Received: 12 July 2024 | Revised: 24 December 2024 | Accepted: 17 January 2025

DOI: 10.1002/bmb.21888

PROTOCOL

Biochemistry and Molecular Biology Education

WILEY

Metagenomics education in a modular CURE format positively affects students' scientific discovery perception and data analytical skills

M. C. Morsink | **E. N. van Schaik** | **K. Bossers** | **D. A. Duijker** | **A. G. C. L. Speksnijder**

Research Group of Environmental Metagenomics, Leiden Centre for Applied Bioscience, Leiden University of Applied Sciences, Leiden, Netherlands

Correspondence
M. C. Morsink, Research Group of Environmental Metagenomics, Leiden Centre for Applied Bioscience, Leiden University of Applied Sciences, Leiden, Netherlands.
Email: maarten.morsink@gmail.com

Abstract

Targeted metagenomics is a rapidly expanding technology to analyze complex biological samples and genetic monitoring of environmental samples. In this research field, data analytical aspects play a crucial role. In order to teach targeted metagenomics data analysis, we developed a 4-week inquiry-driven modular course-based undergraduate research experience (mCURE) using publicly available Australian coral microbiome DNA sequencing data and associated metadata. Since an enormous amount of metadata was provided alongside the DNA sequencing data, groups of students were able to develop their own authentic research questions. Throughout the course, the student groups worked on these research questions and were supported with bioinformatics and statistics lessons. Additionally, practical aspects of data collection and analysis were addressed during hands-on field work on a nearby Dutch beach. Evaluation of the course indicated that the majority of students (1) achieved the intended metagenomics-based learning outcomes and (2) experienced scientific discovery while working on their research projects. In conclusion, the huge amount of data and metadata available in the coral microbiome data set facilitated the development of a strongly inquiry-driven course. Different groups of students were able to develop and conduct their own distinct microbiome research projects and our current mCURE format positively affected students' metagenomics data analytical skills and scientific discovery perception.

KEYWORDS

bioinformatics, coral, CURE, data analysis, inquiry, metabarcoding, microbiome, microbiomics, mixed amplicon sequencing, targeted metagenomics, undergraduate research experience

1 | INTRODUCTION

Research on microbiome composition of a variety of patient and environmental samples has become more accurate due to the emergence of metagenomics technology.[1] Since an enormous amount of data is generated using this technology, data analysis has become a very important aspect. Consequently, we expect an increased demand for research technicians with both technical laboratory and data analytical skills in this research field. Additionally, the increased availability of metagenomic datasets also provides educationalists with many new

opportunities to develop inquiry-driven data analysis courses.[2]

The Department of Biology and Medical Laboratory Research, in close collaboration with the Leiden Centre for Applied Bioscience of the Leiden University of Applied Sciences (The Netherlands), offers an undergraduate (B.Sc.) four-year educational program which trains students to become research laboratory technicians. Currently, data analysis constitutes only a small part of the curriculum, with the majority of the courses focused on practical laboratory skills. In order to facilitate students who have an interest in the field of metagenomics, the department created the opportunity to develop and implement a 140 h elective course with an emphasis on targeted metagenomics data analysis. The elective was implemented into the third year of the curriculum and scheduled throughout a 10-week period. During this period, students also took two other courses, each with a 140 h study load. At this point in the program, students possess basic molecular biology laboratory skills and know how to perform basic statistical tests using R.

A large part of targeted metagenomics consists of several general next generation sequencing (NGS) processing and analysis methods that are also used in other research contexts. Therefore, a targeted metagenomics course equips students with transferable data analytical skills, preparing them for other research areas as well.

When teaching data analysis methodology, using a course based undergraduate research experience (CURE) format has been shown to be highly effective.[3] Since the duration of the elective course had to be confined to 4 weeks, a limited modular course-based undergraduate research experience (mCURE) was developed. Previously, these shorter mCUREs have been shown to positively impact problem-solving ability[4] as well as project ownership.[5] Hence, we aimed to design an inquiry-based course in which students would collaborate in small groups to develop and conduct a targeted metagenomics data research project. Importantly, the a priori outcome would be unknown to both students and teachers. We expected that such a course, in which all the different student groups would perform their own, unique data research projects, could enhance students' perception of scientific discovery.

For this purpose, we used publicly available Australian coral microbiome DNA sequencing data from a targeted metagenomics study performed by Pollock et al.[6] This study was part of the Global Coral Microbiome Project for which a video series is available online.[7,8] In order to enhance students' perception of learning task authenticity, we showed the first video of this series.

In this study, 236 coral colonies were sampled from 21 different sites around Australia. For each sample, up to 162 host and environmental parameters were measured. These metadata included information on photosynthetically active radiation, water temperature, turf contact, and other factors. After sampling, the coral mucus, tissue and skeleton compartments were separated and in total, 691 samples were processed for DNA extraction and partial 16S rRNA gene amplification with generic primers flanking variable regions. Mixed amplicon libraries were sequenced using paired-end Illumina MiSeq technology. This way, bacterial and archaeal components of the coral microbiome were targeted and 1382 fastq files were generated. Subsequently, the published article globally describes the drivers of microbiome diversity in the 3 coral compartments.[6]

This enormous data set provided us with the opportunity to have our students define authentic research questions with regard to the influence of a certain environmental or host factor on the coral microbiome composition. Hence, groups consisting of 3 students were asked to select a host or environmental factor from the available metadata. Next, they had to develop a research question and hypothesis with regard to whether changes in this factor are associated with changes in microbiome diversity or abundancies of specific coral microbiome taxa. Students were expected to use scientific literature to substantiate their hypotheses. Subsequently, students selected samples for which the metadata showed variation in the selected factor. Part of the educational strategy was to address possible confounders at a later stage.

