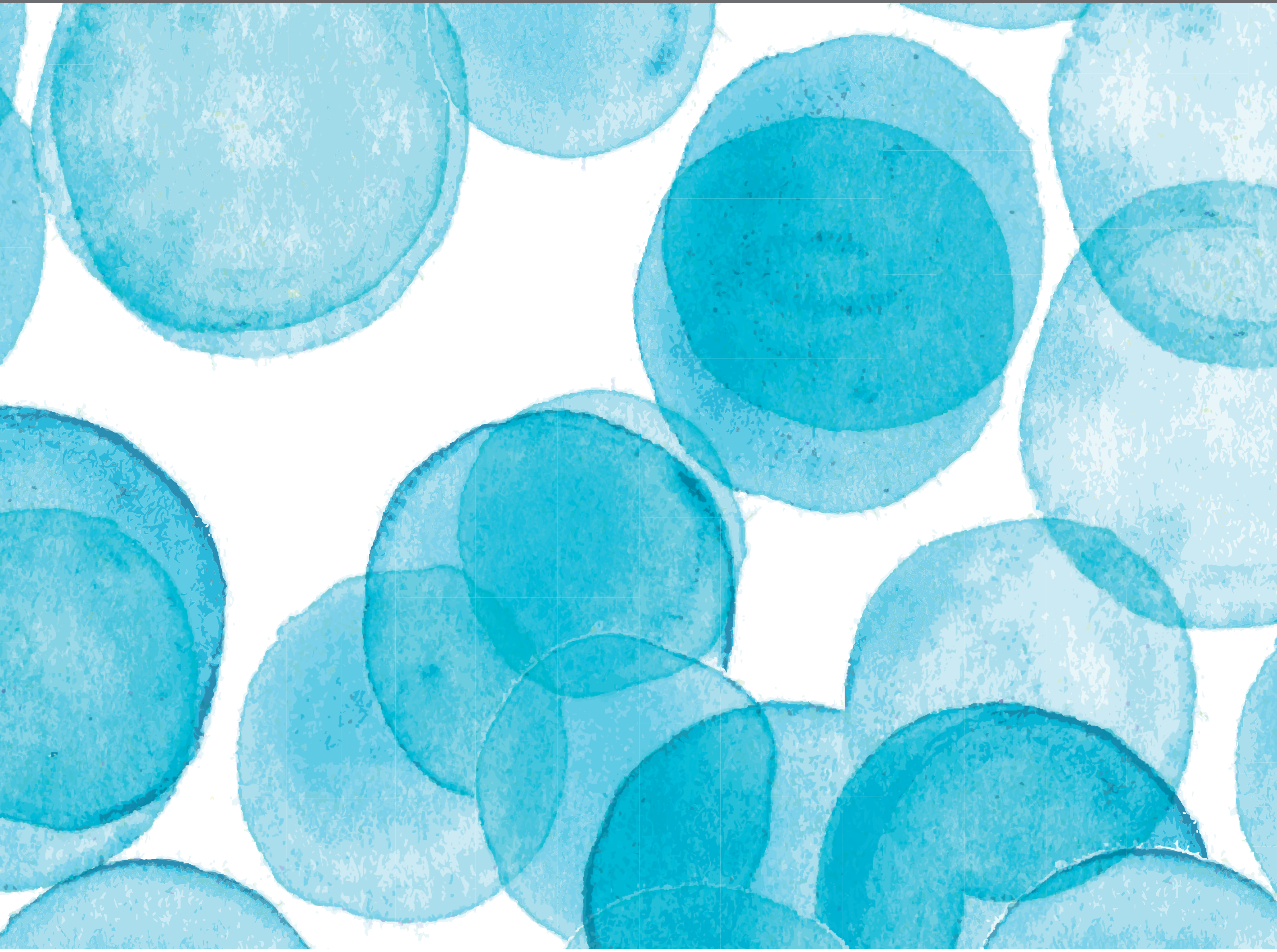
CURRICULUM APPLICATIONS IN MICROBIOLOGY: BIOINFORMATICS IN THE CLASSROOM

EDITED BY: Mel Crystal Melendrez, Brad W. Goodner, Christopher Kvaal,
C. Titus Brown and Sophie Shaw

PUBLISHED IN: Frontiers in Microbiology and Frontiers in Genetics



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88971-268-7
DOI 10.3389/978-2-88971-268-7

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

CURRICULUM APPLICATIONS IN MICROBIOLOGY: BIOINFORMATICS IN THE CLASSROOM

Topic Editors:
Mel Crystal Melendrez, Anoka-Ramsey Community College, United States
Brad W. Goodner, Hiram College, United States
Christopher Kvaal, St. Cloud State University, United States
C. Titus Brown, UC Davis, United States
Sophie Shaw, Cardiff and Vale University Health Board, United Kingdom

Citation: Melendrez, M. C., Goodner, B. W., Kvaal, C., Brown, C. T., Shaw, S., eds. (2021). Curriculum Applications In Microbiology: Bioinformatics in the Classroom. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88971-268-7

Table of Contents

frontiers in Microbiology

EDITORIAL
published: 01 July 2021
doi: 10.3389/fmicb.2021.705233

Editorial: Curriculum Applications in Microbiology: Bioinformatics in the Classroom

Melanie Crystal Melendrez[1]*, *Sophie Shaw*[2], *C. Titus Brown*[3], *Brad W. Goodner*[4] and *Christopher Kvaal*[5]

[1] Anoka-Ramsey Community College, Cambridge, MN, United States, [2] Centre for Genome Enabled Biology and Medicine, University of Aberdeen, Aberdeen, United Kingdom, [3] Department of Population Health and Reproduction, University of California, Davis, Davis, CA, United States, [4] Hiram College, Hiram, OH, United States, [5] Department of Biology, St. Cloud State University, St. Cloud, MN, United States

Keywords: bioinformatics and computational biology, undergraduate research, microbiology, curriculum—undergrad and postgrad, software, genomics, science technology engineering mathematics

OPEN ACCESS

Edited by:
Iain Sutcliffe,
Northumbria University,
United Kingdom

Reviewed by:
Joao Carlos Setubal,
University of São Paulo, Brazil
David W. Ussery,
University of Arkansas for Medical
Sciences, United States

****Correspondence:***
Melanie Crystal Melendrez
melanie.melendrez-vallard@
anokaramsey.edu

Specialty section:
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

Received: *04 May 2021*
Accepted: *07 June 2021*
Published: *01 July 2021*

Citation:
Melendrez MC, Shaw S, Brown CT,
Goodner BW and Kvaal C (2021)
Editorial: Curriculum Applications in
Microbiology: Bioinformatics in the
Classroom.
Front. Microbiol. 12:705233.
doi: 10.3389/fmicb.2021.705233

Editorial on the Research Topic

Curriculum Applications in Microbiology: Bioinformatics in the Classroom

John Naisbitt stated in his 1982 book Megatrends, "We are drowning in information but starved for knowledge." The statement, made nearly 40 years ago, seems acutely applicable in today's scientific and academic world. Reviews by Barba et al. and van Dijk et al., provide a nice historical perspective on the growth of sequencing technology. Over three decades, sequencing technology has improved greatly from 1987 when the first ABI automated sequencing machine went to market up through the mid 2010s when next generation sequencing platforms from 454 Life Sciences, Illumina and other companies were outputting up to 1,800 Gb per run (Barba et al., 2014; van Dijk et al., 2014). Technology has since progressed even further with the development of long-read and single molecule (Pacific Biosciences, Illumina, Oxford Nanopore, and 10X Genomics) sequencing systems that can output terabytes of data, per run, in a matter of days (van Dijk et al., 2018). Specifically, in the areas of genomics, proteomics, and transcriptomics, we are now producing upwards of 1 zetta-bases/year (Stephens et al., 2015). The explosion of data has increased the demand for hardware and software development to manage and analyze the data as well as qualified personnel in bioinformatics to sift through the outputs to draw meaningful conclusions. A report from Reports and Data states the global bioinformatics market is projected to reach 18.96 billion USD by 2026 (Reports and Data, 2019) and this means re-thinking not only how we store data but how we train the next generation of scientists. The greatest needs identified in various surveys compiled by the NSF, ELIXIR-UK, and EMBL-ABR include: (i) data quality and control (ii) data analysis skills in visualization and interpretation, (iii) data mining, manipulation and management, (iv) analysis reproducibility, and (v) statistics (Kanwal et al., 2017; Kim et al., 2018; Attwood et al., 2019).

The large quantity of data available for analysis in many scientific fields is both a strength and a weakness in bioinformatic analysis. There are several databases and repositories available to acquire sequence data (INSDC: ENA+DNA Data bank of Japan and Genbank SRA, GISAID) and it is essential to know that not all data may be handled the same way. This variability in the quality of data that is released to scientists and the public at large can result in low quality data being analyzed and potentially spurious conclusions. A prime example can be seen in the current pandemic where SARS-CoV-2 sequences can be obtained from several different databases and analyzed in real-time. In the haste to make genomes available, the quality of what is released has been variable with challenges in consistent nomenclature (Gozashti and Corbett-Detig, 2021), and genomes containing errors created by sequencing artifacts, sample preparation, consensus calling approaches, or contamination (De Maio et al., 2020a,b; Van Noorden, 2021). Analysis of these genomes, even among experts, can lead to data misinterpretation, over-interpretation, and confusion on important topics such as SARS-CoV-2 origins (Andersen et al., 2020; Zhang et al., 2020; Wacharapluesadee et al., 2021). However, it is important to recognize these challenges to big data quality control, management, analysis, and reproducibility are not unique to SARS-CoV-2 but are systemic in many subfields of bioinformatics such as microbiome analysis (Katsnelson, 2019), metatranscriptomics (Shakya et al., 2019), and RNA-seq (Simoneau et al., 2021) analysis.

Recognition of these short-comings of big data acquisition, quality control, reproducibility, management, and analysis across bioinformatics disciplines have led to improvements in next generation sequencing workflows and quality control (Charre et al., 2020; De Maio et al., 2020b; Van Damme et al., 2021), efforts to use provenance, github, and docker containers to facilitate reproducibility (Kanwal et al., 2017; Kulkarni et al., 2018; Menegidio et al., 2018; Bolyen et al., 2019; Wercelens et al., 2019), nomenclature clarification (Rambaut et al., 2020, 2021), an increased emphasis on workflow automation (Reiter et al., 2021), and database curation and consolidation (Heard et al., 2021). As the tools and refinements to how scientists manage and analyze data continue to move forward, the demand for qualified big data analysts, statisticians, and bioinformaticians is increasing rapidly (Gómez-López et al., 2019; Terry, 2019; Tammi et al., 2020).

To address the need for big data management and analytical skill sets, many university programs have emerged offering certificates, Master's degrees and even Ph.D. degrees in the field of bioinformatics. The most recent guidance on bioinformatics core competencies has highlighted the importance of developing informatics skill sets early in the undergraduate curriculum (Welch et al., 2014; Vincent and Charette, 2015; Mulder et al., 2018; Wilson Sayres et al., 2018; Tractenberg et al., 2019). However, few curricula at the undergraduate level introduce big data analytics and bioinformatics systematically and many students graduate without a full understanding of what bioinformatics is or how it can be used to solve biological problems.

Several bioinformatic disciplines: i.e., metagenomics, genome construction/annotation, pathogen discovery, phylogenetics, metabolomics, and transcriptomics, have well-known workflows that teach valuable skills in data management, analytics, interpretation, and troubleshooting, but have yet to be translated to the classroom. Additionally, while many microbiology instructors recognize the importance of integrating more research, real-world datasets, and informatics into the classroom, they feel their training is inadequate, their curriculum is already over-full, or students do not appear particularly interested or prepared for such topics in the course (Williams et al., 2019). For many instructors, it can be daunting to put together bioinformatics curriculum modules if you are not familiar with the software or general topics within bioinformatics that students can explore.

This Research Topic focuses on bringing both research and educational communities together; encouraging researchers to translate their studies and pipelines into teaching tools and curriculum, and encouraging educators to dive into messy real-world datasets when teaching microbiology. Much of the challenge in implementing research or bioinformatics focused modules in the undergraduate classroom revolves around implementation. Bennet discusses strategies for blending your classroom to incorporate undergraduate research and bioinformatics modules into your curriculum design (CURE). Bennet takes a "workshop" or "project-based" approach to introduce the often complicated and challenging topic of RNA-Seq analysis (Conesa et al., 2016; Bennett) and discusses the long term outcomes for students experiencing this particular CURE as well as educational applications.

Another challenge in implementation of bioinformatics workflows in the classroom is the requirement for background experience in a variety of topics, both biological and computational. While many biology instructors are comfortable introducing and expanding on biological topics related to research and design, they are less comfortable discussing the computing aspects of bioinformatic analysis such as coding languages, data quality control, and data management. Several papers in the special topic discuss data workflows that utilize Microsoft Excel (Mitchell et al.; Hankey et al.; Kruchten). While many individuals working in advanced bioinformatic analysis may cringe at the idea of excel data analysis and tables, this program is well-used in classrooms globally and many instructors are comfortable with implementing data analysis and mathematical functions in the Excel environment. Programs such as Excel can provide a bridge between the user-friendly, GUI-based interfaces and the world of command-line applications (CLI). Krutchen, in particular, offers a nice comparison of the use of Excel vs. the R statistical language when analyzing metagenomic datasets and this may serve as motivation for instructors to explore other programming and CLI-based workflows (Kruchten). Topic papers in the methods category show instructors how to introduce, discuss, and/or implement coding languages and CLI-based bioinformatics in their classroom such as python/R for microbiome analysis (Rosen and Hammrich), basic command line proficiency in analyzing genome scale data for microbial isolates (Petrie and Xie), and how to conduct metagenomic analysis using the R statistical package (Kruchten) or QIIME, which contains its own language and syntax for implementation (Bolyen et al., 2019; Rosen and Hammrich).

Additional topic papers contain curricular designs for introducing and teaching a variety of bioinformatic analysis skills in the classroom without the need for teaching additional modules on coding skills. Topic papers discuss gene discovery and genome annotation using a variety of free web-accessible programs (Amatore et al.; Koury et al.; Martins et al.), microbiome analysis using PUMAA (Mitchell et al.), 16S amplicon identification using DNALC and NCBI-BLAST databases and the DNA Subway software program (Tawde and Williams; Williams et al., 2014), metagenomics analysis using MG-RAST and the MicrobiomeAnalyst program (Meyer et al., 2019; Chong et al., 2020; Baker et al.), phage hunting using PHASTER and iTOL programs (Arndt et al., 2019; Letunic and Bork, 2019; Martinez-Vaz and Mickelson), and Cancer data analysis using The Cancer Genome Atlas (TCGA; Hankey et al.).

To account for variable quality of datasets analyzed in the classroom, special topic studies used already published, curated, data from the Cancer Genome Atlas or GENI-ACT toolkit (Hankey et al.; Koury et al.) or pre-curated genomes for genome prediction exercises rather than raw data from databases (Martins et al.). Studies that made use of raw data or minimally curated data utilized embedded quality assessment tools and discussion modules on data cleanup within their curriculum methods and workflow (Amatore et al.; Kruchten; Tawde and Williams; Petrie and Xie; Baker et al.; Mitchell et al.). However, discussion and training on quality and data management needs to be ongoing; especially given data is being reused for educational purposes. Wilkinson and colleagues proposed the FAIR guiding principles to support the accessibility, findability, interoperability, and reusability of data in science (FAIR principles for data stewardship, 2016; Wilkinson et al., 2016) and there are workshops available on how to get started with "FAIR data" (https://mdibl.org/course/applied-bioinformatics-2021/). These principles should be considered widely in addition to the use of provenance and contained workflows or containers such as those mentioned earlier. In the overwhelming world of big data analysis it will be important for instructors to translate complex analysis techniques to their novice students; a key challenge is balancing quality and rigor with simplicity.

Finally the topic papers extend into existing scientific communities, where skills needed for data analysis are lacking by a large number of current researchers and professionals tasked to conduct bioinformatics analysis and interpretation. Therefore, workshops to educate existing researchers and laboratory personnel, from the level of graduate student to principal investigator, have become more frequent. These professional development and "train the trainer" workshops are attractive in that they are intensive short term experiences that teach very specific skill sets related to computational jobs in the field (McGrath et al., 2019). The Physalia courses (https://www.physalia-courses.org/), Cold Spring Harbor Laboratory Short Courses (https://meetings.cshl.edu/courses.html), and various workshops offered by the Evolution and Genomics training team (http://evomics.org/workshops/) and the MDI Biological Laboratory (https://mdibl.org/course/bioinformatics-t3-2021/) are a few examples of training experiences that undergraduates, graduates and professional personnel can use to augment their skill sets in the field of genetic analysis and computational biology. Internationally, these short term intensive educational opportunities, putting bioinformatics in the classroom, have proven useful in bringing staff and personnel up to date on the latest technologies and analysis capabilities to increase job performance and institute mission output. The BioCANET network in Central America (Orozco et al., 2013), Walter Reed Army Institute of Research (WRAIR) in South America (Pollett et al., 2016), H3Africa consortium in Africa (Aron et al., 2017; Ahmed et al., 2018; Shaffer et al., 2019), and APBioNet in Asia (Khan et al., 2013; Ahmad et al., 2019) are all aimed at increasing capacity for educational and research institutions in the areas of data management, systems administration, biostatistics, genome wide association studies, next generation sequencing analysis, metagenomics, and virology; and all have had success using this educational format. Our topic supports this educational "workshop" format of continued training for professional personnel through a paper by Maljkovic Berry et al., on implementation of a bioinformatics workshop for laboratory and research personnel at a US Department of Defense laboratory located in Kisumu, Kenya.

Special topic papers detail curriculum set up and implementation of bioinformatics modules or coding contain supplemental material to facilitate readers in their own implementation of the module or curriculum design in their classroom. We hope to convey through this topic the versatility of instructional designs that can be used to teach students at all levels of expertise, from high school to established professionals, how to leverage the strength of coding, software, and computational analysis to accomplish their research goals and further scientific teaching and discovery.

AUTHOR CONTRIBUTIONS

MM contributed to the initial concept of the special topic and wrote the initial draft of the editorial. All authors contributed to the design and proposal of the special topic, participated as active guest editors to manage manuscripts and oversee the special topic, and contributed to editorial revision and final approval for submission.

ACKNOWLEDGMENTS

We would like to thank all of the authors who submitted to this special topic, committed to the furthering of academic creativity, excellence, and rigor in the challenging and virtual instructional world of SARS-CoV-2 (COVID-19). To you and all of our educators globally, you are indispensable.

REFERENCES

Ahmad, S., Gromiha, M. M., Raghava, G. P. S., Schönbach, C., and Ranganathan, S. (2019). APBioNet's annual International Conference on Bioinformatics (InCoB) returns to India in 2018. *BMC Genomics* 19:266. doi: 10.1186/s12864-019-5582-8

Ahmed, A. E., Mpangase, P. T., Panji, S., Baichoo, S., Souilmi, Y., Fadlelmola, F. M., et al. (2018). Organizing and running bioinformatics hackathons within Africa: the H3ABioNet cloud computing experience. [version 2; peer review: 2 approved, 1 approved with reservations]. *AAS Open Res.* 1:9. doi: 10.12688/aasopenres.12847.1

Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C., and Garry, R. F. (2020). The proximal origin of SARS-CoV-2. *Nat. Med.* 26, 450–452. doi: 10.1038/s41591-020-0820-9

Arndt, D., Marcu, A., Liang, Y., and Wishart, D. S. (2019). PHAST, PHASTER and PHASTEST: tools for finding prophage in bacterial genomes. *Brief. Bioinformatics* 20, 1560–1567. doi: 10.1093/bib/bbx121

Aron, S., Gurwitz, K., Panji, S., and Mulder, N. (2017). H3abionet: developing sustainable bioinformatics capacity in Africa. *EMBnet J.* 23:886. doi: 10.14806/ej.23.0.886

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., and Schneider, M. V. (2019). A global perspective on evolving bioinformatics and data science training needs. *Brief. Bioinformatics* 20, 398–404. doi: 10.1093/bib/bbx100

Barba, M., Czosnek, H., and Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. *Viruses* 6, 106–136. doi: 10.3390/v6010106

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Charre, C., Ginevra, C., Sabatier, M., Regue, H., Destras, G., Brun, S., et al. (2020). Evaluation of NGS-based approaches for SARS-CoV-2 whole genome characterisation. *Virus Evol.* 6:veaa075. doi: 10.1093/ve/veaa075

Chong, J., Liu, P., Zhou, G., and Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* 15, 799–821. doi: 10.1038/s41596-019-0264-1

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., et al. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biol.* 17:13. doi: 10.1186/s13059-016-0881-8

De Maio, N., Gozashti, L., Turakhia, Y., Walker, C., Lanfear, R., Corbett-Detig, R., et al. (2020a). *Updated Analysis With Data From 12th June 2020*. virological.org. Available online at: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473/11 (accessed May 28, 2021).

De Maio, N., Walker, C., Borges, R., Weilguny, L., Slodkowicz, G., and Goldman, N. (2020b). *Issues With SARS-CoV-2 Sequencing Data*. virological.org. Available online at: https://virological.org/t/issues-with-sars-cov-2-sequencing-data/473 (accessed May 28, 2021).

FAIR principles for data stewardship. (2016). *Nat. Genet.* 48:343. doi: 10.1038/ng.3544

Gómez-López, G., Dopazo, J., Cigudosa, J. C., Valencia, A., and Al-Shahrour, F. (2019). Precision medicine needs pioneering clinical bioinformaticians. *Brief. Bioinformatics* 20, 752–766. doi: 10.1093/bib/bbx144

Gozashti, L., and Corbett-Detig, R. (2021). Shortcomings of SARS-CoV-2 genomic metadata. *BMC Res. Notes* 14:189. doi: 10.1186/s13104-021-05605-9

Heard, E., Birney, E., Apweiler, R., Bloomberg, N., Cochrane, G., Lauer, K., et al. (2021). *Open Letter: Support Data Sharing for COVID-19*. COVID-19 Data Portal. Available online at: https://www.covid19dataportal.org/support-data-sharing-covid19 (accessed May 28, 2021).

Kanwal, S., Khan, F. Z., Lonie, A., and Sinnott, R. O. (2017). Investigating reproducibility and tracking provenance - a genomic workflow case study. *BMC Bioinformatics* 18:337. doi: 10.1186/s12859-017-1747-0

Katsnelson, A. (2019). Standards seekers put the human microbiome in their sights. *ACS Cent. Sci.* 5, 929–932. doi: 10.1021/acscentsci.9b00557

Khan, A. M., Tan, T. W., Schönbach, C., and Ranganathan, S. (2013). APBioNet-transforming bioinformatics in the Asia-Pacific region. *PLoS Comput. Biol.* 9:e1003317. doi: 10.1371/journal.pcbi.1003317

Kim, Y.-M., Poline, J.-B., and Dumas, G. (2018). Experimenting with reproducibility: a case study of robustness in bioinformatics. *Gigascience* 7:giy077. doi: 10.1093/gigascience/giy077

Kulkarni, N., Alessandr,ì, L., Panero, R., Arigoni, M., Olivero, M., Ferrero, G., et al. (2018). Reproducible bioinformatics project: a community for reproducible bioinformatics analysis pipelines. *BMC Bioinformatics* 19:349. doi: 10.1186/s12859-018-2296-x

Letunic, I., and Bork, P. (2019). Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res.* 47, W256–W259. doi: 10.1093/nar/gkz239

McGrath, A., Champ, K., Shang, C. A., van Dam, E., Brooksbank, C., and Morgan, S. L. (2019). From trainees to trainers to instructors: *s*ustainably building a national capacity in bioinformatics training. *PLoS Comput. Biol.* 15:e1006923. doi: 10.1371/journal.pcbi.1006923

Menegidio, F. B., Jabes, D. L., Costa de Oliveira, R., and Nunes, L. R. (2018). Dugong: a Docker image, based on Ubuntu Linux, focused on reproducibility and replicability for bioinformatics analyses. *Bioinformatics* 34, 514–515. doi: 10.1093/bioinformatics/btx554

Meyer, F., Bagchi, S., Chaterji, S., Gerlach, W., Grama, A., Harrison, T., et al. (2019). MG-RAST version 4-lessons learned from a decade of low-budget ultra-high-throughput metagenome analysis. *Brief. Bioinformatics* 20, 1151–1159. doi: 10.1093/bib/bbx105

Mulder, N., Schwartz, R., Brazas, M. D., Brooksbank, C., Gaeta, B., Morgan, S. L., et al. (2018). The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Comput. Biol.* 14:e1005772. doi: 10.1371/journal.pcbi.1005772

Orozco, A., Morera, J., Jiménez, S., and Boza, R. (2013). A review of bioinformatics training applied to research in molecular medicine, agriculture and biodiversity in Costa Rica and Central America. *Brief. Bioinformatics* 14, 661–670. doi: 10.1093/bib/bbt033

Pollett, S., Leguia, M., Nelson, M. I., Maljkovic Berry, I., Rutherford, G., Bausch, D. G., et al. (2016). Feasibility and effectiveness of a brief, intensive phylogenetics workshop in a middle-income country. *Int. J. Infect. Dis.* 42, 24–27. doi: 10.1016/j.ijid.2015.11.001

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., et al. (2021). Addendum: a dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 6:415. doi: 10.1038/s41564-021-00872-5

Rambaut, A., Holmes, E. C., O'Toole, Á., Hill, V., McCrone, J. T., Ruis, C., et al. (2020). A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407. doi: 10.1038/s41564-020-0770-5

Reiter, T., Brooks, P. T., Irber, L., Joslin, S. E. K., Reid, C. M., Scott, C., et al. (2021). Streamlining data-intensive biology with workflow systems. *Gigascience* 10:giaa140. doi: 10.1093/gigascience/giaa140

Reports and Data (2019*). Global Bioinformatics Market To Reach USD 18.96 Billion By 2026 | Reports and Data*. Available online at: https://www.globenewswire.com/news-release/2019/05/14/1823970/0/en/Global-Bioinformatics-Market-To-Reach-USD-18-96-Billion-By-2026-Reports-And-Data.html (accessed July 1, 2020).

Shaffer, J. G., Mather, F. J., Wele, M., Li, J., Tangara, C. O., Kassogue, Y., et al. (2019). Expanding research capacity in Sub-Saharan Africa through informatics, bioinformatics, and data science training programs in Mali. *Front. Genet.* 10:331. doi: 10.3389/fgene.2019.00331

Shakya, M., Lo, C.-C., and Chain, P. S. G. (2019). Advances and challenges in metatranscriptomic analysis. *Front. Genet.* 10:904. doi: 10.3389/fgene.2019.00904

Simoneau, J., Dumontier, S., Gosselin, R., and Scott, M. S. (2021). Current RNA-seq methodology reporting limits reproducibility. *Brief. Bioinformatics* 22, 140–145. doi: 10.1093/bib/bbz124

Stephens, Z. D., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M. J., et al. (2015). Big data: astronomical or genomical? *PLoS Biol.* 13:e1002195. doi: 10.1371/journal.pbio.1002195

Tammi, M., Fei, K. T., and Low, L. (2020). *Career Outlook: Data and Bioinformatics Scientists to 2026. Dr Martti Knowledge is Power*. Available online at: https://bioinformaticshome.com/blog/jobs_landscape_2018_2020.html (accessed June 1, 2021).

Terry, M. (2019). *Careers in Bioinformatics: Hot and Getting Hotter*. BioSpace. Available at: https://www.biospace.com/article/careers-in-bioinformatics-hot-and-getting-hotter/ (accessed June 1, 2021).

Tractenberg, R. E., Wilkinson, M. R., Bull, A. W., Pellathy, T. P., and Riley, J. B. (2019). A developmental trajectory supporting the evaluation and achievement of competencies: articulating the Mastery Rubric for the nurse practitioner (MR-NP) program curriculum. *PLoS ONE* 14:e0224593. doi: 10.1371/journal.pone.0224593

Van Damme, R., Hölzer, M., Viehweger, A., Müller, B., Bongcam-Rudloff, E., and Brandt, C. (2021). Metagenomics workflow for hybrid assembly, differential coverage binning, metatranscriptomics and pathway analysis (MUFFIN). *PLoS Comput. Biol.* 17:e1008716. doi: 10.1371/journal.pcbi.1008716 van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends Genet.* 30, 418–426. doi: 10.1016/j.tig.2014.07.001

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., and Thermes, C. (2018). The third revolution in sequencing technology. *Trends Genet.* 34, 666–681. doi: 10.1016/j.tig.2018.05.008

Van Noorden, R. (2021). Scientists call for fully open sharing of coronavirus genome data. *Nature* 590, 195–196. doi: 10.1038/d41586-021-00305-7

Vincent, A. T., and Charette, S. J. (2015). Who qualifies to be a bioinformatician? *Front. Genet.* 6:164. doi: 10.3389/fgene.2015.00164

Wacharapluesadee, S., Tan, C. W., Maneeorn, P., Duengkae, P., Zhu, F., Joyjinda, Y., et al. (2021). Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. *Nat. Commun.* 12:972. doi: 10.1038/s41467-021-21240-1

Welch, L., Lewitter, F., Schwartz, R., Brooksbank, C., Radivojac, P., Gaeta, B., et al. (2014). Bioinformatics curriculum guidelines: toward a definition of core competencies. *PLoS Comput. Biol.* 10:e1003496. doi: 10.1371/journal.pcbi.1003496

Wercelens, P., da Silva, W., Hondo, F., Castro, K., Walter, M. E., Araújo, A., et al. (2019). Bioinformatics workflows with nosql database in cloud computing. *Evol. Bioinform. Online* 15:1176934319889974. doi: 10.1177/1176934319889974

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

Williams, J., McKay, S., Khalfan, M., Hilgert, U., Lauter, S., Jeong, E.-S., et al. (2014). *DNA Subway - An Educational Bioinformatics Platform for Gene and Genome Analysis: DNA Barcoding, and RNA-Seq.* Vancouver, BC: American Society of Animal Science (ASAS). Available online at: https://www.asas.org/docs/default-source/wcgalp-proceedings-oral/227_paper_9798_manuscript_948_0.pdf (accessed April 20 2021).

Williams, J. J., Drew, J. C., Galindo-Gonzalez, S., Robic, S., Dinsdale, E., Morgan, W. R., et al. (2019). Barriers to integration of bioinformatics into undergraduate life sciences education: a national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. *PLoS ONE* 14:e0224288. doi: 10.1371/journal.pone.0224288

Wilson Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS ONE* 13:e0196878. doi: 10.1101/170993

Zhang, T., Wu, Q., and Zhang, Z. (2020). Probable Pangolin Origin of SARS-CoV-2 associated with the COVID-19 outbreak. *Curr. Biol.* 30, 1346–1351.e2. doi: 10.1016/j.cub.2020.03.022

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

OPINION
published: 12 August 2020
doi: 10.3389/fmicb.2020.01728

The CURE for the Typical Bioinformatics Classroom

Jennifer A. Bennett [1,2]*

[1] Department of Biology and Earth Science, Otterbein University, Westerville, OH, United States, [2] Biochemistry and Molecular Biology Program, Otterbein University, Westerville, OH, United States

Keywords: microbiology, bioinformatics, genomics, STEM, education, CURE, classroom

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Andy Perkins,
Mississippi State University,
United States
Rachel Horak,
American Society for Microbiology,
United States

****Correspondence:***
Jennifer A. Bennett
jbennett@otterbein.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *21 January 2020*
Accepted: *02 July 2020*
Published: *12 August 2020*

Citation:
Bennett JA (2020) The CURE for the Typical Bioinformatics Classroom. Front. Microbiol. 11:1728. doi: 10.3389/fmicb.2020.01728

INTRODUCTION

Bioinformatics is a field that combines biology and computer science to investigate relevant current topics such as annotation of the Human Genome Project and other genomes, protein structure and function, examination of disease processes and personalized medicine, evolutionary relationships and conservation genetics (as reviewed in Luscombe et al., 2001; Can, 2014). Today it is difficult to find a published article in biochemistry and molecular biology/microbiology that does not have a bioinformatics component. Thus, it is important for majors in the life sciences to have exposure to bioinformatics in the curriculum. A blended course format (Garrison and Kanuka, 2004) allows for short lectures and hands-on learning in the classroom combined with computer-based learning outside of the classroom in the form of literature searches, computer tutorials, and independent research that use information from the course applied to specific projects. The topics and techniques that students learn will help them navigate the vast amounts of information that are freely available in databases and inform them about how to manage that data and derive new information.

Bioinformatics is often a course taught in a workshop/computer lab format and in many instances a primarily lecture format. Adding a CURE, or Course-based Undergraduate Research Experience, has many advantages over traditional labs and lectures. CUREs have features of inquiry-based learning and also allow for participation in a larger project and community of researchers (Auchincloss et al., 2014). The Bioinformatics course described here incorporates many of the best teaching practices that have been called upon by numerous professional societies and are included in the 2011 Vision and Change Report (Brewer and Smith, 2011).

OVERVIEW OF THE BLENDED LEARNING WORKSHOP FORMAT

My upper-level bioinformatics course was taught in the format of a workshop that was designed to keep students engaged both inside and outside of the classroom. At the beginning of each week I led the first workshop with what I referred to as a "Bio Byte," a newsworthy current event in the field of bioinformatics that was often in the form of a short video clip. Each class incorporated a mini-lecture of ∼10 min that described the bioinformatics tool(s) to be used that day, descriptions about when and why the particular tool should be used and the relevance to society. A guided demonstration ensued with students practicing the software and later applying it to their gene of interest, in a self-paced, hands-on experience. In the open lab format at the end of each class, students could practice the tools and converse with each other to gain additional knowledge in addition to asking for instructor feedback.

Beginning with student outcomes in mind, my Bioinformatics course was created using a backward design approach (Wiggins and McTighe, 2005). The desired outcomes of the course were for students to (1) Learn how to use various bioinformatics techniques, (2) Apply the techniques to answer pertinent research questions, (3) Explain how bioinformatics is connected to wet-lab experimentation, and (4) Generate and report novel data. The blended learning or hybrid format allowed for the majority of the semester to be taught in two 55-min sessions per week with additional time to work on the project and assignments that were posted to the blackboard learning site. It also allowed for some open computer lab days where students could work and ask questions. The atmosphere was relaxed for the students and fun to teach.

A PROJECT-BASED APPROACH TO BIOINFORMATICS

With the intent of introducing large data sets and independent research, students were given access to RNA Sequencing data generated from my microbial genetics research program. The RNA-Seq experiment compared wild-type bacterial gene expression with that of a mutant under the same conditions (Bennett, unpublished). Students examined the data and chose a gene that was not already annotated in the spreadsheet. The job of each student for the rest of the semester was to characterize this chosen gene by conducting independent research and applying the various bioinformatics tools that they would be learning in the classroom. The fact that each student was assigned a different gene for investigation allowed a unique combination of both creativity and design-sharing within the classroom. Students were able and encouraged to help one another in an environment that promoted improvements to create better quality portfolios and more advanced analyses.

Instead of exams, each student was responsible for submitting a final portfolio showing mastery of the various bioinformatics tools that they learned in class as applied to their specific gene. The portfolio was the culminating project for the bioinformatics course, and required figures and corresponding legends in the format of a publication along with annotations about the techniques used. A minimum set of expectations was given for the gene analysis portfolio, however students had the freedom to characterize their gene in ways that extended beyond these expectations. The result was multiple student portfolios that showcased additional tools not discussed in class and advanced functions of the tools that we had covered.

LONG TERM POSITIVE OUTCOMES

The portfolio specifically allowed students to demonstrate their ability to apply each bioinformatics tool to their chosen gene of interest. The independent research project and resulting portfolio could be listed on student applications, curricula vitae, and resumes. Additionally, each of the 12 students in the class was required to give a short oral presentation, describing what they had learned about their gene and its possible role in the cell. Three of the students also seized the opportunity to present their bioinformatics research, and their abstracts were accepted for poster presentation at the Ohio Branch Meeting of the American Society for Microbiology. Two of the students decided to pursue Ph.D. programs in bioinformatics/computational biology, in large part crediting the experience that they had in the bioinformatics course. Two additional students also entered Ph.D. programs bringing with them bioinformatics knowledge they learned in the course that will be extremely useful to their dissertation research. There was one sophomore student in the spring 2018 course, who is about to graduate and is currently applying to Ph.D. programs where she hopes to combine her skills in bioinformatics and microbiology. Spring 2018 was the second time that I taught Bioinformatics. I first taught the course as an experimental course in Fall 2013 with only six students enrolled. During the first iteration of the course, it was taught in a very similar workshop format with a final portfolio, only without the exposure to the RNA-Seq dataset. Students chose a gene of interest that was uncharacterized from the *Streptomyces* genome (Bentley et al., 2002) and presented on that gene. One of the initial six students chose to enter a graduate program to pursue bioinformatics research based on his bioinformatics experience at Otterbein and is about to graduate with a Ph.D. in Biochemistry with dissertation research entirely in the area of bioinformatics.

The features that make the guided workshop approach with a novel independent research project so successful are hands-on direct application, student ownership of an important project, the ability of students to customize their portfolio and pursue advanced topics, and the ability to communicate their data in both written and oral form. Wilson Sayres et al. (2018) published a set of bioinformatics core competencies in 2018 that are readily achieved in the framework of the bioinformatics course described here. The students must read and evaluate the primary literature and directly apply bioinformatics techniques. Their final product is a source of pride. They produce data that has the possibility of publication and they are part of a larger community of researchers within their classroom and in the field of microbial genetics. The data continues to make an impact as it influences future studies in my research program.

Concepts and techniques that the students learned and applied in the Bioinformatics course included BLAST, multiple sequence alignments (Clustal), phylogenetics, domain mapping (SMART and Pfam), analyses of protein-protein interactions, and protein modeling (RaptorX, Cn3D, and pyMOL). Some of the proteins were also 3D printed in collaboration with engineering students at The Point, Otterbein's STEAM innovation center. Connections to wet-lab experimentation and other disciplines were introduced throughout the semester, including the next steps in the analysis pipeline. For example, after using bioinformatics to identify and begin the characterization of novel genes, some of the gene expression data obtained through RNA-Seq could be verified using Real Time PCR. Genes could be deleted using such tools as the Lambda Red System (Datsenko and Wanner, 2000) or CRISPR-Cas9 (Wang et al., 2016) to determine the mutant phenotype and thus provide experimental evidence for the role of the gene as compared to that predicted using bioinformatics.

The following examples illustrate the long-term impact of bioinformatics skills learned by undergraduates on my research program. Undergraduate research students in my lab have successfully completed Real Time PCR experiments for two genes identified in the RNA-Seq experiments introduced into the bioinformatics class, and a student is currently using the Lambda Red System to delete a gene of interest identified in the bioinformatics course. Another gene identified in our RNA-seq experiment was fortuitously disrupted in a transposon experiment and we are studying it and other similar genes identified using bioinformatics in light of the RNA-Seq data. Hull et al. (2012) serves as a past example of undergraduate co-authors contributing significant research to a published paper using bioinformatic analyses. One student co-author contributed entirely through bioinformatics research, stemming from a portfolio completed as an independent study in my lab. This is the format that I have continued to employ in my bioinformatics course. The published paper incorporates the following skills performed by undergraduate co-authors: construction of genetic maps, primer design, sequence analysis, BLAST to identify similar genes in the genome and to identify orthologues in other species, sequence alignments, and protein domain mapping. As part of my bioinformatics course, students employ these skills that our lab typically uses to present and publish, in addition to many more techniques such as those listed above.

DISCUSSION OF EDUCATIONAL APPLICATIONS

Bioinformatics is a course that lends itself especially well to a blended course and workshop format. The application of techniques to a novel gene of interest in a progressive order kept students engaged with a sense of strong ownership. Only a computer lab or student laptops are required. My entire course made use of databases and software that were freely available to the public. The bioinformatics tools were easily accessible and relatively simple to learn for instructors with little bioinformatics background because the programs employed Graphical User Interfaces (GUIs) that do not require programming knowledge. However, all of these exercises can be easily extended to include introductory scripting for students. The introduction of some command lines into the course is advantageous for students to better understand how their data is being obtained.

The course format used in the bioinformatics course described here can be transferred to other portions of the biology and microbiology curricula. A small portion of a course can be devoted to a bioinformatics analysis of research data using any of the techniques from the full bioinformatics course, allowing students to make important contributions to large projects. The course used a bacterial genome of interest to my research program, but the same techniques can be applied to any organism, based on the interests of the instructor. In my course, I introduced RNA-seq data, but we also have a proteomics project where we used mass spectroscopy to identify binding proteins discovered in bead capture experiments. Students could easily have investigated these proteins instead. Any bioinformatics course or module could also include a functional genomics component where students are involved in complimentary wet-lab experimentation along with the bioinformatics analyses. In summary, this type of bioinformatics CURE can readily be used in full courses or modified for modules within a course to actively engage students in meaningful research with high learning gains.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

FUNDING

Funding for course design was made available through an Otterbein University Blended Learning Course Transformation Grant and training was received through the Otterbein Center for Teaching and Learning Blended Course Transformation Institute.

ACKNOWLEDGMENTS

I gratefully acknowledge the students from my bioinformatics classes, the Otterbein Center for Teaching and Learning, and The Point at Otterbein University. This article is an extension of an abstract published in The FASEB Journal for a poster presentation at the Annual Meeting of the American Society for Biochemistry and Molecular Biology in Orlando (2019).

REFERENCES

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., et al. (2014). Assessment of course-based undergraduate research experiences: a meeting report. *CBE Life Sci. Edu.* 13, 29–40. doi: 10.1187/cbe.14-01-0004

Bentley, S. D., Chater, K. F., Cerdeño-Tárraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., et al. (2002). Complete genome sequence of the model actinomycete *Streptomyces* coelicolor A3(2). *Nature* 417, 141–147. doi: 10.1038/417141a

Brewer, C. A., and Smith, D. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

Can, T. (2014). Introduction to bioinformatics. *Methods Mol. Biol.* 1107, 51–71. doi: 10.1007/978-1-62703-748-8_4

Datsenko, K. A., and Wanner, B. L. (2000). One-step inactivation of chromosomal genes in *Escherichia coli* K-12 using PCR products. *Proc. Natl. Acad. Sci. U. S. A.* 97, 6640–6645. doi: 10.1073/pnas.120163297

Garrison, D., and Kanuka, H. (2004). Blended learning: uncovering its transformative potential in higher education. *Internet Higher Edu.* 7, 95–105. doi: 10.1016/j.iheduc.2004.02.001

Hull, T. D., Ryu, M. H., Sullivan, M. J., Johnson, R. C., Klena, N. T., Geiger, R. M., et al. (2012). Cyclic Di-GMP phosphodiesterases RmdA and RmdB are involved in regulating colony morphology and development in *Streptomyces coelicolor*. *J. Bacteriol.* 194, 4642–4651. doi: 10.1128/JB.00157-12

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? a proposed definition and overview of the field. *Methods Inform. Med.* 40, 346–358. doi: 10.1055/s-0038-1634431

Wang, H., La Russa, M., and Qi, L. S. (2016). CRISPR/Cas9 in genome editing and beyond. *Ann. Rev. Biochem.* 85, 227–264. doi: 10.1146/annurev-biochem-060815-014607

Wiggins, G., and McTighe, J. (2005). *Understanding by Design*. Alexandria, VA: ASCD.

Wilson Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS ONE* 13:e0196878. doi: 10.1371/journal.pone.0196878

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

PERSPECTIVE
published: 10 September 2020
doi: 10.3389/fmicb.2020.578600

A Curricular Bioinformatics Approach to Teaching Undergraduates to Analyze Metagenomic Datasets Using R

*Anne E. Kruchten**

Department of Biology, The College of St. Scholastica, Duluth, MN, United States

Biologists with bioinformatic skills will be better prepared for the job market, but relatively few biology programs require bioinformatics courses. Inclusion in the curriculum may be hindered by several barriers, including lack of faculty expertise, student resistance to computational work, and few examples in the pedagogical literature. An 8-week wet-lab and *in silico* research experience for undergraduates was implemented. Students performed DNA purification and metagenomics analysis to compare the diversity and abundance of microbes in two samples. Students sampled snow from sites in northern Minnesota and purified genomic DNA from the microbes, followed by metagenomic analysis. Students used an existing metagenomic dataset to practice analysis skills, including comparing the use of Excel versus R for analysis and visualization of a large dataset. Upon receipt of the snow data, students applied their recently acquired skills to their new dataset and reported their results via a poster. Several outcomes were achieved as a result of this module. First, YouTube videos demonstrating hands-on metagenomics and R techniques were used as professional development for faculty, leading to broadened research capabilities and comfort with bioinformatics. Second, students were introduced to computational skills in a manner that was intentional, with time for both introduction *and* reinforcement of skills. Finally, the module was effectively included in a biology curriculum because it could function as either a stand-alone course or a module within another course such as microbiology. This module, developed with Course-based Undergraduate Research Experience guidelines in mind, introduces students and faculty to bioinformatics in biology research.

Keywords: bioinformatics, curriculum, undergraduate, R, big data, metagenomics, biology

OPEN ACCESS

Edited by:
Sophie Shaw,
University of Aberdeen,
United Kingdom

Reviewed by:
Carlos Christopher Goller,
North Carolina State University,
United States
Gaurav Arora,
Gallaudet University, United States

****Correspondence:***
Anne E. Kruchten
akruchten@css.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *30 June 2020*
Accepted: *12 August 2020*
Published: *10 September 2020*

Citation:
Kruchten AE (2020) A Curricular Bioinformatics Approach to Teaching Undergraduates to Analyze Metagenomic Datasets Using R. Front. Microbiol. 11:578600. doi: 10.3389/fmicb.2020.578600

INTRODUCTION

In 1920, botanist Hans Winkler coined the term "genome" as a fusion of the words gene and chromosome (Winkler, 1920). Since that time, the "omics" fields have exploded, creating such terms as "pseudome" (the population of pseudogenes), "translatome" (the population of proteins in the cell, weighted by their abundance level), and many others that are increasingly becoming a normal part of the lexicon for biologists[1]. The term "bioinformatics" was defined by

[1] http://bioinfo.mbb.yale.edu/what-is-it/omes/

Luscombe et al. (2001), as "conceptualizing biology in terms of molecules (in the sense of Physical chemistry) and applying "informatics techniques" (derived from disciplines such as applied maths, computer science and statistics) to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications." Undergraduates in biology should be trained in this field to successfully compete in the job market and make vital contributions to the biological sciences as their careers mature.

The *Vision and Change: A Call to Action* report of 2011 (Brewer and Smith, 2011) emphasized that undergraduate biology students should have competence in computational and systems level approaches and the ability to use large databases. Only a small fraction of institutions offer a full undergraduate bioinformatics program (Mellon, 2020), but several offer courses on bioinformatics. In the state of Minnesota, 50% of public and private school biology departments offer a bioinformatics course in their curriculum but none appear to require it for the degree. This may reflect a lack of expertise among faculty to teach the course.

In 2016-17, 116,759 bachelor's degrees in biology were conferred to graduates in the United States (US-DOE, 2020). Among bachelor's degree holders 25–29 years old, biology graduates' annual salaries were not significantly different than the median annual income of all degree holders of $50,600, but computer and information science degree holders had an annual income of $70,100, well above the median income (NCES, 2020). A slight increase in biology-related computer information jobs is predicted, suggesting that biology majors would be well-served to develop computer information skills to complement their biology degrees (Araneo et al., 2017).

Bioinformatics is a broad field that encompasses gene alignment tools, crowdsourcing approaches, metagenomics, and many others. Rather than lecturing about bioinformatics, many groups have chosen to incorporate bioinformatics tools into CUREs (Course-based Undergraduate Research Experiences). In CUREs, students are working in classes on research projects of interest to the broader scientific community (Auchincloss et al., 2014). On the CUREnet website[2], several bioinformatics CUREs have been shared for faculty adoption and participation, including a CRISPR-Cas9 project[3], a study of iron uptake in insects[4], Genome Solver: Microbial Comparative Genomics[5], and the Genomics Education Partnership (GEP)[6]. These programs, and many others across the country, teach students a variety of gene-based bioinformatics approaches including using BLAST, multiple gene alignment, primer design, and many others. Students develop strong gene analysis skills while also contributing to active scientific research projects in the process.

While these CUREs develop students' genome analysis skills, other courses focus on microbiome analyses including mapping microbiomes of the human oral cavity (Wang et al., 2015) and crowdsourcing datasets of antibiotic resistance in microbes (Freeman et al., 2016; Small-World, 2020). Students in these courses develop research skills such as bacterial culturing, sterile technique, PCR, and hypothesis building. Few projects, however, teach undergraduates the computational skills required to statistically analyze "big data" in biological fields.

Computational skills are required to analyze and find patterns in big data, which includes the four Vs: volume of data, velocity of processing the data, variability of data sources, and veracity of the data quality. Applications of big data analysis can be found everywhere, but for biologists especially important applications include genome sequencing, ecological studies (such as of microbiomes), and health care information (Li and Chen, 2014). Graduates of biology programs have opportunities for employment in any of these fields but may not have the important computational skills in parallel with wet lab or field biology skills to be successful in big data fields. There seem to be few CUREs or similar programs published in the literature that provide instructions for how faculty can implement curricular modules to help students develop these big data skills.

Several groups have outlined a series of bioinformatics competencies for life scientists, including CourseSource (the Bioinformatics Learning Framework) (Rosenwald et al., 2016), the Curriculum Task Force of the International Society of Computational Biology (ISCB) Education Committee (Mulder et al., 2018), and the Network for Integrating Bioinformatics into Life Sciences Education (NIBLSE) (Wilson Sayres et al., 2018). Building on previous work from both CourseSource and ISCB, NIBLSE surveyed instructors at US institutions and used the data to develop a list of Core Competencies for Undergraduate Life Scientists. While many of the core competencies focus on genomics-based bioinformatics skills, several of the competencies are addressed by the work in this project. The competencies are listed below (Wilson Sayres et al., 2018), and the bolded ones are addressed by the approach in this project:

- **C1. Explain the role of computation and data mining in addressing hypothesis-driven and hypothesis-generating questions within the life sciences.**
- C2. Summarize key computational concepts, such as algorithms and relational databases, and their applications in the life sciences.
- C3. Apply statistical concepts used in bioinformatics.
- C4. Use bioinformatics tools to examine complex biological problems in evolution, information flow, and other important areas of biology.
- C5. Find, retrieve, and organize various types of biological data.
- **C6. Explore and/or model biological interactions, networks, and data integration using bioinformatics.**
- **C7. Use command-line bioinformatics tools and write simple computer scripts.**
- C8. Describe and manage biological data types, structure, and reproducibility.

[2] https://serc.carleton.edu/curenet/whatis.html

[3] https://serc.carleton.edu/curenet/collection/213026.html

[4] https://serc.carleton.edu/curenet/institutes/boulder/examples/207018.html

[5] https://serc.carleton.edu/curenet/collection/218072.html

[6] https://serc.carleton.edu/curenet/collection/215335.html

- C9. Interpret the ethical, legal, medical, and social implications of biological data.

Importantly, both the CourseSource Bioinformatics Learning Framework and the ISCB Curriculum Task Force recognize that there are different levels of users of bioinformatics curriculum, including bioinformatics *engineers*, bioinformatics *scientists*, and bioinformatics *users*. The approach described here is geared toward bioinformatics *users*, including both faculty who are interested in learning about these tools and students who will be moving forward into a variety of careers in research, medicine, education, and others. This course module is a starting point for introducing students to low level Bloom's taxonomy areas such as knowledge and comprehension of bioinformatics. It is hoped that this introduction will spark an interest in students to learn more about the field and become bioinformatics scientists. This approach is also intended to provide an entry point for faculty to begin developing new courses in bioinformatics within their undergraduate biology programs and collaborate with colleagues in computer science fields to pool interests and resources.

BIOINFORMATICS COURSE MODULE

In response to the need for a big data CURE, I have developed an 8 week course that meets for two 2-hour sessions weekly in which students gain hands-on experience using R and Excel to analyze large datasets. To mimic an authentic research experience as closely as possible, the 10 students work as a research group as they discuss the literature, develop hypotheses, and plan experiments. Individuals or pairs are responsible for collecting samples and performing the actual sample preparation and experiments. Data analysis is completed individually and then discussed and improved in the full research group. While this course was developed as a stand-alone experience, it could easily be incorporated as a module in a broader full length course.

The primary student learning outcome for this course was to develop students' data science skills using Excel and R. The premise of this research course was to perform a metagenomic analysis of the microbiota in two different snow samples. To accomplish this research project, students perform a literature review, develop hypotheses, collect and prepare samples, perform metagenomic sequencing (through a third party vendor), learn data analysis skills, and present their research findings via a poster presentation. Secondary student learning outcomes for this course include those described in the CURE network: making discoveries of interest to the broader scientific community, an iterative work experience, communication of their findings, and development of scientific research skills (CUREnet, 2020).

Weeks 1 and 2: Literature Review, Hypothesis Development, and Sampling

Table 1 highlights the main activities completed in the course, beginning with a literature review. Because the primary learning outcome for this course is the development of R and Excel skills, the instructor can assist in the literature review process by developing the initial research question and providing some preliminary resources to begin the discussion. In this project, I developed the initial research question of "how does the bacterial population vary between two snow samples from different locations on campus?" and provided several primary and secondary articles about microbiomes, microorganisms often found in snow, and bacterial abundance and diversity. Students used these resources as jumping off points to find more sources (usually PDFs, websites, and videos) which were collected in a class Google folder. Students visually mapped these sources into three broad categories on the whiteboard: "snow," "microbiomes," and "microbial diversity." After a group discussion, each student was responsible for developing an individual literature review from these and other sources they found.

This fast-paced literature review process leads to the development of a research question, hypothesis, and sampling procedure. Metagenomic analysis with our vendor takes 3–4 weeks, so it was essential to collect and prepare samples right away to allow time for the primary student learning outcome of developing skills in Excel and R. To this end, after discussion, most of the class agreed upon the same research question and hypothesis, with slight variations that could be accommodated within the sampling and sample preparation processes. Our research question asked if the microbiota of snow samples would differ between an area heavily trafficked by both foot and automobile traffic compared to campus trails primarily traveled by snowshoe. Most students hypothesized that the area with both foot and automobile traffic would have more bacteria overall and more diversity of bacteria. Students demonstrated their understanding of the field and our research question development by submitting a draft of an introduction for their final poster project (see **Supplementary Material Section 5** for teaching materials).

Sampling and sample purification were relatively simple and inexpensive. Students used 50 ml plastic conical tubes (VWR 89039-656) to collect three samples spaced at one meter intervals along a line at each of the two sites for 3 days in a row. To purify microbial DNA from the samples, the snow was melted and filtered through a 0.2 micron polyester membrane using an Aeropress coffee press[7]. The membranes containing the filtered microorganisms were then processed using the Qiagen DNeasy PowerWater kit (Qiagen 14900-50-NF). After confirming the presence of bacterial DNA via PCR with a 16S primer set (idtdna.com; 16S rRNA For #51-01-19-06, 16S rRNA Rev 51-01-19-07), the samples were sent for metagenomic sequencing off campus.

Weeks 3 and 4: Introducing Metagenomics, Big Data, and R

The first step in teaching students about bioinformatics was to guide them through an understanding of how metagenomic sequencing works and how the dataset was generated. A prerequisite for this course was a one semester Foundations in Biology course covering the essentials of molecular biology, including central dogma concepts such as DNA, RNA, base pairing, replication, and transcription.

[7] https://aeropress.com

TABLE 1 | Flow of the bioinformatics course.

Week #	Course topics
1	Literature review *The literature review is initiated by the instructor to save time and is further developed by students.*
2	Hypothesis development, sampling, and sample preparation *Students use their literature review to develop their hypothesis, identify sampling methods, and prepare DNA samples, allowing a pairing of wet lab skills with in silico activities.* *When wet lab resources are unavailable, this step can be completed by the instructor or replaced with an existing publicly available dataset.*
3	Understanding metagenomic sequencing and big data *Students build on foundational knowledge of DNA from prerequisite courses by viewing video material on PCR and sequencing.*
4	Introduction of Data Analysis Skills *Instruction in statistics, Excel, and R using a combination of video material and in class discussions builds a foundation of data analysis skills.*
5	Practice data analysis using an existing dataset *Students use their developing data analysis skills to mimic the instructors' actions using Excel and R to analyze an existing dataset.*
6	Reinforcing data analysis skills with snow dataset *Students apply the data analysis skills they have learned and practiced to a new dataset from samples collected on campus.*
7	Creating a Scientific Poster of Analysis of Big Data *Students showcase all the skills practiced in the course in a poster containing a research question, background material, a hypothesis, methods, results, and discussion.*
8	Poster presentations *Students complete the semester by recording a video presentation of themselves presenting their poster. If available, students also present their posters at a campus-wide research symposium.*

The **Supplementary Material** contain a list of resources used for reviewing foundational DNA and PCR knowledge (**Supplementary Material Section 1**). With this background in mind, students work to understand the polymerase chain reaction, or PCR. This foundational knowledge is essential, in part because it strips away the complexities of how we typically teach replication with emphases on all of the different enzymes (polymerases I and III, primase, ligase, helicase, etc.) and focuses on the simple concept of creating a complement sequence of DNA to the template.

After mastering PCR, students then move on to understanding DNA sequencing, beginning with Sanger sequencing. To do this, students watch a series of YouTube videos on Sanger Sequencing[8], the evolution of next-generation sequencing[9], and finally Illumina sequencing[10] used by our vendor (see **Supplementary Material Section 1** for more details). After watching the video on Illumina sequencing, students usually express a combination of fascination and confusion. To provide further practice in understanding this extremely important process, we break into student pairs and have each pair illustrate the processes of cluster formation on whiteboards using color coding. After performing a similar exercise to better understand base calling, we complete this section of the instruction by discussing how multiple overlapping DNA segments from one organism can be used to generate the sequence for the entire 16S ribosomal RNA gene.

It is common for biology students in our program to have a fear or aversion to mathematical and other quantitative or computational approaches. 65% of traditional undergraduate students enrolled in our college identify as female, 31% identify as first generation college students, 35% have family incomes less than $50,000, and 70% come from rural communities and small cities. Many students have taken the minimum mathematics courses required by the state graduation guidelines. In a study of life sciences majors conducted by Andrews and Aikens (Andrews and Aikens, 2018), both females and first generation students exhibited a lower interest in mathematics topics in biology than their counterparts, and females perceived a higher cost associated with doing math in biology than their male counterparts. They also found that students' likelihood of taking a biostatistics class was positively related to their interest and perceived utility of the course. A goal for this course module is to spark future interest in bioinformatics training, so it was important to demonstrate to students the utility of statistical analysis both for the project and their future careers.

In recognition of these factors, I began the bioinformatics instruction with a review (or novel instruction) of basic statistical analysis. To accomplish this, students first reviewed major statistical functions such as mean, median, standard deviation, standard error, *p*-values, and Student's *t*-test using a freely available resource compiled by MIT[11]. These concepts were practiced using a very simple assignment completed in pairs during class time examining the statistical significance of simple drug treatment data (see **Supplementary Material Section 2** for details). In class discussion helped to sort out problems in understanding before moving on to larger dataset analysis.

Next, students are introduced to fundamental concepts in data analysis, including data clean up and developing the research question. To facilitate this process, I provided the students with a dataset previously collected in the Boundary Waters Canoe Area

[8] https://www.youtube.com/watch?v=Jnk_4Maf5Fk

[9] https://www.youtube.com/watch?v=jFCD8Q6qSTM&t=176s

[10] https://www.youtube.com/watch?v=fCd6B5HRaZ8

[11] https://web.mit.edu/$\sim$csvoss/Public/usabo/stats_handout.pdf

Wilderness (BWCAW). This dataset included triplicate sampling of four different sample sites resulting in 12 columns of data on a spreadsheet. After metagenomic sequencing, 15,000 unique bacterial species or OTUs (operational taxonomic units) were identified in the spreadsheet rows, resulting in 180,000 unique cells of data. Given that most students' experience of using Excel to this point had been in traditional lab courses, this was by far the largest Excel file any of them had ever opened.

To make the experience less overwhelming for the students, I provide them with a version of the dataset that condensed OTUs into phyla, resulting in a dataset with 12 sampling columns and 23 rows of identified phyla. My goal was for them to be able to use Excel to average the triplicate results from each sample site and make comparisons across the data, either between the four individual sample sites or between phyla. To do this, I created a video of myself using Excel to average the sites, perform a *t*-test comparing the data between sites, and then sort the data by increasing *p*-value, thus reordering the data so that the most significant *p*-values were at the top of the list. Students then were required to repeat the actions of this video on both the phyla dataset and the OTU dataset. In doing this, students gained experience cleaning up and renaming columns, writing formulas, accessing the formula bank, sorting, and visualizing data.

Weeks 5 and 6: Practicing and Reinforcing Big Data Analysis Skills

After establishing comfort with analyzing data in Excel, we moved on to R. R is a freely available statistical computing program (The R-Foundation, 2020) used across many fields for the analysis and visualization of data. For the purposes of this course module, I wanted to introduce students to the pros and cons of using the programming language R versus using Excel both for data analysis and for data visualization, particularly for its ability to generate a heat map of large datasets. This includes establishing student knowledge, but not necessarily application, of using a command line and understanding the function of packages, bundles of shareable code created by experts in the field and freely available for use.

When students learned coding was involved, there was an immediate sense of anxiety in the room. To alleviate this stress, I returned to an approach with which the students were familiar: learning by watching videos. Just as they had learned to use Excel functions by watching me perform tasks via video, the basics of R were laid out by watching a series of publicly available YouTube videos. Many videos are available, but I chose the "R Programming for Beginners" playlist from the R Programming 101 YouTube channel[12] (see the **Supplementary Material Section 3** for a complete list of videos). In this series, the host, public health specialist Greg Martin, guides viewers through the whole process of using R, including downloading R and R Studio onto their computers, learning basic commands such as identifying variables and manipulating a preloaded dataset of health characteristics of Star Wars characters, and installing and using R packages. This playlist resonated with the students, both because of the clear instructions and because of the link to public health, a field with which many of the biology students could identify. Students watched this series of videos on their own and their sole assignment was to replicate exactly what the host did and turn in a screenshot of their final R Studio product.

Once the students achieved some initial comfort with R, I gave them a fully composed sheet of code to copy and paste into the script window of R. The code was created by modifying freely available code (Albert and Yoder, 2013), including the packages gplots, vegan, and RColorBrewer to plot data, create the heatmap, and apply a color scheme. I used this approach for three reasons. First, students did not yet have the capability to compose their own code because they didn't have enough knowledge of syntax to do what was needed. Second, because R is an open access community, students and instructors can find existing code for many functions on the internet and modify it to fit their needs. Third, by providing code that was annotated (with # lines explaining each line of code), students were able to walk through each line of code, understand the function, and run the code to achieve a final product of a heat map demonstrating the diversity and abundance of microbial samples across sampling sites in the BWCAW (**Figure 1**; full code in the **Supplementary Material Section 4**). Because the purpose of this course module was to introduce bioinformatics users to command line coding, the ability to generate a finished product was important both to increase their level of confidence in using R and in order to demonstrate the analysis capabilities available in R that were not available in Excel.

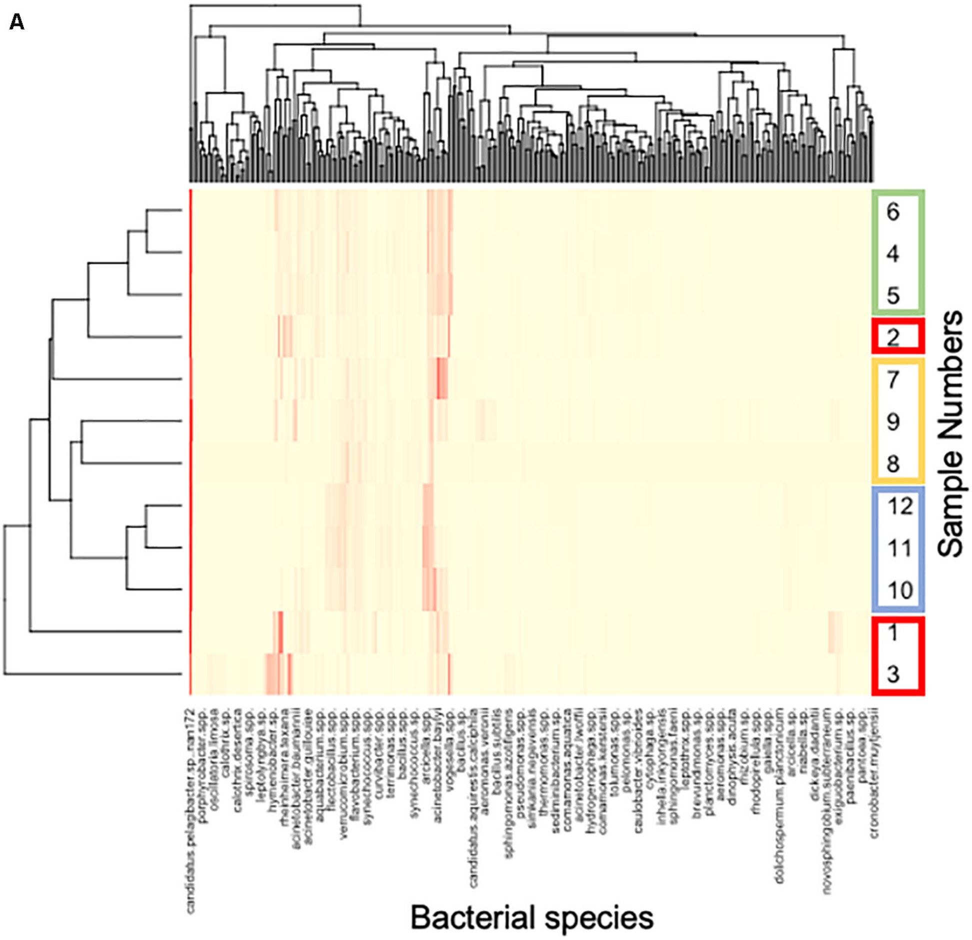
At this point in the course, students had participated in a strong introduction to data analysis using both Excel and R. They had manipulated a dataset larger than any of them had seen before and reflected on the pros and cons of each tool in analyzing the datasets. Each student had observed Excel and R being used via video and followed up with practice completing the work themselves. This iterative approach follows best practices in pedagogy where students are offered multiple opportunities to observe, practice, and learn a skill.

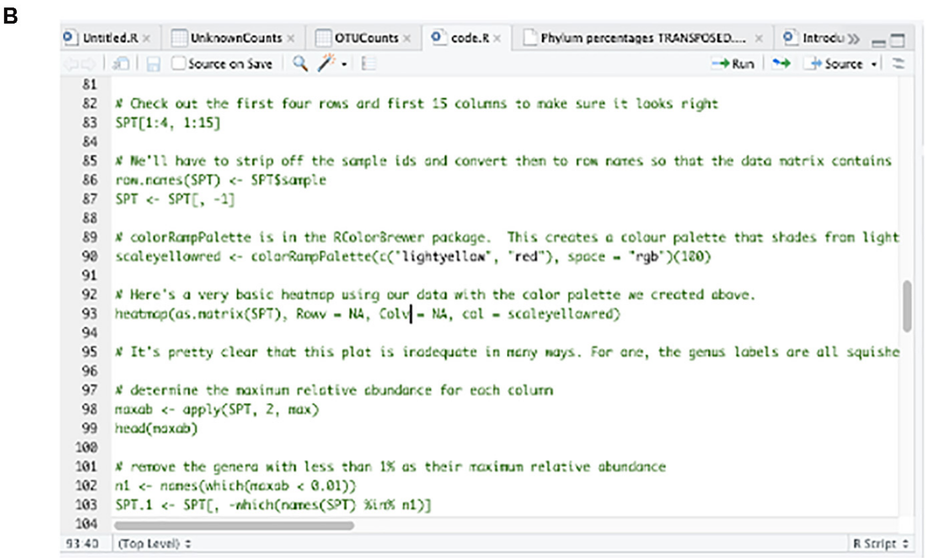
When the data from the metagenomic analysis of the snow samples was returned to us in week six, students were ready to analyze it. The final project was a standard scientific poster presentation of their background, research question, hypothesis, methods, results, and discussion. To accomplish this task, students had to return to the notes they took for the analysis of the BWCAW dataset and apply these approaches for the snow dataset. This task involved cleaning up the data, and properly labeling sample columns, and changing existing lines of code in R to import the proper.csv file, identifying columns correctly and creating an appropriate visualization. By using this iterative approach of first observing, then practicing, and finally applying, all the students were able to successfully assign the right syntax to the code and create a successful project.

Potential for Virtual Course Delivery

As presented, this process allows students to experience both wet bench and *in silico* research. However, it is important to note that the project could be modified to include only the *in silico*

[12] https://www.youtube.com/channel/UCfJyQ3P2k_SuqfxVdqIEQNw





```r

# Check out the first four rows and first 15 columns to make sure it looks right
SPT[1:4, 1:15]

# We'll have to strip off the sample ids and convert them to row names so that the data matrix contains
row.names(SPT) <- SPT$sample
SPT <- SPT[, -1]

# colorRampPalette is in the RColorBrewer package.  This creates a colour palette that shades from light
scaleyellowred <- colorRampPalette(c("lightyellow", "red"), space = "rgb")(100)

# Here's a very basic heatmap using our data with the color palette we created above.
heatmap(as.matrix(SPT), Rowv = NA, Colv = NA, col = scaleyellowred)

# It's pretty clear that this plot is inadequate in many ways. For one, the genus labels are all squishe

# determine the maximum relative abundance for each column
maxab <- apply(SPT, 2, max)
head(maxab)

# remove the genera with less than 1% as their maximum relative abundance
n1 <- names(which(maxab < 0.01))
SPT.1 <- SPT[, -which(names(SPT) %in% n1)]

```

FIGURE 1 | Example heat map and R code. Students used R to generate two heat maps in the course, first with a practice set of data from the Boundary Waters Canoe Area Wilderness which was followed by a heat map of snow sample data to reinforce skills. **(A)** Representative student-generated heat map of the BWCAW data. On the right axis, triplicate samples are boxed with corresponding colors; bacterial species' names are on the bottom axis. R-generated dendrograms are on the left and top axes. **(B)** A snapshot of the script window of R studio showing the code students used to generate the heat maps. A full copy of the code is available in **Supplementary Material**.

experience for students, as was the case in the second iteration of this course in spring 2020 due to the COVID-19 pandemic and the closure of college facilities. It would be possible to provide this experience with the many publicly available datasets, but during the college closure I chose to perform the wet bench portion myself prior to the beginning of the course so that students felt they had a more "personal" sample rather than a dataset to which they had no personal attachment. This approach resonated with students as evidenced in their comments in the course evaluations.

During the COVID-19 pandemic in Spring 2020, the course was delivered using both asynchronous and synchronous (Zoom) methods. The course meeting schedule was altered to limit Zoom fatigue by meeting synchronously on Tuesdays and working asynchronously on course materials during the remainder of the week. Thursday meeting sessions were reserved for open office hours, an approach that well was received by students and widely used. Tuesday synchronous meetings were initially used for discussions of the overall project, research design, and sequencing videos. Breakout rooms in Zoom were used extensively to facilitate small group discussions of research questions and to build comprehension of the sequencing videos. Beginning in week 3, the course took on essentially a "flipped" format. Students viewed and practiced skills introduced in the videos and synchronous class time was used for troubleshooting, comprehension checks, and setting up the "next steps." Students in this virtual course were still able to successfully use R for statistical analysis and visualization of their data.

DISCUSSION

One of the most important take home messages of this work is that we should take advantage of technology both to continue skill development as faculty and to teach resourcefulness to students. Many faculty who teach undergraduate students completed their dissertations before the age of bioinformatics or in an area that did not focus on quantitative skills. These faculty may not currently possess the skills to incorporate a bioinformatics module into a course. YouTube affords faculty the opportunity to learn new skills in a step-by-step manner when the technology and approaches may be wholly new to them. This is a very inexpensive and efficient way to acquire professional development that can serve to enhance both classroom teaching and potential new areas of research.

As part of the course evaluation, students were asked to answer a series of confidence questions about skills developed in the course (**Table 2**). On a scale of 1–10, with 10 being high confidence, all students rated themselves as a ten when asked about confidence in pipetting a variety of liquids with micropipettors, reflecting the skills developed in the wet lab portion of the course. When asked about explaining Sanger sequencing and Next Generation Sequencing to another scientist, the class averages were 6.8 and 6.7 out of 10, respectively, for these new skills learned in the course. The course successfully introduced students to basic knowledge about R, as reflected in an 8.9 average score to "I can copy, perform, and run a simple code in R." As expected from an 8 week introductory module, the students did not feel confident enough to create and run their own R code (average score 5.2).

TABLE 2 | Student confidence in course skills.

How confident are you that you can complete the following tasks?	Class average confidence Scale of 1–10; 10 = high confidence
I can pipet a variety of liquids with micropipetors.	10
I can understand scientific methods and instructions to perform experiments.	9.4
I could explain the process of PCR to another scientist.	8.7
I can explain Sanger sequencing to another scientist.	6.8
I can explain next generation sequencing (NGS) to another scientist.	6.7
I can analyze the bands on a gel to determine if I got the right sized DNA product from PCR.	9.5
I can use Excel to perform a *t*-test and analyze a *p*-value.	8.4
I can pose a question and use a large data set to effectively answer that question.	8.3
I can copy, perform, and run a simple code in R.	8.9
I can create and run my own code in R.	5.2

Students were also asked, "After completing this course, how has your interest in biological research changed?" All of their free response answers are below:

- I am still interested in it, and now realize the importance of being able to effectively use R and excel to convey my data.
- My interest in research has stayed quite high after taking this course. I am planning on working in the more biochemical side, but this was still very interesting and helped me make sure that a career in research is where I belong.
- I feel like I have a better understanding of how questions are being asked in the biological community.
- My interest in biological research has grown even stronger. I knew before that I love research, but every time I continue to do it, my passions grow stronger.
- It greatly raised my interest in biological research. It was cool to see how the experiments we performed gave us numbers, that we could find relationships between.
- I was always curious about how scientists made the figures they did. After using R, examining larger datasets is a lot less frightening.
- I have a greater understanding of the importance of microbiomes and am interested in my own microbiome!
- I was very hesitant about research before this course because I had a few bad experiences, but this class changed my outlook on it. I am definitely more interested and would like to do more.

- My interest has greatly increased in biological research, specifically, on human microbiomes like the gut microbiota. Also, conducting my own biological research and experiencing the challenges of creating a poster has made me appreciate all the hard work scientist do to give us informative papers.
- I am once again excited now about the medical applications of molecular biology and studies! I'm excited to skim new articles and have a better toolbox to understand them after learning about R and how microbiome data can be represented.

While this course had a very small sample size (n = 10), these responses suggest that this approach to using R was positively received by students. Moreover, the students saw a utility in learning R, which research shows may lead to continued interest in participating in mathematical biology experiences (Andrews and Aikens, 2018).

In a short time frame, the course introduced students to bioinformatics and provided an opportunity for further practice. Because of the students' ability to effectively visualize the dataset with R, they were able to think critically about the data and consider future research questions. From the R-generated heat map, the students realized that their initial hypothesis was incorrect. The heavy foot and automobile traffic sample site did have a higher abundance of bacteria but the diversity of bacteria was much lower than the sample site with light traffic. Several students continued their analysis of the data even after the course ended and proposed a new research question for the next offering of the course.

Several outcomes were achieved as a result of this module. First, faculty expertise was enhanced in a time efficient manner using YouTube training videos, leading to broadened research capabilities and comfort. Second, students were introduced to computational skills in a manner that was effective and intentional, with time for both introduction *and* reinforcement of skills. Finally, the module was effectively included in a biology curriculum because it could function as either a stand-alone course or a module within another course such as microbiology, leading to flexibility in the curriculum. This module, developed with CURE guidelines in mind, is an effective and easily implementable way to introduce a broad group of students to bioinformatics in biology research, and also serves as a springboard for interested students to pursue further training and research in bioinformatics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and has approved it for publication.

ACKNOWLEDGMENTS

Special thanks to the students in BIOL 4160 for enthusiastically and successfully attempting a task outside their comfort zone.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.578600/full#supplementary-material

REFERENCES

Albert, A., and Yoder, J. (2013). *Making Heatmaps With R for Microbiome Analysis. The Molecular Ecologist*. Available online at: https://www.molecularecologist.com/2013/08/making-heatmaps-with-r-for-microbiome-analysis/ (accessed August 31, 2020).

Andrews, S. E., and Aikens, M. L. (2018). Life science majors' math-biology task values relate to student characteristics and predict the likelihood of taking quantitative biology courses. *J. Microbiol. Biol. Educ.* 19:jmbe–19–80. doi: 10.1128/jmbe.v19i2.1589

Araneo, K., Schwebach, J. R., and Csikari, M. (2017). Advising biology majors about career choices: resources & information for biology instructors. *Am.Biol. Teach.* 79, 14–21. doi: 10.1525/abt.2017.79.1.14

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., et al. (2014). Assessment of course-based undergraduate research experiences: a meeting report. *CBE—Life Sci. Educ.* 13, 29–40.

Brewer, C. A., and Smith, D. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

CUREnet. (2020). *What is a CURE?*. Washington, DC: Science Education Resource Center at Carleton College.

Freeman, S., Okoroafor, N. O., Gast, C. M., Koval, M., Nowowiejski, D., O'Connor, E., et al. (2016). Crowdsourced data indicate widespread multidrug resistance in skin flora of healthy young adults†. *J. Microbiol. Biol. Educ.* 17, 172–182. doi: 10.1128/jmbe.v17i1.1008

Li, Y., and Chen, L. (2014). Big biological data: challenges and opportunities. *GenomicsProteomics Bioinformatics* 12, 187–189. doi: 10.1016/j.gpb.2014.10.001

Luscombe, N. M., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* 40, 346–358. doi: 10.1055/s-0038-1634431

Mellon, C. (2020). *List of Educational Programs in Computational Biology*. PIttsburgh, PA: Computational Biology Department.

Mulder, N., Schwartz, R., Brazas, M. D., Brooksbank, C., Gaeta, B., Morgan, S. L., et al. (2018). The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Computat. Biol.* 14:e1005772. doi: 10.1371/journal.pcbi.1005772

NCES (2020). "Chapter 3 employment outcomes of bachelor's degree holders," in *The Condition of Education 2020*, (Washington DC: Institute of Education Sciences).

Rosenwald, A. G., Pauley, M. A., Welch, L., Elgin, S. C. R., Wright, R., and Blum, J. (2016). The CourseSource Bioinformatics Learning Framework. *CBE Life Sci.Educ.* 15:le2. doi: 10.1187/cbe.15-10-0217

Small-World (2020). *Small World Initiative: Crowdsourcing Antibiotic Discovery*. Old Greenwich, CT: Small World Initiative Inc.

The R-Foundation (2020). *The R Project for Statistical Computing*. Vienna: The R Foundation.

US-DOE (2020). *Digest of Education Statistics*. Washington, DC: US Department of Education.

Wang, J. T. H., Daly, J. N., Willner, D. L., Patil, J., Hall, R. A., Schembri, M. A., et al. (2015). Do you kiss your mother with that mouth? an authentic large-scale undergraduate research experience in mapping the human oral microbiome†. *J. Microbiol. Biol. Educ.* 16, 50–60. doi: 10.1128/jmbe.v16i1.816

Wilson Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS One* 13:e0196878. doi: 10.1371/journal.pone.0196878

Winkler, H. (1920). *Verbreitung und Ursache der Parthenogenesis im Pflanzen- und Tierreiche*. Jena: Verlag von G. Fischer.

Conflict of Interest: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

PERSPECTIVE
published: 17 September 2020
doi: 10.3389/fmicb.2020.577634

In silico Phage Hunting: Bioinformatics Exercises to Identify and Explore Bacteriophage Genomes

Betsy M. Martinez-Vaz* and Madeline M. Mickelson

Department of Biology, Hamline University, St. Paul, MN, United States

Bioinformatics skills are increasingly relevant to research in most areas of the life sciences. The availability of genome sequences and large data sets provide unique opportunities to incorporate bioinformatics exercises into undergraduate microbiology courses. The goal of this project was to develop a teaching module to investigate the abundance and phylogenetic relationships amongst bacteriophages using a set of freely available bioinformatics tools. Computational identification and examination of bacteriophage genomes, followed by phylogenetic analyses, provides opportunities to incorporate core bioinformatics competencies in microbiology courses and enhance students' bioinformatics skills. The first activity consisted of using PHASTER (PHAge Search Tool Enhanced Release), a bioinformatics tool that identifies bacteriophage sequences within bacterial chromosomes. Further computational analyses were conducted to align bacteriophage proteins, genomes, and determine phylogenetic relationships amongst these viruses. This part of the project was carried out using the Clustal omega, MAFFT (Multiple Alignment using Fast Fourier Transform), and Interactive Tree of Life (iTOL) programs for sequence alignments and phylogenetic analyses. The laboratory activities were field tested in undergraduate directed research, and microbiology classes. The learning objectives were assessed by comparing the scores of pre and post-tests and grading final presentations. Post-tests were higher than pre-test scores at or below $p = 0.002$. The data suggest *in silico* phage hunting improves students' ability to search databases, interpret phylogenetic trees, and use bioinformatics tools to examine genome structure. This activity allows instructors to integrate key bioinformatic concepts in their curriculums and gives students the opportunity to participate in a research-directed learning environment in the classroom.

Keywords: bacteriophages, bioinformatics, genomes, phylogenetic trees, research project, experimental design (study designs)

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Anne Rosenwald,
Georgetown University, United States
Sandeep Tamber,
Health Canada, Canada

****Correspondence:***
Betsy M. Martinez-Vaz
bmartinezvaz01@hamline.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *29 June 2020*
Accepted: *26 August 2020*
Published: *17 September 2020*

Citation:
Martinez-Vaz BM and Mickelson MM (2020) In silico Phage Hunting: Bioinformatics Exercises to Identify and Explore Bacteriophage Genomes. Front. Microbiol. 11:577634. doi: 10.3389/fmicb.2020.577634

INTRODUCTION

There is a distinct need in life science education for educators to adapt their teaching strategies to best support student learning and prepare them for careers in science. Scientific education councils cite incorporating research into the undergraduate curriculum and emphasizing the interdisciplinary nature of biology as major national reform goals (National Research Council, 2009; AAAS, 2010). Research experiences are often interdisciplinary in nature and can teach students to think like scientists (AAAS, 2010; Ballen et al., 2017), greatly increase their chances of entrance into graduate school (Bangera and Brownell, 2014), and allow students to actively engage with current problems in biology. However, the process of obtaining a research experience while at college presents a variety of barriers to historically underrepresented and marginalized student populations (Stebleton and Soria, 2012). As a result, many students are left out of research experiences that can greatly enrich their understanding of biology, open doors for them professionally, and allow them to contribute their perspective in the larger scientific community (Bangera and Brownell, 2014). To make research experiences more accessible, many instructors opt to integrate research-based learning activities into their course curriculums. The activity we describe here is a low-cost research experience that can be implemented as a multi-week laboratory exercise in microbiology and other biology courses. This activity, which we have called *In Silico Phage Hunting*, utilizes bioinformatic tools to learn about bacteriophage genomes and viral proteins. We offer a perspective that this activity can help research experiences be more accessible by being integrated into appropriate elective biology courses, and engages students in inquiry-based learning, critical thinking, and scientific discovery which are all key components of the research process.

In Silico Phage Hunting is an adaptable teaching module that utilizes freely available bioinformatics software to examine the abundance of bacteriophages in bacterial genomes, the viral proteins encoded, and the phylogenetic relationships between phage proteins. Many bacterial genomes remain to be examined for the existence of bacteriophages. For example, only 500 sequenced genomes of bacteriophages that infect the genus *Escherichia* have been isolated compared to the 66,000 *Escherichia* bacterial genomes that are publicly available (Sazinas et al., 2017). Yet, bacteriophages contain a high proportion of novel genetic sequences and are likely to represent the largest reservoir of unexplored genes on earth (Hatfull, 2008). Given the existence of so many bacterial genomes in public databases and the limited knowledge of phages therein, locating phages can help bridge this gap in scientific understanding. *In Silico Phage Hunting* engages students in researching bacteriophage abundance and diversity while they learn about the genetic interplay between viruses and bacteria and develop important bioinformatics skills relevant for careers in STEM fields.

Bacteriophages have been used as model systems to incorporate scientific inquiry and research into undergraduate classrooms across the United States. The SEA-PHAGES (Science Education Alliance Phage Hunting Advancing Genomics and Evolutionary Science) program has successfully engaged thousands of students in authentic research through the isolation, characterization, and genome sequencing of bacteriophages from various bacterial species. Numerous laboratory exercises describing the use of phages to investigate diverse aspects of biology have been reported (Allen and Gyure, 2013; Hyman, 2014). Other initiatives include independent faculty members integrating phage research as multi-week laboratory exercises into their courses. Williamson et al. (2014) utilized a phage research experience in a molecular virology course to isolate novel bacteriophages and perform genetic analyses using computational tools. Recently, the Genome Solver Project utilized phage genomes to create hands-one bioinformatics activities and encourage educators to incorporate computational skills in their biology courses (Mathur et al., 2019). Despite these efforts, the number of phage-based laboratory activities incorporating genome searches, bioinformatics analyses, and phylogeny is still very limited in comparison to wet bench exercises.

