ORIGINAL STRATEGIES FOR TRAINING AND EDUCATIONAL INITIATIVES IN BIOINFORMATICS

EDITED BY: Hugo Verli and Raquel Cardoso de Melo Minardi
PUBLISHED IN: Frontiers in Education and Frontiers in Bioinformatics

frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-83250-183-2
DOI 10.3389/978-2-83250-183-2

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

ORIGINAL STRATEGIES FOR TRAINING AND EDUCATIONAL INITIATIVES IN BIOINFORMATICS

Topic Editors:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil
Raquel Cardoso de Melo Minardi, Minas Gerais State University, Brazil

Citation: Verli, H., de Melo Minardi, R. C., eds. (2022). Original Strategies for Training and Educational Initiatives in Bioinformatics.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-83250-183-2

Table of Contents

frontiers | Frontiers in Education

TYPE Editorial
PUBLISHED 29 August 2022
DOI 10.3389/feduc.2022.1003098

OPEN ACCESS

EDITED AND REVIEWED BY
Lianghuo Fan,
East China Normal University, China

*CORRESPONDENCE
Renato Augusto Corrêa dos Santos
renatoacsantos@gmail.com

SPECIALTY SECTION
This article was submitted to STEM Education, a section of the journal Frontiers in Education

RECEIVED 25 July 2022
ACCEPTED 15 August 2022
PUBLISHED 29 August 2022

CITATION
dos Santos RAC, Verli H and de Melo-Minardi RC (2022) Editorial: Original strategies for training and educational initiatives in bioinformatics.
Front. Educ. 7:1003098.
doi: 10.3389/feduc.2022.1003098

Editorial: Original strategies for training and educational initiatives in bioinformatics

Renato Augusto Corrêa dos Santos[1]*, Hugo Verli[2] and Raquel Cardoso de Melo-Minardi[3]

[1]Centro de Energia Nuclear na Agricultura, Universidade de São Paulo, Piracicaba, Brazil, [2]Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil, [3]Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

KEYWORDS

bioinformatics, education, training, online course, programming

Editorial on the Research Topic
Original strategies for training and educational initiatives in bioinformatics

In the Research Topic "*Original strategies for training and educational initiatives in bioinformatics*," original research (5), perspective (4), opinion (2), a brief research report (1) and an "curriculum, instruction, and pedagogy" (1) article were published. Most of the authors are affiliated in the U.S. (Niepielko and Shumskaya; Tully et al.; Daetwyler et al.; Dow et al.; Robertson et al.; Goller et al.), Africa (Sangeda et al.; Aron et al.), and South America (Ras et al.; Melo-Minardi et al.; Dorn et al.; Braga et al.; da Silva et al.). Several articles involve intercontinental collaborations (Tully et al.; Ras et al.; Sangeda et al.; Braga et al.). This special edition includes manuscripts published in Frontiers in Bioinformatics (Section Computational BioImaging and Protein Bioinformatics) and Frontiers in Education (Section STEM Education).

In their opinion article, Niepielko and Shumskaya reflect on the necessity of data analysis and interpretation skills for life sciences professionals, and that there is a large gap in their training today (e.g., basic statistics knowledge). Authors argue that bioinformatics should be mandatory and taught in a biology course as soon as students have completed classes on general biology and statistics, and that it should be focused on application of statistics in biology, and not in the mathematical theory behind it, making it more accessible. Next, they report their experiences in introducing bioinformatics to undergraduate students at Kean University, including a mandatory course for sophomores of B.S. in Biology, one for students that want to develop more more skills in computational biology and that also works well for students majoring in computer sciences, and another for those pursuing a B.S. in Science and Technology in Molecular Biology (Bioinformatics and Genomic Science). They developed their own materials suitable for beginners with a biology background, including activities that are available as Open Education Resources. Given the common lack of motivation for

students, they emphasize the importance of hands-on courses that employ "real, timely, and relevant data" for students' excitement. Interestingly, at Kean University they offer a 4-day bioinformatics workshop, which includes lectures and hands-on activities.

Goller et al. describe the inquiry-based Molecular Biology Laboratory Education Modules (MBLEMs) developed by the Biotechnology Program at North Carolina State University and how they use bioinformatics tools combined with wet-lab experiments to answer research questions. In this Perspective article, authors explain how MBLEMs are created, a 5D process that includes Designation, Design, Development, Deployment, and Dissemination, and usually take place as part of the program of teaching postdoctoral scholars. Finally, they provide examples of MBLEMs addressing (i) first-year courses for STEM and non-STEM majors, (ii) elective courses for upper-level and graduate students, and (iii) short courses for completing a biotechnology minor, in which bioinformatics tools are included at different levels. A major focus on MBLEMs design is that bioinformatics tools are chosen to offer the best possible pedagogical experience (e.g., user-friendly software), and courses have peer support and employ inquiry-based projects.

Motivated by cancellation during the COVID-19 pandemic of a successful in-person annual hackathon where students were able to solve "real-world" problems and interact with researchers at UT Southwestern, Daetwyler et al. conceived the "U-Hack Med Gap Year" internship, a paid and completely remote experience designed to overcome the lack of exposure to research in a lab. In their perspective article, they described the administrative challenges, how they overcame several of them, in particular those limiting online employment, and the restrictions that remained (e.g., international interns with temporary VISA were not allowed). They described how the recruitment process was and steps involved, and overall subjects of 10 projects that involved themes such as the visualization of 3D data, machine learning, development of a platform in the medical field, and genome sequencing. Finally, they summarized the key features of this novel internship program: a full involvement of the interns in the research lab and in mentor-mentee process including interaction with the principal investigator, graduate students and postdocs of the host laboratory, their access and training to use institutional computing clusters, a biweekly meeting to provide them high-level training, career sessions, interaction with other interns, and support for achievement of the program learning goals.

Aron et al. described a successful sustainable training environment for Bioinformatics carried across African countries by the Pan-African bioinformatics network for H3Africa (H3ABioNet). It initially combined in-person courses, online classes, and hackathon, and later switched to a mixed-model of online and distance learning, where local institutions and instructors supported learning. First courses in this model occurred successfully and effectively (starting in 2016), and remained active even under restrictions imposed by the COVID-19 pandemic. Authors highlighted how important hackathons organized by the H3ABioNet are to continuous applied skills development and peer learning by African scientists, engineers, and systems administrators, including the online edition that occurred imposed by COVID-19. Finally, they described how the H3ABioNet internship program (during which interns analyze their own data) has strengthened knowledge and skills of individuals, initiatives for building additional career-building skills, and bioinformatics communities through online training and support, being gender-inclusive, and promoting curated and accessible training materials. In summary, H3ABioNet initiatives have reached a very wide audience of bioinformatics users, scientists, and engineers, and together constitute a remarkable example for next multidisciplinary training in resource-limited settings worldwide.

An original research article by Dow et al. demonstrated how the U.S. Department of Energy (DOE) Systems Biology Knowledgebase (KBase) can be used by bioinformaticians and biologists to exploit several data and analysis tools. In particular, the manuscript shows how educators can use the platform in the classroom to teach bioinformatics. Authors demonstrated that educators can leverage Narratives that run workflows available in KBase, or create modified versions of existing ones to adapt to their desired learning concepts, and produce teaching resources that are FAIR, reproducible, and shareable. As part of their study, authors (mostly represented by instructors of the Educators KBase Working Group) presented results of surveys that assessed students' perception of use of KBase resources used in several courses in American institutions such as Genomics, Metagenomics, Microbiology, and Molecular Biology, and also educators's perspectives, showing that such resources were useful and valuable for both groups. Additionally, a study case of a metagenomics course at the North Carolina State University and results of a survey showed Kbase as a useful resource for learning new knowledge and metagenomics analysis pipelines. Finally, authors provide the link to KBase documentation, where interested educators can learn how to use or construct a new narrative or join the community for support, networking, and collaboration.