Throughout the course, students worked on their research projects and were supported with bioinformatics and statistics lessons. In order to assess the biodiversity of the microbiome in the selected samples, students used a 'clustering first' approach.[9] Using this approach, the sequenced 16S rRNA reads are first clustered into 'Operational Taxonomic Units (OTUs)'.[10] Subsequently, centroid reads are selected from the OTUs and taxonomically identified using BLAST and a lowest common ancestor algorithm. After obtaining OTU tables with read counts and tables with taxonomical identities for the OTUs, RStudio[11] was used to visualize and statistically analyze the data.

At the end of the course, students presented and discussed their research project with 2 teachers. Subsequently, they were graded according to their performance with respect to (1) data generation, (2) data analysis and (3) application of data analysis to address biological questions. Grading was performed using the different levels of Bloom's revised taxonomy.[12,13]

Since our current focus was on providing students with an authentic research task, the course was evaluated using 2 hallmarks that measure student's perception of doing original scientific research, that is, discovery & iteration.[14] The discovery hallmark indicates the degree to which students are able to generate new scientific

knowledge. The iteration hallmark reflects the opportunities for students to repeat or alter their work to account for errors, fix problems, or to analyze additional data to address new questions or hypotheses that arise during the investigation. These hallmarks were evaluated using student questionnaires.

2 | MATERIALS AND METHODS

2.1 | Course design

The course was performed annually in a period of 2 years. In total, 47 students participated in the course. Students were assigned to project groups consisting of 3 students. Each student group was tasked with developing a research question based on the available meta data.

The course started with a lecture in which a general overview of the research performed by the Research Group of Metagenomics was presented.

The second class, consisting of approximately 4 h of field work, focused on how to measure biodiversity and how to generate species accumulation (rarefaction) curves. When analyzing the composition of microbiomes, these curves give an indication on whether the sampling effort was sufficient, that is, whether enough reads were taxonomically identified to get a good overview of the microbiome diversity of a sample. For that purpose, the students, together with 2 teachers, went to a nearby beach and were tasked with comparing washed up species diversity between 3 different sites. The beach is located at Katwijk aan Zee, The Netherlands, and a river draining sluice is present in close proximity to this beach. Students measured biodiversity (1) directly near the sluice, (2) at approximately 100 meters distance from the sluice and (3) at approximately 200 meters from the sluice. The ObsIdentify app (available at observation.org) was used to identify the washed up species. After data collection at the beach, students went back to the university to perform the analysis and generate rarefaction curves to compare the biodiversity of the 3 different sites at the beach. Afterwards, a class discussion was held in which the rarefaction curves were used to determine whether the sampling efforts were sufficient for each site and whether biodiversity could be compared between the 3 sites.

Subsequently, a tutorial dedicated to general aspects of DNA targeted metagenomics was organized. In targeted metagenomics, DNA markers, such as the 16S rRNA gene, are used to identify the taxonomic groups of interest that are present in a microbiome sample. These markers are first amplified using PCR, after which the resulting mix of amplicons is sequenced using a NextGen sequencing platform. Obtained reads are taxonomically classified using BLAST and a corresponding DNA marker database, such as the 16S rRNA database. The tutorial covered several conceptual and practical aspects of targeted metagenomics. First, students needed to address the question why the term 'mixed amplicon sequencing' is applicable to targeted metagenomics. Then, students used a review article[15] to study the relationship between the gene used as a barcode and the taxonomic groups of interest. Next, students investigated the 16S rRNA barcode gene which is specifically used for bacterial and archeal identifications. This barcode gene contains 9 variable regions and different bacterial genera can be taxonomically identified using different combinations of these regions. Students had to propose a selection of different variable regions for the identification of two different bacterial genera, that is, *Pseudomonas* and *Clostridium*, based on information provided in Johnson et al.[16] Subsequently, students explored the rationale for using degenerate primers during PCR amplification. Finally, since the obtained sequencing reads are taxonomically classified using BLAST, students were tasked with exploring the meaning of different BLAST parameters, that is, query coverage, percentage overlap, max score, total score and E-value.

Next, an introduction into the data analysis research project was given. The targeted metagenomics study performed by Pollock et al.[6] was part of the Global Coral Microbiome Project.[8] This project was introduced by showing the Global Coral Microbiome Project movie, part 1: Great Barrier Reef.[7] Since the original publication of Pollock et al.[6] was written for a highly specialized academic public, we didn't ask the students to read the paper. Instead, we provided the students with a short introduction of the paper, including an explanation of Figure 1c, and guided them through the metadata table stored in the *gcmp16S_map_r25.txt* file under the *GCMP Australia sequence data, OTU tables, and metadata* folder (https://doi.org/10.6084/m9.figshare.c.3855466.v2).