Using computational methods to study bacteriophages is a promising area of biological research, as new insights can be uncovered about phage genomics and proteomics through bioinformatic analysis. Many studies have utilized the data present in public databases and bioinformatic analyses to research phage transcription (Guzina and Djordjevic, 2015), evolutionary classification (Lima-Mendez et al., 2008), and protein function (Carlton et al., 2005). Other authors advocate that undergraduate students can spearhead this research into bacteriophage abundance and diversity (Staub et al., 2017). In the Internet era, many more possibilities for scientific inquiry exist that previously were not accessible. The availability of microbial and viral genomes along with computational tools to assess these genomes gives teachers a unique opportunity to incorporate more inquiry-based and active learning exercises into their classrooms.

Bioinformatics skills are increasingly relevant to research in most areas of the life sciences. A recent nationwide faculty survey led to the development of a set of nine core competencies to guide the integration of bioinformatics in the life sciences curriculum (Wilson-Sayres et al., 2018). These competencies include but are not limited to: (1) understanding the role of data mining and computation in hypothesis-driven processes in the life sciences, (2) summarizing key computational concepts, (3) applying statistical concepts used in bioinformatics, (4) utilizing bioinformatics tools to analyze genomic information, and (5) knowing how to access genomic (Wilson-Sayres et al., 2018). A large number of publicly available bacterial genomes provides an excellent opportunity to teach about the abundance and evolutionary relationships amongst bacteriophages while incorporating five of the nine core bioinformatics competencies in microbiology courses.

In Silico Phage Hunting addresses five of these competencies by accessing genomic data from NCBI (National Center for Biotechnology Information), analyzing bacterial and phage genomes with bioinformatic tools, and facilitating group discussion throughout this process. Additionally, students prepare a final presentation or poster of their findings and complete worksheets along the way, acting as "check points" for their understanding of the concepts underlying the activity. This activity is also designed to promote inquiry in student groups by allowing enough time for students to discuss the concepts they are learning about and practicing the related skills. A sense of scientific discovery is also present in this activity, as students know they are locating phages in bacterial genomes and making genetic interpretations that have not yet been made.

LESSON OVERVIEW

The *In Silico Phage Hunting* lesson described in this report was designed for upper level biology students. The activities can be conducted as part of the laboratory component of upper level biology classes or as independent projects in directed research courses. In traditional biology classes, the activities can be taught as multi-week laboratory exercises or separate classroom assignments. Prior to starting this activity, participants must have background knowledge of the following concepts: (1) principles of microbial and eukaryotic cell structure, (2) the central dogma of biology, (3) basic interpretation of phylogenetic trees, and (4) familiarity with sequence similarities searches and their application to biological research questions. It is recommended that students have completed at least one semester of genetics or cellular biology for effective participation in this project. If the students do not have substantial background in these areas, the instructor should incorporate brief lectures and discussions on common bioinformatics tools, genomes sequences, and any other themes considered necessary for the activity.

The workflow for multi-week laboratory exercises and research projects is illustrated in **Figure 1**. During the first week, students have a pre-test followed by an assigned reading of the laboratory handout and a lecture on phage biology. Bioinformatics searches and phage genome exploration are covered in the second and third week of phage hunting activities. The fourth, fifth, and sixth weeks are used for experimental design, data retrieval, and analysis. Instructors have the choice of using additional weeks for wet-bench experiments or more computational analysis (**Supplementary File S1**). Independent research projects follow a timeline similar to the multi-week laboratory exercises, however, these students have additional weeks to expand their research questions and perform more detailed bioinformatics analyses (**Supplementary File S1**). Both the multi-week laboratory activity and the research project have a final assessment consisting of a poster or an oral presentation.

The *In Silico Phage Hunting* activities were designed to achieve the following learning objectives:

1. Describe the basic structure of a bacteriophage.
2. Explain the cycle of bacteriophage infection and replication.
3. Construct and interpret phylogenetic trees to investigate evolutionary relationships amongst phages.
4. Utilize bioinformatics tools to detect phages in bacterial genome sequences.
5. Retrieve bacteriophage genomes and protein sequences from public databases.
6. Formulate hypotheses regarding the abundance of bacteriophages in microbial genomes.

This research was deemed exempt status by the Hamline University IRB committee as defined by federal regulations (Final Common Rule, 45 CFR §46.104) under normal educational research. The study presented less than minimal risk associated with students' participation and was conducted in an established educational setting using practices that were not likely to adversely impact student learning or assessment of the instructor providing the lesson. The data shown is anonymous and cannot be linked directly or indirectly to any of the participants in the study.

RESULTS AND DISCUSSION

The *In Silico* phage hunting activities were field-tested in two classes, Microbiology (AY 2015–2016), and Research in Biology (AY2017–2018). In lecture-based classes, these activities were conducted as part of multi-week investigative laboratory exercises in the laboratory component of the course. Students taking the Research in Biology course designed and completed independent projects using the *In Silico Phage Hunting* approach over a period of 6–10 weeks. Examples of laboratory and directed research projects are presented in **Supplementary File S2**.

Student learning after the completion of the *In Silico* phage hunting activities was evaluated using multiple assessment tools. For the lecture-based courses, we used the scores of pre and post-tests (**Table 1** and **Supplementary Files S3**) to assess learning objectives (LOs) 1–5. The scores in the post-tests were significantly higher than the pre-tests with a *p*-value at or below 0.002. Learning gains were calculated for learning objectives 1–5. These analyses showed learning objectives 3–5 improved the most with learning gains equal to 0.50, 0.58, and 0.52, respectively. In contrast, LO1 and LO2 showed learning gains of 0.27 and 0.35. These results suggest the introductory lecture, and laboratory exercises completed in the initial portion of the phage-hunting activity improved students' knowledge of bacteriophage biology, computational detection of phages, and interpretation of phylogenetic trees. The data from pre and post-tests indicate students showed the most improvement in the skills related to database searches and interpretation of phylogenetic trees (LOs 3–5).

Learning objectives 3–6 were designed to allow students to formulate hypotheses and employ publicly available data with computational tools to address these propositions. These activities were assessed by grading laboratory worksheets, and final research presentations (**Supplementary Files S2, S8**). Rubrics were developed to evaluate hypothesis statements, interpretation of bioinformatics data, and presentations (**Supplementary File S5**). Evaluation of laboratory worksheets showed students were able to formulate hypotheses and use publicly available genome data to test these propositions. Students were able to formulate hypotheses regarding the abundance of phages in *Escherichia coli*, and several genera of Nitrogen-fixing bacteria. In one of the courses, students carried out additional wet bench experiments to investigate induction of phages using diverse conditions such as temperature shifts, exposure to UV light, and chemicals. When students did *In Silico Phage Hunting* as a summer or semester long research project, they often conducted phage induction and isolation experiments as part of their projects (**Supplementary File S2**). The laboratory worksheets (**Supplementary File S8**) and presentations showed most students met expectations regarding the formulation

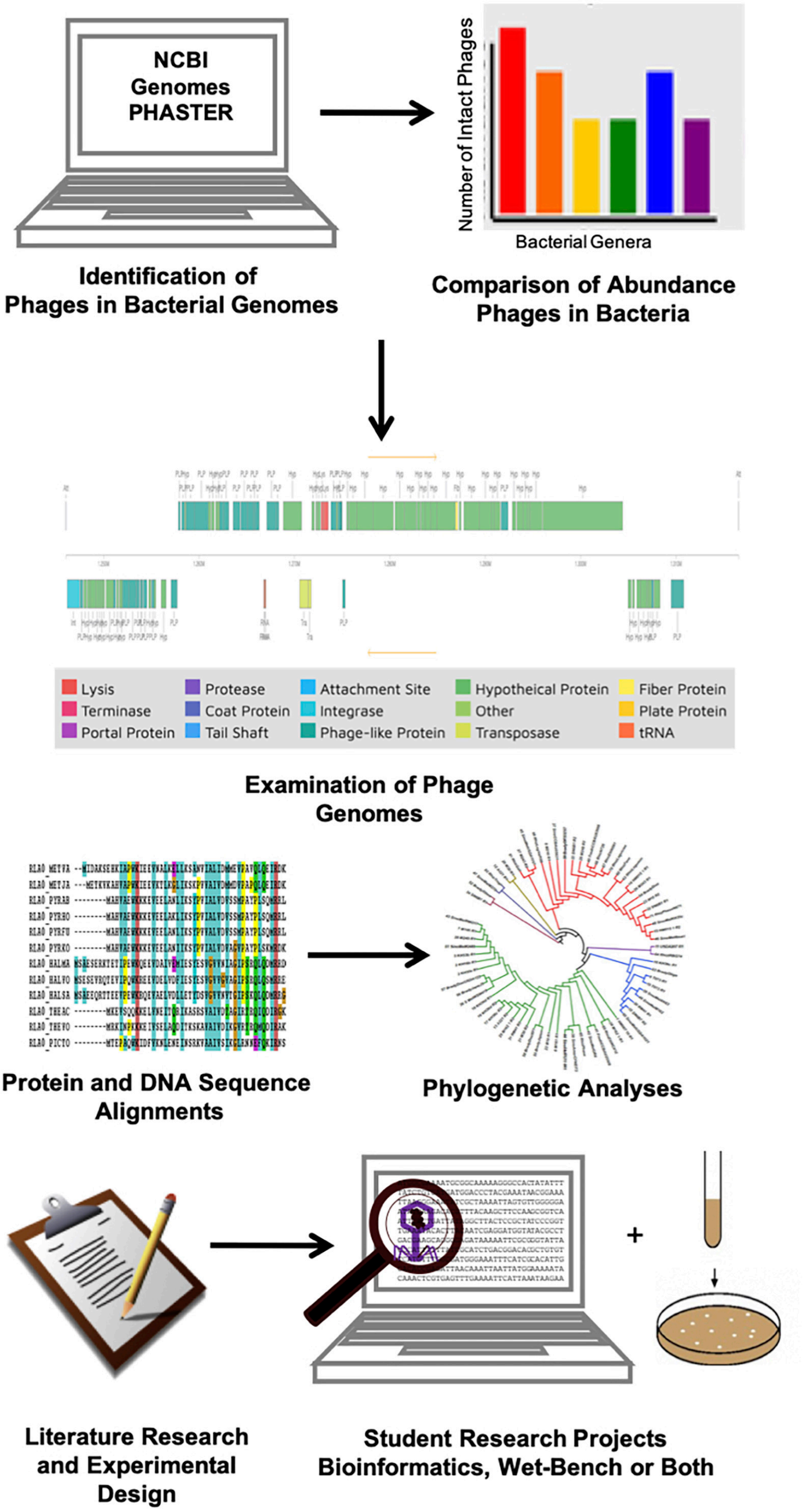

NCBI
Genomes
PHASTER
Identification of Phages in Bacterial Genomes
Number of Intact Phages
Bacterial Genera
Comparison of Abundance Phages in Bacteria
Lysis
Terminase
Portal Protein
Protease
Coat Protein
Tail Shaft
Attachment Site
Integrase
Phage-like Protein
Hypotheical Protein
Other
Transposase
Fiber Protein
Plate Protein
tRNA
Examination of Phage Genomes
Protein and DNA Sequence Alignments
Phylogenetic Analyses
+
Literature Research and Experimental Design
Student Research Projects Bioinformatics, Wet-Bench or Both

FIGURE 1 | Workflow for *in silico* phage-hunting multi-week laboratory project conducted by microbiology students to investigate the presence of bacteriophages in the genomes of different bacteria.

TABLE 1 | Summary of pre/post-test assessment data for *in silico* phage hunting laboratory activities field tested in microbiology courses.

Course	Pre-test score (%)	Post-test score (%)	*p*-value[a]	Normalized learning gains[b]
Microbiology 2015 (*n* = 19)	60.1	71.7	0.00036	0.28
Microbiology 2016 (*n* = 13)	62.9	80.5	0.00179	0.47

[a] *A paired two tail t- test was used to evaluate the mean difference between pre and post-test scores. These scores were significant with p-values at or below 0.002.*
[b] *Average normalized individual student learning gains (G) were calculated for the questions in the pre- and post-tests. G = [(post-test – pre-test)/(100 – pre-test)].*

of hypotheses, retrieval of data from public repositories and interpretation of phylogenetic trees. Students' verbal feedback, overall attitude, and level of engagement with the *In Silico Phage Hunting* activities were very positive. These observations suggest they had an appreciation for bioinformatics and its applications to biology.

In Silico Phage Hunting provides multiple opportunities for students to be exposed to and practice core bioinformatics competencies (**Supplementary File S6**). Formulating hypotheses about the abundance of phages in microbial genomes highlights the role of computation and data mining in addressing questions in the life sciences; this is one of the most important bioinformatics competencies. In addition, during the phage hunting activities, students used multiple databases and evaluated statistical values to assess the accuracy of bioinformatics predictions. These skills support the development of various bioinformatics competencies including the use of computational tools to examine biological problems and applying statistical concepts in bioinformatics.

Students also practice appropriate retrieval and organization of large data sets. These competencies are essential when locating and sorting through different types of biological data to construct sequence alignments and phylogenetic trees.

In Silico Phage Hunting was designed as a series of multi-week laboratory exercises, therefore, the activity has many fundamentally "hands-on" active learning components. For example, students were engaged in downloading data from public databases, employing bioinformatics web tools to analyze bacterial and viral genome sequences, and using the information gathered to discuss the biological relevance and implications of their findings. Many parts of this activity involve students working in teams of 3–4 people. This strategy leads to group discussions focused on the value of the information gathered, and whether the results obtained support or refute the hypothesis posed. These activities are all consistent with the definition of activity learning which states "anything that involves students in doing things and thinking about the things they are doing" (Bornwell, 1991). Group work also encourages effective communication, and higher-order thinking tasks such as critical data analysis, evaluation and synthesis of information.

By using freely available Internet software in research, these activities provide students of diverse backgrounds and academic abilities with an opportunity to learn how to use bioinformatic tools to test hypotheses. Laboratory projects involving data mining and bioinformatics prepare students to participate in summer or course-based research experiences given that addressing modern scientific questions in biology often involves working with large data sets as well as retrieving information from databases. The activities carried out as part of these projects make use of multiple approaches to teach about bacteriophage structure, function, and diversity. Students formulate hypotheses, perform database searches, and explore different ways to analyze and present data. These exercises provide students of diverse learning styles with an opportunity to engage with topics being taught and make contributions to their team.

The phage hunting activities can be easily modified and carried out with any bacterial species, offering research opportunities for students in the classroom that otherwise would be difficult to access. *In Silico* Phage Hunting also provides prospects to discover and investigate novel phages in bacterial genomes. An extension of these activities could include searching and cataloging the abundance of RNA phages in bacterial chromosomes. Once a complete DNA or RNA phage genome is predicted by PHASTER, investigators can isolate the virus by induction of the lytic cycle or cloning. This type of research can contribute to enhancing our understanding of phage biology and viral diversity.

Phage genome analyses can be modified to incorporate other viral bioinformatic tools such as: the MVP (Microbe-Vs.-Phage) database, VIRFAM (Viral Protein Families), and Phage Signature Genes, PhiSiGns (**Supplementary File S7**). *In Silico Phage Hunting* is suitable as a research project or as a laboratory activity for upper-level courses such as virology, microbiology, evolution, and molecular biology. Alternatively, instructors can use the phage hunting activities as individual or stand-alone modules to create assignments and supplement class content. Many of the bioinformatics activities are easy to follow and can be used as classroom or one-time laboratory exercises. These activities can also be used together with commercially available bacteriophage induction and plaque demonstration kits.

PERSPECTIVE

The equitable access of research experiences is of concern in academia. We know students benefit greatly from these experiences. Students report that they have a greater sense of pride in their work (Hekmat-Scafe et al., 2017), learn critical thinking skills by virtue of engaging in the research process, and gain a greater sense of awareness about what scientific research is like. Research experiences are also one of the crucial keys for entrance into graduate programs (Bangera and Brownell, 2014). Given the high importance of research experiences in relation to student learning, appreciation of biology, and gateways to academic and professional opportunities, these experiences should be designed to be more accessible to a wider range of students. Scientific educators occupy a critical space by being able to directly implement research experiences in their classrooms. As individual teachers, we must consider what we can do to create an equitable learning environment for all learners. Teachers have a direct relationship with students' education, and it is this education that can open doors for students into future careers as educators, researchers, doctors, and scientists. Research experiences can open these doors for students and empower them as learners. Teachers can implement a tangible change in their classrooms by incorporating research experiences into their curriculums.

In Silico Phage Hunting is a classroom activity that engages students in important aspects of the research process: hypothesis testing, data mining, interpretation of results, and sharing of findings. It is an adaptable teaching module that incorporates research into laboratories and classrooms with little cost to instructors. A barrier to incorporating research experiences into classrooms is often an issue of lack of resources and funding for departments. The activity we described in this paper does not require advanced equipment or massive amounts of funding to implement and was field-tested at a small liberal arts college showing learning gains made by students. Biological databases offer a new approach to incorporating research experiences into classrooms that would otherwise have been difficult to achieve.

By developing a teaching module that utilizes research into bacteriophage abundance and diversity, we aim to confront the barriers present in academic research. The wealth of biological data that is freely available in databases such as NCBI (National Center for Biotechnology Information) presents a unique opportunity that was not available for past educators to incorporate research into the classroom. Software tools available on the Internet present new modes of inquiry into this data, offering new interpretations and insights into bacteriophage abundance and diversity. The core of this activity is the use of freely available data and tools on the Internet to design research activities in the classroom, and we encourage other educators to get creative with this accessible information.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation, to any qualified researcher.

AUTHOR CONTRIBUTIONS

MM and BM-V contributed to different sections of this work. BM-V devised the study, created the **Supplementary Material**, and authored the sections "Results and Discussion" and the "Lesson Overview." MM performed one the research projects, authored the sections "Introduction" and "Perspective", and managed the references. Both authors contributed to the article and approved the submitted version.

FUNDING

Students who participated in the Research in Biology class for this teaching module received a grant from the Summer Collaborative Research Program (SCUR) to conduct research during the summers of 2017 and 2018.

ACKNOWLEDGMENTS

We would like to warmly thank the Summer Collaborative Research Program (SCUR) at the Hamline University, the Hamline University Biology Department faculty and staff, and most importantly the students who participated in these research activities for their contributions and support of this project.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.577634/full#supplementary-material

REFERENCES

AAAS (2010). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: AAAS.

Allen, M. E., and Gyure, R. A. (2013). An undergraduate laboratory activity demonstrating bacteriophage specificity. *JMBE* 14, 84–92. doi: 10.1128/jmbe.v14i1.534

Ballen, C. J., Blum, J. E., Brownell, S., Hebert, S., Hewlett, J., Klein, J. R., et al. (2017). A call to develop course-based undergraduate research experiences (CUREs) for nonmajors courses. *CBE Life Sci. Educ.* 16:2.

Bangera, G., and Brownell, S. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci. Educ.* 13, 602–606. doi: 10.1187/cbe.14-06-0099

Bornwell, C. C. (1991). *Active Learning: Creating Excitement in the Classroom*. ASHE-ERIC Higher Education Report No. 1. Washington, DC: The George Washington University.

Carlton, R. M., Noordman, W. H., Biswas, B., de Meester, E. D., and Loessner, M. J. (2005). Bacteriophage P100 for control of *Listeria monocytogenes* in foods: genome sequence, bioinformatic analyses, oral toxicity study, and application. *Regul. Toxicol. Pharmacol.* 43, 301–312. doi: 10.1016/j.yrtph.2005.08.005

Guzina, J., and Djordjevic, M. (2015). Bioinformatics as a first-line approach for understanding bacteriophage transcription. *Bacteriophage* 5:e1062588. doi: 10.1080/21597081.2015.1062588

Hatfull, F. G. (2008). Bacteriophage genomics. *Curr. Opin. Microbiol.* 11, 1–10. doi: 10.1002/9783527678679.dg00882

Hekmat-Scafe, D. S., Brownell, S. E., Seawell, P. C., Malladi, S., Imam, J. F. C., Singla, V., et al. (2017). Using yeast to determine the functional consequences of mutations in the human p53 tumor suppressor gene: an introductory course-based undergraduate research experience in molecular and cell biology. *Biochem. Mol. Biol. Educ.* 45, 161–178. doi: 10.1002/bmb.21024

Hyman, P. (2014). Bacteriophage as instructional organisms in introductory biology labs. *Bacteriophage* 4:e27336. doi: 10.4161/bact.27336

Lima-Mendez, G., Van Helden, J., Toussaint, A., and Leplae, R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Mol. Biol. Evol.* 25, 762–777. doi: 10.1093/molbev/msn023

Mathur, V., Arora, G. S., McWilliams, M., Russell, J., and Rosenwald, A. G. (2019). The genome solver project: faculty training and student performance gains in bioinformatics. *J Microbiol Biol Educ.* 20:20.1.4.

National Research Council (2009). *A New Biology for the 21st Century. The National*. Washington, D.C: Academic Press.

Sazinas, P., Redgwell, T., Rihtman, B., Grigonyte, A., Michniewski, S., Scanlon, J. D., et al. (2017). Comparative genomics of bacteriophage of the genus *Seuratvirus*. *Genome Biol. Evol.* 10, 72–26. doi: 10.1093/gbe/evx275

Staub, N. L., Poxleitner, M., Braley, A., Smith-Flores, H., Pribbenow, C. M., Jaworski, L., et al. (2017). Scaling up: adapting a phage-hunting course to increase participation of first-year students in research. *CBE Life Sci. Educ.* 15:ar13. doi: 10.1187/cbe.15-10-0211

Stebleton, M. J., and Soria, K. M. (2012). Breaking down barriers: academic obstacles of first-generation students at research universities. *Learn. Assist. Rev.* 17, 7–19.

Williamson, R. P., Barker, B. T., Drammeh, H., Scott, J., and Lin, J. (2014). Isolation and genetic analysis of an environmental bacteriophage: a 10-session laboratory series in molecular virology. *Biochem. Mol. Biol. Educ.* 42, 480–485. doi: 10.1002/bmb.20829

Wilson-Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS One* 13:e0196878. doi: 10.1371/journal.pone.0196878

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Genetics

METHODS
published: 25 September 2020
doi: 10.3389/fgene.2020.577563

A Department of Defense Laboratory Consortium Approach to Next Generation Sequencing and Bioinformatics Training for Infectious Disease Surveillance in Kenya

Irina Maljkovic Berry[1]*, Wiriya Rutvisuttinunt[1,2], Logan J. Voegtly[3,4], Karla Prieto[5,6], Simon Pollett[1], Regina Z. Cer[3,4], Jeffrey R. Kugelman[6], Kimberly A. Bishop-Lilly[3], Lindsay Morton[7], John Waitumbi[8] and Richard G. Jarman[1]

[1] Viral Diseases Branch, Walter Reed Army Institute of Research, Silver Spring, MD, United States, [2] Office of Genomics and Advanced Technologies National Institute of Allergy and Infectious Diseases, Bethesda, MD, United States, [3] Genomics & Bioinformatics Department, Biological Defense Research Directorate, Naval Medical Research Center-Frederick, Fort Detrick, MD, United States, [4] Leidos, Reston, VA, United States, [5] College of Public Health, University of Nebraska Medical Center, Omaha, NE, United States, [6] Center for Genomic Studies, United States Army Medical Research Institute for Infectious Diseases, Frederick, MD, United States, [7] Global Emerging Infections Surveillance, Armed Forces Health Surveillance Branch, Silver Spring, MD, United States, [8] Basic Science Laboratory, US Army Medical Research Directorate-Africa/Kenya Medical Research Institute, Kisumu, Kenya

OPEN ACCESS

Edited by:
Sophie Shaw,
University of Aberdeen,
United Kingdom

Reviewed by:
Prashanth N. Suravajhala,
Birla Institute of Scientific Research,
India
Paul M. Krzyzanowski,
University Health Network, Canada

****Correspondence:***
Irina Maljkovic Berry
Irina.maljkovicberry.ctr@mail.mil

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Genetics

Received: *29 June 2020*
Accepted: *31 August 2020*
Published: *25 September 2020*

Citation:
Maljkovic Berry I, Rutvisuttinunt W, Voegtly LJ, Prieto K, Pollett S, Cer RZ, Kugelman JR, Bishop-Lilly KA, Morton L, Waitumbi J and Jarman RG (2020) A Department of Defense Laboratory Consortium Approach to Next Generation Sequencing and Bioinformatics Training for Infectious Disease Surveillance in Kenya. Front. Genet. 11:577563. doi: 10.3389/fgene.2020.577563

Epidemics of emerging and re-emerging infectious diseases are a danger to civilian and military populations worldwide. Health security and mitigation of infectious disease threats is a priority of the United States Government and the Department of Defense (DoD). Next generation sequencing (NGS) and Bioinformatics (BI) enhances traditional biosurveillance by providing additional data to understand transmission, identify resistance and virulence factors, make predictions, and update risk assessments. As more and more laboratories adopt NGS and BI technologies they encounter challenges in building local capacity. In addition to choosing the right sequencing platform and approach, considerations must also be made for the complexity of bioinformatics analyses, data storage, as well as personnel and computational requirements. To address these needs, a comprehensive training program was developed covering wet lab and bioinformatics approaches to NGS. The program is meant to be modular and adaptive to meet both common and individualized needs of medical research and public health laboratories across the DoD. The training program was first deployed internationally to the Basic Science Laboratory of the US Army Medical Research Directorate-Africa in Kisumu, Kenya, which is an overseas Lab of the Walter Reed Army Institute of Research (WRAIR). A week-long workshop with intensive focus on targeted sequencing and the bioinformatics of genome assembly (*n* = 24 participants) was held. Post-workshop self-assessment (completed by 21 participants) noted significant median gains in knowledge domains related to NGS targeted sequencing, bioinformatics for genome assembly, and sequence quality assessment. The participants also reported that the information on study design, sample preparation, sequencing quality control, data quality assessment, reporting, and basic and advanced bioinformatics analysis were the most useful information presented in the training. While longer-term evaluations are planned, the training resulted in significant short-term improvement of a laboratory's self-reported wet lab and bioinformatics capabilities. This framework can be used for future DoD laboratory development in the area of NGS and BI for infectious disease surveillance, ultimately enhancing this global DoD capability.

Keywords: NGS, bioinformatics, workshop, infectious disease, DoD

INTRODUCTION

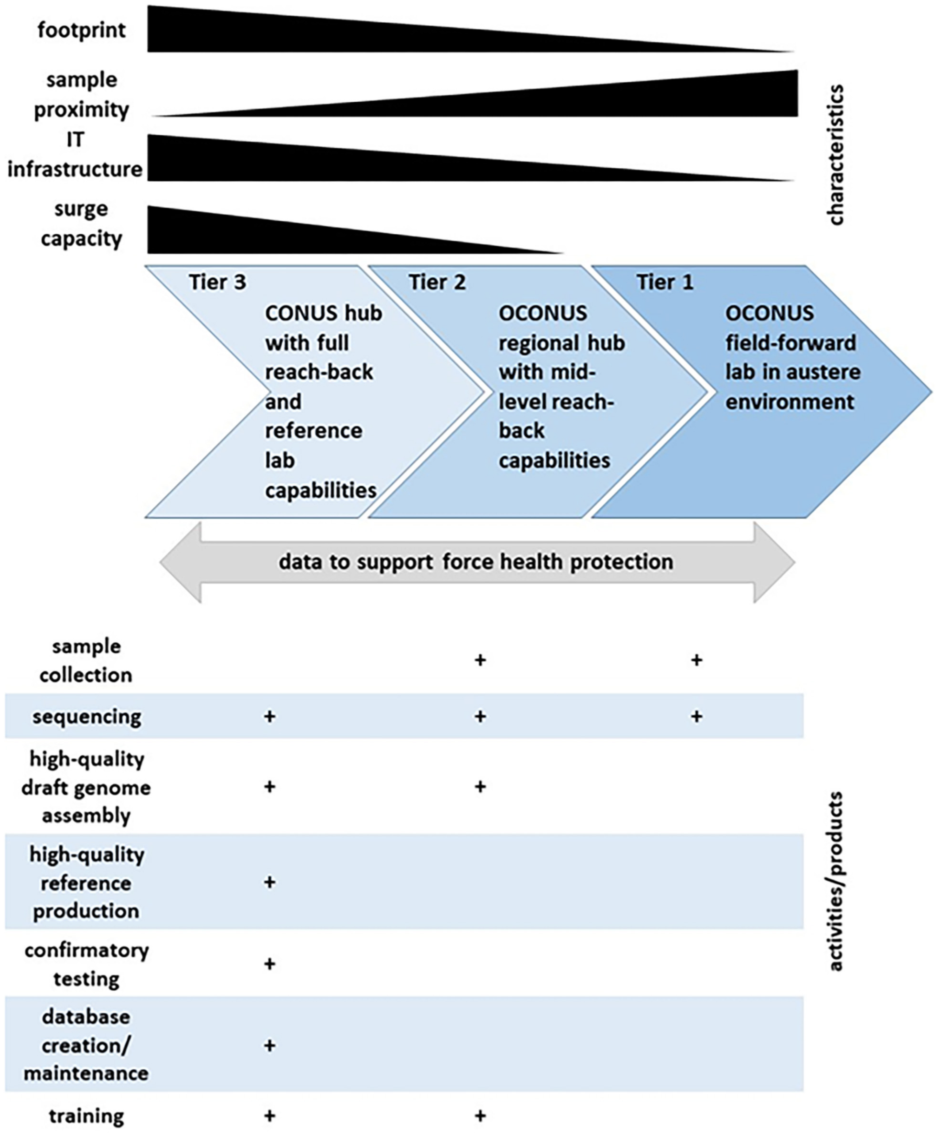
Development of Next-Generation Sequencing (NGS), or High-Throughput Sequencing (HTS), has revolutionized life sciences, dramatically increasing the variety of questions that can be answered using genomic sequence data. With this continuously evolving and growing field, the need for adequate computational hardware resources, software, and expertise to analyze large and complex data is also increasing. The field of bioinformatics has thus experienced substantial growth and advancement in recent years, and the requirement for highly skilled and specialized personnel has surged.

Within the Department of Defense (DoD), NGS and bioinformatics are routinely used to answer many scientific and research questions that ultimately aid in protection of the armed forces, as well as the general population (Kijak et al., 2017; Colby et al., 2018; Ehrenberg et al., 2019; Waickman et al., 2019). Infectious diseases are one area where such research is of high importance. Like the general population, United States forces are vulnerable to many infections commonly occurring within the United States, such as influenza, coronavirus, adenovirus and antibiotic resistant bacterial infections including but not limited to infection by methicillin resistant *Staphylococcus aureus* (MRSA); pathogens that have the ability to negatively impact United States force readiness and mission goals (MacPherson et al., 1923; Beam et al., 1959; Earhart et al., 2001; Shanks and Hodge, 2011; Millar et al., 2017, 2019). In addition, global deployment of the United States forces also puts them at a higher risk for infections that occur more frequently outside the United States, such as Ebola, dengue, Zika, cholera, malaria, leishmaniasis, shigellosis, and many others (Riddle et al., 2011; Murray et al., 2015). The DoD Global Emerging Infections Surveillance (GEIS) program seeks to improve infectious disease surveillance, prevention, and response capability to better protect the health of the military force. Utilizing a global network of partner DoD medical research and public health laboratories, GEIS funds surveillance activities in over 70 countries to inform force health protection through timely and actionable infectious disease surveillance information (Chakhunashvili et al., 2017; Chang et al., 2018; Coleman et al., 2018; Koka et al., 2018; Anyamba et al., 2019; Guerra et al., 2019; Juma et al., 2019; Rivers et al., 2019; Rocha et al., 2019; Sugiharto et al., 2019). Unsurprisingly, development of NGS and bioinformatics methods for infectious disease surveillance and control has enabled a rapid expansion of GEIS partner studies that utilize pathogen genomic information (Frey et al., 2016; Maljkovic Berry et al., 2016, 2019a; Lee et al., 2017; Mullins et al., 2017; Salje et al., 2017; Cowell et al., 2018; LaBreck et al., 2018; Srijan et al., 2018; Grubaugh et al., 2019; Kim et al., 2019; Mbala-Kingebeni et al., 2019; Millar et al., 2019; Pollett et al., 2019; Wiley et al., 2019). However, NGS and bioinformatics can generally be technically challenging, as it requires specific knowledge of complex wet lab and bioinformatics processes (Maljkovic Berry et al., 2019b). Therefore, and in spite of great interest in this technology, only a few partner laboratories have been adequately equipped to utilize these approaches to their full potential.

In 2017, GEIS created a Consortium to address the increasing needs and challenges associated with NGS and bioinformatics at DoD medical research and public health laboratories. The vision of the Consortium is to rapidly detect and characterize known, emerging, and novel infectious disease agents through establishment of a harmonized DoD laboratory NGS and bioinformatics capability to inform force health protection decision making. The Consortium today represents a network of DoD laboratories that use NGS and bioinformatics for infectious disease surveillance. A baseline assessment and initial training effort was led by GEIS and three DoD core sequencing and bioinformatics laboratories: WRAIR-VDB (Walter Reed Army Institute of Research-Viral Diseases Branch), NMRC-BDRD (Naval Medical Research Center-Biological Defense Research Directorate), and USAMRIID-CGS (United States Army Medical Research of Infectious Diseases-Center for Genome Science). The Consortium performed an assessment of the GEIS DoD laboratory partners with access to Illumina MiSeq or other NGS instrument(s), in order to evaluate existing laboratory capabilities in NGS and bioinformatics, and to map gaps and needs in laboratory utilization of these tools to meet their mission goals of infectious disease surveillance. Limited access to experienced and knowledgeable NGS and bioinformatics personnel was one of the main gaps, making basic and advanced bioinformatics analyses a common challenge across the network. Another challenge was the restrictive and limited informatics infrastructure, especially in some of the participating laboratories located in low-and-middle income countries (LMICs). However, the challenge of finding personnel with sufficient training in NGS and bioinformatics was not only observed in laboratories located in LMICs, it was also apparent in domestic laboratories, thus highlighting the need to develop a structured NGS and bioinformatics training for the specific needs of DoD biosurveillance programs. Such training would have to be standardized across the Consortium network, as well as made agile enough to meet different levels of needs and computational resources of the participating DoD laboratories. Using the baseline information from the assessment, desired sequencing capabilities for DoD research and public health laboratories were divided into three tiers (**Figure 1**). Here we present the deployment of NGS and bioinformatics training with our partner laboratory in Kenya, United States Army Medical Research Directorate – Africa (MRD-A). Future iterations of similar trainings and assessments will be used to further strengthen global infectious surveillance for DoD utilizing genomics and bioinformatics.

MATERIALS AND EQUIPMENT

Samples used for the NGS hands-on training included dengue virus 2 (DENV-2) and chikungunya (CHIKV) and were provided on-site. Controls for library preparation, MiSeq


footprint
sample proximity
IT infrastructure
surge capacity
characteristics
Tier 3
CONUS hub with full reach-back and reference lab capabilities
Tier 2
OCONUS regional hub with mid-level reach-back capabilities
Tier 1
OCONUS field-forward lab in austere environment
data to support force health protection
activities/products

	Tier 3	Tier 2	Tier 1
sample collection		+	+
sequencing	+	+	+
high-quality draft genome assembly	+	+	
high-quality reference production	+		
confirmatory testing	+		
database creation/ maintenance	+		
training	+	+	

FIGURE 1 | Tiered next generation sequencing (NGS) and bioinformatics (BI) capabilities for biosurveillance. Relative levels of laboratory and equipment footprint, proximity to source of biosurveillance samples, information technology (IT) infrastructure, and sequencing and bioinformatics surge capacity are displayed by black gradient bars along the **top**. Continuous flow of data back and forth among all three tiers is depicted by gray arrow, and expected types of activities and products by tier are illustrated by plus marks (+) along the **bottom**.

sequencing and TapeStation for both DENV-2 and CHIKV were validated and prepared at VDB-WRAIR in the months prior to the planned NGS&BI training in Kenya. Prior to shipment of controls to Kenya, the control concentrations were measured and documented and the information was sent to MRD-A. Coordination of the reagent and control shipment from VDB-WRAIR to Kisumu, Kenya started a month prior to the training. Four Linux laptops and two Linux servers were prepared for hands-on bioinformatics training. A list of software was prepared by the Consortium and sent out to MRD-A Lab for installation onto the training computers. The software list included ngs_mapper, IGV, Geneious, MEGA7, EDGE (servers only) (Robinson et al., 2011; Kumar et al., 2016; Viral Diseases Branch WRAIR, 2016; Philipson et al., 2017). Three weeks prior to the training, a hands-on genome assembly training dataset was designed, consisting of dengue, chikungunya, and influenza raw fastq data, as well as hands-on performance instructions. The whole dataset was tested at VDB-WRAIR prior to training and saved onto the training computers.

METHODS

Day 1

Lectures and theory included: History of sequencing, overview of NGS, library preparation, quantification, validation and pooling. In detail: (i) List of library preparation kits used by core DoD for different projects and specimens were highlighted; (ii) Several topics on types of kits for viruses, bacteria and parasite work were heavily discussed throughout the lecture; (iii) Specific library preparation kits were highlighted including TruSeq, QIASeq Fx, Kappa, NexteraXT, RNA Access and DNAFlex; (iv) AmpureXP Beads clean up after PCR reactions and library preparation was emphasized as preferred method; (v) Different library validations, including qPCR, Qubit and TapeStation were highlighted as essentials for quality control (QC); (vi) Library pooling based on TapeStation and Qubit were introduced; (vii) Two exercises of how to calculate amount of each library for pooling were conducted. Preparations were made for the upcoming bioinformatics training.

Day 2

Hands-on training for NGS wet lab was performed with 24 participants. The participants were separated into two groups based on their NGS background and interests for hands-on performance. Group 1 prepared the NexteraXT library from the amplicons and assessed amplicons using both Qubit and TapeStation prior to NexteraXT library preparation. The NexteraXT libraries were validated using both Qubit and TapeStation. Group 2 validated the pooling based on the controls from the shipment and prepared sample sheets, the MiSeq instrument and PhiX controls. The libraries were loaded onto the Miseq. Bioinformatics training dataset was prepared on each computer. Server performance was tested for running the pipelines and tools needed for the training, and the training dataset analyses were executed to test functionality prior to the hands-on bioinformatics training.

Day 3

Hands-on wet lab activities from Day 2 were summarized and any questions and concerns were addressed. Lectures on laboratory project experimental design (to include bioinformatics), bioinformatics data cleaning and pre-processing, and genome assembly through reference mapping were performed, as well as exercises in experimental design and genome consensus calling. For hands-on bioinformatics training, the 24 participants were divided into six different groups, each group utilizing one training computer or server. Ngs_mapper was used as the example of a reference mapping pipeline. The first training was performed on the DENV fastq dataset, including training on usage of different stages of the pipeline, setting a desired reference genome and running the pipeline. After ngs-mapper jobs were completed, interpretation of the output, how to utilize data quality scores and depth of coverage, how to assess the performance of the sequencing and the genome assembly were performed. Manual QC and genome curation were performed. The second training dataset consisted of CHIKV fastqs and was used for training on multiple reference usage and reference selection, in addition to repeating the above steps for dataset one.

Day 4

Bioinformatics hands-on training was continued by evaluation of the CHIKV runs for reference genome selection. Based on the best reference choice, the reference mapping run was repeated. The repetition was incorporated on purpose to ensure better knowledge retention. Following reference mapping, the output of CHIKV assembly was evaluated and its genome curated. The data that were used for this training were purposefully chosen to be of lower quality, so that different challenges of genome assembly curation were highlighted, as well as the importance of QC and what consequences a lack of QC might result in. The last reference mapping analysis was performed on CHIKV data but now the participants learned how to change different pipeline thresholds, picking their own requirements for minimum base quality, consensus type output and the like. In addition, lectures were conducted covering theory of *de novo* genome assembly, assembly of bacterial genomes, and troubleshooting and maintenance of the MiSeq platform.

Day 5

A summary of wet lab activities and library pooling to obtain optimal cluster density was presented. An exercise aimed at the evaluation of several MiSeq runs was performed. Management of sequencing libraries and data, and prevention of chimeric sequence data generation and mislabeling were discussed. Bioinformatics training on the influenza dataset was performed separately since influenza virus has a segmented genome and bioinformatically, full genome assembly is slightly more complicated. How to recognize presence of influenza reassortment was covered. A workshop survey was distributed (**Supplementary Material**) and the workshop was concluded.

RESULTS

NGS and Bioinformatics Training Modules

A comprehensive training curriculum was constructed that consisted of standardized wet lab and bioinformatics theory modules (**Figure 2**) as well as hands-on training. The modules could be independently compiled into a set of theoretical lectures that could be adjusted for the existing laboratory tiers and specific knowledge gaps. As they were designed to meet the particular DoD surveillance needs, the modules were divided into two main wet lab sequencing and two main bioinformatics analyses approaches. The wet lab lectures could thus be adjusted to cover: (i) the theory of targeted sequencing, which is mainly used in response to epidemics and outbreaks of known pathogens; and (ii) the theory of metagenomics, which is usually used for pathogen discovery and identification. The bioinformatics lectures focused on: (i) the genome assembly and curation analyses, an essential part of outbreak genomic surveillance; and (ii) the bioinformatics of pathogen discovery, usually the most

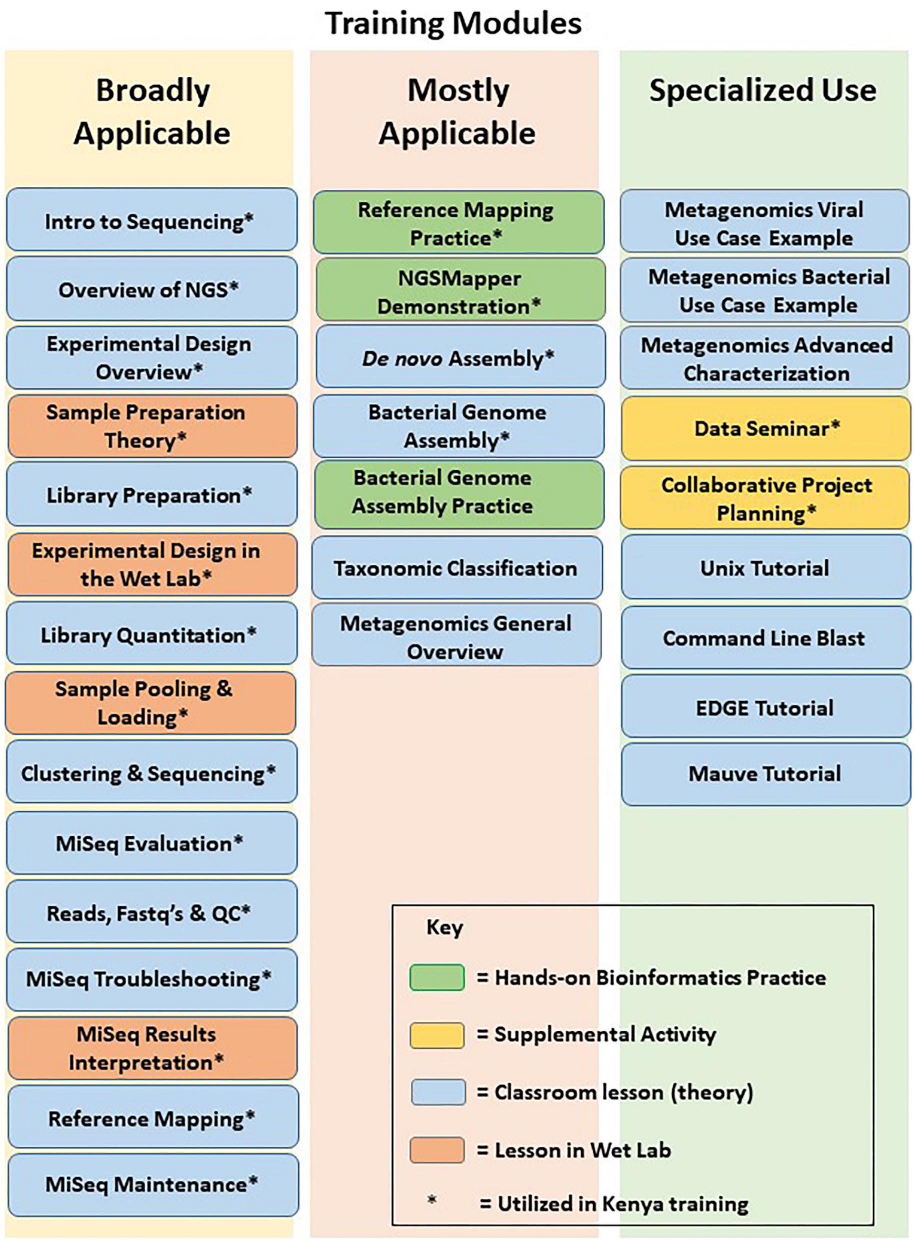

Training Modules
Broadly Applicable
Intro to Sequencing*
Overview of NGS*
Experimental Design Overview*
Sample Preparation Theory*
Library Preparation*
Experimental Design in the Wet Lab*
Library Quantitation*
Sample Pooling & Loading*
Clustering & Sequencing*
MiSeq Evaluation*
Reads, Fastq's & QC*
MiSeq Troubleshooting*
MiSeq Results Interpretation*
Reference Mapping*
MiSeq Maintenance*
Mostly Applicable
Reference Mapping Practice*
NGSMapper Demonstration*
De novo Assembly*
Bacterial Genome Assembly*
Bacterial Genome Assembly Practice
Taxonomic Classification
Metagenomics General Overview
Specialized Use
Metagenomics Viral Use Case Example
Metagenomics Bacterial Use Case Example
Metagenomics Advanced Characterization
Data Seminar*
Collaborative Project Planning*
Unix Tutorial
Command Line Blast
EDGE Tutorial
Mauve Tutorial
Key
🟩 = Hands-on Bioinformatics Practice
🟨 = Supplemental Activity
🟦 = Classroom lesson (theory)
🟧 = Lesson in Wet Lab
* = Utilized in Kenya training

FIGURE 2 | NGS and bioinformatics training modules. Modules used in training of MRD-A are denoted with an asterisk.

challenging aspect of basic sequencing-based biosurveillance. In addition to these, modules covering other parts of NGS and bioinformatics were included, such as theory of experimental design, troubleshooting, and equipment maintenance. The theory modules were complemented with development of corresponding hands-on wet lab and bioinformatics training of the above approaches.

NGS and Bioinformatics Training Deployment

Based on the results of the initial laboratory assessment, training was recommended for the GEIS partner US Army Medical Research Directorate – Africa (MRD-A) laboratories in Kenya. For MRD-A's initial needs, which mainly cover sequencing and analyses of known pathogen outbreaks and epidemics in the region, a 1 week on-site workshop was constructed where the wet lab targeted sequencing was covered in both lectures (specific assembled modules) and hands-on practice, followed by bioinformatics theory (specific assembled modules) and hands-on practice of pathogen genome assembly and curation (**Figure 2**). This approach was specifically designed based on the needs and gaps that were highlighted during the initial assessment of MRD-A capabilities. Participating in the training were representatives from various MRD-A and Kenya Medical Research Institute (KEMRI) laboratory divisions in Kenya: Basic Science, Viral Hemorrhagic Fevers, Entomology, Flu Lab, Antimicrobial Resistance, Sexually Transmitted Infections, Microbiology Hub-Kericho, Influenza, and KEMRI-Centers for Disease Control divisions (**Figure 3**). There was a total of 24 workshop participants.

We undertook a rapid evaluation of participants' self-reported baseline and post-workshop knowledge across ten skill domains related to genomic sequencing (**Supplementary Material**). We also determined individual-level gains in self-reported knowledge after completing the workshop. This was measured with a single hard-copy questionnaire administered after the workshop. This survey asked the participants to self-rate their knowledge in each skill domain on a customized scale of 1–10 (1 = "no prior knowledge", 10 = "high level of experience") before and after the workshop. Median baseline and post-workshop scores are presented in **Table 1**. While interpretation of

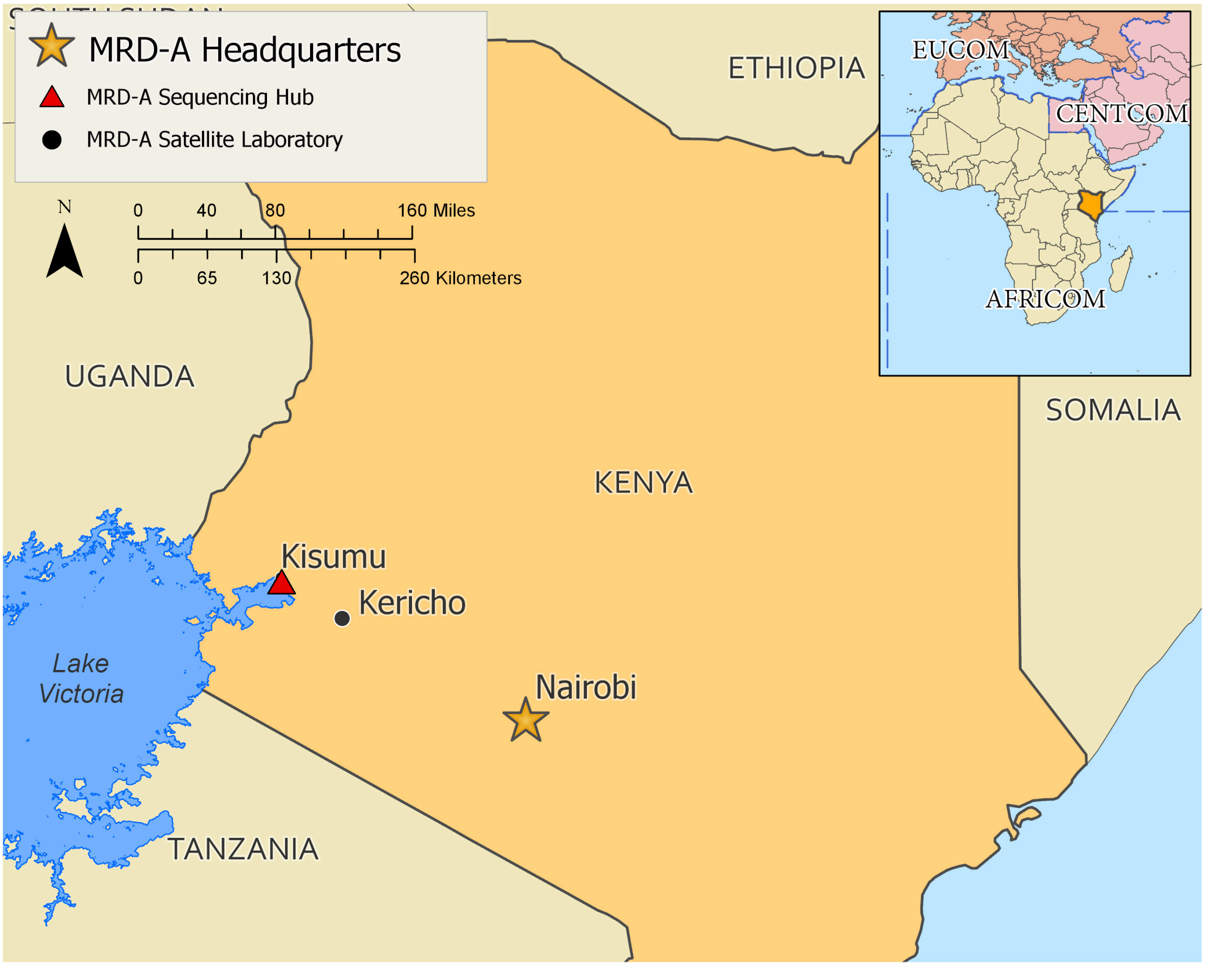


FIGURE 3 | A map of training performance site and participating partner laboratories from Kenya. Red triangle shows where the training was held.

TABLE 1 | Self-reported knowledge across skill domains of genomic sequencing (*n* = 21 respondents).

Knowledge domain	Pre-workshop score[a]	Post-workshop score[a]	Post-workshop gains	
	Median (IQR)	Median (IQR)	Median (IQR)	*p*-value[b]
NGS technology	4 (3,5)	7 (6,8)	3 (2,3)	<0.001
Illumina MiSeq sequencing chemistry	4 (2,5)	7 (6,8)	3 (2,4)	<0.001
NGS library preparation	5 (2,6)	8 (7,9)	3 (2,4)	<0.001
NGS library validation	2 (1,5)	8 (6,8)	4 (3,4)	<0.001
MiSeq run validation	2 (1,3)	6 (4,8)	3 (2,4)	<0.001
Experimental design for bioinformatics analysis	2 (1,4)	6 (5,8)	3 (2,4)	<0.001
FASTQ data cleaning and pre-processing	2 (1,5)	6 (5,8)	4 (2,4)	<0.001
Reference mapping	3 (1,5)	7 (6,9)	4 (1,5)	<0.001
Linux OS use and command line	2 (1,5)	5 (3,7)	1 (0,3)	<0.001
Consensus sequence calling and manual curation	2 (1,3)	6 (5,8)	4 (2,5)	<0.001

NGS, next generation sequencing; OS, operating system. [a]Maximum possible score is 10. [b]Derived from Wilcoxon signed rank test.

these metrics is limited due to the subjectivity of the self-reported knowledge measurements, particularly when measured at a single point in time, the IQR and range around the median reported knowledge scores did suggest that this sample of participants had varying expertise across each of these skill domains. Pre-training baseline scores suggested that the participants had, in particular, less self-reported expertise in NGS library validation, Illumina MiSeq run validation, experimental design for bioinformatics analysis, and FASTQ data cleaning and pre-processing.

There were substantial gains in self-reported knowledge across all skill domains (**Table 1**), with the notable exception of Linux OS and command line skills, suggesting that this is a particular area of residual training need. Indeed, Linux OS and command line skill had the lowest post-workshop self-reported knowledge scores. A module was later developed specifically to fill this gap (**Figure 2**). The questionnaire also measured the participants' perceptions on the most "useful" information learned during the NGS library and bioinformatics components of the workshop. This was measured by free-text open ended questions (**Table 2**).

The participants were also asked in which topics they felt they would like more training and experience (**Table 3**) and how to improve future iterations of this workshop (**Table 4**). The participant's responses all highlight the complexity and the diversity of considerations within NGS and bioinformatics. The many topics that can be covered and trained upon for the fields of infectious disease surveillance and control alone, and the associated time that it would take to train and educate the workforce, would indicate a large gap in the currently existing education programs.

TABLE 2 | Information reported by participants to be the most useful (*n* = 21 respondents)[a].

	n	%[b]
NGS library preparation	12	60
NGS library validation	7	35
Sample pooling	6	30
Tapestation use	5	25
Library normalization	5	25
Qubit use	2	10
QC for sequence reads	1	5
Experimental design considerations	1	5
Sequencing platform overview	1	5
MiSeq runs (hands-on experience)	1	5
Sample pre-processing	1	5
Nextera-XT protocol	1	5
MiSeq run troubleshooting	1	5
Genome assembly (reference mapping and curation)	15	71
Sequence read QC	8	38
De novo sequencing	4	19
Experimental considerations	3	14
Software (including NGS_mapper and IGV)	3	14
Mapping bacterial sequences	1	5
Output analysis	1	5

[a]Derived from open questions: "What information was most useful to you that this NGS library provided?" and "What information was most useful to you that this bioinformatic workshop provided?". [b]Some participants indicated > 1 item of information in response. QC, quality control.

DISCUSSION

The rapid growth and utility of NGS and bioinformatics for research and biosurveillance has resulted in the emergence of DoD requirements for implementation of sequencing and computational technologies, as well as access to highly trained and knowledgeable personnel in the fields of NGS and bioinformatics. Specifically the latter point remains one of the major challenges across the DoD, and even though bioinformatics programs have more recently gained larger momentum in academia, lack of workforce with early-on and/or specialized bioinformatics training is still palpable in the government settings, particularly in government labs outside the continental United States. Therefore, NGS and bioinformatics training programs for infectious disease surveillance have recently been developed by many government agencies or non-governmental organizations. Within the United States Government, Canada, and the European Union, there is movement towards training and coordinated promotion of standardized quality assurance and quality control practices for pathogen genome sequencing using NGS technologies (e.g., Illumina) (Cui et al., 2015; Gargis et al., 2016; Nadon et al., 2017). Some recent examples include the GenomeTrakr program at the Food and Drug Administration, Next Generation PulseNet at Centers for Disease Control, and the Global Microbial Identifier for food-borne pathogen surveillance (Moran-Gilad et al., 2015; Timme et al., 2018; Ribot et al., 2019). More recently, the SARS-CoV-2 Sequencing for Public Health Emergency Response,

TABLE 3 | Suggested topics for more training/experience, as reported by participants (n = 21 respondents)[a].

	n	%[b]
Phylogenetics and other advanced bioinformatics analysis	8	38
Metagenomics for pathogen discovery	4	19
De novo assembly	4	19
Linux OS/cluster	4	19
MiSeq loading and run evaluation	3	14
Bacterial genomics	3	14
Pipeline development (including open source bioinformatic tools)	3	14
Sequence assembly	2	10
Read QC	2	10
Library prep	1	5
16s and 18s molecular analysis	1	5
Sample pre-processing	1	5
Bioinformatic experimental design	1	5
SNP detection and variant calling	1	5
Sample sheet prep	1	5
Reference mapping	1	5
Plasmid sequencing	1	5
Recombination detection	1	5
Comparative genomics	1	5
Outbreak investigations	1	5

[a]Derived from open question: "What topic would you like more training/experience in (if any)?". [b]Some participants indicated more than one line item of information. QC, quality control.

TABLE 4 | Participants' suggestions for workshop improvements (n = 21 respondents)[a].

	n	%[b]
More time (longer workshop >1 week)	4	19
More hands on training/practical sessions (less theory/slides)	4	19
More time on bioinformatic data interpretation and analysis	3	14
Increased frequency of workshop with follow-up training	2	10
More wet lab time	2	10
More laptops	1	5
Split into beginner and advanced classes	1	5

[a]Derived from open question: "Suggestions for workshop improvement?". [b]Some participants had >1 suggestion.

Epidemiology, and Surveillance (SPHERES) national genomics consortium was set up by the Centers for Disease Control, to coordinate SARS-CoV-2 sequencing across the United States (Centers for Disease Control and Prevention, 2019). Within the DoD, the training designed and implemented by the GEIS Consortium aims to develop lasting and sustainable capabilities for pathogen genomic sequencing and bioinformatics at DoD medical research and public health laboratories in overseas locations.

Our experience in deploying a comprehensive yet customizable classroom and hands-on training in NGS and bioinformatics in Kenya was overall successful (see caveats of assessment below) and is a potential model for future training programs in similar environments. This training program consisted of foundational material in sequencing theory and experimental design which formed a basis for more applied modules in targeted sequencing and metagenomics. Additionally, hands-on NGS wet lab and bioinformatics modules were further tailored to meet the needs of the laboratory participants using information obtained from a baseline landscape assessment. This training shows that a highly modular and deployable set of NGS and bioinformatics workshop components can be used within the DoD network of medical research and public health laboratories to improve sequencing wet lab capability, and analysis and interpretation of pathogen genomic data gathered using NGS and bioinformatics.

Embedded within this training workshop was a post-self-assessment questionnaire to gauge immediate improvements in knowledge gained from the workshop materials. It is important to note that this questionnaire has several limitations including a small sample size, the immediate nature of the assessment tool which does not allow one to measure long-term benefits, and the fact that the assessment was only delivered through written evaluation and self-report. Further, more objective measurements of knowledge and skill gains after workshops may not directly translate into effective implementation and retention of these skills. The latter requires medium and longer term evaluations in an implementation science framework (Nilsen, 2015). However, these data do suggest that the participants have perceived that this workshop offered productive training which has led to substantial gains in knowledge. In similar bioinformatics trainings in LMICs, technological limitations were identified as an impediment to knowledge acquisition and long-term improvements in bioinformatics capability (Pollett et al., 2016). This training attempted to overcome these barriers by (a) providing training laptops, (b) providing recommendations for IT upgrades, bioinformatics software, and computer networking, and (c) upgrading local IT equipment for bioinformatics during the workshop.

Following this workshop a mechanism to facilitate reach back support with embedded long-term training and mentorship has been instituted to overcome challenges associated with long-term sustainability of a sequencing capability at MRD-A. Included in this 5-year NGS and bioinformatics implementation plan for MRD-A are: (i) continuous contact and support by the core DoD sequencing laboratories, (ii) repetition of training with focus on real data and troubleshooting, (iii) additional hands-on training in other wet lab and bioinformatics approaches to achieve capability diversification, (iv) development of local computational infrastructure for bioinformatics, and (v) regular assessments of wet lab and bioinformatics knowledge retention. Laboratory-level assessments of proficiency and skill retention 1–2 years post-training have included external review of raw sequence data and consensus genomes generated from GEIS funded surveillance projects. We also anticipate deploying periodic blinded panel of samples or data files for follow-up assessments of knowledge retention and capability development. At the end of this period, the goal is to achieve a high quality diversified portfolio of NGS and bioinformatics capabilities at the site, which then may serve as a central DoD hub for sequencing and advanced characterization of Force Health Protection (FHP) relevant pathogens in Africa.

The current COVID-19 pandemic has further highlighted the importance of access to the NGS and bioinformatics in laboratories throughout the world. This makes the need of workshops such as ours even greater. However, the pandemic has also made travel and in-person learning a challenge, and therefore, GEIS is planning on development of virtual versions of the workshops to continue development of this important DoD-wide capability. In addition, Oxford Nanopore's MinION platform has increasingly been used in pathogen outbreak studies for real-time in-field analyses throughout the world, including analyses of SARS-CoV-2 (Quick et al., 2016; Faria et al., 2018; Moore et al., 2020). Although training in the wet-lab and bioinformatics of this approach was not included in the workshop in Kenya to maintain simplicity and focus, the plan is to apply the modular approach for development and incorporation of a general DoD MinION-focused training for the GEIS partner laboratories. Currently, GEIS has established a separate MinION working group, and has been working in providing basic training in this technology to a subset of partner laboratories.

More broadly, the Consortium goal is the establishment of basic proficiencies and adopted norms in quality assurance and quality control in targeted (hybridization- or amplicon-based) and metagenomic sequencing for viral and bacterial pathogens leading to more reliable results which will ultimately improve DoD public health surveillance and response. An additional objective is the development and maintenance of advanced genomics and bioinformatics capabilities in the United States and priority overseas locations, in order to enhance global health surveillance and facilitate faster response to infectious disease outbreaks. Development of these capabilities with GEIS DoD laboratory partners will require sustained commitment and global coordination. The end results will be the ability to reliably and rapidly sequence, identify, and characterize pathogens of public health importance in order to improve biosurveillance efforts and inform FHP measures throughout the world.

DATA AVAILABILITY STATEMENT

All datasets presented in this study are included in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

IM, WR, KB-L, LM, and RJ designed the training modules and the workshop modules. IM, WR, LV, LM, KP, RJ, JW, SP, and RC prepared and instructed the workshop. IM, WR, LV, KP, SP, RC, JK, KB-L, LM, JW, and RJ performed workshop post-assessment and wrote the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the Armed Forces Health Surveillance Branch Global Emerging Infections Surveillance Section (ProMIS ID: P0149_19_AH) and WUN A1417.

ACKNOWLEDGMENTS

We would like to thank all members of the Global Emerging Infections Surveillance Next-Generation Sequencing and Bioinformatics Consortium for their active engagement and support of this effort. We would also like to thank Dr. Brett Forshey at GEIS for his thoughtful comments on the draft manuscript and Dr. Michael Wiley for his help on workshop design and computational efforts.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.577563/full#supplementary-material

REFERENCES

Anyamba, A., Chretien, J. P., Britch, S. C., Soebiyanto, R. P., Small, J. L., Jepsen, R., et al. (2019). Global disease outbreaks associated with the 2015-2016 El Nino Event. *Sci. Rep.* 9:1930. doi: 10.1038/s41598-018-38034-z

Beam W. E. Jr., Grayston, J. T., and Watten, R. H. (1959). Second Asian influenza epidemics occurring in vaccinated men aboard U.S. Navy vessels. *J. Infect. Dis.* 105, 38–44. doi: 10.1093/infdis/105.1.38

Centers for Disease Control and Prevention (2019). *SARS-CoV-2 Sequencing for Public Health Emergency Response, Epidemiology, and Surveillance 2020.* Available online at: https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/spheres.html (accessed August 6, 2020).

Chakhunashvili, G., Wagner, A. L., Machablishvili, A., Karseladze, I., Tarkhan-Mouravi, O., Zakhashvili, K., et al. (2017). Implementation of a sentinel surveillance system for influenza-like illness (ILI) and severe acute respiratory infection (SARI) in the country of Georgia, 2015-2016. *Int. J. Infect. Dis.* 65, 98–100. doi: 10.1016/j.ijid.2017.09.028

Chang, K. S., Kim, G. H., Ha, Y. R., Jeong, E. K., Kim, H. C., Klein, T. A., et al. (2018). Monitoring and control of *Aedes albopictus*, a vector of Zika Virus, near residences of imported Zika Virus patients during 2016 in South Korea. *Am. J. Trop. Med. Hyg.* 98, 166–172. doi: 10.4269/ajtmh.17-0587

Colby, D. J., Trautmann, L., Pinyakorn, S., Leyre, L., Pagliuzza, A., Kroon, E., et al. (2018). Rapid HIV RNA rebound after antiretroviral treatment interruption in persons durably suppressed in Fiebig I acute HIV infection. *Nat. Med.* 24, 923–926. doi: 10.1038/s41591-018-0026-6

Coleman, R., Eick-Cost, A. A., Hawksworth, A. W., Hu, Z., Lynch, L., Myers, C. A., et al. (2018). Department of defense end-of-season influenza vaccine effectiveness estimates for the 2017-2018 season. *MSMR* 25, 16–20.

Cowell, A. N., Valdivia, H. O., Bishop, D. K., and Winzeler, E. A. (2018). Exploration of *Plasmodium vivax* transmission dynamics and recurrent infections in the Peruvian Amazon using whole genome sequencing. *Genome Med.* 10:52. doi: 10.1186/s13073-018-0563-0

Cui, H. H., Erkkila, T., Chain, P. S., and Vuyisich, M. (2015). Building international genomics collaboration for global health security. *Front. Public Health* 3:264. doi: 10.3389/fpubh.2015.00264

Earhart, K. C., Beadle, C., Miller, L. K., Pruss, M. W., Gray, G. C., Ledbetter, E. K., et al. (2001). Outbreak of influenza in highly vaccinated crew of U.S. Navy ship. *Emerg. Infect. Dis.* 7, 463–465. doi: 10.3201/eid0703.017320

Ehrenberg, P. K., Shangguan, S., Issac, B., Alter, G., Geretz, A., Izumi, T., et al. (2019). A vaccine-induced gene expression signature correlates with protection against SIV and HIV in multiple trials. *Sci. Transl. Med.* 11:507. doi: 10.1126/scitranslmed.aaw4236

Faria, N. R., Kraemer, M. U. G., Hill, S. C., Goes de Jesus, J., Aguiar, R. S., Iani, F. C. M., et al. (2018). Genomic and epidemiological monitoring of yellow fever virus transmission potential. *Science* 361, 894–899. doi: 10.1126/science.aat7115

Frey, K. G., Biser, T., Hamilton, T., Santos, C. J., Pimentel, G., Mokashi, V. P., et al. (2016). Bioinformatic characterization of mosquito Viromes within the Eastern United States and puerto rico: discovery of novel viruses. *Evol. Bioinform.* 12(Suppl 2), 1–12. doi: 10.4137/EBO.S38518

Gargis, A. S., Kalman, L., and Lubin, I. M. (2016). Assuring the quality of next-generation sequencing in clinical microbiology and public health laboratories. *J. Clin. Microbiol.* 54, 2857–2865. doi: 10.1128/jcm.00949-16

Grubaugh, N. D., Saraf, S., Gangavarapu, K., Watts, A., Tan, A. L., Oidtman, R. J., et al. (2019). Travel surveillance and genomics uncover a hidden Zika outbreak during the waning epidemic. *Cell* 178, 1057–1071.e11. doi: 10.1016/j.cell.2019.07.018

Guerra, R. I., Ore, M., Valdivia, H. O., Bishop, D. K., Ramos, M., Mores, C. N., et al. (2019). A cluster of the first reported *Plasmodium ovale* spp. infections in Peru occuring among returning UN peace-keepers, a review of epidemiology, prevention and diagnostic challenges in nonendemic regions. *Malar J.* 18:176. doi: 10.1186/s12936-019-2809-8

Juma, D. W., Muiruri, P., Yuhas, K., John-Stewart, G., Ottichilo, R., Waitumbi, J., et al. (2019). The prevalence and antifolate drug resistance profiles of *Plasmodium falciparum* in study participants randomized to discontinue or continue cotrimoxazole prophylaxis. *PLoS Negl. Trop. Dis.* 13:e0007223. doi: 10.1371/journal.pntd.0007223

Kijak, G. H., Sanders-Buell, E., Chenine, A. L., Eller, M. A., Goonetilleke, N., Thomas, R., et al. (2017). Rare HIV-1 transmitted/founder lineages identified by deep viral sequencing contribute to rapid shifts in dominant quasispecies during acute and early infection. *PLoS Pathog.* 13:e1006510. doi: 10.1371/journal.ppat.1006510

Kim, W. K., No, J. S., Lee, D., Jung, J., Park, H., Yi, Y., et al. (2019). Active targeted surveillance to identify sites of emergence of hantavirus. *Clin. Infect. Dis.* 70, 464–473. doi: 10.1093/cid/ciz234

Koka, H., Sang, R., Kutima, H. L., and Musila, L. (2018). *Coxiella burnetii* detected in tick samples from pastoral communities in Kenya. *Biomed. Res. Int.* 2018:8158102. doi: 10.1155/2018/8158102

Kumar, S., Stecher, G., and Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Mol. Biol. Evol.* 33, 1870–1874. doi: 10.1093/molbev/msw054

LaBreck, P. T., Rice, G. K., Paskey, A. C., Elassal, E. M., Cer, R. Z., Law, N. N., et al. (2018). Conjugative transfer of a novel staphylococcal plasmid encoding the biocide resistance gene, qacA. *Front. Microbiol.* 9:2664. doi: 10.3389/fmicb.2018.02664

Lee, S. H., Kim, W. K., No, J. S., Kim, J. A., Kim, J. I., Gu, S. H., et al. (2017). Dynamic circulation and genetic exchange of a shrew-borne hantavirus, Imjin virus, in the republic of Korea. *Sci. Rep.* 7:44369. doi: 10.1038/srep44369

MacPherson, W., Herringham, W., Elliott, T., and Balfour, A. (1923). *History of the Great War Based On Official Documents: Medical Services Diseases of the War*. London: HMSO.

Maljkovic Berry, I., Eyase, F., Pollett, S., Konongoi, S. L., Joyce, M. G., Figueroa, K., et al. (2019a). Global outbreaks and origins of a Chikungunya Virus variant carrying mutations which may increase fitness for *Aedes aegypti*: revelations from the 2016 Mandera, Kenya Outbreak. *Am. J. Trop. Med. Hyg.* 100, 1249–1257. doi: 10.4269/ajtmh.18-0980

Maljkovic Berry, I., Melendrez, M. C., Bishop-Lilly, K. A., Rutvisuttinunt, W., Pollett, S., Talundzic, E., et al. (2019b). Next generation sequencing and bioinformatics methodologies for infectious disease research and public health: approaches, applications, and considerations for development of laboratory capacity. *J. Infect. Dis.* 221(Suppl. 3), S292–S307. doi: 10.1093/infdis/jiz286

Maljkovic Berry, I., Melendrez, M. C., Li, T., Hawksworth, A. W., Brice, G. T., Blair, P. J., et al. (2016). Frequency of influenza H3N2 intra-subtype reassortment: attributes and implications of reassortant spread. *BMC Biol.* 14:117. doi: 10.1186/s12915-016-0337-3

Mbala-Kingebeni, P., Aziza, A., Di Paola, N., Wiley, M. R., Makiala-Mandanda, S., Caviness, K., et al. (2019). Medical countermeasures during the 2018 Ebola virus disease outbreak in the North Kivu and Ituri provinces of the democratic republic of the Congo: a rapid genomic assessment. *Lancet Infect. Dis.* 19, 648–657. doi: 10.1016/S1473-3099(19)30118-5

Millar, E. V., Rice, G. K., Elassal, E. M., Schlett, C. D., Bennett, J. W., Redden, C. L., et al. (2017). Genomic characterization of USA300 methicillin-resistant *Staphylococcus aureus* (MRSA) to evaluate intraclass transmission and recurrence of skin and soft tissue infection (SSTI) among high-risk military trainees. *Clin. Infect. Dis.* 65, 461–468. doi: 10.1093/cid/cix327

Millar, E. V., Rice, G. K., Schlett, C. D., Elassal, E. M., Cer, R. Z., Frey, K. G., et al. (2019). Genomic epidemiology of MRSA infection and colonization isolates among military trainees with skin and soft tissue infection. *Infection* 47, 729–737. doi: 10.1007/s15010-019-01282-w

Moore, S., Penrice-Randal, R., Alruwaili, M., Dong, X., Pullan, S., Carter, D., et al. (2020). Amplicon based MinION sequencing of SARS-CoV-2 and metagenomic characterisation of nasopharyngeal swabs from patients with COVID-19. *medRxiv*[Preprint] doi: 10.1101/2020.03.05.20032011

Moran-Gilad, J., Sintchenko, V., Pedersen, S. K., Wolfgang, W. J., Pettengill, J., Strain, E., et al. (2015). Proficiency testing for bacterial whole genome sequencing: an end-user survey of current capabilities, requirements and priorities. *BMC Infect. Dis.* 15:174. doi: 10.1186/s12879-015-0902-3

Mullins, K. E., Hang, J., Clifford, R. J., Onmus-Leone, F., Yang, Y., Jiang, J., et al. (2017). Whole-genome analysis of *Bartonella ancashensis*, a novel pathogen causing verruga peruana, rural ancash region, Peru. *Emerg. Infect. Dis.* 23, 430–438. doi: 10.3201/eid2303.161476

Murray, C. K., Yun, H. C., Markelz, A. E., Okulicz, J. F., Vento, T. J., Burgess, T. H., et al. (2015). Operation united assistance: infectious disease threats to deployed military personnel. *Mil. Med.* 180, 626–651. doi: 10.7205/milmed-d-14-00691

Nadon, C., Van Walle, I., Gerner-Smidt, P., Campos, J., Chinen, I., Concepcion-Acevedo, J., et al. (2017). PulseNet International: vision for the implementation of whole genome sequencing (WGS) for global food-borne disease surveillance. *Euro Surveill.* 22:30544. doi: 10.2807/1560-7917.ES.2017.22.23.30544

Nilsen, P. (2015). Making sense of implementation theories, models and frameworks. *Implement Sci.* 10:53. doi: 10.1007/978-3-030-03874-8_3

Philipson, C., Davenport, K., Voegtly, L., Lo, C. C., Li, P. E., Xu, J., et al. (2017). Brief protocol for EDGE bioinformatics: analyzing microbial and metagenomic NGS data. *Bio Protoc.* 7:e2622. doi: 10.21769/BioProtoc.2622

Pollett, S., Fauver, J. R., Maljkovic, B, I., Melendrez, M., Morrison, A., Gillis, L. D., et al. (2019). Genomic epidemiology as a public health tool to combat mosquito-borne virus outbreaks. *J. Infect. Dis.* 221(Suppl. 3), S308–S318.

Pollett, S., Leguia, M., Nelson, M. I., Maljkovic Berry, I., Rutherford, G., Bausch, D. G., et al. (2016). Feasibility and effectiveness of a brief, intensive phylogenetics workshop in a middle-income country. *Int. J. Infect. Dis.* 42, 24–27. doi: 10.1016/j.ijid.2015.11.001

Quick, J., Loman, N. J., Duraffour, S., Simpson, J. T., Severi, E., Cowley, L., et al. (2016). Real-time, portable genome sequencing for Ebola surveillance. *Nature* 530, 228–232.

Ribot, E. M., Freeman, M., Hise, K. B., and Gerner-Smidt, P. (2019). PulseNet: entering the age of next-generation sequencing. *Foodborne Pathog. Dis.* 16, 451–456. doi: 10.1089/fpd.2019.2634

Riddle, M. S., Kaminski, R. W., Williams, C., Porter, C., Baqar, S., Kordis, A., et al. (2011). Safety and immunogenicity of an intranasal *Shigella* flexneri 2a Invaplex 50 vaccine. *Vaccine* 29, 7009–7019. doi: 10.1016/j.vaccine.2011.07.033

Rivers, C., Chretien, J. P., Riley, S., Pavlin, J. A., Woodward, A., Brett-Major, D., et al. (2019). Using "outbreak science" to strengthen the use of models during epidemics. *Nat. Commun.* 10:3102.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., et al. (2011). Integrative genomics viewer. *Nat. Biotechnol.* 29, 24–26. doi: 10.1038/nbt.1754

Rocha, C., Bernal, M., Canal, E., Rios, P., Meza, R., Lopez, M., et al. (2019). First report of New Delhi metallo-beta-lactamase carbapenemase-producing *Acinetobacter baumannii* in Peru. *Am. J. Trop. Med. Hyg.* 100, 529–531. doi: 10.4269/ajtmh.18-0802

Salje, H., Lessler, J., Maljkovic Berry, I., Melendrez, M. C., Endy, T., Kalayanarooj, S., et al. (2017). Dengue diversity across spatial and temporal scales: local structure and the effect of host population size. *Science* 355, 1302–1306. doi: 10.1126/science.aaj9384

Shanks, G. D., and Hodge, J. (2011). The ability of seasonal and pandemic influenza to disrupt military operations. *J. Mil. Veterans Health* 19, 13–18.

Srijan, A., Margulieux, K. R., Ruekit, S., Snesrud, E., Maybank, R., Serichantalergs, O., et al. (2018). Genomic characterization of nonclonal MCR-1-positive multidrug-resistant *Klebsiella pneumoniae* from clinical samples in Thailand. *Microb. Drug Resist.* 24, 403–410. doi: 10.1089/mdr.2017.0400

Sugiharto, V. A., Widjaja, S., Hartman, L. J., Williams, M., Myers, T. E., and Simons, M. P. (2019). Zika virus surveillance in active duty U.S. military and dependents through the Naval Infectious Diseases Diagnostic Laboratory. *MSMR* 26, 18–23.

Timme, R. E., Rand, H., Sanchez Leon, M., Hoffmann, M., Strain, E., Allard, M., et al. (2018). GenomeTrakr proficiency testing for foodborne pathogen surveillance: an exercise from 2015. *Microb. Genom.* 4:e000185.

Viral Diseases Branch WRAIR (2016). *ngs mapper*. Available online at: https://github.com/VDBWRAIR/ngs_mapper (accessed August 6, 2020).

Waickman, A. T., Victor, K., Li, T., Hatch, K., Rutvisuttinunt, W., Medin, C., et al. (2019). Dissecting the heterogeneity of DENV vaccine-elicited cellular immunity using single-cell RNA sequencing and metabolic profiling. *Nat Commun.* 10:3666.

Wiley, M. R., Fakoli, L., Letizia, A. G., Welch, S. R., Ladner, J. T., Prieto, K., et al. (2019). Lassa virus circulating in Liberia: a retrospective genomic characterisation. *Lancet Infect. Dis.* 19, 1371–1378. doi: 10.1016/s1473-3099(19)30486-4

Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Army, Department of the Navy, Department of Defense, or United States Government. Several of the authors are United States Government employees. This work was prepared as part of their official duties. Title 17 U.S.C. §105 provides that "Copyright protection under this title is not available for any work of the United States Government." Title 17 U.S.C. §101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties.

Conflict of Interest: LV and RC were employed by company Leidos, United States.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

ORIGINAL RESEARCH
published: 06 October 2020
doi: 10.3389/fmicb.2020.584699

PUMAA: A Platform for Accessible Microbiome Analysis in the Undergraduate Classroom

Keith Mitchell[1†], Jiem Ronas[1†], Christopher Dao[1], Amanda C. Freise[1], Serghei Mangul[2], Casey Shapiro[3] and Jordan Moberg Parker[1]*

[1] Department of Microbiology, Immunology and Molecular Genetics, University of California, Los Angeles, Los Angeles, CA, United States, [2] Department of Clinical Pharmacy, School of Pharmacy, University of Southern California, Los Angeles, CA, United States, [3] Center for Educational Assessment, Center for the Advancement of Teaching, University of California, Los Angeles, Los Angeles, CA, United States

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Tammy Tobin,
Susquehanna University,
United States
Henrik R. Nilsson,
University of Gothenburg, Sweden

***Correspondence:**
Jordan Moberg Parker
jmobergparker@ucla.edu

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology,
a section of the journal Frontiers in Microbiology

Received: 17 July 2020
Accepted: 14 September 2020
Published: 06 October 2020

Citation:
Mitchell K, Ronas J, Dao C, Freise AC, Mangul S, Shapiro C and Moberg Parker J (2020) PUMAA: A Platform for Accessible Microbiome Analysis in the Undergraduate Classroom. Front. Microbiol. 11:584699. doi: 10.3389/fmicb.2020.584699

Improvements in high-throughput sequencing makes targeted amplicon analysis an ideal method for the study of human and environmental microbiomes by undergraduates. Multiple bioinformatics programs are available to process and interpret raw microbial diversity datasets, and the choice of programs to use in curricula is largely determined by student learning goals. Many of the most commonly used microbiome bioinformatics platforms offer end-to-end data processing and data analysis using a command line interface (CLI), but the downside for novice microbiome researchers is the steep learning curve often required. Alternatively, some sequencing providers include processing of raw data and taxonomy assignments as part of their pipelines. This, when coupled with available web-based or graphical user interface (GUI) analysis and visualization tools, eliminates the need for students or instructors to have extensive CLI experience. However, lack of universal data formats can make integration of these tools challenging. For example, tools for upstream and downstream analyses frequently use multiple different data formats which then require writing custom scripts or hours of manual work to make the files compatible. Here, we describe a microbial ecology bioinformatics curriculum that focuses on data analysis, visualization, and statistical reasoning by taking advantage of existing web-based and GUI tools. We created the Program for Unifying Microbiome Analysis Applications (PUMAA), which solves the problem of inconsistent files by formatting the output files from several raw data processing programs to seamlessly transition to a suite of GUI programs for analysis and visualization of microbiome taxonomic and inferred functional profiles. Additionally, we created a series of tutorials to accompany each of the microbiome analysis curricular modules. From pre- and post-course surveys, students in this curriculum self-reported conceptual and confidence gains in bioinformatics and data analysis skills. Students also demonstrated gains in biologically relevant statistical reasoning based on rubric-guided evaluations of open-ended survey questions and the Statistical Reasoning in Biology Concept Inventory. The PUMAA program and associated analysis tutorials enable students and researchers with no computational experience to effectively analyze real microbiome datasets to investigate real-world research questions.

Keywords: microbiome, 16S rRNA, software tool, GUI (Graphical User Interface), undergraduate education, curriculum, data visualisation, targeted amplicon sequencing

INTRODUCTION

Engaging undergraduates in research has been consistently demonstrated to increase students' performance, attitudes, and retention in sciences (Lopatto, 2004; Russell et al., 2007; Eagan et al., 2013). In particular, course-based undergraduate research experiences (CUREs) have been touted as an inclusive and scalable model to bring these benefits to a diverse set of student populations (Harrison et al., 2011; Bangera and Brownell, 2014; Corwin et al., 2015; Shapiro et al., 2015; Hanauer et al., 2017). Microbiome research using marker gene metabarcoding is an attractive direction for CUREs, as sample collection is relatively straightforward and advances in sequencing technologies and reduced cost have made the acquisition of marker gene microbiome data easier than ever (Clooney et al., 2016; Jovel et al., 2016). The large microbiome datasets using a combination of marker genes targeting bacteria and archaea (16S), eukaryotes (18S), and fungi (ITS) give students an opportunity to ask a variety of questions ranging from the composition of their own oral microbiome to plant–microbe interactions (Rosenwald et al., 2012; Sanders and Hirsch, 2014; Wang et al., 2015; Weber et al., 2018; Parks et al., 2020; Sewall et al., 2020).

We designed a microbial ecology CURE as part of the interdepartmental Competency-Based Research Laboratory Curriculum at the University of California, Los Angeles (Shapiro et al., 2015). In this two-term (two 10-week quarters) curriculum students work in teams to conduct self-directed research projects, with a focus on developing critical thinking and quantitative skills. Under the umbrella of an instructor designated overarching research question, students in the microbial ecology CURE formulate and test hypotheses about the microbiomes of different environments. The functional profiles of microbial communities are just as important as the taxonomic composition (Langille, 2018), and the questions of "who is there?" and "what are they doing there?" are the guiding questions for the curriculum. In the first wet-lab term they use both cultivation-dependent techniques such as isolating bacteria from the soil and characterizing their functional capabilities, and cultivation-independent techniques such as extraction of environmental DNA (eDNA) for 16S rRNA (16S) sequencing. In the second computer-lab term they use a variety of phylogenetics programs and bioinformatics tools for analysis of microbiome taxonomic community profiles and Piphillin predicted functional profiles (Narayan et al., 2020).