In another opinion article, Ras et al. report a set of best practices for delivering Bioinformatics training in Low to Middle Income Countries (LMICs), motivated by known difficulties associated with the organization of effective training, such as access to facilities, infrastructure, and internet, and the lack of local expertise. They divided their best practices into three categories: Planning, Development, and Implementation. In planning, they emphasize the importance of inclusion of topics of interested from the involved countries (e.g., use biodiversity data); getting support from key stakeholders, given

that a particular issue in LMICs is limitation of financial and institutional supports; fostering collaboration between countries and regions with different levels of development of bioinformatics capacity and expertise; to have a project plan, so that both administrative and training goals are achieved on time and effectively. Regarding "Development," authors provide suggestions on organizations that have developed materials and learning resources and emphasize the importance of support, resources, and guides over the process, highlighting how important local trainers and researchers are for the success of events, in particular when these events involve international collaboration; importantly, to consider delivering introductory courses in the language spoken by trainees. Finally, regarding the "Implementation" phase, they comment on how to provide an inclusive course during the selection process, giving tips about timing for visa, accommodation and transportation, in case of face-to-face events. Importantly, authors consider socio-political situations that can lead to postponements and cancellations, so organizers must take them into account to insure costs associated with eventual reimbursements.

At the peak of initial shutdowns at the beginning of the COVID-19 pandemics (March, 2019), Tully et al. founded the Bioinformatics Virtual Coordination Network (BVCN), aiming to provide bioinformatics training resources so that wet-lab researchers could learn how to run their bioinformatic analyses while in isolation. While this was the initial objective, the open and mutable lessons in their Github repository allow learners to access and adapt the material to their own needs. In their perspective manuscript, the authors describe how they implemented the education program (which includes several topics, such as programming languages and bioinformatic analyses), all tools and platforms incorporated in lesson plans, how they planned to foster a diverse and inclusive community (including a code of conduct addressing concerns on acceptable and unacceptable behaviors), and even the evolution of topics that emerged over the process, including discussions on variations in approaches used to address research questions (such lesson discussions are made publicly available on the BVCN Youtube channel).

To map the interest and capacity for doing bioinformatics and related research in Tanzania (Sub-Saharan Africa), including human resources and infrastructure, Sangeda et al. employed self-administered online surveys to staff of public and private institutions in the country. The main purposes of such surveys were to enable leveraging of existing resources and building sustainable expertise in the country. As results, they demonstrate that most participants were early career researchers but even though the level of interest in bioinformatics is high, there is a low level of skilled human resources and a lack of infrastructure. Authors raise the importance of investing in training of undergraduate and graduate students, giving a special emphasis to promotion use of digital resources. They encourage collaboration among local institutions and with global partners and stakeholders, highlight the importance of funding and investments from the government for growth and success of local bioinformatics, and launch other communities like the Tanzania Society of Human Genetics and Tanzania Genome Network to promote the use of bioinformatics.

Robertson et al. provide a background on the curriculum and teaching initiatives with focus on data science and analysis skills in the context of high-throughput (HT) technologies. Since these technologies have application in several areas of biology and have been advancing at an enormous pace recently, authors emphasize the importance of teaching the required basis for HT analysis to undergraduate students. They mention successful educational initiatives that gave students the opportunity to analyze data and get involved in research and refer to the High-throughput Discovery Science & Inquiry-based Case Studies for Today's Students (HITS), which focuses on development of curriculum materials based in case studies. HT case studies are complex enough and have untold stories, providing an enriching place for students to dive into data, develop quantitative skills, hypothesis development, and make discoveries; importantly, they include both the focus on experimental approaches and on quantitative data analysis, highlighting the interdisciplinary nature of the process of HT analysis in biology. Finally, they summarize the goals and achievements of the HITS network model to integrate HT discovery into curricula, that include (i) case fellows, that give opportunity to groups committed to create and validate HT case studies; (ii) an HITS annual conference, that provides an opportunity to be exposed to HT, to interact with experts, and to promote collaboration; and (iii) the HITS steering committee that provides guidance for network expansion, ensures inclusivity, and helps to disseminate products and recruit faculty.

Motivated by the lack of clear definition of bioinformatics and the associated difficulties in training and teaching topics in this steadily growing discipline, for both undergraduates in life sciences and computer science, a brief research report by Dorn et al. described the learning model implemented in the Southernmost Brazilian School of Bioinformatics ("Escola Gaúcha de Bioinformática," or EGB). The school brings together graduate and undergraduate students from different fields and backgrounds in a place where they can learn theory and practice (including lectures and hands-on activities) and interact among themselves and with professionals in different areas of bioinformatics, including the analysis of biological sequences and structure of macromolecules. EGB introduces non-experts to bioinformatics, with focus on transdisciplinarity. Importantly, even though most participants are from the state where the school takes place (Rio Grande do Sul), it receives students from many Brazilian states (including regions where bioinformatic research and courses are scarce) and other Latin American countries.

In their "curriculum, instruction, and pedagogy" article, Braga et al. describe the course "Bioinformatics on the Road," another Brazilian initiative from professors in two different public universities, the Federal University of Pará and the Federal Rural University of Amazônia. Motivated by the lack of skilled human resources in bioinformatics, in particular in the countryside and apart from the main national bioinformatics hubs, the scope of this course involves theoretical and practical workshops to graduate and undergraduate students with focus on the areas of interest in the regions where they take place. They provide details on capacitation events over the last 5 years (starting in 2017), their content, computer resources supporting them, and lastly the demographic information about participants, that included graduate and undergraduate students. Like most scientific events happening in 2020, practical and theoretical sessions started to be carried out online from June 2020. At least 400 students benefited from events described in the article, that happened in two cities or online. Interestingly and as part of the results, authors reported that the initiatives not only provided ways to decentralize and strengthen bioinformatics in the countryside, but also connected interested students to research groups working with bioinformatics, either as scientific initiation projects (undergraduate) or in dissertation and theses involving themes (graduate students).

In their original research article, da Silva et al. compare the in-person bioinformatics course CVBioinfo ("Summer Course in Bioinformatics") with its online version WOB ("Online Workshop in Bioinformatics") introduced in 2020 due to the COVID-19 pandemics. They highlighted the main structural differences between the events, and described the overall subjects covered in both editions and the correlation of interests of participants. They also summarized demographic data for participants, emphasizing an increase in the number of women and expansion of their geographical locations in the online edition (WOB). In summary, authors reported to have achieved a more inclusive and accessible event as an online edition, which was also less complex to organize, but also reported limiting factors such as lower interaction among participants, increased distraction, and difficulties in internet connection.

Closing our special edition, Melo-Minardi et al. reported an approach for evaluating an online distance university extension course—OnlineBioinfo—including its resources, activities, instructional design and student drop out, using learning analytics with machine learning methods. Course modules include an introduction to computer science related topics, basic programming with Python, algorithm complexity, and algorithms for bioinformatics, and most enrolled students belong to life sciences courses such as biology, biomedicine, and pharmaceutical sciences, also including computer science or information systems undergraduate students. Based on data from 245 students, authors were able to predict with high accuracy, even based on information gathered from module zero, whether a student would drop the course. They could also identify types of exercise (e.g., review or programming) that could better predict students' completion of the course. Melo-Minardi et al. also provide enriching information on the perception of students concerning different topics in programming, that can be used to improve the relationship between students' difficulties and knowledge of each topic in their own course or serve as basis to similar initiatives.

