the different microenvironments sampled during the two studies. However, it could also be a result of primer bias, as different primers were used in the two studies, or imprecise assignment by the Greengenes database. Our analysis is, after all, only as good as our datasets and databases.
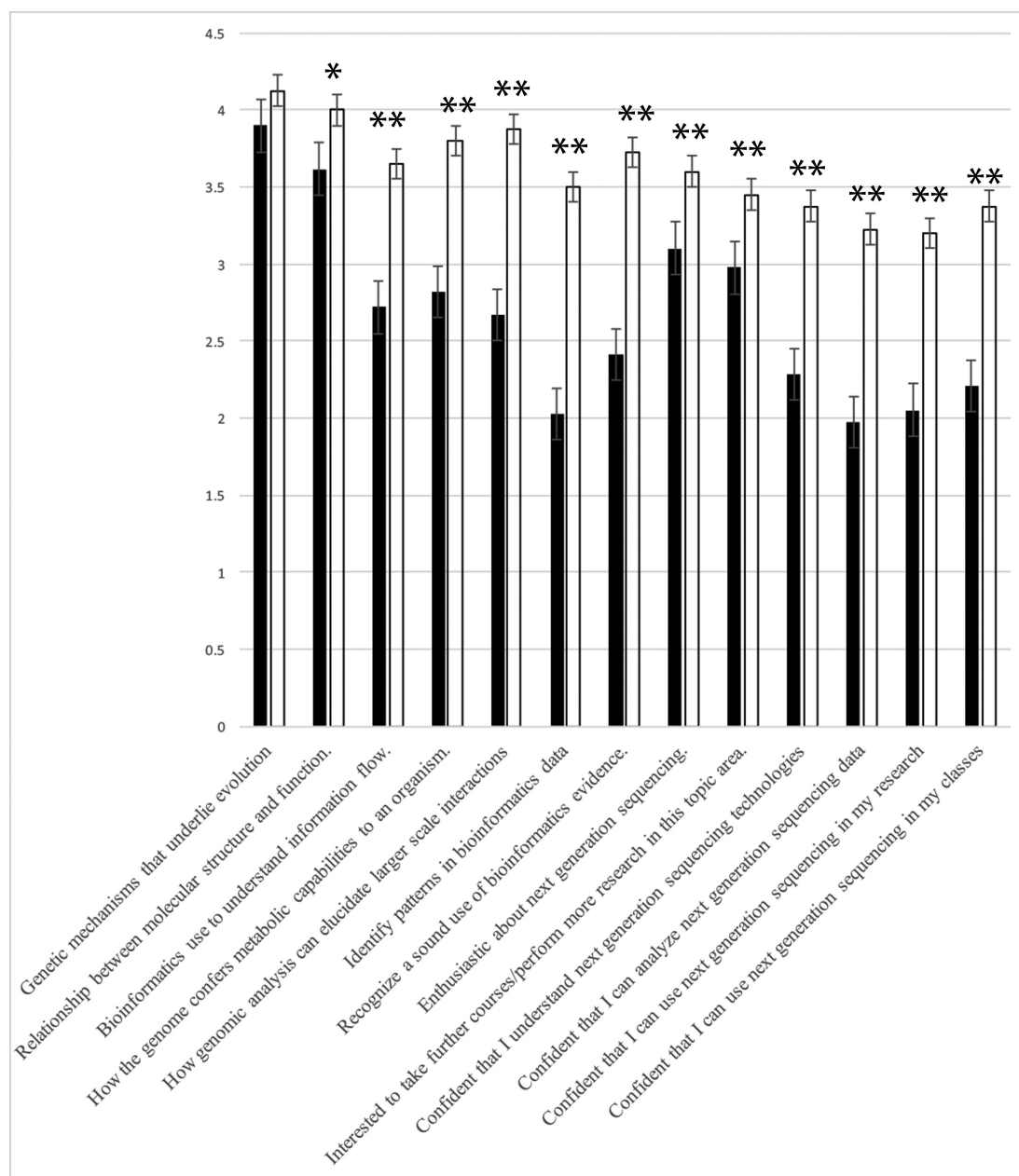
Figure 4. Assessment of student confidence and attitudes. Students (n = 38) were asked to rate their knowledge and attitudes using a 5-point Likert scale via the GCAT-SEEK Survey Instrument. Pre- (black) and post-course (white) responses were tallied. Highly significant (**$P < 0.005$) and significant (*$P < 0.05$) improvements are indicated.