A major challenge for the development of microbiome research for undergraduates is that marker gene amplicon microbiome data provided by sequencing providers requires a number of bioinformatic processing steps before it can be easily analyzed and visualized, a process with which not all instructors or researchers have familiarity (Carey and Papin, 2018; Garcia-Milian et al., 2018). Many of the available end-to-end data analysis packages such as Quantitative Insights Into Microbial Ecology (QIIME/QIIME 2) (Caporaso et al., 2010; Bolyen et al., 2019), mothur, and the Pipeline for Environmental DNA Metabarcoding Analysis (PEMA) (Zafeiropoulos et al., 2020) have steep learning curves, requiring at least some command line interface (CLI) programming skills, or familiarity with R (R: The R Project for Statistical Computing) in the case of phyloseq (McMurdie and Holmes, 2013, 2015) and PEMA, in order to perform data analysis and visualization. Teaching these skills may be outside the scope of the average undergraduate microbiology classroom. Fortunately, there are several microbiome data analysis and visualization tools that do not require command line, such as the Shiny web app ranacapa (Kandlikar et al., 2018) or locally installed programs with graphical user interfaces (GUIs) such as Statistical Analysis of Metagenomic Profiles (STAMP) (Parks and Beiko, 2010; Parks et al., 2014) and Cytoscape (Shannon et al., 2003). These are attractive tools for use in the undergraduate bioinformatics classroom where there is lack of time to devote to the steep learning curve necessary for installation and use of command line programs (Mangul et al., 2017).

Even with the increasing availability of GUI analysis tools, there is still the problem that the data output file formats from QIIME or custom commercial and academic pipelines such as MrDNA (mrdnalab, 2020) and Anacapa (Curd et al., 2019) do not match the data input file formats required for the GUI and web-based analysis and visualization tools. Formatting the different analysis pathway files into a single pipeline is a non-trivial task requiring either running scripts or hours of manual reformatting. To address this problem, we created PUMAA, the Program for Unifying Microbiome Analysis Applications, which takes the output files from QIIME, Anacapa, or MrDNA and reformats them directly for use in downstream GUI or web-based applications for microbiome analysis. Additionally, PUMAA both prepares files for upload to Piphillin for prediction of functional genes from the 16S taxonomy data, and queries the KEGG database to annotate the Piphillin gene predictions (Iwai et al., 2016; Narayan et al., 2020). Inferring functional profiles from 16S rRNA marker genes using programs like PiCRUSt (Langille et al., 2013; Douglas et al., 2020) or Piphillin are accessible options for researchers without the resources to perform full functional metagenomics (Laudadio et al., 2019).

Since classroom time is limited and our curriculum learning objectives focus on microbiome data analysis, visualization, and statistical reasoning rather than learning programming languages, the instructional staff runs the PUMAA program to generate the files necessary for several different GUI or web-based tools and provide them to students. The bioinformatics curriculum is scaffolded such that the students' progress in their microbiome research from phylogenies of individual bacterial isolates, to simple microbial community qualitative analyses, to quantitative diversity metrics, to statistical analysis of the microbial community profiles. We developed accompanying instructional modules, video tutorials, and a lab manual to teach students both the theory behind the analysis tools and the skills needed for visualizing and performing biostatistical methods on the data. The key tools and tutorials include inferring phylogenetic trees, analyzing community profiles and diversity metrics using Microsoft Excel pivot tables and ranacapa, statistical analysis of taxonomic and inferred functional profiles using STAMP, and using KEGG to assign functions to genes.

The curriculum was assessed using entry/exit surveys designed to gauge the students' confidence in integrating computational analysis with microbiology, and the Statistical Reasoning in Biology Concept Inventory (SRBCI) (Deane et al., 2016). Analysis of entry and exit surveys saw an increase in students' self-reported conceptual understanding and confidence levels in using the analysis tools, as well as improved competencies with biostatistics as demonstrated by improvement in the SRBCI post-test. The PUMAA program and associated instructional materials provide a scaffolded learning experience for undergraduate students and make microbiome bioinformatics analyses accessible to novice researchers.

PUMAA – PROGRAM FOR UNIFYING MICROBIOME ANALYSIS APPLICATIONS OVERVIEW

Analyzing metabarcoded microbiome data is a complex multi-step process. Next-generation sequencing produces a variety of data files, which then need to be processed and quality checked before assigning taxonomic profiles (Zhang, 2016; Almeida et al., 2018). Most sequencing providers include basic bioinformatic processing in their pipelines, and provide taxonomic abundance tables and sequence FASTA files along with the raw data. These files can be then used in downstream analysis and visualization applications. However, each taxonomic assignment platform and analysis or visualization tool may have different data input and output formats that need to be reconciled, or have significant data pre-processing steps that need to occur before the various analyses can be performed.

Some sequencing providers, such as MrDNA (mrdnalab, 2020), produce taxonomy abundance tables that must be rearranged in order to be compatible with most visualization programs, but even for those that are in the right general format, many tools have specific formatting requirements. For example, the STAMP tool enforces a "strict hierarchy" requirement where no classification of taxonomy can exist at a lower level than one which was left unclassified. The following classification, from phylum to species: "Proteobacteria, Gammaproteobacteria, Enterobacteriales, unclassified, *Escherichia*, unclassified," will produce errors in STAMP because the family is unclassified even though the genus is classified. In addition, STAMP requires that all unclassified columns must be labeled so and cannot be left blank. Another tool, Cytoscape, requires that each sample identification and taxonomic identification be a unique row where the weight corresponds to the quantity of the given instance in order to create a network type visualization. Web server-based programs such as Piphillin (Iwai et al., 2016) may have file size upload limitations, necessitating sub-setting of the data. These formatting and processing steps need to be carried out independently on the taxonomy or functional data for each of the desired analysis and visualization platforms (**Figure 1A**).

PUMAA, the Program for Unifying Microbiome Analysis Applications, provides the solution to these problems by integrating all of the formatting and pre-processing steps required for the platforms and tools discussed here into a single unified protocol with an easy installation procedure (**Figure 1B**). In addition, PUMAA is easily expandable as it provides the ability to add a new analysis tool or taxonomic ID platform with one added operation. The PUMAA protocol unifies existing data analysis and visualization tools by formatting common amplicon (16S/18S/ITS) taxonomic data outputs from a variety of sources to be compatible with the input formats required for multiple basic and advanced microbiome analysis tools. Additionally, PUMAA integrates Piphillin inferred functional microbiome composition from the 16S taxonomy data. PUMAA provides both a CLI as well as a GUI to accommodate a spectrum of potential users. A CLI version is implemented to allow users with UNIX experience, or those who are interested in learning, to customize their analysis and build upon/automate the provided scripts (Mangul et al., 2017). The GUI is ideal for novice microbiome researchers with little experience on UNIX based systems, who are interested in quickly visualizing their microbiome marker gene amplicon data. Initial installation of the GUI does require running a small set of terminal installation commands, but subsequent usage is straightforward.

PUMAA Supports Input From Various Microbiome Data Pipelines

Currently PUMAA supports three microbiome raw data processing platforms and/or services: MrDNA, Anacapa, and QIIME 2 (Bolyen et al., 2019; Curd et al., 2019; mrdnalab, 2020). PUMAA formats the taxonomic abundance tables and sequence files created by these platforms for any marker gene amplicons, including 16S, 18S, ITS, and others, for downstream analysis and visualization (**Figure 2**).

MrDNA

MrDNA is a commercial full-service next generation sequencing provider that offers 16S, 18S, and ITS amplicon sequencing on a variety of platforms. Regardless of the sequencing platform, MrDNA provides free comprehensive taxonomic analysis in addition to raw data processing using their proprietary pipeline. The pipeline generates operational taxonomic unit (OTU) abundance tables with taxonomic identities and representative FASTA sequence files at each taxonomic level (kingdom, phylum, class, order, family, genus, species).

Anacapa

Anacapa is a software tool kit developed to process environmental DNA (eDNA) sequence data and assign taxonomy data for six marker genes targeting bacteria, archaea, algae, fungi, protozoa, plants, and animals (Curd et al., 2019). Anacapa creates a custom reference library for marker genes, generates amplicon sequence variants (ASV), and assigns taxonomies at each taxonomic level (domain, phylum, class, order, family, genus, species). ASVs have been proposed as a finer resolution replacement for OTU clustering based on sequence similarity (Callahan et al., 2017). Anacapa output includes a detailed taxonomy table with sequences and abundances for each ASV, as well as tables with taxonomies summarized at various percent confidence intervals.

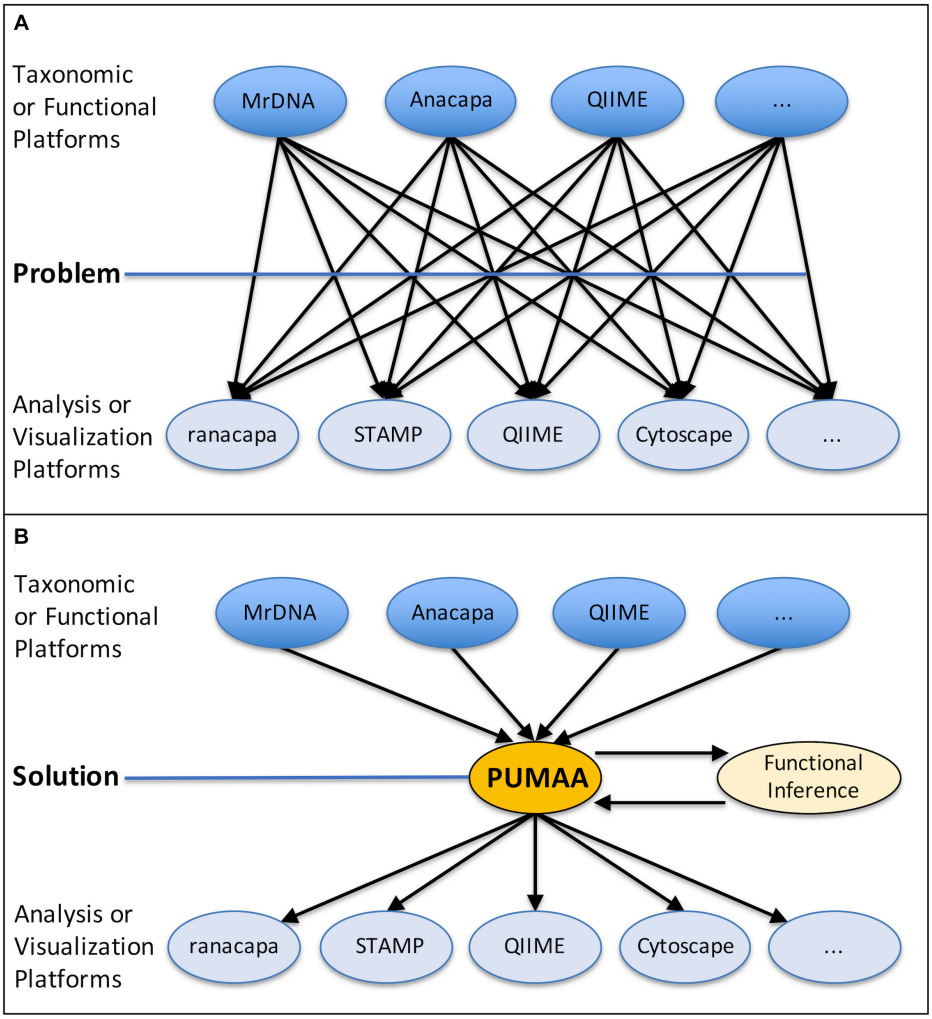


FIGURE 1 | The problem presented and the PUMAA solution. **(A)** The current problem is lack of unification of outputs from different taxonomic identification or functional inference platforms (MrDNA, Anacapa, QIIME, etc.) and the input data required by prospective analysis and visualization tools (ranacapa, STAMP, QIIME, Cytoscape, etc.). **(B)** PUMAA is a streamlined pipeline unifying the output files from multiple platforms and converting them to the input files necessary for varied analysis and visualization tools.

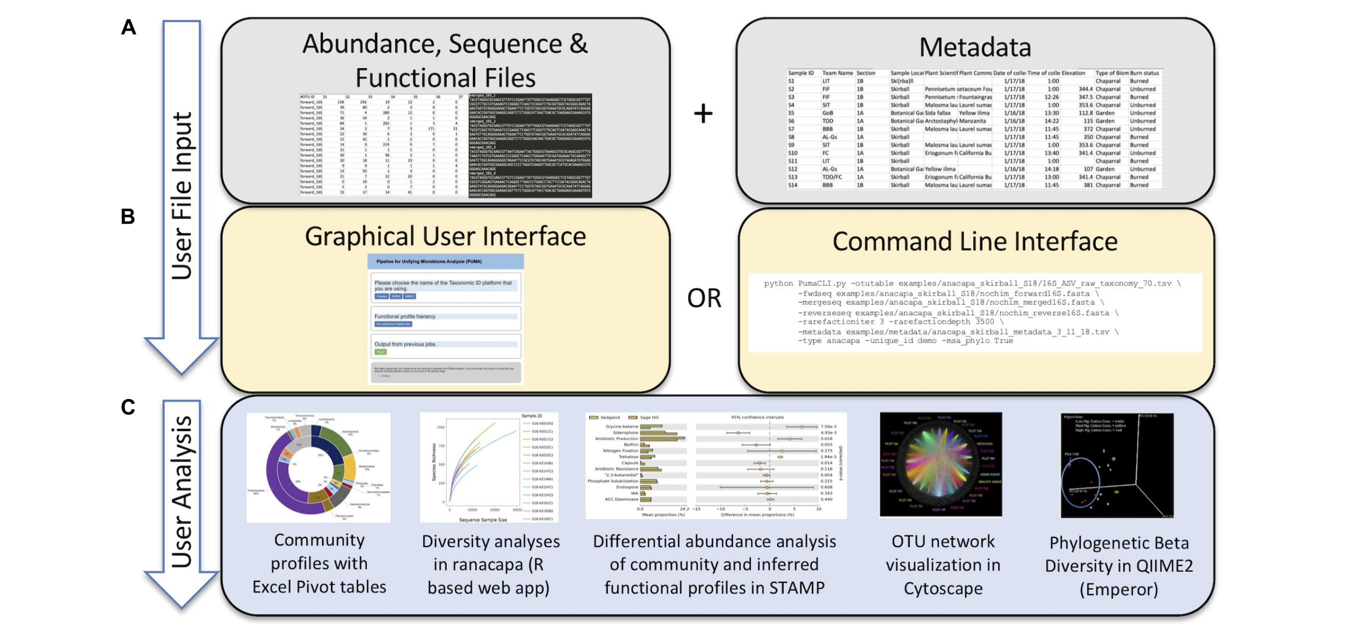
QIIME

QIIME is a powerful and widely adopted package for processing microbiome data, from raw sequences through taxonomy and data visualization. Tutorials and published protocols are available to walk users through standard data processing (Kuczynski et al., 2011), but the scope of QIIME may be daunting for novice users, even with the availability of the QIIME 2 Studio graphical interface (Bolyen et al., 2019). It also remains difficult to convert to other analysis/visualization platforms since QIIME provides users with OTU files and sequence files in the '.qza' format, which is unique to its platform.

PUMAA Supports Piphillin for Inferred Functional Profile Analysis

PUMAA formats taxonomic abundance (OTU or ASV) tables and representative sequence files for prediction of metagenomic content by Piphillin, which uses nearest-neighbor matching of 16S rRNA amplicons and full genomes (Iwai et al., 2016). Piphillin has the added benefits of a web interface and the ability to use any standard abundance table and representative sequence FASTA file, rather than relying on taxonomic assignments assigned from a specific reference phylogenetic tree, as in PiCRUSt (Langille et al., 2013). PiCRUSt2 has an extended database of reference genomes and broader compatibility, but still requires use of the command line for implementation (Douglas et al., 2020). A drawback to Piphillin is the 10 MB limit placed on uploaded file sizes in the web version. PUMAA addresses this by producing subset abundance and FASTA files that comply with these limits. The subset files are uploaded to the Piphillin server[1], and reference database and percent identity cutoffs are chosen [PUMAA currently only supports

[1] https://piphillin.secondgenome.com/


A
User File Input
Abundance, Sequence & Functional Files
+
Metadata
B
Graphical User Interface
OR
Command Line Interface
C
User Analysis
Community profiles with Excel Pivot tables
Diversity analyses in ranacapa (R based web app)
Differential abundance analysis of community and inferred functional profiles in STAMP
OTU network visualization in Cytoscape
Phylogenetic Beta Diversity in QIIME2 (Emperor)

Sample ID	Team Name	Section	Sample Locat	Plant Scientif	Plant Commc	Date of colle	Time of colle	Elevation	Type of Biom	Burn status
S1	LIT	1B	Ski[rba]ll			1/17/18	1:00		Chaparral	Burned
S2	FIF	1B	Skirball	Pennisetum setaceum Fou		1/17/18	1:00	344.4	Chaparral	Unburned
S3	FIF	1B	Skirball	Pennisetum :	Fountaingras	1/17/18	12:26	347.5	Chaparral	Burned
S4	SIT	1B	Skirball	Malosma lau	Laurel sumac	1/17/18	1:00	353.6	Chaparral	Unburned
S5	GoB	1A	Botanical Ga	Sida fallax	Yellow ilima	1/16/18	13:30	112.8	Garden	Unburned
S6	TDD	1A	Botanical Ga	Arctostaphyl	Manzanita	1/16/18	14:22	115	Garden	Unburned
S7	BBB	1B	Skirball	Malosma lau	Laurel sumac	1/17/18	11:45	372	Chaparral	Unburned
S8	AL-Gs	1A	Skirball			1/17/18	11:45	350	Chaparral	Burned
S9	SIT	1B	Skirball	Malosma lau	Laurel sumac	1/17/18	1:00	353.6	Chaparral	Burned
S10	FC	1A	Skirball	Eriogonum fa	California Bu	1/17/18	13:40	341.4	Chaparral	Unburned
S11	LIT	1B	Skirball			1/17/18	1:00		Chaparral	Burned
S12	AL-Gs	1A	Botanical Ga	Yellow ilima		1/16/18	14:18	107	Garden	Unburned
S13	TDD/FC	1A	Skirball	Eriogonum fa	California Bu	1/17/18	13:00	341.4	Chaparral	Burned
S14	BBB	1B	Skirball	Malosma lau	Laurel sumac	1/17/18	11:45	381	Chaparral	Burned

```
python PumaCLI.py -otutable examples/anacapa_skirball_S18/16S_ASV_raw_taxonomy_70.tsv \
       -fwdseq examples/anacapa_skirball_S18/nochim_forward16S.fasta \
       -mergeseq examples/anacapa_skirball_S18/nochim_merged16S.fasta \
       -reverseseq examples/anacapa_skirball_S18/nochim_reverse16S.fasta \
       -rarefactioniter 3 -rarefactiondepth 3500 \
       -metadata examples/metadata/anacapa_skirball_metadata_3_11_18.tsv \
       -type anacapa -unique_id demo -msa_phylo True
```

FIGURE 2 | Protocol of the PUMAA software. **(A)** The first panel as part of the "User File Input" displays the simple protocol to be performed by the user such as uploading metadata and various data formats of supported operational taxonomic unit, sequence, and functional file types. **(B)** The second panel as part of the "User File Input" displays the two forms of user interaction with PUMAA, through the GUI and CLI, which will enable community and functional profile analysis. **(C)** "User Analysis" shows the possible platforms for visualizing community/functional composition data enabled by user input such as STAMP, Excel, QIIME 2, and Cytoscape.

KEGG (Kanehisa, 2000; Kanehisa et al., 2004)], then results are emailed to the user as compressed.tar files. The other drawback to Piphillin is that it provides abundance tables for all predicted genes and pathways (identified by K and KO numbers), but not the associated annotations to assign biological information to the K/KO numbers. To address this, the PUMAA inferred function protocol also performs queries to the KEGG database in order to properly annotate the genes and pathways returned by Piphillin. Prior to PUMAA, this annotation process required command-line experience or labor-intensive manual curation.

PUMAA Supports a Variety of Analysis and Visualization Platforms

There are a wide variety of research questions that can be addressed using amplicon microbiome data, and the methods used for data analysis and visualization will vary based on the needs of the researcher. PUMAA focuses on processing and formatting user data to be compatible with a suite of readily available web-based or GUI data analysis and visualization tools. Using the PUMAA supported tools, researchers can explore data and test hypotheses by linking groups of samples or environmental parameters, otherwise known as metadata, to diversity metrics, community composition, and inferred functional profiles.

We have integrated PUMAA into a broad range of research analysis options (from simple to advanced) and visualization types (from bar charts to network analyses). In addition, PUMAA has options to complete data processing such as rarefaction subsampling to normalize for variation in sequence numbers between samples (McMurdie and Holmes, 2014; Willis, 2019), multiple sequence alignment (MSA) using MUSCLE (Edgar, 2004), and inference of phylogenetic trees using FastTree (Price et al., 2010).

Microsoft Excel

Microsoft Excel pivot tables are an easy way to begin to summarize the massive amounts of data in taxonomic abundance tables for visualizations of the overall community profile of different samples at different taxonomic levels (i.e., kingdom/domain, phylum, class, order, family, genus, species). Excel can also be easily used to make simple (non-statistical) comparisons of sample abundances at different taxonomic levels.

ranacapa

ranacapa (Kandlikar et al., 2018) is a user-friendly Shiny web application designed to explore biodiversity using environmental DNA metabarcoding data. It includes interactive visualizations and brief explanations of sequencing depth, alpha and beta diversity, and taxonomy distribution analyses such as bar plots and heatmaps. ranacapa was developed as an extension of the Anacapa toolkit (Curd et al., 2019), but can prove slightly difficult to access from other taxonomic identification platforms, like that of MrDNA.

STAMP (Statistical Analysis of Metagenomic Profiles)

STAMP (Parks et al., 2014) is a downloadable graphical interface that can quickly generate publication-quality graphics for differential abundance analysis of either taxonomy or functional pathway data without the need to write code or use command-line interface. STAMP supports parametric and nonparametric statistical hypothesis testing for two-sample, two-group, and multiple-group comparisons. It emphasizes the use of effect size and confidence intervals in assessing biological relevance, and supports a variety of visualizations, including heatmaps, PCA plots, extended error bar plots, box plots, and bar plots.

QIIME 2 (Quantitative Insights Into Microbial Ecology)

QIIME 2 (Bolyen et al., 2019) provides numerous interactive and advanced data visualization tools and plugins for evaluation of metagenomic profiles (Caporaso et al., 2010; Kuczynski et al., 2011). Although QIIME can be used for end-to-end data analysis, some researchers may receive data processed by other platforms (e.g., MrDNA or Anacapa) and wish to feed the data back into the QIIME pipeline for analysis.

Cytoscape

Cytoscape (Kohl et al., 2011) is a unique open-source locally downloadable tool that enables the visualization of networks between community and functional profiles. Basic network analysis and visualization can be performed with the core distribution, with many additional features available as Cytoscape Apps.

Methods – PUMAA Protocol

Overview

The user executes a single script for both the GUI and CLI versions in order to execute the program. The PUMAA protocol consists of two key parts: (1) Production of all files for taxonomic community analysis, and (2) production of all files required for inferred functional analysis. PUMAA solves the problem of going from any of the taxonomic identification platforms to the multitude of visualization and analysis tools available by enforcing standardized files as part of the unification process. The user first obtains input files from one of the three supported pipelines (MrDNA, Anacapa, or QIIME2), identifies the metadata necessary for identifying and comparing samples (**Figure 2A**), and chooses to run PUMAA through either the GUI or CLI (**Figure 2B**). PUMAA verifies that the metadata sample IDs match the input data, then produces output files that can be used for a variety of analysis platforms (**Figure 2C**).

Protocol: PUMAA Installation and Requirements

PUMAA is freely available under the Apache-2.0 license at https://github.com/keithgmitchell/PUMAA and is supported by MacOSX and Linux; in addition, PUMAA works on Windows machines after installing the Linux subsystem Comprehensive installation instructions are provided on the Github page. Given software install is handled using conda, all versions of MacOSX and Linux that support the conda environment management software are viable options for usage and make for consistent and user-friendly install (Mangul et al., 2019). Issues or questions with the software can be submitted using the github issues feature: https://github.com/keithgmitchell/PUMAA/issues.

PUMAA is written in Python and the application's GUI is written using the Django web framework running locally. The example datasets all run on a laptop and use <1GB of memory when the MSA and Phylogenetic tree production is set as false. The QIIME 2 and MrDNA datasets run on a laptop and use <1GB of memory when the MSA and Phylogenetic tree production is set as true. The Anacapa dataset was unsuccessful on a laptop with 16GB RAM and was evaluated using a high-performance computing (HPC) cluster with 32GB of RAM and 3 h of runtime. Therefore, to produce a MSA and phylogenetic tree for datasets of this size, access to an HPC cluster, experience with CLI, and experience running jobs on HPC clusters may be required (**Table 1**).

Protocol: PUMAA Verifies Metadata

The user uploads their metadata describing the samples, taxonomy abundance (OTU or ASV) table and sequences from any given supported platform. The first part of the PUMAA protocol verifies the metadata and the taxonomy table to be sure the two files have consistent, alphanumeric sample identifiers which are unique compared to other forms of metadata validation (Rideout et al., 2016). This is a critical step as identifiable metadata is necessary for many downstream analysis steps, and some tools limit the types of characters accepted in the sample identifiers (e.g., underscores, but not periods, are acceptable in sample IDs in ranacapa).

Protocol: PUMAA Produces Files for Community Profile Analysis

PUMAA performs a variety of functions on the taxonomic abundance and sequence files in order to support the suite of tools discussed above. These functions include optional sample rarefaction at a user defined depth and number of iterations (max = 10) (Weiss et al., 2017), multiple sequence alignment by MAFFT (Katoh and Standley, 2013), phylogenetic tree construction via FastTree 2 (Price et al., 2010), and file formatting and annotation for ranacapa, STAMP, QIIME 2, Piphillin, and Excel. The protocol produces files for community profile analysis in the folder 'output,' or some other specified directory as an argument in the CLI. The output folder contains time-stamped subfolders for each PUMAA run, each containing subfolders with ready-to-run files for community profile analyses in Microsoft Excel, STAMP, ranacapa, and Cytoscape. In addition, pre-processed feature table (taxonomy), metadata, and phylogenetic tree files are created that can be imported directly into the QIIME 2 pre-configured virtual machine. A variety of analyses such as alpha- and beta-diversity can be performed in QIIME 2, as well as principal component analysis based on phylogenetic diversity metrics.

Protocol: PUMAA Produces Files for Inferred Functional Profile Analysis

The PUMAA protocol consists of three steps necessary for the generation and visualization of inferred functional profiles. The first step is automatically performed at the same time as the generation of the community profile analysis files. PUMAA creates a "piphillin" subfolder in

TABLE 1 | Dataset size, runtime, and memory usage with no rarefaction performed across the three example datasets.

Dataset	Dataset size (ASV/OTU count *10,000)	Fasta file size (MB)	Runtime (minutes)	FastTree/MAFFT peak memory usage (GB)	Python memory usage (GB)
MrDNA examples	0.3229	0.868	0.0778	0.207	0.02
QIIME 2 examples	0.0759	0.115	0.00517	0.044	0.02
Anacapa examples	3.6	1.789	1.24	12	0.075

the time-stamped output subfolder. This folder contains the original data formatted as a 'phiphillinotu.csv' taxonomic abundance table and a 'phiphillinseqs.fasta' representative sequence file. If the FASTA file exceeds the file size limit of 10 MB enforced by the Piphillin server, PUMAA subsamples the data into the number of necessary file sets of '.fasta' and '.csv' files (e.g., piphillinseqs1.fasta; piphillinseqs2.fasta; piphillinotu.csv1.csv; piphillinotu.csv2.csv). Second, each of the sets of Piphillin files in the output directory are uploaded to the Piphillin functional inference web server, which returns '.tar' files to the user via email.

Finally, the '.tar' files can then be run directly in the PUMAA protocol, which produces files for functional analysis that can be visualized using many of the same tools used for community profile analysis, including STAMP, Excel, and QIIME 2. Importantly, the PUMAA protocol also performs queries to the KEGG database using the KEGG genes to pathway API in order to properly annotate the Piphillin gene estimations (Kawashima et al., 2003). The BRITE hierarchy file of the KEGG database is downloaded and used to evaluate the functional hierarchy based on Piphillin pathway estimations. This ensures that estimated gene expression levels and hierarchy levels are inferred using the actively updated information. Annotating the genes and pathway expression from Piphillin is necessary when producing data visualizations with informative identifiers, and greatly reduces the need for manual querying of KEGG.

PUMAA produces a timestamped output subfolder for the functional profile files, including a gene description and functional hierarchy file designated for use in STAMP and Excel. This file contains annotated gene names and functional pathways, as opposed to just "K number" identifiers, and vastly increases the efficiency and ease of data analysis and visualization. PUMAA also produces weighted functional network files for usage in Cytoscape, which is a platform for visualizing important gene networks between samples.

Sample Data

The sample data used here and in the tutorials was generated by UCLA students in the winter and spring quarters of 2018, where they investigated the effect on rhizosphere microbial communities following the Skirball wildfire of December 2017 (Skirball Fire, 2020). Sample collection kits and sample sequencing were provided by the California Environmental DNA (CALeDNA) program, a community science initiative monitoring California's biodiversity through eDNA (Meyer et al., 2019), and the 16S sequences were processed using the Anacapa toolkit (Curd et al., 2019). The sample data for QIIME 2 is the same as the "moving pictures" human microbiome example dataset available on the QIIME 2 website[2].

Results

PUMAA Input and Output Files

The PUMAA pipeline creates output files formatted specifically for the needed input files for each of the data analysis and visualization platforms described in **Supplementary Table 1**.

PUMAA – CURRICULUM OVERVIEW

The Microbiology, Immunology, and Molecular Genetics (MIMG) undergraduate degree program at UCLA requires the completion of a two-quarter authentic research experience. An option to fulfill this requirement is to take the MIMG 109AL/BL: Research Immersion Laboratory in Microbiology series. This laboratory series is designed to prepare its students with the proper background and training to work in microbiology research, and has been demonstrated to improve their critical thinking and research skills as part of the life science curriculum (Shapiro et al., 2015). The 109AL/BL laboratory curriculum is discovery-based and driven by student-generated hypotheses tested using both cultivation-dependent and cultivation-independent techniques. The first term emphasizes experimental design and isolation of bacteria in a wet lab environment, and the second term focuses on the analysis of 16S sequencing data from individual isolates and 16S rRNA microbial community profiles. Students work in teams to conduct an original research project within the context of an overarching research question for the microbial ecology course, focusing on the interactions between plants and soil-associated bacteria. Recent course projects have involved collaborations with researchers at UCLA and beyond studying plant–microbe interactions in California grasslands (Kandlikar et al., 2020), analysis of the soil microbial communities of a Los Angeles urban farm (St. Clair et al., 2020), and a longitudinal study on the recovery of soil microbial communities following the 2017 Skirball fire in Los Angeles, CA, United States. The Skirball fire project was conducted in conjunction with the California Environmental DNA (CALeDNA) program's efforts to catalog California's biodiversity (Meyer et al., 2019).

In order for the MIMG 109AL/BL lab series to respond to the need for more computationally minded scientists (Bialek and Botstein, 2004; Campbell et al., 2007; Brewer and Smith, 2011), it

[2] https://docs.qiime2.org/2020.2/tutorials/moving-pictures/

was necessary to introduce new modules and tutorials that would sufficiently integrate bioinformatics and statistics with biology in ways that aspiring undergraduate researchers can comprehend (Aikens and Dolan, 2014). We created a comprehensive set of step-by-step tutorials (documents, presentations, and videos) designed to provide students with the necessary theory and skills to use the GUI analysis and visualization tools described in Section 2.3 (Excel, ranacapa, and STAMP), as well as the theory behind inference of metagenomic functional profiles using Piphillin. Although not a biostatistics course, the PUMAA-associated curriculum allows these students to learn about the computational tools available to researchers and the importance of integrating their knowledge of microbiology with statistical and quantitative support.

All tutorials are publicly available at https://sites.google.com/g.ucla.edu/pumaa/home.

First Term – Sample Collection and Bacterial Isolation/Characterization

The first term of the curriculum takes place in the wet lab and closely follows the cultivation-dependent experiments described in units 1–4 of the "I, Microbiologist" (Sanders and Miller, 2010) course textbook and lab manual. In brief, students collect bulk soil and decide on enrichment strategies for isolation of bacteria related to their research questions (e.g., antibiotic production and resistance or plant growth-promoting properties). Students then perform phenotypic characterization of bacterial isolates and 16S rRNA PCR and sequencing. In addition to collecting bulk soil for cultivation-dependent experiments, students also collect separate soil samples for environmental DNA (eDNA) extraction and 16S rRNA high-throughput sequencing for bacterial community profile analysis.

Second Term – Bioinformatics Analysis of 16S rRNA Genes Using PUMAA

In the second term, students use bioinformatics to interpret, expand, or refine 16S rRNA gene datasets generated in MIMG 109AL. Students generate 16S rRNA phylogenetic trees to assign taxonomic identities to their isolates and use statistical tools to make comparisons of the microbial communities from different environments. The course is divided into five Core Concept Modules. The first module (Phylogenetic Trees) concludes the analysis of bacterial isolates, and the other four modules focus on microbiome data analysis and visualization using the PUMAA output files: Community Profiles, Diversity Metrics, Statistical Analysis of Taxonomic Profiles, and Inferring Metagenomic Functional Profiles (**Figure 3A**). Students could also elect to perform optional advanced independent analysis on their data using QIIME or Cytoscape. Each of the modules includes written and/or video tutorials and was assessed with a combination of reading assessments and reflection questions (**Figure 3B**). This bioinformatics course was assessed using pre- and post-course concept inventories and surveys. Learning objectives, activities, and tutorials for each of the Core Concept Modules are outlined in **Supplementary File 1**.

Curriculum Assessment Methods

Study Sample

The study sample consisted of six cohorts of junior and senior level students who enrolled in MIMG 109BL (Advanced Research in Microbiology) in Spring 2016, Spring 2017, Winter 2018, Spring 2018, Winter 2019, and Spring 2019. This yielded an initial population of 143 students. **Table 2** provides a summary of demographic characteristics for these students. Instructor J.M.P. taught the spring cohorts and instructor A.F. taught the winter cohorts. Prerequisites for enrollment in MIMG 109BL included MIMG 109AL (Research Immersion in Microbiology) and either Statistics 13 (Introduction to Statistical Methods for Life and Health Sciences) or Life Sciences 40 (Statistics of Biological Systems).

Assessment Data Collection and Analyses

The study utilized two sources of data: student assignments and self-report surveys. Data collected included qualitative and quantitative measures. UCLA's Institutional Review Board (IRB) gave approval to work with human subjects on all aspects of the assessment (IRB #10-000904).

Administration of Self-Report Surveys

Two self-report surveys were administered to all students in the course. Surveys included a broad collection of open- and closed-ended questions, some developed by the instructors and evaluation team. Students were given the entry survey at the start of the second term and asked to indicate how well they thought they understood key learning goals related to data analysis and their confidence in their ability to analyze data using various visualization plots. The exit survey was completed at the end of the term and had matched questions to the first survey, as well as additional survey questions asking them to assess the quality and usefulness of the tutorials and instructional materials. Both surveys also included open-ended content-related questions. The surveys were piloted in 2016 and 2017 and were given to students anonymously through the course management system as low-stakes (completion points) assessments to increase response rate and reduce response bias (Furnham, 1986). Starting in Winter 2018, these items were added to a comprehensive curricular assessment plan administered electronically by external evaluators (see Shapiro et al., 2015) for details on survey data collection). Of the 143 students who took the course between Spring 2016 and Spring 2019, 141 completed the first survey (98.6% response rate) and 132 completed the second survey (92.3% response rate). The surveys are available as **Supplementary File 2**.

Administration of SRBCI Concept Inventory

The Statistical Reasoning in Biology Concept Inventory (SRBCI) is a series of multiple-choice questions to test students on concepts including statistical significance, basic graph/trend interpretation, and assessing hypotheses based on results (Deane et al., 2016). The twelve questions on the SRBCI pre- and post-tests are designed to identify students' common misconceptions in statistical analysis and track their learning progress as a result of the pedagogical interventions. The concept inventory

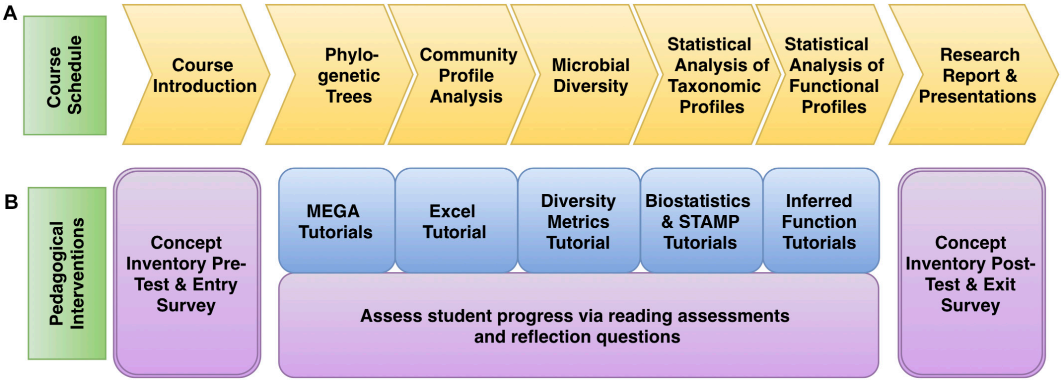

A
Course Schedule
Course Introduction
Phylo-genetic Trees
Community Profile Analysis
Microbial Diversity
Statistical Analysis of Taxonomic Profiles
Statistical Analysis of Functional Profiles
Research Report & Presentations
B
Pedagogical Interventions
Concept Inventory Pre-Test & Entry Survey
MEGA Tutorials
Excel Tutorial
Diversity Metrics Tutorial
Biostatistics & STAMP Tutorials
Inferred Function Tutorials
Assess student progress via reading assessments and reflection questions
Concept Inventory Post-Test & Exit Survey

FIGURE 3 | Microbiome analysis course schedule with pedagogical interventions. **(A)** The progress of the course followed the concept goals as outlined in yellow. **(B)** The pedagogical interventions are described, with tutorials in blue and assessment materials in purple.

TABLE 2 | Study sample demographics.

	Number of students (*N*)	Percent of students (%)
Female	81	56.6%
Transfer student[a]	34	23.8%
URM[b]	34	23.8%
Pell Grant Recipient[c]	53	37.1%
Total	143	100%

Academic terms: Spring 2016, Spring 2017, Winter 2018, Spring 2018, Winter 2019, Spring 2019. [a]Transfer to UCLA, usually from a 2-year institution. [b]Under-Represented Minority (URM) students include American Indian, Native American, Black Non-Hispanic, and Hispanic students. [c]Received Pell Grant for one or more terms while enrolled at UCLA; Pell Grant Recipient is a proxy for low socioeconomic status.

was administered as an anonymous low-stakes (ungraded) in-class activity at the start and end of the second term to the first two cohorts of students in Spring 2016 and Spring 2017. The study design, intended to gauge authentic learning gains across the curriculum by reducing "math anxiety" (Ashcraft and Moore, 2009), necessarily resulted in the inability to assess individual student learning gains using this metric. The pre-test and post-test were administered to a total of 52 and 50 students, respectively. Statistical reasoning gains between the pre-test and post-test groups were assessed using descriptive and Mann–Whitney nonparametric tests to account for variations in sample size.

Analyses of Closed-Ended Quantitative Survey Data

The closed-ended survey questions quantitatively ranked the students' agreement with a statement or confidence with a certain concept using a five-point Likert scale ranging from "Not at all" to "Very well/Very confident." Scores for matched questions were averaged across all participants to compare results from the Entry and Exit Surveys. Survey items asking students about the usefulness of learning activities were rated on a five-point Likert scale where 1 = "Don't remember," 2 = "Not useful," 3 = "Somewhat useful," 4 = "Very useful," and 5 = "Essential." Descriptive analyses of matched pre/post-survey close-ended items were conducted to explore students' change in self-reported confidence and changes in their self-reported levels of understanding. To test for statistical differences between the overall means of the Entry and Exit Survey items, descriptive and Mann–Whitney nonparametric tests were performed on the combined survey data from all cohorts to account for variations in sample size. Because the responses for the Spring 2016 and Spring 2017 surveys were anonymous, we were unable to pair the data by student. Wilcoxon signed ranks (paired nonparametric) tests were conducted on just the surveys administered by the external evaluators from Winter 2018 to Spring 2019, in order to see if there were differences between the all the data and the matched data. Since both sets of tests were significant, we were confident in using the aggregated data and the Mann–Whitney nonparametric tests to report our results.

Analyses of Open-Ended Qualitative Survey Data

Open-ended questions related to course content were included in the Entry and Exit surveys, allowing students to respond in their own words. Of particular interest was a question that asked students to describe the relationship between *p*-value (statistical significance) and effect size (biological significance). A 4-point rubric assessing students' level of proficiency with statistical concepts was used to gather direct evidence of student learning gains (**Supplementary File 3**). Student responses to open-ended questions were scored on a scale of 1 point = no familiarity (i.e., students indicated that they are not familiar with the concept), and 2–4 points for novice, intermediate, and advanced proficiency, respectively. Responses left blank were unscored. All student responses (both pre and post) were randomized and pooled by the external evaluator, then provided to the raters. The rubric was developed and refined by J.R., A.F., and J.M.P. through iterative rounds of scoring a subset of sample responses followed by consensus discussion. All responses were scored independently by all three raters, and interrater reliability (IRR) as determined by Randolph's free-marginal multirater kappa, was 0.49 (61.8% overall agreement) indicating moderate agreement. To account for the IRR variations, the median score for each response was used to assess whether pre-post gains were statistically significant between the groups using both the Mann–Whitney nonparametric test and a t-test.

Curriculum Assessment Results

Conceptual and Confidence Gains From Self-Reported Surveys

We wanted to assess if students would be able to formulate and statistically test hypotheses linking environmental parameters (metadata) to diversity metrics, community composition, and inferred functional profiles. Students were assessed using entry/exit surveys designed to gauge the students' comfort with integrating computational analysis with microbiology. At the beginning of the term the students reported, on average, "very little" understanding of key learning objectives such as how to use and assess the results of bioinformatics databases, and which statistical tests to use and how to interpret them (**Figure 4A**). By the end of the term students reported they understood these

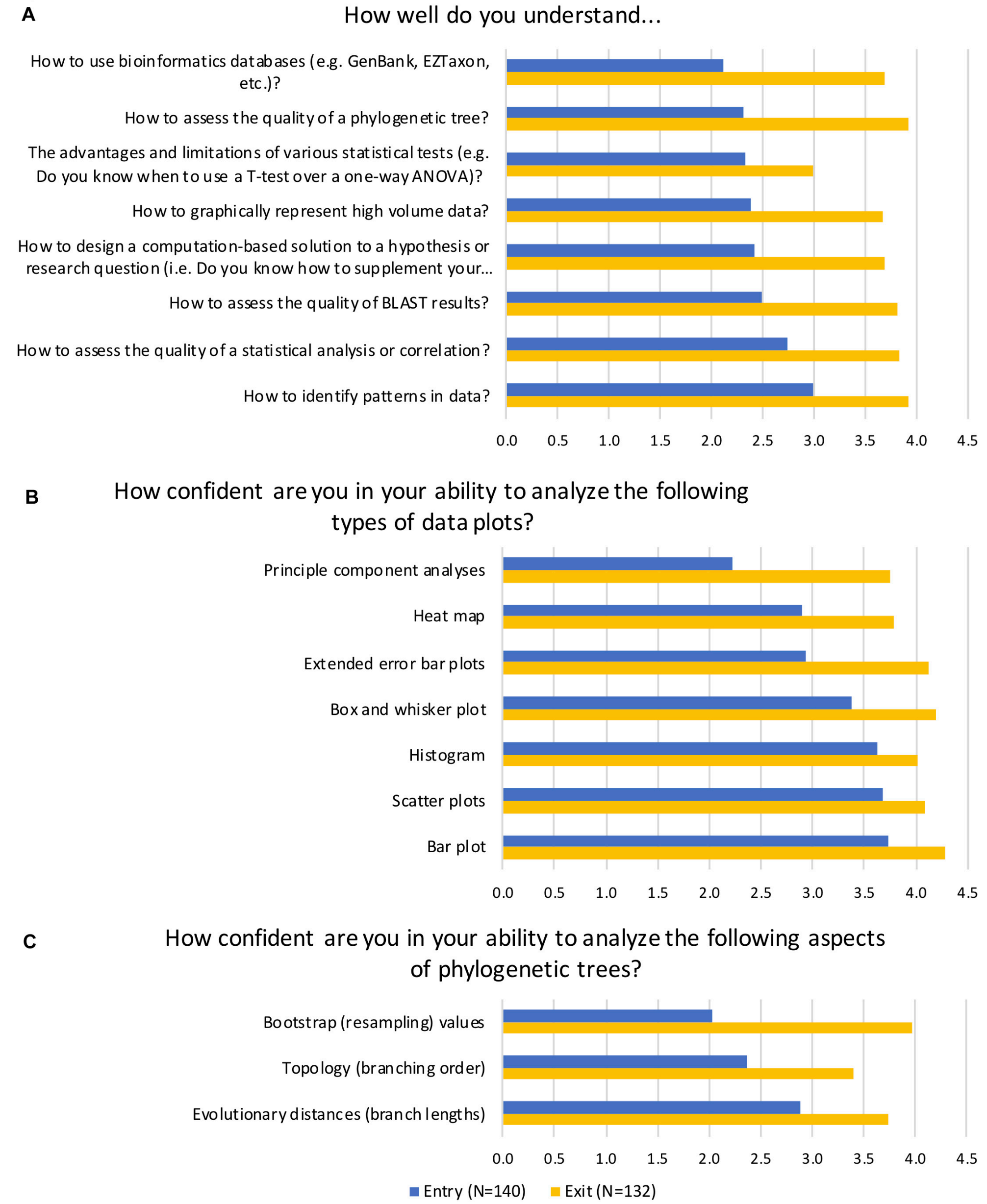


FIGURE 4 | Average ranked responses to selected entry and exit survey questions. In self-reported survey questions, students were asked to indicate **(A)** their level of understanding of key learning goals, **(B)** their confidence in their ability to analyze common data plots, and **(C)** their confidence in their ability to analyze aspects of phylogenetic trees. Average scores on a five-point Likert scale are reported for matched questions. A score of 1 = Not at all, 2 = Very little/Not very, 3 = Fairly well/confident, 4 = Quite well/confident, and 5 = Very well/confident. Students reported significant gains in their understanding and confidence in all categories ($p < 0.001$).

TABLE 3 | Ranked usefulness of STAMP learning activities.

STAMP learning activity	Average score on five-point Likert Scale (*N* = 131)
Hands-on use of the program	4.5
One-on-one discussions with instructional staff	4.3
Tutorials (documents and videos)	3.6
Reading/reading assessment of STAMP user guide or articles	3.0

concepts on average "fairly well" to "quite well," a statistically significant change based on Mann–Whitney nonparametric tests for all measures ($p < 0.001$). Of note, students were generally less confident of their understanding of "the advantages and limitations of various statistical tests (e.g., Do you know when to use a *T*-test over a one-way ANOVA)?" at the end of the term. This result was somewhat to be expected because the statistical analysis tool they used, STAMP, aims to promote best practices by suggesting a statistical hypothesis test based on the input data (Parks and Beiko, 2010). Therefore, students had limited practice with this particular skill.

In addition to performing statistical tests, STAMP generates a variety of data visualization plots, and we wanted to assess how confident students were in their ability to analyze these plots (**Figure 4B**). Mann–Whitney results indicated a statistically significant change in students' self-reported levels of confidence ($p < 0.001$). Specifically, at the start of the term students reported being "fairly" to "quite" confident in their ability to analyze common plots such as scatter plots, bar plots, and histograms. They had much less confidence, however, in their ability to interpret principal component analysis (PCA), heat maps, and extended error bar plots. By the end of the term they were "quite confident" on average in their ability to analyze most of the plots, and had dramatically improved their confidence in PCA, heat map analyses, and extended error bar plots. Another key learning objective of the course was the ability to interpret phylogenetic trees and analyze their statistical support (**Figure 4C**). At the start of the term, students reported being "not very" confident in their ability to assess bootstrap or resampling values, which are an indication of the of statistical confidence in a clade (Efron et al., 1996), and "not very" to "fairly" confident in their ability to interpret topology and evolutionary distances. By the end of the term, students had significantly increased their confidence in their ability to analyze all aspects of phylogenetic trees ($p < 0.001$).

Tutorials

STAMP was an essential component of the curriculum and was central for many of the student data analysis and visualization learning outcomes. We wanted to find out which learning activities the students found to be the most helpful in preparing them to use and interpret data in STAMP. Students reported that tutorials we created were useful, but perhaps unsurprisingly, it was actual use of the program and discussing it with the instructional staff that the students found to be essential (**Table 3**). All tutorials are publicly available at https://sites.google.com/g.ucla.edu/pumaa/home.

Statistical Reasoning and Conceptual Gains Measured by the SRBCI and Open-Ended Survey Responses

We used the SRBCI to directly assess student learning gains in core concepts related to repeatability of results, variations in data, hypotheses and predictions, and sample size. Students took the pre-test in the first week of the term and the post-test at the end of the term following the completion of all of the analysis modules. Scores for the pre-tests and post-tests were binned by number of correct responses and plotted to compare the overall distribution of scores (**Figure 5**). The distribution of the post-test scores is more skewed to the right, demonstrating overall improvement on the SRBCI for the combined cohorts. Statistical reasoning gains between the pre-test and post-test groups were assessed using a Mann–Whitney nonparametric test. There was a statistically significant increase in pre-test (Mean = 58.7%, Mean Rank = 44.3, *N* = 52) to post-test (Mean = 69.3%, Mean Rank = 59.0, *N* = 50) scores ($p = 0.01$) on the SRBCI.

A rubric-guided assessment of an open-ended survey question was used to determine whether the curricular interventions resulted in an increased understanding of the relationship between statistical significance (*p*-value) and biological significance (effect size). At the beginning of the term, 63.9% of students had no familiarity with the concept or held novice understanding, meaning the responses indicated they didn't know, or they had multiple or complete misconceptions (**Figure 6**). By the end of the term, 78.4% of students held intermediate to advanced levels of understanding, and were able to demonstrate conceptual understanding of the relationship to varying degrees. The rubric scores from the Exit survey (Mean = 3.14, Mean Rank = 164.4, *N* = 125) were significantly

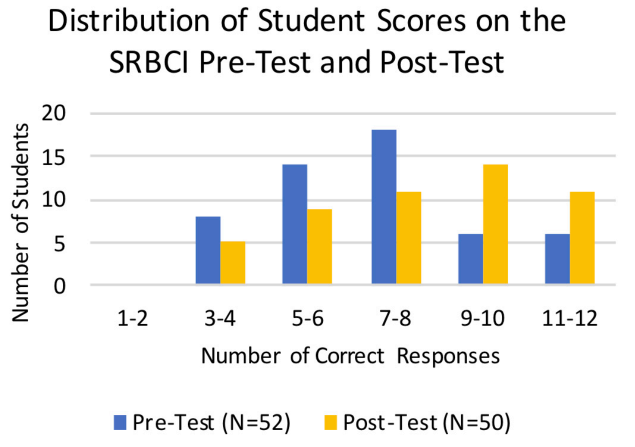


FIGURE 5 | Distribution of Student scores on the SRBCI pre-test and post-test. The number of students plotted on the vertical axis is binned by the number of correct responses shown on the horizontal axis. The blue columns represent the pre-test scores and yellow columns represent the post-test scores. There was a significant increase in the SRBCI scores from the pre-test to the post-test ($p = 0.01$).

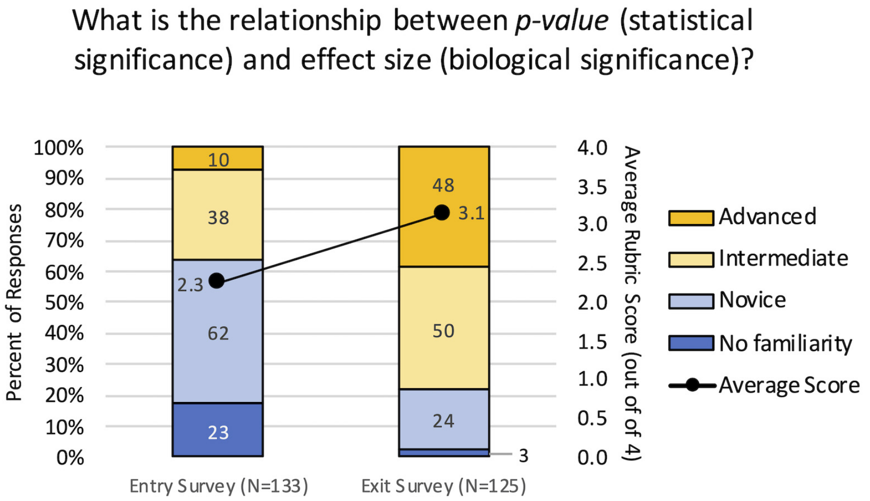


FIGURE 6 | Conceptual gains in understanding the relationship between statistical and biological significance. Student responses to the open-ended question were evaluated using a rubric to assign them a level of competency from 1 = No familiarity, 2 = Novice, 3 = Intermediate, and 4 = Advanced. The primary axis indicates the percent of student responses demonstrating each level of competency for the entry and exit surveys. Blue indicates lower competency levels and yellow indicates higher competency levels. The secondary axis indicates the average score for all responses; there was a significant increase in the average score from the entry surveys to the exit surveys ($p < 0.001$).

higher than the Entry survey (Mean = 2.26, Mean Rank = 96.7, $N = 135$) by both the Mann–Whitney and t-tests ($p < 0.001$). These results demonstrate the shift from lower levels of competency to higher levels of competency in understanding the relationship between statistical and biological significance.

DISCUSSION

The increased availability of microbiome and other "big data" data sets has coincided with calls for life science undergraduates to have bioinformatics "minimum skill sets" or "core competencies" in order to meet the growing demand to analyze that data (Tan et al., 2009; Welch et al., 2016; Mulder et al., 2018; Sayres et al., 2018). PUMAA has been in use in the Research Immersion in Microbiology undergraduate laboratories at UCLA for a number of years, resulting in the development of a suite of instructional materials and tutorials to train students in many of the bioinformatics skills necessary to meet this demand. This curriculum focused on quantitative literacy, which is the intersection of critical thinking, math/statistics, and real-world contexts, and has been highlighted by the Association of American Colleges and Universities as an essential skill for undergraduates (Elrod, 2014). The PUMAA curriculum and associated analysis and visualization tools gave students opportunities to use multiple bioinformatic approaches to analyzing their data. Repeated practice with tools and integration of said tools into student-driven research projects increased self-reported confidence with data visualization and analysis. For example, use of STAMP enabled students to perform statistical tests on microbiome community and functional profiles, and improved their competence with statistical concepts such as statistical significance and biological significance. This was of particular interest due to the tendency of notice researchers to over interpret p-values and disregard the importance of effect sizes and confidence intervals (Nakagawa and Cuthill, 2007; Martínez-Abraín, 2008).

PUMAA presents a user-friendly, time-and-cost-effective approach to processing, analyzing, and visualizing marker gene microbiome data. It improves the accessibility and range of available microbiome investigations by providing users with a simple way to unify the output of various taxonomic identification platforms with a suite of tools for data analysis and visualization. The protocol accomplishes this by producing properly configured, formatted, and annotated files for analysis of taxonomic community profiles and inferred functional profiles. This process of data manipulation can often be performed by sequencing services for additional fees or completed by users with significant time commitment, both of which could be barriers for those with funding or time constraints. PUMAA is an open-source solution which is highly accessible to a wide spectrum of users, including undergraduates or other researchers interested in learning to conduct microbiome analyses, as it can be used as a GUI as well as a CLI. It provides an easy and flexible interface for a variety of users requiring a clear and brief interface for production of files needed for diversity analysis and data visualization for analysis of targeted amplicon sequencing studies. The demand for tools that meet this need is evidenced by the recent development of DNA metabarcoding data processing tools like the web-based SLIM (Dufresne et al., 2019) and minimal coding-required PEMA (Zafeiropoulos et al., 2020). Both of these tools produce OTU and/or ASV tables from raw metabarcode data that could be incorporated into the PUMA input pipeline for downstream data analysis and visualization.

In practice, the instructional staff runs the PUMAA program and provides students with files ready for use in Excel, ranacapa, STAMP, and other tools. One limitation of this approach is that students do not get direct experience with command-line bioinformatics, which is one of the core competencies for undergraduate life sciences education described by several different bioinformatics curriculum committees (Tan et al., 2009; Welch et al., 2016; Mulder et al., 2018; Sayres et al., 2018). However, the International Society for Computational Biology's Curriculum Task Force has refined their core competencies and designated different user profiles requiring different *levels* of competency (Mulder et al., 2018). For example, an undergraduate in a 10-week microbial ecology course may be considered a "bioinformatics user," rather than a "bioinformatics scientist" or "bioinformatics engineer," and the steep learning curve required to gain CLI skills may not be practical with the limited time available. We focused instead on training students to perform all of the bioinformatic analyses needed for an authentic course-based undergraduate research experience in microbial ecology. PUMAA is not intended to replace comprehensive CLI tools such as QIIME or mothur, but rather serve as an entry point for novice researchers to analyze and visualize their datasets. Students that express interest in expanding their bioinformatics skills can be directed to a wealth of tutorials and resources for learning to code.

The PUMAA program and the curriculum described here have the potential to have a wide impact by making marker gene microbiome research accessible to researchers with multiple levels of experience, and with the included instructional module documents, it can be practically implemented in a classroom setting for undergraduates.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the UCLA Institutional Review Board (UCLA IRB). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

KM designed the PUMAA program and wrote the manuscript. JR created instructional materials, designed assessments, collected and analyzed assessment data, and contributed to the manuscript. CD contributed to writing the program, created instructional materials, and contributed to the manuscript. CS collected and analyzed assessment data, and contributed to the manuscript. SM consulted on the program and contributed to the manuscript. AF created instructional materials, designed assessments, and contributed to the manuscript. JMP designed the curriculum, created assessments, conceptualized the PUMAA program, and wrote the manuscript. All authors have reviewed and approved the manuscript.

FUNDING

Curricular development and assessment were supported by a grant from the UCLA Center for the Advancement of Teaching (IIP#15-02). Institutional support for the laboratory curriculum is provided by the Division of Life Sciences in the College of Letters and Science and the Department of Microbiology, Immunology and Molecular Genetics at UCLA.

ACKNOWLEDGMENTS

This manuscript has been released as a pre-print at https://www.biorxiv.org/ (Mitchell et al., 2018). We would like to thank Annabel Beichman for writing early versions of the Python file conversion scripts and Joseph Olivier for contributions to analysis tutorials. We would like to thank Kris Reddi, without whom the MIMG instructional laboratories curriculum wouldn't be possible. We would like to acknowledge the Institute for Quantitative and Computational Biology (QCBio) at UCLA. We would like to acknowledge Nathan Kraft and Gaurav Kandlikar in the Ecology and Evolutionary Biology Department at UCLA, and Savanah St. Clair and Marcie Sakadjian from Pierce College, for collaborating on our research projects. We would like to thank Rachel Meyer, Emily Curd, Teia Schweizer, Miroslava Munguia Ramos, and Ana Garcia Vedrenne from the University of California Conservation Genomics Consortium CALeDNA Project for help with project design, preparation of sample collection kits, data collection and storage, eDNA extraction and sequencing library preparation, and Anacapa sequence data processing. And finally, we would like to thank and acknowledge all of the undergraduate researchers in the Research Immersion in Microbiology curriculum.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.584699/full#supplementary-material

Supplementary Table 1 | PUMMA_Input and Output files.

Supplementary File 1 | PUMAA_Curriculum details.

Supplementary File 2 | PUMAA_Surveys.

Supplementary File 3 | PUMAA_Rubric_Revised.

REFERENCES

Aikens, M. L., and Dolan, E. L. (2014). Teaching quantitative biology: goals, assessments, and resources. *Mol. Biol. Cell* 25, 3478–3481. doi: 10.1091/mbc.E14-06-1045

Almeida, A., Mitchell, A. L., Tarkowska, A., and Finn, R. D. (2018). Benchmarking taxonomic assignments based on 16S rRNA gene profiling of the microbiota from commonly sampled environments. *GigaScience* 7:giy054. doi: 10.1093/gigascience/giy054

Ashcraft, M. H., and Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *J. Psychoeduc. Assess.* 27, 197–205. doi: 10.1177/0734282908330580

Bangera, G., and Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci. Educ.* 13, 602–606. doi: 10.1187/cbe.14-06-0099

Bialek, W., and Botstein, D. (2004). Introductory science and mathematics education for 21st-Century biologists. *Science* 303, 788–790. doi: 10.1126/science.1095480

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857. doi: 10.1038/s41587-019-0209-9

Brewer, C. A., and Smith, D. (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washigntone, DC: Am. Assoc. Adv. Sci.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *ISME J.* 11, 2639–2643. doi: 10.1038/ismej.2017.119

Campbell, A. M., Ledbetter, M. L. S., Hoopes, L. L. M., Eckdahl, T. T., Heyer, L. J., Rosenwald, A., et al. (2007). Genome Consortium for Active Teaching: meeting the goals of BIO2010. *CBE Life Sci. Educ.* 6, 109–118. doi: 10.1187/cbe.06-10-0196

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Carey, M. A., and Papin, J. A. (2018). Ten simple rules for biologists learning to program. *PLoS Comput. Biol.* 14:e1005871. doi: 10.1371/journal.pcbi.1005871

Clooney, A. G., Fouhy, F., Sleator, R. D., O' Driscoll, A., Stanton, C., Cotter, P. D., et al. (2016). Comparing apples and oranges: next generation sequencing and its impact on microbiome analysis. *PLoS One* 11:e0148028. doi: 10.1371/journal.pone.0148028

Corwin, L. A., Graham, M. J., and Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: an agenda for future research and evaluation. *CBE Life Sci. Educ.* 14:es1. doi: 10.1187/cbe.14-10-0167

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., et al. (2019). Anacapa Toolkit: an environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods Ecol. Evol.* 10, 1469–1475. doi: 10.1111/2041-210X.13214

Deane, T., Nomme, K., Jeffery, E., Pollock, C., and Birol, G. (2016). Development of the statistical reasoning in biology concept inventory (SRBCI). *CBE Life Sci. Educ.* 15:ar5. doi: 10.1187/cbe.15-06-0131

Douglas, G. M., Maffei, V. J., Zaneveld, J. R., Yurgel, S. N., Brown, J. R., Taylor, C. M., et al. (2020). PICRUSt2 for prediction of metagenome functions. *Nat. Biotechnol.* 38, 685–688. doi: 10.1038/s41587-020-0548-6

Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J., and Cordier, T. (2019). SLIM: a flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics* 20:88. doi: 10.1186/s12859-019-2663-2

Eagan, M. K., Hurtado, S., Chang, M. J., Garcia, G. A., Herrera, F. A., and Garibay, J. C. (2013). Making a difference in science education the impact of undergraduate research programs. *Am. Educ. Res. J.* 50, 683–713. doi: 10.3102/0002831213482038

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32, 1792–1797. doi: 10.1093/nar/gkh340

Efron, B., Halloran, E., and Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proc. Natl. Acad. Sci. U.S.A.* 93, 13429–13429. doi: 10.1073/pnas.93.23.13429

Elrod, S. (2014). *Quantitative Reasoning: The Next "Across the Curriculum" Movement*. Available online at: https://www.aacu.org/peerreview/2014/summer/elrod (accessed June 25, 2020).

Furnham, A. (1986). Response bias, social desirability and dissimulation. *Personal. Individ. Differ.* 7, 385–400. doi: 10.1016/0191-8869(86)90014-0

Garcia-Milian, R., Hersey, D., Vukmirovic, M., and Duprilot, F. (2018). Data challenges of biomedical researchers in the age of omics. *PeerJ* 6:e5553. doi: 10.7717/peerj.5553

Hanauer, D. I., Graham, M. J., SEA-PHAGES, Betancur, L., Bobrownicki, A., Cresawn, S. G., et al. (2017). An inclusive research education community (iREC): impact of the SEA-PHAGES program on research outcomes and student learning. *Proc. Natl. Acad. Sci. U.S.A.* 114, 13531–13536. doi: 10.1073/pnas.1718188115

Harrison, M., Dunbar, D., Ratmansky, L., Boyd, K., and Lopatto, D. (2011). Classroom-based science research at the introductory level: changes in career choices and attitude. *CBE Life Sci. Educ.* 10, 279–286. doi: 10.1187/cbe.10-12-0151

Iwai, S., Weinmaier, T., Schmidt, B. L., Albertson, D. G., Poloso, N. J., Dabbagh, K., et al. (2016). Piphillin: improved prediction of metagenomic content by direct inference from human microbiomes. *PLoS One* 11:e0166104. doi: 10.1371/journal.pone.0166104

Jovel, J., Patterson, J., Wang, W., Hotte, N., O'Keefe, S., Mitchel, T., et al. (2016). Characterization of the gut microbiome using 16S or shotgun metagenomics. *Front. Microbiol.* 7:459. doi: 10.3389/fmicb.2016.00459

Kandlikar, G. S., Gold, Z. J., Cowen, M. C., Meyer, R. S., Freise, A. C., Kraft, N. J. B., et al. (2018). ranacapa: an R package and Shiny web app to explore environmental DNA data with exploratory statistics and interactive visualizations. *F1000Research* 7:1734. doi: 10.12688/f1000research.16680.1

Kandlikar, G. S., Yan, X., Levine, J. M., and Kraft, N. J. B. (2020). Quantifying microbially mediated fitness differences reveals the tendency for plant-soil feedbacks to drive species exclusion among California annual plants. *bioRxiv* [Preprint]. doi: 10.1101/2020.02.13.948679

Kanehisa, M. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi: 10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32, D277–D280.

Katoh, K., and Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30, 772–780. doi: 10.1093/molbev/mst010

Kawashima, S., Katayama, T., Sato, Y., and Kanehisa, M. (2003). KEGG API: a web service using SOAP/WSDL to access the KEGG system. *Genome Inform.* 14, 673–674.

Kohl, M., Wiese, S., and Warscheid, B. (2011). "Cytoscape: software for visualization and analysis of biological networks," in *Data Mining in Proteomics: From Standards to Applications Methods in Molecular Biology*, eds M. Hamacher, M. Eisenacher, and C. Stephan (Totowa, NJ: Humana Press), 291–303. doi: 10.1007/978-1-60761-987-1_18

Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., and Knight, R. (2011). "Using QIIME to analyze 16S rRNA gene sequences from microbial communities," in *Current Protocols in Bioinformatics*, ed. A. D. Baxevanis (Hoboken, NJ: John Wiley & Sons, Inc).

Langille, M. G. I. (2018). Exploring Linkages between Taxonomic and Functional Profiles of the Human Microbiome. *mSystems* 3:e00163-17. doi: 10.1128/mSystems.00163-17

Langille, M. G. I., Zaneveld, J., Caporaso, J. G., McDonald, D., Knights, D., Reyes, J. A., et al. (2013). Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat. Biotechnol.* 31, 814–821. doi: 10.1038/nbt.2676

Laudadio, I., Fulci, V., Stronati, L., and Carissimi, C. (2019). Next-generation metagenomics: methodological challenges and opportunities. *OMICS J. Integr. Biol.* 23, 327–333. doi: 10.1089/omi.2019.0073

Lopatto, D. (2004). Survey of undergraduate research experiences (SURE): first findings. *Cell Biol. Educ.* 3, 270–277. doi: 10.1187/cbe.04-07-0045

Mangul, S., Martin, L. S., Hoffmann, A., Pellegrini, M., and Eskin, E. (2017). Addressing the digital divide in contemporary biology: lessons from teaching UNIX. *Trends Biotechnol.* 35, 901–903. doi: 10.1016/j.tibtech.2017.06.007

Mangul, S., Mosqueiro, T., Abdill, R. J., Duong, D., Mitchell, K., Sarwal, V., et al. (2019). Challenges and recommendations to improve the installability and archival stability of omics computational tools. *PLoS Biol.* 17:e3000333. doi: 10.1371/journal.pbio.3000333

Martínez-Abraín, A. (2008). Statistical significance and biological relevance: a call for a more cautious interpretation of results in ecology. *Acta Oecol.* 34, 9–11. doi: 10.1016/j.actao.2008.02.004

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* 8:e61217. doi: 10.1371/journal.pone.0061217

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

McMurdie, P. J., and Holmes, S. (2015). Shiny-phyloseq: web application for interactive microbiome analysis with provenance tracking. *Bioinformatics* 31, 282–283. doi: 10.1093/bioinformatics/btu616

Meyer, R. S., Curd, E. E., Schweizer, T., Gold, Z., Ramos, D. R., Shirazi, S., et al. (2019). The California environmental DNA "CALeDNA" program. *bioRxiv* [Preprint]. doi: 10.1101/503383

Mitchell, K., Dao, C., Freise, A., Mangul, S., and Parker, J. M. (2018). PUMA: a tool for processing 16S rRNA taxonomy data for analysis and visualization. *bioRxiv* [Preprint]. doi: 10.1101/482380

mrdnalab (2020). Available online at: http://www.mrdnalab.com/16s-ribosomal-sequencing.html (Accessed June 24, 2020).

Mulder, N., Schwartz, R., Brazas, M. D., Brooksbank, C., Gaeta, B., Morgan, S. L., et al. (2018). The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Comput. Biol.* 14:e1005772. doi: 10.1371/journal.pcbi.1005772

Nakagawa, S., and Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol. Rev.* 82, 591–605. doi: 10.1111/j.1469-185X.2007.00027.x

Narayan, N. R., Weinmaier, T., Laserna-Mendieta, E. J., Claesson, M. J., Shanahan, F., Dabbagh, K., et al. (2020). Piphillin predicts metagenomic composition and dynamics from DADA2-corrected 16S rDNA sequences. *BMC Genomics* 21:56. doi: 10.1186/s12864-019-6427-1

Parks, D. H., and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics* 26, 715–721. doi: 10.1093/bioinformatics/btq041

Parks, D. H., Tyson, G. W., Hugenholtz, P., and Beiko, R. G. (2014). STAMP: statistical analysis of taxonomic and functional profiles. *Bioinformatics* 30, 3123–3124. doi: 10.1093/bioinformatics/btu494

Parks, S., Joyner, J. L., and Nusnbaum, M. (2020). Reaching a large urban undergraduate population through microbial ecology course-based research experiences. *J. Microbiol. Biol. Educ.* 21:21.1.17. doi: 10.1128/jmbe.v21i1.2047

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010). FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5:e9490. doi: 10.1371/journal.pone.0009490

Rideout, J. R., Chase, J. H., Bolyen, E., Ackermann, G., González, A., Knight, R., et al. (2016). Keemei: cloud-based validation of tabular bioinformatics file formats in google sheets. *GigaScience* 5:27. doi: 10.1186/s13742-016-0133-6

Rosenwald, A. G., Arora, G. S., Madupu, R., Roecklein-Canfield, J., and Russell, J. S. (2012). The human microbiome project: an opportunity to engage undergraduates in research. *Proc. Comput. Sci.* 9, 540–549. doi: 10.1016/j.procs.2012.04.058

Russell, S. H., Hancock, M. P., and McCullough, J. (2007). Benefits of undergraduate research experiences. *Science* 316, 548–549. doi: 10.1126/science.1140384

Sanders, E. R., and Hirsch, A. M. (2014). Immersing undergraduate students into research on the metagenomics of the plant rhizosphere: a pedagogical strategy to engage civic-mindedness and retain undergraduates in STEM. *Front. Plant Sci.* 5:157. doi: 10.3389/fpls.2014.00157

Sanders, E. R., and Miller, J. H. (2010). *I, Microbiologist: A Discovery-Based Course in Microbial Ecology and Molecular Evolution*. Washington, DC: ASM Press.

Sayres, M. A. W., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS One* 13:e0196878. doi: 10.1371/journal.pone.0196878

Sewall, J. M., Oliver, A., Denaro, K., Chase, A. B., Weihe, C., Lay, M., et al. (2020). Fiber force: a fiber diet intervention in an advanced course-based undergraduate research experience (CURE) course. *J. Microbiol. Biol. Educ.* 21:21.1.40. doi: 10.1128/jmbe.v21i1.1991

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 13, 2498–2504. doi: 10.1101/gr.1239303

Shapiro, C., Moberg-Parker, J., Toma, S., Ayon, C., Zimmerman, H., Roth-Johnson, E. A., et al. (2015). Comparing the impact of course-based and apprentice-based research experiences in a life science laboratory curriculum. *J. Microbiol. Biol. Educ.* 16, 186–197. doi: 10.1128/jmbe.v16i2.1045

Skirball Fire (2020). *Wikipedia*. Available online at: https://en.wikipedia.org/w/index.php?title =Skirball_Fire&oldid =948253595(Accessed June 27, 2020).

St. Clair, S., Saraylou, M., Melendez, D., Senn, N., Reitz, S., Kananipour, D., et al. (2020). Analysis of the soil microbiome of a Los Angeles urban farm. *Appl. Environ. Soil Sci.* 2020:e5738237. doi: 10.1155/2020/5738237

Tan, T. W., Lim, S. J., Khan, A. M., and Ranganathan, S. (2009). A proposed minimum skill set for university graduates to meet the informatics needs and challenges of the "-omics" era. *BMC Genomics* 10:S36. doi: 10.1186/1471-2164-10-S3-S36

Wang, J. T. H., Daly, J. N., Willner, D. L., Patil, J., Hall, R. A., Schembri, M. A., et al. (2015). Do you kiss your mother with that mouth? An authentic large-scale undergraduate research experience in mapping the human oral microbiome†. *J. Microbiol. Biol. Educ.* 16, 50–60. doi: 10.1128/jmbe.v16i1.816

Weber, K. S., Bridgewater, L. C., Jensen, J. L., Breakwell, D. P., Nielsen, B. L., and Johnson, S. M. (2018). Personal microbiome analysis improves student engagement and interest in Immunology, Molecular Biology, and Genomics undergraduate courses. *PLoS One* 13:e0193696. doi: 10.1371/journal.pone.0193696

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Welch, L., Brooksbank, C., Schwartz, R., Morgan, S. L., Gaeta, B., Kilpatrick, A. M., et al. (2016). Applying, evaluating and refining bioinformatics core competencies (an update from the curriculum task force of ISCB's education committee). *PLoS Comput. Biol.* 12:e1004943. doi: 10.1371/journal.pcbi.1004943

Willis, A. D. (2019). Rarefaction, alpha diversity, and statistics. *Front. Microbiol.* 10:2407. doi: 10.3389/fmicb.2019.02407

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., et al. (2020). PEMA: a flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience* 9:giaa022. doi: 10.1093/gigascience/giaa022

Zhang, H. (2016). Overview of sequence data formats. *Methods Mol. Biol.* 1418, 3–17. doi: 10.1007/978-1-4939-3578-9_1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 14 October 2020
doi: 10.3389/fmicb.2020.578099

Bioinformatics-Based Activities in High School: Fostering Students' Literacy, Interest, and Attitudes on Gene Regulation, Genomics, and Evolution

Ana Martins[1,2]*, *Maria João Fonseca*[3], *Marina Lemos*[4], *Leonor Lencastre*[4] and *Fernando Tavares*[1,2]*

[1] Departamento de Biologia, FCUP-Faculdade de Ciências, Universidade do Porto, Porto, Portugal, [2] CIBIO-Centro de Investigação em Biodiversidade e Recursos Genéticos, InBIO-Laboratório Associado, Universidade do Porto, Vairão, Portugal, [3] MHNC-UP-Museu de História Natural e da Ciência, Universidade do Porto, Porto, Portugal, [4] FPCEUP-Faculdade de Psicologia e Ciências da Educação, Universidade do Porto, Porto, Portugal

OPEN ACCESS

Edited by:
Ludmila Chistoserdova,
University of Washington,
United States

Reviewed by:
Yi Xiong,
Shanghai Jiao Tong University, China
Liang Leon Wang,
Xuzhou Medical University, China

***Correspondence:**
Ana Martins
asmartins@cibio.up.pt
Fernando Tavares
ftavares@fc.up.pt

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: 30 June 2020
Accepted: 08 September 2020
Published: 14 October 2020

Citation:
Martins A, Fonseca MJ, Lemos M, Lencastre L and Tavares F (2020) Bioinformatics-Based Activities in High School: Fostering Students' Literacy, Interest, and Attitudes on Gene Regulation, Genomics, and Evolution. Front. Microbiol. 11:578099. doi: 10.3389/fmicb.2020.578099

The key role of bioinformatics in explaining biological phenomena calls for the need to rethink didactic approaches at high school aligned with a new scientific reality. Despite several initiatives to introduce bioinformatics in the classroom, there is still a lack of knowledge on their impact on students' learning gains, engagement, and motivation. In this study, we detail the effects of four bioinformatics laboratories tailored for high school biology classes named "Mining the Genome: Using Bioinformatics Tools in the Classroom to Support Student Discovery of Genes" on literacy, interest, and attitudes on 387 high school students. By exploring these laboratories, students get acquainted with bioinformatics and acknowledge that many bioinformatics tools can be intuitive for beginners. Furthermore, introducing comparative genomics in their learning practices contributed for a better understanding of curricular contents regarding the identification of genes, their regulation, and how to make evolutionary assumptions. Following the intervention, students were able to pinpoint bioinformatics tools required to identify genes in a genomics sequence, and most importantly, they were able to solve genomics-related misconceptions. Overall, students revealed a positive attitude regarding the integration of bioinformatics-based approaches in their learning practices, reinforcing their added value in educational approaches.

Keywords: bioinformatics, comparative genomics, gene regulation, high school, genomic literacy

INTRODUCTION

Bioinformatics, understood as the use of computational resources to categorize massive raw data and retrieve meaningful information from datasets, has gained a primordial utility in scientists' daily routine (Sadek, 2004). This paradigm of biological research cannot be disregarded when seeking to promote a scientifically informed society. Indeed, it demands the improvement of curricular and educational resources at middle and high school educational levels based on initiatives validated by focused science education research.

Learning by accessing online bioinformatics resources in the classroom has already proven to have a beneficial impact on students' ability to build up and mobilize scientific contents, namely, related to drug resistance, phylogenetic trees, or genetic expression (Amenkhienan and Smith, 2006; Taylor et al., 2014; Newman et al., 2016; Machluf et al., 2017). In addition, the introduction of bioinformatics at high school enhances the learning of new information through novel technologies and recruits resources used in research laboratories, serving as a stimulus to spark students' future interest in scientific careers (Kovarik et al., 2013; Machluf et al., 2017).

Despite the various initiatives across Europe to support teachers and students to integrate bioinformatics-based approaches in their classes, these remain sporadic and are still not implemented consistently. Recent studies have called attention to the importance of a joint effort by all stakeholders (e.g., research institutions, governmental entities, teachers, trainers, and researchers) to deliver an action plan that can lead to bioinformatics dissemination in schools in a wider, more structured and cohesive manner (Koch and Fuellen, 2008; Campbell and Nehm, 2013; Attwood et al., 2017). Recent reports call for more educational assessments to strengthen the positive impact of bioinformatics-based activities on students' scientific and digital literacy, providing a rationale to incorporate bioinformatics in the curriculum (Dudley and Butte, 2009; Campbell and Nehm, 2013; Machluf and Yarden, 2013; Magana et al., 2014; Marques et al., 2014; Machluf et al., 2017).

This study aims to address the educational impact on high school students of a set of activities developed to introduce basic bioinformatics analysis used to deconstruct a bacterial genomic sequence into its coding genes (Martins et al., 2018a), using purposely tailored evaluation instruments. The main research question driving this investigation was: *are there significant changes in high school students' scientific and digital literacy, interest, and attitudes toward gene regulation, genomics, and evolution after performing bioinformatics-based activities?*

MATERIALS AND METHODS

Participants

The sample studied included a group of 387 students and 11 teachers from five public and private schools in Porto and Lisboa, Portugal. Fourteen 11th grade biology and geology classes (students' age: 16–17 years old) and five 12th grade biology classes (students' age: 17–18 years old), comprising 167 male and 220 female students, were involved in this study. Students' average age was 16.34 ± 0.67 years. The study included an experimental group (n = 292) with 123 male students and 169 female students (average age: 16.27 ± 0.68 years) from 14 classes and a control group (n = 95) including 44 male students and 51 female students (average age: 16.54 ± 0.62 years) from five classes.

Students participated in the project as part of their science classes, and taking into account all ethical requirements, the project was institutionally approved by each school's Directive Board. Upon entering the project, the participants were invited to take part in the study and informed of its nature and aims, being assured that all the data collected were to be processed and analyzed anonymously. Students were given the chance to participate in the project without participating in this specific study.

Didactic Instrument: Bioinformatics Laboratories

A set of bioinformatics-based activities previously proposed by Martins et al. (2018a) to identify genes from a bacterial genomic sequence and disclose their genomic context in different species was chosen as the didactic instrument. A tutorial video[1] provides teachers and students with a detailed road map of the sequential bioinformatics resources needed to deconstruct a 2 kb genomic region of *Escherichia coli* and determine its occurrence across different bacteria taxa and hypothesize about its evolution. Participants were initially instructed to select a particular *E. coli* strain (*E. coli* str. K-12 substr. MG1655, Accession number: NC_000913.3) and a specific 2 kb genomic region, to ensure that all of them would be working with the same genomic sequence, allowing for a more efficient teacher supervision and facilitating subsequent analysis. In fact, the 2 kb sequence proposed includes the *lac* operon, which is the paradigm used to introduce gene expression and regulation at the high school. This provides a meaningful curricular framing for these activities and is aligned with students' previous knowledge. Furthermore, it is important to emphasize that implementing bioinformatics exercises framed within the curriculum was the main concern of the participant teachers. Currently, the Portuguese biology curricula for the 11th and 12th grades include contents related with DNA and protein synthesis (for example, transcription, translation, and start and stop codons), as well as evolution (Mendes et al., 2003), and genetic expression (Mendes et al., 2004). These topics are also comprised in the Next Generation Science Standards (NGSS; National Research Council, 2013). While these curricular topics are frequently focused on eukaryotic models, bacterial genomes were chosen as an educational instrument for this study having in mind that bacteria stand for the most represented domain in genome databases, reflecting its high taxonomic diversity, and may be easily recruited by ingenious bioinformatics platforms with graphical and user-friendly interfaces using a Windows or Mac browser. In addition, bacterial genomes are frequently restricted to a single replicon (i.e., the chromosome), besides having a small-sized haploid genome that favors comparative genomics and contributes to strengthen students' knowledge on bacteria, fostering their motivation and interest on microbiology-related topics, presently poorly explored in high school.

The bioinformatics resources used include the genome database, Open Reading Frames Finder (ORFfinder) and Basic Local Alignment Search Tool (BLAST) from the National Center for Biotechnology Information (NCBI; Altschul et al., 1990;

[1] https://drive.google.com/file/u/1/d/1WrtjzLHzKI7nLALtVnmqy6WhPnkUsQSR/view

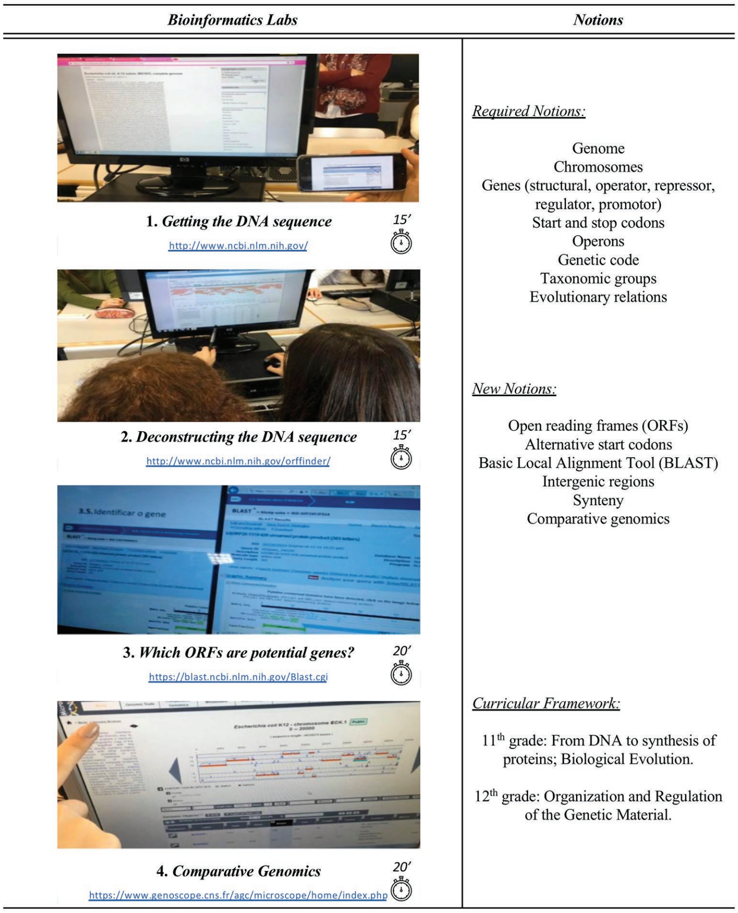
Agarwala et al., 2018), and the genome browser of Magnifying Genomes (MaGe) that is part of MicroScope, a comparative genomics platform (Vallenet et al., 2013). Before starting with the *in silico* laboratories, the teachers work through basic and already known concepts, such as genome, genes, start codons, stop codons, and operons with the class, and introduce new notions, such as Open Reading Frames (ORFs), synteny, and comparative genomics (**Figure 1**; Martins and Tavares, 2018). This is particularly important since these new notions, presently absent from the curricula, are instrumental to understand the data retrieved by the students when performing the bioinformatics exercises proposed (Martins and Tavares, 2018).