Next, students worked on the development of their own research question and were able to ask the teacher for advice. Afterwards, the students presented and discussed their research question, hypothesis and data analysis approach in a classroom presentation during the next lesson.

The following lesson was focused on the data analysis pipeline used on the Galaxy platform.[17] Students obtained accounts for our custom-made Galaxy instance and were tasked with uploading the data files needed for their research projects.

Since the sequencing data in Pollock et al.[6] was generated using the Illumina MiSeq sequence platform, subsequent tutorials were dedicated to (a) the principles of

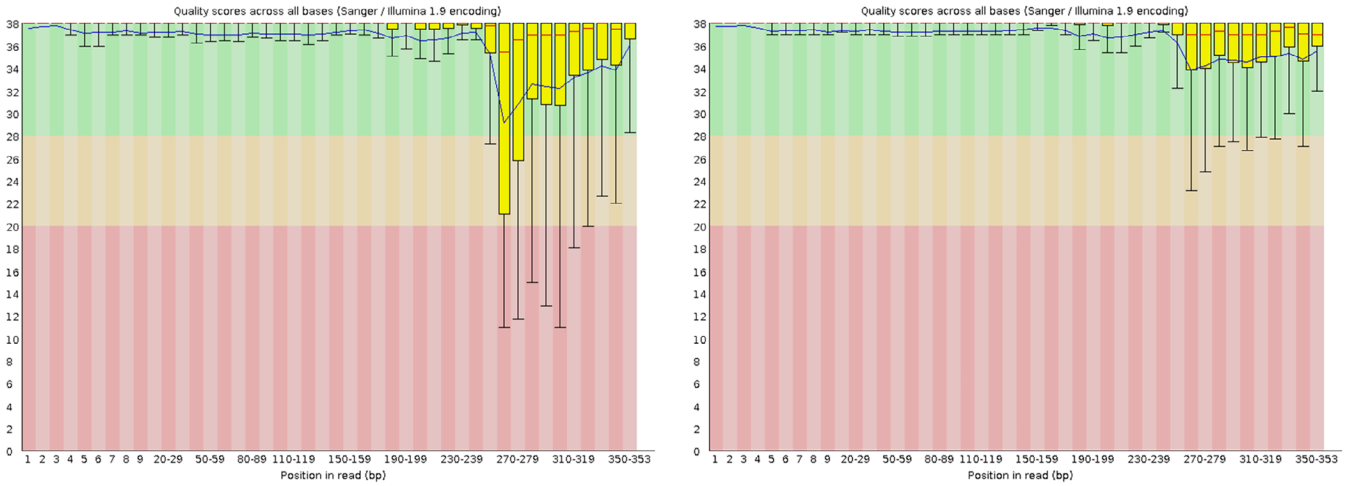

FIGURE 1 FastQC results after removing PCR-primers with Cutadapt from 2 different student groups are displayed (left and right). On the x-axis, the position of the nucleotides are displayed, on the y-axis the quality scores. Green indicates good quality, yellow mediocre quality and red low quality. In the left panel, lower quality scores are obtained in the 3′ end of the reads compared to the right panel.

multiplex paired-end sequencing, (b) quality control using FastQC,[18] (c) merging forward and reverse reads with FLASH,[19] (d) removal of primers using CutAdapt[20] and (e) sequence trimming with PRINSEQ.[21]

Next, an entire lesson was dedicated to the algorithm for clustering reads into OTUs.[22] Students were given DNA sequence reads printed on small stretches of paper (see Data S1, S2) and had to use a description of the algorithm to produce an OTU table. Then, students used BLAST to assign taxonomy to the OTU centroids.

After OTU tables and taxonomies were obtained, several tutorials were dedicated on how to use RStudio[11] to perform subsequent data analysis steps. These tutorials covered the topics of rarefaction curves, alpha and beta diversity and several statistical tests, among which t-test and ANOVA. Additionally, the concept of confounders and how to deal with these was also included in these tutorials.

Finally, students presented and discussed their research findings and grading was performed with a single point rubric.[23,24] In this rubric, the procedural knowledge aspects of data generation, data analysis and application of the data to biological questions were graded using 4 performance criteria. These included the ability to (1) interpret the biological meaning of obtained results, (2) sketch laboratory procedures for targeted metagenomics, (3) use bioinformatic methods necessary for targeted metagenomics and (4) employ relevant statistical analyses. Generally, in The Netherlands, a 1 to 10 point grading system is used and 6 is the minimum score required to pass. To obtain a 6 in our current course, students needed to fulfill each of the 4 criteria on Bloom's cognitive process level of 'application' for procedural knowledge.[12,13] Additionally, students could obtain bonus points for each criterium if they increased their performance to higher Bloom levels of analysis, evaluation or creation for procedural knowledge. This way, a maximum of 4 bonus points could be obtained, adding up to a maximum final grade of 10.

All lesson plans and the grading rubric are available in Supplementary Information Files 1 and 2.

Since human subjects were involved, this study was approved ethically by the Committee on Applied Research of the Leiden Centre for Applied Bioscience, Leiden University of Applied Sciences.