In summary, this collection highlights important contributions to the field of bioinformatics education, and suggests future directions for research to advance the training of the current and future generations of bioinformaticians, and the curricula of professionals and students in several disciplines dealing with biological data. Importantly, our Research Topic covered authors spread across different regions of the globe. We conclude this editorial with some observed trends in the published articles:

- *The importance of building sustainable expertise in bioinformatics, including sustainable educational networks;*
- *Interpretation of "real-world" data, including COVID-19 data;*
- *The influence of COVID-19 pandemics, including movement of courses and internships that were previously in-person to an online version;*
- *The emergence of short bioinformatics courses, specially in Brazil;*
- *Materials developed and courses provided online and with manageable costs, increasingly available in repositories and whose instructors/developers are more concerned about making them FAIR (Findable, Accessible, Interoperable, and Reusable);*
- *Changes in administration of courses, internships, and training to a model adapted to online resources;*
- *Initiatives to provide access to bioinformatics skills in Low-Income countries, including recognition of challenges regarding access to technology, internet connection, and infrastructure.*

Author contributions

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

Acknowledgments

We thank Patricia Carvajal-López (European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge,

United Kingdom) who helped to compose the original proposal and descriptive page of our Research Topic.

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.

Publisher's note

frontiers in Bioinformatics

OPINION
published: 23 April 2021
doi: 10.3389/fbinf.2021.656531

Early Requirement for Bioinformatics in Undergraduate Biology Curricula

Matthew G. Niepielko[1] *and Maria Shumskaya*[2]*

[1]New Jersey Center for Science, Technology, and Mathematics, Kean University, Union, NJ, United States, [2]School of Natural Sciences, Biology, Kean University, Union, NJ, United States

Keywords: undergraduate education, computational biology, statistics for biologists, R, stem education

INTRODUCTION

As the world unravels its most impactful event of the century so far – the COVID-19 pandemic, - billions of people turn on televisions, tune into radios, and browse websites trying to understand what the epidemiologic graphs are saying; and in most cases, they turn to media and friends asking to explain what these graphs mean. The COVID-19 pandemic has confirmed: there are huge gaps in the ability for the general population to interpret statistical analyses and graphical representation of biological data (Andrew, 2020; Leybzon, 2020; Tracy, 2020).

In the current situation, understanding data being a prerogative for only data specialists is gone; every health care professional, biologist, chemist, or any natural scientist have taken on the responsibility for the evaluation of massive amount of the pandemic data and delivering the conclusions to their friends and families. But are we really prepared for such work and responsibility? As classically trained in US biologists, we were never required by undergraduate college programs to dive deep into the quantitative analysis world (Cheesman et al., 2007). We were dealing with enzyme kinetics graphs in biochemistry, but not once did we touch "big data"–with exception to some of our friends who were brave enough to enroll into a biostatistics class as an elective. Only later, in graduate school, some had an opportunity to take bioinformatics courses such as computational biology, systems biology, or statistical programming.

Today, there is enough data generated by sequencing, gene expression, bench work on DNA, proteins, and metabolites, thus bioinformaticians have plenty of work to do in phylogenetics, gene expression analysis, genome analysis or an interactome prediction (Hagen, 2000; Gauthier et al., 2019). Many of bioinformaticians either have a computer science background or learned computational analysis in their graduate programs. With the overwhelming amount of data that is available today on any topic, including ecology, biodiversity, epidemiology, we believe that all biologists should receive mandatory training in bioinformatics during their undergraduate years, just as they receive training in organic chemistry or physics. For almost two decades, it has been documented that there is need for undergraduate life science majors to graduate with competency in bioinformatics to help not only scientific progression, but students' careers as well (Bialek and Botstein, 2004; Pevzner and Shamir, 2009; Levine, 2014; American Association for the Advancement of Science, 2015; Sayres et al., 2018), with attempts to implement data science across life sciences curriculum (Dill-McFarland et al., 2021). However, there are still barriers preventing successful inclusion of bioinformatics into undergraduate life sciences education, including lack of student interest, overly full curricula, lack of student preparation, and faculty members belonging to underrepresented groups (Williams et al., 2019). Here, we discuss our opinions and experiences regarding the inclusion of bioinformatics early in undergraduate life science curricula at Kean University, a Hispanic-serving Institution (HSI).

Courses that cover bioinformatics skills should be offered as early as sophomore year, immediately after students complete two semesters of general biology courses and introductory

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Renato Augusto Corrêa Dos Santos, State University of Campinas, Brazil

***Correspondence:**
Maria Shumskaya
mshumska@kean.edu

Specialty section:
This article was submitted to Computational BioImaging, a section of the journal Frontiers in Bioinformatics

Received: *21 January 2021*
Accepted: *08 April 2021*
Published: *23 April 2021*

Citation:
Niepielko MG and Shumskaya M (2021) Early Requirement for Bioinformatics in Undergraduate Biology Curricula. Front. Bioinform. 1:656531. doi: 10.3389/fbinf.2021.656531

statistics. We advocate for this program improvement because biologists need to understand basic data analysis and be familiar with the methods applied, their pros and cons. The lack of important skills necessary to evaluate the validity of a data analysis or draw a critical conclusion from a graph we observe in undergraduate biology majors is devastating. "What is p-value, why do we need it here and what does it tell us? Are graphs scaled correctly for comparison? Do the error bars represent standard error or standard deviation? Is all the data represented on the graph or is there just the averages?"–"How would I know?" These common conversations with students make it clear: the opinion of a person with a college degree in biology can be easily manipulated with some invalid data. We would not expect a biology major to excel in math; rather, we want all students to accrue basic computational analysis skills to deal with the data by embracing the Jim Frost idea: "I'll help you intuitively understand statistics by focusing on concepts and using plain English so you can concentrate on understanding your results" (Allison Loves Math Podcast, 2021; Frost, 2021).

Possessing essential computational skills and bioinformatics tools are no longer for special people talented "in computers;" it is a required competency for a biologist (Pevzner and Shamir, 2009; White et al., 2013; Sayres et al., 2018). Computational classes for sophomore level biology majors should focus on biological application, rather than the math theory behind it. A biologist should need to know how and when to use a statistical test and how to interpret the data; the statistical equations behind it should be secondary. Such a course should be taught by a biologist who understands data analysis requirements, who is trained in R, Python, MATLAB, MEGA, PyMOL (Van Rossum and Drake, 1995; MatLab, 2010; PyMOL, 2010; Core R Team, 2017; Kumar et al., 2018) and can design multiple practical exercises for the course. Data on gene expression, heart rates, cholesterol levels, drug efficacy, biodiversity and community structure, evolutionary relationships, selection in a population, and mutations in a gene can be incorporated into class exercises thanks to easily accessible and free online databases and publications.

OUR EXPERIENCE

Computational Courses in a Biology Undergraduate Curriculum

To start filling the gaps in data analysis skills for future biologists, we offer a course on Bioinformatics (3-credits) for sophomores majoring in BS Biology at Kean University. The course is mandatory for the program option in Cell and Molecular Biology and is designed as a set of hands-on exercises on data analysis. Students use Excel and R statistical programming to work with biodiversity data, MEGA to study alignments and phylogenetics, and PyMOL for protein modeling. Introduction to R is taught in an online game form using DataCamp platform (datacamp.com). Since we found that most textbooks are too advanced for our sophomore undergraduates, we developed our own teaching activities. The activities are based on data available from NCBI, NEON, PDB databases, or our own research data (Shumskaya et al., 2019), and some even using the early nucleic acid and protein structure data on SARS-CoV-2 virus (Lorusso and Shumskaya, 2020; Shumskaya and Lorusso, 2020). The activities cited are published online as Open Education Resources to help promote our teaching philosophy surrounding bioinformatics.

For students interested in developing more skills in computational biology, we have developed a minor in Bioinformatics which includes an advanced course on biostatistics and a set of courses on basic computer programming that would count as free electives. This minor works for students majoring in computer sciences or informational technology as well; such students are required to have passed two semesters of general biology and genetics in addition to bioinformatics and statistics.