Research Design and Methodology

To implement bioinformatics-based activities as a successful didactic instrument, it is crucial to engage both teachers and students in the selection of the activities to ensure that these are meaningful and adjusted to the curricular contents (Marques et al., 2014). In this regard, the design of the bioinformatics-based activities proposed by Martins et al. (2018a) took into account teachers' contribution in revising and piloting the proposed educational resources with their students (**Figure 2**). To lighten the burden for teachers, a dedicated webpage[2] was developed to provide them with resources that introduced the bioinformatics tools and the new concepts to be addressed.

The workflow of the bioinformatics activities includes four parts (**Figure 2**). Firstly, teachers provide the knowledge background about gene regulation, genomics, and evolution. Secondly, students are introduced to the bioinformatics databases and tools to be used, namely, NCBI database, NCBI ORFfinder, NCBI Blast, and MicroScope (MaGe). And thirdly, the bioinformatics exercises are performed. These exercises were set to meet the curricular requirements for the topic, and given the novelty of bioinformatics for these students (and teachers), guidelines were prepared to provide a comprehensible workflow to address the research questions outlined. This allowed to prevent students from becoming overwhelmed by the wide plethora of choices of links and commands available

[2] https://bioinformaticaaula.wixsite.com/bioinformatica-pt



Bioinformatics Labs

1. *Getting the DNA sequence* 15′
http://www.ncbi.nlm.nih.gov/

2. *Deconstructing the DNA sequence* 15′
http://www.ncbi.nlm.nih.gov/orffinder/

3. *Which ORFs are potential genes?* 20′
https://blast.ncbi.nlm.nih.gov/Blast.cgi

4. *Comparative Genomics* 20′
https://www.genoscope.cns.fr/agc/microscope/home/index.php

Notions

Required Notions:

Genome
Chromosomes
Genes (structural, operator, repressor, regulator, promotor)
Start and stop codons
Operons
Genetic code
Taxonomic groups
Evolutionary relations

New Notions:

Open reading frames (ORFs)
Alternative start codons
Basic Local Alignment Tool (BLAST)
Intergenic regions
Synteny
Comparative genomics

Curricular Framework:

11th grade: From DNA to synthesis of proteins; Biological Evolution.

12th grade: Organization and Regulation of the Genetic Material.

FIGURE 1 | Bioinformatics laboratories framed within the curricular biology contents for high school to reinforce genomics topics currently required and to introduce new core concepts.

in the platforms mentioned before. In the fourth and final stage of implementation, the results obtained in each exercise were discussed with the students, and conclusions were drawn.

During the implementation of the activities, a member of the research team (Martins) was present to identify misconceptions and reasoning difficulties, as well as to check the participants' engagement and interaction, and to carry out qualitative observations useful to improve the robustness of the interpretations made.

A quasi-experimental pre-/post-design, with a control and an experimental group, was set up. The control group included classes exclusively exposed to the first two parts of the intervention, i.e., the introductory lectures about the scientific questions and the bioinformatics databases and resources (**Figure 2** – workflow I and II). In turn, the experimental group was exposed to the full set of the bioinformatics activities, i.e., from the introductory lectures to the bioinformatics laboratories and the interpretation of the results (**Figure 2** – workflow I–IV). To mitigate possible bias effects, the control group classes were from the same schools, from the same education levels, and taught by the same teachers as the experimental group classes. The comparison between the performance of students in the control group and the experimental group was intended to test the educational impact of the practical bioinformatics-based activities. In this regard, the control group was taught only through expositive teaching (**Figure 2** – workflow I and II), and the experimental group was exposed to the same lectures as the control group plus the practical component (**Figure 2** – workflow I–IV).

The Questionnaire

To assess the educational impact of integrating the mentioned bioinformatics-based activities in high school, a specific and comprehensive questionnaire including open-ended questions, dichotomous questions, and Likert-type scales was designed (**Figure 3**).

The questionnaire was structured according to three main dimensions: knowledge, interest, and attitudes. The knowledge-related questions (Q1, Q2, Q4, Q5, Q6, Q7, and Q8.5) aimed to characterize students' literacy regarding gene regulation, comparative genomics, bioinformatics, and its usefulness for scientific research. Students' interest (Q3 and Q9) was measured by their perception of the role of bioinformatics in tackling different biology research questions and by their awareness about the scientific disciplines addressed in the *in silico* activities, namely, genetics, genomics, and evolution. Students presently attending high school are part of the so called *iGeneration* (iGen), which is characterized by being highly motivated to use technology in their daily lives (Rosen et al., 2010; Quinn and Oldmeadow, 2013). Having this in mind, a question (Q8) was added to depict students' attitudes toward the use of computer/technological devices to study and to assess their motivation to access bioinformatics tools inside or outside the classroom.

The questionnaire developed was piloted in two high school classes (*n* = 43 students; **Figure 2**), which, as recommended by several authors (Treece and Treece, 1982; Connelly, 2008; Johanson and Brooks, 2010), represent slightly over 10% of the universe of students included in the main research study. This procedure allowed to ensure that the students' responses were not biased by a lack of comprehension of the questionnaire

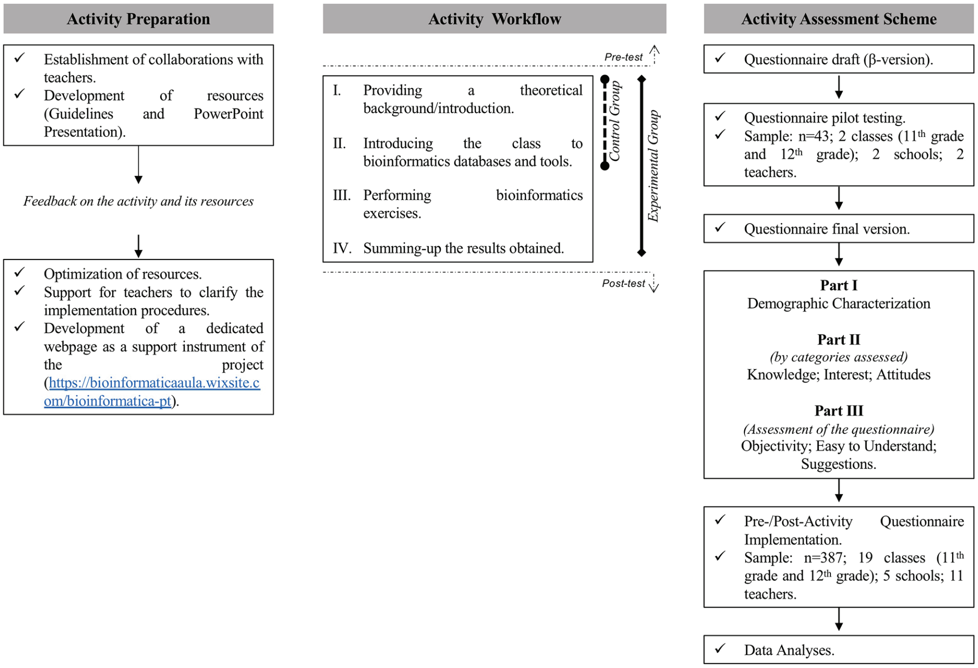
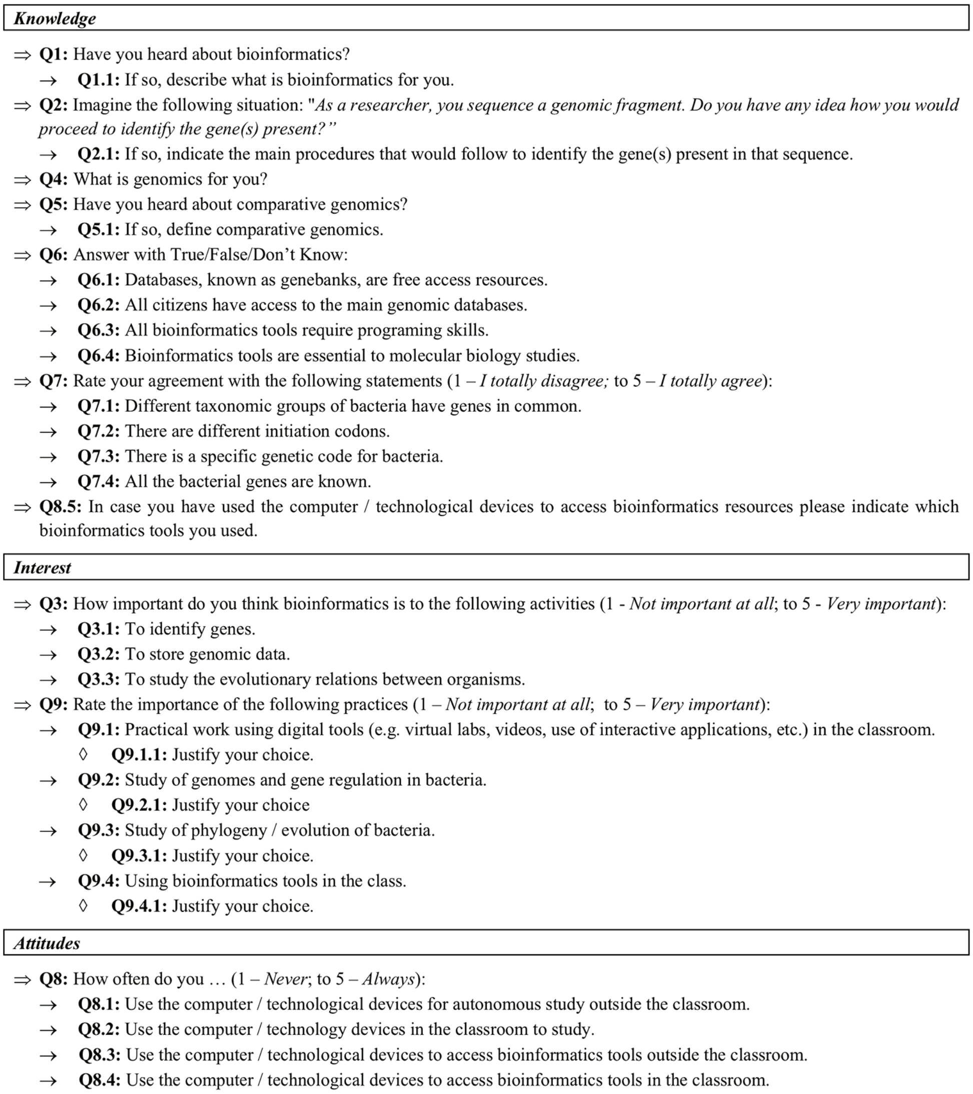

Activity Preparation
✓ Establishment of collaborations with teachers.
✓ Development of resources (Guidelines and PowerPoint Presentation).
Feedback on the activity and its resources
✓ Optimization of resources.
✓ Support for teachers to clarify the implementation procedures.
✓ Development of a dedicated webpage as a support instrument of the project (https://bioinformaticaaula.wixsite.com/bioinformatica-pt).
Activity Workflow
Pre-test
I. Providing a theoretical background/introduction.
II. Introducing the class to bioinformatics databases and tools.
III. Performing bioinformatics exercises.
IV. Summing-up the results obtained.
Control Group
Experimental Group
Post-test
Activity Assessment Scheme
✓ Questionnaire draft (β-version).
✓ Questionnaire pilot testing.
✓ Sample: n=43; 2 classes (11th grade and 12th grade); 2 schools; 2 teachers.
✓ Questionnaire final version.
Part I
Demographic Characterization
Part II
(by categories assessed)
Knowledge; Interest; Attitudes
Part III
(Assessment of the questionnaire)
Objectivity; Easy to Understand; Suggestions.
✓ Pre-/Post-Activity Questionnaire Implementation.
✓ Sample: n=387; 19 classes (11th grade and 12th grade); 5 schools; 11 teachers.
✓ Data Analyses.

FIGURE 2 | Experimental design for preparation, implementation, and assessment of bioinformatics-based activities.



Knowledge

⇒ **Q1:** Have you heard about bioinformatics?

→ **Q1.1:** If so, describe what is bioinformatics for you.

⇒ **Q2:** Imagine the following situation: "*As a researcher, you sequence a genomic fragment. Do you have any idea how you would proceed to identify the gene(s) present?"*

→ **Q2.1:** If so, indicate the main procedures that would follow to identify the gene(s) present in that sequence.

⇒ **Q4:** What is genomics for you?

⇒ **Q5:** Have you heard about comparative genomics?

→ **Q5.1:** If so, define comparative genomics.

⇒ **Q6:** Answer with True/False/Don't Know:

→ **Q6.1:** Databases, known as genebanks, are free access resources.

→ **Q6.2:** All citizens have access to the main genomic databases.

→ **Q6.3:** All bioinformatics tools require programing skills.

→ **Q6.4:** Bioinformatics tools are essential to molecular biology studies.

⇒ **Q7:** Rate your agreement with the following statements (1 – *I totally disagree;* to 5 – *I totally agree*):

→ **Q7.1:** Different taxonomic groups of bacteria have genes in common.

→ **Q7.2:** There are different initiation codons.

→ **Q7.3:** There is a specific genetic code for bacteria.

→ **Q7.4:** All the bacterial genes are known.

⇒ **Q8.5:** In case you have used the computer / technological devices to access bioinformatics resources please indicate which bioinformatics tools you used.

Interest

⇒ **Q3:** How important do you think bioinformatics is to the following activities (1 - *Not important at all*; to 5 - *Very important*):

→ **Q3.1:** To identify genes.

→ **Q3.2:** To store genomic data.

→ **Q3.3:** To study the evolutionary relations between organisms.

⇒ **Q9:** Rate the importance of the following practices (1 – *Not important at all*; to 5 – *Very important*):

→ **Q9.1:** Practical work using digital tools (e.g. virtual labs, videos, use of interactive applications, etc.) in the classroom.

◊ **Q9.1.1:** Justify your choice.

→ **Q9.2:** Study of genomes and gene regulation in bacteria.

◊ **Q9.2.1:** Justify your choice

→ **Q9.3:** Study of phylogeny / evolution of bacteria.

◊ **Q9.3.1:** Justify your choice.

→ **Q9.4:** Using bioinformatics tools in the class.

◊ **Q9.4.1:** Justify your choice.

Attitudes

⇒ **Q8:** How often do you … (1 – *Never*; to 5 – *Always*):

→ **Q8.1:** Use the computer / technological devices for autonomous study outside the classroom.

→ **Q8.2:** Use the computer / technology devices in the classroom to study.

→ **Q8.3:** Use the computer / technological devices to access bioinformatics tools outside the classroom.

→ **Q8.4:** Use the computer / technological devices to access bioinformatics tools in the classroom.

FIGURE 3 | The questionnaire used in this study included demographic characterization of the participants and items to assess students' knowledge, interest, and attitudes.

and also to prevent difficulties in deconstructing the answers to open-ended questions during the content analysis. Furthermore, it is important to highlight that in the final version of the measurement instrument, students were invited to rate the questionnaire regarding its objectivity and intelligibility, to guarantee that the questions were clear and well understood by all respondents.

Lastly, students from both the control group and the experimental group rated the questionnaire as being objective and easy to understand, which further emphasizes the adequacy of the validated version of the questionnaire.

Data Analyses

Methods of descriptive and inferential statistics were used to analyze the pre-/post-test data. All statistical analyses were carried out using IBM's Statistical Package for the Social Sciences (SPSS) version 24.

Independent samples *t*-tests and paired samples *t*-tests for a 95% confidence interval were used for five-point Likert-type scale data, and the effect size of mean differences registered with *t*-test was measured using Cohen's *d* (Cohen, 1988). Data gathered through open-ended questions and dichotomous variables were analyzed using Chi-square and the McNemar tests, respectively, and considering the phi coefficient as the effect size measure (Pallant, 2007). Furthermore, to obtain a broader, more inclusive depiction of the effectiveness of the activities, while strengthening the interpretation of the outcomes of the analyses performed (Punch, 2009), it was decided to combine quantitative and qualitative methods of analysis, as has been suggested in similar studies (Gelbart et al., 2009; Fonseca et al., 2012; Machluf and Yarden, 2013). This methodology would avoid missing detailed information that cannot be retrieved exclusively from quantitative data (Johnson and Christensen, 2012).

In what concerns the qualitative data, a thematic content analysis of the participants' responses to open-ended questions was performed with the purpose of producing a systematic description of the meaning of specific information gathered through the definition of coding categories (Schreier, 2012). This allowed to organize extensive answers to open-ended questions into fewer and more focused content categories (Weber, 1990; Krippendorff, 2004; Hsieh and Shannon, 2005). The analysis of the answers to the open-ended questions was performed according to the framework previously created by the authors in which specific categories of answers have been defined (**Supplementary Figure 1**). Regarding the open-ended question Q9, aimed to assess students' interest, the subjective task value of Eccles and Wigfield (2002), Eccles (2005) that characterizes an expectancy–value model of achievement motivation was used as the theoretical framework underlying data analysis. Task value is related with the quality of the task, which influences the probability of it being select by an individual. In this study, the intrinsic/interest value (i.e., expected enjoyment of engaging in the task), the utility value (i.e., possible rewards from the task), and the cost of engaging in the activities were the dimensions considered when analyzing the students' answers.

RESULTS AND DISCUSSION

Students' Literacy on Bioinformatics and Its Applications

It is consensual that an updated and edifying high school level education requires an attentive revision of the curricula aligned with the challenges of NGSS and capable to meet Science, Technology, Engineering and Mathematics (STEM) education (Wefer and Sheppard, 2008; Kovarik et al., 2013; National Research Council, 2013; Champagne Queloz et al., 2017). In this regard, bioinformatics is in a privileged position, due to the transdisciplinary approach it entails, by seeking a level of integration of different disciplines, such as biology, computer science, and mathematics, beyond the mere interdisciplinary relationship between them. It is therefore reasonable to acknowledge the importance of integrating bioinformatics in high school, as emphasized in several studies (Dudley and Butte, 2009; Machluf and Yarden, 2013; Magana et al., 2014; Marques et al., 2014; Machluf et al., 2017), even though there is scarce research on how to do it (Campbell and Nehm, 2013; Magana et al., 2014; Machluf et al., 2017). To measure the impact of educational initiatives using bioinformatics resources on high school students and to emend misconceptions and tailor adequate bioinformatics activities for successful learning, it is important to diagnose the knowledge students perceive to have about bioinformatics- and genomics-related concepts (Gelbart and Yarden, 2006; Gelbart et al., 2009; Form and Lewitter, 2011; Champagne Queloz et al., 2017).

In the universe of 387 high school students enquired in the present study, only a modest percentage (40.1% of the experimental group, 24.2% of the control group) revealed to have heard about bioinformatics in the pre-test (Q1), and most of the ones who did so could not define bioinformatics, admitting that their answer reflected the etymological meaning of the word. Following an expositive teaching session on bioinformatics and associated resources, such as databases and applications, in the post-test, the percentage of the students who revealed to have heard about bioinformatics raised consistently for both the experimental group (99.0%) and the control group (99.0%; **Figure 4**). Regardless of the fact that in the post-test most of these students linked bioinformatics to the etymology of the word: bio + informatics (60.9% of the experimental group, 73.6% of the control group), which undermines a truly sensible diagnostic of their understanding of bioinformatics, some students did mention specific aspects, such as data analysis, storage, and comparative genomics. The difference observed in this regard between the experimental and control groups (31.0% of the experimental group, 22.0% of the control group) may be explained by the fact that students in the experimental group carried out a set of bioinformatics exercises using the mentioned resources and used bioinformatics platform for comparative genomics, contrarily to their counterparts in the control group. This is particularly evident regarding comparative genomics, a completely new notion for the majority of the students, which was mentioned by 6.6% of the experimental group students and only by 1.1% of the control group students. Furthermore, students from both groups recognized that genebanks are open-access resources (Q6.1; 81.2% of the experimental group, 59.0% of the control group) and generally accessible to all citizens (Q6.2; 78.8% of the experimental group, 62.1% of the control group), suggesting an enhanced perception of what comprises a bioinformatics scientific toolbox and of their empowerment to access it (**Figure 4**). These findings, observed in other studies (Kovarik et al., 2013; Machluf et al., 2017), report for a motivational trigger of scientific literacy and STEM education. The higher percentage scores obtained with the experimental group indicate that complementing expositive teaching with hands-on *in silico* laboratories favors the acquisition of structural knowledge. This was a particularly relevant outcome that allows to dismiss the common misconception that bioinformatics analysis always requires programing skills. In fact, while initially, i.e., before the intervention, students from both groups (62.9% of the experimental group, 69.5% of the control group) agreed that programing skills would be needed to use bioinformatics tools (Q6.3; **Figure 4**), after the intervention, only 27.1% of the students from the experimental group and 32.6% of the control group agreed with this statement (**Figure 4**). These data indicate that while initially students associated bioinformatics analysis to a set of complex computer codes, after they were challenged with bioinformatics activities, they were able to acknowledge the panoply of bioinformatics applications with user-friendly interfaces tailored for web browsers that do not require programming competencies as has been highlighted

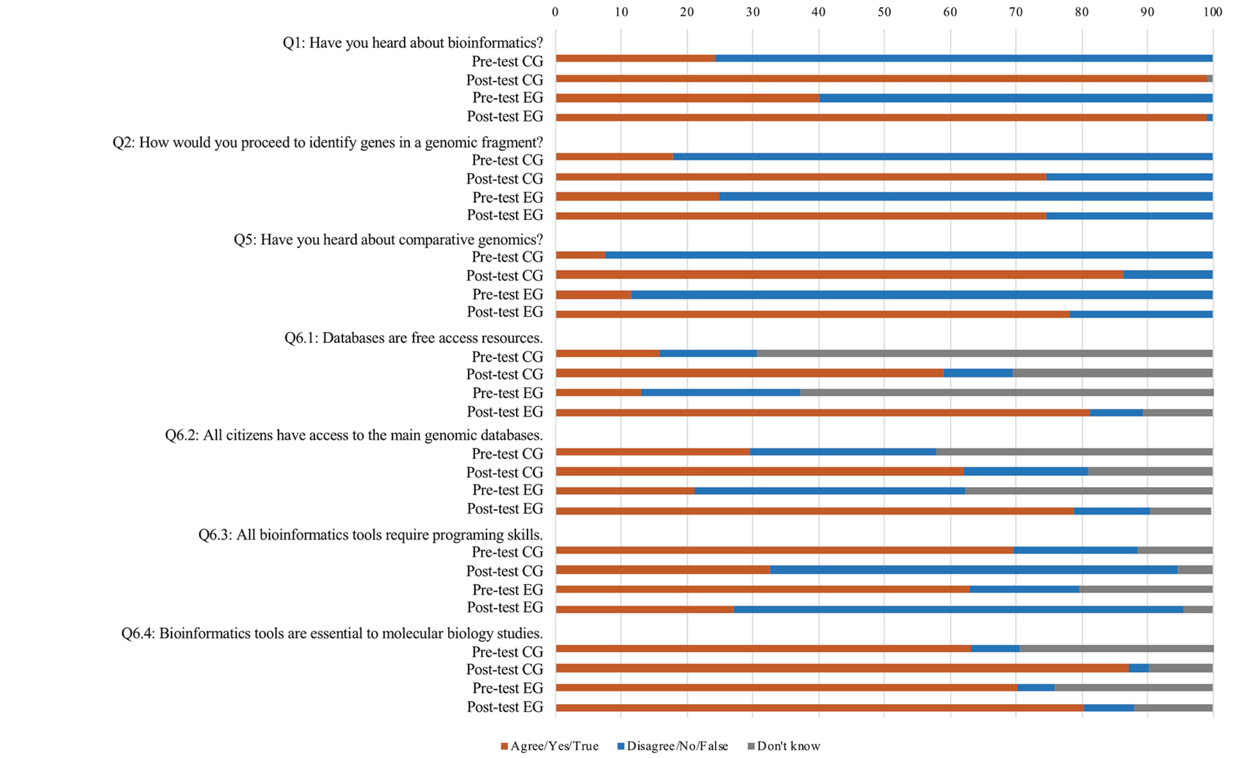


FIGURE 4 | Students' knowledge toward bioinformatics, gene regulation, and genomics.

by Martins et al. (2018b). Students were shown to be aware that bioinformatics tools are essential to molecular biology studies (Q6.4), in both the pre- and post-test (**Figure 4**). Still, in the post-test, there was a slight increment in the percentage of students who agree with this statement, suggesting that they confirmed their previous idea about the role of bioinformatics in molecular biology.

Following the intervention (i.e., post-test), when the participants were asked to "Indicate which bioinformatics platforms [they] used" (Q8.5), 16.7% of the students in the control group failed to mention any of the expected resources used during the intervention, namely, NCBI, NCBI ORFfinder, NCBI BLAST, and MaGe. This percentage dropped to 1.7% in the experimental group (**Figure 5**), indicating the positive impact of bioinformatics laboratories on students' knowledge.

The bioinformatics exercises used in this study aimed to train the students on key procedures to identify genes from a genome sequence, as proposed by Martins et al. (2018a). Since the bioinformatics exercises were supported by a tutorial video comprising detailed guidelines and instructions,[3] it was important to determine if the students' performance actually contributed to enhance their knowledge on basics genome mining and did not resume to a mere mechanical procedure of following a recipe step by step. To address this question, the students were asked to describe the procedures that can be used to identify putative genes within a genomic DNA sequence (Q2, Q2.1). While in the pre-test, only a minority of the students in both groups (24.9% of the experimental group, 17.9% of the control group) claimed to know the procedures to deconstruct a DNA sequence into putative coding sequences, in the post-test, this percentage increased significantly (74.6% of the experimental group, 74.7% of the control group; **Figure 4**). As expected, the change between pre- and post-test is statistically significant for both groups (**Supplementary Table 1**). To fully elucidate if the students' perceptions were aligned with their knowledge, a content analysis was carried out.

In this regard, a framework with three expected bioinformatics-related notions was defined: (1) "Getting the target DNA sequence in a database," (2) "Looking for Open Reading Frames," and (3) "Deciding which of the retrieved ORFs are likely to be genes running a BLAST."

The pre-test content analysis regarding the answers to Q2.1 showed that students who admitted knowing how to identify putative genes from a genomic DNA sequence failed to mention any of the three notions. Instead, they mentioned, for instance, that "To unveil a DNA sequence we can perform an electrophoresis to determine the genes, looking at the gel bands in comparison to a known gene. Restriction enzymes may be needed in this procedure," which was the most frequently recorded notion in the experimental group, and that it is possible to "Use the genetic code to identify the codons in a

[3] https://drive.google.com/file/u/1/d/1WrtjzLHzKI7nLALtVnmqy6WhPnkUsQSR/view

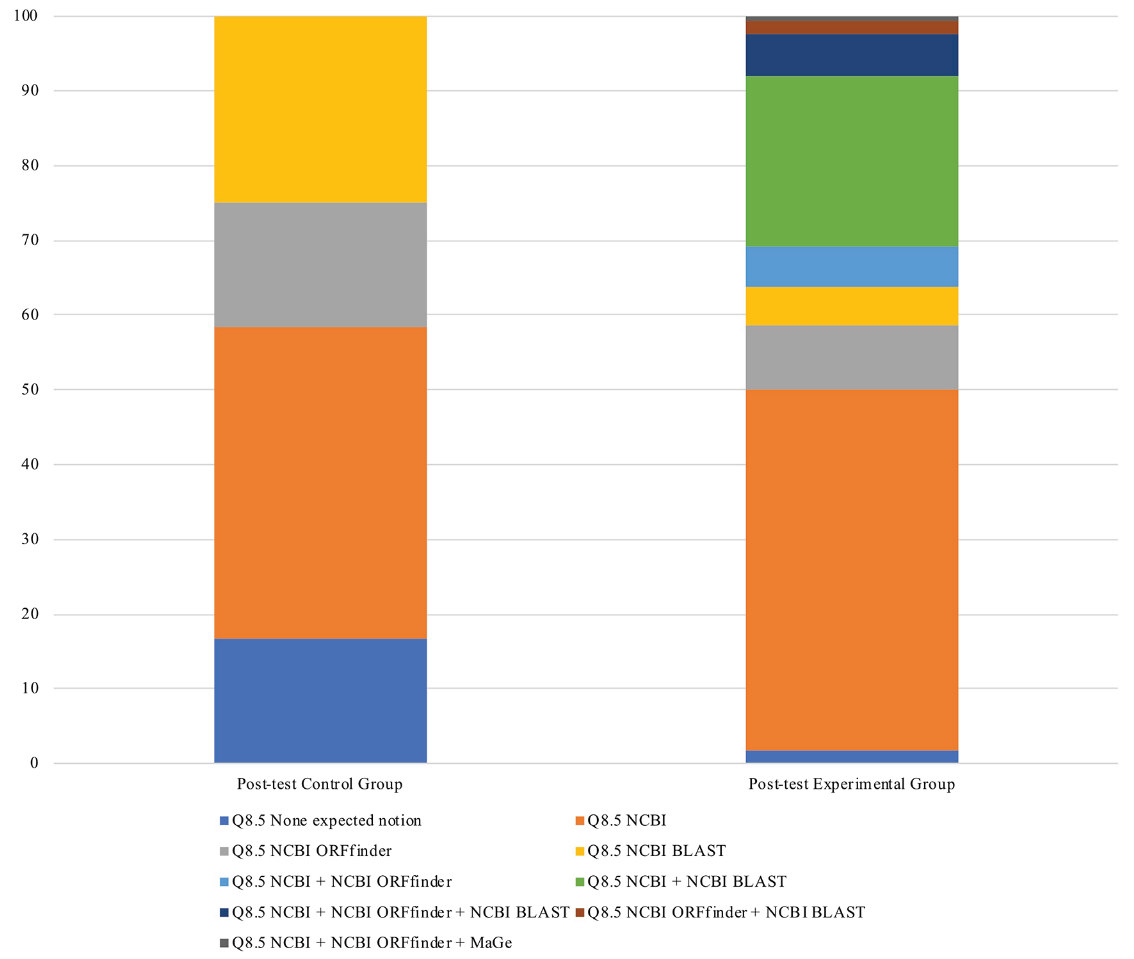


FIGURE 5 | Bioinformatics tools mentioned by students to unveil genes from bacterial genomics sequences.

DNA sequence," which was the most frequently recorded notion in the control group.

The post-test content analysis for the answers to Q2.1 revealed that 47.7% of the students in the control group did not mention any of the expected answers, 9.0% mentioned one of the expected answers, 41.8% mentioned two expected notions, and 1.5% mentioned all three expected notions. This trend improved in the experimental group, for which the percentage of students who mentioned one expected notion (14.3%) and all the expected notions (11.1%) was higher. Furthermore, the percentage of students who did not mention expected notions was lower in the experimental group than in the control group (38.6%).

Contrary to what was observed in the pre-test, in the post-test, students from both groups mentioned bioinformatics approaches, rather than wet laboratory techniques currently mentioned in their biology classes, such as electrophoresis and restriction enzymes. This outcome highlights the notion that, following a bioinformatics laboratory, most of the experimental designs envisioned by students to address a research question are based on a bioinformatics approach, instead of involving wet laboratory techniques that were already known to them. More than suggesting an enrichment of students' scientific toolbox and the development of thinking skills, the intervention seems to narrow the gap between students' school reality and what are common research practices nowadays, which is consistent with the educational benefits of bioinformatics reported in the literature (Gelbart and Yarden, 2006; Flanagan, 2013; Wood and Gebhardt, 2013). The data further suggest that when students are guided in the use of a wide variety of resources, they show to be capable to explore ideas and to interpret results in order to answer questions raised by the teacher (Kuhlthau et al., 2007).

Students' Knowledge on Gene Regulation and Genomics

According to the educational theories proposed by Ausubel (1968) and Vygotskiĭ and Cole (1978), students' prior knowledge, and in particular students' misconceptions, is of crucial importance when learning a new issue. Several diagnostic instruments are available, in published research studies, that can be used to obtain guidelines for specific interventions to address these misconceptions (Klymkowsky et al., 2010;

Tsui and Treagust, 2010; Gurel et al., 2015). Examples of students' key misconceptions regarding basic genetic and genomics notions are already described in the literature and include the use of gene and genome as synonyms, the misunderstanding of the relationship between a gene and DNA, a misinterpretation of the association between a gene and gene regulation, and the idea that some organisms, such as bacteria and fungi, often do not have DNA (Lewis and Kattmann, 2004; Mills Shaw et al., 2008). Adding the relevance of addressing these misconceptions, the Portuguese biology curriculum for the 11th grade (Mendes et al., 2003) recommends the discussion of the concept of "codogene" – part of a gene, i.e., a triplet of DNA, which is contributing to mislead students on the definition of gene. Having in mind the reported misconceptions, the activities implemented in this study aimed to tackle notions related with genes, genomes, alternative start codons, and the genetic code. Participants of both groups agreed that different bacteria groups have genes in common (Q7.1) and were shown to be aware that not all bacteria genes are identified and characterized, and that genomics information is still missing for many species (Q7.4; **Figure 6**). Conversely, misconceptions related with gene structure and the features of the genetic code did not seem to be overcome following the activities. In fact, students of both groups tended to disagree with the existence of different start codons (Q7.2) and were shown to be unaware of the existence of a bacterial genetic code (Q7.3), in both pre- and post-test (**Figure 6**).

These two questions were conceived having in mind that in high school, it is commonly taught that there is a unique start codon, a misconception that is reinforced in most textbooks. During the practical activities, students from the experimental group explored different start codons and worked with a specific bacteria-dedicated genetic code when using the tool NCBI ORFfinder, which was expected to make them aware of the specifications of the genetic code. However, surprisingly, the acquisition of this knowledge was not confirmed, which can be explained by reported evidence that even after being taught and accurately updated on a given scientific content for which misconceptions are observed, many students do not reconstruct their thinking (Mills Shaw et al., 2008). In this study, the practical component designed to address this particular misconception was also not effective. In fact, the use of misleading terms, simplified explanations that induce erroneous interpretations, adapted language, and everyday examples to explain the biological phenomena is often the origin of students' misconceptions, which can be tenacious and quite difficult to be overcome, ending up being perpetuated all through their high school education (Cho et al., 1985; Soyibo, 1995; Tekkaya, 2003; Mills Shaw et al., 2008). These data call for further attention and suggest that exercises specifically dedicated to exploring different start codons and distinct genetic codes according to the taxa of interest are needed to successfully overcome these deep-rooted misconceptions.

Other knowledge dimensions analyzed in this study include the concepts of genomics (Q4) and comparative genomics (Q5) aimed to acknowledge the importance of genomics in nowadays science and how it is impacting common societal sectors, such as human health and biotechnology. The results recorded for these two questions (Q4 and Q5) revealed a noticeable lack of knowledge about these concepts as previously described (Kirkpatrick et al., 2002; Mills Shaw et al., 2008; Baumler et al., 2012; Chen and Kim, 2014), which bares implications when trying to use bioinformatics tools.

In the post-test, 54.7% of the students in the experimental group provided a correct definition of genomics, i.e., "The field of science that studies genomes," trend that was not registered in the control group in which only 29.1% of the students were able to define this concept correctly. Zooming in the answers to identify the reported misuse of gene and genomics in an interchangeable way evidences a significant difference between the control and the experimental groups. In the pre-test, 1.5% of the control group students mentioned that genomics is a field of science that studies genes and/or genomes, a frequency that increased in the post-test (5.1%). In turn, in the experimental group, the trend was opposite, with the frequency of these notions decreasing from the pre- to the post-test (8.2 vs. 0.5%). These differences suggest an improvement of the quality of the answers of the students who carried out the bioinformatics exercises, i.e., the experimental group, apparently denoting that the expository teaching failed to clearly teach the difference between genomics and genetics. This may have resulted in the lack of accuracy witnessed in students' replies to question Q4, in what relates to the reference genome instead of gene. It is important to mention that in the particular case of the Portuguese science curriculum and in the NGSS, genomics is not at all mentioned; the topic addressed when referring to gene- and genome-related issues is genetics. In this regard, before the intervention, only a few students mentioned that they had heard about comparative genomics (Q5; **Figure 4**), an important concept that currently is not addressed in science classes (Martins and Tavares, 2018).

When students were asked to define comparative genomics in the post-test (Q5.1), the majority was able to do so correctly (79.5% in the experimental group, 75.3% in the control group). They associated the field with "genomic characteristics/genomes/genes/DNA sequences/homologous between different organisms," which suggests that the expository teaching on comparative genomics was efficient in fostering an accurate understanding about comparative genomics in students in both groups. As comparative genomics was a notion new to students, it was not conditioned by their previous perceptions, contrary to what happened with the concepts of genetics and genomics. Despite this general trend, question 5.1 was also aimed to depict more misconceptions that could be associated with the definition of comparative genomics. In the pre-test, 3.6% of the students in the experimental group mentioned that comparative genomics could be defined as comparisons between genes and phenotype, claiming that comparative genomics is the comparison between genetic sequences. The percentages of students with these misconceptions in the experimental and control groups lowered significantly in the post-test (2.4 and 1.4%, respectively). At this stage, i.e., in the post-test, a new notion was identified, with the experimental group students associating comparative genomics with the "Comparison of genomes of two or more species aiming to investigate phylogenetic relations" (6.7%).

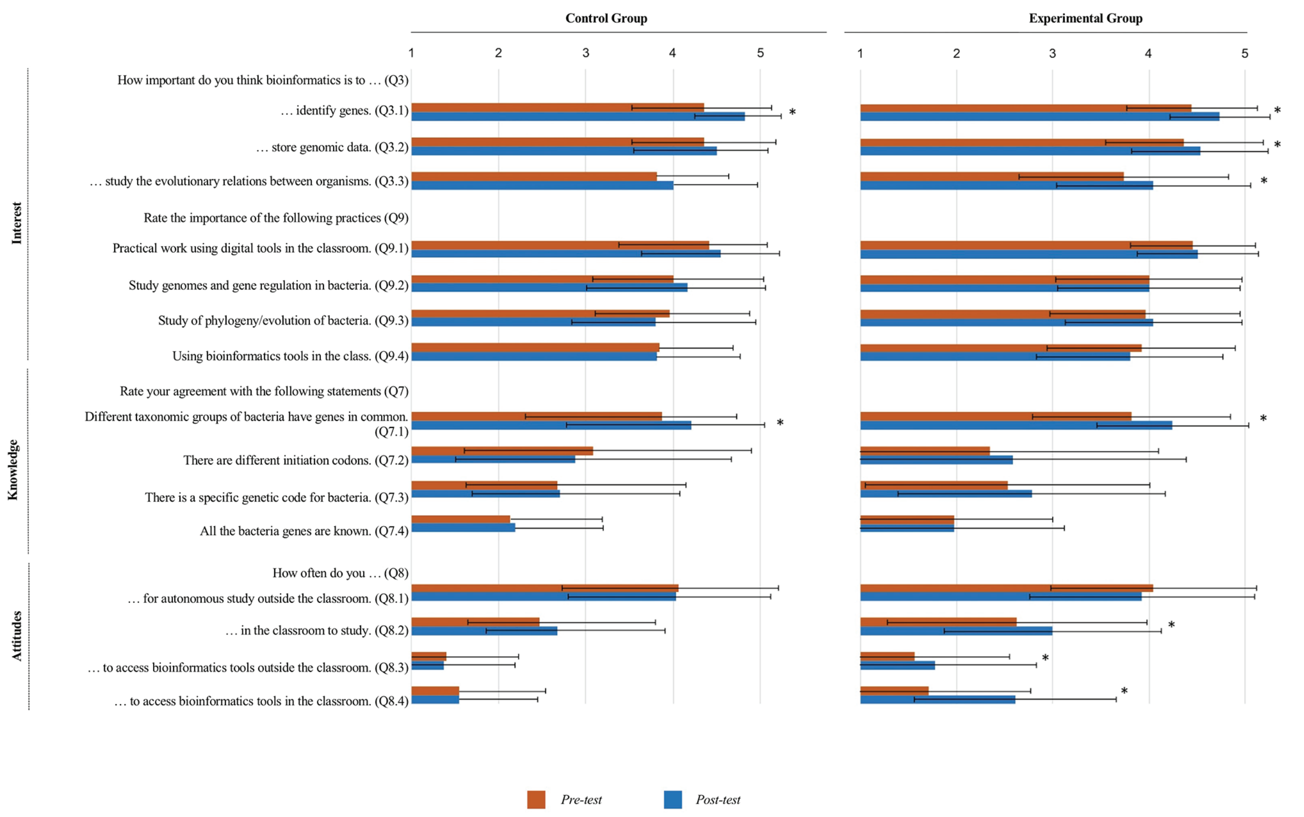


FIGURE 6 | Students' knowledge, interest, and attitudes toward the integration of bioinformatics in science curricula.

Having these outcomes in mind, it can be noted that the quality of answers of students in the experimental group improved after the intervention. It is worth mentioning that in the post-test, 5.2% of the control group students also recognized that comparative genomics can be associated with phylogenetic studies, which can be justified by the expositive teaching.

Attitudes and Interest

Together with the characterization of the students' knowledge regarding bioinformatics, gene regulation, and genomics, as described in the previous section, a depiction of their attitudes and interest toward bioinformatics was also carried out. As previously mentioned, in the context of this study, interest was interpreted according to Eccles' expectancy–value model (Eccles, 2005), which foresees motivation as a result of the combination of expectancy and value. The value given by students to a specific task is extremely important because they are more likely to pursue an activity if they acknowledge its worth. The model further differentiates task value into four components: attainment value (importance of doing it correctly), intrinsic value (personal enjoyment), utility value (perceived usefulness for future goals), and cost (competition with other goals; Eccles and Wigfield, 2002; Eccles, 2005; Leaper, 2011).

From the start, students were shown to be aware about the importance of bioinformatics to identify genes (Q3.1). Nevertheless, the classroom discussion that followed the expository teaching session about the need of bioinformatics tools to efficiently mine the huge genomics datasets contributed to reinforce this belief as demonstrated by the statistically significant difference between pre- and post-test results (**Figure 6**; **Table 1**).

Regarding the role of bioinformatics to store genomic data (Q3.2) and to study evolution (Q3.3), a statistically significant difference was observed from pre- to post-test in the experimental group (**Figure 6**; **Table 1**), but not in the control group. As the bioinformatics laboratories entailed the recruitment of bioinformatics resources particularly suited to access large datasets and address evolutionary inferences through synteny maps, these results highlight the direct impact of the intervention, which sustains identical results detailed in other studies (Luscombe et al., 2001; Kremer et al., 2005).

When asked to rate the importance of studying gene regulation (Q9.2) and evolution in bacteria (Q9.3), students in both groups agreed on its importance in both assessment moments (**Figure 6**). In what concerns the study of gene regulation (Q9.2.1), in the control group, its perceived importance was mainly connected with its usefulness from an instrumental point of view (60.3%; utility value), as suggested by expressions that linked its importance with the goals, such as "To get in touch with the world around us" or "To improve human life quality." Interestingly, in the experimental group, adding to the utilitarian value

TABLE 1 | Pre- and post-test comparison of students' knowledge, interest, and attitudes toward bioinformatics.

			Control group				Experimental group			
			t	***df***	***p***	**|*d*|**	***t***	***df***	***p***	**|*d*|**
Interest	How important do you think bioinformatics is to…(Q3)	…identify genes. (Q3.1)	−6.27	89	<0.01*	0.77	−5.94	274	<0.01*	0.48
		…store genomic data. (Q3.2)	−1.59	90	0.12	0.21	−2.66	271	0.01*	0.21
		…study the evolutionary relations between organisms. (Q3.3)	−1.87	78	0.07	0.22	−3.62	230	<0.01*	0.30
	Rate the importance of the following practices (Q9)	Practical work using digital tools in the classroom. (Q9.1)	−1.78	94	0.08	0.18	−1.32	281	0.19	0.08
		Study of genomes and gene regulation in bacteria. (Q9.2)	−1.65	72	0.10	0.17	0.07	209	0.94	0.00
		Study of phylogeny/ evolution of bacteria. (Q9.3)	0.48	62	0.63	0.15	−1.64	221	0.10	0.09
		Using bioinformatics tools in the class. (Q9.4)	0.00	41	1.00	0.03	1.33	181	0.18	0.12
Knowledge	Rate your agreement with the following statements (Q7)	Different taxonomic groups of bacteria have genes in common. (Q7.1)	−4.12	66	<0.01*	0.40	−6.08	182	<0.01*	0.47
		There are different initiation codons. (Q7.2)	0.99	92	0.33	0.11	−1.78	279	0.08	0.13
		There is a specific genetic code for bacteria. (Q7.3)	0.32	65	0.75	0.02	−1.61	191	0.11	0.17
		All the bacterial genes are known. (Q7.4)	−0.22	66	0.83	0.05	0.00	207	1.00	0.00
Attitudes	How often do you use the computer/ technological devices…(Q8)	…for autonomous study outside the classroom. (Q8.1)	0.20	94	0.84	0.02	1.72	290	0.09	0.11
		…in the classroom to study. (Q8.2)	−1.86	94	0.07	0.16	−5.03	290	<0.01*	0.30
		…to access bioinformatics tools outside the classroom. (Q8.3)	0.28	94	0.78	0.04	−2.82	285	0.01*	0.20
		…to access bioinformatics tools in the classroom. (Q8.4)	0.00	94	1.00	0.00	−11.89	286	<0.01*	0.85

*t, paired samples t-test for a 95% confidence interval (p); df, degrees of freedom; d, Cohen's d measure of effect size. *Indicates significant differences between pre- and post-test to each group.*

(42.7%), a more knowledge-related intrinsic worth (intrinsic value) was also well represented (42.7%), as shown by statements, such as "When we study bacteria it is interesting to have the chance to better understand this group and to get information about their metabolism in different environments." These results indicate that the scientific topic chosen for these activities is of interest to the students, and that the bioinformatics exercises carried out by the experimental group contributed to a more focused appraisal of the relevance of genomics and gene regulation. An identical trend was observed concerning the interest of evolutionary studies in bacteria (Q9.3.1), with 59.0% of the students in the experimental group and 50.9% of the students in the control group mentioning notions that reflect their motivation to explore the scientific topic, which emphasizes the importance of adding comparative genomic tools to the activities proposed.

As expected, students considered the practical work using digital tools important, engaging and motivating, raising their intrinsic interest (Q9.1, Q9.1.1; **Figure 6**). Concerning students' interest on the use of bioinformatics tools in the classroom, even before the *in silico* laboratories, they had already shown to be motivated in this regard (Q9.4; **Figure 6**). Despite the lack of statistically significant differences (**Table 1**), in the post-test, the students from both groups agreed that the integration of bioinformatics laboratories in the classroom (Q9.4) can have a beneficial impact to increase their intrinsic interest. This suggests their curiosity and awareness about the potential of using these tools in the classroom, regardless of whether they carried out (experimental group) or not (control group) the bioinformatics exercises.

Interesting remarks on the participants' engagement and interaction can be made based on the observations carried out during the implementation of the activities. For instance, the students were very surprised when they realized the incredible amount of open-access biological data, as translated by questions of amazement, such as "Can I access these bioinformatics resources for free at home?" and "Nice! Everyone can do it?." Having in mind we are now living in the post-genomic era, these reflections are crucial for students to get acquainted with genomics data sharing and to become aware of the social benefits and ethical implications of open access data (Foster and Sharp, 2007; Oliver et al., 2012).

Another aspect that students stated as being truly interesting pertained to the fact that they were sharing the exact same platforms used by professional researchers. These findings meet the reported importance of exposing science students to real-world phenomena and data, since this kind of activities can increase their interest and better prepare them for engaging in careers in science (Gelbart and Yarden, 2006; Flanagan, 2013). Furthermore, the observations showed that after completing the activities, students looked forward to exploring other tools in the platforms suggested, making comments, such as "What is the size of the genome of a spider?," "Are virus – such as HIV, genomes also available at this database?," or "Let us search for the gene coding for insulin." While this enhanced enthusiasm and curiosity have been reported for university science students (Chapman et al., 2006; Madlung, 2018), it has been poorly described in pre-university levels of education, which makes this finding even more interesting.

Confirming the participants' interest in learning science with bioinformatics tools is the fact that only a low percentage of students (13.5% in the experimental group, 9.3% in the control group) associated the integration of bioinformatics in the class (Q9.4.1) with a cost, according to Eccles' framework (Eccles, 2005). These students mentioned that incorporating bioinformatics in the classroom "is not that important once there are similar ways of obtaining the same results" or that "According to the Portuguese curricula for science in high school there is no need of using such complex tools" and also "This kind of activities can make classes more confusing since students are not used to working with these applications." These comments seem to reveal a lack of sympathy for innovative learning challenges.

As it is well-known, nowadays, youths are particularly at ease with digital resources (Rosen et al., 2010; Quinn and Oldmeadow, 2013), and indeed, students from both the experimental and the control groups admitted that they often take advantage of the technologies at their disposal in their autonomous study outside the classroom (Q8.1; **Figure 6**). Despite this reality, students from both groups stated that they do not use computers or other technological devices in the classroom (Q8.2, **Figure 6**). The statistically significant pre- to post-test increase observed in the answers to this question among the experimental group students is likely due to the unique opportunity created by this study for them to join bioinformatics laboratories (**Table 1**). Recent studies reported that although schools apparently have the necessary conditions to successfully integrate Information and Communications Technology (ICT) in the classroom, there are still barriers, such as teachers' pedagogical beliefs, which prevent the use of computers in classroom settings (Marcinkiewicz, 1993; Ertmer, 2005; Sang et al., 2010). Interestingly, some informal comments made by the students revealed that their teachers often feel discouraged to use technology in the classroom because they do not feel comfortable with it, which meets the constraints mentioned by the teacher, the majority of whom acknowledged their anxiety regarding the use of technology in this setting (Machluf and Yarden, 2013; Martins et al., 2018c, 2020).

Even though students of both groups also revealed (Q8.3) that they usually do not access bioinformatics tools outside the classroom, there is a significant pre- to post-test difference for the experimental group, which may suggest that these students decided to take advantage of the bioinformatics resources explored after the activities (**Figure 6**; **Table 1**). Regarding the specific use of bioinformatics tools in the classroom (Q8.4), while in the pre-test, students from both groups answered negatively to this question, as expected, in the post-test, the students from the experimental group reported that they used bioinformatics in their classes (**Figure 6**; **Table 1**).

Having in mind that the students who took part in this study belong to a highly technological society, one can anticipate that their performance in manipulating computer-based tools was efficient (Rosen et al., 2010; Quinn and Oldmeadow, 2013). Indeed, and regardless that most of the students had never experienced working with bioinformatics tools before, during the implementation of the bioinformatics laboratories, no major difficulties to follow the guidelines and discussing the issues raised were reported to the teacher. The observations showed that students were completely able to manage the platforms and did not feel the need to use printed out guidelines. Instead, they looked for solutions and alternatives together with their classmates and took advantage of the technological resources available, namely, smartphones. In spite of the expectable side talk, the participants' behavior and their questions and comments suggest their engagement in every task that they were asked to perform.

CONCLUSION

The findings obtained in this study demonstrate an improvement in students' knowledge of concepts, such as gene, protein synthesis, nucleic acid (DNA, RNA), start and stop codons, genome, evolutionary relations, and genomic or comparative genomics, following their participation in bioinformatics-based activities "Mining the Genome: Using Bioinformatics Tools in the Classroom to Support Students Discovery of Genes" (Martins et al., 2018a). By the end of the activities, students were also shown to be more aware of the applications and potential of bioinformatics.

This study also raises several questions that are worth pursuing in future research, namely, related with misconceptions that were addressed in this intervention. In addition, future focus on other school levels (namely, middle school) and other curricular topics might be relevant to cross-examine and more widely and consistently depict the impact of bioinformatics-based activities in the classroom. Likely pertinent will be to assess the influence of the "teacher" in students' performance through a nested effect analysis.

Beyond the evidence of the educational benefits of incorporating practical activities in science education programs, overall, this study represents a contribution to introduce a top-notch research area – bioinformatics – in school and to inform stakeholders about its potential from not only educational but also scientific and other social points of view.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The project was institutionally approved by each school's Directive Board. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

AM, MJF, and FT designed the research plan. AM followed the implementation of the instruments in the classroom and collected, organized, and analyzed the data. MJF, ML, LL, and FT participated in the data analysis and interpretation. All authors contributed to the writing of this article and approved the submitted version.

FUNDING

AM is supported by a fellowship from Fundação para a Ciência e Tecnologia – FCT (SFRH/BD/112038/2015). This work is supported by National Funds through FCT – Fundação para a Ciência e a Tecnologia in the scope of the project UIDB/50027/2020.

ACKNOWLEDGMENTS

The authors are grateful to all participants (teachers, students, and schools) for creating the opportunity to implement these bioinformatics laboratories and carry out this study.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.578099/full#supplementary-material

REFERENCES

Agarwala, R., Barrett, T., Beck, J., Benson, D. A., Bollin, C., Bolton, E., et al. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* 46, D8–D13. doi: 10.1093/nar/gkx1095

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410. doi: 10.1016/S0022-2836(05)80360-2

Amenkhienan, E., and Smith, E. J. (2006). A web-based genetic polymorphism learning approach for high school students and science teachers. *Biochem. Mol. Biol. Educ.* 34, 30–33. doi: 10.1002/bmb.2006.49403401030

Attwood, T. K., Blackford, S., Brazas, M. D., Davies, A., and Schneider, M. V. (2017). A global perspective on evolving bioinformatics and data science training needs. *Brief. Bioinform.* 20, 398–404. doi: 10.1093/bib/bbx100

Ausubel, D. P. (1968). *Educational psychology: A cognitive view*. Michigan: Holt, Rinehart and Winston.

Baumler, D. J., Banta, L. M., Hung, K. F., Schwarz, J. A., Cabot, E. L., Glasner, J. D., et al. (2012). Using comparative genomics for inquiry-based learning to dissect virulence of *Escherichia coli* O157:H7 and *Yersinia pestis*. *CBE Life Sci. Educ.* 11, 81–93. doi: 10.1187/cbe.10-04-0057

Campbell, C. E., and Nehm, R. H. (2013). A critical analysis of assessment quality in genomics and bioinformatics education research. *CBE Life Sci. Educ.* 12, 530–541. doi: 10.1187/cbe.12-06-0073

Champagne Queloz, A., Klymkowsky, M. W., Stern, E., Hafen, E., and Köhler, K. (2017). Diagnostic of students' misconceptions using the Biological Concepts Instrument (BCI): a method for conducting an educational needs assessment. *PLoS One* 12:e0176906. doi: 10.1371/journal.pone.0176906

Chapman, B. S., Christmann, J. L., and Thatcher, E. F. (2006). Bioinformatics for undergraduates: steps toward a quantitative bioscience curriculum. *Biochem. Mol. Biol. Educ.* 34, 180–186. doi: 10.1002/bmb.2006.49403403180

Chen, L. -S., and Kim, M. (2014). Needs assessment in genomic education. *Health Promot. Pract.* 15, 592–598. doi: 10.1177/1524839913483470

Cho, H. -H., Kahle, J. B., and Nordland, F. H. (1985). An investigation of high school biology textbooks as sources of misconceptions and difficulties in genetics and some suggestions for teaching genetics. *Sci. Educ.* 69, 707–719. doi: 10.1002/sce.3730690512

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Lawrence Erlbaum Associates.

Connelly, L. M. (2008). Pilot studies. *Medsurg Nurs.* 17, 411–413.

Dudley, J. T., and Butte, A. J. (2009). A quick guide for developing effective bioinformatics programming skills. *PLoS Comput. Biol.* 5:e1000589. doi: 10.1371/journal.pcbi.1000589

Eccles, J. S. (2005). "Subjective task value and the Eccles et al. model of achievement-related choices" in *Handbook of competence and motivation*. eds. A. J. Elliot and C. S. Dweck (New York: The Guilford Press), 105–121.

Eccles, J. S., and Wigfield, A. (2002). Motivational beliefs, values and goals. *Annu. Rev. Psychol.* 53, 109–132. doi: 10.1146/annurev.psych.53.100901.135153

Ertmer, P. A. (2005). Teacher pedagogical beliefs: the final frontier in our quest for technology integration? *Educ. Technol. Res. Dev.* 53, 25–39. doi: 10.1007/BF02504683

Flanagan, J. (2013). Open data for science education. PLoS Blogs. doi: 10.1525/bio.2010.60.5.2

Fonseca, M. J., Costa, P., Lencastre, L., and Tavares, F. (2012). Multidimensional analysis of high-school students' perceptions about biotechnology. *J. Biol. Educ.* 46, 129–139. doi: 10.1080/00219266.2011.634019

Form, D., and Lewitter, F. (2011). Ten simple rules for teaching bioinformatics at the high school level. *PLoS Comput. Biol.* 7:e1002243. doi: 10.1371/journal.pcbi.1002243

Foster, M., and Sharp, R. (2007). Share and share alike: deciding how to distribute the scientific and social benefits of genomic data. *Nat. Rev. Genet.* 8, 633–639. doi: 10.1038/nrg2124

Gelbart, H., Brill, G., and Yarden, A. (2009). The impact of a web-based research simulation in bioinformatics on students' understanding of genetics. *Res. Sci. Educ.* 39, 725–751. doi: 10.1007/s11165-008-9101-1

Gelbart, H., and Yarden, A. (2006). Learning genetics through an authentic research simulation in bioinformatics. *J. Biol. Educ.* 40, 107–112. doi: 10.1080/00219266.2006.9656026

Gurel, D. K., Eryilmaz, A., and McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *Eurasia J. Math. Sci. Technol. Educ.* 11, 989–1008. doi: 10.12973/eurasia.2015.1369a

Hsieh, H. -F., and Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qual. Health Res.* 15, 1277–1288. doi: 10.1177/1049732305276687

Johanson, G. A., and Brooks, G. P. (2010). Initial scale development: sample size for pilot studies. *Educ. Psychol. Meas.* 70, 394–400. doi: 10.1177/0013164409355692

Johnson, B., and Christensen, L. B. (2012). *Educational research: Quantitative, qualitative, and mixed approaches.* Thousand Oaks, California: SAGE Publications.

Kirkpatrick, G., Orvis, K., and Pittendrigh, B. (2002). A teaching model for biotechnology and genomics education. *J. Biol. Educ.* 37, 31–35. doi: 10.1080/00219266.2002.9655843

Klymkowsky, M. W., Underwood, S. M., and Garvin-Doxas, R. K. (2010). Biological Concepts Instrument (BCI): a diagnostic tool for revealing student thinking. Available at: http://arxiv.org/abs/1012.4501 (Accessed June 23, 2020).

Koch, I., and Fuellen, G. (2008). A review of bioinformatics education in Germany. *Brief. Bioinform.* 9, 232–242. doi: 10.1093/bib/bbn006

Kovarik, D., Patterson, D., Cohen, C., Sanders, E., Peterson, K., Porter, S., et al. (2013). Bioinformatics education in high school: implications for promoting science, technology, engineering, and mathematics careers. *CBE Life Sci. Educ.* 12, 441–459. doi: 10.1187/cbe.12-11-0193

Kremer, A., Schneider, R., and Terstappen, G. C. (2005). A bioinformatics perspective on proteomics: data storage, analysis, and integration. *Biosci. Rep.* 25, 95–106. doi: 10.1007/s10540-005-2850-4

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, California: SAGE Publications.

Kuhlthau, C. C., Caspari, A. K., and Maniotes, L. K. (2007). *Guided inquiry: Learning in the 21st century*. New York: Libraries Unlimited.

Leaper, C. (2011). More similarities than differences in contemporary theories of social development?: a plea for theory bridging. *Adv. Child Dev. Behav*. 40, 337–378. doi: 10.1016/b978-0-12-386491-8.00009-8

Lewis, J., and Kattmann, U. (2004). Traits, genes, particles and information: re-visiting students' understandings of genetics. *Int. J. Sci. Educ.* 26, 195–206. doi: 10.1080/0950069032000072782

Luscombe, N., Greenbaum, D., and Gerstein, M. (2001). What is bioinformatics? A proposed definition and overview of the field. *Methods Inf. Med.* 40, 346–358. doi: 10.1053/j.ro.2009.03.010

Machluf, Y., Gelbart, H., Ben-Dor, S., and Yarden, A. (2017). Making authentic science accessible-the benefits and challenges of integrating bioinformatics into a high-school science curriculum. *Brief. Bioinform.* 18, 145–159. doi: 10.1093/bib/bbv113

Machluf, Y., and Yarden, A. (2013). Integrating bioinformatics into senior high school: design principles and implications. *Brief. Bioinform.* 14, 648–660. doi: 10.1093/bib/bbt030

Madlung, A. (2018). Assessing an effective undergraduate module teaching applied bioinformatics to biology students. *PLoS Comput. Biol.* 14:e1005872. doi: 10.1371/journal.pcbi.1005872

Magana, A. J., Taleyarkhan, M., Alvarado, D. R., Kane, M., Springer, J., and Clase, K. (2014). A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research. *CBE Life Sci. Educ.* 13, 607–623. doi: 10.1187/cbe.13-10-0193

Marcinkiewicz, H. R. (1993). Computers and teachers. *J. Res. Comput. Educ.* 26, 220–237. doi: 10.1080/08886504.1993.10782088

Marques, I., Almeida, P., Alves, R., Dias, M., Godinho, A., and Pereira-Leal, J. (2014). Bioinformatics projects supporting life-sciences learning in high schools. *PLoS Comput. Biol.* 10:e1003404. doi: 10.1371/journal.pcbi.1003404

Martins, A., Fonseca, M. J., and Tavares, F. (2018a). Mining the genome: using bioinformatics tools in the classroom to support student discovery of genes. *Am. Biol. Teach.* 80, 619–624. doi: 10.1525/abt.2018.80.8.619

Martins, A., Lencastre, L., and Tavares, F. (2018b). "Predictive microbiology in a non- formal science education context: understanding food preservation techniques" in *Hands-on science*. *Advancing science. Improving education*. eds. M. Costa, B. Dorrío and J. Fernandez Novell. July 16–20, 2018 (Barcelona: Hands-on Science Network), 309–317.

Martins, A., Lencastre, L., and Tavares, F. (2018c). "Integrating bioinformatics in elementary and secondary education: teacher's perceptions" in *3rd International Conference on Teacher Education (INCTE)*; May 4–5, 2018 (Bragança: Instituto Politécnico de Bragança).

Martins, A., Lencastre, L., and Tavares, F. (2020). "Bioinformatics, a befitting tool for e-learning: potential and constrains according teachers' perceptions" in *Hands-on science*. *Science education. Discovering and understanding the wonders of nature.* eds. M. F. Costa and J. B. Dorrío. July 13–17, 2020 (Hands-on Science Network), 97–105.

Martins, A., and Tavares, F. (2018). "Genomics education: update core concepts in high school" in *Hands-on science. Advancing science. Improving education*. eds. M. Costa, B. Dorrío and J. Fernandez-Novell (Barcelona: Hands-on Science Network), 145–150.

Mendes, A., Rebelo, D., and Pinheiro, E. (2003). Programa de Biologia e Geologia 11º ou 12ºano(s).

Mendes, A., Rebelo, D., and Pinheiro, E. (2004). Biologia 12ºano—Curso Científico Humanístico de Ciências e Tecnologias.

Mills Shaw, K. R., Van Horne, K., Zhang, H., and Boughman, J. (2008). Essay contest reveals misconceptions of high school students in genetics content. *Genetics* 178, 1157–1168. doi: 10.1534/genetics.107.084194

National Research Council (2013). *Next generation science standards*. Washington, DC: National Academies Press.

Newman, L., Duffus, A. L. J., and Lee, C. (2016). Using the free program MEGA to build phylogenetic trees from molecular data. *Am. Biol. Teach.* 78, 608–612. doi: 10.1525/abt.2016.78.7.608

Oliver, J., Slashinski, M., Wang, T., Kelly, P., Hilsenbeck, S., and McGuire, A. (2012). Balancing the risks and benefits of genomic data sharing: genome research participants' perspectives. *Public Health Genomics* 15, 106–114. doi: 10.1159/000334718

Pallant, J. (2007). *SPSS—survival guide to data analysis using SPSS for windows*. Maidenhead: Open University Press/McGraw-Hill.

Punch, K. F. (2009). *Introduction to research methods in education*. Los Angeles: SAGE Publications.

Quinn, S., and Oldmeadow, J. A. (2013). Is the iGeneration a "we" generation? Social networking use among 9- to 13-year-olds and belonging. *Br. J. Dev. Psychol.* 31, 136–142. doi: 10.1111/bjdp.12007

Rosen, L. D., Carrier, M. L., and Cheever, N. A. (2010). *Rewired: Understanding the iGeneration and the way they learn*. New York: Palgrave Macmillan.

Sadek, H. (2004). *Bioinformatics: Principles, basic internet applications*. Canada: Trafford Publishing.

Sang, G., Valcke, M., Braak, J.van, and Tondeur, J. (2010). Student teachers' thinking processes and ICT integration: predictors of prospective teaching behaviors with educational technology. *Comput. Educ.* 54, 103–112. doi: 10.1016/J.COMPEDU.2009.07.010

Schreier, M. (2012). *Qualitative content analysis in practice*. London, United Kingdom: SAGE Publications.

Soyibo, K. (1995). A review of some sources of students' misconceptions in biology. *Singapore J. Educ.* 15, 1–11. doi: 10.1080/02188799508548576

Taylor, J. M., Davidson, R. M., and Strong, M. (2014). Drug-resistant tuberculosis. *Am. Biol. Teach.* 76, 386–394. doi: 10.1525/abt.2014.76.6.6

Tekkaya, C. (2003). Remediating high school students' misconceptions concerning diffusion and osmosis through concept mapping and conceptual change text. *Res. Sci. Technol. Educ.* 21, 5–16. doi: 10.1080/02635140308340

Treece, E. W., and Treece, J. W. Jr. (1982). *Elements of research in nursing*. St. Louis: Mosby.

Tsui, C., and Treagust, D. (2010). Evaluating secondary students' scientific reasoning in genetics using a two-tier diagnostic instrument. *Int. J. Sci. Educ.* 32, 1073–1098. doi: 10.1080/09500690902951429

Vallenet, D., Belda, E., Calteau, A., Cruveiller, S., Engelen, S., Lajus, A., et al. (2013). MicroScope—an integrated microbial resource for the curation and comparative analysis of genomic and metabolic data. *Nucleic Acids Res.* 41, D636–D647. doi: 10.1093/nar/gks1194

Vygotskiĭ, L. S., and Cole, M. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, Mass, United States: Harvard University Press.

Weber, R. (1990). *Basic content analysis*. Thousand Oaks, California: SAGE Publications.

Wefer, S. H., and Sheppard, K. (2008). Bioinformatics in high school biology curricula: a study of state science standards. *CBE Life Sci. Educ.* 7, 155–162. doi: 10.1187/cbe.07-05-0026

Wood, L., and Gebhardt, P. (2013). Bioinformatics goes to school—new avenues for teaching contemporary biology. *PLoS Comput. Biol.* 9:e1003089. doi: 10.1371/journal.pcbi.1003089

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 26 October 2020
doi: 10.3389/fmicb.2020.578810

Antibiotic Resistance in Environmental Microbes: Implementing Authentic Research in the Microbiology Classroom

Mangala Tawde and Marianne Williams*

Queensborough Community College, CUNY, Bayside, NY, United States

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Felipe Hernandes Coutinho,
Miguel Hernández University of Elche,
Spain
Justin Joseph Donato,
University of St. Thomas,
United States

****Correspondence:***
Mangala Tawde
mtawde@qcc.cuny.edu

Specialty section:
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

Received: *01 July 2020*
Accepted: *05 October 2020*
Published: *26 October 2020*

Citation:
Tawde M and Williams M (2020)
Antibiotic Resistance in Environmental
Microbes: Implementing Authentic
Research in the Microbiology
Classroom.
Front. Microbiol. 11:578810.
doi: 10.3389/fmicb.2020.578810

Incorporating Undergraduate Research Experience in Microbiology Classroom. Dr. Mangala Tawde, Associate Professor, Department of Biological Sciences and Geology, Queensborough Community College, CUNY. Undergraduate Research (UR) experience is increasingly being recognized as one of the most transforming experiences students can have in their undergraduate years of education. To make it accessible to all students, incorporating authentic research experiences in the classroom is important and it is a major initiative at Queensborough community college; where we have institutionalized UR as a High Impact Practice. We incorporated an authentic research project into the Microbiology course curriculum for allied health majors. The research project was to isolate and identify antibiotic-resistant microbes from diverse environments. As students are aware of antibiotic resistance being a serious concern in today's medicine, they get interested and are enthusiastically engaged in the research project. Students collect soil samples from various environments and locations of their choice and then they isolate and identify bacteria that may exhibit antibiotic resistance. The microbes isolated from diverse environments are identified based on the 16s rRNA sequence analysis as well as biochemical tests. The research experience is relevant and aligns well with the course curricula, course learning objectives as well as the college's General Education objectives.

Keywords: undergraduate research experience, course based undergraduate research experiences, antibiotic resistance, environmental microbiome, community college undergraduate courses

INTRODUCTION

Inquiry-based team learning is shown to be vital for developing skills such as critical-thinking, scientific problem-solving ability, and acquiring scientific content knowledge in undergraduate biology education (Lord, 2001; Apedoe et al., 2006; Hunter et al., 2007; Kuh, 2008). Many recent studies have shown that research experiences for students early on during their undergraduate years, result in improved learning outcomes, and science career decisions leading to a stronger Science, Technology, Engineering and Mathematics (STEM) workforce (Lopatto, 2004, 2007; Kuh, 2008). Thus Undergraduate Research (UR) experience is considered as one of the best practices to engage and motivate students in undergraduate education (Lopatto, 2004, 2007; Kuh, 2008; Lopatto and Tobias, 2010). Though the traditional one-on-one apprenticeship model with a specific mentor for research internship is known to transform students' lives and careers, its accessibility is limited to a few students (National Academies of Sciences Engineering and Medicine, 2015). In order to make the pedagogy of undergraduate research accessible to all students, authentic research experiences need to be implemented and incorporated in the undergraduate classroom setting. Thus course-based undergraduate research experiences (CUREs) incorporated in the classroom setting are the response to national "Call for Action" (National Research Council, 2003; American Association for the Advancement of Science, 2011; Ballen et al., 2017) to reform the undergraduate Biology curriculum (Handelsman et al., 2004; Woodin et al., 2010; Lopatto et al., 2011; Wei and Woodin, 2011; Dolan, 2012; Caplan and MacLachlan, 2014; Brownell and Kloser, 2015; Ballen et al., 2017). Students involved in research-based courses are more engaged, more likely to complete their courses, show a greater appreciation of science and inclination toward STEM careers and are more likely to pursue them as compared to those taking traditional courses (Handelsman et al., 2004; Woodin et al., 2010; Lopatto et al., 2011; Wei and Woodin, 2011; Dolan, 2012; Caplan and MacLachlan, 2014; Brownell and Kloser, 2015). There are numerous CUREs that have been proposed as inclusive models to make these experiences accessible to all students (Handelsman et al., 2004; Woodin et al., 2010; Lopatto et al., 2011; Wei and Woodin, 2011; Dolan, 2012; Auchincloss et al., 2014; Caplan and MacLachlan, 2014; Brownell and Kloser, 2015; Brownell et al., 2015; Corwin et al., 2015a,b; Bangera and Brownell, 2017; Mader et al., 2017). However, at institutions without a strong research infrastructure or resources such as community colleges, it is a totally different beast of a challenge for the faculty to convert an entire semester-long course into a CURE. Here we describe a course-based research experience where we incorporated an authentic research experience of studying antibiotic resistance in bacteria isolated from environmental samples into a microbiology lab course that is required for allied health majors.

The student body at Queensborough Community College (QCC) at City University of New York (CUNY) is extremely diverse in its ethnic, cultural and financial backgrounds as well as levels of college preparedness. The unique demographics and needs of CUNY's community college student population present multiple barriers to students success. Most students come from lower income households, they juggle work, school and family obligations in one of the nation's most expensive cities. Many have not had science classes in high schools or are returning to school after a hiatus. Understandably, these students are highly unprepared for college-level learning experiences leading to attrition rates of over 30% in our science classes. Therefore incorporating UR experience in classroom is a vital strategy to engage these students, retain and motivate them for rewarding and meaningful educational experiences especially in STEM.

Queensborough CC institutionalized Undergraduate Research (UR) as a High Impact Practice (HIP) in 2013–2014. UR as a HIP is a learning-centered and student centered practice supported by student learning outcomes, assessments, and professional development. Since spring 2014, over 60 faculty members have participated in UR professional development. Close to 100 UR experiences have been offered reaching over 800 students –in addition to the students who engage in the more traditional, dedicated research experiences of the apprenticeship model (QCC Fact book 2018–2019).

The undergraduate research experience in Microbiology course started as a "Research in the Classroom (RIC)" grant initiative that was awarded to M. Tawde by CUNY's Office of Research. We teach a one-semester Microbiology course (BI 311) that is offered to students seeking to pursue allied health careers and programs. The students typically are rushing to finish the course to get into Nursing, Physician's Assistant or other programs or may already be in their desired programs. Hence undergraduate research is usually not on their radar and they are not planning to participate in any research program or internship. Most students in our courses have never had any prior UR experience. M. Tawde also teaches one section of Environmental Health class (BI 501) every spring semester. The research experience was implemented in one section of BI 501 and 4 sections of BI 311 lab courses thus involving about 80 students.

Course Description of Microbiology (BI 311)

A one semester, 4- credit course, Microbiology is intended for Nursing and Allied Health students. The course involves a systematic study of the bacteria, viruses, fungi and helminths with an emphasis on those associated with infectious diseases. Laboratory work includes microbiological techniques and procedures for control.

Course Description of Environmental Health (BI 501)

A one semester, 4- credit course. An introduction to our environment and its influence on human health; emphasis on scientific principles needed to understand environmental requirements of life; role of air, water, food, energy; studies of effect of human activity on environment and effect of modified environment on human health.

As both classes have a common focus on human health, it is imperative to study the effect environmental microbes may have on human health. Antibiotic resistance is a grave concern in the fields of medicine and healthcare (Allen et al., 2010; Centers for Disease Control and Prevention, 2013; World Health Organization, 2014). Biopharmaceutical agencies are trying to keep up with the growing demand for novel drugs to defeat the antibiotic-resistant pathogens. Hence, we decided to bring this research into our classroom by integrating it into the course curriculum.

Typically in a Microbiology laboratory, students start to learn basic microbiology concepts and standard techniques such as aseptic technique, isolation of bacteria from mixed cultures, staining techniques etc., and then continue to learn how to identify bacteria using Gram staining and various metabolic tests. The midterm practical is conducted over a period of 4–6 weeks and involves identification of "unknown" bacteria. Students learn all the standard "cookbook" microbiology techniques needed to identify the "unknown" bacteria which are actually pure cultures of known bacteria provided to them as unknowns. Thus the students do not receive an authentic research experience.

The research project that we implemented in this course was titled "Research in the Classroom: Antibiotic Resistance in Environmental Microbes." The goal of the project was to provide an authentic research investigation experience to students as part of their Microbiology laboratory curriculum while they isolate and identify novel microorganisms from the environment and study their resistance/susceptibility to most commonly used antibiotics. Students are aware that antibiotic resistance is a serious concern in the field of health care today. So they are immediately interested and enthusiastic about participating in the research project. The laboratory course syllabi were modified to incorporate the non-traditional activities such as DNA extraction, PCR and DNA analysis by Agarose Gel. A similar CURE has been developed at a larger scale as the PARE project (Genné-Bacon and Bascom-Slack, 2018) as we were developing ours. It is a crowd-sourcing monitoring system that engages students across the country to systematically test and report the prevalence of tetracycline-resistant bacteria from soil at diverse geographic sites. However, our model involved testing antibiotic susceptibility against 12 different antibiotics; not just tetracycline and targets a student population who would not have a research experience otherwise. The majority of students in the above classes usually focus on learning just the techniques but not the concepts behind the techniques or their applications in the real world. Since humans and microorganisms co-exist in dynamic relationships in nature and these relationship critically affects human health; it is crucial that the applications of the microbial genomics are emphasized and understood.

As we implemented the research experience, we attempted to ask more specific questions-

(a) What type of microbes exist at various environments for example soil vs. water vs. surfaces of objects. Do you find more number/types of bacteria in environments with higher human activity as compared to natural environments?

(b) Are the microbes from crowded areas more resistant to antibiotics (or wider variety of antibiotics) compared to those that are isolated from natural environments? Does the environment have any effect on antibiotic susceptibility of organisms that reside in it?

MATERIALS AND METHODS

Timeline to incorporate research lab activities during a 15-week semester of a microbiology lab. Laboratory class of Environmental Health class will have similar outline:

	Traditional Laboratory Outline	**Laboratory Outline with implemented research experience**
Week 1	Use and care of the microscope; diversity of microbial life; bacterial shapes	Use and care of the microscope; microbial diversity;Introduction of research project, sample collection.
Week 2	Basic aseptic technique; isolation of single colonies; culturing microbes from the environment; selective and differential media	Basic aseptic technique, culturing environmental samples and isolation of single colonies; selective and differential media
Week 3	Introduction to smear preparation; staining techniques, Gram staining and special stains	Introduction to smear preparation; staining techniques Gram staining and special stains
Week 4	Acid-fast stain; endospore stain; Practice for Gram stain	DNA extraction of unknown environment isolates, set up PCR, practice Gram stain
Week 5	Mid-term Lab practical- part 1: Gram stain of unknowns; Inoculate for Metabolic activities	Mid-term Lab practical: Gram stain unknowns and unknown environmental isolates
Week 6	Analysis of metabolic activities, Preparation of dichotomous key for Lab Practical I unknowns	Analysis of metabolic activities, Running Agarose gels, prepare samples for sequencing
Week 7	Physical control of microorganisms: temperature, UV radiation, moisture, Inoculate for Practical I: Part 2 (inoculate metabolic tests)	Physical control of microorganisms: temperature, UV radiation, moisture; Practical I—Part 2 (inoculate metabolic tests)
Week 8	Lab Practical I - Part 2: Analysis of metabolic tests for unknowns, Chemical control of microorganisms: disinfectants and antibiotics	Lab Practical I - Part 2: Analysis of metabolic tests for unknowns, Chemical control of microorganisms: Test for antibiotics resistance
Week 9	Quantification of bacteria in food- milk and chicken broth	Quantification of bacteria in food-milk/chicken broth, Unknowns sequences Analysis
Week 10	Lab reports for unknown due	Lab reports for unknown due

Description of the Research Activity

The research component was implemented during the spring and fall of 2017 and 2018 semesters. This authentic microbiology wet-lab, hands-on research experience was carried out in groups of 4–5 students each. The students needed to meet twice during the semester outside the class time, (typically during the club hours) each for a block of 1–2 h. These meetings were typically followed by the regular lab class. The first meeting is for DNA extraction and setting up PCR while the second meeting is held to analyze the sequencing data and identification of bacterial species.

Students formulated a hypothesis as to which environmental site may contain the most harmful or highest number of bacteria. Based on their hypothesis, they selected sites for sample collection and went around to swab a small area from the sites such as cafeteria, gym, bathroom, bus-stops, nature trails and botanical garden etc. Some samples came from students' cell phones. Students were provided with sterile wet swabs to collect the samples of choice. They were asked to bring in the samples at the second class meeting and possibly collect the sample right before the class. After students brought in the soil/surface samples, they streaked them on to sterile Tryptic Soy agar plates and incubated further for growth. At next class meeting, single isolated bacterial colonies were picked and grown in Tryptic Soy broth to confluent cultures. DNA extraction was carried out by using the MoBio DNA PowerSoil DNA isolation kit or Qiagen DNeasy PowerSoil kit and the kit protocols. DNA extraction was followed by setting up a 50 μl polymerase chain reaction (PCR) to amplify the 16s rRNA gene. Once amplified, small amount (10 μl) of amplicons were analyzed by running an agarose gel in the class to ensure amplification of the correct gene product. Remaining amplified product was sent to external sequencing facility GENEWIZ, Inc.,[1] for sequencing. When the sequencing data was received, it was analyzed using NCBI or DNA Learning Center (DNALC) databases. Students determined the identity of the bacteria by doing a BLAST (Basic Local Alignment Search Tool) search from the National Center for Biotechnology Information (NCBI) database. Alternatively, students used a user friendly version of BLAST – the "DNA Subway" program which is hosted by the DNA Learning Center of Cold Spring Harbor laboratory[2] (**Supplementary Material Part II**).

After identifying bacterial species students streaked some of the isolates on the Mueller-Hinton agar plates to form uniform bacterial lawns and carried out Kirby Bauer disk diffusion assay for testing antibiotic susceptibility of select isolates. A BBL disk dispenser was used to dispense commercially available disks impregnated with 12 antibiotics- penicillin, vancomycin, polymyxin B, nitrofurantoin, tobramycin, streptomycin, ciprofloxacin, oxacillin, piperacillin, gentamicin, neomycin, and ampicillin.

All the laboratory procedures were carried out in a BSL 2 laboratory with two hand washing stations, an eye-wash station, an emergency shower, fire blanket etc. Students performed bacterial culturing procedures using aseptic techniques with Bunsen burners and mandatory lab coats. For all bacterial isolates that showed antibiotic resistance, students were supervised closely for all the following procedures performed.

The students submit a comprehensive lab activity report at the end of the semester. The entire research project makes up 10% of the course grade for the students. Other course sections involve a variety of other course activities since 10% of the course grade is at the discretion of the individual instructor.

Guidelines for Writing the 10% Project Report

- What was the research project that you participated in?
  {Antibiotic resistance (susceptibility) of environmental microbes}
- Describe the procedures and methods
- Sample Collection- location (where did you pick your sample from? Home/outside/kitchen/cafeteria/Gym etc.) How did you collect sample? (By swabbing/picking soil?)
- Growing bacteria (You streaked the swab on an TSA agar plate and incubated it for 24–48 h)
- Genomic DNA extraction by using a MoBio PowerSoil/Qiagen DNeasy PowerSoil kit (describe briefly)- 1–2 paragraphs
- Kirby-Baur assay for Antibiotic testing- which antibiotics did you test for? Which antibiotics was your bacterium found to be sensitive or resistant to?
- Analysis/viewing of genomic DNA or PCR amplified 16s rRNA product on Agarose gel by gel Electrophoresis
- What are your thoughts about the research project? (Interesting/Not Interesting/Hmm?)

Surveying Students' Attitudes Toward the CURE

Though we were not able to perform a formative assessment of the impact of integrating research experience into the course, students were surveyed for their attitudes toward and feedback about their UR experience using following questions/reflection pointers -

1. The Research project as UR experience helped me understand the course material better.
2. I think I can apply the learned knowledge to newer concepts.
3. After participating in the Research, I am able to comprehend my course material better.
4. How much did the research experience help you to integrate the course concepts in your learning process?
5. How much do you think the course materials were integrated into the research project?
6. How well do you think the course materials were integrated into the research project?
7. How did you like doing the research activities in hands-on form/in laboratory?
8. How did you like doing the research activities online, downloading information from other resources?

[1] https://www.genewiz.com/

[2] https://dnasubway.cyverse.org

9. Has your appreciation for science as it relates to everyday life increased?
10. Would you like to participate in a science research project in other classes at QCC?

RESULTS

Some of the bacterial species identified were not surprisingly those commonly found on human skin such as *Staphylococcus epidermidis, Staphylococcus aureus* and *Staphylococcus haemolyticus.* Other bacteria that were isolated included various strains of *Bacillus subtilis, Bacillus cereus* and *Escherichia coli.* Some novel species such as *Staphylococcus caprae, Bacillus circulans* were identified as well. Students were intrigued to observe that majority of isolates showed high resistance to many commonly used antibiotics such as penicillin, oxacillin and ampicillin (**Figure 1**). However, bacteria isolated from crowded places were not necessarily found to be more resistant to tested antibiotics (data not shown).