2.2 | Technical details

Coral microbiome DNA sequencing data was obtained from a study performed by Pollock et al. *(6 Pollock 2018)*. We used 16S V4 amplicon forward & reverse Illumina MiSeq reads which are publicly available at https://doi.org/10.6084/m9.figshare.c.3855466.v2. These reads are stored in the *raw_fastqs.tgz* file under the *GCMP Australia sequence data, OTU tables, and metadata* folder (posted on 2017-08-22).

In order to process the FASTQ-files and obtain OTU tables, a custom-made pipeline on a Galaxy instance[17] was used. This pipeline contained data analytical tools obtained from the Naturalis Biodiversity Center GitHub environment. In order to align with the newest Galaxy best practices, we updated the tools and links to the original as well as the updated GitHub and toolshed versions are included in the references list.

The names of the original raw FASTQ-files contain many dots which can hamper file-type recognition during

downstream analyses in Galaxy. Therefore, prior to distributing the FASTQ-files to the students, these names were renamed using the ManageZIP tool.[25] Dots were replaced with underscores, with an exception made for the last 2 file type extension dots. For example, the file name *E1.2.Tur.pelt.1.20140814.M_S71_L001_R1_001.fastq.gz* was changed into *E1_2_Tur_pelt_1_20,140,814_M_S71_L001_R1_001_1.fastq.gz*. Similarly, the filenames in the metadata table were adjusted accordingly. The total collection of renamed file names is available at Zenodo (https://zenodo.org/records/12723661 or DOI https://doi.org/10.5281/zenodo.12723661).

After selecting their files of interest based on the metadata, students used FLASH[19] to merge forward and reverse reads. Cutadapt[20] was used to remove primers from the reads and PRINSEQ[21] to further trim the reads if necessary. FastQC[18] was used to perform quality control throughout the procedure.

Subsequently, reads were clustered into OTUs using VSEARCH,[22] generating an OTU table. This OTU table contained OTU numbers and read counts per OTU number for each sample. Column names were adjusted using Microsoft Excel, removing the # signs and removing all the text behind the _M, _S and _T in the sample names.

The sequences of all OTU centroids were displayed in a second, multifasta file. These centroids were used for taxonomic identification with BLAST[26] after which the Lowest Common Ancestor tool was applied,[27] resulting in the LCA table.

Afterwards, both tables were used for further downstream analysis in R[28] and, together with the metadata table, concatenated into 1 data frame.

Then, students considered whether sampling depth was sufficient to estimate sample biodiversity. For this purpose, rarefaction curves were generated using the Vegan package.[29] Biodiversity was quantified using alpha diversity (Shannon index) and / or beta diversity (Bray–Curtis dissimilarity and Jaccard distance) estimators. Finally, dependent on the research questions raised by the students, different statistical tests were performed, including T-tests and ANOVA. These tests were performed on bacterial counts or diversity indices.

All R scripts are available in the supplementary information R markdown stored at Zenodo.

3 | RESULTS

The course was performed annually in a period of 2 years. Due to the largeness of the coral microbiome data set and the availability of data on many host and environmental parameters, a wide range of different research questions were proposed by the different student groups. Examples included (1) the effect of water temperature on the balance between beneficial and pathogenic bacteria in the microbiome, (2) the relation between coral skeletal density and presence of oxygen producing microorganisms, (3) the effect of contact with cyanobacteria on microbiome composition and (4) the relation between depth and presence of cyanobacteria in the microbiome.

Since each student group conducted its own distinct research project and used different data samples, subsequent analyses yielded very different results for the different student groups. For instance, we observed high variability in the DNA sequence quality results after FastQC analysis, necessitating different trimming steps for different student groups (Figure 1).

Another example of data variability included the rarefaction curves. Different groups obtained different curves which indicated different sampling depths (Figure 2).

Rarefaction curves reaching the plateau phase indicate that enough reads were sequenced and taxonomically classified to assess the total biodiversity within a sample. Thus, dependent on the rarefaction curves, some student groups had to discard samples for further analyses.

At the end of the course, student project groups presented their findings, which was followed by a discussion with 2 teachers. Each student project group presented and discussed their findings separately from the other groups. During these presentations and discussions, both teachers graded each of the performance criteria for every individual student ($n = 47$).

In Figure 3, scores for the performance criteria for application, data analysis and data generation are displayed. Performance criteria were graded as sufficient if the student demonstrated mastery of the criterion at the application level of Bloom's revised taxonomy.[12,13] Overall, during the presentations, each student presented a part of the research project. To assess individual performance of each student for every performance criterion, teachers asked questions about the other criteria as well. For instance, when a student had presented the bioinformatics part, care was taken to test this student for the other criteria in the assessment rubric. This way, each student was individually tested at the application level of Bloom's revised taxonomy for each criterion. On some occasions, one or more students within a student group did not demonstrate mastery at the application level of one or more performance criteria. Hence, these students received insufficient scores and had to retake the assessment.

In general, students performed well, with the fraction of students receiving sufficient or higher scores over 70% for all performance criteria after the first assessment.