Additionally, we offer a Bioinformatics and Genomic Science track for our students pursuing a B.S. in Science and Technology in Molecular Biology. Students in this track complete courses in computer programming, statistical programming, and a bioinformatics elective by the end of their sophomore year. The responses from sophomores that complete our undergraduate bioinformatics course and related courses are overwhelmingly positive. In general, none of the students even considered computational biology as a field prior to participating in our bioinformatics course, being unfamiliar with this option. Course surveys reveal that most students become interested in the field when working with data, especially now when a lot of data on current COVID pandemic is available to practice (Johns Hopkins University Center for Systems Science and Engineering (CSSE), 2021) and appreciate learning new software that can help them succeed in multiple courses. A lack of student interest is a clear barrier that prevents students from pursuing higher-level courses that cover topics in bioinformatics (Williams et al., 2019). Our opinion is that a key component in getting students excited and interested in bioinformatics includes hands-on course exercises that cover real, timely, and relevant data (Shumskaya et al., 2019; Lorusso and Shumskaya, 2020; Shumskaya and Lorusso, 2020).

Updating of Pre-Requisite Courses

Biology majors at Kean University are required to take a 3-credits course on statistics. This course traditionally has a lecture component with a heavy math approach; however, at Kean there is an option to add a 1 credit "Probabilistic Methods Lab" taught by a biologist. In this lab, students learn the R statistical programming language in the first half of the semester using the "Undergraduate Guide to R" tutorial (Martin, 2009). Programming activities are designed to reiterate key concepts such as data structures, functions, normalization of data, and graphs using the ggplot R package (Wickham, 2009), followed by the analysis of data such as drug efficacy and RNA expression levels. What separates this lab from a traditional computer programming courses is surrounding algorithms and equations are not detailed; rather, concepts and application of statistical tests using R

are the focus. In this lab, students are "tool users" rather than "tool makers" (Pevsner, 2015). This enables students to analyze data and interpret results without the intimidation of complex equations. Because this course is given during their sophomore year, the exposure to basic computer programming has motivated some students to pursue more advanced computer programming courses during their junior or senior years. This approach helps us introduce bioinformatics into a lower level course, alleviating the "lack of student preparation" barrier (Williams et al., 2019).

Computational Biology for Undergraduate Research

At Kean University, undergraduates can participate in undergraduate research courses [CUREs (Corwin et al., 2015; Rodenbusch et al., 2016; Shortlidge and Brownell, 2016)] such as Research-First-Initiative (RFI, freshmen) or Research Experience in Biology (REB, juniors). Such courses offer students an opportunity to join a faculty-led research project. The 2-3 credits courses are counted as a part of students' 120-college credit program and are often used to jumpstart future independent research projects. Computational biology research projects are offered as part of CUREs. One RFI project focuses on understanding how mRNA localize within a developing cell. Specifically, freshmen are trained in quantifying real mRNA localization data from confocal images using custom MATLAB scripts from (Niepielko et al., 2018), and how to statistically analyze data and compare how mRNA localization changes in various genetic backgrounds. In two back-to-back semesters, students learn biological research such as single molecule *in situ* hybridization and confocal microscopy followed by computational analysis using MATLAB and R statistical programming. One REB course focuses on molecular biodiversity of dead wood decomposing fungi. Students work with Next Generation Sequencing to assess mycobiome gathered from environmental samples, and then employ a bioinformatics pipeline to analyze NGS data. This research course option finishes with students performing ordination and other statistical analyses to study microbial communities identified on dead wood. Research has shown that students engaging in research as undergraduates had the greatest benefit (Russell et al., 2007; Russell et al., 2017). By offering computational biology research and hands on training opportunities built into life science curricula, we believe that this addresses multiple educational barriers including lack of student interest, overly full curricula, and lack of student preparation. Together, we feel that our approach creates an environment that promotes bioinformatics while benefiting students (Levine, 2014) and faculty research projects.

Summer Workshop for High School Students

We believe that bioinformatics and computational biology should be offered to students as early as possible. At Kean University, we offer a bioinformatics workshop for high school students that are interested in any STEM field. The 4 day remote workshop is offered during the summer months as part of Kean University's Group Summer Scholars Research Program. The course is structured with a 2 h morning session and 2 h afternoon session which allows for students to receive a lecture on all the relevant background information in the morning and apply that knowledge by completing hands-on exercises in the afternoon. The hands-on activities cover RNA, DNA, and protein databases, BLAST searches, sequence analysis using MEGA, and protein structure analysis using PyMOL. Based on our experience, introducing bioinformatics to high school students has been overwhelmingly positive. Regardless of their diverse STEM interests, students are receptive to learning about the field and are proficient at completing all the workshop activities which include articulating key findings and developing hypotheses. Although Kean's workshop is not part of a research study, we believe that offering general introduction to bioinformatics at a high school level will help relieve the "lack of student preparation" barrier identified in research studies (Williams et al., 2019). Furthermore, the feasibility and success of the workshop supports our opinion that early exposure to bioinformatics course material should be a strategy integrated into biology curricula and can be accomplished by including

TABLE 1 | Introduction of computational skills in the biology curriculum.

Level	Biology major	Recommended intervention	Course content	Mandatory
Junior high school students	No	16 h summer workshop	Intro into RNA, DNA, and protein databases, BLAST searches, sequence analysis using MEGA, and protein analysis using PyMOL	No
Any college undergraduate level	Yes	Introductory research course on a specific topic, 1–2 credits	Hands-on analysis of real biological data acquired in lab or from publications	No
College freshmen/ sophomores	Yes	1 semester of a lab course on statistics, 1 credit	Basics of R functions, ggplot graphing, interpreting graphs	Highly recommended
College sophomores	Yes	1 semester course on essentials of bioinformatics, 3 credits	Basics of working with biological data in Excel, R, MEGA, PyMOL. t-test, standard deviation and errors, ordination, BLAST, multiple sequence alignment, phylogeny, protein modeling, selection. Online databases NCBI, PDB etc.	Yes
College juniors, seniors	Yes	1 semester course on biostatistics, 3 credits	Advanced data analysis in Excel and R	Yes for minors in bioinformatics; major elective for others

more background information into the course design rather than relying on mandatory pre-requisite courses, which should help alleviate overly full curricula issues.

DISCUSSION

It is no secret that many barriers exist that prevent exposing students to computational biology and bioinformatics, hence introduction of a special course on computational skills into undergraduate biology curricula is in dire need (Sayres et al., 2018; Williams et al., 2019). Our experience shows that the early introduction and a careful planning of computational biology courses has a positive influence on our diverse undergraduate student population. We summarized our steps on introducing computational biology in a biology curriculum in **Table 1**. Our goal is to promote and teach computational skills as early as possible so that students become comfortable with topics such as "How do I analyze data?" "When do I do a certain statistical test?" "What does the p-value actually mean?" In our opinion, biology students learning computer skills from other biologists helps students embrace quantitative biology without fear of overwhelming complex equations and computational algorithms.

From our experience, providing an early opportunity for students to get involved with computational biology spikes their interest to continue to more advanced independent research projects, especially if they participate in CUREs. In a broader sense, such training would have a huge impact on our society. As documented with COVID-19 analyses we discussed above, scientific data can be misrepresented very easily, leading towards rapid spread of misinformation and poor policy choices. The more specialists that receive training in data analysis and data interpretation, the better, regardless of their specialized background. In the future, perhaps general education courses on data analysis and data interpretation can be designed and made a requirement for all student majors.

AUTHOR CONTRIBUTIONS

Both MS and MN contributed equally to writing the manuscript.

REFERENCES

Allison Loves Math Podcast (2021). #36 How to Make Statistics Exciting with Jim Frost. Available at: https://allisonlovesmath.mykajabi.com/blog/JimFrost (Accessed on April 2, 2021).

American Association for the Advancement of Science (2015). Vision and Change in Undergraduate Biology Education: Chronicling Change, Inspiring the Future. Available at: https://visionandchange.org/about-v-c-chronicling-the-changes/ (Accessed April 2, 2021).