Assessment

Assessment of student skills and attitudes was primarily performed using the GCAT-SEEK survey. GCAT SEEK is an NSF-funded initiative that provides access to high-throughput sequencing technologies for faculty in primarily undergraduate and minority-serving institutions. It also helps these faculty to develop new curricular modules that use that sequence data. (Buonaccorsi *et al.* 2011, 2014), and this 16S rRNA gene amplicon module was developed during a GCAT-SEEK workshop in 2015. Pre- and post-module surveys were administered to all students in the class during 2015 and 2017.

A total of 38 students responded to this survey. In general, as shown in Fig. 4, students' self-reported enthusiasm for

and confidence in their ability to do bioinformatics analyses increased across the board, with the most notable increases seen in questions that involved analyzing bioinformatics data. The one question in which no significant increase was seen was 'I understand genetic mechanisms that underlie evolution'. Because most students enrolled in microbiology during those 2 years had either taken an introductory ecology and evolution or a systematic biology class, or both, it was not surprising to see very high student confidence even in the pre-test, leading to a small but insignificant increase during the post-test.

Direct assessment of students' actual progress on sequencing and bioinformatics skills and knowledge showed much more modest gains. The GCAT-SEEK instrument contained five multiple choice questions that directly pertained to the material

Table 2. Student performance on GCAT-SEEK direct assessment questions.

Question	Percent correct		Percent increase
	Pre	Post	
Which of the following answers correctly represents the order of steps by which a typical next generation sequencing project is performed?	7.9	8.5	0.60
Which of the rationales below correctly explains a step in next generation sequencing?	5.3	33	28
Which is not true about PHRED quality scores?	0	44	44
Which of the following bioinformatics workflows would you use to analyze your sequence data?	0	67	67
Your sequence data is in the 'pg1' directory. In the terminal, what would you type to get to the directory with your sequence data. (Assume you start in home directory).	44	72	28

learned in this module, and the students' gains on these questions are shown in Table 2. The first two questions, which involved sequencing techniques that the students did not actually do, showed the most modest (in one case, essentially zero) gains and the lowest overall post-module scores. By contrast, student improvement on the last three questions that related to skills or knowledge they directly utilized in the module had increases ranging from 28% to 67%, with the % correct ranging from 44% to 72% during the post-test.

In their final papers, students again displayed these same learning trends, demonstrating a sophisticated comprehension of microbial roles in biogeochemical cycling, and the ways in which sequence data helped them to understand those roles. By way of example, one team chose *Nitrospira moscoviensis* as their target species based on the high levels of nitrate and ammonium in the soil samples. *Nitrospira moscoviensis* is a urease positive nitrite oxidizing bacterium (Koch *et al.* 2015). As the students concluded, 'These two traits virtually guarantee *N. moscoviensis* is an organism that forms the foundation of the nitrogen cycle in many environments, including extreme environments like the Centralia soils overlying the fire. In this cycle, *N. moscoviensis* cleaves urea to ammonia and CO_2, freeing up ammonia for urease-negative, ammonia oxidizing microbes. The non-ureolytic ammonia oxidizing microbes concurrently oxidize the ammonia into nitrite, forming a reciprocal feeding pattern in which *Nitrospira* can then fully oxidize nitrite into nitrate'.

Their comprehension of the techniques used to generate 16S rRNA gene amplicon sequences, however, was much more simplistic. When they were asked to demonstrate a clear understanding of the sequencing methods used, the students frequently scrambled the steps, 'The (PCR) products were then used as the template for Roche 454 pyrosequencing after which a metagenomic 16S rRNA gene library was created', or displayed a lack of comprehension regarding the underlying science, 'A gene library consisting of all the bacterial 16S rRNA sequences present at the sample site was generated from the isolated RNA'. In the absence of wet-lab activities that are not possible in this lecture class, it will be necessary to develop future classroom activities that require the students to reflect on these techniques more deliberately (Prince 2004).

During course evaluations, the students suggested one further area of improvement. Apparently, teams of four were too large, and when that many students were gathered around a single computer, only one student really got to do the analysis. Students remarked that they felt they would learn more if they actually had time with the keyboard. Thus, in future years I plan to split the teams of four in half after they generate their hypotheses, have them work in pairs during the QIIME analyses, and have them get back together to analyze the data and write their final papers.

CONCLUSIONS

The 'A Town on Fire!' module successfully integrated 16S rRNA gene amplicon analysis into an undergraduate microbiology lecture course. Student attitudes towards and confidence in their ability to carry out bioinformatics analyses improved, as did their ability to draw ecological conclusions from bioinformatics data. In particular, the use of datasets with known flaws helped students to understand the role that primer design and data quality play in the analysis of molecular data. Student learning gains were strongest in the areas in which they actually performed the analyses (bioinformatics analyses) and weakest in the areas in which lectures alone covered the material (sequencing). In lecture courses, this weakness could be addressed by developing classroom activities that require students to reflect upon those areas much more deliberately.