Bonus points were given when students displayed criterium mastery at the analysis, evaluation or creation levels of Bloom's Taxonomy. Generally, the presentation

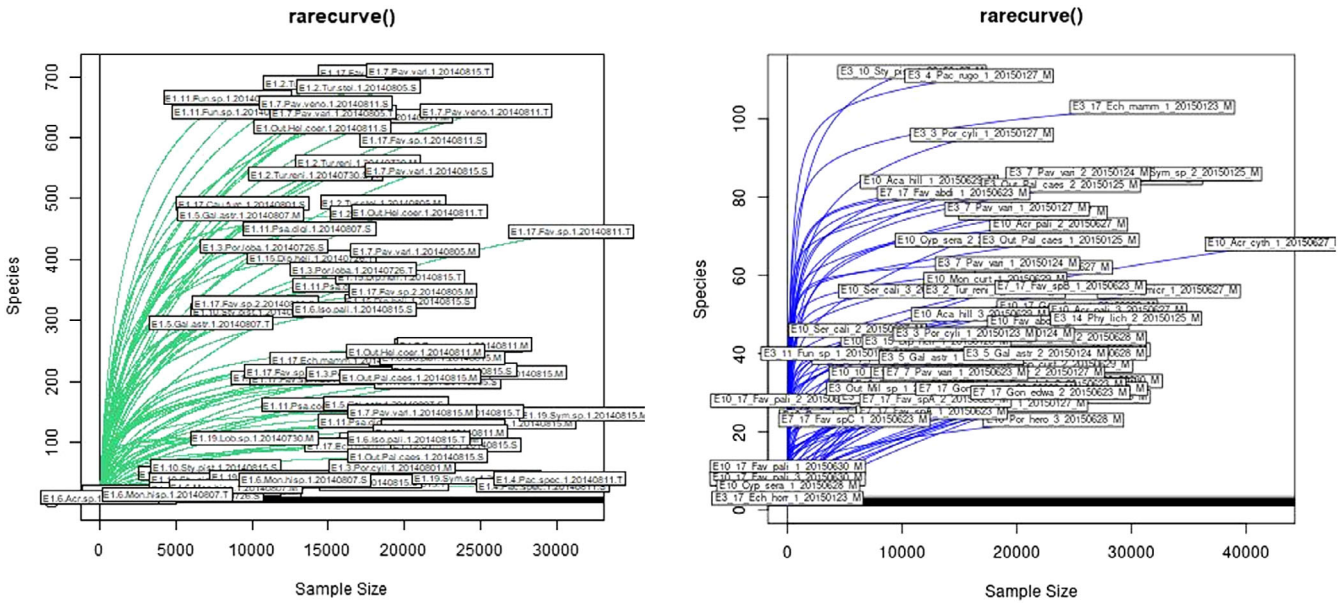

FIGURE 2 Rarefaction curves from 2 different student groups (left and right panel) for different samples after taxonomy assignment to sequenced reads. On the x-axis the number of sequenced reads is displayed, on the y-axis the number of obtained unique species. In the left panel, several samples do not reach the plateau phase, indicating underestimation of biodiversity in these samples.

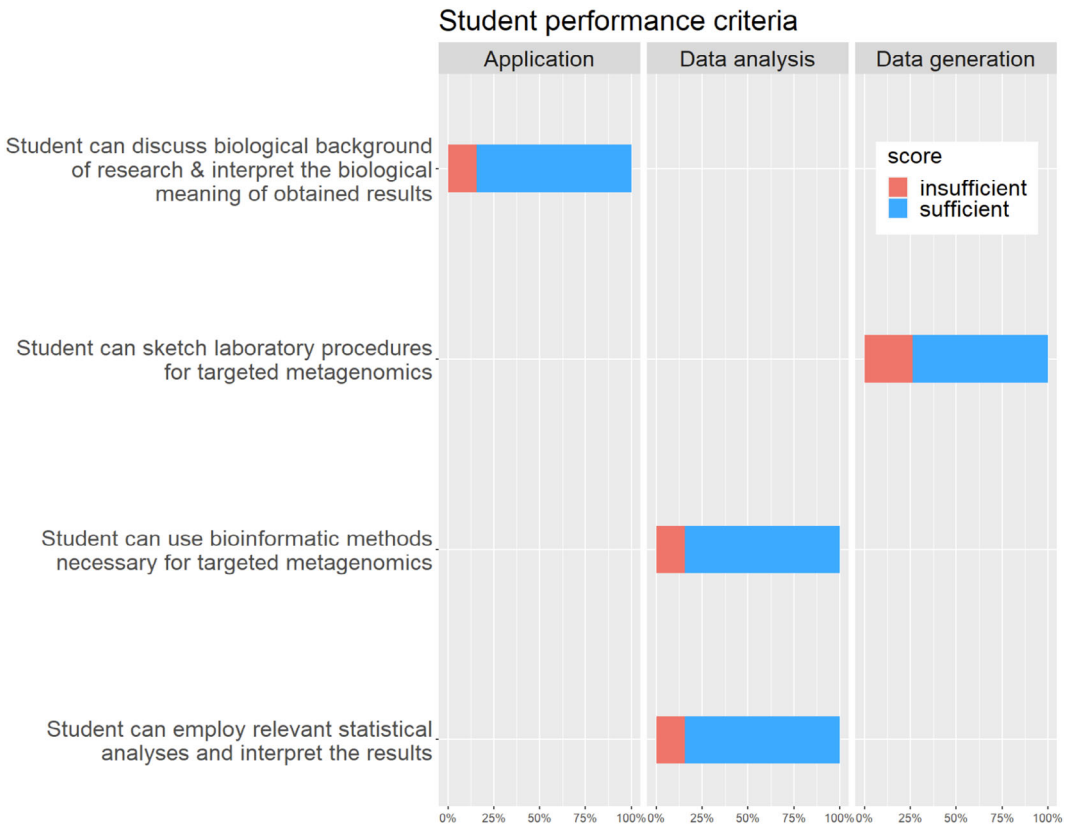

FIGURE 3 Percentages of (in)sufficient scores for the performance criteria of application, data analysis and data generation after the first assessment. Results are displayed for both years the course was offered ($n = 47$).