Andrew, G. (2020). Hey, I Think Something's Wrong with This Graph!. Available at: https://statmodeling.stat.columbia.edu/2020/05/18/hey-i-think-somethings-wrong-with-this-graph (Accessed on April 2, 2021).

Bialek, W., and Botstein, D. (2004). Introductory Science and Mathematics Education for 21st-Century Biologists. *Science* 303 (5659), 788–790. doi:10.1126/science.1095480

Cheesman, K., French, D., Cheesman, I., Swails, N., and Thomas, J. (2007). Is There Any Common Curriculum for Undergraduate Biology Majors in the 21st Century?. *Bioscience* 57 (6), 516–522. doi:10.1641/b570609

Core, R. Team. (2017). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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 (1), es1. doi:10.1187/cbe.14-10-0167

Dill-McFarland, K. A., König, S. G., Mazel, F., Oliver, D. C., McEwen, L. M., Hong, K. Y., et al. (2021). An Integrated, Modular Approach to Data Science Education in Microbiology. *PLoS Comput. Biol.* 17 (2), e1008661. doi:10.1371/journal.pcbi.1008661

Frost, J. (2021). Statistics by Jim. Available at: https://statisticsbyjim.com/ (Accessed April 2, 2021).

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A Brief History of Bioinformatics. *Brief. Bioinform.* 20 (6), 1981–1996. doi:10.1093/bib/bby063

Hagen, J. B. (2000). The Origins of Bioinformatics. *Nat. Rev. Genet.* 1 (3), 231–236. doi:10.1038/35042090

Johns Hopkins University Center for Systems Science and Engineering (CSSE) (2021). COVID-19 Data Repository. Available at: https://github.com/CSSEGISandData/COVID-19 (Accessed March 1, 2021).

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35 (6), 1547–1549. doi:10.1093/molbev/msy096

Levine, A. G. (2014). An Explosion of Bioinformatics Careers. *Science*. 344 (6189), 1303–1306. doi:10.1126/science.344.6189.1303

Leybzon, D. D. (2020). Bad Data Visualization in the Time of COVID-19. Available at: https://medium.com/nightingale/bad-data-visualization-in-the-time-of-covid-19-5a9f8198ce3e (Accessed on April 2, 2021).

Lorusso, N. S., and Shumskaya, M. (2020). Online Laboratory Exercise on Computational Biology: Phylogenetic Analyses and Protein Modeling Based on SARS-CoV -2 Data during COVID -19 Remote Instruction. *Biochem. Mol. Biol. Educ.* 48 (5), 526–527. doi:10.1002/bmb.21438

Martin, T. (2009). *The Undergraduate Guide to R*. Princeton, NJ: Princeton University.

MatLab (2010). *Natick*. Massachusetts: The MathWorks Inc.

Niepielko, M. G., Eagle, W. V. I., and Gavis, E. R. (2018). Stochastic Seeding Coupled with mRNA Self-Recruitment Generates Heterogeneous Drosophila Germ Granules. *Curr. Biol.*, 28(12), 1872–1881. doi:10.1016/j.cub.2018.04.037

Pevsner, J. (2015). *Bioinformatics and Functional Genomics*. Hoboken, United States: Wiley.

Pevzner, P., and Shamir, R. (2009). Computing Has Changed Biology-Biology Education Must Catch up. *Science*. 325 (5940), 541–542. doi:10.1126/science.1173876

Rodenbusch, S. E., Hernandez, P. R., Simmons, S. L., and Dolan, E. L. (2016). Early Engagement in Course-Based Research Increases Graduation Rates and Completion of Science, Engineering, and Mathematics Degrees. *CBE Life Sci. Educ.* 15 (2), ar20. doi:10.1187/cbe.16-03-0117

Russell, J. E., D'Costa, A. R., Runck, C., Barnes, D. W., Barrera, A. L., Hurst-Kennedy, J., et al. (2017). Correction for Bridging the Undergraduate Curriculum Using an Integrated Course-Embedded Undergraduate Research Experience (ICURE). *CBE Life Sci. Educ.* 16 (1), co3. doi:10.1187/cbe.14-09-0151-corr

Russell, S. H., Hancock, M. P., and McCullough, J. (2007). THE PIPELINE: Benefits of Undergraduate Research Experiences. *Science* 316 (5824), 548–549. doi:10.1126/science.1140384

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 (6), e0196878. doi:10.1371/journal.pone.0196878

Shortlidge, E. E., and Brownell, S. E. (2016). How to Assess Your CURE: A Practical Guide for Instructors of Course-Based Undergraduate Research Experiences †. *J. Microbiol. Biol. Educ.* 17 (3), 399–408. doi:10.1128/jmbe.v17i3.1103

Shumskaya, M., and Lorusso, N. (2020). Introduction to Nucleotide Sequence Analysis and Protein Modeling in MEGA and PyMol Using Coronavirus SARS-CoV-2. *QUBES Educ. Resour.* doi:10.25334/37N4-SW29

Shumskaya., M., and Zambell, C. (2019). NMDS to Study Dead Wood Fungi Communities in Parks of New Jersey. NEON Faculty Mentoring Network. *QUBES Educ. Resour.* doi:10.25334/A2ME-QH70

The PyMOL Molecular Graphics System (Version 1.3r1 edu). (2010). *Schrödinger, LLC*, https://pymol.org/2/.

Tracy, S. (2020). COVID-19 in Charts: Examples of Good & Bad Data Visualization. https://analythical.com/blog/covid19-in-charts (Accessed on April 2, 2021).

Van Rossum, G., and Drake, J. F. L. (1995). *Python Reference Manual*. Amsterdam: Centrum voor Wiskunde en Informatica.

White, H. B., Benore, M. A., Sumter, T. F., Caldwell, B. D., and Bell, E. (2013). What Skills Should Students of Undergraduate Biochemistry and Molecular Biology Programs Have upon Graduation?. *Biochem. Mol. Biol. Educ.* 41 (5), 297–301. doi:10.1002/bmb.20729

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

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 (11), e0224288. doi:10.1371/journal.pone.0224288

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

frontiers in Education

ORIGINAL RESEARCH
published: 25 June 2021
doi: 10.3389/feduc.2021.665313

A Baseline Evaluation of Bioinformatics Capacity in Tanzania Reveals Areas for Training

Raphael Zozimus Sangeda[1,2], Aneth David Mwakilili[3,4], Upendo Masamu[2], Siana Nkya[2,5], Liberata Alexander Mwita[1,2], Deogracious Protas Massawe[6], Sylvester Leonard Lyantagaye[7] and Julie Makani[2]*

[1]Department of Pharmaceutical Microbiology, Muhimbili University of Healthy and Allied Sciences, Dar es Salaam, Tanzania, [2]Muhimbili Sickle Cell Program, Department of Haematology and Blood Transfusion, Muhimbili University of Healthy and Allied Sciences, Dar es Salaam, Tanzania, [3]Department of Molecular Biology and Biotechnology, University of Dar es Salaam, Dar es Salaam, Tanzania, [4]Plant Protection Department, Swedish University of Agricultural Sciences, Alnarp, Sweden, [5]Department of Biological Sciences, Dar es Salaam University College of Education, Dar es Salaam, Tanzania, [6]Department of Crop Sciences and Horticulture, Sokoine University of Agriculture, Morogoro, Tanzania, [7]Department of Biochemistry, Mbeya College of Health and Allied Sciences, University of Dar es Salaam, Mbeya, Tanzania