Though most students had never had any research experience, all the students in the class displayed mostly positive attitude toward participating in all types of research experiences. Most said they were able to comprehend the course material better, and integrate course concepts in learning process as the concepts were integrated well in the research project. Many liked doing the research activities in hands-on format in laboratory compared to research online or in the library. Their appreciation for science as it relates to everyday life has increased. Most reported that they would like to

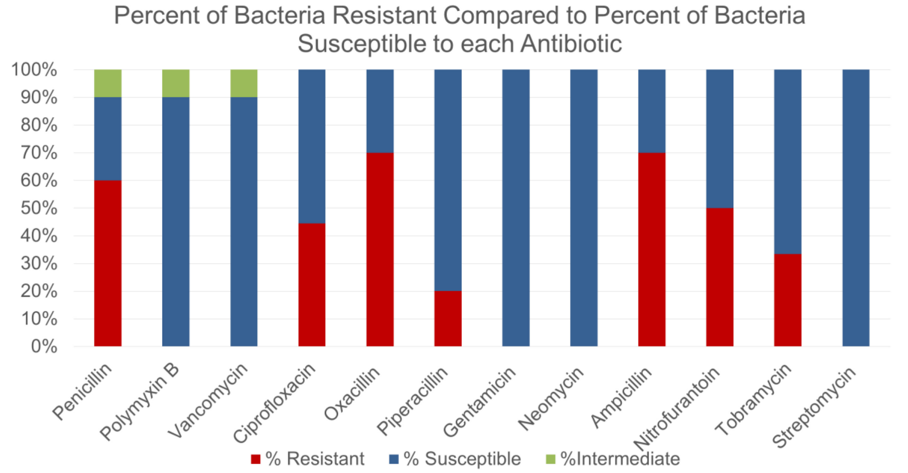


FIGURE 1 | Comparing the susceptibility or resistance of environmental isolates against antibiotics most commonly tested in Microbiology laboratory.

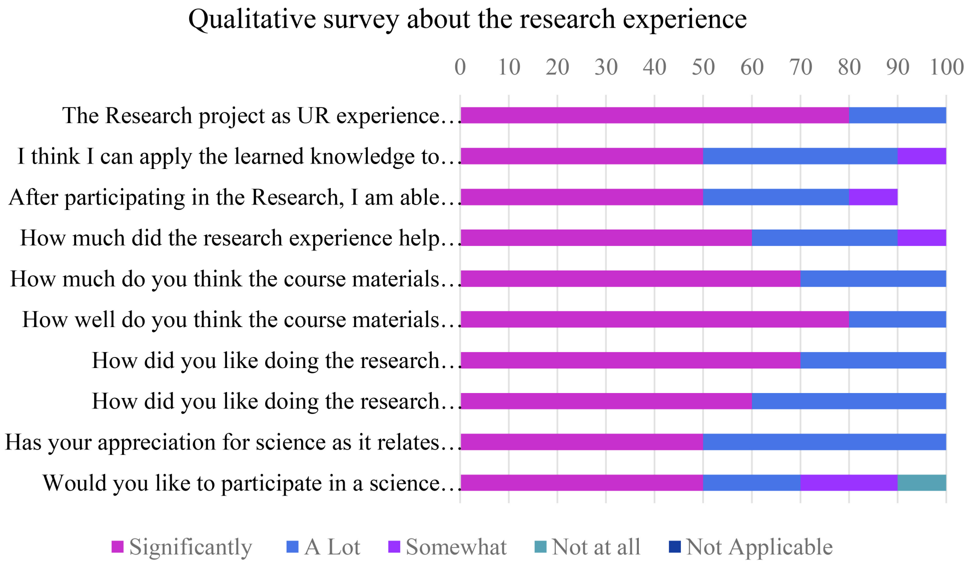


FIGURE 2 | Qualitative student response survey about the research in classroom experience.

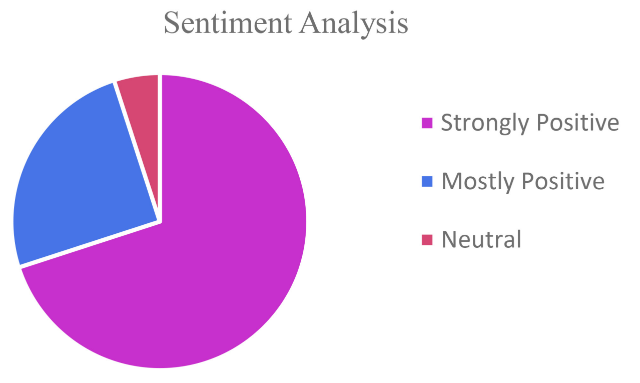


FIGURE 3 | Student Reflections: Sentiment analysis.

participate in a science research project in other classes at QCC (**Figure 2**).

Students performed all the laboratory procedures successfully including sample collection, streaking on media plates, isolation and culturing/growing bacteria from the environmental sample, DNA extraction from bacterial isolates, setting up PCR, performing Agarose gel electrophoresis and analyzing the 16s RNA sequence data to identify bacteria isolated from their environmental samples. They displayed increased engagement while learning the procedures and techniques as well as relevance of the research experience to real life situations as is evident from the student response survey (**Figure 2**) and the student reflections (**Figure 3** and **Supplementary Material Part II**).

Thus the research experience aligned well with the following course learning objectives.

1. Students will understand the general principles of Microbiology with practical emphasis on pathogenic microorganisms.
2. Students will develop the skills necessary to perform various microbiological laboratory procedures.
3. To create an incentive for further investigations in the field and to acquire sufficient background to understand the technical terminology in current publications.
4. To correlate the principles of Microbiology with the students' own interest and future as a health practitioner.

DISCUSSION

By integrating a research component directly into an existing Microbiology laboratory course, not just a select few, but ALL students in the class had the opportunity to participate in an inquiry-based real-world application of genomics in Microbiology experience. Incorporating the UR as high impact practice into a course that is required for allied-health career pathway, many students were successfully introduced to biology research concepts and practices, including DNA isolation, amplifying DNA using Polymerase Chain Reactions, DNA sequencing, and genomic/bioinformatics concepts. The vast majority of the students would have never been introduced to these practices had it not been incorporated into a required course. The survey results demonstrate an overwhelmingly positive response and experience for all of the students (**Figure 3**). The students enjoyed performing the research, recognized the applicability of it to their lives and future careers, and stated that the research experience was valuable. The UR experience helped students make a solid connection between what they learn in class and how it can be applied to the environment around them in real life. It made the students aware of the wide diversity of microbial species in their surroundings as well as introduced them to the technology in the fields of Microbiology and Biotechnology. The students who participated in the project reported significant gain in their knowledge and confidence. They expressed interest in pursuing STEM careers.

Nevertheless, we did face some challenges. There is always time constraint from the instructor point of view as we struggle to "cover" the course content. There is time constraint for students as they are juggling too many classes and work/family responsibilities. These hurdles are prominent especially in community college students. It is extremely challenging to motivate all of the students in a class.

CONCLUSION

Here we describe a model CURE that was successfully implemented in a biology lab course at an institution with minimal research infrastructure and limited funding resources. Though it is extremely challenging to incorporate a CURE in a community college science class, it has been a highly rewarding experience for students as we look at the student reflections. It has been a gratifying experience for the faculty as well. We think that this model of CURE can be successfully implemented in other Biology lab courses at other small and large schools alike without too much efforts.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by City University of New York IRB. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

MT is the principal and corresponding author on the submission and MW was a contributing author. MT conceived and conceptualized the idea, conducted, organized the research project, carried out the assessment, and wrote the manuscript. MW implemented the research project in her class sections and assisted in writing the results. Both authors contributed to the article and approved the submitted version.

FUNDING

MT is grateful for the funding support from the Research in the Classroom Idea grant award from City University of New York (CUNY) Office of research (award # RIC256), the administrative support from QCC's Office of Grants and Sponsored programs and Department of Biological Sciences and Geology.

ACKNOWLEDGMENTS

We thank Laura Rachiele and Terence Rohan for providing technical support and assistance with student laboratory equipment, materials and maintaining safe laboratory environment.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.578810/full#supplementary-material

REFERENCES

Allen, H. K., Donato, J., Wang, H. H., Cloud-Hansen, K. A., Davies, J., and Handelsman, J. (2010). Call of the wild: antibiotic resistances genes in natural environments. *Nat. Rev. Microbiol.* 8, 251–259. doi: 10.1038/nrmicro2312

American Association for the Advancement of Science (2011). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: AAAS.

Apedoe, X. S., Walker, S. E., and Reeves, T. C. (2006). Integrating inquiry-based learning into undergraduate geology. *J. Geosci. Educ.* 54, 414–421. doi: 10.5408/1089-9995-54.3.414

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., et al. (2014). Assessment of course-based undergraduate research experiences: a meeting report. *CBE Life Sci. Educ.* 13, 29–40. doi: 10.1187/cbe.14-01-0004

Ballen, C. J., Blum, J. E., Brownell, S., Hebert, S., Hewlett, J., Klein, J. R., et al. (2017). A call to develop course-based undergraduate research experiences (CUREs) for non-majors courses. *CBE Life Sci. Educ.* 16:mr2. doi: 10.1187/cbe.16-12-0352

Bangera, G., and Brownell, S. E. (2017). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci. Educ.* 13, 573–738.

Brownell, S. E., Hekmat-Scafe, D. S., Singla, V., Seawell, P. C., Conklin Imam, J. F., Eddy, S. L., et al. (2015). A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. *CBE Life Sci. Educ.* 14:ar21. doi: 10.1187/cbe.14-05-0092

Brownell, S. E., and Kloser, M. J. (2015). Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. *Stud. High. Educ.* 40, 525–544. doi: 10.1080/03075079.2015.1004234

Caplan, A. J., and MacLachlan, E. S. (2014). *CUR Tapping the Potential of All: Undergraduate Research at Community Colleges*, eds B. Cejda and N. Hensel (Washington, DC: Council on Undergraduate Research).

Centers for Disease Control and Prevention (2013). *Antibiotic Resistance Threats in the United States, 2013*. Atlanta, GA: Centers for Disease Control and Prevention.

Corwin, L. A., Graham, M. J., and Dolan, E. L. (2015a). Modeling course-based undergraduate research experiences: an agenda for future research and evaluation. *CBE Life Sci. Educ.* 14:es1. doi: 10.1187/cbe.14-10-0167

Corwin, L. A., Runyon, C., Robinson, A., and Dolan, E. L. (2015b). The laboratory course assessment survey: a tool to measure three dimensions of research-course design. *CBE Life Sci. Educ.* 14:ar37. doi: 10.1187/cbe.15-03-0073

Dolan, E. L. (2012). Next steps for vision and change: moving from setting the vision to change. *CBE Life Sci. Educ.* 11, 201–202. doi: 10.1187/cbe.12-06-0082

Genné-Bacon, E. A., and Bascom-Slack, C. A. (2018). The PARE project: a short course-based research project for national surveillance of antibiotic-resistant microbes in environmental samples. *J. Microbiol. Biol. Educ.* 19:19.3.97.

Handelsman, J., Beichner, R., Bruns, P. J., Chang, A., DeHaan, R., Gentile, J., et al. (2004). Scientific teaching. *Science* 304, 521–522.

Hunter, A., Laursen, S. L., and Seymour, E. (2007). Becoming a scientist: the role of undergraduate research in students' cognitive, personal, and professional development. *Sci. Educ.* 91, 36–74. doi: 10.1002/sce.20173

Kuh, G. D. (2008). *High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter*. Washington, DC: Association of American Colleges and Universities.

Lopatto, D. (2004). Survey of undergraduate research experiences (SURE): first findings. *Cell Biol. Educ.* 3, 270–277. doi: 10.1187/cbe.04-07-0045

Lopatto, D. (2007). Undergraduate research experiences support science career decisions and active learning. *CBE- Life Sci. Educ.* 6, 297–306. doi: 10.1187/cbe.07-06-0039

Lopatto, D., Harrison, M., Dunbar, D., Ratmansky, L., and Boyd, K. (2011). Classroom-based science research at the introductory level: changes in career choices and attitude. *CBE Life Sci. Educ.* 10, 279–286. doi: 10.1187/cbe.10-12-0151

Lopatto, D., and Tobias, S. (2010). *Science in Solution: The Impact of Undergraduate Research on Student Learning*. Washington, DC: Council on Undergraduate Research.

Lord, T. R. (2001). 101 reasons for using cooperative learning in biology teaching. *Am. Biol. Teach.* 63, 30–38. doi: 10.2307/4451027

Mader, C. M., Beck, C. W., Grillo, W. H., Hollowell, G. P., Hennington, B. S., Staub, N. L., et al. (2017). Multi-institutional, multidisciplinary study of the impact of course-based research experiences. *J. Microbiol. Biol. Educ.* 18:18.2.44.

National Academies of Sciences, Engineering, and Medicine (2015). *Integrating Discovery-Based Research into the Undergraduate Curriculum: Report of a Convocation*. Washington, DC: The National Academies Press.

National Research Council (2003). *BIO2010: Transforming Undergraduate Education for Future Research Biologists*. Washington, DC: National Academic Press.

Wei, C. A., and Woodin, T. (2011). Undergraduate research experiences in biology: alternatives to the apprenticeship model. *CBE Life Sci. Educ.* 10, 123–131. doi: 10.1187/cbe.11-03-0028

Woodin, T., Carter, C. V., and Fletcher, L. (2010). Vision and change in biology undergraduate education, a call for action—initial responses. *CBE Life Sci. Educ.* 9, 71–73. doi: 10.1187/cbe.10-03-0044

World Health Organization (2014). *Antimicrobial Resistance: Global Report on Surveillance*. Geneva: World Health Organization.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 29 October 2020
doi: 10.3389/fmicb.2020.528051

Teaching Microbiome Analysis: From Design to Computation Through Inquiry

Gail L. Rosen[1] and Penny Hammrich[2]*

[1] Ecological and Evolutionary Signal-processing and Informatics (EESI) Laboratory, Electrical and Computer Engineering, Drexel University, Philadelphia, PA, United States, [2] School of Education, Drexel University, Philadelphia, PA, United States

In this article, we present our three-class course sequence to educate students about microbiome analysis and metagenomics through experiential learning by taking them from inquiry to analysis of the microbiome: Molecular Ecology Lab, Bioinformatics, and Computational Microbiome Analysis. Students developed hypotheses, designed lab experiments, sequenced the DNA from microbiomes, learned basic python/R scripting, became proficient in at least one microbiome analysis software, and were able to analyze data generated from the microbiome experiments. While over 150 students (graduate and undergraduate) were impacted by the development of the series of courses, our assessment was only on undergraduate learning, where 45 students enrolled in at least one of the three courses and 4 students took all three. Students gained skills in bioinformatics through the courses, and several positive comments were received through surveys and private correspondence. Through a summative assessment, general trends show that students became more proficient in comparative genomic techniques and had positive attitudes toward their abilities to bridge biology and bioinformatics. While most students took individual or 2 of the courses, we show that pre- and post-surveys of these individual classes still showed progress toward learning objectives. It is expected that students trained will enter the workforce with skills needed to innovate in the biotechnology, health, and environmental industries. Students are trained to maximize impact and tackle real world problems in biology and medicine with their learned knowledge of data science and machine learning. The course materials for the new microbiome analysis course are available on Github: https://github.com/EESI/Comp_Metagenomics_resources.

Keywords: bioinformatics, microbiome, metagenomics, microbial ecology, multidisciplinary education

OPEN ACCESS

Edited by:
C. Titus Brown,
Michigan State University,
United States

Reviewed by:
Jonathan Badger,
National Cancer Institute (NCI),
United States
Karen Word,
The Carpentries, United States

****Correspondence:***
Gail L. Rosen
glr26@drexel.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *18 January 2020*
Accepted: *09 September 2020*
Published: *29 October 2020*

Citation:
Rosen GL and Hammrich P (2020) Teaching Microbiome Analysis: From Design to Computation Through Inquiry. Front. Microbiol. 11:528051. doi: 10.3389/fmicb.2020.528051

INTRODUCTION

In recent years, there has been a call for greater data literacy in life science education (Gibson and Mourad, 2018). Bioinformatics core competencies have been identified by various organizations. Competencies include a combination of biology, understanding of technologies, statistics, and computational methods in addition to teamwork, communication, and the scientific discovery process. Also, researchers have found that while learning the breadth of biology, computation, and math, it is important to start early and maintain depth and focus on a multidisciplinary topic (Anton Feenstra et al., 2018). Thus, it is concluded a series of courses, if not whole training program, is needed to effectively train students in bioinformatics. Also, an iterative teaching approach allows students to incorporate feedback, especially from multiple sources (e.g., biology and computation) (Marbach-Ad and Marr, 2018).

Metagenomics has been introduced in the undergraduate and graduate curriculums, but usually as a short course (Falana et al., 2015; Bolyen et al., 2019), research module in a larger course (Muth and McEntee, 2014; Gibbens et al., 2015; Lentz et al., 2017), or a single course (Edwards et al., 2013). Also, there is an issue of students from more biological disciplines and from more computational/engineering disciplines both gaining valuable knowledge from these courses.

To address some of these issues, we introduce three interdisciplinary courses to educate students in the realms of genomics, molecular evolution, and the bioinformatics analyses of genes and genomes. Students participating in these courses come from biology, biomedical engineering, electrical engineering, and computer science, providing a diverse multidisciplinary environment with great potential for peer learning. While developing hypotheses, students gain hands-on skills in DNA sample preparation and sequence analysis in the Molecular Ecology Laboratory and Bioinformatics courses. They analyze amplicon and metagenomic datasets that they helped to generate, using these to test hypotheses about microbial ecology, symbiosis, and the roles of microbes in nutrition and disease. Through the thematic activities, we actively engage students in the learning process, helping them to develop as critical-thinkers who understand the scientific method. The course sequence is complementary in its approaches, with the Molecular Ecology Lab being hypothesis generating and learning lab techniques, while the Bioinformatics course builds skills through a more traditional format, and the sequence finally culminates in the Computational Microbiome Analysis course where students share and learn about cutting-edge tools. Specifically, in the microbiome course, students conduct tutorials to learn cutting-edge tools by (1) independently following or composing tutorials, demonstrating what they learned, and sharing with the tutorial and results others, (2) learn from peers' tutorials, and (3) learn the steps to analyze their project data. We attempt to reach out to heterogeneous backgrounds by having students take a hands-on lab course (rather than bio theory), by teaching bioinformatic algorithms through demonstration, by teaching coding through example and debugging, and through group work in two of the courses. We are the first to broaden training in microbiome data analysis so that students gain deeper understanding from learning bioinformatics basics to more advanced analysis via inquiry. Quantitative assessments of knowledge gain of 45 undergraduate students showed that students generally improved knowledge in several bioinformatics areas.

THE STRUCTURE OF THE 3-COURSE SEQUENCE

Drexel university has 3 quarters (approximately 10 weeks each) per year. The course sequence is as follows: Molecular Ecology Lab and Bioinformatics are concurrently offered in the first quarter, followed by Computational Microbiome Analysis in a second quarter. Due to some life events, we offered the course sequence twice—once in the 2015–2016 and again in the 2016–2017 school years. In 2015–2016, the concurrent Molecular Ecology lab and Bioinformatics was offered in the Fall with the Computational Microbiome Analysis course in the Spring, while the second time, it was offered in the Fall/Winter. The specific learning objectives of each course are (1) Molecular Ecology: Proficiency in molecular lab techniques and knowledge of technologies, mastery of knowledge of computational analyses of ecology, and understand an application, methods, and synthesize hypotheses; (2) Bioinformatics: Be able to modify python code, introduced to bash scripting, learn algorithms such as dynamic programming, hidden Markov models, phylogenetics, and learn about their implementations (e.g., BLAST); and (3) Computational Microbiome Analysis: working knowledge of bioinformatics programming, proficiency in bioinformatics pipeline development, and learning how and when to use comparative genomics tools.

With the three courses, we were able to address 11 out of 16 core competencies identified by the Intl. Consortium for Systems Biology (ICSB) curriculum task force (Mulder et al., 2018) and 11 out of the 15 core competencies identified by Network for Integrating Bioinformatics into Life Sciences Education (NIBLSE) (Wilson Sayres et al., 2018). This course series teaches ICSB core competences—B: Depth in at least one area of biology, C: Biological data generation technologies, D: Details of the scientific discovery process and the role of bioinformatics in it, E: (at a high-level due to undergraduate curriculum): statistical research methods, F: bioinformatics tools and methods, G: ability of a computer-based system to meet scientific problem, J: Command line skills and scripting, K: Web-based Bioinformatics, L: Impacts of bioinformatics/genomics, N: (partial) communication of results to peers, and O: Effective Teamwork. We also address NIBLSE's core competencies: S1: Role of Bioinformatics in hypothesis-drive biology, S2: Bioinformatic computational concepts, S3: Statistics, S4: Accessing genomics, S5: Using genomic tools, S11 (partial through functional prediction module): Using pathway prediction tools using expression tools, S12: Metagenomics, S13: Scripting, S14: Using software packages, and S15: operate different computing environments. A summary of the core competencies targeted in each course are shown in **Figure 1**.

Molecular Ecology Lab

The Molecular Ecology Lab course (first quarter class in the sequence) was designed to train students in basic laboratory techniques and technologies from the field of molecular biology, applying these to enable research on microbial symbionts of animals. The course was also designed to emphasize the design of hypotheses and experiments using amplicon and meta-genomic/transcriptomic sequencing to ask questions about host-microbe interactions that are challenging to study in other ways. The timeline for the course project instructions is shown in **Figure 2**.

In this course, students were graded on: (1) two quizzes, which emphasized their understanding of methods/technologies

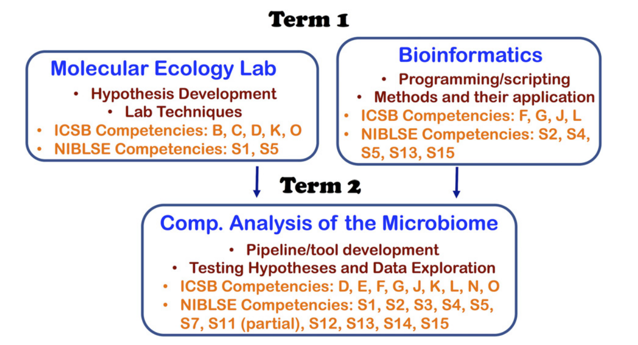

Term 1
Molecular Ecology Lab
• Hypothesis Development
• Lab Techniques
• ICSB Competencies: B, C, D, K, O
• NIBLSE Competencies: S1, S5
Bioinformatics
• Programming/scripting
• Methods and their application
• ICSB Competencies: F, G, J, L
• NIBLSE Competencies: S2, S4, S5, S13, S15
Term 2
Comp. Analysis of the Microbiome
• Pipeline/tool development
• Testing Hypotheses and Data Exploration
• ICSB Competencies: D, E, F, G, J, K, L, N, O
• NIBLSE Competencies: S1, S2, S3, S4, S5, S7, S11 (partial), S12, S13, S14, S15

FIGURE 1 | Each course in the sequence and its mapping to ICSB and NIBLSE competencies.

and situations in which to apply them; (2) course participation, which included a requirement that the students demonstrate competency in DNA extraction, PCR amplification, PCR primer design, and gel electrophoresis; (3) an 8 page paper in which they analyzed and reported data that they generated on a bacterial endosymbiont of ants, showing competency in DNA sequence alignments, BLAST searches, and phylogenetics; and (4) their 4–6 page microbiome analysis proposal. Skills emphasized in the class were, thus, not only related to lab techniques but also thinking like a scientist and analyzing and interpreting data.

Molecular Ecology Project Proposals

For the microbiome analysis proposal students submitted one outline and one rough draft, using instructor feedback to improve their ideas, hypotheses, justification, and methodologies. We focused on five research programs that were put forth as areas where the students could develop questions that they could then test through a follow-up course: (1) reciprocal impacts between non-alcoholic fatty liver disease and gut bacteria; (2) identifying function of ancient gut symbionts of predatory army ants; (3) studies of ant gut microbiome gene expression in response to dietary variation; (4) microbial source tracking in the Delaware River watershed; and (5) studies on bacteria co-colonizing bioreactors with algae.

Scientists from labs supporting these projects delivered 20–30 min presentations at the start of the course, helping to establish the "menu." They put forth knowns and unknowns for their systems, helping to make clear the motivations for study. For each presentation one or more articles from the primary research literature were assigned for background reading, helping students to develop further understanding of these subdisciplines.

Students were given some guidance in narrowing down the list of potential projects. As an example, see the below excerpt from the microbiome analysis proposal guidelines provided to the students:

"The best hypotheses will combine a mixture of novelty and realism, with clear links to mechanism as a guiding force or focus. For instance:

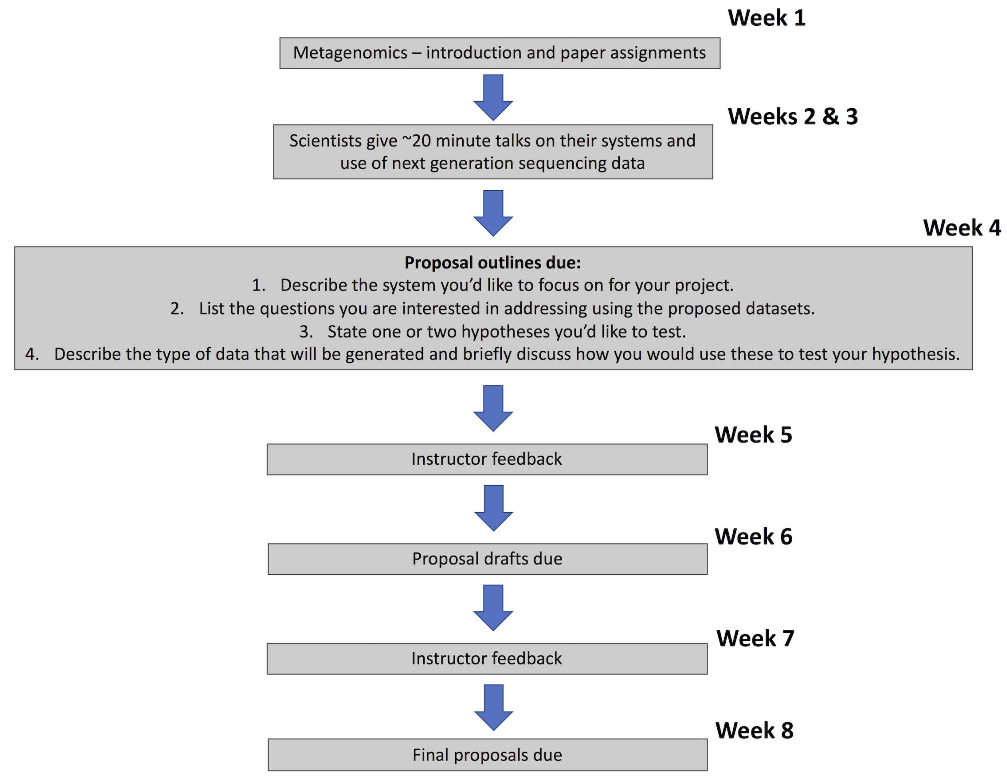

Week 1
Metagenomics – introduction and paper assignments
Weeks 2 & 3
Scientists give ~20 minute talks on their systems and use of next generation sequencing data
Week 4
Proposal outlines due:
1. Describe the system you'd like to focus on for your project.
2. List the questions you are interested in addressing using the proposed datasets.
3. State one or two hypotheses you'd like to test.
4. Describe the type of data that will be generated and briefly discuss how you would use these to test your hypothesis.
Week 5
Instructor feedback
Week 6
Proposal drafts due
Week 7
Instructor feedback
Week 8
Final proposals due

FIGURE 2 | Timeline of the molecular ecology lab projects.

1. For the *Cephalotes* transcriptome project (project 3), one might hypothesize particular genes and pathways that should show transcriptional responses to the various diets if bacteria do indeed use substances contained within. One might also hypothesize which organisms to be involved.
2. For the army ant project (project 2), one might hypothesize functions expected to be common among gut symbionts of carnivorous animals. One might also propose functions that should differ between closely related strains of bacteria hosted by sibling ants belonging to the same colonies".

While biologically-inquisitive students went through several rounds of hypothesis development with the instructor, those who were less-developed to choose hypotheses were given a specific problem with limited choices on hypotheses. Groups were encouraged to be heterogeneous, meaning that groups that contained at least bioscience and one engineering/comp student were encouraged for peer learning. All groups were required to submit a 4–6 page proposal draft that utilized metagenomics, metatranscriptomics, or 16S rRNA amplicon sequencing to study one of the potential projects presented in class. Students learned about the subject area through independent study and interaction with the instructors to learn more about these systems and techniques.

Examples of Specific Aims and hypotheses from undergraduate projects included:

Project 1

> "Hypothesis 1: Non-alcoholic fatty liver disease development will correlate with changes associated with increased short chain fatty acid production."
>
> "Hypothesis 2: Non-alcoholic steatohepatitis progression may correlate with endogenous alcohol production."

Projects 2 and 3 – One student combined two of the projects on the menus.

> "I predict that different amounts of Enzyme Commission numbers (E.C.s) associated with in (*sic*) digestion will be present in ants with different feeding types, as was found in Muegge et al. (2011). . .. enzymes used in amino acid synthesis will be more common in *Cephalotes* than army ants because of the nitrogen poor diets in *Cephalotes*" (*Student is using precedent from a prior publication and knowledge of ant biology to predict differences in the devotion of gut microbes to particular digestive processes.*)

Project 3

> "The main aim for this project is to find whether particular genes are highly expressed based on the diet. In this project, we'll analyze metabolic pathways that should show transcriptional responses to various diets."

Project 4

> "The primary objective of the study is to identify microbes present in the watershed that correspond to specific sources of fecal contamination for MST. To achieve this, fecal samples have been gathered from a variety of microbial hosts at different times of the year, and water samples have been collected upstream and downstream of the potential contamination sites."

Bioinformatics

While students engaged in the Molecular Ecology lab, students took Bioinformatics, which was co-taught by Dr. Rosen (Engineering) and Dr. Russell (Biology). Most of this course was developed prior to the grant, except for the first 2-week coding bootcamp. Previously, the course had lacked some of the more practical data wrangling and retrieval necessary to start in bioinformatics. So, for the grant, we introduced an intensive introduction to bash and Biopython (Cock et al., 2009). The first 2 weeks were a review of molecular evolution, and a "coding bootcamp" that was an introduction to Biopython and the bash environment/job queuing system on Proteus, Drexel's campus computing cluster (over 2000 CPU-cores offered to the campus community in 2014) (URCF, 2019). One of the programming assignments was to debug Biopython code to NCBI retrieve sequences, where intentional errors were introduced into the code that students had to correct. This exercise was specifically designed for the course and reinforced the idea that most bioinformatics programming is not coded from scratch, but that "related code" can often be found online (e.g., on a forum) and that it must be manipulated for specific solution to solve a specific problem. Subsequently to the coding bootcamp format, the biological goals and algorithmic foundations of dynamic programming/BLAST, hidden Markov models, phylogenetics, and sequence logos to represent DNA variation, were taught. Our lectures were structured so that the biological application and goals were laid out, followed by the computational and mathematical underpinnings of the algorithms. The course contains 3 homeworks, one midterm, and one final.

Computational Microbiome Analysis

Computational Microbiome Analysis (also listed as "Statistical Analysis of Genomics" to enroll a wider audience) is the flagship course developed for the project. The course generally teaches fundamentals in the first 3–4 weeks; first, there is a review of shell scripting, Biopython, and running code in a cluster queuing environment (overlap with Bioinformatics for students that repeat). Then, an introduction to the microbiome (including the significance of the 16S rRNA gene), microbial ecology, and metagenomics is introduced. Large-scale databases and meta-analysis programs for both amplicon sequencing and metagenomics datasets [like QIIME (Bolyen et al., 2019) and MEGAN (Bağcı et al., 2019)] are covered. These fundamentals are expected to get students comfortable with automating code and using third party software, with both being necessary for the individualized course projects. Students also sign up for one or two tutorials, in which they must learn a particular package/method in-depth and present a summary of how the method works and give an example of how to run the software and the output that one can expect. While undergraduates present on 1 tutorial and graduate students present on 2 tutorials in groups of 2–3, most of the quarter (6–7 weeks) is consumed by the 10–12 tutorials from groupings of all the students. Usually, the instructor gives a 30 min lecture to give background on the analysis theme for the week, such as "Metagenome assembly," which would explain the need and challenges of the area. Then, the rest of the week is 2 tutorials (usually 30 min in length on average) to talk about the algorithms and show how the various methods work, with added time for discussions. For our example theme week, this would include a review of IDBA-UD and Metaspades (depending on the year). The students work on instructor-selected datasets to demonstrate the tools in their tutorials and compare metrics, such as N50/min and max contig lengths for our example theme week. The students use online materials about the associated tools to develop the 10–15 min algorithm discussion followed by a 15–20 min tutorial demonstration. While a few groups do take the class through a real-time tutorial, usually 15–20 min is not enough, and the students, who are teaching, usually point the students, who are learning, to a Github repository where they can view and run the code themselves. This course focus on tutorials of important microbiome analysis tools allows the course to update itself and keep up with the quickly-moving field of microbial community analysis. Tutorials have included High-throughput Phylogenetics [using alignment and tree methods on CIPRES (Miller et al., 2012), learning microbial ecology comparison techniques (like diversity metrics, distance measures between samples like Unifrac (Lozupone and Knight, 2005) etc., ordination, etc.], assembly and binning of genomes from metagenomics, taxonomic identification from metagenomics, functional annotation of metagenomes, functional prediction of amplicon data, metatranscriptomic analysis (differential abundance comparisons), and even basic statistics (like ANOVA/MANOVA/correction for multiple comparisons) and analysis like gene set enrichment analysis. The tools that are reviewed can change from course iteration to course iteration. For example, tutorials on taxonomic classification methods went from Metaphlan2 (Segata et al., 2012) in the first year to Kraken2 (Wood et al., 2019) and Kaiju (Menzel et al., 2016) in the latest iteration.

The course projects are the most important aspect of this course. Students who take the Molecular Ecology lab will analyze a dataset that they set out to investigate to verify a hypothesis. Students, who did not take the Molecular Ecology lab, can choose from a menu of datasets and project ideas, some of which may be investigating algorithms and comparing methods (which appeal to the engineering and computer science students in the course.) Students received detailed guidance from the PIs and teaching assistants (TAs). Also, we made a concerted effort to pair graduate students with undergraduates, so that each team had a balance of levels. Projects titles include (results and project findings can be found on the course Github page):

1. "*Metatranscriptomic Analysis of Laboratory-reared Cephalotes varians RNA Dataset and Comparison across Four Dietary Treatments*.
2. "*Metagenomic analysis of and comparison between the photosynthetic microbial communities in two photobioreactors*".
3. "*A Metagenomic Analysis of Healthy Mice vs. Fatty Liver Disease Induced Mice on Both Control and High Fat Diets*".
4. "*Finding Patterns in Time-course Metagenomic Data*".
5. "*Metagenomic Analysis of Army Ant Guts*".
6. "*Building Ensembles of Taxonomic Classifiers*".

TABLE 1 | Student self-reported knowledge and skills (*n* = 45).

Level of skill	No skill	Somewhat skilled	Very skilled
Genetics	31%	49%	20%
Ecology	51%	38%	11%
Bioinformatics	51%	42%	7%
Metagenomics	80%	13%	7%
Hypothesis development	31%	47%	22%
Experimental design	16%	51%	33%
Programming	18%	53%	29%

Each week, students had to compose quiz questions (with corresponding answers), which we found acted as a formative assessment, to understand what students were absorbing from the lectures and tutorials since this forced students reflect on the material in weekly intervals. Undergraduate students learn one tool in-depth by teaching a tutorial, and finally, most of the skills are learned from a data analysis project. In order to keep this projects on-track, we have learned that students need to submit a project declaration, proposal, progress report, and final report throughout the short 10-week quarter.

PROJECT OUTCOMES

A total of ∼150 students enrolled in all three courses for the two offerings. However, we performed formative and summative instruments (a demographic questionnaire, de-identified but non-blind comparison of pre- and post-surveys; and bi-weekly administered surveys) only for the undergraduates. The surveys were administered under instruments approved under Drexel IRB #1211001675, and we obtained student consent at the beginning of each course. *Forty*-five undergraduates enrolled in at least one of the three courses, with 4 taking all three (there were substantially more graduate students that took all 3 courses). We surveyed demographics of the 45 undergraduates that took at least one of the courses, with 62% of them identified as male and 6% identifying with an ethnic group that was not Caucasian or Asian.

From a pre-course survey, students were asked to rate their abilities/skills of different subjects. In **Table 1**, Most students rated themselves with no skills in metagenomics, bioinformatics, genetics, and hypothesis development. This has identified that focusing the course on such skills is much needed.

Reflections From the Molecular Ecology Lab

From the Molecular Ecology Lab course, we generated four new next-generation sequencing datasets. These were presented to students in the Computational Microbiome Analysis follow-up course, a class whose roster included several students who participated in the lab.

Beyond serving as a prelude to the Computational Microbiome Analysis course, and an introduction to how the ‘omics revolution has revolutionized microbiology, the microbiome analysis proposal served to allow students to “demonstrate a capacity to synthesize and integrate results into the broader context of the field,” an objective from the course syllabus (all syllabi can be seen in the **Supplementary Material**). Through in class discussions, rough draft feedback, it was clear that students were able to do this to some extent. While some strongly mimicked documents disseminated from the scientists leading these projects, others demonstrated a strong vocabulary and independent thinking in areas they had not previously studied.

Through assessments of student quizzes and papers, it was clear that all developed a deeper understanding of microbial ecology and the applications of DNA/RNA sequencing to study microbes in their natural habitats. Several showed clear proficiency in developing well-justified hypotheses and aims. At minimum, all were able to develop a coherent and reasonable set of research activities.

Challenges included the fact that students often deviated from directives to limit their proposed work to suit the available/pending datasets. This meant that for those moving on to the subsequent Computational Microbiome Analysis course, several could not directly test their hypotheses.

Another challenge was the very steep learning curve required for students to develop a good understanding of bacterial metabolism. This was key to formulating strong hypotheses for several of the projects and more time devoted to this area during the course would have been immensely helpful.

Reflections From Bioinformatics

The Bioinformatics class was the most standard class of the three, with homeworks and tests. The biology students found the coding challenging but rewarding, with the statement “*. . .coding activities most difficult to understand but most rewarding” and “. . .use of NCBI was great.”* Others wanted to see more coding and did not want the theory behind the algorithms – “*I expected to learn more practical skills that I can use such as a script to sequence alignments but this course taught a lot about background theory of these algorithms.*”

Many students were satisfied with the course – *“The fusion of disciplines is readily apparent”, “This course is more hybrid than all other engineering science courses I‘m taken. Requiring understanding of two fields to apply them in bioinformatics”.* There was a trend that students with backgrounds in biology found programming part challenging and the students with programming background found biology challenging.

Reflections From Computational Microbiome Analysis

In the computational microbiome analysis course, students learned about state-of-the-art methods and tools used for microbiome and metagenomic analyses through hands-on tutorials and projects. Because each tool could possibly elicit a few weeks to itself alone, it is perceived that too much is covered in the class. We required that each student group spend half of a 30 min slot on describing how the method/tool works and half the time showing how to operate the tool and interpret its results. We did notice that computational students seemed to spend more time on the methods while biological students spent more time on results interpretation, which is to be expected. The hope is that the tutorial will give a basic introduction to the students, so that they can be aware of its existence in the vast toolbox of microbiome analysis to reference and learn more in-depth when needed.

The tutorials, each learned in-depth by a few students, were reinforced to the rest of the class through reflection – students were required to hand in 3 mock quiz questions and answers, some of which would be selected (or reshaped into more cohesive questions) for a quiz given the following week. The weekly quizzes were a good mechanism, as it induced a “studying for the quiz” reinforcement of the material. In our second iteration of the three-course sequence, we limited quiz content to conceptual understanding of the tool’s purpose and interpretation of their function. This way, students could focus their studying and understand the fundamental concepts of each week’s theme.

While students are excited by no tests or finals, they soon realize the curse of a project-based course, as it is 50% of their grade. As with all projects, students struggle to maintain a schedule, so we have found that 10-week project-based classes need multiple hard deadlines throughout the course to keep students on track. Having four deadlines is perfect. The “Project declaration” (due in week 2) is where the students must decide which topic they are interested in and demonstrate that they can gather the data. Demonstrating that students can import data structures and objects is pivotal, as we have found that many groups delay actually working with the data. Then the “Project proposal” (due in the week 5) must (1) describe the problem they are interested in (they would be able to take this hypothesis development directly from the Molecular Ecology Lab if enrolled in this class prior or if not, detail their hypothesis or design idea) and (2) propose the analysis steps and timeline of how they will test their hypothesis or build a tool. Then, the “Progress Report” (week 7) gives a deadline that students must report on some analysis steps, any issues encountered, and gives them the final chance to modify their proposed analysis design. Around week 10, students must give an oral presentation on their final results, and the following week, a written report is due. These spaced deadlines keep students thinking and working on the project in a timely manner.

Many undergraduates find that the freedom from tests and finals is more challenging than they expect, because they must now “get things to work” and peruse literature to understand concepts and tools. Varying quality of the tutorials and projects result. However, instead of teaching and testing on methods that are in constant flux, the focus is software pipeline design to test hypotheses or make tools, which builds critical thinking. Some students realize that this course helps build skills needed in the workforce. A spontaneous email that was received approximately 6 months after the Computational Microbiome Analysis course by a graduate student, who went on to work in the pharmaceutical industry, wrote:

> *"Dr. Rosen,*
> *I would like to thank you in the strongest possible terms for your course in the Spring term of '15: ECES 690.*
> *Without a doubt it is the single most applicable course I have taken, not only at Drexel, but in my entire academic career, to my current endeavors.*
> *At the time I expected it to be useful, but now I am discovering that the lessons learned there are *completely indispensable* to my occupation.*
> *I encourage you to keep up the amazing work with that class, and more like it, so that a new class of students can benefit from such instruction as I have had".*

Assessment of Learning Outcomes

We can show that bioinformatic competencies generally improved upon completion of any of courses in the three-course sequence. Pre- and Post- surveys of the Bioinformatics and Computational Microbiome Analysis classes included 20 content questions; the full list of questions can be found in the **Supplementary Material**. Quantitative data was collected by using a pre- and post-survey that was administered at the beginning and the end of the course and were coded so only the evaluator knew the identities. The questionnaire consisted of 20 open-ended content questions (seen in questions.docx in the **Supplementary Material**) on the microbiome, metagenomics and molecular ecology. The student responses in both the pre- and post- surveys were graded independently by two subject matter experts on a scale 1–5, with 1 meaning that the student demonstrated no knowledge and 5 meaning that the student demonstrated excellent mastery of the material. The pre- and post- surveys were collected from the 45 undergraduate students who agreed to participate in the study with 12 pre- and post-matched surveys that were near-completely filled out (due to student absences or incomplete surveys on either end since the surveys were lengthy). There were 7 questions that received more than 10 responses on the pre- and post- surveys and were statistically significant (as determined by a 2-tailed T-test). Other questions either received less than or equal to 10 responses or they were not significant (meaning that there was no statistical difference between the pre- or post-survey answers). The content questions that were statistically significant are:

2. What is a Standard Flowgram File and what type of DNA sequencer outputs it?
3. How would you convert a SFF file to a FASTA file?
4. What is the difference between PCA (Principal components analysis) and PCoA (Principal coordinates analysis)?
5. What are the trade-offs of supervised learning algorithms (trade-off of random forests vs. support vector machines vs. bayes classifiers)?
9. Genome sizes for a given species or taxon vary, often considerably. Describe why metatranscriptomic reads need to be normalized, especially for downstream analysis.
14. Name at least two ways that you can annotate WGS (whole-genome shot sequencing) reads with functional annotation?
15. Describe the difference between phylogenetic tree reconstruction methods?

As seen in **Figure 3**, Question 5 (about machine learning algorithms learned in Comp. Microbiome Analysis) has the biggest increase in understanding. Questions 2, 3, and 15 were learned in Bioinformatics, and Questions 4, 5, 9, 14, and 15 were learned in Computational Microbiome Analysis (note that question 15 was taught in both classes). Students completed the lab assignments, proposal report, computational assignments, tutorial demonstrations, and project demonstrations that meet the criteria in **Figure 1**. Students gained knowledge of wet lab and programming techniques, although proficiency was lacking for students from the opposite discipline, and this was a challenge. However, most students gained an appreciation for algorithms through hands-on calculations and learning how to use a tool through tutorials. Finally, microbiome analysis skills through group projects were facilitated through peer learning, and students gained at least some skills/knowledge that they did not have before. This demonstrates that knowledge of bioinformatics and metagenomics analysis increased for some topics. We believe that knowledge increased for other questions, but the sample size was too small (due to content question changes and not as many students answered those questions).

We have also included a qualitative report on student perceptions, experiences, and understandings (seen in the Evaluator_report.pdf in the **Supplementary Material**) that can elucidate more detail on how the learning outcomes were realized by the students.

DISCUSSION

We describe a 3-course sequence in microbiome analysis training via a Molecular Ecology Lab, Bioinformatics, and Computational Microbiome Analysis. A summative analysis and student feedback demonstrate that the course sequence and individual courses had some beneficial impact on student bioinformatic competencies. In a world where data is becoming ever abundant, students need to be equipped with the knowledge to handle it. Our training sequence helps to meet those training goals. Yet, there is still the challenge of

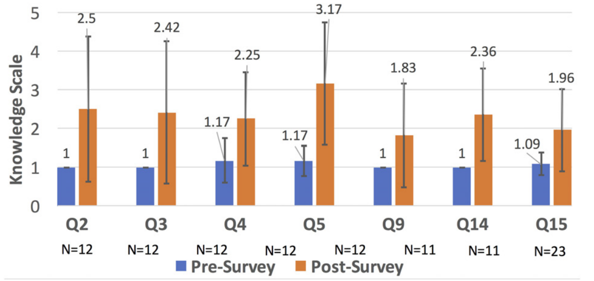


FIGURE 3 | Bar chart comparison of the knowledge scale for different bioinformatic topics (that were statistically significant). In around 7 areas (many related to microbiome analysis), there was improved knowledge. Other areas, see **Supplementary Material**, were not noticeably improved due to removal because of curricular changes, lack of enough responses, or no significance between the pre- and post- surveys.

educating students from heterogeneous backgrounds (biology and computation/engineering), so that students can (1) come to a level playing field or (2) speak each other's languages to work together and learn from each other. Future work may involve iterative differentiated coursework, adding more peer learning to the bioinformatics class, offering short courses (or bootcamps) to facilitate interdisciplinary communication for peer learning (computational students to get up to speed on biology and biology students to improve their programming).

Training in an emerging multidisciplinary field, that has great potential, importance, and need, has both its advantages and challenges. We have found that students who have bioinformatic skills and understand the domain science are urgently needed in the workforce. We encourage faculty and administration at universities to look past immediate barriers (such as financial constraints and/or politics) and foster interdisciplinary teaching and courses. When successful, we can train a new generation of scientists and engineers who will push the boundaries of discovery.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author. The Computational Microbiome Analysis course materials developed plus student projects and tutorials can be found at: https://github.com/EESI/Comp_Metagenomics_resources.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Drexel University Institutional Review Board under project #1211001675.

AUTHOR CONTRIBUTIONS

GR formulated the concept of the three-course sequence, designed and offered the courses, and wrote most of the manuscript. PH conducted and summarized the summative student findings. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by the NSF grant #1245632. This grant utilized computing support of Proteus through Drexel's University Research Computing Facility.

ACKNOWLEDGMENTS

Many modules and guest lectures were authored and taught by graduate students who were involved in the project and performed as teaching assistants for the classes. We would like to thank: Gregory Ditzler, Yemin Lan, Erin Reichenberger, Stephen Woloszynek (https://sw1.github.io/teaching/), and Zhengqiao Zhao for their ideas, guidance of projects, and education of the next-generation of bioinformatics researchers. We thank Dr. Jacob Russell for his involvement in the Molecular Ecology lab and Bioinformatics courses and his contributions toward sections "Molecular Ecology Lab" and "Reflections from the Molecular Ecology Lab".

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.528051/full#supplementary-material

Supplementary Syllabus 1 | Molecular ecology lab syllabus.

Supplementary Syllabus 2 | Bioinformatics syllabus.

Supplementary Syllabus 3 | Computational microbiome analysis syllabus (sometimes called Statistical Analysis of Genomics).

Supplementary Questions | The questions that were a part of the IRB process at the beginning of the project. Some questions that assessed outdated topics were omitted. Also, questions that did not have a statistically significant difference between groups were omitted.

Supplementary Evaluator Report | A qualitative report of formative assessments of the classes.

REFERENCES

Anton Feenstra, K., Abeln, S., Westerhuis, J. A., Brancos Dos Santos, F., Molenaar, D., Teusink, B., et al. (2018). Training for translation between disciplines: a philosophy for life and data sciences curricula. *Bioinformatics* 34:i4–i12.

Bağcı, C., Beier, S., Górska, A., and Huson, D. H. (2019). Introduction to the analysis of environmental sequences: metagenomics with MEGAN. *Evol. Genom.* 856, 415–429. doi: 10.1007/978-1-61779-585-5_17

Bolyen, E., Ram Rideout, J., Dillion, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25, 1422–1423. doi: 10.1093/bioinformatics/btp163

Edwards, R. A., Haggerty, J. M., Cassman, N., Busch, J. C., Aguinaldo, K., Chinta, S., et al. (2013). Microbes, metagenomes and marine mammals: enabling the next generation of scientist to enter the genomic era. *BMC Genomics* 14:600. doi: 10.1186/1471-2164-14-600

Falana, K., Knight, R., Martin, C. R., Goldszmid, R., Greathouse, K. L., Gere, J., et al. (2015). Short course in the microbiome. *J. Circ. Biomark.* 4:8. doi: 10.5772/61257

Gibbens, B. B., Scott, C. L., Hoff, C. D., and Schottel, J. L. (2015). Exploring metagenomics in the laboratory of an introductory biology course. *J. Microbiol. Biol. Educ.* 16, 34–40. doi: 10.1128/jmbe.v16i1.780

Gibson, J. P., and Mourad, T. (2018). The growing importance of data literacy in life science education. *Am. J. Bot.* 105, 1953–1956. doi: 10.1002/ajb2.1195

Lentz, T. B., Ott, L. E., Robertson, S. D., Windsor, S. C., Kelley, J. B., Wollenberg, M. S., et al. (2017). Unique down to our microbes—assessment of an inquiry-based metagenomics activity. *J. Microbiol. Biol. Educ.* 18:18.2.33.

Lozupone, C., and Knight, R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl. Environ. Microbiol.* 71:8228. doi: 10.1128/aem.71.12.8228-8235.2005

Marbach-Ad, G., and Marr, J. (2018). Enhancing graduate students' ability to conduct and communicate research through an interdisciplinary lens. *J. Microbiol. Biol. Educ.* 19:19.3.104.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nat. Commun.* 7:11257.

Miller, M. A., Pfeiffer, W., and Schwartz, T. (2012). "The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources," in *Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the eXtreme to the Campus and Beyond*, New York, NY.

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., et al. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* 20, 970–974. doi: 10.1126/science.1198719

Mulder, N., Schwartz, R., Brazas, M. D., Brooksbank, C., Gaeta, B., Morgan, S. L., et al. (2018). The development and application of bioinformatics core competencies to improve bioinformatics training and education. *PLoS Comput. Biol.* 14:e1005772. doi: 10.1371/journal.pcbi.1005772

Muth, T. R., and McEntee, C. M. (2014). Undergraduate urban metagenomics research module. *J. Microbiol. Biol. Educ.* 15, 38–40. doi: 10.1128/jmbe.v15i1.645

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., and Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* 9, 811–814. doi: 10.1038/nmeth.2066

URCF (2019). *Proteus Cluster @ Drexel URCF*. Available online at: https://drexel.edu/research/urcf/services/cluster/ (accessed December 31, 2019).

Wilson Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS One* 13:e0196878. doi: 10.1371/journal.pcbi.0196878

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol.* 20:257.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Microbiology

METHODS
published: 07 December 2020
doi: 10.3389/fmicb.2020.577497

An Educational Bioinformatics Project to Improve Genome Annotation

*Zoie Amatore[1], Susan Gunn[2] and Laura K. Harris[1]**

[1] Science Department, Harris Interdisciplinary Research, Davenport University, Lansing, MI, United States, [2] College of Urban Education, Davenport University, Grand Rapids, MI, United States

Scientific advancement is hindered without proper genome annotation because biologists lack a complete understanding of cellular protein functions. In bacterial cells, hypothetical proteins (HPs) are open reading frames with unknown functions. HPs result from either an outdated database or insufficient experimental evidence (*i.e.*, indeterminate annotation). While automated annotation reviews help keep genome annotation up to date, often manual reviews are needed to verify proper annotation. Students can provide the manual review necessary to improve genome annotation. This paper outlines an innovative classroom project that determines if HPs have outdated or indeterminate annotation. The Hypothetical Protein Characterization Project uses multiple well-documented, freely available, web-based, bioinformatics resources that analyze an amino acid sequence to (1) detect sequence similarities to other proteins, (2) identify domains, (3) predict tertiary structure including active site characterization and potential binding ligands, and (4) determine cellular location. Enough evidence can be generated from these analyses to support re-annotation of HPs or prioritize HPs for experimental examinations such as structural determination via X-ray crystallography. Additionally, this paper details several approaches for selecting HPs to characterize using the Hypothetical Protein Characterization Project. These approaches include student- and instructor-directed random selection, selection using differential gene expression from mRNA expression data, and selection based on phylogenetic relations. This paper also provides additional resources to support instructional use of the Hypothetical Protein Characterization Project, such as example assignment instructions with grading rubrics, links to training videos in YouTube, and several step-by-step example projects to demonstrate and interpret the range of achievable results that students might encounter. Educational use of the Hypothetical Protein Characterization Project provides students with an opportunity to learn and apply knowledge of bioinformatic programs to address scientific questions. The project is highly customizable in that HP selection and analysis can be specifically formulated based on the scope and purpose of each student's investigations. Programs used for HP analysis can be easily adapted to course learning objectives. The project can be used in both online and in-seat instruction for a wide variety of undergraduate and graduate classes as well as undergraduate capstone, honor's, and experiential learning projects.

Keywords: bioinformatics, hypothetical protein, genome annotation, education, classroom, undergraduate

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
João Marcelo Pereira Alves,
University of São Paulo, Brazil
Luis Carlos Guimarães,
Federal University of Pará, Brazil

****Correspondence:***
Laura K. Harris
laura.harris@davenport.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *30 June 2020*
Accepted: *27 October 2020*
Published: *07 December 2020*

Citation:
Amatore Z, Gunn S and Harris LK (2020) An Educational Bioinformatics Project to Improve Genome Annotation. Front. Microbiol. 11:577497. doi: 10.3389/fmicb.2020.577497

INTRODUCTION

Nucleic acid sequencing has become so inexpensive that researchers are generating a plethora of fully sequenced genomes annually through massive initiatives such as the Earth BioGenome Project which aims to sequence the genomes of 1.5 million eukaryotic species by 2050 (Yandell and Ence, 2012; Lewin et al., 2018). Once a genome sequence is determined, it must be annotated to identify the locations and functions of genes (Koonin and Galperin, 2003). In bacteria, the first step in genome annotation is identifying open reading frames (ORFs). An ORF is a continuous stretch of DNA that begins with a start codon and ends at a stop codon and has the proper number of nucleotides to potentially encode a functional protein (Brown, 2002). Due to the lack of introns and exons in bacterial genes, an ORF is usually synonymous with a gene in bacteriology. The amino acid (*i.e.*, primary protein) sequence for each ORF is used to search several databases to predict gene function. These databases include (1) sequence databases to identify sequence similarities with established sequences, (2) domain databases to detect conserved domains, (3) genome-oriented databases for identification of orthologous relationships for refined functional prediction, and/or (4) metabolic databases for metabolic pathway reconstruction (Koonin and Galperin, 2003). From these data, a public knowledgebase record for each ORF is generated which typically includes nucleic acid and amino acid sequences, gene and protein sizes, any identified domains, and a predicted function. The record is easily retrievable via a unique identifier (*i.e.*, locus tag) which is consistently used across knowledgebases (Brown et al., 2015; Tatusova et al., 2016; Coordinators, 2018). These public records are used for a wide variety of gene analyses, such as pathway enrichment, so having proper genome annotation is important to draw accurate and complete scientific conclusions (Goad and Harris, 2018; Smits, 2019).

Unfortunately, many genomes have a substantial number (up to 70%) of hypothetical proteins (HPs), which are ORFs with unknown functions (Sivashankari and Shanmughavel, 2006; Mohan and Venugopal, 2012; Bharat Siva Varma et al., 2015; Ijaq et al., 2015; Islam et al., 2015; School et al., 2016). Reports estimated that around 33% of National Center for Biotechnology Information (NCBI) knowledgebase sequences in 2006 were HPs (Kolker et al., 2004; Sivashankari and Shanmughavel, 2006; Omeershffudin and Kumar, 2019). While the exact number of HPs in today's NCBI is unknown, recent papers on *Mycobacterium tuberculosis* and *Exiguobacterium antarcticum* strain B7 genomes report around 27% HPs (da Costa et al., 2018; Yang et al., 2019) with 16% HPs in *Shigella flexneri* (Gazi et al., 2018). Assuming 20% of the current 218,642,238 GenBank sequences are HPs, over 43 million proteins need proper annotation, and this number continues to grow exponentially as sequences continue to be deposited. A hypothetical protein (HP) can be the result of either outdated or indeterminate annotation. Outdated HPs result from an out-of-date knowledgebase. Older genomes are more likely to have outdated HPs since experimental work to determine function of HPs is ongoing and annotation for older genomes was completed prior to the characterization of a similar sequence with known function. Automated and manual curation of public knowledgebases is needed to improve genome annotation and identify sequences with out-of-date annotation. For example, function was successfully attributed to approximately 17% of HPs in *E. antarcticum* strain B7 using computational methods (da Costa et al., 2018). If computational approaches can re-annotate just 10% of current HPs, then annotation will be improved for over 4 million proteins, which would substantially improve public knowledgebases overall. Alternatively, indeterminate annotation is the result of true HPs whose amino acid sequence has low similarity to proteins with known function. Experimental work is needed to properly annotate true HPs and improve genome annotation, but once completed manual inspection is needed to further discover, analyze, and correct erroneous annotation.

Several previously reported studies have used computational approaches to assign functional annotation to HPs in a wide range of bacterial and viral species, including but not limited to *Staphylococcus aureus* (Mohan and Venugopal, 2012; School et al., 2016), *M. tuberculosis* (Raj et al., 2017; Yang et al., 2019), *Vibrio cholerae* (Islam et al., 2015), *Klebsiella pneumoniae* (Pranavathiyani et al., 2020), *Mycoplasma pneumoniae* (Shahbaaz et al., 2015), *Orientia tsutsugamushi* (Imam et al., 2019), *Corynebacterium pseudotuberculosis* (Araujo et al., 2020), human adenovirus (Dorden and Mahadevan, 2015; Naveed et al., 2017), and vaccinia virus (Mahmood et al., 2016). These studies utilize some combination of the various computational tools and databases available to analyze the physiochemical, functional, and structural properties of an HP (**Table 1**) since results generated from a single server cannot provide a complete functional determination currently (Dorden and Mahadevan, 2015). While these computational resources are continually changing, due to their wide application in research it would be beneficial for undergraduate microbiology students to be familiar using some of the more enduring and commonly referenced resources. Therefore, this paper introduces a Hypothetical Protein Characterization Project based off commonly referenced resources in previously reported *in silico* HP characterization studies that students use while learning interdisciplinary concepts in bioinformatics, microbiology, biochemistry, and genetics (**Figure 1**). This educational, inquiry-based bioinformatics project familiarizes students with multiple free web-accessible programs that identify and predict HP characteristics, such as sequence similarities to other proteins, protein domains, tertiary (*i.e.*, 3D) protein structure, ligand binding partners, and cellular location. Critical thinking skills applied by the student to results obtained from the Hypothetical Protein Characterization Project are used to determine whether an HP has outdated or indeterminate annotation. This determination can be useful for improving public knowledgebase annotation and prioritizing experimental examination of true HPs.

HYPOTHETICAL PROTEIN SELECTION

The first step in the Hypothetical Protein Characterization Project is the selection of HPs to be characterized. This section

TABLE 1 | Example studies considered in the development of the Hypothetical Protein Characterization Project.

Species	Citation	No. HPs	Resources Used
Staphylococcus aureus	Mohan and Venugopal, 2012	10	CDD-BLAST, Pfam, PS^2, STRING, QFinder, ExPASy ProtParam, SOSUI, DISULFIND
	School et al., 2016	35	PSI-BLAST, ExPASy ProtParam, CDD-BLAST, Pfam, PS^2, 3DLigandSite, STITCH, STRING, PSORTb, SOSUI, DISULFIND
Mycoplasma pneumoniae	Shahbaaz et al., 2015	204 (41%)	BLAST, FASTA, HMMER, SBASE, CATH, SUPERFAMILY, InterPro, SYSTERS, CDART, SMART, GPCRpred, Discovery Studio, STITCH, STRING, iPfam, ExPASy ProtParam, PSORTb, PSLpred, LOCTree3, TMHMM, HMMTOP, SignalP 4.1, SecretomeP, VirulentPred, DBETH server
Mycobacterium tuberculosis	Raj et al., 2017	1055 (55%)	BLASTP, ExPASy ProtParam, PSORTb, CELLO, TMHMM, SignalP 4.1, HHPred, HMMSCAN, Pfam, InterPro, SUPERFAMILY, VirulentPred, VICMPred
Klebsiella pneumoniae	Pranavathiyani et al., 2020	540	InterPro, Pfam, BLASTP, CELLO2GO, GO FEAT, STRING, ExPASy ProtParam, VICMpred, MP3, I-TASSER
Corynebacterium pseudotuberculosis	Araujo et al., 2020	172 (47%)	GO FEAT, Pfam, CATH, SUPERFAMILY, VICMPred, CDART, CDD-BLAST, ExPASy ProtParam, PSORTb, TopHat, Gipsy, VirlentPred, STRING, PSIPRED, Modeler
Vibrio cholerae	Islam et al., 2015	6	CDD-BLAST, Pfam, PS^2, STRING, QFinder, ExPASy ProtParam, PSORTb, DISULFIND
Orientia tsutsugamushi	Imam et al., 2019	344	BLASTP, ExPASy ProtParam, PSLpred, CELLO, ScanProsite, SMART, Motif Scan, PFP-FunDSeqE, VirulentPred, PFP, Argot2, PSIPred, Modeler
Vaccinia virus	Mahmood et al., 2016	1 (100%)	BLAST, GOR IV server, I-TASSER, ExPASy ProtParam PSI-BLAST and Clustal Omega used to select model template for I-TASSER
Human adenovirus	Dorden and Mahadevan, 2015	28	BLASTP, Pfam, SMART, Phyre2, SWISS-MODEL, MuFOLD, PFP, ESG, Argot2, BAR+, PSIPred, ProtFun, dcGO, 3d2GO
	Naveed et al., 2017	38 (16%)	BLASTP, Pfam, CATH, SUPERFAMILY, INETRPRO, MOTIF, CDART, SMART, SVMPort, ProtoNet, I-TASSER, ExPASy ProtParam, Virus PLoc, TMHMM, HMMTOP, DISULFIND

No. HPs, Total number of hypothetical proteins examined (percent of hypothetical proteins with proposed annotation revisions if available).

details three general approaches for HP selection (**Table 2**). HPs can be selected randomly or targeted through differential gene expression analysis or phylogenetic relations.

Random Selection

Depending on instructor preference and learning objectives, students can be allowed to select HPs themselves (*i.e.*, student-directed) or selection can be partially or completely directed by the instructor (*i.e.*, instructor-directed). Students can find HPs easily by searching the NCBI knowledgebase for the term "hypothetical protein" to generate a list for selection, as done previously (Bharat Siva Varma et al., 2015). Further, if the student is interested in a specific organism, HPs can be selected randomly using NCBI's Genome database.

Alternatively, instructors may choose to partially or completely direct HP selection. One way a project can be partially instructor-directed is by requiring the class to designate a class pet microbe. The instructor then provides a list of available HPs from the class-appointed pet microbe for student selection. The class pet microbe technique is based on early published computational characterization studies that limited focus to HPs that were randomly selected from several hundred HPs in one highly pathogenic bacterial species (Mohan and Venugopal, 2012; School et al., 2016). To reduce the number of potential HPs for selection, a protein size cut-off can be imposed also (Shahbaaz et al., 2015).

Differential Gene Expression

The differential gene expression approach requires gene expression data, such as those produced by microarray or RNAseq procedures, containing at least two groups (*i.e.*, experimental and control) that are useful for comparison. HPs that have the greatest change in gene expression between groups (*i.e.*, differential gene expression) are given the highest priority for HP selection. Gene expression datasets that measure expression for nucleotide sequences associated with HPs can be generated by the student in the laboratory or found in the Gene Expression Omnibus (GEO) database (Edgar et al., 2002; Barrett et al., 2011, 2013).

If only two groups are available, HPs can be selected using single-gene analysis approach which requires meeting a statistical cut-off, like a T-test p-value <0.05. This approach can produce long lists of differentially expressed HPs that may contain redundancy and cannot be prioritized based on biological relevance, thus prioritization of HPs for characterization, require utilization of statistical methods. For example, volcano plots (*i.e.*, scatter plot that compares a gene's statistical significance via T-test p-value to its biological relevance via fold change) are frequently used to identify differentially expressed genes (Li, 2012; Kumar et al., 2018). Differentially expressed HPs with the best statistical significance (*i.e.*, lowest p-value) and biological relevance (*i.e.*, highest fold change) are given selection priority for the Hypothetical Protein Characterization Project (**Figure 2A**).

If more than two experimental groups are available, HPs can be selected by gene enrichment analysis (Goad and Harris, 2018). HPs can be selected by either singular enrichment analysis or gene set enrichment analysis (Huang et al., 2009; Tipney and Hunter, 2010). In singular enrichment analysis, each gene is considered individually via single-gene analysis, generating multiple lists of statistically significant HPs, one from the differential expression comparison of each experimental group relative to the control. HP lists are then examined for overlapping

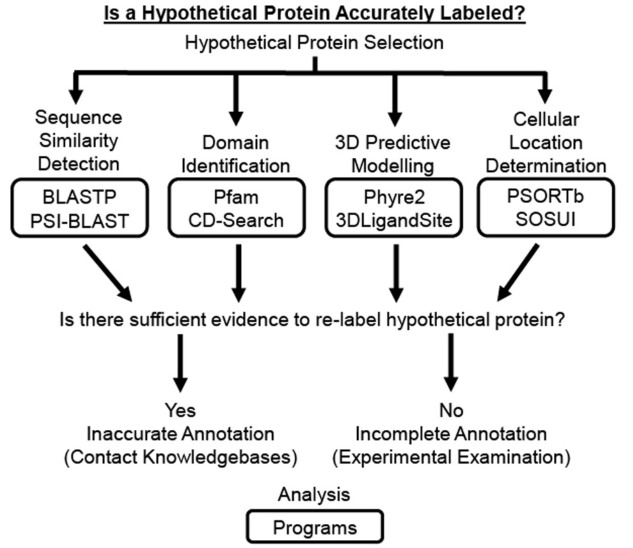

Is a Hypothetical Protein Accurately Labeled?
Hypothetical Protein Selection
Sequence Similarity Detection
BLASTP
PSI-BLAST
Domain Identification
Pfam
CD-Search
3D Predictive Modelling
Phyre2
3DLigandSite
Cellular Location Determination
PSORTb
SOSUI
Is there sufficient evidence to re-label hypothetical protein?
Yes
Inaccurate Annotation
(Contact Knowledgebases)
No
Incomplete Annotation
(Experimental Examination)
Analysis
Programs

FIGURE 1 | Schematic of Hypothetical Protein Characterization Project. The Hypothetical Protein Characterization Project provides students with a process that generates evidence to address if a hypothetical protein (HP) is accurately labeled. The HP can be selected randomly, through differential gene expression analysis using established statistical methods, or phylogenetic relations established through sequence similarity. Once selected, the HP's amino acid sequence is analyzed by web-accessible individual programs for (1) detection of sequence similarities, (2) identification of protein domains, (3) 3D predictive modeling of the HP's structure including active site and potential ligand binding partners, and (4) determination of protein cellular location. If results from these analyses provide sufficient evidence to support a function for the HP, the results can be provided directly to knowledgebases so the protein's public record can be updated. Otherwise, the HP needs experimental examination before a function could be assigned.

HPs, which are considered most relevant to the phenotypic variation under examination (**Figure 2B**).

Alternatively, gene set enrichment analysis (GSEA) compares gene signatures (*i.e.*, list of genes ranked by their differential expression based on an appropriate statistic method such as *T*-test or fold change) rather than individual genes. To do this, one gene signature is used as reference (*i.e.*, all genes are used) and the other signature is used to generate two separate query gene sets derived from the signature's positive and negative tails (*i.e.*, representing the most over-or under-expressed genes in the gene signature, respectively). Query gene sets must include between 15 and 500 genes for GSEA to properly function (Subramanian et al., 2005), and to maximize potential HPs for selection we recommend using a 500 gene inclusion size. GSEA compares the reference signature to each query gene set individually to calculate an enrichment score (**Figure 2C**). Genes that contribute most to reaching the maximum enrichment score for GSEA are called leading-edge genes and are thought to contribute to the phenotypic difference under examination. HPs included among identified leading-edge genes are given the highest priority in HP selection. GSEA requires use of specialized software with a JAVA-based, user-friendly desktop version freely available at Broad Institute (Subramanian et al., 2005).

Sequence Similarity to a Protein With Determined Structure

The sequence similarity to a protein with determined structure approach can find outdated HPs for characterization, as we demonstrate in section 4.1. To select HPs using this approach, students begin by finding established proteins that have already undergone some experimental examination, such as protein structure determination via X-ray Crystallography, and therefore have accurate annotation. The Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) is a rich resource for finding established proteins since it is the largest free and publicly available archives of macromolecular structural data (Bank, 1971; Berman et al., 2000, 2014; Burley et al., 2017). Next, amino acid sequences from established proteins undergo sequence similarity searches using programs such as the Position-Specific Iterated Basic Local Alignment Search Tool (PSI-BLAST) to select HPs for the Hypothetical Protein Characterization Project.

ANALYSIS OF HYPOTHETICAL PROTEINS

After an HP is selected for characterization, the amino acid sequence in FASTA format is acquired from a public knowledgebase like NCBI or UniProt, and used to detect sequence similarities, identify protein domains, predict protein tertiary structure including active site and potential ligand binding partners, and determine cellular location (**Table 3**). Instructional videos for use of each program discussed in this section are available on our "Online Faculty Mentoring Network to Develop Video Tutorials" YouTube channel[1].

Sequence Similarity Detection

Detecting sequences that share significant similarity to an HP is an important first step in analysis since similar sequences are thought to be homologous and likely share a common ancestor (Pearson, 2013). Widely used similarity search programs, like the Basic Local Alignment Search Tool (BLAST), are used to estimate similarity between sequences (Altschul et al., 1990). Results from any BLAST program includes the percentage of query (*e.g.*, amino acid) coverage and identity to individual sequences, with high percentages of query coverage and identify to sequences with known function indicating an outdated HP. Further, a bit-score indicates the required size of the database needed to find the same sequence similarity by random chance with a high bit-score indicating sequence similarity. To estimate the statistical significance of detected similarities, the bit-score is used to calculate an Expect-value (*E*-value), representing the number of closely matched sequences that are anticipated by random change when searching a database of certain size (*i.e.*, random background noise). *E*-values close to zero highlight similar sequences.

At NCBI's website there are several BLAST programs available for use. Nucleotide BLAST (BLASTN) and Protein BLAST

[1] https://www.youtube.com/channel/UCEE6oecA8YKQip9VaqOOHbg

TABLE 2 | Selected approaches for hypothetical protein selection.

Approach	Sub-approach	Description	Level[1]	Setting(s)[2]
Random	Student-directed	Complete student autonomy to select HPs for characterization	Beginner	C
	Instructor-directed	Instructors limit student ability to select HPs for characterization (*e.g.*, students select HPs from genome of "class pet microbe")	Beginner	C
Differential Gene Expression	Single-gene Analysis	Use of statistical method(s) (*e.g.*, *T*-test and/or fold change) on gene expression data to find and prioritize individual differentially expressed HPs for characterization	Intermediate	C, E, H, G
	Singular Enrichment Analysis	Gene enrichment analysis comparing groups of significant HPs with similar differentially expression as defined by single-gene analysis	Intermediate	C, E, H, G
	Gene Set Enrichment Analysis	Gene enrichment analysis comparing a group of the most differentially expressed HPs to a gene signature (*i.e.*, gene list ranked by differential expression based on a statistical method)	Advanced	E, H, G
Phylogenetic Relations	N/A	HPs for characterization are selected for their sequence similarities to proteins with established tertiary structures	Intermediate	E, H, G

[1] *Level definitions: Beginner, does not require additional steps or prior knowledge of statistics; Intermediate, may require prior knowledge of statistics and/or additional steps using free web-accessible programs; Advanced, requires prior knowledge of statistics and additional steps using publicly available free downloadable programs.*

[2] *C, classroom; E, experiential learning courses; G, graduate projects; H, undergraduate honors and capstone projects.*

HPs, hypothetical proteins.

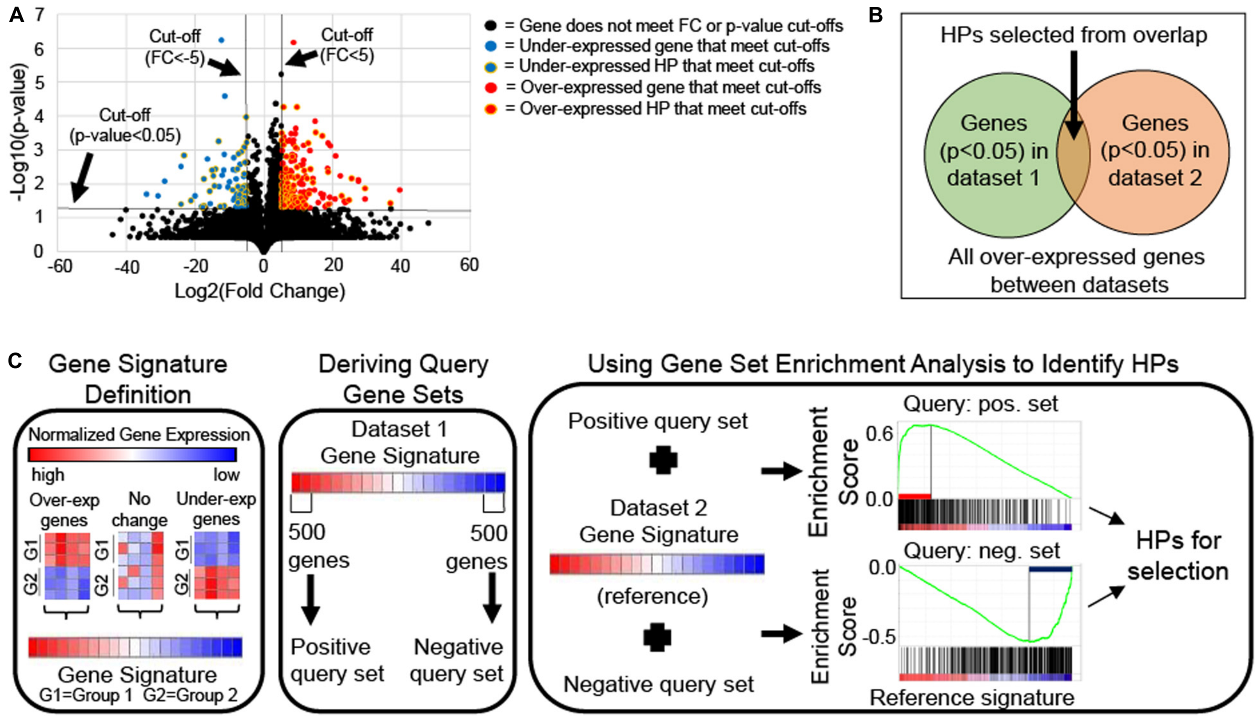


FIGURE 2 | Schematics of differential gene expression approaches for hypothetical protein (HP) selection. **(A)** Volcano plot of mRNA expression data from Gene Expression Omnibus accession number GSE46687 identified HPs with statistical (two-tailed Welch's *T*-test *p*-value < 0.05) and biological relevance [fold change (FC) > 5 for over-expressed or <–5 for under-expressed genes in experimental compared to control groups] to antibiotic resistance in *Staphylococcus aureus* that could be selected for the Hypothetical Protein Characterization Project. **(B)** Venn diagram illustrates conceptually how HPs are selected from singular enrichment analysis using the overlap of statistically significant (e.g., *T*-test *p*-value < 0.05) over-expressed genes between two mRNA expression datasets. The same concept applies to selecting under-expressed HPs also. **(C)** Schematic shows how HPs can be selected from gene signature comparison using Gene Set Enrichment Analysis (GSEA). Gene signatures are gene lists ranked by their differential expression based single-gene analysis (e.g., T-score or FC). A gene signature for each of two mRNA expression datasets are generated. One signature is chosen from which the 500 most over- and under-expressed genes are taken to derive positive and negative query gene sets, respectively. Each query gene set is compared individually to the second gene signature, which is used as reference for GSEA. GSEA calculates an enrichment plot with a maximum enrichment score. GSEA identifies leading-edge genes, which are genes that contribute most to reaching the maximum enrichment score. HPs among leading-edge genes are selected for the Hypothetical Protein Characterization Project.

TABLE 3 | Selected analysis programs for Hypothetical Protein Characterization Project.

Objective	Program	Citation	Description
Sequence Similarity Detection	BLASTP	Altschul et al., 1990	Encompasses similarities between relevant sequences to predict the functionality and evolutionary aspect of sequences between gene families.
	PSI-BLAST	Altschul et al., 1997; Altschul and Koonin, 1998	Provides means of detection to note distant relationships between proteins.
Domain Identification	Pfam	Sonnhammer et al., 1998; El-Gebali et al., 2019	Database of functional proteins that are called domains. Provides the students with structure of the protein, family annotation, and protein search against database models.
	CD-Search	Marchler-Bauer and Bryant, 2004; Lu et al., 2020	Protein annotation that contains annotated sequence alignment models along with complete proteins. The output allows for identification of domains in the form of matrices.
3D Predictive Modeling	PHYRE2	Kelley et al., 2015	Provides affiliation of proteins to predict protein structure, function, and mutation. Software uses a detection method through homologs to build 3D models, note binding sites, and analyze amino acids.
	3DLigandSite	Wass et al., 2010	Allows for the prediction of ligand binding sites by using the predicted protein structure.
Cellular Location Determination	SOSUI	Hirokawa et al., 1998	Provides transmembrane domain prediction of a single alpha helix. This process occurs through scanning through protein sequence to identify hydrophobic regions.
	PSORTb	Yu et al., 2010	Contains multiple modules to analyze biological features of known characteristics pertaining to subcellular localization. Thus, the database may predict a protein localization site. Database also encompasses Gram-negative and Gram-positive localization features.

(BLASTP) detect sequence similarities between other nucleotide and amino acid sequences, respectively. While either BLAST program can be used and comparing between BLASTN and BLASTP would generate a good educational discussion, the Hypothetical Protein Characterization Project uses BLASTP to reduce student confusion by providing input consistency across HP analysis. The Hypothetical Protein Characterization Project also looks at results from Position-Specific Iterated BLAST (PSI-BLAST). PSI-BLAST first generates the same results as BLASTP sequence alignments to establish a specialized position-specific scoring matrix (PSSM) from all user-selected sequences, representing what the group of sequences might look like on a positional basis. Use of PSSM allows for the comparison of local amino acid sequence patterns between proteins rather than direct comparison of amino acid sequences themselves. Therefore, through several rounds of computational analysis (*i.e.*, iterations), PSI-BLAST refines the PSSM for an HP based on PSSM alignments with user-selected sequences identified within each iteration. This process combines underlying conservation information from a range of related sequence into a single score matrix (Altschul et al., 1997; Bhagwat and Aravind, 2007). By using this PSSM methodology, PSI-BLAST can detect less similar sequences and is more likely to identify HPs. True HPs, by definition, cannot have similar sequences with established function. Thus, identification of similar sequences with known function using BLAST can strongly indicate outdated annotation for the HP being analyzed.

Domain Identification

Protein domains are spatially distinct and compact regions of a protein that can fold into a stable structure that may be integral to the protein's function (Yegambaram et al., 2013). Domains are often conserved across proteins with similar function across diverse species. There are several protein domains databases that are readily available. For example, the Pfam database has been collecting protein information since 1995 and now contains more than 17,000 entries (Sammut et al., 2008; Finn et al., 2010; El-Gebali et al., 2019; Lu et al., 2020). Pfam has a large collection of protein domains, which are individually represented by hidden Markov model (HMM) based profiles and multiple sequence alignments (Sonnhammer et al., 1998). While Pfam is a trusted resource, it can be expanded upon. NCBI's Conserved Domain Database (CDD) is a collection of multiple sequence alignment models for full-length proteins and ancient domains that includes NCBI-curated domains, which use 3D-structure information to define domains, and domain models imported from several external databases including Pfam (Lu et al., 2020). The CDD can be searched using the CD-Search tool which is easily accessible from NCBI's Protein Database. Conserved domain (CD)-Search uses RPS-BLAST, a PSI-BLAST variant, to scan a protein for any sets of pre-calculated position-specific scoring matrices (Marchler-Bauer and Bryant, 2004). CD-Search results are presented as an annotation of protein domains with high confidence associations. These associations are determined by calculating the *E*-value between the protein's sequence and any domains are shown as specific hits using similar methods to those previously described for BLAST programs. The Structural Classification of Proteins (SCOP) database of proteins with known structures that organizes protein domains by their evolutionary and structural relationships, providing a broad overview of established protein folds, detailed information about any close relatives to an HP, and a general framework for future protein classification (Andreeva et al., 2014, 2020). SUPERFAMILY is a database of structural and functional protein annotation based on a collection of HMMs representing SCOP superfamily structural domains

(Gough et al., 2001). The Conserved Domain Architecture Retrieval Tool (CDART) and Simple Modular Architecture Research Tool (SMART) can be used to identify similarities across significant evolutionary distances through comparing domain architecture (*i.e.*, sequential order of conserved domains in a protein sequence) for protein (Geer et al., 2002)and genetically mobile domains (Schultz et al., 1998; Letunic and Bork, 2018), respectively, both using PSI-BLAST. Further, the CATH protein domain database classifies protein secondary structures from the PDB and collects domains into superfamilies only when there is enough evidence of divergence from a common ancestor (Sillitoe et al., 2019). The CATH database is paired with Gene3D which uses CATH's information to predict structural domain locations for protein sequences available in public databases, allowing for functional information and active site residue annotations (Lewis et al., 2018). Since domains are distinct regions of a protein, it is not uncommon for a protein to have more than one identified domain, ergo results from searching these domain databases also usually identify the range of amino acids associated with domains of HPs under investigation. HPs containing at least one domain with an established function likely have outdated annotation.

3D Predictive Modeling

3D predictive modeling gives students the ability to consider an HP's tertiary structure and potential binding partners. To do this, the Structural Bioinformatics Group at Imperial College London developed a suite of integrative modeling programs, Protein Homology/analogY Recognition Engine V 2.0 (Phyre2), with free web portal access (Kelley et al., 2015). Phyre2 uses template-based modeling (*i.e.*, homology and comparative modeling) based on a three-step procedure. First, homologous sequences are gathered by scanning a query sequence against specially curated protein sequence database with HHblits. This produces a multiple-sequence alignment that is used by PSIPRED to predict secondary structure before both the alignment and secondary structure prediction combined into a query HMM. Next, the query model is scanned against a database of HMMs of proteins of known structure. From this search, top-scoring alignments are used to generate an unrefined backbone-only model. Finally, the model is refined via loop modeling and side-chain placement. Template-based modeling as used by Phyre2 is a good approach assuming homology exists between a user-supplied sequence and at least one sequence of known structure, meaning Phyre2 and any other template-based modeling programs are unable to model true HPs. If the Phyre2 generated model is assigned a $>90\%$ confidence and does not contain substantial disorder ($<50\%$), Phyre2 automatically submits the model and its corresponding amino acid sequence to the 3DLigandSite server for ligand binding site prediction (Wass et al., 2010). In a similar approach to template-based modeling, 3DLigandSite identifies structures like the one generated by Phyre2 model and superimposes bound ligands from identified structures onto the model. This is done multiple times to establish a cluster of the highest number of ligands for active site prediction. It may take several hours for Phyre2 and 3DLigandSite to generate results, however, those results include: (1) tables of identified ligand clusters and binding-site residues, (2) visual representations of the model, and (3) predicted binding site and any ligand clusters. Thus, 3D predictive modeling can identify outdated HPs due to theoretical tertiary structure homologies with proteins of known function.