itself and the discussion with the entire student group were used to determine whether bonus points were awarded. For example, a group of students received a bonus point for the *bioinformatic methods criterium* since they suggested that increasing the percentage of similarity in the OTU clustering algorithm could lead to more accurate taxonomic identifications. Other students indicated that beside using the V4 region, sequencing other variable regions of the 16S gene could also increase the accuracy of taxonomic identifications, resulting in a bonus point for the *laboratory procedures criterium*. Additionally, several students developed research questions based on metadata that was not present in the original publication of Pollock[6] but which could be found in online databases. An example included the effect of microplastic pollution on the composition of the microbiome. Hence, a bonus point was granted for the *biological background criterium*.

After the presentations were held, a questionnaire was distributed to the students. This questionnaire contained items corresponding to 2 hallmarks which measure students' perception of doing original scientific research, that is, 'discovery' and 'iteration'. In Figure 4, the results of these questionnaires are displayed for both years the course was offered ($n = 28$).

Interestingly, high agreement scores were obtained for the discovery hallmark, indicating high levels of student perception of scientific discovery. However, 2 out of 3 iteration dimension items showed lower agreement scores, that is, (1) time to collect & analyze additional data and (2) time to revise or repeat analyses. When asked, students indicated that they would appreciate to have more time during the lessons to work on their projects.

In addition to the survey, students generally acknowledged the added value of working with big data sets using the Galaxy and RStudio platforms, indicating that they could use these skills in other research contexts as well.

On several occasions, students complained about encountering errors when working with these platforms. However, students appreciated the fact that in the end they had learned how to cope with these errors. For instance, errors were obtained when trying to perform FastQC on unmerged read files, most likely caused by a software bug. Errors were also encountered when mistakes were made when providing FLASH with filenames containing a trailing space at the end of the fastq extension. Similarly, errors were obtained if such a trailing space was present at the rear end of primer sequences which were imported into CutAdapt.

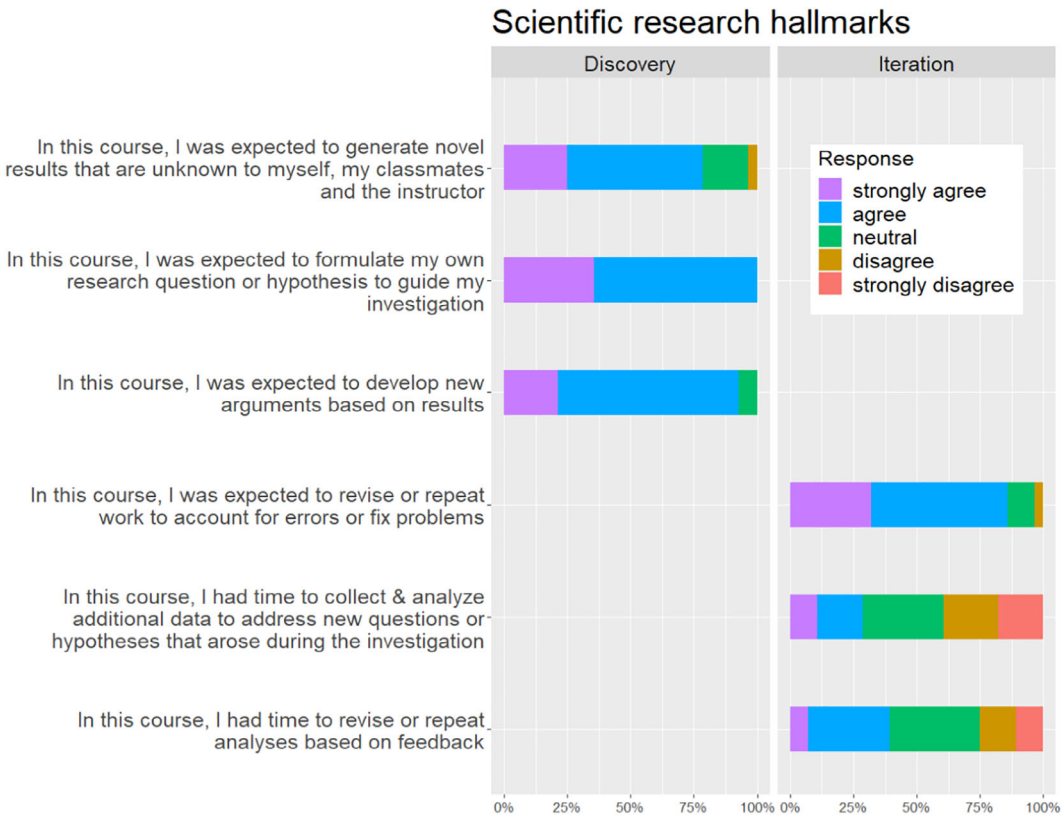

FIGURE 4 Results of the questionnaires measuring the discovery and iteration hallmark. Results are displayed for both years the course was offered ($n = 28$). Students indicated how much they agreed with the different statements provided in the questionnaire.