Due to the insufficient human and infrastructure capacity to use novel genomics and bioinformatics technologies, Sub-Saharan Africa countries have not entirely ripped the benefits of these technologies in health and other sectors. The main objective of this study was to map out the interest and capacity for conducting bioinformatics and related research in Tanzania. The survey collected demographic information like age group, experience, seniority level, gender, number of respondents per institution, number of publications, and willingness to join the community of practice. The survey also investigated the capacity of individuals and institutions about computing infrastructure, operating system use, statistical packages in use, the basic Microsoft packages experience, programming language experience, bioinformatics tools and resources usage, and type of analyses performed. Moreover, respondents were surveyed about the challenges they faced in implementing bioinformatics and their willingness to join the bioinformatics community of practice in Tanzania. Out of 84 respondents, 50 (59.5%) were males. More than half of these 44 (52.4%) were between 26–32 years. The majority, 41 (48.8%), were master's degree holders with at least one publication related to bioinformatics. Eighty (95.2%) were willing to join the bioinformatics network and initiative in Tanzania. The major challenge faced by 22 (26.2%) respondents was the lack of training and skills. The most used resources for bioinformatics analyses were BLAST, PubMed, and GenBank. Most respondents who performed analyses included sequence alignment and phylogenetics, which was reported by 57 (67.9%) and 42 (50%) of the respondents, respectively. The most frequently used statistical software packages were SPSS and R. A quarter of the respondents were conversant with computer programming. Early career and young scientists were the largest groups of responders engaged in bioinformatics research and activities across surveyed institutions in Tanzania. The use of bioinformatics tools for analysis is still low, including basic analysis tools such as BLAST, GenBank, sequence alignment software, Swiss-prot and TrEMBL. There is also poor access to resources and tools for bioinformatics analyses. To address the skills and

OPEN ACCESS

Edited by:
Raquel Cardoso de Melo Minardi, Minas Gerais State University, Brazil

Reviewed by:
Renato Augusto Corrêa Dos Santos, State University of Campinas, Brazil
Sabrina Silveira, Universidade Federal de Viçosa, Brazil

***Correspondence:**
Raphael Zozimus Sangeda
sangeda@gmail.com

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *07 February 2021*
Accepted: *15 June 2021*
Published: *25 June 2021*

Citation:
Sangeda RZ, Mwakilili AD, Masamu U, Nkya S, Mwita LA, Massawe DP, Lyantagaye SL and Makani J (2021) A Baseline Evaluation of Bioinformatics Capacity in Tanzania Reveals Areas for Training. Front. Educ. 6:665313. doi: 10.3389/feduc.2021.665313

resources gaps, we recommend various modes of training and capacity building of relevant bioinformatics skills and infrastructure to improve bioinformatics capacity in Tanzania.

Keywords: bioinformatics, Tanzania, Tanzania genome network, Tanzania society of human genomics, bioinformatics education, bioinformatics capacity

INTRODUCTION

Recently, the field of genomics has become instrumental in medical research and provision of healthcare diagnosis, understanding prevention, and treatment of several disease conditions (Adedokun et al., 2016; Shoko et al., 2018). This was fuelled by the increased ability to generate data and perform bioinformatics analysis, which has become critical for biomedical scientists (Mulder et al., 2016a), particularly in the face of the continued fall of the cost of data generation and analysis using trending technologies (Mulder et al., 2016a).

Despite the decreasing cost of using bioinformatics technologies for research at the global level, Sub-Saharan Africa (SSA) countries, including Tanzania, face difficulties accessing quality health despite the significant disease burden in these countries (de Martel et al., 2020). The lag in SSA is due to the lack of human and technological capacity to run and interpret such bioinformatics analysis effectively, thus hindering the benefits of applying genomics in medicine and other research areas (Karikari et al., 2015; Adedokun et al., 2016; Mulder et al., 2017). The hurdle in health research also extends into leveraging new technologies such as genomics and bioinformatics to resolve some significant issues such as food insecurity and poverty (Lyantagaye, 2013) by focusing on human health, agriculture and animals production research.

Several initiatives have been established to address the gap. One of the initiatives is the Human Heredity and Health in Africa (H3Africa, 2021a) Pan Africa Bioinformatics Network (H3ABioNet) (Mulder et al., 2017). This African initiative was established to facilitate bioinformatics capacity in the continent and health genomics research (H3Africa, 2021b; Mulder et al., 2016a). H3Africa has successfully mobilized resources and developed researchers' networks and capacity for various research resources, including biobanks, developing researchers networks, and capacity to analyze genomics data through H3ABioNet (Mulder et al., 2017). Through this network, both human and infrastructure bioinformatics needs have been addressed through training, setting up of standardized bioinformatics analysis workflows, access to expertise in various domains and data harmonization have been put in place in Africa (Mulder et al., 2017).

However, individual countries may not have fully embraced the collaborative efforts to strengthen the bioinformatics capacity. For example, in Tanzania, only three nodes became members of the network. These were initially the University of Dar es Salaam (UDSM), Muhimbili University of Health and Allied Sciences (MUHAS) and the Management and Development for Health (MDH). MUHAS is still an active member since 2012, with a renewed new grant for the year 2017–2022 (H3ABionet, 2021). Tanzania is also a part of other international and regional bioinformatics networks and consortia, including the Eastern Africa Network for Bioinformatics Training (EANBiT) (Hernández-de-Diego et al., 2017) and the African Society for Bioinformatics and Computational Biology (ASBCB) (http://www.asbcb.org/). Several local initiatives are geared towards advancing the capacity to conduct bioinformatics and related research, such as the Tanzania Genome Network, an association of bioinformaticians from public and private research institutions, and the Tanzania Society of Human Genetics (TSHG).

Currently, only the University of Dar es Salaam (UDSM) and Sokoine University of Agriculture (SUA) in Tanzania offer some form of training in Bioinformatics. The UDSM and SUA undergraduate prospectus 2018/2019 include selected programs offering Bioinformatics courses at undergraduate and postgraduate levels (Lyantagaye, 2013). These bioinformatics courses are embedded in other programs and none of the Tanzanian universities offer a pure bioinformatics program at undergraduate or postgraduate levels.

There is a need to document the existing human capacity for conducting bioinformatics-related research and analyses. This will enable effective leveraging of existing resources and strategizing to build sustainable expertise in the country further. There is no documentation on the existing bioinformatics capacity in the country to the best of our knowledge.

This study aimed to evaluate the existing human expertise and capacity to use bioinformatics tools for research in public and private institutions to address this challenge. On the one hand, the documentation is hoped to guide the leveraging of present resources and identify areas for improvement and training. On the other hand, it will also support the H3Africa and H3ABioNet and other projects' efforts to build bioinformatics capacity in Africa. The study findings may help to make recommendations for improvement in bioinformatics training and research in Tanzania, a model that can be emulated in other SSA countries. Training people in Bioinformatics will also provide the critical mass to manage the local resources such as computing infrastructure, data centers and high-performance computers.

MATERIALS AND METHODS

This is a cross-sectional, explorative, descriptive study among researchers and academics in Tanzania's public and private academic and research institutions. The study employed a self-administered online survey to gather information regarding the baseline status of bioinformatics practices in Tanzania.

TABLE 1 | Demographic attributes of the respondents surveyed about bioinformatics practice in Tanzania (*N* = 84).