There are several other computational resources available to predict an HP's tertiary structure from its primary (*i.e.*, amino acid) sequence and predict its potential binding partners. Alternatives to Phyre2 include but are not limited to SWISS-MODEL (Schwede et al., 2003; Waterhouse et al., 2018), PS^2 (Chen et al., 2006, 2009), and the Iterative Threading Assembly Refinement (I-TASSER) program (Roy et al., 2010; Yang and Zhang, 2015). SWISS-MODEL is the original fully automated protein homology modeling server. In its most recent version, SWISS-MODEL uses a ProMod3 that differs from prior versions and other programs like Phrye2 by replacing *ab-initio* techniques to resolve insertions and deletions in the aligned template structure with structural database searches for viable candidate fragments. PS^2 is another automatic homology modeling server that uses a substitution matrix, S2A2, to combine sequence and secondary structure information to detect established proteins with remote similarity before the 3D structure is generated via the MODELER modeling package (Sali and Blundell, 1993; Webb and Sali, 2014). MODELER uses an alignment between the HP's sequence and known related structures to generate a model containing all non-hydrogen atoms based on satisfying atomic spatial restraints. The I-TASSER is an integrated platform for automated protein structure and function prediction from an amino acid sequence that is based on a sequence-to-structure-to-function paradigm. To accomplish this, I-TASSER begins by using multiple threading alignments and iterative structural assembly simulations to generate 3D atomic models. The HP's function is inferred from these 3D models by structurally matching them with known proteins. Phyre2, SWISS-MODEL, PS^2, and I-TASSER all measure the quality of their resulting models though differences exist in how models are measured for quality. I-TASSER also provides functional annotations on ligand-binding (*i.e.*, active) sites, Gene Ontology terms, and Enzyme Commission numbers not provided by the other programs, though 3DLigandSite competes by providing active site characterization and ligand predictions for models produced by Phrye2. Further, potential binding partners for HPs can be predicted from programs separate from 3D modeling programs. For example, STRING (Snel et al., 2000; Szklarczyk et al., 2019) and STITCH (Kuhn et al., 2008; Szklarczyk et al., 2016) are databases of protein-protein and protein-chemical interactions, respectively. An HP's function can be inferred from the network of proteins and chemicals identified from searching its amino acid sequence in the STRING and STITCH databases.

Cellular Location Determination

Students finally consider the cellular environment in which their HP may exist. For classroom purposes, students focus on determining the cellular location of their HP using two programs, PSORTb and the SOSUI server. PSORTb consists of several analytical modules that each analyze one biological feature known to impact or be characteristic of a subcellular localization. PSORTb combines the results from each module to assess the likelihood of a protein being assigned a specific localization. Based on these likelihood assessments, a probability value between 0 and 10 for each of the five localization sites is determined. PSORTb considers 7.5 a good cutoff for assignment of a protein to a single cellular location (Yu et al., 2010). Similarly, SOSUI distinguishes between membrane and soluble proteins and predicts transmembrane helices in potential membrane proteins (Hirokawa et al., 1998; Mitaku and Hirokawa, 1999; Mitaku et al., 2002). To do this, SOSUI considers four physicochemical parameters (amphiphilicity index, hydropathy index, index of amino acid charges, and length of each sequence) to calculate grand averages of hydropathy (GRAVY). Positive GRAVY values indicate hydrophobic; negative values mean hydrophilic (Chang and Yang, 2013). For a more detailed analysis, ExPASy ProtParam can be used to calculate physicochemical parameters individually including aliphatic index, index of amino acid composition, length of each sequence, and GRAVY (Gasteiger et al., 2005; Artimo et al., 2012). ExPASy ProtParam also provides experimentally useful information such as instability index (*i.e.*, estimate of HP stability in a test tube), extinction coefficient (*i.e.*, measure of light absorbance at 280 nm wavelength), estimated half-life in mammalian reticulocytes, yeast, and *Escherichia coli*, and theoretical pI (*i.e.*, isoelectric point, pH where the HP is electrically neutral). While the ability to determine cellular location for an HP does not distinguish outdated annotation from true HPs, cellular location can support re-annotation conclusions for outdated HPs drawn from other results generated from the Hypothetical Protein Characterization Project.

EXAMPLE HYPOTHETICAL PROTEIN CHARACTERIZATION PROJECTS

The following section contains examples to demonstrate possible Hypothetical Protein Characterization Project results that might be encountered in educational settings. The examples presented here utilized FASTA-formatted amino acid sequences acquired from the NCBI Protein database (Coordinators, 2018). The UniProt knowledgebase (UniProt, 2019) was consulted to highlight differences between knowledgebases. For consistency across projects, the following program parameters were used: (1) Default program settings for all programs, (2) The most similar non-HP sequence was reported from BLASTP analysis, making it the most relevant description for potential re-annotation, (3) PSI-BLAST results were generated from three iterations of each sequence to capture similar sequences more extensively as no significant change resulted from running additional iterations, and (4) The least similar non-HP sequence resulting from PSI-BLAST analysis was reported. Data for these example projects were collected between March 15–23, 2020.

AUH26_00140 Should Be Re-annotated as an ABC Transporter Permease

To find an example of an HP with outdated annotation, the sequence similarity to a protein with determined structure approach to select HPs was used. Since we previously used this approach to examine HPs related to major facilitator superfamily proteins related to antibiotic resistance in *S. aureus* (Marklevitz and Harris, 2016), we browsed the PDB for multidrug resistance transporters related to antimicrobial resistance. We performed PSI-BLAST on approximately five randomly selected transporters before finding a transporter with HPs, a process taking less than 30 min, demonstrating the feasibility of sequence similarity to a protein with determined structure approach to identify outdated HPs. We found PSI-BLAST of the multidrug ABC transporter Sav1866 from *S. aureus* (PDB accession: 2ONJ) identified HPs. We selected AUH26_00140 (96% query coverage, 38.89% identity, *E*-value = 6.0 × 10^{-142}) over three other HPs with lesser similarity (W538_02582 from *S. aureus* VET0261R, W475_02351 from *S. aureus* VET0166R, and V089_02512 from *S. aureus* GD2010-115). We noted that AUH26_00140 was not included in the UniProt knowledgebase. The 592-amino acid sequence for AUH26_00140 is below:

>OLC18526.1 hypothetical protein AUH26_00140 [*Candidatus Rokubacteria* bacterium 13_1_40CM_69_96]

```
MPLGPYRRLFVYLRPHVPVLVLGACLALIVSGMEGLTAWLV
KPVMDDIFIRRDGLMLKLIPLALLAVYVVKGVARYLQSYLM
AAVGERVVARLRRELYTHIQSMPLSFFSDVHSADLMSRILTD
VTRLARLSSGVLVMGVRQLGTIAALLVVMLAREWALTLTA
LVAFPAIALIVRTIGRRLYTINKRTQERVAQLAVLLHESFSGTK
IVKAFGRERHEQARFDALNDRLLNLSLKNVRADEITEPLME
IAGALGIMAVLWYGGYRVIEGHMTPGTLFSFTAAALMLYG
PVRRLSRSLNVVQQSTASVERVFHILELPPAITDRPGATRLET
FTRALAFERVDFRYGDADEMTLKEISLEIRKGEVVAFVGMS
GAGKSTLMDLVPRFHDVTAGRITLDGRDLRDVTQASLRAQ
LGVVTQETFLFSDTIRYNIAYGRPDATFEEIVRAARQAHAH
DFTLACPDGYDTLVGERGVRLSGGQRQRIAIARAFLKNPPIL
ILDEATSDLDAESEFMVQQALAELMHGRTVFVIAHRLATVR
NADRIVVVHDGRIAEIGRHEELIARDGIYRRLYALQMEGFPG
EQVGGPGGPLRPR
```

When AUH26_00140 was used as query for BLASTP, the most similar non-HP sequence was an ABC transporter permease from *Candidatus rokubacteria* bacterium (97% query coverage, 98.96% identity, *E*-value = 0.0), which is a strong indicator that AUH26_00140 has outdated annotation. PSI-BLAST results included mostly lipid A export permease protein MsbA (98% query coverage, *E*-value = 0.0, 49.06% identity) and no HPs, further supporting BLASTP results.

The NCBI Protein database did not list any domains. CD-Search identified COG1132 (*E*-value = 0.00), a domain that spans most of AUH26_00140 (amino acids 3 to 576) which is associated with the ATPase and permease component of the ABC-type multidrug transport system. Pfam also found two matches: (1) an ABC transporter transmembrane region (CL0241, *E*-value = 3.2 × 10^{-52}) spanning amino acids 21 to 291, and an ABC transporter domain (CL0023, *E*-value = 3.3 × 10^{-33}) that spans amino acids 354 to 503, supporting results identified by CD-Search.

Phyre2 generated a tertiary structure model for AUH26_00140 with 100% confidence from part of an X-ray diffraction structure of a heterodimeric ABC transporter from *Thermotoga maritima* (model template c3qf4A) whose protein sequence covered 96% of AUH26_00140's sequence with 31% identity (**Figure 3A**). From this model, 3DLigandSite predicted a 14-amino acid binding site that could bind to adenosine triphosphate (ATP), adenosine diphosphate (ADP), and magnesium.

PSORTb predicted that AUH26_00140 is a cytoplasmic membrane protein (localization score = 10). These results are supported by SOSUI, which calculated AUH26_00140 to be a membrane protein (GRAVY = 0.168920) with five transmembrane helices. While additional analysis, such as comparison of physiochemical properties, multiple sequence alignment, and phylogenetic tree analysis, are needed to fully support re-annotation, these results here suggest AUH26_00140 likely has outdated annotation and should be re-labeled to be a ABC transporter permease in keeping with its closest similar sequence.

L2624_01843 Should Be Re-annotated as a DUF871-Containing Outer Surface Protein

L2624_01843 from *Listeria monocytogenes* was originally characterized as part of student's Hypothetical Protein Characterization Project using the student-directed approach for HP selection. NCBI Protein database listed L2624_01843 as an HP. L2624_01843 was not included in the UniProt knowledgebase. The 362-amino acid sequence for L2624_01843 is provided below:

>AKI46902.1 hypothetical protein L2624_01843 [*L. monocytogenes*] MRKLGISVFPQHVALEESL
EYIETAAKYGFSRIFTCLISANDEAEFAKLETICKRAKELGFD
VIADVDPTVFESLNITYKELDRFKELGLAGLRLDLGFSGSEE
AAMSFDDTDLKIELNISNGTRYVENILSYQANVGNIIGCHN
FYPRKYTGLSRKHFLRTSKQFKDLNLRTAAFVSSNSGEFGPW
FVVDGGLPTMEEHRGVDITVQAKDLWNTGLIDDVIVGNM
FASEDELRALSELNRNELQLAVEFLDGATDVEKEIVLTQKHF
NRGDASEYVLRSTMTRVNFKQFDFPAHDTNTIAKGDVTID
NDGYERYKGEMQVALQEMENSGNTNIVARIVPEERYLLDTI
LPWQHFRLVEKKK

When L2624_01843 was used as query for BLASTP, all identified similar sequences had DUF871 domain-containing protein annotation (100% query coverage, 99.17% identity, *E*-value = 0.0). While most similar sequences identified by PSI-BLAST for L2624_01843 are DUF871 domain-containing proteins, a few sequences had outer surface protein descriptions with the closest sequence being EFR87458.1 which is found in *Listeria marthii* FSL S4-120 (100% query coverage, 98.90% identity, *E*-value = 0.0).

The NCBI Protein database showed that L2624_01843 contains a conserved COG3589 region that has an unknown function that spans 361 amino acids (99.7% of the protein). CD-Search showed COG3589 was similar (covering amino acids 1 to 361, *E*-value = 0.00) to the DUF871 domain superfamily, which was confirmed by Pfam that found DUF871 was the only significant match (covering amino acids 1 to 357, *E*-value = 3.1×10^{-136}).

Next, the tertiary structure and potential ligand binding partners for L2624_01843 were predicted. Phyre2 generated a protein model for L2624_01843 with 100% confidence from the crystal structure of an outer surface protein from *Bacillus cereus* (model template c1x7fA, PDB accession 1X7F_A) whose protein sequence covered 95% of L2624_01843's sequence with 51% identity (**Figure 3B**). Interestingly, according to NCBI's Protein database, 1X7F_A is 385 amino acids long and contains a DUF871 domain spanning across amino acids 28 to 384. 3DLigandSite predicted a binding site involving 32 amino acids, mostly comprised of residues 176–185 and 222–228, that bound with the following heterogens: NADPH dihydro-nicotinamide-adenine-dinucleotide phosphate (NDP), flavin mononucleotide (FMN), magnesium, NADP nicotinamide-adenine-dinucleotide phosphate (NAP), zinc, b-D-mannose (BMA), a-D-mannose (MAN), and calcium.

SOSUI calculated L2624_01843 to be a soluble protein (GRAVY = −0.328453) with no transmembrane helices, which supported PSORTb predictions that L2624_01843 was a cytoplasmic protein (localization score = 7.50). We noted that PSORTb is unable to detect outer surface as a cellular location (Yu et al., 2010). Taken together, these data suggested that L2624_01843 should be re-labeled as a DUF871-containing Outer Surface Protein though experimental examination of DUF871 is needed to further refine L2624_01843's annotation.

WP_002214142 Is a True Hypothetical Protein

WP_002214142 from *Yersinia pestis* plasmid pMT1 was originally characterized as part of student's Hypothetical Protein Characterization Project using the instructor-directed class pet microbe approach for HP selection. WP_002214142 was labeled as a hypothetical (*i.e.*, uncharacterized) protein in both NCBI Protein and UniProtKB databases. The 77-amino acid sequence is provided below:

>WP_002214142.1 MULTISPECIES: hypothetical protein [Bacteria] MAQAIPSTSVCSTKRTRPPMLVALNGH
PVSRRLKTPTSYRQATEQPSDSLQATICRNRTLGRLMRVAIIK
PTRKQIV

BLASTP identified several HPs from various species with similar sequences to WP_002214142. PSI-BLAST was not able to identify similar sequences for WP_002214142 that were not HPs and new sequences could not be detected above the 0.005 threshold from the second iteration of PSI-BLAST. In summary, no sequences from non-HPs were identified.

WP_002214142 contains no documented domains according to NCBI's knowledgebase, either Protein database or the CDD. Pfam also could not detect any domains. Lack of identified domains is a good indication that the HP under characterization is a true HP.

Phyre2 generated a tertiary structure model for WP_002214142 with 31.8% confidence from part of an X-ray diffraction interferon-induced RNA binding protein from *Homo sapiens* (model template c6c6kD) whose protein sequence covered 30% of WP_002211802's sequence with 52% identity (**Figure 3C**). Low model confidence and similarity to the template supports the conclusion that WP_002214142 is a true HP. To further support this conclusion, 3DLigandSite

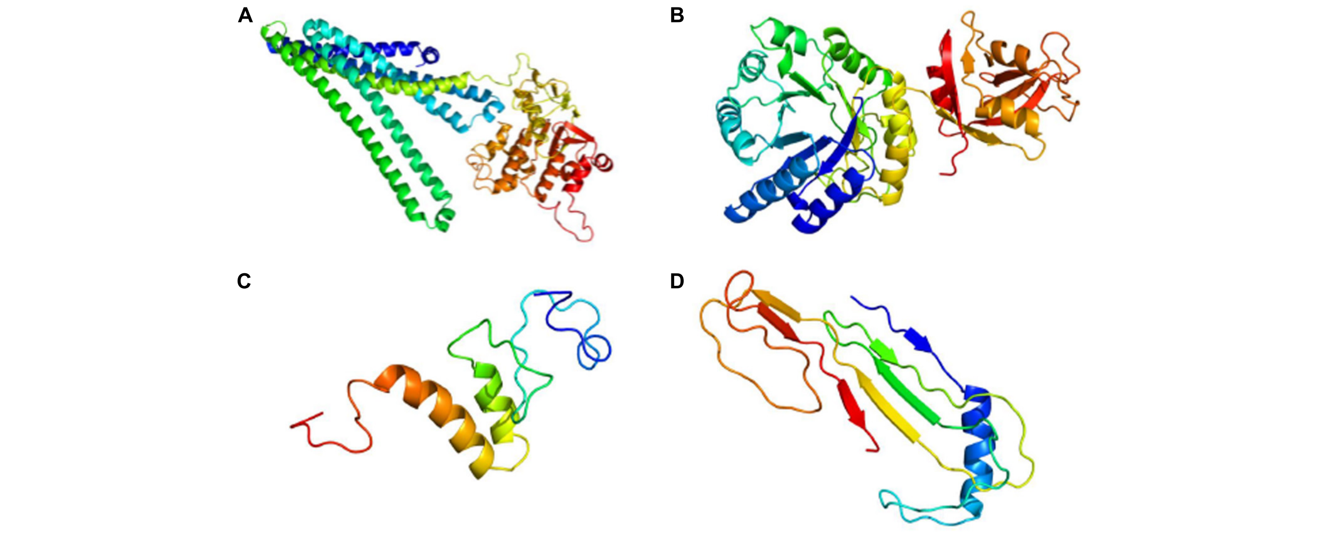


FIGURE 3 | Predictive 3D Models for Hypothetical Protein Characterization Project Examples. **(A)** Completeness of Phyre2 model of AUH26_00140 shows AUH26_00140 has outdated annotation. **(B)** Completeness of Phyre2 model of L2624_01843 suggests L2624_01843 has outdated annotation. **(C)** Lack of completeness of Phyre2 model of WP_002214142 supports the conclusion that WP_002214142 is an example of indeterminate annotation. **(D)** Lack of completeness of Phyre2 model of YP_009724396 indicates YP_009724396 is an example of indeterminate annotation. All images are colored by rainbow from N terminus to C terminus.

was unable to predict a binding site or ligand binding partners from this model.

SOSUI calculated WP_002214142 to be a soluble protein (GRAVY = −0.425), though PSORTb could not predict a cellular location for WP_002214142 (localization score = 2.00). Project results taken together do not provide sufficient evidence to re-label WP_002214142 in public knowledgebases. Therefore, experimental examination is needed before WP_002214142's annotation can be improved.

ORF8 (YP_009724396.1) Is a Viral Example of a True Hypothetical Protein

While the Hypothetical Protein Characterization Project was optimized for use on bacterial species, students frequently want to apply it to other organisms. A virus that students have recently want to use for their projects is Severe Acute Respiratory Syndrome coronavirus 2 (*i.e.*, SARS-CoV-2), the causative agent of COVID-19 (Wang et al., 2020). So, for this example, ORF8 (*i.e.*, ORF8) was randomly selected from the SARS-CoV-2 genome. When this example was prepared, ORF8 was labeled as an HP in the NCBI Protein database and not found in UniProt. The 121-amino acid sequence is provided below:

>YP_009724396.1 ORF8 protein [Severe acute respiratory syndrome coronavirus 2] MKFLVFLGIITTVAAFHQE CSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIELC VDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGSLVVRCSF YEDFLEYHDVRVVLDFI

All but one protein identified by BLASTP had ORF8 annotation and came from SARS-CoV-2. The one sequence that was not an ORF8 was a HP from Bat SARS-like coronavirus (100% query coverage, 94.21% identity, *E*-value = 8×10^{-81}). Most similar sequences identified by PSI-BLAST for ORF8 were also HPs or proteins with vague descriptions (*e.g.*, ORF8a or ORF10). However, one sequence (AAP51236.1), which came from Human SARS coronavirus (SARS Co-V) GD01, had a BGI-PUP(GZ29-nt-Ins) description (98% query coverage, 29.03% identity, *E*-value = 4×10^{-42}). The BGI-PUP(GZ29-nt-Ins) description is associated with a SARS-CoV isolate with a 29 nucleotide insertion at the relative position 27,995 in its genome (Pavlovic-Lazetic et al., 2005).

The NCBI Protein database listed no domains for ORF8. However, CD-Search showed a functionally uncharacterized corona_NS8 superfamily domain conserved in coronaviruses (100% query coverage, *E*-value = 1.87×10^{-39}). CD-Search results were confirmed by Pfam that found Coronavirus NS8 protein was the only significant match (*E*-value = 3.8×10^{-44}). Both CD-Search and Pfam aligned the corona_NS8 superfamily domain to residues 1 to 118 in ORF8.

To predict the tertiary structure for ORF8, Phyre2 generated a protein model for ORF8 with 33.3% confidence from the immunoglobulin-like beta-sandwich fold of an X-ray diffraction of the ORF7a accessory protein from SARS-CoV (model template d1xaka) whose protein sequence covered 17% of ORF8's sequence with 30% identity (**Figure 3D**). From this limited model, 3DLigandSite was unable to predict potential binding site or ligand binding partners.

With regards to cellular location, SOSUI calculated ORF8 as a soluble protein (GRAVY = 0.219). PSORTb could not predict a cellular location for ORF8 because PSORTb cannot analyze viral sequences. Taken together, these data suggested that more experimental examination is needed before ORF8's annotation can be improved, which is not surprising given the novelty of SARS-CoV-2 at this time.

DISCUSSION

The Hypothetical Protein Characterization Project is a valuable educational tool where students learn and apply knowledge of computational programs that can assist with ongoing manual curation efforts to improve genome annotation (**Figure 1**). This project incorporates interdisciplinary concepts to identify and predict HP characteristics, such as sequence similarities, domains, 3D structure, ligand binding partners, and cellular location. Project results are used to determine whether an HP has outdated or indeterminate annotation. Individual and collective results from student projects can be used to improve public database annotation. While current NCBI knowledgebase protocols dictate that only the research group that deposited the genome can change its annotation, depositor contact information is usually provided. While contact information may need to be updated, students are encouraged to use internet search resources to find and share their HPCP results for outdated HPs with the genome's depositor(s). This provides students with an opportunity to establish and develop professional connections that could benefit them throughout their careers. Further, individual and collective results from student projects are often welcomed for scientific conference poster presentations, which further stimulates student motivation, learning opportunities, and ideally scientific employability.

The project is versatile and customizable to accommodate a wide variety of learning objectives. The project can be used in both online and in-seat educational settings for undergraduate and graduate classes in microbiology, bioinformatics, genetics, and/or biochemistry. HP analysis objectives and programs can be modified based on the instructor's learning objectives, and we recommend instructors test programs immediately prior to classroom use to ensure functionality as programs are often temporarily taken off-line for maintenance and updates. Further, this project can be expanded through advanced approaches to HP selection, such as differential gene expression or phylogenetic relations, and additional HP analysis to provide an advanced, research-oriented project that is well suited for undergraduate capstone, honor's, and experiential learning projects as well as Master level theses (**Table 2**). Given the variety of potential HP selection approaches and programs for HP analysis, students and instructors are encouraged to find, develop, and/or use these and other methods of selecting and analyzing HPs to best suit their specific needs.

Further, the project was designed to stimulate classroom discussion based on the methodology and interpretation of variations in results from different knowledgebases and HP analysis programs (**Table 3**). Classroom discussion can begin with comparing and contrasting information found on the HP between NCBI Protein database (Coordinators, 2018) and UniProt knowledgebase (UniProt, 2019). As seen from examples provided in this paper, in some cases like WP_002214142, HP information provided is the same between Protein and UniProt. In other cases, like AUH26_00140 there are differences in HP inclusion and/or provided information. Similar discussions that compare analysis programs can be applied to each objective. For example, if an instructor wants to examine program methodology differences, students can discuss why results first iteration PSI-BLAST results are the same as BLASTP results and how PSI-BLAST uses BLASTP results to identify distant similar sequences. An instructor that wants to continue discussing impacts of knowledgebase inclusion could similarly emphasize program inclusion by discussing similarities and differences in methodology and generated results between Pfam and CDD, which includes a number of external source databases including Pfam (Marchler-Bauer et al., 2017; Lu et al., 2020). Instructors may decide to have students explore other bioinformatic resources to supplement or replace analysis databases and programs described in this paper to stimulate student discussion. Finally, though we used default settings for our examples here, student discussion can be generated around how and why variations from default settings change results of program analysis. Taken together, this discussion highlights the educational aptitude of the Hypothetical Protein Characterization Project.

Random Selection of Hypothetical Proteins Is Best for Classroom Use

Random selection of HPs for the Hypothetical Protein Characterization Project is optimal for beginning students with no prior experience in bioinformatics or statistics (**Table 2**). Random selection is the easiest HP selection method since it does not require extra computational analysis. This makes random selection of HPs good for undergraduate classroom use, particularly as a multi-step individual assignment. Example assignment instructions with grading rubrics and their 15-week course schedule designed for use in student-directed random HP selection are included in **Supplementary Materials**.

Giving students complete autonomy in HP selection (*i.e.*, student-directed) empowers them to take ownership of their projects. Students will naturally select HPs from a wide range of species, the student-directed approach is good for identifying both outdated and true HPs that can be used as examples in large-class discussions. However, programs can vary in their ability to generate accurate results from diverse species. For example, PSORTb requires its users to provide the type of microbe (*i.e.*, Gram-negative or Gram-positive) that the amino acid sequence came from. If the student selects an HP from a *Mycobacterium* that has an advanced cell wall, PSORTb may struggle to provide clear and accurate results. Further, PSORTb was not designed to analyze eukaryotic HPs, though its complementary program WoLF PSORT can analyze eukaryotic HPs (Horton et al., 2007), which can cause confusion and frustration among students and instructors alike if the student selects a eukaryotic protein for study. To avoid such complications, we recommend some instructor-imposed limitations in HP selection (*i.e.*, instructor-directed) for classroom use. Partially instructor-directed approaches, such as the class pet microbe discussed earlier, are better than the instructor simply assigning HPs to students directly (*i.e.*, completely instructor-directed) as this approach allows students to retain some autonomy in the selection process while still reducing the confusion that can result from interpreting results across diverse species. However, both partial and complete instructor-directed HP selection approaches may not generate ample examples of outdated HPs needed for large-class discussions unless the instructor is careful to select HPs from older genomes that are more likely to have outdated annotation compared to recently published genomes.

Hypothetical Protein Selection via Differential Gene Expression Is Best for Advanced Students With the Ability to Conduct Laboratory Experimentation

Selecting HPs based on differential gene expression is a great approach that expands the Hypothetical Protein Characterization Project by incorporating statistical analysis of gene expression data to identify HPs that have a specific biological relevance. Analysis of gene expression differences adds more scientific rationale to the project, which makes true HPs identified by the project using the differential gene expression approach potentially valuable in addressing serious biological questions, allowing a priority to be placed on their experimental examination. While the differential gene expression approach can be used in upper-level undergraduate and graduate classrooms where statistics is a pre-requisite, without laboratory access students cannot fully realize their educational potential (**Table 2**). For this, advanced educational applications such as first-year experiential learning courses, undergraduate honor's and capstone projects, or graduate work where students have access to laboratory resources to experimentally examine true HPs identified from this approach are needed. Further, having a laboratory component to the project can be helpful if the instructor wants to share student project results within the broader biological sciences community.

This paper discussed three progressively more challenging ways to identify HPs using differential gene expression. Single-gene analysis, the easiest way to use differential gene expression to identify HPs, requires an understanding of statistics since it uses statistical methods such as a Student's T-test to select HPs through via differential gene expression. Singular enrichment analysis improves upon single-gene analysis by selecting overlapping HPs between differential expression comparisons so that HPs can be grouped based on their potential biological relevance. However, due to its dependence on single-gene analysis for HP selection, singular enrichment analysis only considers HPs that meet a specific statistical cut-off, producing long lists of differentially expressed HPs that may contain redundancy. To overcome these limitations, GSEA considers all genes during analysis by removing the need for a statistical cut-off (Tipney and Hunter, 2010). GSEA is extremely complex, and best for advanced educational projects such as a Master thesis, where the goal is to identify true HPs whose immediate experimental examination could directly enhance scientific understanding of a variety of biological mechanisms (Goad and Harris, 2018).

Further Computational Analysis Expands the HPCP for Advanced Students Without Laboratory Access

As mentioned earlier, selection of HPs via sequence similarity to a protein with determined structure is inherently useful for finding outdated HPs that do not require further experimental examination (Marklevitz and Harris, 2016). Results generated from HPs selected by this approach become supporting evidence toward the conclusion that the selected HPs should be re-annotated in keeping with similar sequences with established annotation. Due to this, 3D predictive models generated from this project, like the one we provided for AUH26_00140, should be further validated for accuracy. Procheck and other free web-based programs check the stereochemical quality of a model's structure, such as deviations from ideal bonding angles and bond length, and produce a Ramachandran plot identifying outliers and clashing contacts which is a standard part of structure analysis before deposition (Praznikar et al., 2019). Further, after completion of the project, selected HPs and identified similarly sequenced proteins with established annotation should undergo additional comparisons to support re-annotation conclusions. Examples of additional computational analyses include multiple sequence alignment, physiochemical properties, and phylogeny tree builder, performed by programs such as PROMALS3D (Pei et al., 2008) or CLUSTAL Omega (Thompson et al., 1994; Madeira et al., 2019), ExPASy ProtParam (Artimo et al., 2012), and the PHYLIP suite (Lim and Zhang, 1999; Retief, 2000; Abdennadher and Boesch, 2007), respectively. These additional analyses make the phylogenetic relations approach for selecting HPs a complete bioinformatics project that is ideal for undergraduate honor's and capstone projects or as part of graduate work where scientific rationale for the study is needed but students lack access to a laboratory for further experimental examination.

Knowledgebases Are Constantly Improving

The overall goal of the Hypothetical Protein Characterization Project from a student perspective is to assist in improving genome annotation. To emphasize the speed at which knowledgebases update as well as the importance of improving genome annotation, we re-ran the project on ORF8 on June 10, 2020, to see how results may have changed in a short time under substantial pressure to computationally and experimentally characterize SARS-CoV2 due to the COVID-19 pandemic. We found that NCBI Protein database updated the protein's description in the public record from HP to ORF8 protein (Severe acute respiratory syndrome coronavirus 2). The record now shows a corona_NS8 domain for ORF8 where it was not listed in March despite previous CDD and Pfam identification. In March, CDD and Pfam described the corona_NS8 domain as a functionally uncharacterized superfamily domain conserved in coronaviruses. While the statistical values have not changed, now the description details a superfamily of immunoglobulin (Ig) domain proteins without mention of anything still being uncharacterized. While UniProt did not have an entry for ORF8 in March and still does not have one using the same identifies as NCBI, UniProt has now added ORF8 as a 121 amino acid long, non-structural protein 8 under the identifier P0DTC8 (NS8_SARS2). We used the WayBack Machine web archival site[2] to confirm

[2] https://archive.org/web/

P0DTC8 did not exist in UniProt in March. 3D predictive modeling and cellular location results did not change between March and June, though we expect modeling for ORF8 to improve when the structure of ORF8 or one of its homologs has been elucidated.

Given the high number of newly sequenced genomes deposited regularly to public knowledgebases, there will be plenty of HPs for use in the Hypothetical Protein Characterization Project for years to come. Further, proteins with vague annotation descriptions (*e.g.*, membrane protein) and no gene symbol may also benefit from characterization using this project. The quick update in the annotation of ORF8 due to the COVID-19 pandemic highlights how manual review can improve genome annotation when ample resources are available. This paper provides a tool that turns students into manual reviewers of genome annotation while learning valuable interdisciplinary concepts. Application of the Hypothetical Protein Characterization Project in educational settings worldwide has the potential to significantly improve public knowledgebases and the scientific conclusions derived from their information.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: https://www.ncbi.nlm.nih.gov/, YP_009724396.1; https://www.ncbi.nlm.nih.gov/, WP_002214142; https://www.ncbi.nlm.nih.gov/, AKI46902.1; https://www.ncbi.nlm.nih.gov/, OLC18526.1.

AUTHOR CONTRIBUTIONS

LH conceived the presented idea, developed the theory, and performed the computations. ZA verified the computations and manuscript citations. LH took the lead in writing the manuscript in consultation with SG. All authors contributed to the article and approved the submitted version.

FUNDING

Support for this work has been generously provided by M.E. Davenport Legacy Endowment Grants.

ACKNOWLEDGMENTS

The authors would like to acknowledge Quantitative Undergraduate Biology Education and Synthesis Consortium for the mentorship and support to establish the "Online Faculty Mentoring Network to Develop Video Tutorials for Computational Genomics" YouTube channel. Further, the authors also would like to thank Kuana School, Jessica Marklevitz, Blue Goad, and Dr. Masayuki Shibata for their contributions toward the development of the concepts presented in this manuscript. And also thank you to Ahlam Kader for her manuscript review.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.577497/full#supplementary-material

REFERENCES

Abdennadher, N., and Boesch, R. (2007). Porting PHYLIP phylogenetic package on the desktop GRID platform XtremWeb-CH. *Stud. Health Technol. Inform.* 126, 55–64.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Altschul, S. F., and Koonin, E. V. (1998). Iterated profile searches with PSI-BLAST–a tool for discovery in protein databases. *Trends Biochem. Sci.* 23, 444–447.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402. doi: 10.1093/nar/25.17.3389

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014). SCOP2 prototype: a new approach to protein structure mining. *Nucleic Acids Res.* 42, D310–D314. doi: 10.1093/nar/gkt1242

Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2020). The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic Acids Res.* 48, D376–D382. doi: 10.1093/nar/gkz1064

Araujo, C. L., Blanco, I., Souza, L., Tiwari, S., Pereira, L. C., Ghosh, P., et al. (2020). In silico functional prediction of hypothetical proteins from the core genome of *Corynebacterium pseudotuberculosis* biovar *ovis*. *PeerJ* 8:e9643. doi: 10.7717/peerj.9643

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., de Castro, E., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* 40, W597–W603. doi: 10.1093/nar/gks400

Bank, P. D. (1971). Protein data bank. *Nat. New Biol.* 233:223.

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., et al. (2011). NCBI GEO: archive for functional genomics data sets–10 years on. *Nucleic Acids Res.* 39, D1005–D1010. doi: 10.1093/nar/gkq1184

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Berman, H. M., Kleywegt, G. J., Nakamura, H., and Markley, J. L. (2014). The protein data bank archive as an open data resource. *J. Comput. Aided Mol. Des.* 28, 1009–1014. doi: 10.1007/s10822-014-9770-y

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242. doi: 10.1093/nar/28.1.235

Bhagwat, M., and Aravind, L. (2007). "Psi-blast tutorial," in *Comparative Genomics*, ed. N. H. Bergman, (Berlin: Springer), 177–186.

Bharat Siva Varma, P., Adimulam, Y. B., and Kodukula, S. (2015). In silico functional annotation of a hypothetical protein from *Staphylococcus aureus*. *J. Infect. Public Health* 8, 526–532. doi: 10.1016/j.jiph.2015.03.007

Brown, G. R., Hem, V., Katz, K. S., Ovetsky, M., Wallin, C., Ermolaeva, O., et al. (2015). Gene: a gene-centered information resource at NCBI. *Nucleic Acids Res.* 43, D36–D42. doi: 10.1093/nar/gku1055

Brown, T. A. (ed.). (2002). "Understanding a genome sequence," in *Genomes*, 2nd Edn, (Oxford: Wiley-Liss).

Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., and Velankar, S. (2017). Protein data bank (PDB): the single global macromolecular structure archive. *Methods Mol. Biol.* 1607, 627–641. doi: 10.1007/978-1-4939-7000-1_26

Chang, K. Y., and Yang, J. R. (2013). Analysis and prediction of highly effective antiviral peptides based on random forests. *PLoS One* 8:e70166. doi: 10.1371/journal.pone.0070166

Chen, C. C., Hwang, J. K., and Yang, J. M. (2006). (PS)2: protein structure prediction server. *Nucleic Acids Res.* 34, W152–W157. doi: 10.1093/nar/gkl187

Chen, C. C., Hwang, J. K., and Yang, J. M. (2009). (PS)2-v2: template-based protein structure prediction server. *BMC Bioinformatics* 10:366. doi: 10.1186/1471-2105-10-366

Coordinators, N. R. (2018). Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 46, D8–D13. doi: 10.1093/nar/gkx1095

da Costa, W. L. O., Araujo, C. L. A., Dias, L. M., Pereira, L. C. S., Alves, J. T. C., Araujo, F. A., et al. (2018). Functional annotation of hypothetical proteins from the *Exiguobacterium antarcticum* strain B7 reveals proteins involved in adaptation to extreme environments, including high arsenic resistance. *PLoS One* 13:e0198965. doi: 10.1371/journal.pone.0198965

Dorden, S., and Mahadevan, P. (2015). Functional prediction of hypothetical proteins in human adenoviruses. *Bioinformation* 11, 466–473. doi: 10.6026/97320630011466

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30, 207–210.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., et al. (2010). The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222. doi: 10.1093/nar/gkp985

Gasteiger, E., Hoogland, C., Gattiker, A., Wilkins, M. R., Appel, R. D., and Bairoch, A. (2005). "Protein identification and analysis tools on the ExPASy server," in *The Proteomics Protocols Handbook*, ed. J. M. Walker, (Berlin: Springer), 571–607.

Gazi, M. A., Mahmud, S., Fahim, S. M., Kibria, M. G., Palit, P., Islam, M. R., et al. (2018). Functional prediction of hypothetical proteins from *Shigella flexneri* and validation of the predicted models by using ROC curve analysis. *Genomics Inform.* 16:e26. doi: 10.5808/GI.2018.16.4.e26

Geer, L. Y., Domrachev, M., Lipman, D. J., and Bryant, S. H. (2002). CDART: protein homology by domain architecture. *Genome Res.* 12, 1619–1623. doi: 10.1101/gr.278202

Goad, B., and Harris, L. K. (2018). Identification and prioritization of macrolide resistance genes with hypothetical annotation in *Streptococcus pneumoniae*. *Bioinformation* 14, 488–498.

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001). Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* 313, 903–919. doi: 10.1006/jmbi.2001.5080

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14, 378–379. doi: 10.1093/bioinformatics/14.4.378

Horton, P., Park, K. J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C. J., et al. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35, W585–W587. doi: 10.1093/nar/gkm259

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* 4, 44–57. doi: 10.1038/nprot.2008.211

Ijaq, J., Chandrasekharan, M., Poddar, R., Bethi, N., and Sundararajan, V. S. (2015). Annotation and curation of uncharacterized proteins- challenges. *Front. Genet.* 6:119. doi: 10.3389/fgene.2015.00119

Imam, N., Alam, A., Ali, R., Siddiqui, M. F., Ali, S., Malik, M. Z., et al. (2019). In silico characterization of hypothetical proteins from *Orientia tsutsugamushi* str. Karp uncovers virulence genes. *Heliyon* 5:e02734. doi: 10.1016/j.heliyon.2019.e02734

Islam, M. S., Shahik, S. M., Sohel, M., Patwary, N. I., and Hasan, M. A. (2015). In silico structural and functional annotation of hypothetical proteins of *Vibrio cholerae* O139. *Genomics Inform.* 13, 53–59. doi: 10.5808/GI.2015.13.2.53

Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., and Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* 10, 845–858. doi: 10.1038/nprot.2015.053

Kolker, E., Makarova, K. S., Shabalina, S., Picone, A. F., Purvine, S., Holzman, T., et al. (2004). Identification and functional analysis of 'hypothetical' genes expressed in *Haemophilus influenzae*. *Nucleic Acids Res.* 32, 2353–2361. doi: 10.1093/nar/gkh555

Koonin, E. V., and Galperin, M. Y. (eds). (2003). "Genome annotation and analysis," in *Sequence—Evolution—Function*, (Boston, MA: Springer), 193–226.

Kuhn, M., von Mering, C., Campillos, M., Jensen, L. J., and Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Res.* 36, D684–D688. doi: 10.1093/nar/gkm795

Kumar, N., Hoque, M. A., and Sugimoto, M. (2018). Robust volcano plot: identification of differential metabolites in the presence of outliers. *BMC Bioinformatics* 19:128. doi: 10.1186/s12859-018-2117-2

Letunic, I., and Bork, P. (2018). 20 years of the SMART protein domain annotation resource. *Nucleic Acids Res.* 46, D493–D496. doi: 10.1093/nar/gkx922

Lewin, H. A., Robinson, G. E., Kress, W. J., Baker, W. J., Coddington, J., Crandall, K. A., et al. (2018). Earth BioGenome project: sequencing life for the future of life. *Proc. Natl. Acad. Sci. U.S.A.* 115, 4325–4333. doi: 10.1073/pnas.1720115115

Lewis, T. E., Sillitoe, I., Dawson, N., Lam, S. D., Clarke, T., Lee, D., et al. (2018). Gene3D: extensive prediction of globular domains in proteins. *Nucleic Acids Res.* 46, D435–D439. doi: 10.1093/nar/gkx1069

Li, W. (2012). Volcano plots in analyzing differential expressions with mRNA microarrays. *J. Bioinform. Comput. Biol.* 10:1231003. doi: 10.1142/S0219720012310038

Lim, A., and Zhang, L. (1999). WebPHYLIP: a web interface to PHYLIP. *Bioinformatics* 15, 1068–1069. doi: 10.1093/bioinformatics/15.12.1068

Lu, S., Wang, J., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2020). CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.* 48, D265–D268. doi: 10.1093/nar/gkz991

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., et al. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic Acids Res.* 47, W636–W641. doi: 10.1093/nar/gkz268

Mahmood, M. S., Ashraf, N. M., Bilal, M., Ashraf, F., Hussain, A., Zubair, M., et al. (2016). In silico structural and functional characterization of a hypothetical protein of *Vaccinia virus*. *J. Biochem. Biotechnol. Biomater.* 1, 28–35.

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., et al. (2017). CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.* 45, D200–D203. doi: 10.1093/nar/gkw1129

Marchler-Bauer, A., and Bryant, S. H. (2004). CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32, W327–W331. doi: 10.1093/nar/gkh454

Marklevitz, J., and Harris, L. K. (2016). Prediction driven functional annotation of hypothetical proteins in the major facilitator superfamily of *S. aureus* NCTC 8325. *Bioinformation* 12, 254–262. doi: 10.6026/97320630012254

Mitaku, S., and Hirokawa, T. (1999). Physicochemical factors for discriminating between soluble and membrane proteins: hydrophobicity of helical segments and protein length. *Protein Eng.* 12, 953–957. doi: 10.1093/protein/12.11.953

Mitaku, S., Hirokawa, T., and Tsuji, T. (2002). Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* 18, 608–616. doi: 10.1093/bioinformatics/18.4.608

Mohan, R., and Venugopal, S. (2012). Computational structural and functional analysis of hypothetical proteins of *Staphylococcus aureus*. *Bioinformation* 8, 722–728. doi: 10.6026/97320630008722

Naveed, M., Tehreem, S., Usman, M., Chaudhry, Z., and Abbas, G. (2017). Structural and functional annotation of hypothetical proteins of human adenovirus: prioritizing the novel drug targets. *BMC Res. Notes* 10:706. doi: 10.1186/s13104-017-2992-z

Omeershffudin, U. N. M., and Kumar, S. (2019). In silico approach for mining of potential drug targets from hypothetical proteins of bacterial proteome. *Int. J. Mol. Biol. Open Access* 4, 145–152.

Pavlovic-Lazetic, G. M., Mitic, N. S., Tomovic, A. M., Pavlovic, M. D., and Beljanski, M. V. (2005). SARS-CoV genome polymorphism: a bioinformatics study. *Genomics Proteomics Bioinformatics* 3, 18–35.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Curr. Protoc. Bioinformatics* 42, 3.1.1–3.1.8. doi: 10.1002/0471250953.bi0301s42

Pei, J., Kim, B. H., and Grishin, N. V. (2008). PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucleic Acids Res.* 36, 2295–2300. doi: 10.1093/nar/gkn072

Pranavathiyani, G., Prava, J., Rajeev, A. C., and Pan, A. (2020). Novel target exploration from hypothetical proteins of *Klebsiella pneumoniae* MGH 78578 reveals a protein involved in host-pathogen interaction. *Front. Cell. Infect. Microbiol.* 10:109. doi: 10.3389/fcimb.2020.00109

Praznikar, J., Tomic, M., and Turk, D. (2019). Validation and quality assessment of macromolecular structures using complex network analysis. *Sci. Rep.* 9:1678.

Raj, U., Sharma, A. K., Aier, I., and Varadwaj, P. K. (2017). In silico characterization of hypothetical proteins obtained from *Mycobacterium tuberculosis* H37Rv. *Netw. Model. Anal. Health Inform. Bioinform.* 6:5. doi: 10.1007/s13721-017-0147-8

Retief, J. D. (2000). Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* 132, 243–258.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. *Nat. Protoc.* 5, 725–738. doi: 10.1038/nprot.2010.5

Sali, A., and Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779–815. doi: 10.1006/jmbi.1993.1626

Sammut, S. J., Finn, R. D., and Bateman, A. (2008). Pfam 10 years on: 10,000 families and still growing. *Brief. Bioinform.* 9, 210–219. doi: 10.1093/bib/bbn010

School, K., Marklevitz, J., Schram, W. K., and Harris, L. K. (2016). Predictive characterization of hypothetical proteins in *Staphylococcus aureus* NCTC 8325. *Bioinformation* 12, 209–220. doi: 10.6026/97320630012209

Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* 95, 5857–5864. doi: 10.1073/pnas.95.11.5857

Schwede, T., Kopp, J., Guex, N., and Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* 31, 3381–3385. doi: 10.1093/nar/gkg520

Shahbaaz, M., Bisetty, K., Ahmad, F., and Hassan, M. I. (2015). In silico approaches for the identification of virulence candidates amongst hypothetical proteins of *Mycoplasma pneumoniae* 309. *Comput. Biol. Chem.* 59(Pt A), 67–80. doi: 10.1016/j.compbiolchem.2015.09.007

Sillitoe, I., Dawson, N., Lewis, T. E., Das, S., Lees, J. G., Ashford, P., et al. (2019). CATH: expanding the horizons of structure-based functional annotations for genome sequences. *Nucleic Acids Res.* 47, D280–D284. doi: 10.1093/nar/gky1097

Sivashankari, S., and Shanmughavel, P. (2006). Functional annotation of hypothetical proteins - a review. *Bioinformation* 1, 335–338. doi: 10.6026/97320630001335

Smits, T. H. M. (2019). The importance of genome sequence quality to microbial comparative genomics. *BMC Genomics* 20:662. doi: 10.1186/s12864-019-6014-5

Snel, B., Lehmann, G., Bork, P., and Huynen, M. A. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. *Nucleic Acids Res.* 28, 3442–3444. doi: 10.1093/nar/28.18.3442

Sonnhammer, E. L., Eddy, S. R., Birney, E., Bateman, A., and Durbin, R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* 26, 320–322. doi: 10.1093/nar/26.1.320

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., et al. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* 47, D607–D613. doi: 10.1093/nar/gky1131

Szklarczyk, D., Santos, A., von Mering, C., Jensen, L. J., Bork, P., and Kuhn, M. (2016). STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* 44, D380–D384. doi: 10.1093/nar/gkv1277

Tatusova, T., DiCuccio, M., Badretdin, A., Chetvernin, V., Nawrocki, E. P., Zaslavsky, L., et al. (2016). NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* 44, 6614–6624. doi: 10.1093/nar/gkw569

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673–4680. doi: 10.1093/nar/22.22.4673

Tipney, H., and Hunter, L. (2010). An introduction to effective use of enrichment analysis software. *Hum. Genomics* 4, 202–206.

UniProt, C. (2019). UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* 47, D506–D515. doi: 10.1093/nar/gky1049

Wang, H., Li, X., Li, T., Zhang, S., Wang, L., Wu, X., et al. (2020). The genetic sequence, origin, and diagnosis of SARS-CoV-2. *Eur. J. Clin. Microbiol. Infect. Dis.* doi: 10.1007/s10096-020-03899-4 [Epub ahead of print].

Wass, M. N., Kelley, L. A., and Sternberg, M. J. (2010). 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.* 38, W469–W473. doi: 10.1093/nar/gkq406

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., et al. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* 46, W296–W303. doi: 10.1093/nar/gky427

Webb, B., and Sali, A. (2014). Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics* 47, 5.6.1–5.6.37. doi: 10.1002/0471250953.bi0506s47

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13, 329–342. doi: 10.1038/nrg3174

Yang, J., and Zhang, Y. (2015). I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res.* 43, W174–W181. doi: 10.1093/nar/gkv342

Yang, Z., Zeng, X., and Tsui, S. K. (2019). Investigating function roles of hypothetical proteins encoded by the *Mycobacterium tuberculosis* H37Rv genome. *BMC Genomics* 20:394. doi: 10.1186/s12864-019-5746-6

Yegambaram, K., Bulloch, E. M., and Kingston, R. L. (2013). Protein domain definition should allow for conditional disorder. *Protein Sci.* 22, 1502–1518. doi: 10.1002/pro.2336

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615. doi: 10.1093/bioinformatics/btq249

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Genetics

PERSPECTIVE
published: 16 December 2020
doi: 10.3389/fgene.2020.573992

Using The Cancer Genome Atlas as an Inquiry Tool in the Undergraduate Classroom

William Hankey[1]*, Nicholas Zanghi[2], Mackenzie M. Crow[2], Whitney H. Dow[2], Austin Kratz[3,4], Ashley M. Robinson[2], Meaghan R. Robinson[2] and Verónica A. Segarra[2]*

[1] Department of Pathology, Duke Cancer Center, Duke University, Durham, NC, United States, [2] Department of Biology, High Point University, High Point, NC, United States, [3] Department of Chemistry, High Point University, High Point, NC, United States, [4] Department of Physics, High Point University, High Point, NC, United States

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Alexander V. Favorov,
Johns Hopkins University,
United States
Cyriac Kandoth,
Cornell University, United States

***Correspondence:**
William Hankey
william.hankey@duke.edu
Verónica A. Segarra
vsegarra@highpoint.edu

Specialty section:
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Genetics

Received: 18 June 2020
Accepted: 24 November 2020
Published: 16 December 2020

Citation:
Hankey W, Zanghi N, Crow MM,
Dow WH, Kratz A, Robinson AM,
Robinson MR and Segarra VA (2020)
Using The Cancer Genome Atlas as
an Inquiry Tool in the Undergraduate
Classroom. Front. Genet. 11:573992.
doi: 10.3389/fgene.2020.573992

Undergraduate students in the biomedical sciences are often interested in future health-focused careers. This presents opportunities for instructors in genetics, molecular biology, and cancer biology to capture their attention using lab experiences built around clinically relevant data. As biomedical science in general becomes increasingly dependent on high-throughput data, well-established scientific databases such as The Cancer Genome Atlas (TCGA) have become publicly available tools for medically relevant inquiry. The best feature of this database is that it bridges the molecular features of cancer to human clinical outcomes—allowing students to see a direct connection between the molecular sciences and their future professions. We have developed and tested a learning module that leverages the power of TCGA datasets to engage students to use the data to generate and test hypotheses and to apply statistical tests to evaluate significance.

Keywords: bioinformatics, cancer, genomics, cancer genomics, undergraduate teaching and learning

INTRODUCTION

While many undergraduates are interested in becoming medical doctors and declare "pre-med" early in their academic careers, it is predicted that by 2032 the United States will face a shortage of between 46,900 and 121,900 physicians (Dall et al., 2019). One of the factors likely to exacerbate this projected shortage is the high attrition rates of undergraduates from the premedical academic track (Lin et al., 2013). In fact, many of the empirical studies recorded in the scientific literature related to undergraduate premedical students are focused on documenting and better understanding attrition from the premedical track (Lin et al., 2013). High attrition rates in undergraduate premedical tracks have been found to be influenced by a variety of factors including loss of interest and negative experiences in required courses (Lin et al., 2013).

Student interest and persistence in STEM careers can be increased and strengthened through participation in Course-based Undergraduate Research Experiences (CUREs) as part of the curriculum (Estrada et al., 2016). These findings suggest that one of the ways in which student persistence in undergraduate premedical programs can be increased is through relevant CURE experiences that highlight clinically relevant data and its applications.

While undergraduate access to clinical research experiences is limited, the biomedical sciences are becoming increasingly dependent on high-throughput data, and well-established scientific databases such as The Cancer Genome Atlas (TCGA) have become publicly available tools for medically relevant inquiry (Cancer Genome Atlas Network, 2012; Cerami et al., 2012; Gao et al., 2013). These databases are increasingly being recognized as resources available for undergraduate teaching (Coughlan, 2020).

Furthermore, there is currently a need for physicians and health professionals to recognize and use the power of cutting edge genomics to inform diagnosis and treatments for their patients (Rubanovich et al., 2018). Through the use of clinically relevant genomic datasets like the ones found in TCGA in the undergraduate classroom, we can raise awareness for the relevance of these resources in medicine early on in the training of these individuals (Schoenborn et al., 2019).

It is also important to point out the increasing need for scientific literacy, pro-science attitudes, and evidence-based decision-making among non-majors in a variety of different disciplines (Ballen et al., 2017). These skills, including scientific literacy, can be developed using CURE experiences and inquiry-based modules in the non-majors classroom (Ballen et al., 2017; Segarra et al., 2018).

We have developed and tested a learning module that leverages the power of TCGA datasets to engage students in inquiry-based clinical research in the context of cancer—a human disease that is of universal relevance. Our module allows students to not only generate and test hypotheses with clinical relevance, but also apply statistical tests to evaluate significance. Continuing to refine such activities to better cultivate engagement in and comfort with data-based decision-making will better position us to foster interest, persistence, and scientific literacy among undergraduate science majors both inside and outside of the premedical track, as well as non-majors preparing to enter an increasingly data-driven workplace.

METHODOLOGY

Accessing TCGA Datasets

The Cancer Genome Atlas data were accessed by the course instructor through cBioportal[1], a widely used web interface that provides access to public cancer genomics datasets (Cancer Genome Atlas Network, 2012; Cerami et al., 2012; Gao et al., 2013). Breast cancer was selected as a focus because of the increased likelihood for the intended audience members to make personal connections to a highly prevalent cancer type with a significant impact on human health, and because of the convenience of introducing the genomic data starting with familiar genes such as *BRCA1*, *BRCA2*, and the gene encoding p53 (*TP53*) that had previously been discussed in the lecture component of the class. TCGA was chosen as a data source for the combination of high-quality genomic and associated clinical data characteristic of TCGA datasets in general and the high sample size of the available datasets. The TCGA Breast Invasive Carcinoma dataset associated with the 2015 publication in *Cell* (Ciriello et al., 2015) was specifically chosen from among the four available TCGA Breast Invasive Carcinoma datasets because of its combination of mutation data and copy number alteration data, as well as its inclusion of stage among the clinical data variables (study ID "brca_tcga_pub2015"; https://www.cbioportal.org/study/summary?id==brca_tcga_pub2015). It should be noted that the original data set was composed of a total of 818 patient samples—817 from primary and 1 from metastatic tumors. Only data from the 817 primary tumor samples were included in the student analysis. The metastatic sample was excluded in order to present the students with a comparable and consistent group of samples for analysis.

While mutation and copy number data were available in the dataset for more than 20,000 genes, a more focused subset of 16 total genes was selected to provide to the students. This subset was narrow enough facilitate visualization of the complete dataset and analysis by first-time bioinformaticists in Microsoft Excel, but diverse enough to include examples fitting several different patterns. The list began with well-known cancer-associated genes previously discussed in the course (*BRCA1*, *BRCA2*, *TP53*), then added genes that were among the most frequently targeted in breast cancer by known pathogenic mutations (*PIK3CA*, *CDH1*, *GATA3*, *MAP3K1*, *KMT2C*, and *AKT1*), amplifications (*MYC*, *CCND1*, and *ERBB2*), or deletions (*RB1*, *PTEN*). These high-frequency targets of mutations and copy number alterations were identified by selecting the dataset of interest (Cancer Genome Atlas Network, 2012) from the cBioportal menu and using the Explore Selected Studies function to view the Summary of findings. The genes encoding β-actin (*ACTB*) and hemoglobin subunit β (*HBB*) were added to the list in order to function as recognizable negative control genes generally not associated with cancer. Once the list of sixteen breast cancer-relevant genes and controls was determined, the 16 gene names were entered as a list into the cBioportal website to access genomics data for this subset using the Query by Gene function. For each gene of interest, genetic mutation data and copy number alteration data were separately accessed for all 817 tumors in the dataset from the Download section of the site, selecting the Tab Delimited Format option. Clinical data were accessed through the cBioportal site using the Explore Selected Studies function and the Clinical Data tab. A limited subset of clinical characteristics were downloaded, with each characteristic chosen to help illustrate a different point or to enable the students to test a different hypothesis. The majority of clinical variables were categorical, facilitating the use of 2 × 2 tables to test association between the clinical category and the status of a gene as mutated/unmutated, etc. The 15 characteristics were Informed Consent by Patient (Yes/No), Diagnosis Age, Cancer Type, Race Category, Ethnicity Category, Sex, Disease Stage (I–IV), Treatment Outcome (Living Disease-Free/Living with Tumor/Recurred, or Progressed/Deceased), Time from Treatment to Recurrence (Months), Time from Treatment to Death (Months), Time from Treatment to Most Recent Contact (Months), ER Status (by Staining), PR Status (by Staining), HER2 Status (by Staining), and Total Number of Mutations. Similar to

[1] https://www.cbioportal.org/

the mutation and copy number data files, the clinical data were arrayed so that the clinical variables were each assigned a different column, while the 817 tumors were each assigned a different row (**Supplementary Appendix 1**).

Combining Genetic and Clinical TCGA Data in Microsoft Excel

Initially, the separate Mutations and Copy-Number Alterations files were integrated into a single Excel file by alphabetizing the list of samples in each file by Patient ID and integrating the columns along matching rows. The instructor then sought to integrate the mutation status and the copy number status into a single column for each gene, stating only the change in that gene most relevant to the disease. For example, if the Copy Number Alteration column for *TP53* listed the gene as Amplified in a particular tumor, while the Mutations column for *TP53* listed it as a known Pathogenic Mutant in that same tumor, the merging of those two columns into one *TP53* Status column listed it as Pathogenic Mutant for that tumor. On the other hand, if the Copy Number Alteration column for *TP53* listed the gene as Amplified in a particular tumor, while the Mutations column for *TP53* listed it as a Mutant of Unknown Significance in that same tumor, the merging of those two columns into one *TP53* Status column listed it as Amplified for that tumor. The resulting Excel file containing gene status data was then integrated with the Clinical Data file into a single Excel file by alphabetizing the list samples in each file by Patient ID and integrating them along matching rows. The resulting file contained 16 columns of genetic data and 15 columns of clinical data, with 817 rows of tumor samples, each representing a different patient (**Supplementary Appendix 1**).

Generation of the Worksheet

The instructor designed an assignment to introduce students to the kinds of research hypotheses that are testable using the combination of genetic and clinical data. The initial assignment was generated in the form of a worksheet (**Supplementary Appendix 2**), which consisted of five different tables. Categorical clinical and/or genetic characteristics were listed along the *x*- and *y*-axes, and students were asked to count how many tumors from the dataset possessed each combination of characteristics. Students first determined how many of the patients classified as Living Disease-Free, Living with Tumor, Recurred or Progressed, and Deceased were diagnosed with Stage I vs. II vs. III vs. IV tumors. This comparison of stage and outcome was selected to illustrate a well-known clinical association and presented students with an opportunity to test whether the counts matched their expectations. Students then determined how many of the patients classified as Living Disease-Free, Living with Tumor, Recurred or Progressed, and Deceased harbored vs. did not harbor pathogenic mutations/deletions in *TP53*, *BRCA1*, or *BRCA2*. Students were already familiar with all three genes as well-known tumor suppressors in breast cancer, and were able to formulate hypotheses about how mutations in each gene might associate with clinical outcome. In the final table, students were asked to calculate the total number of tumors with pathogenic mutant, mutant of unknown significance, amplified, and deleted genotypes, for each of the sixteen genes. Since most of these genes were less familiar, students would have the opportunity to collect the data without bias, and then to use them to form a hypothesis about each gene's status as an oncogene, tumor suppressor gene, or neither.

Generation of Instructions for Sorting Tumors in Microsoft Excel

Students came into the assignment with heterogeneous backgrounds using Microsoft Excel for similar tasks, and were provided with general instructions to help them complete the worksheet (**Supplementary Appendix 3**). The Sort and Filter function in Excel was recommended as a critical tool for organizing data into subsets according to a particular genomic or clinical characteristic. Within each subset, students were recommended to count occurrences of the associated characteristic using the COUNTIF function in combination with quotation marks around the text of interest.

Generation of a Microsoft Excel File to Support Statistical Analysis

As a follow-up assignment, students were asked to use the counts data from their completed worksheet to generate one hypothesis about the association of two variables. They would then construct a 2 × 2 table and perform a test for statistically significant association. The chi-square test of independence was recommended as an applicable statistical test that can be performed using Excel. To facilitate their introduction to this statistical test, a template Excel file was constructed into which the students could enter their 2 × 2 table (**Supplementary Appendix 4**). The file would then use these observed counts to calculate the expected counts, determine the test statistic, and generate a *p*-value.

CLASSROOM IMPLEMENTATION

The documents/data described above (also see **Supplementary Materials**) were used to create and implement a bioinformatics laboratory experience during two 3-h lab periods near the conclusion of an upper-level undergraduate Cancer Biology course. This activity can also be implemented in a bioinformatics or genetics course and is particularly well suited to be implemented remotely in the context of online teaching.

Step 1: *Introduce students to the Microsoft Excel file containing data subset of interest.*

Students were introduced to the data subset of interest, including the kind of information each column and row contained (**Supplementary Appendix 1**).

Step 2: *Students complete a worksheet composed of 2 × 2 tables that measure associations between presence/absence of a mutation and categorical clinical phenotypes.*

Students were given the opportunity to increase their familiarity with the dataset of interest (**Supplementary Appendix 1**) by completing an Excel worksheet (**Supplementary Appendix 2**) that required them to identify the data relevant to different categories. To help students sift through the data, they were provided with tips for sorting tumor data in Excel (**Supplementary Appendix 3**).

Step 3: *Students articulate a new association of interest to test (research question), create/complete the appropriate 2 × 2 tables, and calculate statistical significance of association.*

Using the data as a guide, students were given the opportunity to come up with their own association or research question to test (**Table 1**). Students had to examine the data provided and decide which two categorical variables they wanted to use to test an association. Students were introduced to the chi-square test of independence and its relevance to categorical data. To facilitate students performing the relevant statistics, an Excel file template was provided (**Supplementary Appendix 4**). Before beginning this portion of the assignment, the instructor demonstrated the process from selection of an association of interest and 2 × 2 table construction, all the way to statistical analysis.

TABLE 1 | Representative research questions answered by students using TCGA Breast Invasive Carcinoma datasets.

Research question	Categorical variables being compared	*p*-value
Are pathogenic *PIK3CA* gene mutations associated with poor clinical outcomes (not living disease-free) for breast cancer?	Wildtype *PIK3CA* vs. pathogenic mutations in *PIK3CA* Good clinical outcome (living disease-free) vs. Poor clinical outcome (not living disease-free)	0.24
Is the wildtype *BRCA2* gene associated with good (living disease-free) breast cancer clinical outcomes?	Wildtype *BRCA2* vs. pathogenic mutant *BRCA2* Good clinical outcome (living disease-free) vs. Poor clinical outcome (not living disease-free)	0.93
Are *BRCA1* gene tumor mutations associated with poor (not living disease-free) breast cancer outcomes?	Wildtype *BRCA1* vs. mutated BRCA1 gene Good clinical outcome (living disease-free) vs. Poor clinical outcome (not living disease-free)	0.60
Are pathogenic *TP53* mutations associated with more advanced (Stages II/III/IV) stages of cancer?	Wildtype *TP53* vs. pathogenic *TP53* gene mutations Early (Stage I) vs. advanced stages of cancer (Stages II/III/IV)	0.14
Are pathogenic *BRCA1* mutations associated with breast cancer recurrence?	Wildtype *BRCA1* vs. pathogenic *BRCA1* mutants Good clinical outcome (living disease-free) vs. Poor clinical outcome (living but tumor recurred/progressed)	0.01
Are patients living disease-free more likely to have been diagnosed early stage breast cancer (Stages I/II)?	Living disease-free vs. Not living disease free Early stage (Stages I/II) vs. late stage (III/IV) cancer	2×10^{-6}

For Step 3 in Classroom Implementation, students articulate a new association of interest to test (research question), create/complete 2 × 2 tables, and calculate its statistical significance. Shown in this table are representative research questions (associations being tested), including the categorical variables being tested and the determined statistical significance (p-value) of the association. Associations that were not independent from each other have a p-value less or equal than 0.05.

Microsoft Excel was selected for this activity due to its familiarity to the majority of undergraduate students as both a calculator and a tool for generating scientific figures. Thus, it serves as a comfortable starting point in which the dimensions of the dataset can be visualized and new functions and calculations for data analysis can be introduced. At the same time, it is important to note the caveat that Microsoft Excel is increasingly recognized as a flawed platform for statistical analysis. In comparison to the open-source programming language R, which has become a preferred platform for many research applications of statistics, Excel is considered the less reproducible and more error-prone option (Ziemann et al., 2016). A key advantage of R is the ability to record and share in a transparent way the

TABLE 2 | Student feedback in response to each of the steps of the TCGA module.

Step 1 of the module: *Introduce students to the Microsoft Excel file containing data subset of interest.*

Student feedback

Spreadsheet with TCGA data made it clear how large the pool of genome data from cancer patients can be and how these data can be used to determine relationships between mutations and clinical patient outcomes.

Humbling to think about the data on the Excel spreadsheet coming from actual patients, some who died, and some who recovered and were able to continue living cancer-free

While spreadsheet was well organized, it took some time and exploring to understand and get a feel for the information in it.

I finally understood what it means for a patient to have “triple negative” breast cancer at the molecular level. Seeing all the potential options for these receptors lined up on the spreadsheet drove the point home.

I would be interested in learning how to create the spreadsheet with data entirely from scratch using information posted in TCGA.

Step 2 of the module: *Students complete a worksheet composed of 2 × 2 tables that measure associations between presence/absence of a mutation and categorical clinical phenotypes.*

Student feedback

Completing the worksheet helped with understanding information on dataset.

I learned new easy excel functions (like COUNTIF function) that will likely be useful later on in data and statistical analysis.

Completing the worksheet was time consuming and could easily be combined with the research question creation and analysis. This would have allowed me to come up with a question while the information in the data set is still fresh in my head.

I liked the worksheet because I was able to turn the data into relationships and percentages that were applicable to real human disease.

Step 3 of the module: *Students articulate a new association of interest to test (research question), create/complete the appropriate 2 × 2 tables, and calculate statistical significance of association.*

Student feedback

You always hear of the statistics of different cancers and stages, but with the data we were able to see the actual outcomes of real patients for our own research question, which made it more real than reading about it in a textbook.

This is the first time I have actually gotten to make my own experiment with clinical data from real humans.

I was overwhelmed at first by the amount of research questions that could be addressed with the data provided.

I tried testing several associations in the hopes of getting a statistically significant difference, but was not successful.

For similar student feedback or statements, one representative comment was chosen and listed on this table.

steps taken to organize and analyze the data (Incerti et al., 2019). While we felt that the benefits of Excel outweighed the caveats in this particular application, future adaptations of this exercise might consider introducing students instead to programming in R or to other commercial software packages for statistics and data science, such as Stata, SPSS, SAS, or JMP. Substitution of these tools for Excel might create an additional obstacle to the accessibility of key concepts to students, but would likely benefit those students who might continue to use these programs in their future research.

DISCUSSION

While, at first, students had difficulty managing the large amount of information that was provided, sharing strategies to sort and count data using Excel helped them gain confidence in using the dataset to complete Steps 2 and 3 described above. In fact, all students were ultimately able to get perfect grades on their practice worksheets (**Supplementary Appendix 2**).

Table 1 provides representative research questions that were answered using the breast cancer tumor data available. In general, many of the associations tested were not statistically significant. This is likely due to shortcomings of the dataset that have been noted and described by others (Huo et al., 2017; Liu et al., 2018). For example, clinical annotation of TCGA datasets with patient survival and treatment outcomes is incomplete—follow-up times are short (TCGA only stayed in touch with clinicians regarding their patients' clinical outcomes for a short period of time) and data is unclear at times about what the cause of death actually was (may not have been cancer; Huo et al., 2017). Moreover, breast cancer is a less aggressive cancer type, and can take 10 years or more to recur (Liu et al., 2018). So given the relatively short window of follow-up time during which TCGA outcomes were measured (reported by clinicians following up on their patients), overall survival is not a suitable clinical outcome to use (Liu et al., 2018). Overall survival is also complicated by other causes of death besides breast cancer. Disease-free survival/recurrence might have been a better endpoint to use (Liu et al., 2018). While these factors may compromise the accuracy of correlations to survival and staging, they do not affect the primary goal of using these data as a tool for learning in the classroom.

Table 2 provides student feedback that captures their attitudes and perceptions about the TCGA modules described in this paper. While students reported being initially taken aback by the size of the dataset, they reported that completing the worksheet and learning new Excel commands like the COUNTIF function helped them navigate the data effectively. Some students pointed out wanting to learn how to download data directly from the TCGA database. Others reported that working with real patient data made an impression on them.

All in all, we find this is an effective way for students to experience clinically relevant inquiry in the classroom. This bioinformatics activity can also be expanded by having the students selecting the cancer of interest and pulling relevant data from TCGA.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

AUTHOR CONTRIBUTIONS

WH and VS designed the activities outlined in this manuscript. WH, NZ, MC, WD, AK, AR, MR, and VS implemented the activities and wrote and revised the manuscript. All authors accept and are in agreement with the content of the manuscript.

ACKNOWLEDGMENTS

We thank the High Point University, its Wanek School of Natural Sciences, and its Department of Biology for the resources to carry out this work.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.573992/full#supplementary-material

Supplementary Appendix 1 | TCGA breast cancer data subset sheet: Microsoft Excel file.

Supplementary Appendix 2 | Instructional student data analysis worksheet: Microsoft Excel file.

Supplementary Appendix 3 | Tips for sorting tumors in Microsoft Excel.

Supplementary Appendix 4 | Microsoft Excel file to support statistical analysis.

REFERENCES

Ballen, C. J., Blum, J. E., Brownell, S., Hebert, S., Hewlett, J., Klein, J. R. (2017). A Call to Develop Course-Based Undergraduate Research Experiences (CUREs) for Nonmajors Courses. *CBE Life Sci. Educat.* 16:mr2. doi: 10.1187/cbe.16-12-0352

Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature* 490, 61-70. doi: 10.1038/nature11412

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., et al. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2, 401-4. doi: 10.1158/2159-8290.CD-12-0095

Ciriello, G., Gatza, M. L., Beck, A. H., Wilkerson, M. D., Rhie, S. K., Pastore, A., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* 163, 506-19. doi: 10.1016/j.cell.2015.09.033

Coughlan, T. (2020). The use of open data as a material for learning. *Educ. Technol. Res. Dev.* 68, 383–411. doi: 10.1007/s11423-019-09706-y

Dall, T., West, T., Chakrabarti, R., Reynolds, R., and Iacobucci, W. (2019). *The complexities of physician supply and demand: projections from 2015 to 2032.* Washington, DC: I. H. S. Markit Ltd.

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., et al. (2016). Improving underrepresented minority student persistence in STEM. *CBE Life Sci. Educ.* 15:es5. doi: 10.1187/cbe.16-01-0038

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal*. 6:l1. doi: 10.1126/scisignal.2004088

Huo, D., Hu, H., Rhie, S. K., Gamazon, E. R., Cherniack, A. D., Liu, J., et al. (2017). Comparison of Breast Cancer Molecular Features and Survival by African and European Ancestry in The Cancer Genome Atlas. *JAMA Oncol.* 3, 1654–1662. doi: 10.1001/jamaoncol.2017.0595

Incerti, D., Thom, H., Baio, G., and Jansen, J. P. (2019). R you still using excel? The advantages of modern software tools for health technology assessment. *Value Health* 22, 575–579. doi: 10.1016/j.jval.2019.01.003

Lin, K. Y., Parnami, S., Fuhrel-Forbis, A., Anspach, R. R., Crawford, B., and De Vries, R. G. (2013). The undergraduate premedical experience in the United States: a critical review. *Int. J. Med. Educ.* 4, 26-37.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., et al. (2018). An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell* 173, 400–416.

Rubanovich, C. K., Cheung, C., Mandel, J., and Bloss, C. S. (2018). Physician preparedness for big genomic data: a review of genomic medicine education initiatives in the United States. *Hum. Mol. Genet*. 27, R250-R258.

Schoenborn, P., Osborne, R., Toms, N., Johnstone, K., Milsom, C., Muneer, R., and Belshaw, R. (2019). OncoSim and OncoWiki: an authentic learning approach to teaching cancer genomics. *BMC Med. Educ.* 19:407.

Segarra, V. A., Hughes, N. M., Ackerman, K. M., and Grider, MH, Lyda, T, Vigueira, PA. (2018). Student performance on the Test of Scientific Literacy Skills (TOSLS) does not change with assignment of a low-stakes grade. *BMC Res. Notes* 11:422. doi: 10.1186/s13104-018-3545-9

Ziemann, M., Eren, Y., and El-Osta, A. (2016). Gene name errors are widespread in the scientific literature. *Genome Biol.* 17:177. doi: 10.1186/s13059-016-1044-7

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers
in Microbiology

PERSPECTIVE
published: 15 January 2021
doi: 10.3389/fmicb.2020.578747

Gene Annotation in High Schools: Successful Student Pipeline and Teacher Professional Development in Bioscience Using GENI-ACT

Stephen T. Koury[1]*, *Shannon Carlin-Menter*[2], *Rama Dey-Rao*[1] *and Kimberle Kelly*[3]

[1] Department of Biotechnical and Clinical Laboratory Sciences, State University of New York at Buffalo, Buffalo, NY, United States, [2] Department of Family Medicine, State University of New York at Buffalo, Buffalo, NY, United States, [3] Oak Ridge Associated Universities, Oak Ridge, TN, United States

Knowledge of genomics is an essential component of science for high school student health literacy. However, few high school teachers have received genomics training or any guidance on how to teach the subject to their students. This project explored the impact of a genomics and bioinformatics research pipeline for high school teachers and students using an introduction to genome annotation research as the catalyst. The Western New York-based project had three major components: (1) a summer teacher professional development workshop to introduce genome annotation research, (2) teacher-guided student genome annotation group projects during the school year, (3) with an end of the academic year capstone symposium to showcase student work in a poster session. Both teachers and students performed manual gene annotations using an online annotation toolkit known as Genomics Education National Initiative-Annotation Collaboration Toolkit (GENI-ACT), originally developed for use in a college undergraduate teaching environment. During the school year, students were asked to evaluate the data they had collected, formulate a hypothesis about the correctness of the computer pipeline annotation, and present the data to support their conclusions in poster form at the symposium. Evaluation of the project documented increased content knowledge in basic genomics and bioinformatics as well as increased confidence in using tools and the scientific process using GENI-ACT, thus demonstrating that high school students are capable of using the same tools as scientists to conduct a real-world research task.

Keywords: professional development, STEM education and careers, curriculum development (education), high school (9–12), bioinformatics, gene annotation

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Rachael P. Huntley,
SciBite, United Kingdom
Trinity L. Hamilton,
University of Minnesota Twin Cities,
United States

****Correspondence:***
Stephen T. Koury
stvkoury@buffalo.edu

Specialty section:
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

Received: *01 July 2020*
Accepted: *30 November 2020*
Published: *15 January 2021*

Citation:
Koury ST, Carlin-Menter S,
Dey-Rao R and Kelly K (2021) Gene
Annotation in High Schools:
Successful Student Pipeline
and Teacher Professional
Development in Bioscience Using
GENI-ACT.
Front. Microbiol. 11:578747.
doi: 10.3389/fmicb.2020.578747

INTRODUCTION

With the continuing expansion of genomic databases, discovery of rare disease-causing genetic variations and reports of drug efficacy-genotype associations, genomics has ever-increasing relevance to everyday life. It is important that the education of everyone, from doctors to patients, include genomics and bioinformatics for the continued successful integration of genomics into healthcare (Green et al., 2011). At the same time, career opportunities for students trained in genomics are growing and the recruitment and retention of talent in genomics is important for

United States economic growth (Grand View Research, 2019). This growth is due to technical advances, with DNA sequence data being generated at a much faster rate, which has created a gap between the actual generation of data and its analysis (Li et al., 2016).

While a thorough knowledge of genomics is an essential component of science and health literacy required for students to become informed citizens, consumer and professionals, educational resources and curricula fail to address this need, as few high school teachers have received genomics training or any guidance on how to teach the subject to their students (Wray, 2017). Even fewer resources are available to high school teachers to address the newer, nuanced understanding of genome structure and function and emerging genomic technologies, such as genome sequencing (National Human Genome Research Institute, 2018). The Next Generation Science Standards (NGSS) promote a three-dimensional learning approach focused on core ideas intertwined with science and engineering practices and cross-cutting concepts such as "structure and function" (Next Generation Science Standards, 2019) and the AP Biology curriculum has been redesigned to incorporate inquiry-driven scientific practices in the core (Anon, 2019). These changes in standards provide an opportunity to embed more genomics into the high school classroom, involving students in applications of genomics in real-world problem-solving settings. Incorporating inquiry-based genetic sequencing science projects into the high school curriculum is a way to narrow this knowledge gap and to educate, inspire and encourage the development of technical research skills that are needed within healthcare and personalized genomics (Ditty et al., 2010; Moitra, 2017).

Project Background

Beginning in 2013 and funded by a 3-years NSF Innovative Technology Experiences for Students and Teachers (ITEST) Grant, we developed the Western New York Genetics in Research Partnership (WNYGRP). The partnership was comprised of the University at Buffalo, including the departments of Biotechnical and Clinical Laboratory Sciences and Family Medicine; the NYS Center of Excellence in Bioinformatics and Life Sciences (CBLS); the New York State Area Health Education Center System (NYSAHEC), including Erie-Niagara (EN AHEC) and Western New York Rural (R-AHEC); Oak Ridge Associated Universities (ORAU); UB faculty with expertise in genome annotation; and a NYS STEM Master High School Teacher. The project introduced high school teachers and students to genomics and bioinformatics through the use of freely available, hands-on, state-of-the-art bioinformatics tools.

This ITEST research project developed partnerships with disadvantaged high schools across a 14-county region in Western New York, forming a pipeline for teacher and student recruitment. The details of the development of the partnership will be presented elsewhere. Grades 9–12 biology teachers were trained on the use of the Genomics Education National Initiative-Annotation Collaboration Toolkit (GENI-ACT)[1]. This innovative technology experience increased high school students' and teachers' knowledge of bioinformatics and allowed teachers to gain experience with bioinformatics software tools for classroom use through real-world research experiences.

PROGRAM COMPONENTS

The ITEST project had three major components outlined below, consisting of a summer teacher professional development (PD) workshop, teacher-guided student genome annotation projects during the school year, and a capstone symposium at the end of the school year. High school Biology teachers recruited from the targeted schools signed-up for the summer workshop for a variety of reasons, including learning something new, using the training hours to count toward their mandatory staff development, the stipend they received for their involvement, and/or the ability to offer their students something new to add to their portfolios or highlight during college interviews. One teacher commented, "The idea of exposing students to real science was very enticing to me and I feel like the idea of being a scientist and being able to handle Big Data is a skill that we need to start teaching our students." Overall, we recruited 74 Biology teachers over the 3 years to take part in the summer professional development training.

Summer PD Workshop

During the 5-day Summer Workshop, teachers were trained using nine modules customized by project faculty that were based on those in GENI-ACT (9, **Table 1**). After the training, the teachers worked with their students on the same modules during the school year. GENI-ACT and the online bioinformatics tools utilized during the training were free, so only computer and internet access were needed to take part in the project. First, we presented teachers with background knowledge that provided them with an understanding of genomics, DNA structure, and transcription/translation relevant to gene annotation. Teachers were then instructed on how to log into GENI-ACT and navigate the website.

Faculty instructors assigned the teachers a set of demonstration genes to annotate that illustrated positive and negative results obtained from the tools in the modules. Teachers were shown how to use each tool and interpret results using such parameters as scores and *e*-values and then allowed to apply it on their own during the week of training. The relative strengths and drawbacks of results obtained from different databases were stressed to inform the development of hypotheses about genes under investigation.

A manual with background information and complete step-by-step instructions for completing all modules was developed during the project is freely available on our website (NSF, 2020). The gene annotation work was interspersed with talks from project faculty on personalized genomics and program evaluation. Teachers completed pre and post-workshop surveys to evaluate gains in content knowledge about bioinformatics related to genome annotation and their comfort level with teaching bioinformatics concepts.

[1] https://geni-act.org

TABLE 1 | The modules used in GENI-ACT.

Modules	Activities	Questions investigated
Basic information	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of the gene and protein? Where is it located in the genome?
Sequence-based similarity	Blast (Altschul et al., 1990), COG (Tatusov et al., 1997), T-Coffee (Di Tommaso et al., 2011), WebLogo (Crooks et al., 2004)	How similar is the sequence of the protein under investigation to other proteins in GenBank?
Structure-based similarity	TIGRFAM (Haft et al., 2001), Pfam (El-Gebali et al., 2019), PDB (Berman et al., 2000)	What functional domains are present in the protein under investigation?
Cellular localization	Gram Stain, TMHMM (Krogh et al., 2001), SignalP (Almagro Armenteros et al., 2019), PSORTb (Yu et al., 2010), Phobius (Käll et al., 2007)	Is the protein under investigation located in the cytoplasm, secreted, in the periplasm, or embedded in the cell membrane or cell wall?
Enzymatic function	KEGG (Kanehisa and Goto, 2000), MetaCyc (Karp et al., 2002), E.C. Number (Expasy, 2020)	In what process or structure is the protein under investigation involved?
Duplication and degradation	Paralog, Pseudogene	Are there other forms of the protein under investigation in the same genome? Is it functional?
Horizontal gene transfer	Phylogenetic Tree, Gene Neighborhood, GC Content	Has the protein under investigation co-evolved with the rest of the genome or has it been obtained in a different way?
RNA family	Rfam (Kalvari et al., 2018)	Does the gene under investigation encode a functional RNA?
Final annotation	Evaluate data from all modules	Has the gene been correctly called by the pipeline annotation?

GENI-ACT was undergoing a transition at the time the project was initiated, resulting in creation of customized notebooks and instructions for this project (NSF, 2020).

Academic Year Annotation Projects

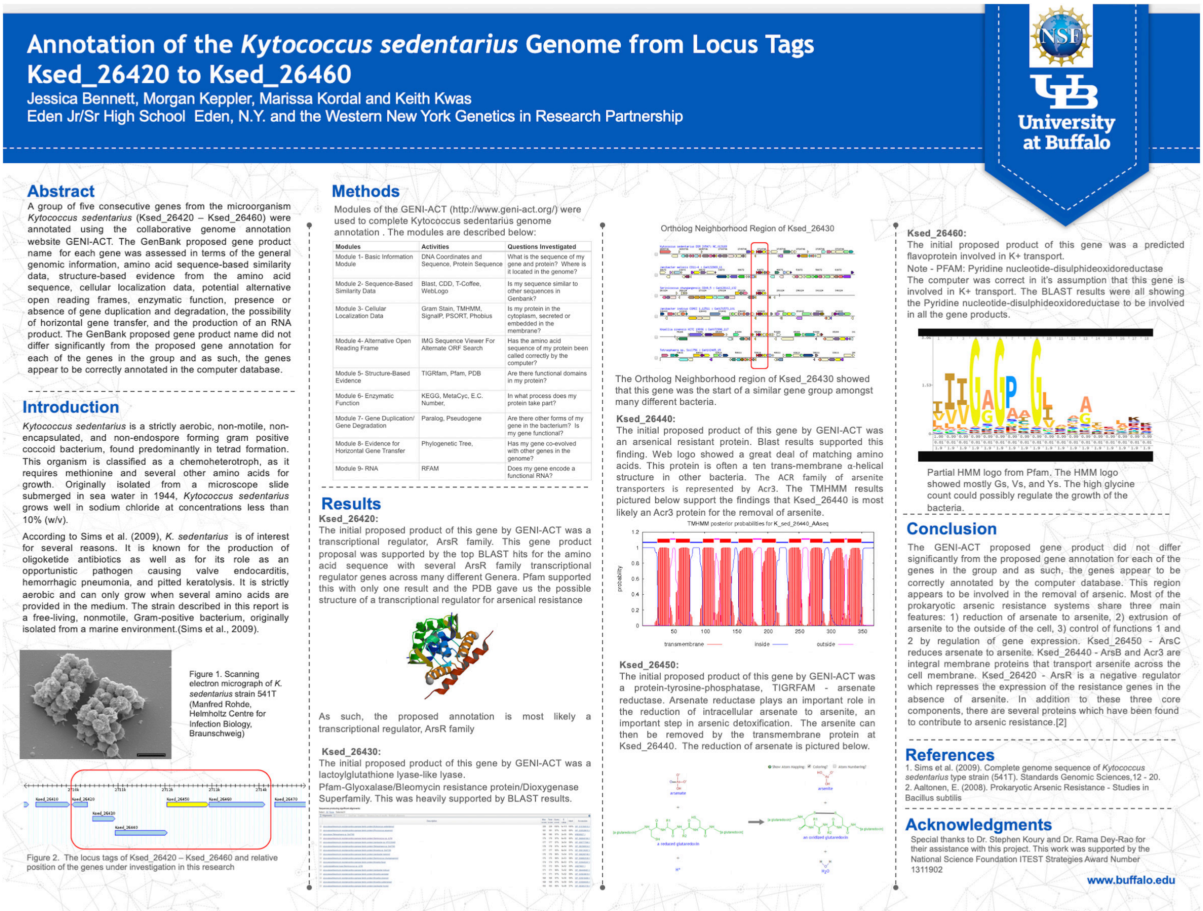
As the teachers returned to school in September, they recruited student participants and trained them using the nine GENI-ACT modules. All interested students were offered career counseling and exposure to genomics activities to encourage the recruitment of student participants. Activity 1, College and Career Exploration, was facilitated by AHEC coordinators from the school's local center, R-AHEC or EN-AHEC, and provided students with STEM college and career guidance. Activity 2, also facilitated by AHEC, explored bioinformatics and genomic careers in more detail. Activity 3, facilitated by University of Buffalo faculty, provided students with an introduction to genome annotation. A total of 1,948 high school students attended at least one of the three activities over the 3 years of the program.