Additionally, students indicated that they would appreciate it if more time was allocated for learning to work with RStudio and for learning about statistical analyses.

Throughout the course, the teachers noted that students experienced difficulties dealing with potential confounders in their research project.

4 | DISCUSSION/CONCLUSIONS

We developed an inquiry-based course using an mCURE format. The availability of a large number of host and environmental coral microbiome metadata enabled us to have students perform self-designed research projects. Hence, the current course contained an authentic research context in which the answer to the research question was unknown to both students and teachers. Most likely, this facilitated the scientific discovery process, as shown by the high scores for this hallmark.

On the other hand, (1) the lower agreement scores for the iteration hallmark and (2) remarks made by students to allocate more time for working on their projects during the scheduled classes, showed that students were less able to address new questions and formulate new hypotheses. This indicated a perception of time limitation to revise analyses and research questions. However, the study load of the course included time for working on the projects outside class contact hours. Therefore, stimulating students to invest more time outside of class may improve the iteration process. This is an important aspect of performing scientific research and can potentially be achieved by scheduling out-of-class sessions for student project groups to work on their projects. In order to stimulate time investment, we aim to enhance students' project ownership by incorporating peer feedback in our course. This could be organized by setting up several research meetings in which student groups present their progress to each other in the presence of a teaching assistant. Due to the variety in research objectives and subsequent data analysis in the projects, each student group will present different findings and conclusions. We expect this to strongly drive peer feedback on all 4 performance criteria and augment iteration.

Furthermore, one major discussion topic during the presentations involved the concept of confounders. Students found it difficult to comprehend how to statistically correct for confounders. If more time is committed to address confounders during class, iteration may be enhanced as well.

Our course aimed to provide students with data analytical skills for targeted metagenomics. The results indicate that this was facilitated by the current course since the majority of students achieved the intended learning outcomes. A large part of metagenomics data analysis consists of several general next generation sequencing (NGS) processing and analysis methods that are also used in other research contexts. Hence, the current course and infrastructure equips students with transferable data analytical skills, preparing them for other research areas in which NGS data is used as well. Additionally, the graphic user interface bypasses the command line programming skills, allowing our students to fully focus on data analysis rather than programming. However, both Galaxy and RStudio platforms occasionally generated errors, illustrating the fact that troubleshooting is also necessary when using graphic interfaces and data analytical platforms. Thus, our course facilitates expectation management with regard to the ease of use of data analytical platforms and graphic user interfaces.

Although microbiome data are compositional and should be transformed with log ratio transformations,[30] we used very basic statistical analyses throughout the course since it was designed for undergraduate students. However, course complexity could be increased for a more advanced audience by expanding the statistical analyses with methods that are more appropriate for compositional data. These include ALDEx2[31] and ANCOM-BC.[32] Additionally, at present we used OTU clustering for data reduction. Complementing this analysis with a denoising method in which Amplicon Sequence Variants (ASVs) are generated[33] could also increase course complexity.

In conclusion, the publicly available coral microbiome data set enabled us to develop an inquiry-based course in which students performed authentic data analysis research projects. The mCURE format enhanced students' metagenomics data analytical skills and stimulated students' perception of scientific discovery as well. In the future, attention should be paid to the iteration hallmark. Stimulating the students to more closely examine confounders, adjust research questions and formulate new hypotheses based on data analysis results could stimulate iteration.

DATA AVAILABILITY STATEMENT

The data that were used in this study are available in The Global Coral Microbiome Project—Australia at https://doi.org/10.6084/m9.figshare.c.3855466.v2, Reference 6. These data were derived from the following resources available in the public domain: Nat Commun 9, 4921 (2018), https://doi.org/10.1038/s41467-018-07275-x.