Attribute of respondent		N	%
Gender	Male	50	59.5
	Female	34	40.5
Age group	19–25	6	7.1
	26–32	44	52.4
	33–39	19	22.6
	40–46	5	6.0
	46–52	8	9.5
	53–59	1	1.2
	≥60	1	1.2
Highest education attained	Master's degree	41	48.8
	Bachelor degree	25	29.8
	PhD	18	21.4
Area of research and practice ±	Molecular biology	18	21.4
	Medical	15	17.9
	Microbiology	13	15.5
	Biotechnology	13	15.5
	Agriculture	11	13.1
	Genomics and bioinformatics	7	8.3
	Biochemistry	2	2.4
Where did the respondents get bioinformatics training? ±	Bachelor training	40	47.6
	Master training	27	32.1
	Conference or workshop	24	28.6
	PhD training	18	21.4
	Short course	17	20.2
	Online course	13	15.5
	Reading articles	11	13.1
Which computing facilities do you have access to? ±	Personal laptop or desktop	81	96.4
	Institutional computer server in Tanzania	34	40.5
	Institutional computer server abroad	7	8.3
	High-performance computer in Tanzania	9	10.7
	High-performance computer abroad	7	8.3
	Cloud computing	6	7.1
Number of publications related to bioinformatics	0	56	66.7
	1–4	24	28.6
	5–8	4	4.8
Operating system used ±	Windows 8 or above	57	67.9
	Windows 7 or below	21	25.0
	macOS	19	22.6
	Linux	12	14.3
Know computer programming?			
	Yes	21	25.0
	No	63	75.0
The programming languages used ±	Python	8	9.5
	Bash or another scripting	5	6.0
	Java	4	4.8
	C++	2	2.4
	C	2	2.4
	Perl	2	2.4
	Basic	1	1.2
	Visual basic	1	1.2
	Haskell	1	1.2
	Octave	1	1.2
	Erlang	1	1.2
	Smalltalk	1	1.2
	LISP	1	1.2
	Prolog	1	1.2
	Yes	15	17.9
Do you know any computer database management system	No	69	82.1
Computer database management system used ±	Microsoft access	14	16.7
	MariaDB/MySQL	6	7.1
	PostgressPro	3	3.6
	Foxpro	2	2.4
	Ms SQL	2	2.4
	Dbase	1	1.2
	Clipper	1	1.2
Willing to join the bioinformatics network and genomics initiative in Tanzania	Yes	80	95.2
	No	4	4.8

(Continued in next column)

TABLE 1 | (*Continued*) Demographic attributes of the respondents surveyed about bioinformatics practice in Tanzania (*N* = 84).

Key±: multiple responses were possible for this question.

An online survey was developed and distributed using REDCap tool (Harris et al., 2009; Harris et al., 2019). The study population included staff from research institutions and academic institutions offering education in health, agriculture and other natural sciences. A survey link was sent to scientists in Tanzania's academic and research institutions, including research, education, and commercial institutions. The survey was distributed through individual emails, mailing lists in relevant groups such as Tanzania Genome Network (TGN) and institutional mailing lists and social media platforms. The English language was used for the survey since it is the official language of communication in academia and research in Tanzania. The survey was conducted between September 2018 and November 2018.

The survey began with an introduction to the bioinformatics research, an explanation of the study's objectives and the information expected from the participant. Participants were assured of anonymity and privacy of collected data by reporting it in an aggregated format. Information captured included respondents' demographics such as employment institution, age group, gender, level of seniority, and area of research. The level of seniority question intended to capture self-perceived positioning of seniority in the profession where 0–50 was the early carrier, 50 was a mid-carrier and 51–100 was a senior. Other questions related to the years of work experience, number of publications in bioinformatics, and the highest level of education attained. The sections that followed investigated access to and knowledge about infrastructure and software tools for bioinformatics analysis. We also asked questions about access to computing facilities and computer operating systems in regular use by the respondents.

We evaluated the skill levels of the selected Microsoft Office tools and selected statistical packages as well as the frequency of use of some basic bioinformatics resources such as PubMed (Ossom Williamson and Minter, 2019), Swiss-prot and TrEMBL—Protein sequence databases (Bairoch, 1996), National Center for Biotechnology Information (NCBI) 's BLAST search (McGinnis and Madden, 2004), GenBank (Clark et al., 2016), European Bioinformatics Institute (EMBL-EBI) (Li et al., 2015; Madeira et al., 2019), DNA Data Bank of Japan (DDBJ) (Mashima et al., 2017), Entrez Genome Browser, Human Genome Browser from UCSC (Kent et al., 2002), Protein Data Bank (PDB) (Berman, 2000),

sequence alignments software such as Muscle (Edgar, 2004), T-coffee (Di Tommaso et al., 2011) and CLC Workbench (a QIAGEN product for DNA, RNA and protein sequence data analysis) (Smith, 2015). Lastly, we asked questions intending to understand frequently analyzed tasks, ranging from sequence alignment, phylogenetic, 16s data analysis, genome-wide association studies (GWAS), internal transcribed spacer (ITS) data analysis, variant calling, genome annotation, RNASeq, proteomics and other tasks as specified.

We also asked questions intended to investigate the participants' knowledge and type of computer programming languages and the computer database management systems preferred. We sought out to identify the challenges that respondents face in bioinformatics research. The broader problems were re-categorized into electric power and internet, mentorship and research network, computer infrastructure, and training skills.

Finally, we interrogated the participants' willingness to join the bioinformatics network and initiative in Tanzania under the TGN.

Total Bioinformatics Analysis Knowledge Score

The total analysis score was calculated based on scoring knowledge of nine essential bioinformatics skills. These included sequence alignment, phylogenetics, 16s analysis, GWAS, ITS, variant calling, genome annotation, RNASeq or proteomics. A respondent scored a '1′ for each skill they knew and then a total score was calculated.

Statistical Analysis

The survey responses were exported from REDCap into a comma-separated file for analysis. Analysis of the results was conducted using R (R Development Core Team, 2020) software integrated into R Studio version 1.2.5033.

Descriptive statistics, including frequency tables and bar plots, were used to summarize the responses. The Pearson chi-square test was employed to determine the association between the publication status and knowledge of Bioinformatics analysis tools, taking a p-value < 0.05 as a significant cutoff at a 95% confidence interval.

Ethical Approval and Consent to Participate

Participant's consent was requested before conducting the survey. The survey was halted if the participants opted not participating in the survey. The Muhimbili University of Health and Allied Sciences (MUHAS) Research Ethics Committee granted a study waiver of informed consent. No identifying information was collected.

RESULTS

Demographic Characteristics of Respondents

A total of 90 respondents from academic and non-academic institutions participated in the survey. Six respondents were removed because they acknowledged that they do not know anything about bioinformatics at the beginning of the survey. The majority of respondents (**Table 1**) were male participants, 50 (59.5%), while females were only 34 (40.5%). When asked to self-rate their seniority on a scale of 0–100, respondents rated themselves with mean seniority of 39.1 [Interquartile range (IQR) 8.0–53.0]. The mean work experience of the respondents in years was 6.2 (IQR 2.0–8.0). Concerning the participants' age groups, the majority of respondents, 44 (52.4%), were aged between 26 and 32 years (**Table 1**). The highest education level attained by most respondents were master's degree holders 41 (45.8%) followed by bachelor degree holders 25 (29.8%) (**Table 1**).

The number of publications related to bioinformatics by the respondents was mainly in the range of 1–4, as reported by 24 (28.57) respondents (**Table 1**). Altogether, only 28 (33.3%) of the surveyed respondents have at least one publication about bioinformatics. In comparison, 56 (67.7%) did not have any publications in bioinformatics. We did not find any association between the number of publications and the total score of bioinformatics analysis knowledge (Chi-square p-value = 0.360) (**Supplementary Table S1**).

The area of research or practice for the majority of the respondents was molecular biology 18 (21.4%), followed by 15 (17.9%) from the field of medicine (**Table 1**).

Most of the respondents reported learning bioinformatics at bachelor 40 (47.6%), followed by master's training 27 (32.1%) and other sources (**Table 1**).

Access to Infrastructure for Bioinformatics Analysis

Eighty-one (96.4%) of the respondents used their personal computers (laptops) for bioinformatics work. A small percentage (less than 10%) indicated having access to institutional servers abroad or computer cloud (**Table 1**). Fifty-seven (67.9%) of these respondents run their computers on Windows 8. Only twelve (14.3%) of these respondents have the Linux operating system on their computer systems (**Table 1**).

Knowledge and use of computer programming language and database management systems.