To evaluate the effectiveness of the program, informed consent was obtained from all participating students, and pre and post surveys assessed gains of student knowledge and changes related to their attitudes about careers in STEM. An experimental design was used, which randomized the 667 students recruited by the teachers into two groups: 343 were randomized into the intervention group (received GENI-ACT training) and the other 324 into the comparison group (no GENI-ACT training). Comparison group activities included various topics, which included researching bioethics or doing background research on genes identified by the annotators and/or the organism under study. Each student group in the intervention (GENI-ACT trained) was assigned a unique gene from the bacterium *Kytococcus sedentarius.* The students worked on this gene in the modules, along with a demonstration gene that teachers could use in a "show one, do one" model of teaching. Most teachers worked with their students through an after school club, as teachers were compensated for their time outside the classroom. Since a randomized design was utilized, the control and intervention students' work were separated and easier to control outside of the regular classroom in an after school program. On average, teachers met with their intervention students once a week from January through April of the school year. Each teacher worked with a group averaging about seven students, assisting their work on the modules and recording data in their online notebooks. The students enjoyed the GENI-ACT modules. As one student explained "the modules themselves along with the paper manual really made the program easy to follow, which was great for first time students." Students also appreciated that each of the genes they were assigned were different and that the modules allowed them find something unique about their particular gene. One student commented that the aspect of the uniquely assigned genes helped to fuel their love of research.

Refresher trainings were offered to teachers on three different Saturdays during the school year. The third refresher training, offered in April, dealt with preparing the teachers for their students' research poster preparation and presentation at the project culminating Capstone Symposium held in May. Using a poster template that could be populated with data generated by their students, teachers submitted the completed posters to program faculty approximately 1 week before the capstone, and faculty edited them for formatting only (**Figure 1**). The content was left as submitted (unless a glaring error was noted) to ensure that the posters represented student work and data interpretation. All posters were printed with dimensions of 4 × 3 feet and displayed at the capstone symposium.

Capstone Symposium

In all, four student capstones were hosted. A total of 136 posters were prepared and presented during capstone symposia from 2014 to 2017 and are viewable on our online website (NSF, 2020). Annual Capstone Symposia took place at the end of each project year at the University at Buffalo, and, on two different occasions in academic institutions outside of the immediate Buffalo area, with participant numbers increasing each successive year. The capstone provided each student participant with the experience of attending a scientific meeting to present their data and to network



Annotation of the *Kytococcus sedentarius* Genome from Locus Tags Ksed_26420 to Ksed_26460

Jessica Bennett, Morgan Keppler, Marissa Kordal and Keith Kwas
Eden Jr/Sr High School Eden, N.Y. and the Western New York Genetics in Research Partnership

NSF
University at Buffalo

Abstract

A group of five consecutive genes from the microorganism *Kytococcus sedentarius* (Ksed_26420 – Ksed_26460) were annotated using the collaborative genome annotation website GENI-ACT. The GenBank proposed gene product name for each gene was assessed in terms of the general genomic information, amino acid sequence-based similarity data, structure-based evidence from the amino acid sequence, cellular localization data, potential alternative open reading frames, enzymatic function, presence or absence of gene duplication and degradation, the possibility of horizontal gene transfer, and the production of an RNA product. The GenBank proposed gene product name did not differ significantly from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated in the computer database.

Introduction

Kytococcus sedentarius is a strictly aerobic, non-motile, non-encapsulated, and non-endospore forming gram positive coccoid bacterium, found predominantly in tetrad formation. This organism is classified as a chemoheterotroph, as it requires methionine and several other amino acids for growth. Originally isolated from a microscope slide submerged in sea water in 1944, *Kytococcus sedentarius* grows well in sodium chloride at concentrations less than 10% (w/v).

According to Sims et al. (2009), *K. sedentarius* is of interest for several reasons. It is known for the production of oligoketide antibiotics as well as for its role as an opportunistic pathogen causing valve endocarditis, hemorrhagic pneumonia, and pitted keratolysis. It is strictly aerobic and can only grow when several amino acids are provided in the medium. The strain described in this report is a free-living, nonmotile, Gram-positive bacterium, originally isolated from a marine environment.(Sims et al., 2009).

Figure 1. Scanning electron micrograph of *K. sedentarius* strain 541T (Manfred Rohde, Helmholtz Centre for Infection Biology, Braunschweig)

Figure 2. The locus tags of Ksed_26420 – Ksed_26460 and relative position of the genes under investigation in this research

Methods

Modules of the GENI-ACT (http://www.geni-act.org/) were used to complete Kytococcus sedentarius genome annotation . The modules are described below:

Modules	Activities	Questions Investigated
Module 1- Basic Information Module	DNA Coordinates and Sequence, Protein Sequence	What is the sequence of my gene and protein? Where is it located in the genome?
Module 2- Sequence-Based Similarity Data	Blast, CDD, T-Coffee, WebLogo	Is my sequence similar to other sequences in Genbank?
Module 3- Cellular Localization Data	Gram Stain, TMHMM, SignalP, PSORT, Phobius	Is my protein in the cytoplasm, secreted or embedded in the membrane?
Module 4- Alternative Open Reading Frame	IMG Sequence Viewer For Alternate ORF Search	Has the amino acid sequence of my protein been called correctly by the computer?
Module 5- Structure-Based Evidence	TIGRfam, Pfam, PDB	Are there functional domains in my protein?
Module 6- Enzymatic Function	KEGG, MetaCyc, E.C. Number,	In what process does my protein take part?
Module 7- Gene Duplication/ Gene Degradation	Paralog, Pseudogene	Are there other forms of my gene in the bacterium? Is my gene functional?
Module 8- Evidence for Horizontal Gene Transfer	Phylogenetic Tree,	Has my gene co-evolved with other genes in the genome?
Module 9- RNA	RFAM	Does my gene encode a functional RNA?

Results

Ksed_26420:
The initial proposed product of this gene by GENI-ACT was a transcriptional regulator, ArsR family. This gene product proposal was supported by the top BLAST hits for the amino acid sequence with several ArsR family transcriptional regulator genes across many different Genera. Pfam supported this with only one result and the PDB gave us the possible structure of a transcriptional regulator for arsenical resistance

As such, the proposed annotation is most likely a transcriptional regulator, ArsR family

Ksed_26430:
The initial proposed product of this gene by GENI-ACT was a lactoylglutathione lyase-like lyase.
Pfam-Glyoxalase/Bleomycin resistance protein/Dioxygenase Superfamily. This was heavily supported by BLAST results.

Ortholog Neighborhood Region of Ksed_26430

The Ortholog Neighborhood region of Ksed_26430 showed that this gene was the start of a similar gene group amongst many different bacteria.

Ksed_26440:
The initial proposed product of this gene by GENI-ACT was an arsenical resistant protein. Blast results supported this finding. Web logo showed a great deal of matching amino acids. This protein is often a ten trans-membrane α-helical structure in other bacteria. The ACR family of arsenite transporters is represented by Acr3. The TMHMM results pictured below support the findings that Ksed_26440 is most likely an Acr3 protein for the removal of arsenite.

Ksed_26450:
The initial proposed product of this gene by GENI-ACT was a protein-tyrosine-phosphatase, TIGRFAM - arsenate reductase. Arsenate reductase plays an important role in the reduction of intracellular arsenate to arsenite, an important step in arsenic detoxification. The arsenite can then be removed by the transmembrane protein at Ksed_26440. The reduction of arsenate is pictured below.

Ksed_26460:
The initial proposed product of this gene was a predicted flavoprotein involved in K+ transport.
Note - PFAM: Pyridine nucleotide-disulphideoxidoreductase
The computer was correct in it's assumption that this gene is involved in K+ transport. The BLAST results were all showing the Pyridine nucleotide-disulphideoxidoreductase to be involved in all the gene products.

Partial HMM logo from Pfam. The HMM logo showed mostly Gs, Vs, and Ys. The high glycine count could possibly regulate the growth of the bacteria.

Conclusion

The GENI-ACT proposed gene product did not differ significantly from the proposed gene annotation for each of the genes in the group and as such, the genes appear to be correctly annotated by the computer database. This region appears to be involved in the removal of arsenic. Most of the prokaryotic arsenic resistance systems share three main features: 1) reduction of arsenate to arsenite, 2) extrusion of arsenite to the outside of the cell, 3) control of functions 1 and 2 by regulation of gene expression. Ksed_26450 - ArsC reduces arsenate to arsenite. Ksed_26440 - ArsB and Acr3 are integral membrane proteins that transport arsenite across the cell membrane. Ksed_26420 - ArsR is a negative regulator which represses the expression of the resistance genes in the absence of arsenite. In addition to these three core components, there are several proteins which have been found to contribute to arsenic resistance.[2]

References

1. Sims et al. (2009). Complete genome sequence of *Kytococcus sedentarius* type strain (541T). Standards Genomic Sciences,12 - 20.
2. Aaltonen, E. (2008). Prokaryotic Arsenic Resistance - Studies in Bacillus subtilis

Acknowledgments

Special thanks to Dr. Stephen Koury and Dr. Rama Dey-Rao for their assistance with this project. This work was supported by the National Science Foundation ITEST Strategies Award Number 1311902

www.buffalo.edu

FIGURE 1 | Example of a capstone event poster. Two to three students typically pooled data to prepared a poster in most instances and took turns presenting the poster at the capstone. High resolution versions of all posters presented at capstone events are available on the Student Research page of our project website (NSF, 2020).

with other teacher/student participants and program faculty. The capstone poster session was broken into two sections, allowing students to visit and interact with students from other schools.

A luncheon also allowed for informal interaction among students, followed by a series of speakers highlighting current topics in bioinformatics and genomics. The capstones concluded with a ceremony recognizing each student and teacher participant with a certificate of participation. Teachers were encouraged to take their posters back to their school and display them in the hallway or classroom. One teacher commented that their students "are very proud of those posters hanging up there in the hallway." Another teacher noted that the capstone is "a nice program for the high school students to see what's going on at the college level and the poster event is something unique, and something we don't usually do at the high school level."

Program Outcomes

Teacher Content Knowledge was measured before and after the workshop. Teachers were asked to complete two sets of 10 True/False questions to assess their knowledge of bioinformatics and genome annotations at the start and end of the summer training workshop. The ten questions included in Set 1 were developed by the Microbial Genome Annotation Network (MGAN) to assess learning in students who used GENI-ACT within their courses. Set 2 includes 10 supplemental items developed by Faculty to help assess learning specific to the program. Mixed ANOVAs produced a significant increase in content knowledge scores from the pre workshop survey to the post workshop survey [$F(1,31) = 37.86$, $p < 0.001$, $\eta^2 p = 0.55$], confirming that teachers increased their content knowledge of bioinformatics and gene annotation by the end of the workshop, as predicted. The content knowledge questions, scoring, and example teacher responses are available in the educational resources section of our project website (NSF, 2020a).

Teaching Behaviors around bioinformatics and gene annotation were also expected to increase as a result of training. As a way of gauging their comfort with teaching the material, teachers were asked to rate their confidence in teaching GENI-ACT content topics. Specifically, teachers rated 28 topics on a percentage scale, from 10 to 100% in 10-percentage point increments. Their pre and post workshop ratings were compared using paired t-tests. In the case of every single topic, there was a significant increase by the end of the workshop. The mean increase in confidence from pre workshop to post workshop across all 28 content topics was 56%. The workshops clearly prepared teachers to use the GENI-ACT content and software tools with their students. However, not all teachers went on to work with students during the following academic year, with reasons including perceived difficulty of the project activities, difficulty implementing the study using the control group model or that they personally did not want to participate in the project.

Student content knowledge was projected to increase by the end of program in the intervention group, or those students receiving training on the GENI-ACT modules. Students completed the same content knowledge assessment as the teachers, measured twice as part of pre and post student surveys. Students were asked to complete two sets of 10 True/False questions to assess their knowledge of bioinformatics and genome annotations. In independent t-tests, Intervention students significantly increased their content knowledge of bioinformatics and gene annotation by the end of the project, while comparison students did not, on both Set 1, $t(173) = 3.19$, $p = 0.002$ and Set 2, $t(173) = 8.40$, $p < 0.001$. Moreover, the scores in the Treatment group increased by well over 50%, especially in Knowledge Set 2.

Participant Perspectives

Impact of the project could be seen in student participants when it came to college applications, choosing a major and college interviews. One student said that "After participating in the ITEST program I knew that I wanted to become a chemical engineer. Furthermore, I knew that I wanted to attend the University at Buffalo because of how research-oriented the university is. Lastly, I knew that I wanted to attempt to pursue applications of chemical engineering in medicine and specifically the genomic medicine field. Over the next 4 years and beyond, I plan to pursue a career in this field." Another student, who was accepted into RIT after participating in this program, was able to petition to be allowed into a Bioinformatics course that was only available for seniors as an elective. He was able to take the course as a Sophomore because he was able to prove through his Capstone poster that he had all the background knowledge to take the course.

Other teacher and student perspectives on performing gene annotations as a part of this project are available in an NSF STEM For All Video Showcase presentation (Videohall, 2016).

DISCUSSION

The results of this project informed different approaches to gene annotation with high school students and teachers that were utilized in another recently completed NIH Science Education Partnership Award (manuscript in preparation). The valuable partnership relationships developed have continued to expand since completion of the ITEST project described here and continue for the foreseeable future through another recently funded project. This project demonstrated that grade 9–12 students could grasp gene annotation and bioinformatics tools and use them appropriately.

The major limitation of this project for teachers was the use of the control group design. With this design, teachers could not include the gene annotation activities within their regular classes due to the need of having some students in a control group. This restricted most teachers to working with students before or after regular school hours, resulting in competition with other after-school student activities (sports/clubs). Another limitation of the control group design was the amount of time needed to recruit and randomize students before they could begin working with students on their annotations. As such, most teachers could not to begin work with their students until well after winter break and were only able to work through the first four modules before the end of the school year.

Sustained use of the bioinformatics tools by teacher participants after project completion is being explored and will be reported in more detail elsewhere. While complete gene annotation is not a common theme, teachers have been able to pick and choose tools from modules to integrate into their curriculum with relative ease. Some teachers have continued to pursue complete gene annotations and have their students present at the annual capstone event tied to another project, as they feel the poster presentation is a great experience for their students. One past participating teacher has integrated all nine GENI-ACT modules into his Honors Biology class by putting together PowerPoint presentations based on the Modules and meeting with the students every day in a lab situation. Future research might aim to determine the effect of taking part in gene annotation on academic performance related to biology and genetics. A study performed at the community college level demonstrated that students taking part in gene annotation in a cell biology lab exhibited clear gains in understanding of topics related to molecular biology in a lecture course (Beagley, 2013), suggesting similar gains could be expected in the high school classroom as well. Additional research is needed to identify topics most appropriate for, and learned most optimally by, high school students. For example, which aspects of bioinformatics-based research would most easily be integrated into high school biology curricula guided by NGSS? NGSS-friendly curricula will make it easier for teachers to introduce more students to bioinformatics. While bioinformatics software tools are complex and their use is challenging to teach, this study shows they can be successfully used by high school teachers with their students. Furthermore, utilizing the same bioinformatics tools used by scientists to conduct authentic research promotes student interest in science by seeing that they too can apply the scientific method to study real-world problems.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the University at Buffalo (IRB). Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

AUTHOR CONTRIBUTIONS

SK: training teachers, working with high school students, editing student posters, and writing the manuscript. RD-R: training teachers, working with high school students, and editing student posters. SC-M: program evaluation, writing the manuscript, and supervision of program manager. KK: program evaluation and writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work supported by the National Science Foundation – United States: Innovative Technology Experiences for Students and Teachers. Award Number: 1311902 NSF funded all aspects of this award. This work was also supported by the University at Buffalo, Department of Biotechnical and Clinical Laboratory Sciences and Family Medicine – will pay page charges associated with publication.

REFERENCES

Almagro Armenteros, J. J., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., et al. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* 37, 420–423. doi: 10.1038/s41587-019-0036-z

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.

Anon (2019). *AP Biology Investigative Labs: An Inquiry-Based Approach. 2012.* Available from: https://secure-media.collegeboard.org/digitalServices/pdf/ap/APBioTeacherLabManual2012_2ndPrt_lkd.pdf (accessed June 29, 2019).

Beagley, C. T. (2013). Genome annotation in a community college cell biology lab. *Biochem. Mol. Biol. Educ.* 41, 44–49. doi: 10.1002/bmb.20669

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. *Genome Res.* 14, 1188–1190. doi: 10.1101/gr.849004

Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J. M., et al. (2011). T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. *Nucleic Acids Res.* 39, W13–W17. doi: 10.1093/nar/gkr245

Ditty, J. L., Kvaal, C. A., Goodner, B., Freyermuth, S. K., Bailey, C., Britton, R. A., et al. (2010). Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biol.* 8:e1000448. doi: 10.1371/journal.pbio.1000448

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S. R., Luciani, A., Potter, S. C., et al. (2019). The Pfam protein families database in 2019. *Nucleic Acids Res.* 47, D427–D432. doi: 10.1093/nar/gky995

Expasy (2020). *Expasy Enzyme.* Available online at: https://enzyme.expasy.org (accessed June 30, 2020).

Grand View Research (2019). *Genomics Market Size, Share & Trends Analysis Report By Application And Technology (Pathway Analysis, Functional Genomics), By Deliverables (Instruments, Consumables, Services), By End Use, And Segment Forecasts, 2019 - 2025. 2019.* Available online at: https://www.grandviewresearch.com/industry-analysis/genomics-market (accessed June 25, 2020).

Green, E. D., Guyer, M. S., and National Human Genome Research Institute (2011). Charting a course for genomic medicine from base pairs to bedside. *Nature* 470:204. doi: 10.1038/nature09764

Haft, D. H., Loftus, B. J., Richardson, D. L., Yang, F., Eisen, J. A., Paulsen, I. T., et al. (2001). TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.* 29, 41–43. doi: 10.1093/nar/29.1.41

Käll, L., Krogh, A., and Sonnhammer, E. L. (2007). Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.* 35, W429–W432. doi: 10.1093/nar/gkm256

Kalvari, J., Argasinska, J., Quinones-Olvera, N., Nawrocki, E. P., Rivas, E., Eddy, S. R., et al. (2018). Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res.* 37, 420–423. doi: 10.1093/nar/gkx1038

Kanehisa, M., and Goto, S. (2000). KEGG. Kyoto rncyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

Karp, P. D., Riley, M., Paley, S. M., and Pellegrini-Toole, A. (2002). The MetaCyc database. *Nucleic Acids Res.* 30, 59–61. doi: 10.1093/nar/30.1.59

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305, 567–580. doi: 10.1006/jmbi.2000.4315

Li, J., Batcha, A. M., Grüning, B., and Mansmann, U. R. (2016). An NGS workflow blueprint for DNA sequencing data and its application in individualized molecular oncology. *Cancer Inform.* 14(Suppl. 5), 87–107.

Moitra, K. (2017). Releasing the "GENI": integrating authentic microbial genomics research into the classroom through GENI-ACT. *FEMS Microbiol. Lett.* 364:fnx215.

National Human Genome Research Institute (2018). *Genomic Literacy, Education, And Engagement (Glee) Initiative 2017 Strategic Visioning Meeting: K-16 Working Group. 2017 2018.* Available online at: https://www.genome.gov/Pages/About/OD/ECIB/GLEE/GLEE_white_paper_K-16_WG.pdf (accessed June 25, 2018).

Next Generation Science Standards (2019). Available from: https://www.nextgenscience.org (accessed June 29, 2019).

NSF (2020). *Western New York Genetics in Research Partnership Educational Resources.* Available from: http://ubwp.buffalo.edu/wnygirp/educational-resources/ (accessed June 26, 2020).

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. *Science* 278, 631–637. doi: 10.1126/science.278.5338.631

Videohall (2016). *Western New York Genetice in Research Partnership NSF STEM For All Video Showcase.* Avaialble at: https://stemforall2016.videohall.com/presentations/709 (accessed June 26, 2020).

Wray, C. G. (2017). Introducing students to the genome: brave new world or the red Queen's wonderland? *Am. Biol. Teach.* 79, 253–253. doi: 10.1525/abt.2017.79.4.253

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., et al. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* 26, 1608–1615. doi: 10.1093/bioinformatics/btq249

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ORIGINAL RESEARCH
published: 18 February 2021
doi: 10.3389/fmicb.2021.579325

Students in a Course-Based Undergraduate Research Experience Course Discovered Dramatic Changes in the Bacterial Community Composition Between Summer and Winter Lake Samples

Stokes S. Baker[1]*, Mohamed S. Alhassan[1], Kristian Z. Asenov[1], Joyce J. Choi[1,2], Griffin E. Craig[1], Zayn A. Dastidar[1,3], Saleh J. Karim[1], Erin E. Sheardy[1], Salameh Z. Sloulin[1], Nitish Aggarwal[1], Zahraa M. Al-Habib[1], Valentina Camaj[1], Dennis D. Cleminte[1], Mira H. Hamady[1], Mike Jaafar[1], Marcel L. Jones[1], Zayan M. Khan[1], Evileen S. Khoshaba[1], Rita Khoshaba[1], Sarah S. Ko[1], Abdulmalik T. Mashrah[1], Pujan A. Patel[1], Rabeeh Rajab[1] and Sahil Tandon[1]

[1] Biology Department, University of Detroit Mercy, Detroit, MI, United States, [2] School of Environment and Sustainability, University of Michigan, Ann Arbor, MI, United States, [3] Mike Ilitch School of Business, Wayne State University, Detroit, MI, United States

OPEN ACCESS

Edited by:
Mel Crystal Melendrez,
Anoka-Ramsey Community College,
United States

Reviewed by:
Jitendra Keshri,
Texas A&M University, United States
Eric Daniel Becraft,
University of North Alabama,
United States
Madhav P. Nepal,
South Dakota State University,
United States

***Correspondence:**
Stokes S. Baker
bakerss@udmercy.edu

Specialty section:
This article was submitted to
Evolutionary and Genomic
Microbiology,
a section of the journal
Frontiers in Microbiology

Received: 02 July 2020
Accepted: 20 January 2021
Published: 18 February 2021

Citation:
Baker SS, Alhassan MS,
Asenov KZ, Choi JJ, Craig GE,
Dastidar ZA, Karim SJ, Sheardy EE,
Sloulin SZ, Aggarwal N, Al-Habib ZM,
Camaj V, Cleminte DD, Hamady MH,
Jaafar M, Jones ML, Khan ZM,
Khoshaba ES, Khoshaba R, Ko SS,
Mashrah AT, Patel PA, Rajab R and
Tandon S (2021) Students in a
Course-Based Undergraduate
Research Experience Course
Discovered Dramatic Changes
in the Bacterial Community
Composition Between Summer
and Winter Lake Samples.
Front. Microbiol. 12:579325.
doi: 10.3389/fmicb.2021.579325

Course-based undergraduate research experience (CURE) courses incorporate high-impact pedagogies that have been shown to increase undergraduate retention among underrepresented minorities and women. As part of the Building Infrastructure Leading to Diversity program at the University of Detroit Mercy, a CURE metagenomics course was established in the winter of 2019. Students investigated the bacterial community composition in a eutrophic cove in Lake Saint Clair (Harrison Township, MI, United States) from water samples taken in the summer and winter. The students created 16S rRNA libraries that were sequenced using next-generation sequencing technology. They used a public web-based supercomputing resource to process their raw sequencing data and web-based tools to perform advanced statistical analysis. The students discovered that the most common operational taxonomic unit, representing 31% of the prokaryotic sequences in both summer and winter samples, corresponded to an organism that belongs to a previously unidentified phylum. This result showed the students the power of metagenomics because the approach was able to detect unclassified organisms. Principal Coordinates Analysis of Bray–Curtis dissimilarity index data showed that the winter community was distinct from the summer community [Analysis of Similarities (ANOSIM) $r = 0.59829$, $n = 18$, and $p < 0.001$]. Dendrograms based on hierarchically clustered Pearson correlation coefficients of phyla were divided into a winter clade and a summer clade. The conclusion is that the winter bacterial population was fundamentally different from the summer population, even though the samples were taken from the same locations in a protected cove. Because of the small class sizes, qualitative as well as statistical methods were used to evaluate the course's impact on student attitudes. Results from the Laboratory Course Assessment Survey showed that most of the respondents felt they were contributing to scientific knowledge and the course fostered student collaboration. The majority of respondents agreed or strongly agreed that the course incorporated iteration aspects of scientific investigations, such as repeating procedures to fix problems. In summary, the metagenomics CURE course was able to add to scientific knowledge and allowed students to participate in authentic research.

Keywords: 16S rRNA, aquatic, Course-based Undergraduate Research Experience, eutrophic, ice, Laboratory Course Assessment Survey, metabarcoding, metagenomics

INTRODUCTION

For over a quarter of a century, reports from science, technology, engineering and mathematics (STEM) advisory organizations have been calling for reform of undergraduate STEM curricula to focus on developing analytical skills instead of memorizing content (Project Kaleidoscope, 1991; Howard Hughes Medical Institute, 1996; National Research Council, 1996, 2003; National Science Foundation, 1996; Bauerle et al., 2009). These same reports have called for teaching innovations that will increase the participation of underrepresented minority students in STEM. Programs that have met this goal have some of the following attributes: experience with authentic research, active learning, collaborative learning communities where students share an intellectual experience, and involvement in research that directly impacts their communities (Graham et al., 2013; Toven-Lindsey et al., 2015; Estrada et al., 2016; Provost, 2016). Faculty-supervised undergraduate research is a well-established approach to provide these high-impact activities. Unfortunately, the approach has limited capacity (i.e., only a few students can be effectively taught using an apprentice model). One strategy to overcome the bottleneck is to provide course-based undergraduate research experience (CURE) instruction (Provost, 2016; Bell et al., 2017).

Course-based undergraduate research experiences are defined as laboratory courses that incorporate the following attributes (Auchincloss et al., 2014; Provost, 2016):

1. Scientific Process: Conducting research as practiced by professional scientists.
2. Discovery: Investigating novel questions.
3. Relevance: Having impacts beyond the classroom because the research advances scientific knowledge.
4. Collaboration: Collectively tackling difficult problems.
5. Iteration: Conducting research built upon existing knowledge, learning by failure and retrying, and revising thinking after self-analysis and peer-critique.

Several CURE courses have been successfully implemented that involved microbiology, virology, molecular biology, bioinformatics, and other life science disciplines (Wang, 2017), including metagenomics (CUREnet, 2013; Lentz et al., 2017; Wang, 2017). One strength of CUREs is they can support distributive approaches to address large biological questions (Hatfull, 2015; Wang, 2017). Because the microbial world is so diverse and vast, the National Research Council has called for the incorporation of metagenomics into undergraduate biology instruction because it can be an effective distributive strategy to advance scientific knowledge (Jurkowski et al., 2007). An example of a successful distributive-science CURE is the Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES) program (Hatfull, 2015).

The University of Detroit Mercy's ReBUILD Detroit program (Snyder and Kumar, 2019) is part of a National Institutes of Health initiative to increase the pipeline of underrepresented minority undergraduates entering biomedical STEM research careers (National Institutes of Health, 2019). To recruit and retain the target population, ReBUILD Detroit is using a "persistence model" (Graham et al., 2013; Toven-Lindsey et al., 2015) which involves having the students participate in research activities every semester, including the first semester of their freshman year. To increase the availability of authentic research experiences for undergraduates and to support ReBUILD Detroit's retention strategy, a CURE course entitled, "Applied Metagenomics" was established in the winter of 2019 and repeated in the winter of 2020. The course investigations focused on aquatic microbiology because water quality issues are important community concerns in metropolitan Detroit (Bolsenga and Herdendorf, 1993). Because Detroit is a large industrial city within the Great Lakes Basin, the students have a myriad of water quality issues they can investigate.

Background Related to the Environmental Question Investigated by the Students

Metagenomics, as defined by the National Research Council (2007) and Wooley et al. (2010), is the study of uncultured microorganisms found in environmental samples, by use of massively parallel sequencing. The environmental DNA (eDNA) sequences can be bulk DNA (a.k.a., shotgun metagenomics) or amplicons from specific loci (a.k.a., metabarcoding). Metagenomic studies have shown that freshwater ecosystems appear to have a distinct assemblage of prokaryotes in the epilimnia. Metanalysis studies of 16S rRNA gene sequences obtained from diverse lakes (e.g., oligotrophic to highly eutrophic) on different continents have shown that freshwater lakes have an assemblage of prokaryotes that are distinct from marine and terrestrial habitats (Zwart et al., 2002; Newton et al., 2011).

Some 16S rRNA metabarcoding studies have shown that freshwater trophic status can impact the composition of prokaryote communities. For example, a study of human-impacted tributaries of the Great Lakes showed greater species richness in oligotrophic lake samples (Newton and McLellan, 2015). A similar pattern was observed in a separate study of the Great Lakes, canals, and streams of the Niagara Peninsula (Mohiuddin et al., 2019). In contrast, a study of oligotrophic versus eutrophic lakes in Greece showed greater species diversity in the eutrophic samples (Karayanni et al., 2019). These results suggest that trophic status can alter the freshwater prokaryote diversity, but a general rule on the relationship between nutrient level and prokaryote community diversity has not been established.

Many metagenomic investigations of aquatic ecosystems only sample water during ice-free months (for examples, see Shade et al., 2007; Mohiuddin et al., 2019). As a result, less information on the nature of aquatic bacterial communities in seasonally freezing lakes is available in the literature. Vigneron et al. (2019) observed that ice-covered tundra lakes had a rich prokaryotic community with similar cell densities to the ice-free water. However, the composition of the prokaryotic community changed with the seasons. Metabolic pathways deduced from shotgun metagenomic sequencing showed the prokaryotic community shifted from phototrophic and aerobic metabolism in the summer to reductive metabolism that could degrade aromatic organics in the winter. Tran et al. (2018) observed similar results in their investigations of Verrucomicrobia communities of taiga lakes. These results suggest that winter prokaryotic communities in ice-covered lakes contain a rich biota distinct from their open water counterparts. With these observations in mind, the goal of the students in Applied Metagenomics was to determine if the prokaryotic community in an ice-covered versus open-water temperate lake exhibited changes in community composition similar to those observed in tundra and boreal lakes.

MATERIALS AND METHODS

Human Subject Statement

This study was carried out in accordance with the recommendations of National Institutes of Health's Human Subjects Research Guidelines. The protocol was approved by the University of Detroit Mercy's Institutional Review Board (Protocol Number 1718-53) on March 10, 2018. All subjects gave written informed consent in accordance with the Declaration of Helsinki.

Class Description and Assessment

Applied Metagenomics (BIO3201) was offered at the University of Detroit Mercy during the winter terms of 2019 and 2020. The prerequisite for the course was genetics, cell and molecular biology, or biochemistry. In 2019, eight students were enrolled in the 15 week course. Their self-reported demographics were as follows: Gender: 75% males, 25% female; Ethnicity/Race: 75% white, 25% Asian/Pacific Islander. In 2020, 16 students took the course. Their self-reported demographics were as follows: Gender: 50% males, 43.75% female, 6.25% prefer not to answer. Ethnicity/Race: 37.5% White (Middle Eastern descent), 25% White (European descent), 31.25% Asian/Pacific Islander, 6.25% Black African American, and 6.25% prefer not to disclose. The sum is greater than 100% because some students reported themselves in more than one category. The course was taught twice weekly in 2 to 3 h sessions. During the first 2 weeks of the course, students performed skills-building activities involving accurate micro-pipetting, sterile technique, and basic bacteriology (i.e., pouring Petri plates, streak plates, and liquid transfers). After completing the skills-building portion of the course, the students conducted their investigations. Students' grades were based on written laboratory reports and exams. In 2019, students elected to conduct a study to compare the prokaryote composition of summer versus winter aquatic communities. In 2020, students chose to study the prokaryotic community of two park ponds. In both terms, the students performed dilution plate count assays, field-collected water samples, and isolated eDNA. Due to the COVID-19 epidemic, the students in 2020 were unable to complete their study because the course was switched to an online format during the last 5 weeks. For the online component, the students independently analyzed the data generated by the 2019 students. Both years, students were taught how to interpret species accumulation curves (Knell, 2018), principal component analysis (Starmer, 2015, 2017), and hierarchically clustered heatmaps (Starmer, 2016) by watching online videos. In 2020, the instructor created a video tutorial on how to use MicrobiomeAnalyst (Dhariwal et al., 2017; Chong et al., 2020), which was posted on a course-management website.

To determine if the course provided the expected outcome of a CURE, the Laboratory Course Assessment Survey (LCAS) was administered during the last week of the course (Corwin et al., 2015). The LCAS is a validated psychometric instrument that assesses students' views of the frequency of collaboration, perception of creating new scientific knowledge, and frequency they needed to repeat and evaluate their experimental results. To assess student attitudes regarding next-generation sequencing technologies, the Genome Consortium for Active Teaching – Sequencing Group (GCAT-SEEK) questionnaire (Buonaccorsi et al., 2011; Tobin and Shade, 2018) was administered the first week of the course and the last week of the course. Additionally, an end-of-term survey written by the instructor was given to the students as a qualitative assessment. All the surveys and questionnaires were taken anonymously.

Study Site

Samples were taken from an artificial cove in Lake Saint Clair (Harrison Township, MI, United States; latitude 42.561496, longitude −82.843249; **Figure 1**). The cove was created when a stone and earth breakwater was installed to create a boat harbor. The cove is located next to the mouth of the Clinton River Bypass, a flood-control canal that can carry Clinton River sediments (Francis and Haas, 2006; Healy et al., 2008). The harbor was abandoned when the property was acquired by the Michigan Department of Natural Resources. Natural successional processes have been allowed to occur in the cove for several years. Sediments from the Clinton River Bypass have been

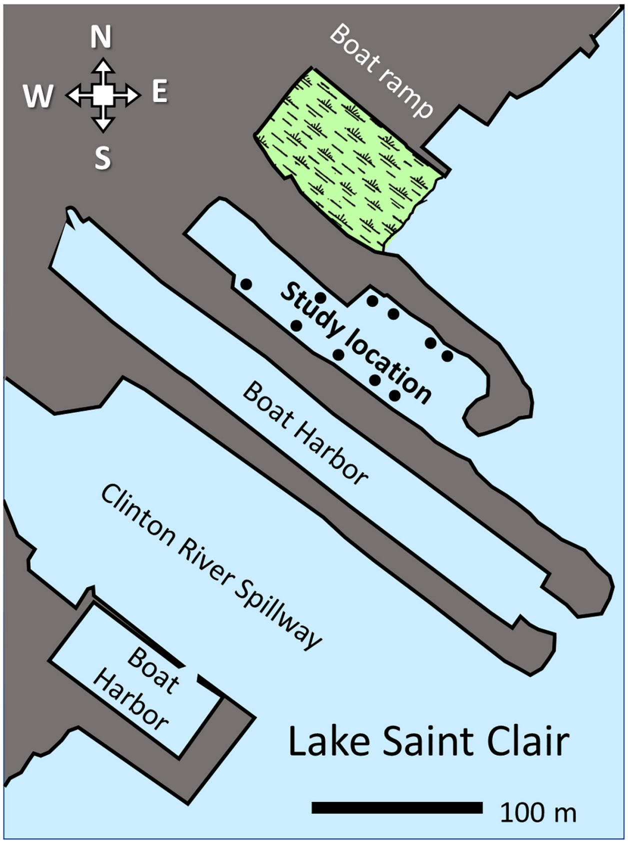


FIGURE 1 | Map of the study's location. Dots represent water sampling locations. Wetland contiguous to the study site is shaded in green. The range bar is 100 m.

accumulating. As a result, the water depth was approximately 1 to 2.5 meters, with the shallowest portion near the mouth of the harbor. A rich community of aquatic vegetation, invertebrates, fish, and turtles resided in the cove.

Water Sampling

Water samples were collected in gamma-irradiated sterile bottles placed on ice and transported back to the laboratory. Surface water samples were collected in summer (June 22, 2016) from ten different locations (**Figure 1**). To exclude floating plant material, the water was filtered through autoclaved rayon polyester mesh (22–25 μm pore size) during collection. After collection, the bottles were capped with an airtight closure. The water temperature was 23°C. To collect water in the winter (February 5, 2019), an autoclaved ice auger (15 cm diameter) was used to drill holes through 10 cm to 61 cm of ice. The auger was sterilized with 95% ethanol between samplings. The holes were drilled near the same location as the ten summer water samples. The water temperature underneath the ice was 0.8°C. A surface sterilized pole was used to lower the collection bottle below the ice. An ethanol sterilized rubber stopper was removed from the mouth of the sampling bottles by pulling an attached string. Once recovered, the bottles were closed with an airtight cap. The samples were stored on ice until DNA extraction.

Water and Sediment Analysis

After microfiltration (see section "DNA Extraction"), the sterile cove-water samples were placed into a −20°C freezer until analysis. On the day of analysis, the water samples were thawed in a room temperature water bath. Orthophosphate concentrations were measured using Hanna Instruments (Woonsocket, RI, United States) Ultra Low Range Phosphate Reagent kit, which is based on the ammonium molybdate-ascorbic acid method (Environmental Protection Agency, 1978). For orthophosphate analysis, 10 mL of water was transferred to virgin sterile polypropylene tubes. The content of the reagent packet was dissolved into the samples. After a 3 min room temperature incubation, the samples were transferred to a 5 cm long cuvette. The absorbance at 708 nm was measured. Winter samples were measured by Applied Metagenomics students. Summer samples were measured by students enrolled in Ecology Laboratory during the fall of 2019. Water hardness, ammonia, and nitrate levels were measured using Hanna Instruments model HI83200 Multiparameter Photometer kits.

During the fall 2018 semester, students enrolled in Ecology Laboratory performed chemical assays on the cove's benthic sediments. Samples were collected by attaching a plastic beaker to a 3 m pipe. To remove the excess water from the sample, small colanders were lined with coffee filter paper and allowed to drip. The LaMotte (Chestertown, MD, United States) Soil Analysis Kit (5010-01) was used to measure phosphorous, potassium, nitrogen, and pH.

DNA Extraction

Within 2 h of sampling, bacteria were isolated by passing the samples through gamma-irradiated disposable microfiltration (pore size 0.2 μm, diameter 47 mm) apparatuses. The apparatuses had closures to prevent contamination. Immediately after vacuum filtration, the apparatuses were moved to a laminar flow hood. Membranes were cut out using sterile scalpels, transferred to gamma-irradiated polystyrene Petri plates, and cut into small fragments. To prevent cross contamination, virgin sterile scalpel blades were used for each membrane filter. The eDNA was isolated using the Zymo Research (Irvine, CA, United States) Quick-DNA Fecal/Soil Microbe Miniprep Kit (Catalog number D6010). As a control, membranes were wetted with 100 μL of the kit's elution buffer and processed like the other filters. Cell disruption and lysis were performed by placing membrane fragments into the kit's lysis tubes. A Bead Bug Homogenizer (Benchmark Scientific, Sayreville, NJ, United States) shaken at 4,000 cycles per minute was used for 180 s. The manufacturer's instructions were followed for the remaining DNA purification steps. To remove contaminating RNA, isolated DNA was treated with 1/10 volume of 10 mg/mL RNase A (37°C for 30 min). The DNA was purified and size selected (>500 pb) using 0.65X volume of Mag-Bind Total Pure NGS magnetic beads (Omega Bio-tek, Norcross, GA, United States) per the manufacturer's instructions. DNA purity was assessed by measuring the 260 nm/280 nm optical density

(OD) absorption ratio with a NanoDrop Lite Spectrophotometer (ThermoFisher, Waltham, MA, United States). All samples had an OD260/280 ratio of less than 1.9. DNA concentrations were measured with an Invitrogen Qubit fluorimeter (ThermoFisher; Double Stranded DNA Broad Range Assay Kit, Catalog number Q32853). The size of the RNase A treated eDNA was evaluated using a rapid gel electrophoresis system (1.2% DNA FlashGel, Lonza Group, Basel, Switzerland).

16S Amplicon Library Construction and Sequencing

Library preparations and sequencing were performed by a commercial service (Molecular Research Laboratory, Shallowater, TX, United States). The 16S rRNA gene variable region V4 (Gray et al., 1984) was amplified using Illumina (San Diego, CA, United States) barcoded oligonucleotides that contain the priming sequences 515F-GTGYCAGCMGCCGCGGTAA (Parada et al., 2016) and 806R-GGACTACNVGGGTWTCTAAT (Apprill et al., 2015). Polymerase chain reaction (PCR) was performed using the HotStarTaq Plus Master Mix Kit (Qiagen, Hilden, Germany). The thermocycling protocol was as follows: polymerase activation by heating at 94°C for 3 min; 28 cycles of melting at 94°C for 30 s; annealing at 53°C for 40 s; and primer extension at 72°C for 1 min. An additional elongation step of 72°C for 5 min was added to the last cycle. After 2% agarose gel electrophoresis, successfully produced amplicons were pooled in equal molar amounts and purified using Ampure XP beads (Beckman Coulter Life Sciences, Indianapolis, IN, United States). The library was sequenced with an Illumina MiSeq using the manufacturer's protocol. After sequencing, barcodes were removed. Sequences shorter than 150 pb were purged, and chimeras were removed. Ten of ten samples were successfully sequenced from the winter samples while nine of ten samples were successfully sequenced from the summer samples.

Analysis Pipeline

To facilitate data processing by undergraduates with no command-line computing experience, software pipelines with web-based graphical user interfaces (GUI) were used. A flow-chart of the data analysis steps used is shown in **Figure 2** and a detailed description of how the students completed the steps is presented in **Supplementary Table 1**. Metagenomics Rapid Annotation using Subsystem Technology (MG-RAST; Meyer et al., 2008) version 4.0.3 (Argonne National Laboratory, 2017) web-based pipeline was used as a sequence data repository, to perform data quality control, and to query the 16S rRNA databases (Quast et al., 2012).

Students in the 2020 course performed data analysis by accessing the web-based MicrobiomeAnalyst pipeline (Dhariwal et al., 2017). After data uploading, the students used the Projection with Marker Data Profiling (PPD) pipeline. To deal with data paucity in low abundance taxa (Weiss et al., 2017), a filter was used to remove OTUs with fewer than four counts in 20% of the data cells. After filtering, 341 OTU's were assessed. To deal with variability in library sizes, the data was rarefied without replacement to the minimum library size. The data was normalized by total sum scaling (Weiss et al., 2017). The pipeline was used to analyze the data with rarefaction curves, alpha-diversity tools, and beta-diversity tools. Additionally, differential abundance was evaluated by using built-in RNAseq tools [DEseq2 algorithm (Love et al., 2014)]. An MA-plot (Love et al., 2014) was created by using a spreadsheet to merge the log2-fold change data (M) calculated by DEseq2 with average OTU count data (A). The larger abundance average (summer versus winter) was used to plot the *A*-axis.

RESULTS

Limnology

To evaluate the trophic status of the study site, nutrient concentrations of the water and benthic sediments were measured (**Table 1**). Notably high concentrations of orthophosphate were observed in the water. Additionally, high levels of phosphorous were detected in the sediments.

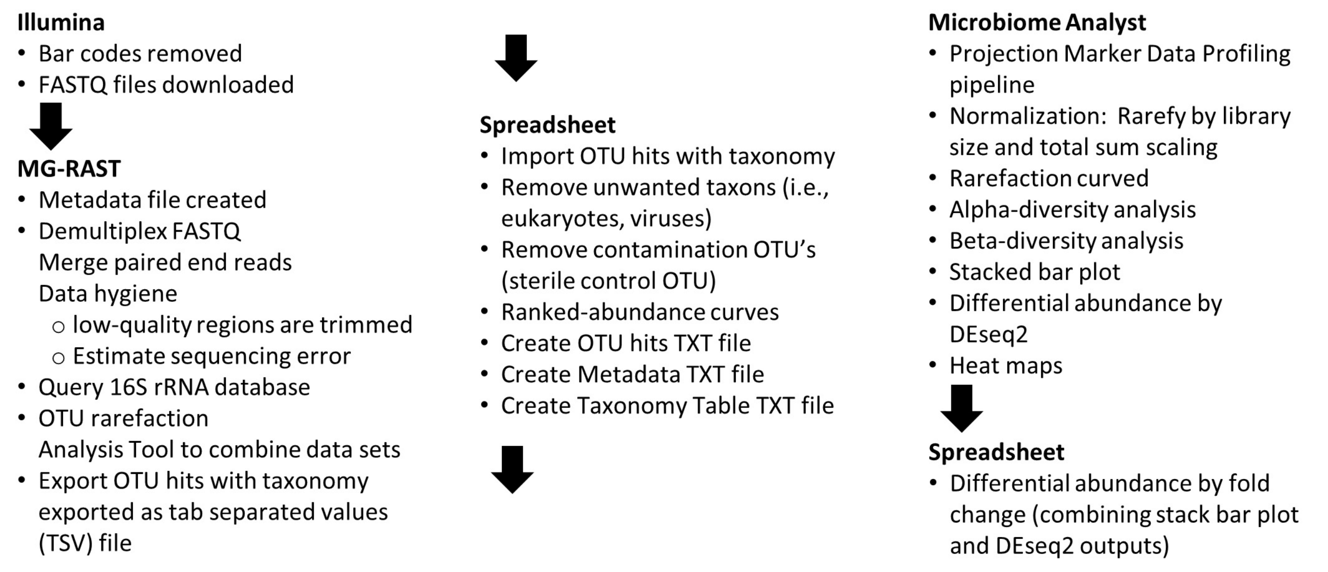
16S Metagenomic Libraries

The students were successful in isolating high quality eDNA from bacteria sampled from the frozen cove. After RNaseA treatment, the DNA had a modal size of ≥4 kbp (**Supplementary Figure 1A**) and was successfully used to create libraries containing 16S rRNA encoding amplicons. After Illumina MiSeq sequencing, the SILVA 16S rRNA gene database was queried by the students. The hits observed from the libraries ranged from 14,995 hits to 178,120 hits, with the median being 62,317 hits.

Species accumulation curves were used to determine if the sequenced libraries were representative of the species richness of the prokaryotic communities (**Figure 3**). In both the summer samples and the winter samples, the slopes of the curves of the low-count unfiltered datasets did not produce an asymptote, even library S3 that produced over 120,000 hits to the SILVA 16S rRNA database. The lack of an asymptote indicates that the libraries did not capture the complete species richness of the prokaryotic community. Additionally, the graphs show that the sequencing runs did not produce datasets of equal sampling efforts, especially the libraries made from the summer samples. As a result, the data sets were rarefied before subsequent analysis. The bottom panels of **Figure 3** showed that the filtering and data rarefaction produced datasets representing equal sampling efforts, making the data amenable to statistical analysis.

α-Diversity

The students used three methods to compare the α-diversity (i.e., taxonomic diversity within a habitat) of the winter and summer prokaryote communities. Unfiltered data was used to produce ranked abundance curves. Summer versus winter data sets of nearly equal sizes were compared (**Figure 4A** and **Supplementary Figures 1B,C**). The analysis showed that both the summer and winter bacteria populations produced nearly identical genera abundance structure. Even on a log-scale, the distribution produced a steep negative-sloping curve. Analysis



Illumina
- Bar codes removed
- FASTQ files downloaded

MG-RAST
- Metadata file created
- Demultiplex FASTQ
  Merge paired end reads
  Data hygiene
  - o low-quality regions are trimmed
  - o Estimate sequencing error
- Query 16S rRNA database
- OTU rarefaction
  Analysis Tool to combine data sets
- Export OTU hits with taxonomy exported as tab separated values (TSV) file

Spreadsheet
- Import OTU hits with taxonomy
- Remove unwanted taxons (i.e., eukaryotes, viruses)
- Remove contamination OTU's (sterile control OTU)
- Ranked-abundance curves
- Create OTU hits TXT file
- Create Metadata TXT file
- Create Taxonomy Table TXT file

Microbiome Analyst
- Projection Marker Data Profiling pipeline
- Normalization: Rarefy by library size and total sum scaling
- Rarefaction curved
- Alpha-diversity analysis
- Beta-diversity analysis
- Stacked bar plot
- Differential abundance by DEseq2
- Heat maps

Spreadsheet
- Differential abundance by fold change (combining stack bar plot and DEseq2 outputs)

FIGURE 2 | Flow-chart of data analysis steps. The order of activities and the software tools used to accomplish the corresponding tasks are described.

TABLE 1 | Nutrient data.

Water chemistry (Winter 2019)

Parameter	Sample size	Mean	Standard deviation	95% confidence interval		Units
Orthophosphate	9	31.2	11.0	39.6	22.8	mg/L
Hardness, Ca^{2+}	9	98.0	4.6	101.5	94.5	mg/L
Ammonium	3	0.23	0.09	0.46	0.01	mg/L
Nitrate	9	3.3	2.2	5.0	1.5	mg/L
pH	9	6.70	0.27	6.90	6.49	

Sediments (Summer 2018)

Parameter	Sample size	Mean	Standard deviation	Units
Phosphorous	3	224	0	kg/ha
Potassium	3	477	0	kg/ha
Nitrogen	3	17	0	kg/ha
pH	3	7	0	

using Simpson's diversity index indicated that both communities showed similar genera diversity (**Figure 4B**). The difference in the diversity indices was not statistically significant (two-sample t-test assuming unequal variance: winter $\bar{x}$ = 0.8349, s = 0.0367; $\bar{x}$ = 0.7941 and s = 0.1135; pooled degrees of freedom = 8, t = 0.9754, p = 0.3579; and Shapiro Wilk test of normality: winter p = 0.9712, summer p = 0.2081). Similar results were obtained when using Shannon's diversity index (**Figure 4C**, two-sample t-test assuming unequal variance: winter $\bar{x}$ = 2.633, s = 0.1292; summer $\bar{x}$ = 2.650, s = 0.3803; pooled degrees of freedom = 7, t = −0.2495, p = 0.8101; and Shapiro Wilk test of normality: winter p = 0.5777, summer p = 0.5215).

β-Diversity

To evaluate β-diversity (i.e., comparison of taxonomic diversity between habitats), the students used principal coordinate analysis of Bray–Curtis dissimilarity indexes (**Figure 5**). The summer samples and winter samples produced two distinct clusters. ANOSIM showed the clustering to be statistically significant (r = 0.59829, p < 0.001). To determine which phyla were responsible for the observed differences in β-diversity, changes in abundance were analyzed.

To visualize which phyla were associated with winter versus summer communities, the students created stacked bar charts (**Figure 6A**). The graph shows that the majority of the observed phyla were present in low abundance. Only

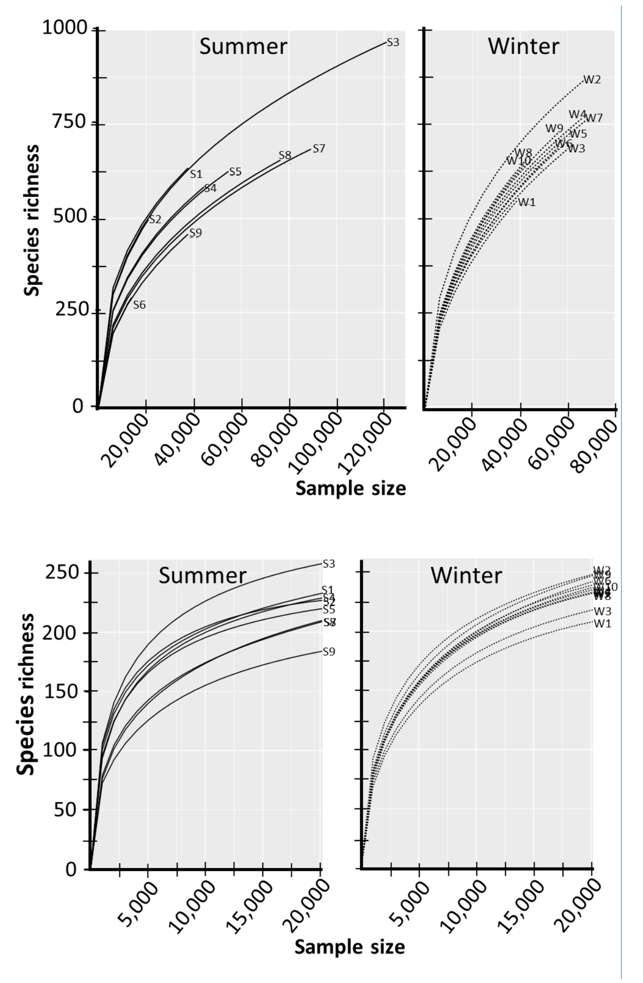


FIGURE 3 | Species curves of unaltered and filtered-rarefied datasets. Raw OTU counts that represent species richness are presented in the top two panels. The filtered and rarefied datasets are presented in the bottom two panels. Libraries from summer-collected samples are labeled with S and winter-collected samples with W.

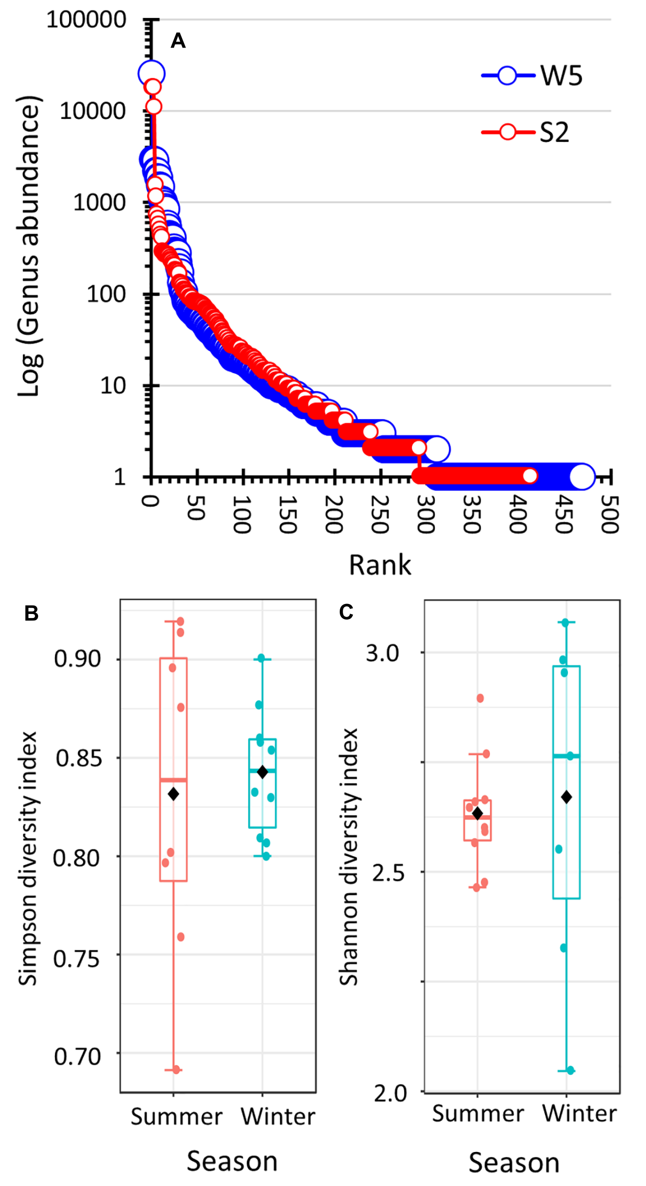


FIGURE 4 | Comparison of the alpha-diversity of summer and winter prokaryote populations. **(A)** Ranked-abundance curve by genera of summer sample (S2) and winter sample (W5). Both libraries produced nearly identical sampling efforts (62,317 hits and 63,700 hits to the SILVA 16S rRNA database, respectively). The S2 data was normalized to 63,700 hits by multiplying the hit count for each genus by 1.022. **(B,C)** Box-and-whisker plots comparing Simpson diversity and Shannon diversity, respectively, of the summer versus winter prokaryote populations. The data set was filtered and rarefied.

one phylum, Proteobacteria, was highly prevalent and showed increased abundance in the winter. Additionally, only one phylum, Verrucomicrobia, was highly prevalent and showed increased abundance in the summer. Dendrograms with differential abundance heat-maps (**Figure 6B**) produced a distinct summer clade and a winter clade. Similar results were also produced when taxonomic orders were analyzed (**Supplementary Figure 1D**).

An MA-plot (**Figure 7**) was used to show the differential abundance of genera. Of the 230 genera in the analysis, 80 had greater abundance in the winter samples and 59 were more abundant in the summer samples (**Supplementary Table 2**). Six genera showed substantially increased winter abundance: unclassified within Betaproteobacteria, *Prolixibacter*, unclassified within the *Sphingobacteriaceae*, *Delftia*, and *Pedobacter* (descending order). Five genera showed substantially increased summer abundance: *Clostridium*, *Cryobacterium*, *Rubritalea*, unclassified within the Gamaproteobacteria, *Terrimonas*, and *Chthoniobacter*.

Course Assessment

Students' perceived experiences in conducting authentic research were assessed using the LCAS (**Figure 8**). The Collaboration component of the LCAS assesses the frequency that collaborative

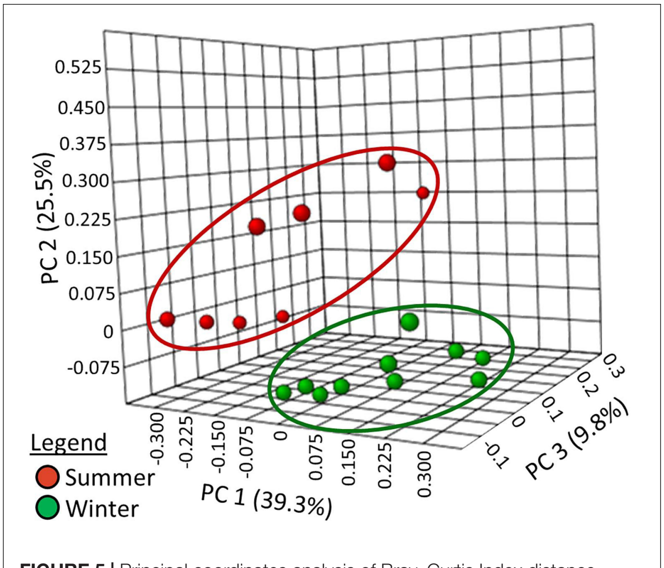


FIGURE 5 | Principal coordinates analysis of Bray–Curtis Index distance measurements at the taxonomic level of the genera.

activities occurred during the course. Two-tail sign-tests were used to assess the null hypothesis that Collaborative activities occurred monthly. In 2019 (**Figure 8A**), responses to questions C1, C2, C4, C5, and C6 were statistically significant ($p < 0.05$). The results indicated that collaborative activities were perceived to occur more frequently than monthly, with the median response corresponding to weekly. The null hypothesis was accepted for C3 (6 positives, 1 negative, and $p = 0.1250$). In 2020 (**Figure 8C**), responses to all the Collaboration questions were statistically significant ($p < 0.01$). The results indicated that the students' perceived collaborative activities more frequently than monthly, with the median value being weekly.

The Discovery section of the LCAS (**Figures 8B,D**) assesses students' perceptions of their experiments contributing to new scientific knowledge. The Iteration section assesses student perceptions of the frequency that procedures were duplicated and the frequency that experiments were repeated to resolve problems with their data. Both sections used a six-point Likert scale. The students' responses were evaluated with two-tail sign-tests, using the null hypothesis median = 3.5. For the 2019 class, all the Discovery questions and Iteration questions produced statistically significant responses ($p < 0.05$). For the 2020 class, all the Discovery questions and Iteration questions produced statically significant responses at $p < 0.001$. The results indicated that the students perceived that they participated in iteration-processes associated with the scientific method and their research activities were scientifically relevant.

In addition to the LCAS, a survey created by GCAT-SEEK was used to evaluate the students' attitudes and perceptions related to next-generation sequencing. The results from the 2019 course (**Figure 9**) indicated that the students felt their understanding of genetics, biochemistry, and bioinformatics increased after completing the course. Analyses of the Understanding questions with Mann–Whitney *U*-tests detected statically significant ($p < 0.05$) increases of median response scores for all questions. Additionally, the two questions related to students' bioinformatics skills showed statistically significant increases. The students also showed a statistically significant increase in their "enthusiasm" regarding next-generation sequencing (question A1). They also indicated increased confidence (questions A3 to A5) in their ability to use next-generation sequencing in future research. There was no change in students' interest in taking additional courses (question A2), possibly because their initial interest was already high (median = 4.5 on a 5-point scale). Students answering the questionnaire in 2020 reported high scores in all categories of the questionnaire. As a result, no statistically significant changes were observed in the pre-course/post-course median responses. Comparisons of pre-course questionnaire responses from 2019 to 2020 showed the 2020 students had statistically greater median scores ($p < 0.05$) for questions U1 to U4, S1, S2, and A5. The results indicated that the 2020 students felt they had a greater understanding of the concepts and better analytical skills at the beginning of the course than their 2019 counterparts.

For qualitative assessment, an anonymous end-of-course student questionnaire was given to the students. Their verbatim responses are presented in **Supplementary Table 3**. In 2019 and 2020, students' responses to the question, "What aspects of the course did you like?" were longer than their responses to "What changes can be made to improve the course." Some noteworthy comments made in 2019 related to the students' positive attitudes toward field collections during the winter. Because the course switched to an online format due to the COVID-19 epidemic in 2020, those students did not have the opportunity to participate in field collection. A theme observed in both the 2019 and 2020 surveys was student comments on the hands-on nature of their experience, collaborations with their peers, repeating procedures that did not work in the first attempt and performing experiments where the answers were not already known. Some students noted that having cycles of draft and revision of their laboratory reports was beneficial to their learning.

DISCUSSION

Students Conducted Authentic Research

Undergraduates enrolled in the CURE course Applied Metagenomics were successful in conducting an authentic scientific investigation. The students' chemical analysis of the water samples (**Table 1**) showed high orthophosphate concentrations. As a result, the cove-water can be classified as eutrophic (Carlson, 1977). The presence of dense mats of three duckweed species (Baker, 2018) also indicates that the water is eutrophic (Landesman et al., 2011). The likely source of nutrients is the Clinton River Bypass (**Figure 1**), a waterway high in nutrient and sediment pollution (Healy et al., 2008). Sediments from the bypass can be observed entering the cove (personal observation). Analysis of the benthic sediments collected from the cove showed high nutrient levels, including phosphorous (**Table 1**).

The students used some standard computational approaches (Gotelli and Chao, 2013) to evaluate community diversity.

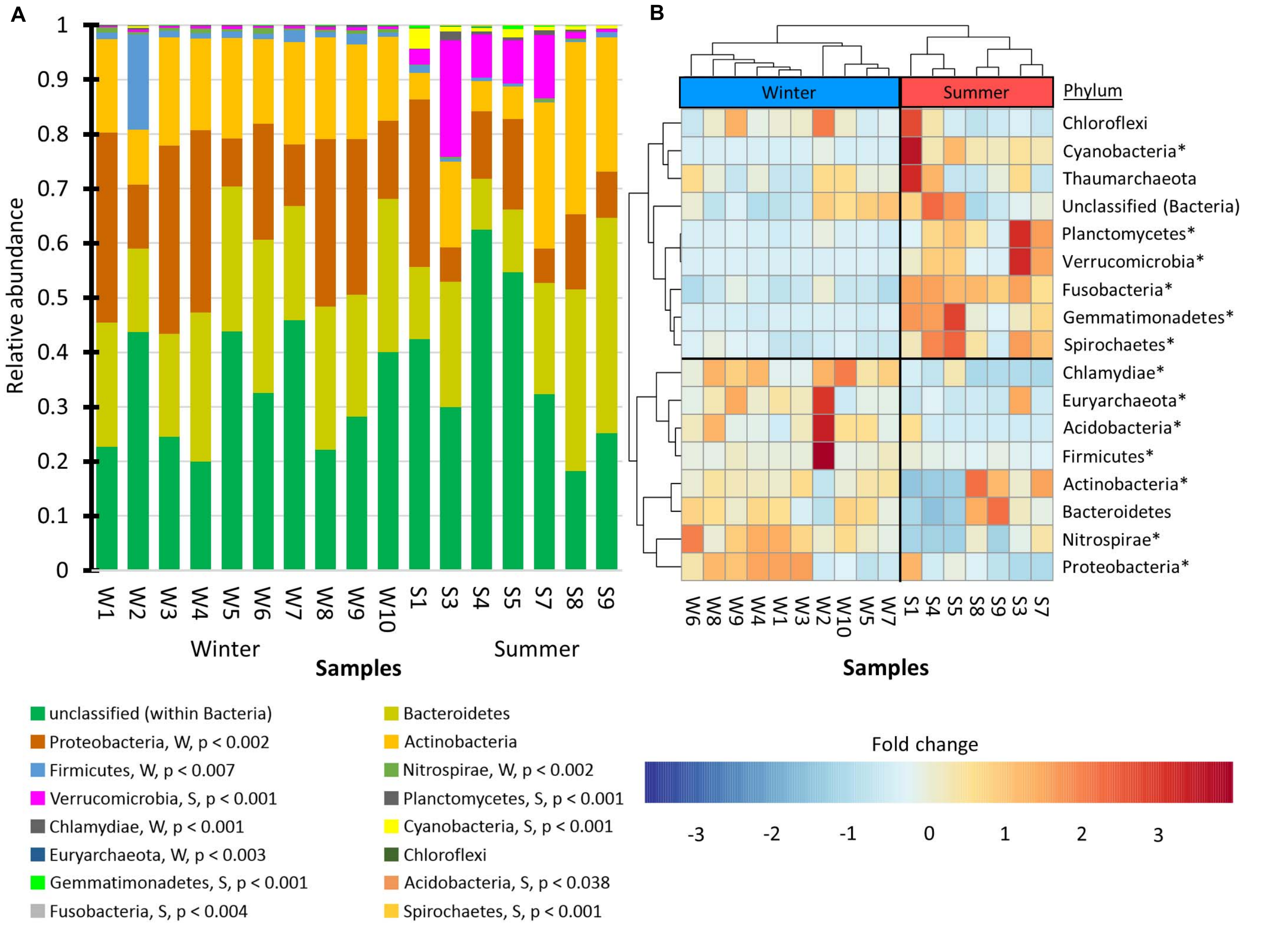


FIGURE 6 | Changes in prokaryote abundance by phyla. **(A)** Stacked bar chart of relative abundance. The phyla were arranged from greatest mean winter abundance to lowest mean winter abundance. DEseq2, with a false discovery rate set at 0.05, was used to assess the statistical significance of the fold changes. Phyla that showed greater winter abundance are indicated with a "W" while those with greater summer abundance are indicated with "S." The adjusted *p*-values are shown. **(B)** Dendrogram with heat-maps that were hierarchically clustered by average Pearson correlation coefficient. Phyla that were statistically significant in **(A)** are marked with asterisks in **(B)**.

Species accumulation curves (**Figure 3**) indicated that the 16S rRNA sequence data sets did not sample all the species present in the summer and winter samples. The largest sequencing library (S3) detected 1,032 OTUs (**Supplementary Table 2**). Because detected species richness is a function of sampling effort (Gotelli and Chao, 2013) and the libraries had over a 10-fold difference in sequencing depth (Weiss et al., 2017), the students conducted most of the subsequent data analysis with rarefied datasets.

Multiple approaches were used to evaluate α-diversity by the students. One approach was to use spreadsheets to create ranked abundance curves (Smith and Smith, 2015) using data from libraries of equal sequencing depth (**Figure 4A** and **Supplementary Figures 1B,C**). Since 16S rRNA barcoding cannot reliably classify bacteria to the species level (Lebonah et al., 2014), the ranked abundance curves were created at the genera level as defined by the SILVA 16S rRNA databases (Quast et al., 2012). The graphs had backwards-J shapes indicating the communities were comprized of one to three highly abundant genera. The winter and summer lines on the graphs overlapped, which indicated that the amount of prokaryote diversity in the winter samples was the same as in the summer samples. This conclusion was supported by calculating diversity indices. Simpson's and Shannon's diversity indices measure diversity by considering the number of taxa and the evenness of distribution of the taxa (Smith and Smith, 2015; Kim et al., 2017). The box-and-whisker plots of both diversity indices overlapped, thus showing no difference in α-diversity between summer and winter samples.

The students used the Bray–Curtis dissimilarity index to evaluate β-diversity. This index was chosen because it is the complement of the Sørensen similarity index, a community comparison index presented in many undergraduate ecology textbooks (Smith and Smith, 2015). Principle coordinate analysis (**Figure 5**) showed that the winter and summer community compositions were distinct. Dendrograms with heat-maps were used to display the differential abundance of phyla (**Figure 6**) and

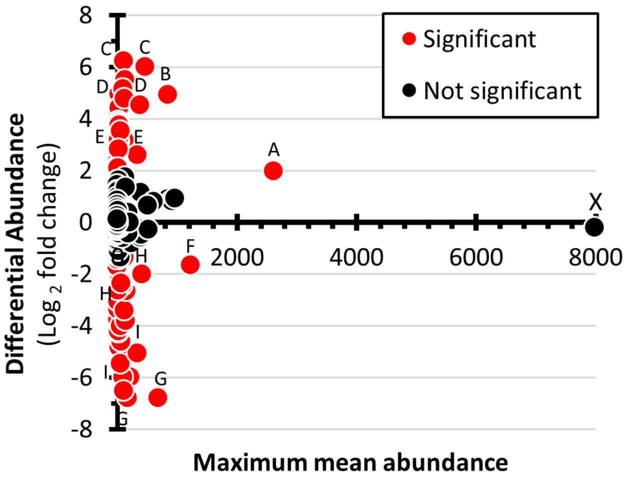


FIGURE 7 | Differential abundance of prokaryote genera. Maximum mean abundance (summer versus winter) is presented on the *X*-axis. Differential abundance expressed as $\log_2$ [(mean winter abundance) – (mean summer abundance)] is presented on the *Y*-axis. Genera with greater winter relative abundance are given positive values and those with greater summer abundance are given negative values. Red data points represent genera that have statistically significant change in abundance, as determined by DEseq2, with a false discovery rate set at 0.05. The labeled data points correspond to the following genera: A, unclassified within Betaproteobacteria; B, *Prolixibacter*; C, unclassified within the Sphingobacteriaceae; D, *Delftia*; E, *Clostridium*; F, *Cryobacterium*; G, *Rubritalea*; H, *Terrimonas*; I, *Chthoniobacter*; and X, unidentified phylum in Bacteria domain.

orders (**Supplementary Figure 1D**). The data clearly showed that the taxonomic composition of the winter prokaryotic community was different than that of the summer community.

The community compositions of the cove (**Figure 6**) contained the same phyla identified as ubiquitous freshwater bacteria by Zwart et al. (2002) and Newton et al. (2011). They are Proteobacteria, Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia. As a result, the students concluded that the composition of the prokaryote community in the cove was typical of freshwater ecosystems.

In contrast, the students concluded that the composition of the frozen cove community was unlike communities in frozen tundra lakes described by Vigneron et al. (2019). When frozen, the Methanogens, Planctomycetes, Chloroflexi, and Deltaproteobacteria became abundant in tundra lakes. In contrast, no Methanogens or Deltaproteobacteria in any of the lake samples were observed by the students (**Figure 6**). The undergraduates did observe Chloroflexi and Planctomycetes, but they were more abundant in the summer samples. In the summer, Actinobacteria and Betaproteobacteria were the predominant phyla in the tundra lake. In contrast, in the Lake Saint Clair samples, Betaproteobacteria were not predominant, and Actinobacteria were more abundant in the winter samples. These results indicated that the community composition of the eutrophic temperate water was distinctly different than the community composition observed in a tundra lake.

The most abundant phylum detected by the students in all water samples was classified as unidentified (**Figure 6**). This phylum contained a single OTU (**Supplementary Table 2**). Thus, this organism likely has not been described by science. OTU2675 represented 31% of the counts in the dataset. Although it is the most prevalent bacterium in the community, it likely does not grow on tryptic soy agar or minimal media. Over 30 bacteria strains have been isolated as pure cultures by students taking an ecology laboratory course. 16S rRNA barcodes of these isolates did not correspond to OTU2675 (personal observation). The probability of obtaining this result due solely to chance is 1.46×10^{-5}. This result showed the students that culture-based methods can miss environmentally prominent organisms, thus illustrating one of the strengths of using metagenomics to study microbial ecology.

Because of their great metabolic diversity, it is difficult to determine the ecological role of prokaryotes by just evaluating higher-level taxa. Thus, differential abundance level was analyzed at the level of genera. With 224 genera in the data set (**Supplementary Table 2**), stacked bar charts and hierarchically arranged heat-maps were inadequate methods of presenting the data. To solve the problem, the students used spreadsheet software to create an MA plot to analyze differential abundance at the level of the genera (**Figure 7**). Four genera stood out as having increased abundance in the winter samples; unclassified within Betaproteobacteria; *Prolixibacter*; unclassified within the Sphingobacteriaceae; *Delftia*; and *Clostridium*. Four genera showed prominently increased abundance in the summer samples; *Cryobacterium*, *Rubritalea*, *Terrimonas,* and *Chthoniobacter*. The most abundant genera corresponded to the unidentified OTU2675 bacteria. Its relative abundance was nearly identical in winter and summer samples. Based on its position on the *A*-axis, this bacterium was the dominant prokaryote in the cove. One possible line of future student investigation is to determine the prevalence of the species in other locations within Lake Saint Clair and other waterways of the Great Lakes Region.

Analyzing the natural history of prominent genera may provide insights into the ecology of the frozen lake and become a basis for students to develop testable hypotheses. For example, datapoint-A (**Figure 7**) corresponds to an unclassified genus within Betaproteobacteria. Betaproteobacteria are often numerically dominant in lake epilimnia, have rapid growth rates, are major components in microbial grazing food chains, and prefer nutrient-rich environments (Newton et al., 2011). Thus, organism-A may have increased its relative winter abundance due to the exploitation of winter-abundant resources. Another example is the genera *Prolixibacter* (datapoint-B). Members of this taxon are non-cellulosic fermenting facultative anaerobes that have been isolated from marine sediments (Holmes et al., 2007) and cold (5°C) peat bogs (Schmidt et al., 2015). Often, biological oxygen demands cause hypoxia in ice-covered lakes (Ellis and Stefan, 1989). Thus, the increased prevalence of *Prolixibacter* may be due to its being adapted to cold low oxygen environments. To test this hypothesis, dissolved oxygen measurements can be conducted of water samples collected from under the ice.

Nutrient availability may be a factor causing an increased abundance of some genera. For example, *Delftia* abundance increased 32-fold in the winter samples. The two corresponding OTUs had high homology to *D. acidovorans* and *D. tsuruhatensis*. The type specimens for these species were isolated from high nutrient environments (Han et al., 2005; Yilmaz and Icgen, 2014). Another genus, *Clostridium*, had a 6-fold greater prevalence in the winter samples. Members of this genera have been isolated from activated sludge (Gumaelius et al., 2001). Many strains are aerobic denitrifiers. The presence of *Clostridium* suggests an active role in nutrient turnover.

The pattern observed in one of the differentially abundant genera is puzzling. *Cryobacterium*, represented by a single OTU, showed a 2.8-fold increased abundance in the summer samples (**Figure 7**, datapoint F). The *Cryobacterium* OTU had high homology to *C. psychrophilum* and was the 2nd most abundant OTU in the summer dataset (**Supplementary Table 2**). The type specimen of *C. psychrophilum* was isolated from samples in Iceland. It grew best in cool water (9 to 12°C) and stopped growing when the temperature reached 18°C (Suzuki et al., 1997). When the water samples were collected in the summer, the surface temperature was 23°C. Thus, the increased prevalence of the *C. psychrophilum*-like bacteria in the summer sample is unexplained and warrants further investigation.

Student Data Analysis Workflow

One of the goals in the development of the Applied Metagenomics CURE course was to overcome computing-barriers in processing metagenomics data. The data presented in this manuscript show that undergraduates without knowledge of computer coding or command-line computing can complete a metabarcoding investigation. However, the students did find some of the computing tasks difficult to accomplish. The nature of the difficulties and strategies used to overcome the bottlenecks are presented in **Supplementary Table 4**.

The approach of using MG-RAST in combination with MicrobiomeAnalyst can be used to analyze shotgun metagenomic sequence data as well since both portals support this type of data. Additionally, undergraduates can use other pipelines to analyze metabarcoding data sets. Recently, CyVerse has beta-released the Purple Line of its DNA Subway (CyVerse, 2019), a GUI-based version of the QIIME 2 pipeline (Bolyen et al., 2019). As a result, students can use more than one approach to process 16S rRNA metabarcoding data.

Limitations When Using CUREs

Though CUREs can contribute to scientific knowledge, there are inherent limitations on the nature of the investigations that can be conducted. For example, undergraduate students do not have access to the array of resources often available in research laboratories. In this course, the students wanted to obtain water samples that were as representative of the prokaryotic community as possible. However, they did not have access to a mobile field laboratory to perform immediate microbial isolation. Though the collection vessels were filled to the lip, closed with an air-tight cap, and kept on wet-ice for less than 2 h, some organisms, such as obligate anaerobes, may have been lost. Many other metagenomic investigations of environmental water samples have stored samples on wet-ice before microbe isolation (Yannarell et al., 2003; Shade et al., 2007; Van Rossum et al., 2015; Uyaguari-Diaz et al., 2016; Linz et al., 2017; Karayanni et al., 2019; Mohiuddin et al., 2019). Thus, the collection procedure that were used by the students is within the norms of basic research.

Another limitation to CURE studies is the timeframe of the investigation. Ideally, a longitudinal study like this one would be conducted over consecutive seasons and multiple years. However, the CURE course only lasted one semester (15 weeks). The students were able to compare different seasons because they were able to utilize a data set created 3 years earlier. Though the primary conclusions are valid (i.e., the microbial community from the ice cover lake samples were as diverse as the open water summer samples, and the compositions of the two communities were strikingly different), the students could not determine the variability of the community structure from one year to the next. Finally, budget constraints limit the number of samples analyzed. For this course, the maximum number of samples, including controls, that could be used in the experimental design was limited to 12 sequencing runs.

Course Assessment

The course was assessed to determine if the goals of a typical CURE were accomplished. The LCAS (Corwin et al., 2015) is designed to measure three attributes of CUREs. Students were asked six questions regarding their perceptions of collaborative activity frequency. The results in **Figure 8** showed that the students felt that they discussed with their peers or the instructor elements of their investigations, reflected on their learning, contributed to discussions, collaborated on data analysis, and collaborated on resolving problems on a weekly basis.

Five questions on the LCAS evaluated the students' perceptions of their research as they relate to scientific discovery and scientific relevance. All questions produced statistically significant responses ($p < 0.05$ in 2019; $p < 0.01$ in 2020) to the null hypothesis of neutral attitude (i.e., Median = 3.5). The lowest median response observed was to question D1, generating novel results unknown to the instructor or scientific community. The lead author (SSB) was surprised by this result because the discovery-nature of the course was explicitly conveyed to the students. In contrast, questions addressing students' perception of their investigating something previously unknown (D2), formulating a hypothesis (D3), developing an argument based on evidence (D4), and creating new scientific knowledge (D5), the median responses were "highly agree" or "agree." To resolve the dichotomy in the students' attitudes, open-response questions will need to be added to future surveys. In total, the students' responses to this section of the LCAS indicate that the students felt their research contributed to scientific knowledge and was scientifically relevant.

Six LCAS questions evaluated the iteration processes used in scientific investigations. All questions produced statistically significant responses ($p < 0.05$ in 2019; $p < 0.01$ in 2020) with the median responses corresponding to "agree" or to "strongly agree." These results indicate that the students felt they repeated

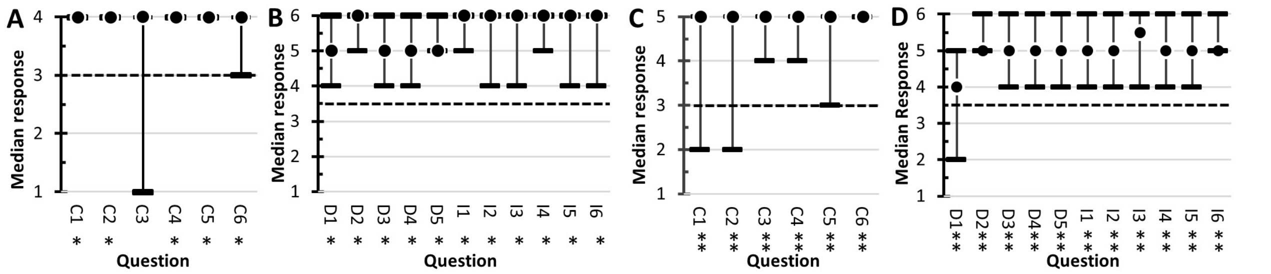


FIGURE 8 | Student responses to the Laboratory Course Assessment Survey. Results from the 2019 survey are shown in **(A,B)**. The 2020 results are in **(C,D)**. The dot represents the students' median response. The range bars are the range of responses. The dotted line corresponds to the null hypothesis used in the sign-tests. All responses statistically significant at α = 0.05 are marked with single asterisks (*) and significance at α = 0.01 are marked with a double asterisk (**). The questions for the Collaboration section **(A,C)** started with, In this course, I was encouraged to, and ended with, (C1) discuss elements of my investigation with classmates or instructors, (C2) reflect on what I was learning, (C3) contribute my ideas and suggestions during class discussions, (C4) help other students collect or analyze data, (C5) provide constructive criticism to classmates and challenge each other's interpretations, and (C6) share the problems I encountered during my investigation and seek input on how to address them. The answer options for the 2019 survey **(A)** were never (1), one or two times (2), monthly (3), and weekly (4). In the 2020 survey **(C)**, the options were never (1), one or two times (2), monthly (3), every other week (5), and weekly (6). The Discovery and Relevance questions **(B,D)** started with, In this course, I was expected to, and ended with (D1) generate novel results that are unknown to the instructor and that could be of interest to the broader scientific community or others outside the class, (D2) conduct an investigation to find something previously unknown to myself, other students, and the instructor, (D3) formulate my own research question or hypothesis to guide an investigation, (D4) develop new arguments based on data, and (D5) explain how my work has resulted in new scientific knowledge. The Iteration section questions started with, In this course, I had time to, and ended with, (I1) revise or repeat work to account for errors or fix problems, (I2) change the methods of the investigation if it was not unfolding as predicted, (I3) share and compare data with other students, (I4) collect and analyze additional data to address new questions or further test hypotheses that arose during the investigation, (I5) revise or repeat analyses based on feedback, and (I6) revise drafts of papers or presentations about my investigation based on feedback. The answer options were, strongly disagree (1), disagree (2), somewhat disagree (3), somewhat agree (4), agree (5), and strongly agree (6).

work to fix problems with their results, changed methods in response to unanticipated results, compared their results to the results of their peers, collected additional data to help revise hypotheses, responded to feedback from others, and revised their written work.

The GCAT-SEEK opinion questionnaire was used to assess students' attitudes to next-generation sequencing technologies (**Figure 9**). In 2019, students reported an increased understanding of the genetic mechanisms related to evolution, the relationships of molecular structure and functions, genome information flow, and how genomes control metabolism ($p < 0.05$). The same students felt their skills in using bioinformatics to identify patterns and making arguments increased after completing the course. The students also indicated a more positive attitude toward research involving next-generation sequencing. The median "enthusiasm" (A1) response increased from 3.5 (a neutral value) to 5 (highly agree). They also reported increased confidence in understanding (A3) and using (A4 and A5) next-generation sequencing data. The students indicated they had a "high" (median = 4.5) interest in performing more research with next-generation sequencing at the start of the course and maintained this interest after the course ($p = 0.610$).

A different response pattern was observed in the 2020 data. The students indicated they had a strong understanding of core concepts (U1 to U5) in the pre-survey, with a median value corresponding to "agree" or "strongly agree." They maintained this opinion after they completed the course [$p = 0.056$ (due to a small increase in the post-survey) to 0.608]. The same patterns were observed with the skills questions ($p = 0.082$ to 0.110) and attitude questions ($p = 0.154$ to 0.984). The results indicate that the students maintained their positive attitudes regarding next-generation sequencing technologies after completing the course. Comparison of the pre-course responses of the 2019 and 2020 classes showed the 2020 class reported higher median scores for the understanding questions and skill questions. The differences were statistically significant for questions A1 to A4 and S1 to S2. These results suggest that the students taking the course in 2020 felt more intellectually prepared for the coursework than did the students in 2019.