ORCID

M. C. Morsink https://orcid.org/0009-0001-5876-923X

REFERENCES

1. Ingliss LK, Edwards RA. How metagenomics has transformed our understanding of bacteriophages in microbiome research. Microorganisms. 2022;10:1671.
2. Avramovska O, Rokop ME. A low-cost cure for CUREs: an undergraduate microbiology course engaging students in authentic research using publicly available datasets. Biochem Mol Biol Educ. 2023;52:106–16.
3. Gao L, Guo M. A course-based undergraduate research experience for bioinformatics education in undergraduate students. Biochem Mol Biol Educ. 2023;51:189–99.
4. Dahlberg CL, Wiggins BL, Lee SR, Leaf DS, Lily LS, Jordt H, et al. A short, course-based research module provides metacognitive benefits in the form of more sophisticated problem solving. J Coll Sci Teach. 2019;48(4):22–30.
5. Hanauer DI, Nicholes J, Liao F-Y, Beasley A, Henter H. Short-term research experience (SRE) in the traditional lab: qualitative and quantitative data on outcomes. CBE Life Sci Educ. 2018;17:1–14.
6. Pollock FJ, McMinds R, Smith S, Bourne DG, Willis BL, Medina M, et al. Coral-associated bacteria demonstrate phylosymbiosis and cophylogeny. Nat Commun. 2018;9:4921.
7. Oregon State University. Global Coral Microbiome Video Series. Oregon State Productions. 2017 Available from: https://films.oregonstate.edu/global-coral-microbiome-video-series
8. Vega Thurber Lab. The global coral microbiome project (GCMP). Oregon State University. 2014 Available from: https://vegathurberlab.wixsite.com/microbiology/research
9. Siegwald L, Touzet H, Lemoine Y, Hot D, Audebert C, Caboche S. Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. PLoS One. 2017;12(1):e0169563.
10. Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, et al. Defining operational taxonomic units using DNA barcode data. Philos Trans R Soc Lond Ser B Biol Sci. 2005;360(1462):1935–43.
11. RStudio Team. RStudio: Integrated Development for R. RStudio. PBC, Boston, Massachusetts. 2020 Available from: http://www.rstudio.com/
12. Krathwohl DR. A revision of Bloom's taxonomy: an overview. Theory Pract. 2002;41(4):212–8.
13. Stapleton-Corcoran E. Bloom's taxonomy of educational objectives. 2023 [accessed 2024 Jun 17]. Available from: https://teaching.uic.edu/blooms-taxonomy-of-educational-objectives/
14. Corwin LA, Runyon C, Robinson A, Dolan EL. The laboratory course assessment survey: a tool to measure three dimensions of research-course design. CBE Life Sci Educ. 2015;14:1–11.
15. Pavan-Kumar A, Gireesh-Babu P, Lakra WS. DNA metabarcoding: a new approach for rapid biodiversity assessment. J Cell Sci Molecul Biol. 2015;2(1):111.
16. Johnson JS, Spakowicz DJ, Hong BY, Petersen LM, Demkowicz P, Chen L, et al. Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis. Nat Commun. 2019;10:5029.
17. Community G. The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 2022;50:345–51.
18. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data 2010. Original tool available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Updated version available from https://github.com/daduijker/fastqc_jb_naturalis or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_fastqc/5ada1d599cd7
19. Magoč T, Salzberg SL. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics. 2011;27(21):2957–63. Original tool available from: https://github.com/naturalis/galaxy-tool-flash. Updated version available from https://github.com/daduijker/galaxy-tool-flash or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_flash/5b5fc04db237
20. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;17(1):10–2. Original tool available from: https://github.com/naturalis/galaxy-tool-cutadapt. Updated version available from https://github.com/daduijker/galaxy-tool-cutadapt or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_cutadapt/c3717346bb6f
21. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27(6):863–4. Original tool available from https://github.com/naturalis/galaxy-tool-prinseq-sequence-trimmer. Updated version available from https://github.com/daduijker/galaxy-tool-prinseq-sequence-trimmer or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_prinseq_sequence_trimmer/5b1bee1bf320
22. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ. 2016;4:e2584 Original tool available from https://github.com/naturalis/galaxy-tool-vsearch-pipeline/. Updated version available from https://github.com/daduijker/galaxy-tool-make-otu-table or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_make_otu_table/ea9885dddff4
23. Dietz M. Single point rubric idea presented at INTASC academy. Milwaukee, Wisconsin: Alverno College; 2000.
24. Fluckiger J. Single point rubric: a tool for responsible student self-assessment. Teach Edu Fac Publ. 2010;76(4):18–25.
25. The BLFS Development Team. ManageZIP. 1999–2023 Original tool available from: https://github.com/naturalis/galaxy-tool-manage-zip. Updated version available from https://github.com/daduijker/galaxy-tool-manage-zip or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_manage_zip/d55ee156b272
26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinf. 2009;10:421 Original tool available from https://github.com/naturalis/galaxy-tool-BLAST. Updated version available from https://github.com/daduijker/galaxy-tool-BLAST or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_blast/309a4f560f31
27. Beentjes KK, Speksnijder AGCL, Schilthuizen M, Hoogeveen M, Pastoor R, van der Hoorn BB. Increased performance of DNA metabarcoding of macroinvertebrates by taxonomic sorting. PLoS One. 2019;14(12):e0226527 Original tool available from https://github.com/naturalis/galaxy-tool-lca. Updated version available from https://github.com/daduijker/galaxy-tool-lca or https://toolshed.g2.bx.psu.edu/view/duijker.d/naturalis_lca/33460e756f42
28. R Core Team. R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria. 2021 Available from: https://www.R-project.org/
29. Oksanen J, Simpson G, Blanchet F, Kindt R, Legendre P, Minchin P, et al. Vegan: community ecology package 2022. R

package version 2.6–4. Available from: https://CRAN.R-project.org/package=vegan
30. Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Modeling and analysis of compositional data. London: JohnWiley & Sons; 2015.
31. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2:15.
32. Lin H, Peddada SD. Analysis of compositions of microbiomes with bias correction. Nat Commun. 2020;11:3514.
33. Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639–43.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Morsink MC, van Schaik EN, Bossers K, Duijker DA, Speksnijder AGCL. Metagenomics education in a modular CURE format positively affects students' scientific discovery perception and data analytical skills. Biochem Mol Biol Educ. 2025;53(3):311–20. https://doi.org/10.1002/bmb.21888