SUPPLEMENTARY DATA

Supplementary data are available at *FEMSLE* online.

ACKNOWLEDGEMENTS

FUNDING

Conflict of interest. None declared.

REFERENCES

Buonaccorsi V, Boyle M, Grove D *et al*. GCAT-SEEKquence: Genome Consortium for Active Teaching of Undergraduates through Increased Faculty Access to Next-Generation Sequencing Data. *CBE Life Sci Educ* 2011;**10**:342–5.

Buonaccorsi V, Peterson M, Lamendella G *et al*. Vision and change through the Genome Consortium for Active Teaching Using Next-Generation Sequencing (GCAT-SEEK). *CBE Life Sci Educ* 2014;**13**:1–2.

Caporaso JG, Bittinger K, Bushman FD *et al*. PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 2010a;**26**:266–7.

Caporaso JG, Kuczynski J, Stombaugh J *et al*. QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 2010b;**7**:335–6.

DeSantis TZ, Hugenholtz P, Larsen N *et al*. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* 2006;**72**:5069–72.

Edwards U, Rogall T, Bloecker H *et al*. Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucl Acids Res* 1989;**17**:7843–53.

Elick J. Mapping the coal fire at Centralia, Pa using thermal infrared imagery. *Int J Coal Geol* 2011;**87**:197–203.

Hedlund B, Cole J, Williams A *et al*. A review of the microbiology of the Rehai geothermal field in Tengchong, Yunnan Province, China. *Geosci Front* 2012;**3**:273–88.

Janzen C, Tobin-Janzen T. 2008. Microbial communities in fire-affected soils. In: Dion P, Nautiyal C (eds). *Microbiology of Extreme Soils*. Heidelberg; Springer-Verlag, 2008, 299–316.

Kim S-J, Weon H-Y, Kim Y-S *et al*. Nevskia terrae sp. nov., isolated from soil. *Int J Syst Evol Microbiol* 2011;**61**:1226–9.

Koch H, Lücker S, Albertsen M *et al*. Expanded metabolic versatility of ubiquitous nitrite-oxidizing bacteria from the genus Nitrospira. *P Natl Acad Sci USA* 2015;**112**:11371–6.

Leandro T, França L, Nobre M *et al*. *Nevskia aquatilis* sp. nov. and *Nevskia persephonica* sp. nov., isolated from a mineral water aquifer and the emended description of the genus *Nevskia*. *Syst Appl Microbiol* 2012;**35**:297–301.

Lee S-H, Sorensen JW, Grady KL *et al*. Divergent extremes but convergent recovery of bacterial and archaeal soil communities to an ongoing subterranean coal mine fire. *ISME J* 2017;**11**:1447–59.

McDonald D, Clemente JC, Kuczynski J *et al*. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *GigaSci* 2012;**1**:7.

National Research Council BIO2010. *Transforming Undergraduate Education for Future Research Biologists*. Washington, DC: The National Academies Press, 2003, DOI: 10.17226/10497.

Nikkari S, Lopez FA, Lepp PW *et al*. Broad-range bacterial detection and the analysis of unexplained death and critical illness. *Emerg Infect Dis* 2002;**8**:188–94.

Nolter M, Vice D. Looking back at the Centralia coal fire: a synopsis of its present status. *Int J Coal Geol* 2004;**59**: 99–106.

Price MN, Dehal PS, Arkin AP. FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS One* 2010;**5**:e9490.

Prince M. Does active learning work? A review of the research. *J Eng Educ* 2004;**93**:223–31.

Quigley J. *The Day The Earth Caved In: An American Mining Tragedy*. New York: Random House, 2007.

Shade A, Teal T. Computing Workflows for biologists: a roadmap. *PLoS Biol* 2015;**13**:e1002303.

Rambaut A. *FigTree v1. 4. Molecular Evolution, Phylogenetics and Epidemiology*. Edinburgh, UK: University of Edinburgh, Institute of Evolutionary Biology, 2012. http://tree.bio.ed.ac.uk/ (21 December 2017, date last accessed).

Tobin-Janzen T, Shade A, Marshall L *et al*. Nitrogen changes and domain bacteria ribotype diversity in soils overlying the Centralia, Pennsylvania underground coal mine fire. *Soil Sci* 2005;**170**:191–201.

Vision and Change. *Vision and Change in Undergraduate Biology Education: A Call to Action*. American Association for the Advancement of Science, 2011, http://visionandchange.org/finalreport/ (21 December 2017, date last accessed).

Wang Q, Garrity GM, Tiedje JM *et al*. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* 2007;**73**:5261–7.

Yang B, Wang Y, Qian P-Y. Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC Bioinformatics* 2016;**17**:135.