Qualitative assessment involved the instructor giving the students anonymous open-ended survey questions (**Supplementary Table 3**). Major themes observed in the student comments indicated that the course contained some of the major elements of CUREs (i.e., Scientific Process, Discovery, Collaboration, and Iteration). Their comments aligned well with the response observed in the LCAS survey (**Figure 9**). In terms of areas for improvement, some students felt that the open-ended nature of the laboratory was "disorganized," and the procedures were too time intensive. In total, the responses in the open-ended survey indicated that the students found the CURE elements (Auchincloss et al., 2014) of the course helpful to their learning.

Supporting the Goals of *Vision and Change*

Vision and Change is a joint policy statement of the American Association for the Advancement of Science, the National Academy of Sciences, and other organizations on how undergraduate biology curricula should be reformed during the 21st century (Bauerle et al., 2009). Because of the ever-expanding nature of science, *Vision and Change* calls for biology education to focus on a few key concepts, develop student investigative

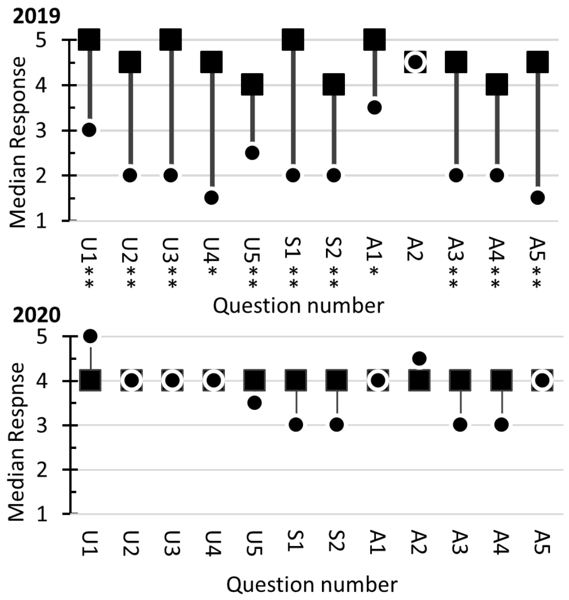


FIGURE 9 | Student responses to the GCAT-SEEK opinion survey in 2019 and 2020. The survey used a Likert scale response system (1 "not at all" to 5 "a great deal"). The dots (●) represent the students' median response in the pre-course survey. The squares (■) indicate the students' median response in the post-course survey. A two-tailed Mann–Whitney *U*-test for two independent samples was used to assess the null hypothesis $\text{Median}_{Pre-course} = \text{Median}_{Post-course}$. Statistically significant differences at $\alpha = 0.05$ and $\alpha = 0.01$ are marked with single and double asterisks, respectively. The questions related to student perception of their Understanding (U) started with the phrase, Presently, I understand, and ended with U1, the genetic mechanisms that underlie evolution (mutation, selection, migration, drift, and etcetera.); U2, the relationship between basic units of molecular structure and their function; U3, how bioinformatics can be used to understand the flow, exchange, and storage of information from genome to phenotype; U4, how the genome confers metabolic capabilities to an organism; and U5, how genomic analysis can elucidate larger scale interactions within organisms, between organisms, and/or between organisms and ecosystems. The Skills (S) questions assess students' perception of their abilities and started with the phrase, Presently, I can, and ended with S1, identify patterns in bioinformatics data; S2, recognize a sound argument based on the appropriate use of bioinformatics evidence. The Attitudes (A) questions started with the phrase, Presently, I am, and ended with A1, enthusiastic about next-generation sequencing; A2, interested, if the opportunity is available, in taking further courses/performing more research in this topic area; A3, confident that I understand next-generation sequencing technologies; A4, confident that I can analyze next-generation sequencing data; and A5, confident that I can incorporate next-generation sequencing technologies into my research.

competencies and enhance student engagement in the scientific process (Woodin et al., 2010). The Applied Metagenomics course incorporates many of the *Vision and Change* recommendations. For example, students used the concept of evolution and biological information flow to analyze the results of their experiment. Additionally, the students developed competencies in using large data sets and computational analysis. Moreover, mathematical and communication skills were developed by having students write formal laboratory reports where they had to interpret their numeric data and clearly present their results with graphs. Finally, the students were fully engaged in the scientific process, because the research they performed was authentic and contributed to the knowledgebase of society.

The development of educational strategies that help retain undergraduate underrepresented minority students is identified as one of the "pressing needs" in *Vision and Change*. Large-scale studies of 6-year graduation rates showed that CUREs increase retention of underrepresented minority students (Jones et al., 2010; Schultz et al., 2011). CUREs may increase retention because the self-efficacy of underrepresented minorities increases when they participate in research (Hurtado et al., 2009). The results of the course assessment (**Figures 8**, **9**) indicate that this metagenomic CURE course had a positive impact on the students' attitudes toward research and thus has the potential of improving retention of underrepresented minority students.

CONCLUSION

The instructional approach utilized in the Applied Metagenomics course can be used as a template to foster the development of additional CURE courses. The course was designed to overcome potential computational barriers (Maloney et al., 2010) by using publicly available web-based resources. Additionally, the data-analysis workflow used did not require students to learn command-line computing or programming. The students' research was relevant because the sequence data was posted in a data repository and their research findings are published here. Additionally, the students' data (i.e., posted sequence data and the OTU count data) can be used to develop additional *in-silico* activities for undergraduate instruction.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The name of the repository and accession numbers can be found in **Supplementary Table 2**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by University of Detroit Mercy's Institutional Review Board (Protocol Number 1718-53). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SB conceived of the project, created all instructional materials, confirmed all student calculations, created all figures, and wrote the manuscript. MA, KA, JC, GC, ZD, SK, ES, and SS collected environmental samples, isolated DNA, and did the primary data analysis with MG-RAST. ES, GC, ZD, NA, ZA-H, VC, DC, MH, MJ, MLJ, ZK, EK, RK, SK, AM, PP, RR, and ST performed additional data analysis with MicrobiomeAnalyst. All authors contributed to the article and approved the submitted version.

FUNDING

Work reported in this publication was supported by the National Institutes of Health Common Fund and Office of Scientific Workforce Diversity under three linked awards RL5GM118981, TL4GM118983, and 1UL1GM118982 administered by the National Institute of General Medical Sciences.

ACKNOWLEDGMENTS

The authors would like to thank Jonathan Finkel, Emily F. Brown Baker, and Brian M. Napoletano for their careful reading of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.579325/full#supplementary-material

Supplementary Figure 1 | Supplemental student data.

Supplementary Table 1 | Detailed description of the students' data processing.

Supplementary Table 2 | Prokaryotic count data and abundance analysis.

Supplementary Table 3 | Results of anonymous end-of-course student questionnaire.

Supplementary Table 4 | Overcoming bottlenecks in student workflow.

REFERENCES

Apprill, A., McNally, S., Parsons, R., and Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquat. Microb. Ecol.* 75, 129–137. doi: 10.3354/ame01753

Argonne National Laboratory (2017). *MG-RAST Metagenomics Analysis Server*. Available online at: http://www.mg-rast.org (accessed January 20, 2020)

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., et al. (2014). Assessment of course-based undergraduate research experiences: a meeting report. *CBE Life Sci. Educ.* 13, 29–40.

Baker, S. S. (2018). "Using DNA barcoding to identify duckweed species as part of an undergraduate ecology course," in *Environmental Chemistry: Undergraduate and Graduate Classroom, Laboratory, and Local Community Learning Experiences*, eds E. S. Roberts-Kirchhoff and M. A. Benvenuto (Washington, DC: American Chemical Society), 67–79. doi: 10.1021/bk-2018-1276.ch005

Bauerle, C., DePass, A., Lynn, D., O'Connor, C., Singer, S., Withers, M., et al. (2009). *Vision and Change in Undergraduate Biology Education: A Call to Action*. Washington, DC: American Association for the Advancement of Science.

Bell, J. K., Eckdahl, T. T., Hecht, D. A., Killion, P. J., Latzer, J., Mans, T. L., et al. (2017). CUREs in biochemistry—where we are and where we should go. *Biochem. and Mol. Biol. Edu.* 45, 7–12. doi: 10.1002/bmb.20989

Bolsenga, S. J., and Herdendorf, C. E. (1993). *Lake Erie and Lake Saint Clair Handbook*. Detroit, MI: Wayne State University Press.

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat. Biotechnol.* 37, 852–857.

Buonaccorsi, V. P., Boyle, M. D., Grove, D., Praul, C., Sakk, E., Stuart, A., et al. (2011). GCAT-SEEKquence: genome consortium for active teaching of undergraduates through Increased faculty access to next-generation sequencing data. *CBE Life Sci. Educ.* 10, 342–345. doi: 10.1187/cbe.11-08-0065

Carlson, R. E. (1977). A trophic state index for lakes. *Limnol. Oceanogr.* 22, 361–369. doi: 10.4319/lo.1977.22.2.0361

Chong, J., Liu, P., Zhou, G., and Xia, J. (2020). Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. *Nat. Protoc.* 15, 799–821. doi: 10.1038/s41596-019-0264-1

Corwin, L. A., Runyon, C., Robinson, A., and Dolan, E. L. (2015). The laboratory course assessment survey: a tool to measure three dimensions of research-course design. *CBE Life Sci. Educ.* 14:ar37. doi: 10.1187/cbe.15-03-0073

CUREnet (2013). *Urban Microbial Dynamics and Metagenomics*. Available online at: https://curenet.cns.utexas.edu/projects/urban-microbial-dynamics-and-metagenomics (accessed January 6, 2018).

CyVerse (2019). *DNA Subway: Fast Track to Gene Annotation and Genome Analysis Home Page*. Available online at: https://dnasubway.cyverse.org/ (accessed January 5, 2021).

Dhariwal, A., Chong, J., Habib, S., King, I. L., Agellon, L. B., and Xia, J. (2017). MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. *Nucleic Acids Res.* 45, W180–W188.

Ellis, C. R., and Stefan, H. G. (1989). Oxygen demand in ice covered lakes as it pertains to winter aeration. *J. Am. Water Resour. Assoc.* 25, 1169–1176.

Environmental Protection Agency (1978). "Method 365.3: Phosphorous, all forms (colorimetric, ascorbic acid, two reagent)," in *Clean Water Act Analytical Methods*, (Washington, DC: United States Environmental Protection Agency).

Estrada, M., Burnett, M., Campbell, A. G., Campbell, P. B., Denetclaw, W. F., Gutiérrez, C. G., et al. (2016). Improving underrepresented minority student persistence in STEM. *CBE Life Sci. Educ.* 15:es5. doi: 10.1187/cbe.16-01-0038

Francis, J. T., and Haas, R. C. (2006). *Clinton River Assessment*. Fisheries Special Report 39. Ann Arbor, MI: Michigan Department of Natural Resources.

Gotelli, N. J., and Chao, A. (2013). "Measuring and estimating species richness, species diversity, and biotic similarity from sampling data," in *Encyclopedia of Biodiversity*, 2nd Edn, ed. S. Levin (Waltham, MA: Academic Press), 195–211. doi: 10.1016/b978-0-12-384719-5.00424-x

Graham, M. J., Frederick, J., Byars-Winston, A., Hunter, A.-B., and Handelsman, J. (2013). Increasing persistence of college students in STEM. *Science* 341, 1455–1456. doi: 10.1126/science.1240487

Gray, M. W., Sankoff, D., and Cedergren, R. J. (1984). On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. *Nucleic Acids Res.* 12, 5837–5852. doi: 10.1093/nar/12.14.5837

Gumaelius, L., Magnusson, G., Pettersson, B., and Dalhammar, G.(2001). *Comamonas denitrificans* sp. nov., an efficient denitrifying bacterium isolated from activated sludge. *Int. J. of Syst. Evol. Microbiol.* 51, 999–1006.

Han, J., Sun, L., Dong, X., Cai, Z., Sun, X., Yang, H., et al. (2005). Characterization of a novel plant growth-promoting bacteria strain *Delftia tsuruhatensis* HR4 both as a diazotroph and a potential biocontrol agent against various plant pathogens. *Syst. and Appl. Microbiol.* 28, 66–76. doi: 10.1016/j.syapm.2004.09.003

Hatfull, G. F. (2015). Innovations in undergraduate science education: going viral. *J. Virol.* 89, 8111–8113. doi: 10.1128/jvi.03003-14

Healy, D. F., Chambers, D. B., Rachol, C. M., and Jodoin, R. S. (2008). *Water Quality of the St. Clair River, Lake St. Clair, and their U.S. Tributaries, 1946–2005*. Scientific Investigations Report 2007–5172. Reston, VA: U.S. Geological Survey.

Holmes, D. E., Nevin, K. P., Woodard, T. L., Peacock, A. D., and Lovley, D. R. (2007). *Prolixibacter bellariivorans* gen. nov., sp. nov., a sugar-fermenting, psychrotolerant anaerobe of the phylum Bacteroidetes, isolated from a marine-sediment fuel cell. *Int. J. of Syst. Evol. Micr.* 57, 701–707. doi: 10.1099/ijs.0.64296-0

Howard Hughes Medical Institute (1996). *Beyond Bio 101*. Chevy Case, MD: Howard Hughes Medical Institute.

Hurtado, S., Cabrera, N. L., Lin, M. H., Arellano, L., and Espinosa, L. L. (2009). Diversifying science: underrepresented student experiences in structured research programs. *Res. High. Educ.* 50, 189–214. doi: 10.1007/s11162-008-9114-7

Jones, M. T., Barlow, A., and Villarejo, M. (2010). The importance of undergraduate research for minority persistence and achievement in biology. *J. High. Educ.* 81, 82–115. doi: 10.1080/00221546.2010.11778971

Jurkowski, A., Reid, A. H., and Labov, J. B. (2007). Metagenomics: a call for bringing a new science into the classroom (While it's still new). *CBE Life Sci. Educ.* 6, 260–265. doi: 10.1187/cbe.07-09-0075

Karayanni, H., Macingo, S. C., Tolis, V., and Alivertis, D. (2019). Diversity of bacteria in lakes with different chlorophyll content and investigation of their respiratory activity through a long-term microcosm experiment. *Water* 11:467. doi: 10.3390/w11030467

Kim, B.-R., Shin, J., Guevarra, R. B., Lee, J. H., Kim, D. W., Seol, K.-H., et al. (2017). Deciphering diversity indices for a better understanding of microbial communities. *J. Microbiol. Biotechnol.* 27, 2089–2093. doi: 10.4014/jmb.1709.09027

Knell, R. (2018). *Species Accumulation Curves*. Available online at: https://youtu.be/Jj7LYrU_6RA (accessed June 28, 2020).

Landesman, L., Fedler, C., and Duan, R. (2011). "Plant nutrient phytoremediation using duckweed," in *Eutrophication: Causes, Conseqeucnes and Control*, eds A. A. Ansari, S. S. Gill, G. R. Lanza, and W. Rast (Dordrecht: Springer), 341–354. doi: 10.1007/978-90-481-9625-8_17

Lebonah, D. E., Dileep, A., Chandrasekhar, K., Sreevani, S., Sreedevi, B., and Pramoda Kumari, J. (2014). DNA barcoding on bacteria: a review. *Adv. Biol.* 2014:541787. doi: 10.1155/2014/541787

Lentz, T. B., Ott, L. E., Robertson, S. D., Windsor, S. C., Kelley, J. B., Wollenberg, M. S., et al. (2017). Unique down to our microbes - Assessment of an inquiry-based metagenomics activity. *J. Microbiol. Biol. Edu* .18:18.12.33.

Linz, A. M., Crary, B. C., Shade, A., Owens, S., Gilbert, J. A., Knight, R., et al. (2017). Bacterial community composition and dynamics spanning five years in freshwater bog lakes. *mSphere* 2:3. doi: 10.1128/mSphere.00169-17

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* 15:550. doi: 10.1186/s13059-014-0550-8

Maloney, M., Parker, J., LeBlanc, M., Woodard, C. T., Glackin, M., and Hanrahan, M. (2010). Bioinformatics and the undergraduate curriculum. *CBE Life Sci. Educ.* 9, 172–174.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., and Kubal, M. (2008). The metagenomics RAST server: a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinform.* 9:386. doi: 10.1186/1471-2105-9-386

Mohiuddin, M. M., Botts, S. R., Paschos, A., and Schellhorn, H. E. (2019). Temporal and spatial changes in bacterial diversity in mixed use watersheds of the Great Lakes region. *J. Great Lakes Res.* 45, 109–118. doi: 10.1016/j.jglr.2018.10.007

National Institutes of Health (2019). *Building Infrastructure Leading to Diversity (BUILD) Initiative*. Bethesday, MD: National Institutes of Health.

National Research Council (1996). *From Analysis to Action: Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Washington, DC: National Academy Press.

National Research Council (2003). *BIO2010: Transforming Undergraduate Education for Future Research Biologists*. Washington, DC: The National Academies Press.

National Research Council (2007). *The New Science of Metagenomics: Revealing the Secrets of Our Microbial Planet*. Washington, DC: The National Academies Press.

National Science Foundation (1996). *Shaping the Future: New Expectations for Undergraduate Education in Science, Mathematics, Engineering, and Technology*. Arlington, VA: National Science Foundation.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., and Bertilsson, S. (2011). A guide to the natural history of freshwater lake bacteria. *Microbiol. Mol. Biol. R.* 75, 14–49. doi: 10.1128/mmbr.00028-10

Newton, R. J., and McLellan, S. L. (2015). A unique assemblage of cosmopolitan freshwater bacteria and higher community diversity differentiate an urbanized estuary from oligotrophic Lake Michigan. *Front. Microbiol.* 6:1028. doi: 10.3389/fmicb.2015.01028

Parada, A. E., Needham, D. M., and Fuhrman, J. A. (2016). Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ. Micro.* 18, 1403–1414. doi: 10.1111/1462-2920.13023

Project Kaleidoscope (1991). *What Works, Building Natural Science Communities: A Plan for Strengthening Undergraduate Mathematics and Sciences*. Washington, D.C: Independent College Office.

Provost, J. (2016). Research for all: a CURE for undergraduates. *ASBMB Today* 15, 30–31.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., et al. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596.

Schmidt, O., Horn, M. A., Kolb, S., and Drake, H. L. (2015). Temperature impacts differentially on the methanogenic food web of cellulose-supplemented peatland soil. *Environ. Microbiol.* 17, 720–734. doi: 10.1111/1462-2920.12507

Schultz, P. W., Hernandez, P. R., Woodcock, A., Estrada, M., Chance, R. C., Aguilar, M., et al. (2011). Patching the pipeline: reducing educational disparities in the sciences through minority training programs. *Educ. Eval. Policy Anal.* 33:10. doi: 10.3102/0162373710392371

Shade, A., Kent, A. D., Jones, S. E., Newton, R. J., Triplett, E. W., and McMahon, K. D. (2007). Interannual dynamics and phenology of bacterial communities in a eutrophic lake. *Limnol. Oceanogr.* 52, 487–494. doi: 10.4319/lo.2007.52.2.0487

Smith, T. M., and Smith, R. L. (2015). *Elements of Ecology*, 9th.edn Edn. New York, NY: Pearson.

Snyder, K., and Kumar, A. (2019). *ReBUILDetroit: Building Infrastructure Leading to Diversity*. Detroit. Available online at: http://rebuildetroit.org/ (accessed January 3 2020)

Starmer, J. (2015). *Principal Component Analysis (PCA) Clearly Explained*. Available online at: https://youtu.be/_UVHneBUBW0 (accessed June 28, 2020).

Starmer, J. (2016). *Drawing and Interpreting Heatmaps*. Available online at: https://youtu.be/oMtDyOn2TCc (accessed June 28, 2020).

Starmer, J. (2017). *StatQuest: PCA Main Ideas in only 5 minutes!*. Available online at: https://youtu.be/HMOI_lkzW08 (accessed June 28, 2020).

Suzuki, K.-I., Sasaki, J., Uramoto, M., Nakase, T., and Komagata, K. (1997). *Cryobacterium psychrophilum* gen. nov., sp. nov., nom. rev., comb. nov., an obligately psychrophilic actinomycete to accommodate "*Curtobacterium psychrophilum*" Inoue and Komagata 1976. *Int. J. Syst. Evol.* 47, 474–478. doi: 10.1099/00207713-47-2-474

Tobin, T. C., and Shade, A. (2018). A town on fire! Integrating 16S rRNA gene amplicon analyses into an undergraduate microbiology lecture class. *FEMS Microbiol. Lett.* 365:fny104. doi: 10.1093/femsle/fny104

Toven-Lindsey, B., Levis-Fitzgerald, M., Barber, P. H., and Hasson, T. (2015). Increasing persistence in undergraduate science majors: A model for institutional support of underrepresented students. *CBE Life Sci. Educ.* 14:ar12. doi: 10.1187/cbe.14-05-0082

Tran, P., Ramachandran, A., Khawasik, O., Beisner, B. E., Rautio, M., Huot, Y., et al. (2018). Microbial life under ice: metagenome diversity and in situ activity of Verrucomicrobia in seasonally ice-covered lakes. *Environ. Microbiol.* 20, 2568–2584. doi: 10.1111/1462-2920.14283

Uyaguari-Diaz, M. I., Chan, M., Chaban, B. L., Croxen, M. A., Finke, J. F., Hill, J. E., et al. (2016). A comprehensive method for amplicon-based and metagenomic characterization of viruses, bacteria, and eukaryotes in freshwater samples. *Microbiome* 4:20. doi: 10.1186/s40168-016-0166-1

Van Rossum, T., Peabody, M. A., Uyaguari-Diaz, M. I., Cronin, K. I., Chan, M., Slobodan, J. R., et al. (2015). Year-Long metagenomic study of river microbiomes across land use and water quality. *Front. Microbiol.* 6:1405. doi: 10.3389/fmicb.2015.01405

Vigneron, A., Lovejoy, C., Cruaud, P., Kalenitchenko, D., Culley, A., and Vincent, W. F. (2019). Contrasting winter versus summer microbial communities and metabolic functions in a permafrost thaw lake. *Front. Microbiol.* 10:1656. doi: 10.3389/fmicb.2019.01656

Wang, J. T. H. (2017). Course-based undergraduate research experiences in molecular biosciences - patterns, trends, and faculty support. *FEMS Microbiol. Lett.* 364:fnx157. doi: 10.1093/femsle/fnx157

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Woodin, T., Carter, V. C., and Fletcher, L. (2010). Vision and change in biology undergraduate education, a call for action–initial responses. *CBE Life Sci. Educ.* 9, 71–73. doi: 10.1187/cbe.10-03-0044

Wooley, J. C., Godzik, A., and Friedberg, I. (2010). A primer on metagenomics. *PLoS Comput. Biol.* 6:e1000667. doi: 10.1371/journal.pcbi.1000667

Yannarell, A. C., Kent, A. D., Lauster, G. H., Kratz, T. K., and Triplett, E. W. (2003). Temporal patterns in bacterial communities in three temperate lakes of different trophic status. *Microb. Ecol.* 46, 391–405. doi: 10.1007/s00248-003-1008-9

Yilmaz, F., and Icgen, B. (2014). Characterization of SDS-degrading *Delftia acidovorans* and *in situ* monitoring of its temporal succession in SDS-contaminated surface waters. *Environ. Sci. Pollut. Res. Int.* 21, 7413–7424.

Zwart, G., Crump, B. C., Kamst-van Agterveld, M. P., Hagen, F., and Han, S.-K. (2002). Typical freshwater bacteria: an analysis of available 16S rRNA gene sequences from plankton of lakes and rivers. *Aquat. Microb. Ecol.* 28, 141–155. doi: 10.3354/ame028141

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

METHODS
published: 16 March 2021
doi: 10.3389/fmicb.2021.578859

Resequencing of Microbial Isolates: A Lab Module to Introduce Novices to Command-Line Bioinformatics

Katherine Lynn Petrie[1,2]* and Rujia Xie*[1]

[1] Division of Biological Sciences, University of California, San Diego, La Jolla, CA, United States, [2] Earth-Life Science Institute, Tokyo Institute of Technology, Tokyo, Japan

Familiarity with genome-scale data and the bioinformatic skills to analyze it have become essential for understanding and advancing modern biology and human health, yet many undergraduate biology majors are never exposed to hands-on bioinformatics. This paper presents a module that introduces students to applied bioinformatic analysis within the context of a research-based microbiology lab course. One of the most commonly used genomic analyses in biology is resequencing: determining the sequence of DNA bases in a derived strain of some organism, and comparing it to the known ancestral genome of that organism to better understand the phenotypic differences between them. Many existing CUREs — Course Based Undergraduate Research Experiences — evolve or select new strains of bacteria and compare them phenotypically to ancestral strains. This paper covers standardized strategies and procedures, accessible to undergraduates, for preparing and analyzing microbial whole-genome resequencing data to examine the genotypic differences between such strains. Wet-lab protocols and computational tutorials are provided, along with additional guidelines for educators, providing instructors without a next-generation sequencing or bioinformatics background the necessary information to incorporate whole-genome sequencing and command-line analysis into their class. This module introduces novice students to running software at the command-line, giving them exposure and familiarity with the types of tools that make up the vast majority of open-source scientific software used in contemporary biology. Completion of the module improves student attitudes toward computing, which may make them more likely to pursue further bioinformatics study.

Keywords: CURE, bioinformatics, bioinformatics tutorial, resequencing, breseq, next-generation sequencing

Edited by:
Brad W. Goodner,
Hiram College, United States

Reviewed by:
Sophie Shaw,
University of Aberdeen,
United Kingdom
Ludmila Chistoserdova,
University of Washington,
United States

****Correspondence:***
Katherine Lynn Petrie
kpetrie@ucsd.edu

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *01 July 2020*
Accepted: *16 February 2021*
Published: *16 March 2021*

Citation:
Petrie KL and Xie R (2021) Resequencing of Microbial Isolates: A Lab Module to Introduce Novices to Command-Line Bioinformatics. Front. Microbiol. 12:578859. doi: 10.3389/fmicb.2021.578859

INTRODUCTION

The Need for Bioinformatics in the Undergraduate Biology Curriculum

Bioinformatics is increasingly an important part of research in any biological discipline (Barone et al., 2017), and there is widespread agreement that bioinformatics should be incorporated into the undergraduate biology curriculum (Pevzner and Shamir, 2009; Wilson Sayres et al., 2018). However, barriers to this exist at both the instructor and student level. Instructors report lack of training as the primary barrier to shifting their curricula (Williams et al., 2017), while research has suggested that student anxiety about computing and lack of confidence in their capabilities may act as a barrier to learning computing (Doyle et al., 2005).

The paper presents a way to introduce complete novices to bioinformatics as part of a module in an undergraduate biology laboratory course. This module is not as extensive as a full bioinformatics class, but could be part of an effort to incorporate bioinformatics throughout the curriculum, to reach students who wouldn't otherwise complete any bioinformatics or computer science coursework. The goal of this module is to get undergraduate students engaged with bioinformatics in the context of a broader course, where they can connect the analysis of their data to something tangible they are exploring in another context.

How does this module address a gap in bioinformatics education? The vast majority of bioinformatics software used by researchers to analyze next-generation sequencing data is open-source and run at the command-line. This means that users interact with the software by typing commands into a text-based window (called a terminal), rather than through a point-and-click graphical user interface (GUI). Excellent short workshops to teach this type of command-line bioinformatics to researchers exist (Wilson, 2014; Teal et al., 2015; ANGUS, 2019), but they are primarily aimed at graduate students and researchers beyond the undergraduate level. There are several well-known efforts to introduce undergraduate students to bioinformatics including the Genomic Education Partnership (Elgin et al., 2017) and SEA-PHAGES (Hanauer et al., 2017). These efforts create genuine research opportunities for undergraduate students in classrooms around the world to contribute to scientific understanding and even earn authorship on scientific publications (Leung et al., 2015). However, they focus primarily on aspects of bioinformatics that do not require command-line skills. Students in these programs typically start with an assembled genome sequence that has already been processed from raw data, and they generally use GUI-based software or websites to finish and annotate the sequence (Genomics Education Partnership, 2020; SEA PHAGES, 2020). While finishing and annotation are certainly important components of genome bioinformatics, there is still a need for instruction focused on the command-line skills to needed to work with raw sequence data.

Working at the command-line can be difficult and intimidating for novices, so several GUI-based platforms that simulate command-line bioinformatics pipelines have been developed (Hilgert et al., 2014; Batut et al., 2018). While these can be used to perform real analysis and introduce the underlying concepts, alone, GUI-based platforms cannot fully prepare students to handle working with bioinformatics data the way it is done by most researchers. There are curriculum modules in the literature that focus on quantitative analysis of sequencing data using statistics-focused computing languages like R (i.e., Peterson et al., 2015; Kruchten, 2020). This module complements those modules by focusing on the data processing and analysis steps that would need to be run before (or in lieu of) that type of quantitative statistical analysis. This module aims to a fill a need in bioinformatics curricula by showing students how command-line software tools are used to go from raw sequencing data to interpretable outputs.

What Types of Courses Could Use This Module to Bring Bioinformatics Into the Classroom?

What types of courses would be a good fit for this module? An undergraduate microbiology lab class that includes, or is thinking of including, a CURE would be ideal. CUREs, or Course-Based Undergraduate Research Experiences, incorporate genuine open-ended research of potential relevance to the scientific community (Auchincloss et al., 2014). They have been lauded as a way to answer calls to incorporate more of the skills used in science into the undergraduate curriculum (American Association for the Advancement of Science [AAAS], 2011), and they contribute to making science more inclusive (Bangera and Brownell, 2017). There are many CUREs that have been developed for microbiology labs which select or evolve a novel variant of a known microbe (overviewed in the methods, below). This module would allow students to sequence the genome of that variant and compare it to an ancestor genome that has already been sequenced, an approach called resequencing. The paper combines a guide for the wet-lab preparation of microbial DNA for next-generation resequencing with a guide to the dry-lab analysis of the resulting data.

This module would be ideal in a microbiology lab, or molecular biology lab which uses microbes as a model system. Why are microbes the ideal organism for this module? Although the costs of next-generation sequencing continue to drop, it is still prohibitively expensive and computationally time-consuming to sequence and analyze most eukaryotic genomes. Microbes, on the other hand, have genomes which are generally short enough to facilitate multiplexing – combining multiple samples together so that data for an entire class of student-generated variants can be analyzed on a single sequencing run. Microbial genome datasets are also small enough that analysis of them they can be completed in reasonable time-frames with desktop or laptop computers; they do not require high-performance computing clusters or supercomputer access.

Organization and Goals of This Guide

The methods section contains background information and guidelines for setting up and teaching a resequencing module. The first part of the methods describes how to get from a derived microbial isolate to DNA ready for next-generation sequencing. The second part of the methods introduces the bioinformatics skills needed to computationally analyze next-generation sequencing data.

Neither the preparation of DNA for next-generation sequencing, nor the computational analysis of genomic data are novel methods, however, this article attempts to bring all of the relevant information together in one place in an accessible, easy-to-use format. We have provided a detailed lab manual with bioinformatics tutorials, lecture slides, and lecture notes in the **Supplementary Materials**. Instructors can use the module as-is, or they can use it as a starting point to be adapted to their own particular purposes. Although specific details of sequencing methodology and software may change over time, this article covers several universal considerations that should guide any instructor thinking of incorporating a resequencing and bioinformatics module into their class.

This article is intended as a guide to help course designers and instructors who do not have prior next-generation sequencing or bioinformatics experience bring a resequencing module into their own course. This approach has been vetted in the classroom over several quarters of a microbiology lab course by an instructional team consisting of a lead instructor (the author), two additional instructors, multiple graduate instructional assistants, and laboratory support staff. We show that this module can improve student attitudes toward computing, which could make students more likely to engage and persist in further opportunities to use bioinformatics.

METHODS FOR IMPLEMENTING RESEQUENCING MODULE AS PART OF A COURSE

The following methods provide a general guide for instructors, covering key considerations and pitfalls to avoid for each step of the module. Detailed, step-by-step instructions for students, including protocols adapted from kit manufacturers' instructions, as well as bioinformatics tutorials, are provided in the **Supplementary Materials**. For instructor testing, or for use in a course that is only incorporating the bioinformatics portion of the course (see section "Dry Lab Methods: Analyzing Genomic Re-sequencing Data"), a sample dataset has been provided in the **Supplementary Materials**. For courses incorporating the wet lab methods, the methods assume that each student, or group of students, has isolated a unique microbe of interest that they will characterize. **Figure 1** shows an overview of the individual lab sessions, along with suggested lessons for down time or for lectures between labs. **Figure 2** shows a suggested preparation timeline for instructors planning to add all or part of this module to a class.

Wet Lab Methods: Preparing DNA for Genome Sequencing

One can imagine many possible experiments students could run to generate or isolate novel derivatives of a microbial strain. One popular example is the isolation of antibiotic resistant microbes in laboratory selection experiments. Another is culturing the fast-evolving *Pseudomonas fluorescens* SBW25 strain in static microcosms to study the evolution of biofilm-forming phenotypes (this is, in fact, what we did in our implementation of the class). Details on how to set up those experiments for the classroom are provided elsewhere (Green et al., 2011; Spiers, 2014; Johnson and Lark, 2018; Van den Bergh et al., 2018), so they will not be included here. Once a strain of interest has been generated, the following steps provide an overview for educators of how to prepare DNA from that strain for sequencing.

Generally, it is important that students use good sterile technique, especially when they are working with the bacterial strain itself. Once cultures are grown and DNA is extracted, there are not as many opportunities for exponential amplification of contaminants, but students should still practice clean laboratory technique to avoid cross contamination and prevent the introduction of DNases (Students should wear gloves at all times, always use fresh micropipettor tips for every step of a procedure, and take care not to touch the inner caps or rims of microfuge tubes).

For preparation steps which use kits, the instructional team found it useful to pre-aliquot reagents per pair of students or per group (providing a slightly higher volume than required) to speed up time in the classroom and avoid cross-contamination. The adapted protocols provided in the manual are aimed at students; for any kits utilized, we recommend instructors also read the kit manuals provided by the manufacturer; they contain important details, like storage conditions, not provided here.

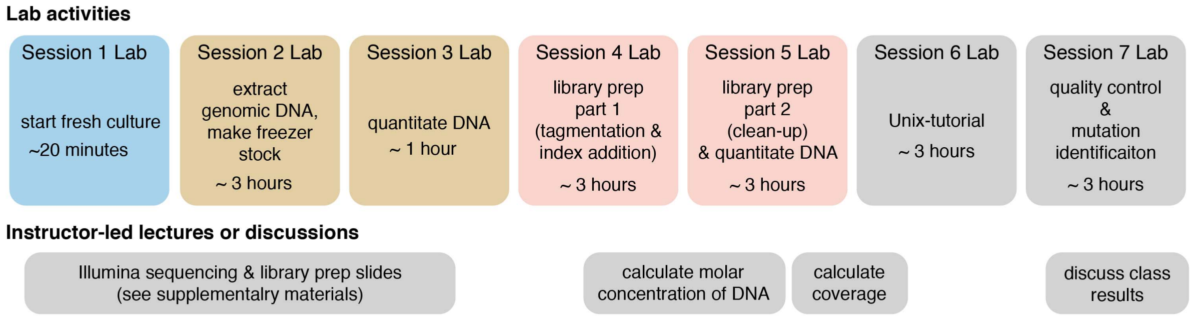

Lab activities
Session 1 Lab
start fresh culture
~20 minutes
Session 2 Lab
extract genomic DNA, make freezer stock
~ 3 hours
Session 3 Lab
quantitate DNA
~ 1 hour
Session 4 Lab
library prep part 1 (tagmentation & index addition)
~ 3 hours
Session 5 Lab
library prep part 2 (clean-up) & quantitate DNA
~ 3 hours
Session 6 Lab
Unix-tutorial
~ 3 hours
Session 7 Lab
quality control & mutation identificaiton
~ 3 hours
Instructor-led lectures or discussions
Illumina sequencing & library prep slides (see supplementalry materials)
calculate molar concentration of DNA
calculate coverage
discuss class results

FIGURE 1 | Overview of each lab session and suggested lessons for downtime or in between lab session. Lab sessions are broken up into 3 h (or less) blocks, though adjacent sessions of identical color could be combined into a single session if timing allows. The lower track shows suggested instructor-led lectures or discussions. Two sets of slides, covering Illumina sequencing and Illumina library preparation, are provided in the **Supplementary Materials**, and they include active-learning questions for students to attempt in-class and for peer-instruction (these can be administered through digital or informal polling methods).

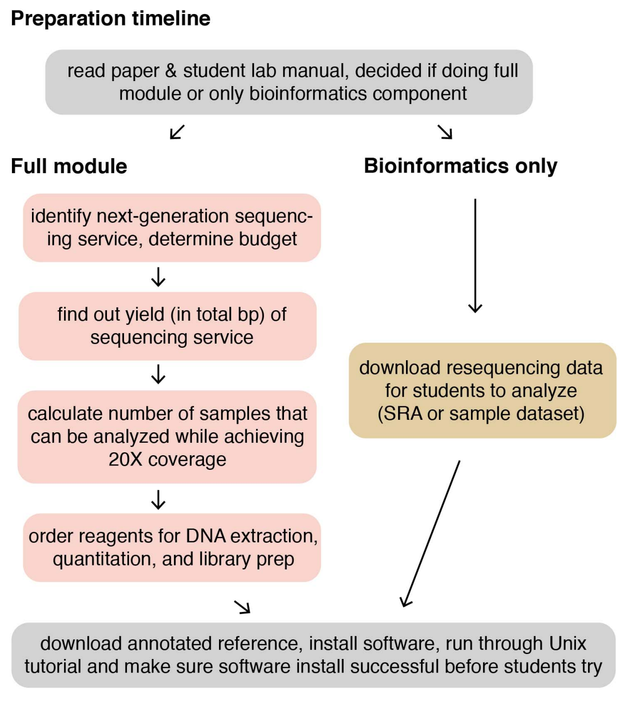

Preparation timeline
read paper & student lab manual, decided if doing full module or only bioinformatics component
Full module
Bioinformatics only
identify next-generation sequencing service, determine budget
find out yield (in total bp) of sequencing service
calculate number of samples that can be analyzed while achieving 20X coverage
order reagents for DNA extraction, quantitation, and library prep
download resequencing data for students to analyze (SRA or sample dataset)
download annotated reference, install software, run through Unix tutorial and make sure software install successful before students try

FIGURE 2 | Suggested preparation timeline for instructors.

Initial Considerations – How Many Samples Can You Sequence?

The biggest consideration for introducing a resequencing module into a laboratory class is how many individual isolates to analyze. One of the benefits of a CURE is that students have the opportunity to make a genuine scientific contribution. The more isolates there are, the more there is to potentially learn, and students working with their own unique microbe may have a greater sense of ownership over their project. However, these goals must be balanced by cost, time, and the amount of data required per strain to get meaningful results. In our implementation, the cost, from start to finish, was approximately $100 per isolate sequenced, though this may vary according to the specific kits and sequencing service used (many institutions have a core facility that provides next-generation sequencing services, there are also several commercial vendors). In terms of time, several of the processing steps are fairly work intensive, so we recommend students work in groups of 2–4 students (with one isolate per group) so they can assist one another.

The amount of data required per strain is a key consideration in determining how many variants can be sequenced. Most sequencing services require users to buy an entire sequencing "run," where all the samples loaded onto the machine are submitted by a single user. In the author's implementation of the module, every pair of students works with their own unique isolate, and 24 isolates were pooled together into an Illumina sequencing run. Illumina machines produce the largest share of contemporary sequence data. It would certainly be possible to carry out resequencing projects with third generation sequencing technologies, like Oxford Nanopore's MinION sequencing or Pacific Bioscience's SMRT sequencing, but a detailed guide for those technologies is beyond the scope of this article. For an excellent review of all sequencing technologies, see Slatko et al. (2018). For our course, samples were sequenced on an Illumina MiSeq instrument.

Why is the MiSeq an appropriate instrument for the job? The answer requires a basic understanding of how Illumina sequencing works (Bentley et al., 2008), briefly reviewed here. To prepare DNA for sequencing, genomes are randomly fragmented into smaller pieces and Illumina-specific adaptor oligos are attached. The ~600 bp-long fragments are then loaded onto a flow cell coated with a lawn of oligos complementary to the adaptors. Fragments are loaded at low concentration so they are well-isolated from one another when they anneal, and then they are clonally amplified in a 2-dimensional PCR-like process to produce DNA clusters. Each cluster contains enough template to make visualization of base-specific dyes possible in a subsequent sequencing-by-synthesis step. All clusters are visualized simultaneously, with images taken after each base is added. The resulting stack of images is then converted into a digital sequence, called a "read," corresponding to each individual cluster. The determining factor for how much data a particular instrument can put out is how many individual clusters can be visualized in a single run on that instrument. The MiSeq instrument can generate ~22 million reads per run (Illumina, 2018).

How do we use that information to figure out how many samples to run? First, we need more information about the read-length — how many bp of the DNA strands in each cluster are actually sequenced. Usually, the entire length is not read; with the most recent reagent kits, 75 or 300 bases are read from each end (Illumina, 2018). This results in a pair of reads for each DNA fragment. Once we know the read-length and the number of reads, it is possible to estimate the coverage — how many times, on average, each position in the genome will be represented in the data. Coverage is calculated as the total number of bases sequenced/the genome size of the organism. For resequencing, 20-fold (or "20X") coverage is a safe bet, as coverage is not always uniformly distributed across the genome, and extra coverage can help distinguish rare sequencing errors from genuine variants. In our implementation of the class, we sequence 24 samples of *P. fluorescens* SBW25 in a single, 75 bp paired-end run; this corresponds to ~20-fold coverage. [(22,000,000 reads × 150 bp per read)/6,722,539 bp in SBW25 genome] = 491-fold coverage/24 samples = 20-fold coverage).

Genomic DNA Extraction – The Fresher, the Better

Students should begin with a clonal isolate of their strain of interest — ideally in the form of a well-isolated colony on an agar plate. For laboratory evolution and selection experiments, the starting, or ancestral, strain should also be processed as described here, even if a reference is available (Strains may accumulate mutations as they are propagated and stored in labs over time, and knowing the precise sequence of what you are starting with is crucial for interpreting mutations. The sample dataset illustrates this point; the SBW25 strain we used to initiate evolution — the ancestor — is slightly different than the public reference available in GenBank).

While it is possible to isolate genomic DNA directly from a colony, it is easier to achieve the high yield and quality of DNA necessary for next-generation sequencing by first preparing a fresh saturated culture (~10^9 cfu/mL, typically from overnight incubation). If lab sessions are scheduled so that students will not be able to come in on contiguous days, inoculated cultures should be held at 4°C and only transferred to an incubator for growth the afternoon or evening before students will return. Over-incubation or prolonged storage in the stationary phase can lead to the accumulation of GASP mutations (Finkel, 2006), which could make interpretation of sequencing results difficult. LB (lysogeny broth) media is recommended for this overnight growth as LB has been widely used to amplify bacteria without downstream issues in next-generation sequencing. Any formulation should work, though we have most recently used LB-Miller (Miller, 1972).

For the actual genomic DNA extraction itself, there are many different commercial kits available. A column-based kit is recommended; in student hands they were both easier and higher-yield than those which rely on phase-separation. Many genomic extraction kits have an "elution buffer" designed for the final step of eluting or resuspending genomic DNA; these should not be used, as some elution buffers, especially those that contain EDTA, can interfere with downstream steps. Instead, MilliQ or molecular grade water should be used for elution. We have had success with the QIAGEN DNeasy Ultra Clean Microbial Kit (Cat No. 12224).

Genomic DNA Quantitation – OD_{260} Is a No-Go

For next-generation sequencing, it is important to measure the concentration of DNA as precisely as possible. To that end, it is recommended that fluorescence-based quantitation methods be used (as opposed to UV absorbance-based methods). In our implementation of the class (as described in the **Supplementary Materials**), we used a Qubit fluorimeter and Qubit ds DNA HS assay kit according to the manufacturers' instructions, but many other dyes/reagent kits are available, and they can be used on any instrument with the appropriate excitation wavelength and emission detection spectrum. One important consideration is sensitivity. The amount of genomic DNA required for library preparation, using the approach described in the next step, is 1 – 500 ng (in 2–30 μL). Given the elution volume, this translates to a minimum DNA concentration of ~33 pg/μL, however, students have had the most success with DNA concentrations at or above 16.7 ng/μL (concentrated samples can always be diluted). It may be useful to have a backup sample of DNA available for students who do not extract the minimum amount.

DNA Library Preparation – Your Students Can Handle It!

There are many library preparation protocols and kits for Illumina sequencing. Regardless of the specific approach used, all protocols break the genomic DNA into smaller fragments and attach oligonucleotide adaptor sequences to the ends of each fragment. This collection of prepared fragments is called a sequencing library. The adaptor sequences help each fragment bind to the flow cell and generate clusters, and they are complementary to primers used in sequencing by synthesis. Library prep can also include a step that adds a unique oligonucleotide — called an index — to every fragment in the sample. This acts as a barcode or a tag, so that when multiple samples are pooled and sequenced together, the output can be computationally sorted by the unique index sequences.

Library preparation is the most difficult part of next-generation sequencing, and even experienced scientists in research labs sometimes elect to outsource library prep as an additional fee paid to the sequencing service provider. However, this is (often prohibitively) costly. The kit used in the our implementation of this module (Illumina DNA Prep, 20018705, from Illumina) was chosen primarily because of its ease of use. Fragmentation and adaptor ligation are carried out in a single step (cleverly called "tagmentation"), and the bead-based purification is somewhat self-normalizing, in that the beads can only bind a certain maximum amount of DNA; so as long as they are saturated, different students should get fairly similar yields.

The **Supplementary Materials** contains detailed directions adapted from the kit's manual. Normally, library prep kits are designed to allow a single researcher to process up to 96 samples at once, using multichannel micropipettors and 96-well plates. Here, they have been rewritten to allow students (or groups of students) to process their samples individually, with standard micropipettors. Below are a few key pointers:

Amount of input DNA: the kit is designed to accommodate 1–500 ng of DNA, added in anywhere from 2–30 μL of liquid volume. As mentioned before, students should add the maximum amount of DNA possible. However, students new to the laboratory may have difficulty calculating what volume to add, so it is useful to have instructors or assistants check student's calculations before proceeding (see the Illumina Library Prep slide deck in the **Supplementary Materials** for examples of this calculation). If less than 50 ng was added, students will need extra amplification steps, so they should pay attention to the note in step 21 of the session four protocol.

Magnetic bead-based purification: most kits rely on ferrous microbeads that bind the DNA. When tubes are placed in magnetic racks, the beads are immobilized while solutions are exchanged. Magnetic stands typically use strong, rare-earth magnetics to speed up the separation process. We have had success with eight-well magnetic stands shared among groups with four students each. We use stands which orient the magnet on the side of the tube (rather than at the bottom or in a ring), as this allows students to rest the pipette tip against the opposite wall without disrupting the beads. The particular stands we use are not currently available from the supplier, but there are many different commercial sources and DIY plans for constructing your own (Oberacker et al., 2019).

Index addition: Typically, samples will get two unique indexes: one for each end of the fragments. We have successfully used the Nextera CD indexes (Illumina 20018708). Indices are typically supplied in trays or in a limited set of tubes, which are difficult

TABLE 1 | The five components of student attitude toward computing (For a complete list of the items in each factor, see Dorn and Tew, 2015).

	Factor	Description	Sample item from CAS (expert consensus)
1	Problem Solving – Transfer	Ability to see/apply connection between concepts and ideas to solve problems	Errors generated by computers are random, and when they happen there's not much I can do to understand why (disagree)
2	Problem Solving – Strategies	Attitude toward problem-solving strategies in computer science	When I solve a computer science problem, I break it into smaller parts and solve them one at a time (agree)
3	Problem Solving – Growth Mindset	Belief in ability to improve skill or understanding with practice	If I get stuck on a computer science problem, there is no chance I'll figure it out on my own (disagree)
4	Real-World Connections	Belief in real-world relevance of computer science discipline	Tools and techniques from computer science can be useful in the study of other disciplines (e.g., biology, art, business) (agree)
5	Personal Interest and Enjoyment	Personal interest, motivation, and engagement with computer science	I am interested in learning more about computer science (agree)

to share with students. We have students bring samples to the instructor or an instructional assistant to receive their unique indices one sample at a time. This prevents cross contamination and allows the instructor to record which samples get which indices, which is useful if students misplace this information. To attach the index oligos, the number of PCR cycles needed varies depending on the amount of DNA originally used as input; it is important to make sure students use the correct number.

Stopping points: the complete library preparation process is fairly time consuming, however, it can be broken into two lab sessions, with the DNA stored at 4°C after the index addition and amplification step. We have stored DNA at this stopping point for up to 5 days with no problems, however, there are no other recommended stopping points during library preparation.

Figuring Out the Molar Concentration

Although the Illumina DNA Prep kit is designed to normalize the yield of library DNA, when carried out by many different student groups, we tend to see a fairly wide range in the library yields. So the library DNA should be quantified using a fluorescent-dye based method, as described above. Additionally, it is typically recommended that the average fragment size of the library be measured with either a TapeStation (Agilent) or BioAnalyzer (Agilent) instrument. This is because the tagmentation may not always produce fragments of the exact same size. However, if you do not have access to one of these instruments, it is acceptable to use the average expected fragment size, which for the Illumina DNA Prep kit is 600 bp. The average fragment length is used to calculate the molar concentration of DNA, using an average atomic mass of 660 g/mol for one basepair. The molar concentration will be used in the next step.

DNA Pooling for Sample Submission – It's All About Balance

To take advantage of small microbial genome sizes and maximize data yield, samples are multiplexed: pooled together and run on the same machine. To make sure that each sample is equally represented in the sequencing data, it is important that an equal number of DNA fragments are added from each sample. This will require students to dilute their DNA to a universal concentration before their sample is added to the pool (Alternatively, different volumes of each sample can be added to achieve the same final concentration). The sequencing service provider will specify the required total concentration of DNA in the pool; it is typically at least 10 nM. Because some students may have lower than expected amounts of DNA, we recommend instructors be the ones to collect the final concentration of each library from students and calculate how the DNA should be pooled. It may be necessary to add a little less of some high-concentration libraries to "make room" for low-concentration libraries, and some very low concentration libraries may have to be dropped altogether, if they fall too far below the threshold required by the sequencing center. Once a pooling scheme has been established, we recommend that students bring their samples to the instructor or an assistant to be added to the pool one-at-a time. This prevents cross contamination and lets the instructor "check off" each sample as it is added. If you are exceptionally lucky, your sequencing service provider may offer to pool your samples for you, but if they do, you should verify whether they will account for different sample concentrations, or else the data may be dominated by the highest-concentration libraries.

Dry Lab Methods: Analyzing Genomic Re-sequencing Data

It may take several weeks to get data back from your sequencing service provider, so it is important to start the wet lab portion of this module early in the course. Most sequencing providers will demultiplex the data for you (separate it into individual files according to the unique index barcodes for each sample). Typically, the size of sequencing data files is large enough that sharing via email or an LMS may be problematic, so the service provider may provide an ftp link that could be shared with students, or data can be distributed from a central source using USB storage devices.

To provide a universal computing environment for all students, ideally analysis would be carried out in a computer lab with software preinstalled. While it is possible to set up virtual machines that can be downloaded or accessed through cloud computing so that students can use their own devices, instructions on how to do so are beyond the scope of this article. A Unix-based operating system (OS) is required, as most open-source bioinformatics software cannot be run directly on PCs. This means that you can use Unix OS, Linux OS, or Mac OS. In principle, you can use a simulated Unix environment on a PC through the use of an interface like Cygwin[1], though this will be more challenging. If you plan to have students use their own machines, we recommend setting aside at least an entire lab session to help students configure them, and we recommend skipping the "Quality Control" section below.

Working at the Command-Line – A Guide for the Complete Beginner

The **Supplementary Materials** include a brief tutorial that introduces students to the basic commands used to navigate through directories (folders) in Unix-based terminals. It is essential that students try it out on their own (rather than follow along with the projector while an instructor demonstrates), as engaging with the activity and seeing for themselves what actually happens is key to understanding some essential rules about working in a command-line terminal.

Many instructors or teaching assistants may be apprehensive about teaching bioinformatics if they are themselves new to working at the command-line. However, trying the tutorial ahead of time and seeing what common mistakes occur is sufficient preparation for most of the problems students might encounter. The vast majority of errors stem from typos or from commands that try to use a file not located in the current working directory. A quick check of the command that students entered and a look at the file contents of their current location reveals most problems. We have provided a troubleshooting guide expanding on this and other common problems in the **Supplementary Materials**. More complicated issues can usually be solved by reading the error messages, and occasionally using a search engine to find out what they mean. Additional information can be found through discussion forums focused on computing (Stack Overflow, 2020), bioinformatics in general (Biostars, 2020), next-generation sequencing (SEQanswers, 2020), and on the support pages for individual software tools. Even experienced bioinformatics researchers have to troubleshoot software, so it is good to adopt a collaborative outlook to helping students solve problems, encouraging them to be resourceful and not get discouraged if things don't work out the first time.

Installing Software – Use a Package Manager if Possible!

Software installation is probably the most difficult part of next-generation sequence analysis. Many open source software programs are not self-contained; they require other, previously developed software programs to function. This is the nature of high-throughput sequencing analysis – newer, more specialized programs build on earlier algorithms and data processing tools. The software tools required by a particular program are called dependencies, and up until about a few years ago, there was little else to do but install each dependency – and the dependencies of that dependency – one at a time. One would have to hope that all of the versions of each piece of software were compatible with one another, and if not, just keep troubleshooting until it all worked. Software developers and users refer to this problem as "dependency hell," and it is particularly vexing in bioinformatics, since software tools have been developed by independent research teams over nearly a decade and a half of next-generation sequencing history.

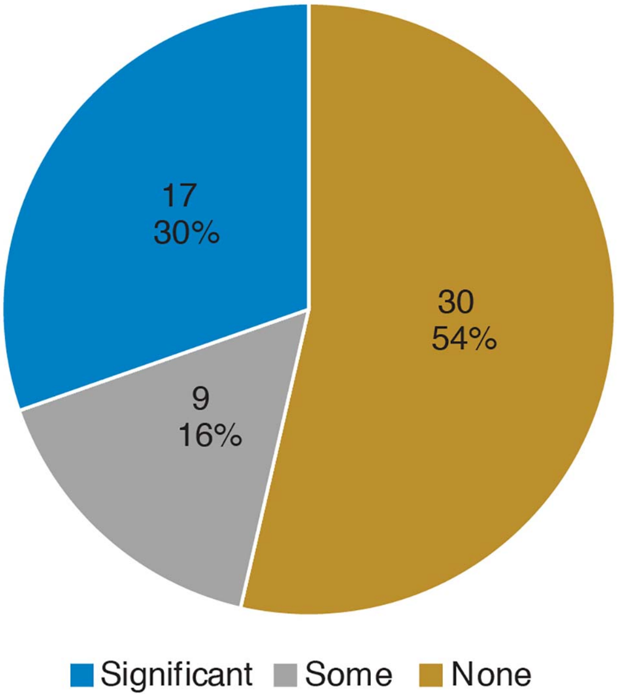


FIGURE 3 | Prior computing experience of students. In the pre-survey, students were given free-response space to respond to the following prompt: "*Please describe any experience you have in computer science. Write "none" if you don't have any experience. Some examples of computer science are: using a computer to analyze data, taking a CSE course, programming, writing software, writing computer scripts, analyzing data in R, debugging, running computer programs via command-line interfaces (where you type directions to your computer instead of using a mouse to point and click).*" Student responses were coded according to the following scheme: "None" if the student wrote none or indicated no experience, "Some" if the student described MATLAB or limited use of command-line (i.e., in personal time), and "Significant" if the student mentioned a bioinformatics or computer science course, or if they mentioned "knowing" a computer language. The number and percentage of students in each category are shown.

So, how do we make this easy and accessible? Fortunately, tools called package managers have been developed to make software installation easier. With a single command, they can install the desired software program and all of its dependencies automatically, and package managers can be used to create "environments" – workspaces for individual projects with defined collections of software. In recent years, the bioinformatics community has assembled bioconda, a collection of packages (software, dependencies, and directions that tell the computer how to install them) for over 7,000 bioinformatics software tools

[1] https://www.cygwin.com/

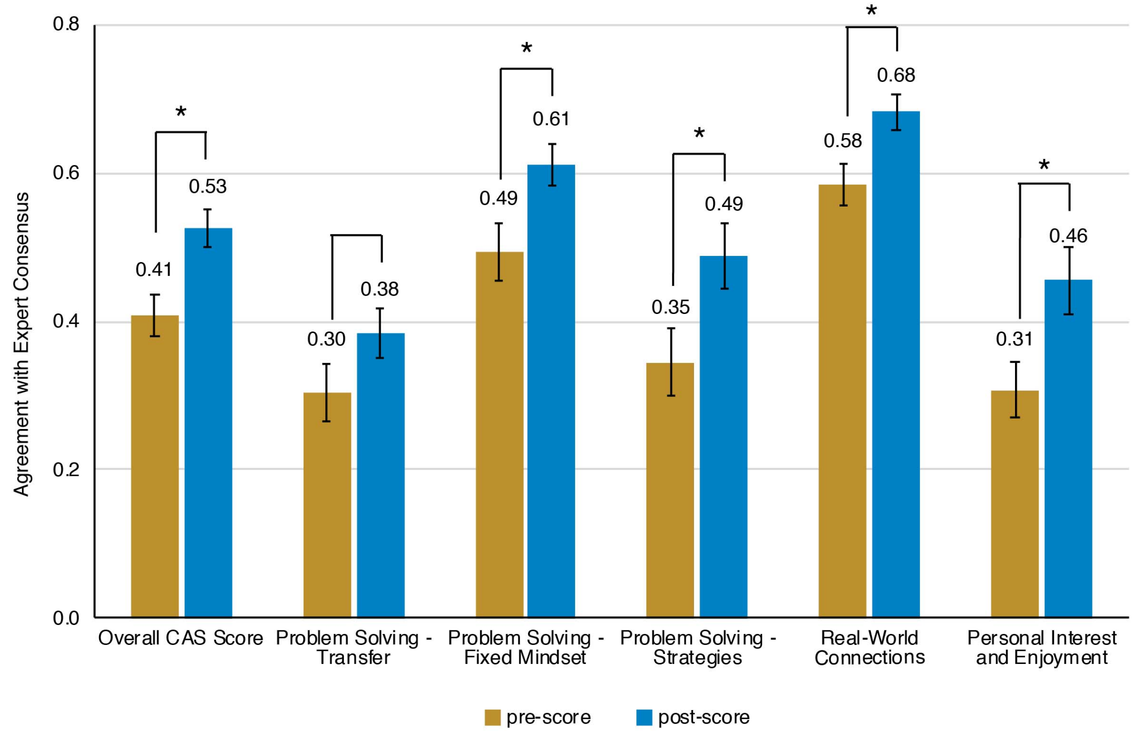


FIGURE 4 | Scores on Computing Attitude Survey before and after completing bioinformatics module. Mean pre- and post- scores are shown. Bars indicate standard error. Significant differences between pre- and post- scores (Wilcoxon Signed Rank, n = 56) are indicated with a star.

(Grüning et al., 2018). The packages in bioconda can be installed with the popular package manager, conda[2] or miniconda, a lightweight version of conda.

All of the bioinformatics software used in this module (fastqc, fastx_toolkit, and breseq) can be installed through bioconda using the conda or miniconda package manager. To use bioconda, carefully follow all the directions in the "Getting Started" section of the bioconda user documents found at[3] (For additional information, see the notes accompanying the sample dataset). It is also possible to install the software without a package manager, by following the individual installation instructions for each individual software tool (see citations below for links to the user support). Finally, a streamlined version of this module can be completed with just breseq, to minimize the software requirements, though the breseq installation instructions must be followed carefully to make sure all dependencies are also installed. If possible, we recommend that you work with your institution's technology support staff to facilitate software installation, especially if you will be installing software in a computer lab (most institution-managed computers do not grant regular users permission to install software by default). Note: if students will be running this module on PCs (i.e., via Cygwin), it may be easiest to skip the quality control steps and install breseq directly according to the instructions in the breseq user documentation.

Examining Data – An Introduction to the FASTQ Format

Illumina sequencing outputs data files in the FASTQ format. FASTQ files contain information on all of the "reads" corresponding to that sample. A "read" is the information derived from an individual genome fragment, and contains the sequence of bases, as well as a quality score for each base call.

The quality score, Q, estimates the probability that the base is incorrect (P, probability of error), according to the formula $Q = -10 \log_{10} (P)$ (Ewing and Green, 1998; Cock et al., 2010). This conversion takes potentially long character strings (i.e., a high-quality base call like P = 0.0001, or 99.99% accuracy) and reduces them to one or two digits (i.e., Q = 40). To compress the quality score even further, in the FASTQ file, Q is reported as single ASCIII keyboard character (ASCII characters are numbered, for example, the letter "I" is ASCII character 73). To get Q, you subtract 33 from the ASCII value, however, older Illumina data (only a concern if you are using previously collected data from several years ago) had an offset of 64 (Cock et al., 2010). The tutorial leads students through an exploration of the FASTQ format and how to interpret the "two-layer" code of compressed quality scores.

Cleaning Up Data – Optional Here, but Good Practice for Students

All sequencers occasionally produce low quality base-calls, and in many bioinformatics applications, it is important to filter

[2] https://docs.conda.io/

[3] https://bioconda.github.io/

TABLE 2 | Pre- and post- CAS scores.

Measure	Pre-score		Post-score		Shift	Wilcoxon signed-rank	
	Mean	SD	Mean	SD		*W*	*p*
Overall CAS Score	0.409	0.212	0.526	0.183	0.117	195	<0.001*
Problem Solving – Transfer	0.304	0.289	0.384	0.257	0.080	254	0.026
Problem Solving – Growth Mindset	0.493	0.288	0.612	0.213	0.119	179	0.003*
Problem Solving – Strategies	0.345	0.339	0.488	0.332	0.143	106	<0.001*
Real-World Connections	0.585	0.220	0.683	0.181	0.098	106	0.007*
Personal interest and Enjoyment	0.308	0.274	0.455	0.341	0.147	150	0.002*

*The shift of each score was calculated as the post-score minus the pre-score. The average and standard deviation for all students are shown. To see if there was a significant difference between the pre and post scores, a Wilcoxon signed-rank (paired, non-parametric, n = 56) test was used. Significant differences are starred with * (for the overall CAS score, α = 0.05, for the individual factors, a Bonferroni correction was applied; α = 0.01).*

TABLE 3 | Gain in overall CAS score by prior level of computing experience.

Prior computing experience	*N*	Mean shift	SD
None	30	0.169	0.193
Some	9	0.0756	0.0581
Significant	17	0.0471	0.143

The mean and standard deviation are shown.

these out. Here, the tutorial guides students in the use of two software tools, FastQC (FastQC, 2015), which produces statistics on the quality score distribution of a FASTQ file, and the FASTX-Toolkit (Hannon, 2010), which can be used to remove reads where a specified proportion of the bases fall below a specified quality score. For this module, this filter is not strictly necessary, as the next step of analysis actually takes quality score into account, so it can be safely skipped if time is a limiting factor. However, it is good bioinformatics practice to examine the quality of the data, and removing low-quality reads can make subsequent steps of the analysis run faster.

The FASTX-Toolkit filter can only be run on one file at a time. This means students with paired-end sequencing data must run it twice, once on the forward reads file and once on the reverse reads file. Because the different files may have a different number of reads passing the filter, the filtered files may be different sizes. This is not a problem for breseq, as it treats the forward and reverse reads as if they were two independent lists of single-end data. However, we would be remiss not to mention that other bioinformatics software make use of pair linkage information (the fact that forward and reverse reads are ∼600 bp apart) to guide analysis, and in other applications, it is critical that every read in the forward file has its corresponding pair in the reverse file. If you are considering other analyses, you may need to use a filter designed to work with paired-end data, such as sickle (Joshi and Fass, 2011).

Running Breseq to Identify Mutations – The Software That Does It All!

In order to identify mutations in the sequenced strains, the reads need to be compared to an existing reference sequence (of the ancestor or a closely related strain). First, the reads are mapped to the reference (each read is scanned against the genome to see where it belongs), and then each position is examined to see if the majority of the reads there have the same base at that position as the reference does. There is a huge variety of software tools capable of performing these steps (alignment and variant identification), but this module uses a tool called breseq (short for bacterial resequencing) (Deatherage and Barrick, 2014). A detailed guide to all of breseq's capabilities is available elsewhere (Deatherage and Barrick, 2014); here, we cover a few important pointers.

Breseq is ideal for students new to bioinformatics, as it outputs results as easy-to-navigate html files that can be opened in a web-browser. A key feature of breseq is that it can report not just the genomic position and identity of any mutations, but also whether a mutation is synonymous or non-synonymous as well as the name of the gene it is in or near. To get this information, you must use an annotated reference in a gff3 or GenBank (.gbk) format, which includes the location and name of genes. Annotated references for many microbial species can be obtained from the NCBI. For our classes, we provide the reference file to students to ensure that they are all using the same one. To find a reference sequence at NCBI[4], restrict the search to the "Genomes" database, and type your species of interest in the search bar. This will display the landing page for your species, and you can click a link to browse all available genomes for the species. Locate your strain (or a close relative), click the link in the "strain" or "organism name" column, and you'll be taken to its genome assembly and annotation report. From there, you can click "download genome annotation in GenBank format." You can also get to the GenBank record by clicking on the RefSeq ID, though to ensure that you download the full file, you will have to set the "customize view" to show all features before you download it with the "Send to" link.

Running breseq will take a considerable amount of time, as aligning millions of short reads to a genome that is millions of bp long is not a trivial task. The time required depends on the genome size, the amount of data, and the computer itself. We have found that on a typical desktop or laptop computer, it takes

[4] https://www.ncbi.nlm.nih.gov

TABLE 4 | No significant difference in overall CAS pre-score between different demographic groups (Mann–Whitney *U*-Test, non-parametric, independent, *n* = 56).

					Mann–Whitney *U*	
		N	**Pre-score**	**SD**	***U***	***p***
Applicant type	New-Freshman	43	0.422	0.198	227	0.312
	Transfer	13	0.366	0.259		
First-Generation Status	Non-First-Generation	38	0.397	0.207	324	0.751
	First-Generation	18	0.436	0.227		
Binary Gender	Female	42	0.386	0.204	231	0.236
	Male	14	0.480	0.229		

TABLE 5 | No significant difference in overall CAS score improvement between different demographic groups (Mann–Whitney *U*-Test, non-parametric, independent, *n* = 56).

					Mann–Whitney *U*	
		N	**Mean shift**	**SD**	***U***	***p***
Applicant type	New-Freshman	43	0.1060	0.156	239	0.431
	Transfer	13	0.1538	0.217		
First-Generation status	Non-First-generation	38	0.1147	0.133	318	0.679
	First-Generation	18	0.1222	0.238		
Binary gender	Female	42	0.1190	0.176	285	0.864
	Male	14	0.1114	0.163		

The mean shift is the average difference in pre- and post-score.

10–20 min for breseq to analyze an ∼7 million bp genome with ∼20-fold coverage.

RESULTS

Implementation of Module

This module was incorporated into a Microbiology Laboratory course taken primarily by juniors and seniors. It has since been taught by three different instructors (including the author) to ∼450 students (in person). In Spring 2020, the class was held remotely for ∼100 students, and we implemented only the bioinformatics analysis, relying on data generated by previous classes. Instructors wishing to run only the bioinformatics portion of the module can use the sample dataset provided (see **Supplementary Materials**), or browse publicly available resequencing data in the NCBI's Sequence Read Archive (Sequence Read Archive Submissions Staff, 2011).

In every offering of the course, variants have been successfully analyzed. Identifying mutations is only the first step of the analysis; the bigger challenge for students lies in interpreting them. Students have to predict which mutations are responsible for the observed phenotype of their variant, and which mutations are neutral, acquired by random chance. This requires students to dive into the literature to learn more about the genes or regulatory regions where they find mutations. Students can also see if similar mutations have been previously observed. In our implementation of the class, we have students write up their findings in a lab report, though other forms of assessment are possible.

Assessment of Bioinformatics Module's Impact on Student Attitudes Toward Computing

For a subset of classes in which this module was offered, we carried out a focused assessment of the bioinformatics portion of the module. Before and after the computing module, we administered a validated instrument, the Computing Attitudes Survey (CAS), designed to measure student attitudes toward learning the practices and skills of computing (Dorn and Tew, 2015). Why focus on student attitudes? After a brief introduction like the one in this module, students will likely need further practice to really master bioinformatics content. However, if the module can positively impact their attitude toward computing, they may be more likely to persist in future opportunities to learn and use bioinformatics.

The CAS is a 26-item Likert scale that assesses the beliefs people have about the process of computing and learning computational skills (Dorn and Tew, 2015). Within the scale, items are divided into subscales, called factors, that relate to different components of student attitude. The scale includes three factors connected to problem solving: belief that concepts and ideas can transfer to new problems, attitude toward problem solving strategies, and adoption of a growth mindset (the idea that skills and understanding are not fixed and can be improved with practice). Another factor relates to belief in the real-world relevance of computer science, and the final factor assess personal interest in and enjoyment of computer science (see **Table 1** for a detailed description of the five factors and sample items). Each item in the CAS has a "correct," or expert-like rating, based on the consensus opinion (agreement or disagreement) for each item when administered to a group of computing faculty as described in Dorn and Tew, 2015. Students are scored based on their level of agreement with the expert consensus, providing a measure of how students may shift from holding more novice-like attitudes toward more expert-like attitudes (Dorn and Tew, 2015).

We administered the CAS immediately before and immediately after the bioinformatics ("dry-lab") portion of the module, so that each student has a pre- and a post- score. We also asked students to describe their previous experience in computing. Students' demographic factors, including applicant type, first-generation student status, and gender, were added and student responses were deidentified. All responses were collected with approval of the UC San Diego Institutional Review Board. Only students who completed both the pre- and post-survey once and on time were included. Students who did not respond to more than five items were removed. Students who did not correctly respond to the control statement ("We use this statement to discard the surveys of people who are not reading the questions. Please select "Agree" for this question to preserve your answers) were removed, and this item was not used in subsequent analysis.

Student responses were scored according to the method described by the survey's developer (Dorn and Tew, 2015). For each item, students selected "Strongly Disagree," "Disagree," "Neutral," "Agree," or "Strongly Agree." The responses were first collapsed into a 3-point scale by replacing "strongly agree" by "agree" and replacing "strongly disagree" by "disagree," then scored based on their agreement with expert opinion. Each item receive a "1" if the student agreed with the expert opinion, and a "0" if their response was "Neutral" or they disagreed with expert opinion. The score for each student was calculated as the average of their responses to all items (to get an overall score) or only those from the relevant subscale for each factor. A score of 1 represents student agreement with expert opinion on all items, and a score of 0 represents disagreement with expert opinion on all items.

Prior to the bioinformatics module, over half the students surveyed did not have any experience with computing (**Figure 3**). When looking at all students, there was a significant improvement in overall computing attitude scores after completion of the bioinformatics module (**Figure 4** and **Table 2**, Wilcoxon signed rank test, $p < 0.001$, $n = 56$), suggesting that even this short module can improve student attitudes toward computing. Looking at all items in the survey, students went from an average of 41% agreement with expert opinion to 53% agreement with expert opinion. A significant improvement was seen in four out of the five factors, with "Problem Solving – Transfer" as the only factor with no significant improvement. Shifts in the overall CAS score were greatest for the students with no prior computing experience, though students at all levels of experience showed a gain (**Table 3**).

Computer science remains one of the STEM majors with the biggest gender gap [only about 20% of CS majors are female (Sax et al., 2016)]. Studies have attributed this gap to differences in attitudes (Dorn and Tew, 2015; Sax et al., 2016). Here, we sought to explore whether there were differences in computing attitudes between male and female students in the context of a biology course. We also explored whether there were differences between first-generation college students and students with at least 1 parent with a four-year degree, and between students who enrolled directly as new freshman and students who transferred from other (typically community) colleges. In contrast to previous studies, we did not see a significant difference in pre-scores on the CAS between any of the demographic groups (**Table 4**). We also did not see any significant differences between demographic groups in how much the CAS scores improved after the module (**Table 5**). Possible explanations for the difference between our observations and previous work are explored in the discussion.

All statistics were computed in jamovi (The jamovi project, 2020).

DISCUSSION

This methods paper serves as a guide for instructors who are thinking of adding a next-generation resequencing project into their courses. We hope short modules like this can act as a bridge for novice students with no prior command-line experience. Though only a small amount of specific knowledge of particular software programs is covered, the familiarity with command-line that students develop, and the positive impact of the module on student attitudes toward computing, may serve as a bridge to future learning.

This type of exposure may be particularly important for students from populations that are underrepresented in computing fields. Studies have shown that key barriers for participation are student attitudes toward computing, including their confidence about their computing ability and their perception of belonging in computer science (Cheryan et al., 2009; Dorn and Tew, 2015; Sax et al., 2016). Completion of the bioinformatics portion of the module improved student attitudes, but there was no difference in the magnitude of this shift among the different demographic groups we analyzed, nor was there a difference in incoming attitudes as measured by the pre-score alone. There are two possible explanations for this.

First, the CAS, which focuses primarily on attitudes toward the practice of computing itself, may not capture attitudes about belonging and identity as someone who does computer science, and these factors may be the ones that better explain demographic differences in attitudes. In future implementations of the course we plan to include assessments that measure these other components of student attitude.

Second, it may be that negative attitudes are more strongly held in environments in which students are the minority group. In contrast to computer science, where there is a strong gender imbalance, biology majors typically have greater gender equity in their cohorts. At the institution where we collected data, only 17% of computer science majors were women; by comparison, 60% of biology majors were women (institutional research, 2017/2018 school year), and in the student responses we analyzed, 75% of the students were women. This may create a more welcoming environment for female students, and suggests that teaching computing in the context of biology may be a way to better reach underrepresented students.

In the future, we hope to assess the impact of this module on other student outcomes, including content knowledge and understanding in bioinformatics, as well as potential gains in other related areas, like microbiology and evolution. Additionally, we plan to explore how this module, or any introductory bioinformatics module, could be improved in ways that lead to an even greater shift in student attitudes toward computing. This module incorporates a tutorial to walk students through the mechanics of command-line work, but there could be other potential learning activities or self-reflections focused on students' self-efficacy and capacity for growth that might improve outcomes, especially for populations underrepresented in STEM and bioinformatics.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the **Supplementary Material** and via figshare, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by UC San Diego Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

KP developed this course module, wrote instructional materials for it, collected assessment data and provided guidance on its analysis and wrote the manuscript. RX analyzed the assessment data and wrote the manuscript. Both authors contributed to the article and approved the submitted version.

FUNDING

This work was possible, in part, due to the UC San Diego Division of Biological Sciences Teaching Professor Summer Research Fellowship, awarded in 2018, 2019, and 2020.

ACKNOWLEDGMENTS

The authors wish to thank the UC San Diego instructional lab team, including staff, instructional assistants, and instructors, with special thanks to instructors Dr. Giorgia Pirino and Dr. Aubrie O'Rourke for teaching this module in their classes, instructional assistant Daniel Tyler for his help in piloting the class, and instructional lab support staff Ana Gomez and Kritin Karkare for preparing equipment and reagents and running the sequencer. We also thank the UC San Diego Teaching and Learning Commons for removing any students who opted out of the research, adding demographic information, and deidentifying the student response data. Finally, we are grateful to Illumina for providing the equipment and reagents used for this teaching activity.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2021.578859/full#supplementary-material

There several supplementary files accompanying this manuscript, listed below.

1. Resequencing Manual. This manual for students contains step-by-step directions for both the wet and dry lab components of this module, a detailed list of materials, and background information.
2. Illumina Sequencing slides. Lecture slides with active learning activities that can be incorporated into lectures that prepare students for this laboratory module. All images in the slides were generated by the author.
3. Illumina Library Prep slides. Lecture slides with active learning activities that can be incorporated into lectures that prepare students for this laboratory module. All images in the slides were generated by the author.
4. Troubleshooting Guide.
5. Sample Dataset Overview & Expected Results.
6. Sample Dataset, available via figshare DOI: 10.6084/m9.figshare.13426889.

REFERENCES

American Association for the Advancement of Science [AAAS] (2011). *Vision and Change in Undergraduate Biology Education*. Washington, DC: American 939 Association for the Advancement of Science.

ANGUS (2019). *Analyzing Next-Generation Sequencing Data Workshop*. California: ANGUS.

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., et al. (2014). Assessment of course-based undergraduate research experiences: a meeting report. *CBE Life Sci. Educ.* 13, 29–40. doi: 10.1187/cbe.14-01-0004

Bangera, G., and Brownell, S. E. (2017). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE Life Sci. Educ.* 13, 602–606. doi: 10.1187/cbe.14-06-0099

Barone, L., Williams, J., and Micklos, D. (2017). Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators. *PLoS Computat. Biol.* 13:e1005755. doi: 10.1371/journal.pcbi.1005755

Batut, B., Hiltemann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Clemens, B., et al. (2018). Community-Driven Data Analysis Training for Biology. *Cell Syst.* 6, 752–758. doi: 10.1016/j.cels.2018.05.012

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 456, 53–59. doi: 10.1038/nature07517

Biostars (2020). *Biostars*. Available online at: https://www.biostars.org (accessed September 10, 2020).

Cheryan, S., Plaut, V. C., Davies, P. G., and Steele, C. M. (2009). Ambient belonging: How stereotypical cues impact gender participation in computer science. *J. Personal. Soc. Psychol.* 97, 1045–1060. doi: 10.1037/a0016239

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. (2010). Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 38, 1767–1771. doi: 10.1093/nar/gkp1137

Deatherage, D. E., and Barrick, J. E. (2014). Identification of mutations in laboratory-evolved microbes from next-generation sequencing data using breseq. *Methods Mol. Biol.* 1151, 165–188. doi: 10.1007/978-1-4939-0554-6_12

Dorn, B., and Tew, A. E. (2015). Empirical validation and application of the computing attitudes survey. *Comput. Sci. Educat.* 25, 1–36. doi: 10.1080/08993408.2015.1014142

Doyle, E., Stamouli, I., and Huggard, M. (2005). Computer anxiety, self-efficacy, computer experience: an investigation throughout a computer science degree. *Proc. Front. Educat. 35th Annu. Confer.* 1, 2H–3H. doi: 10.1109/FIE.2005.1612246

Elgin, S. C. R., Hauser, C., Holzen, T. M., Jones, C., Kleinschmit, A., and Leatherman, J. (2017). The GEP: Crowd-Sourcing Big Data Analysis with Undergraduates. *Trends Genet.* 33, 81–85. doi: 10.1016/j.tig.2016.11.004

Ewing, B., and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res.* 8, 186–194. doi: 10.1101/gr.8.3.186

FastQC (2015). *A Quality Control Tool for High Throughput Sequence Data*. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed June 30, 2020).

Finkel, S. E. (2006). Long-term survival during stationary phase: evolution and the GASP phenotype. *Nat. Rev. Microbiol.* 4, 113–120. doi: 10.1038/nrmicro1340

Genomics Education Partnership (2020). *Genomics Education Partnership*. Available online at: https://gep.wustl.edu (accessed September 9, 2020).

Green, J. H., Koza, A., Moshynets, O., Pajor, R., Ritchie, M. R., and Spiers, A. J. (2011). Evolution in a test tube: rise of the Wrinkly Spreaders. *J. Biol. Educat.* 45, 54–59. doi: 10.1080/00219266.2011.537842

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., et al. (2018). Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nat. Methods* 15, 475–476. doi: 10.1038/s41592-018-0046-7

Hanauer, D. I., Graham, M. J., Sea-Phages, Betancur, L., Bobrownicki, A., Cresawn, S. G., et al. (2017). An inclusive Research Education Community (iREC): Impact of the SEA-PHAGES program on research outcomes and student learning. *PNAS* 114, 13531–13536. doi: 10.1073/pnas.1718188115

Hannon, G. J. (2010). *FASTX-Toolkit*. Available online at: http://hannonlab.cshl.edu/fastx_toolkit (accessed June 30, 2020).

Hilgert, U., McKay, S., Khalfan, M., Williams, J., Ghiban, C., and Micklos, D. (2014). DNA Subway: Making Genome Analysis Egalitarian. *XSEDE 14 Proc.* 70, 1–3. doi: 10.1145/2616498.2616575/

Illumina (2018). *MiSeq System: Datasheet [Specification Sheet]*. San Diego, CA: Illumina.

Johnson, W. R., and Lark, A. (2018). Evolution in Action in the Classroom: Engaging Students in Science Practices to Investigate and Explain Evolution by Natural Selection. *Am. Biol. Teacher* 80, 92–99. doi: 10.1525/abt.2018.80.2.92

Joshi, N. A., and Fass, J. N. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33)*. San Francisco: github.

Kruchten, A. E. (2020). A Curricular Bioinformatics Approach to Teaching Undergraduates to Analyze Metagenomic Datasets Using R. *Front. Microbiol.* 11:2135. doi: 10.3389/fmicb.2020.578600

Leung, W., Shaffer, C. D., Reed, L. K., Smith, S. T., Barshop, W., Dirkes, W., et al. (2015). *Drosophila* Muller F Elements Maintain a Distinct Set of Genomic Properties Over 40 Million Years of Evolution. *G3 Genes Genomes Genet.* 5, 719–740. doi: 10.1534/g3.114.015966

Miller, J. H. (1972). *Experiments in molecular genetics*. New York, NY: Cold Spring Harbor Laboratory.

Oberacker, P., Stepper, P., Bond, D. M., Höhn, S., Focken, J., Meyer, V., et al. (2019). Bio-On-Magnetic-Beads (BOMB): Open platform for high-throughput nucleic acid extraction and manipulation. *PLoS Biol.* 17:e3000107. doi: 10.1371/journal.pbio.3000107

Peterson, M. P., Malloy, J. T., Buonaccorsi, V. P., and Marden, J. H. (2015). *Teaching RNAseq at Undergraduate Institutions: A tutorial and R package from the Genome Consortium for Active Teaching*. vienna: R Core Team.

Pevzner, P., and Shamir, R. (2009). Computing Has Changed Biology—Biology Education Must Catch Up. *Science* 325, 541–542. doi: 10.1126/science.1173876

Sax, L. J., Allison Kanny, M., Jacobs, J.A., Whang, H., Weintraub, D. S., and Hroch, A. (2016). Understanding the Changing Dynamics of the Gender Gap in Undergraduate Engineering Majors: 1971–2011. *Res. High Educ.* 57, 570–600. doi: 10.1007/s11162-015-9396-5

SEA PHAGES (2020). *The SEA-PHAGES Program*. Available online at: https://seaphages.org (accessed September 9, 2020).

SEQanswers (2020). *SEQanswers*. Available online at: http://seqanswers.com (accessed September 10, 2020).

Sequence Read Archive Submissions Staff. (2011). "Understanding SRA Search Results," in *SRA Knowledge Base [Internet]*. Bethesda, MD: National Center for Biotechnology Information. Available online at: https://www.ncbi.nlm.nih.gov/books/NBK56913/ (accessed September 10, 2020).

Slatko, B. E., Gardner, A. F., and Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. *Curr. Protoc. Mol. Biol.* 122:e59. doi: 10.1002/cpmb.59

Spiers, A. J. (2014). Getting Wrinkly Spreaders to demonstrate evolution in schools. *Trends Microbiol.* 22, 301–303. doi: 10.1016/j.tim.2014.03.007

Stack Overflow (2020). *Stack Overflow*. Available online at: https://stackoverflow.com (accessed September 10, 2020).

Teal, T. K., Cranston, K. A., Lapp, H., White, E., Wilson, G., Ram, K., et al. (2015). Data Carpentry: Workshops to Increase Data Literacy for Researchers. *Int. J. Digital Curat.* 10:351. doi: 10.2218/ijdc.v10i1.351

The jamovi project (2020). *jamovi. (Version 1.2)*. Available online at: https://www.jamovi.org

Van den Bergh, B., Swings, T., Fauvart, M., and Michiels, J. (2018). Experimental Design, Population Dynamics, and Diversity in Microbial Experimental Evolution. *Microbiol. Mol. Biol. Rev.* 82:e00008–e18. doi: 10.1128/MMBR.00008-18

Williams, J. J., Drew, J. C., Galindo-Gonzalez, S., Robic, S., Dinsdale, E., Morgan, W., et al. (2017). Barriers to Integration of Bioinformatics into Undergraduate Life Sciences Education. *bioRxiv* 2017:204420.

Wilson Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2018). Bioinformatics core competencies for undergraduate life sciences education. *PLoS One* 13:e0196878. doi: 10.1371/journal.pone.0196878

Wilson, G. (2014). Software Carpentry: lessons learned. *F1000Research* 3:62. doi: 10.12688/f1000research.3-62.v2 doi: 10.12688/f1000research.3-62.v1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read
for greatest visibility
and readership

FAST PUBLICATION
Around 90 days
from submission
to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative,
and constructive
peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers
acknowledged by name
on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data
and methods to enhance
research reproducibility

DIGITAL PUBLISHING
Articles designed
for optimal readership
across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics
track visibility across
digital media

EXTENSIVE PROMOTION
Marketing
and promotion
of impactful research

LOOP RESEARCH NETWORK
Our network
increases your
article's readership