Only a quarter of the respondents reported using computer programming language and 15 (17.9%) use a database management system. The most used programming language is Python by 8 (9.5%) of the respondents. The widely used database management systems were Microsoft Access and MariaDB/MySQL, which were used by 14 (16.7%) and 6 (7.1%) of the respondents, respectively (**Table 1**).

Out of the 84 respondents confirmed to know bioinformatics, 80 (95.2%) (**Table 1**) were willing and ready to join the bioinformatics network and initiative in Tanzania under the TGN.

The majority of respondents were from research institutions, 50 (59.5%) (**Table 2**). The respondents were from a total of 33 institutions (**Supplementary Table S2**).

Challenges Facing Bioinformatics in Tanzania

More than half of the respondents reported one or more problems in Tanzania's bioinformatics practice (**Table 3**). The majority, 22 (26.2%), reported a lack of training and skills as a

TABLE 2 | Distribution of respondents per type of institution in Tanzania.

Type of institution	N	%
Research institutions	50	59.5
Academic institutions	26	31.0
Others	7	8.3
Private/commercial	1	1.2

significant problem. Only 2 (2.4%) of the respondents reported inadequate electrical power supply and lack of internet access (**Table 3**). All except one of the 51 respondents faced challenges running bioinformatics analyses use personal computers or laptops. However, even those who did not face challenges, only one out of 24 (Chi-Square p-value = 0.338) used infrastructure other than personal computers or laptops (**Supplementary Table S3**).

Many challenges were given by participants in the categories shown in **Table 3**. Here are two examples of challenges stated by the responses that were given as free text. A female respondent replied, "Yes, I face challenges. I used JoinMap (Ooijen, 2021) (a Microsoft-Windows program for the calculation of genetic linkage maps in experimental populations of diploid species) when I was in a (university in the United States) doing DNA sequence alignment, linkage mapping and quantitative trait locus (QTL) analysis which was under (the university in the United States) license. When I came back to a (University in Tanzania). I started facing difficulties because the [University in Tanzania] does not have such a program. In addition, there are only a few individuals working on research involving sequencing at the (University in Tanzania). Due to this problem, I had to send back my data to the (University in the United States) for assistance in performing the analysis instead of doing it by myself in the (University in Tanzania)". Another male respondent said, "Yes, I face challenges in bioinformatics. We do not have a well-recognized, reputable center for training on bioinformatics in Tanzania. During our studies, the bioinformatics training was merely an overview and a few practical demonstrations. At least we can do partial sequence analyses on data such as sequence alignment and phylogeny. However, extensive proteomics analysis is still a challenge. Besides, whole-genome sequence analysis is a challenge in many institutions in Tanzania.

Nevertheless, the world is moving toward whole-genome approaches. Therefore, Tanzanian experts need to disseminate their knowledge to their global counterparts. For instance, many PhD students plan to undertake whole-genome analysis in their research at the (University in Tanzania). However, almost all of them plan to go to the International Livestock Research Institute (ILRI) in Kenya to train on bioinformatics and perform whole genome sequencing and analysis".

Usage of Bioinformatics Tools and Genomics and Bioinformatics Analyses Performed by the Respondents

Of the surveyed bioinformatics tools and resources, the seldom-used ones were QIAGEN CLC Main Workbench, where 57 (67.9%) respondents reported that they never used the program. This was followed by the DNA Data Bank of Japan (DDBJ), where 52 (61.9%) never used the resource. The most used resources were BLAST, PubMed and GenBank (**Figure 1**).

The majority, 57 (67.9%) of the surveyed participants, did perform sequence alignment, followed by 42 (50%) who carried out phylogenetics analysis (**Figure 2**).

Software Usage of Statistical Package and Microsoft Office Products by Respondents

Regarding statistical software packages, the least use of statistical software packages was reported by 78 (92.9%) in WinBUGS followed by 73 (86.9%) in MedCalc (**Figure 3**). The frequently used were SPSS and R, where respondents report expert, high and intermediate skills in these tools (**Figure 3**).

On the one hand, respondents reported more Microsoft Word expertise 27 (32.1%), followed by Microsoft PowerPoint 19 (23.2%). On the other hand, less expertise 1 (1.3%) was noted in Microsoft Access (**Figure 4**). These numbers self-reporting high skills are slightly higher for with 44 (52.4%) and 38 (46.3%) and 13 (16.3%), in Microsoft Word, Microsoft PowerPoint and Microsoft Access, respectively (**Figure 4**).

DISCUSSION

This is the first study that assesses the level of bioinformatics capacity in Tanzania to the best of our knowledge. We found out that the majority of the respondents were males, had a master's degree and were in the age group 26–32 years. The mean work experience of the respondents in years was 6.2, indicating a young group of scientists. The highest education level for most respondents was a master's degree, followed by a bachelor's degree. When asked to rate their

TABLE 3 | Challenges that the respondent face in bioinformatics practice in Tanzania.

Question response		N	%
Do you face any challenges in your bioinformatics research	Yes	51	60.7
	No	33	39.3
Challenges faced in bioinformatics research ±	Lack of training and skills	22	26.2
	Lack of reliable computer infrastructure	21	25.0
	Lack of mentorship and network partners	8	9.5
	Insufficient electrical power and poor internet access	2	2.4

Key±: multiple responses were possible for this question.

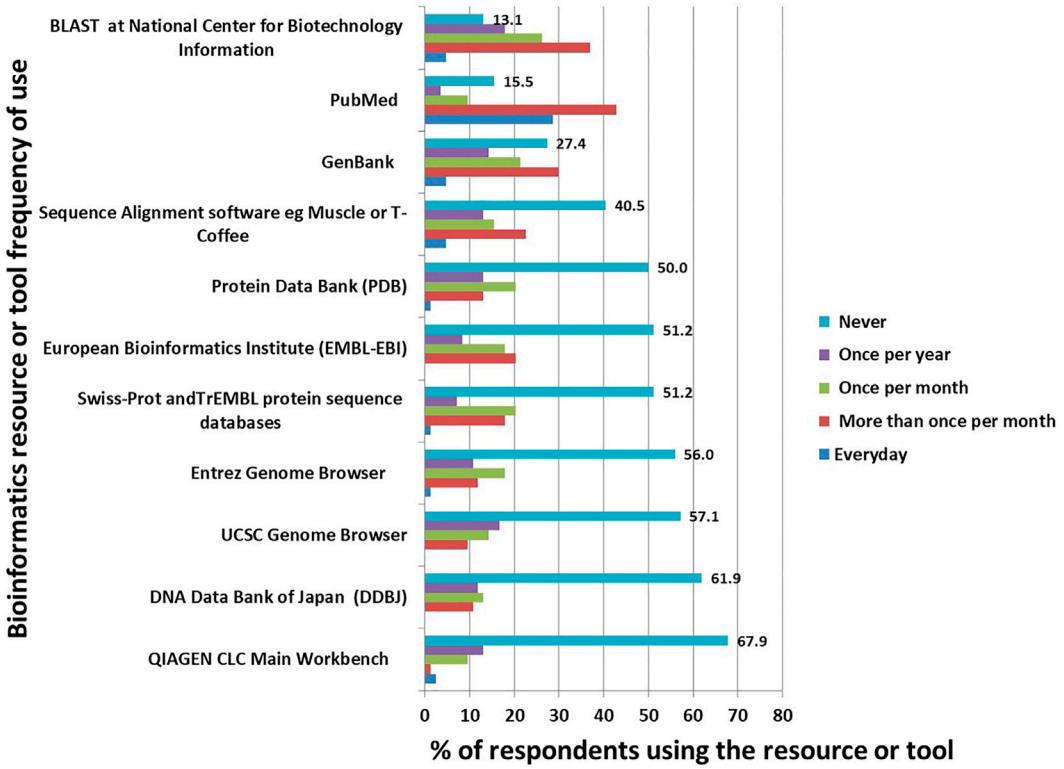


FIGURE 1 | Frequency of use of common bioinformatics resources and tools.

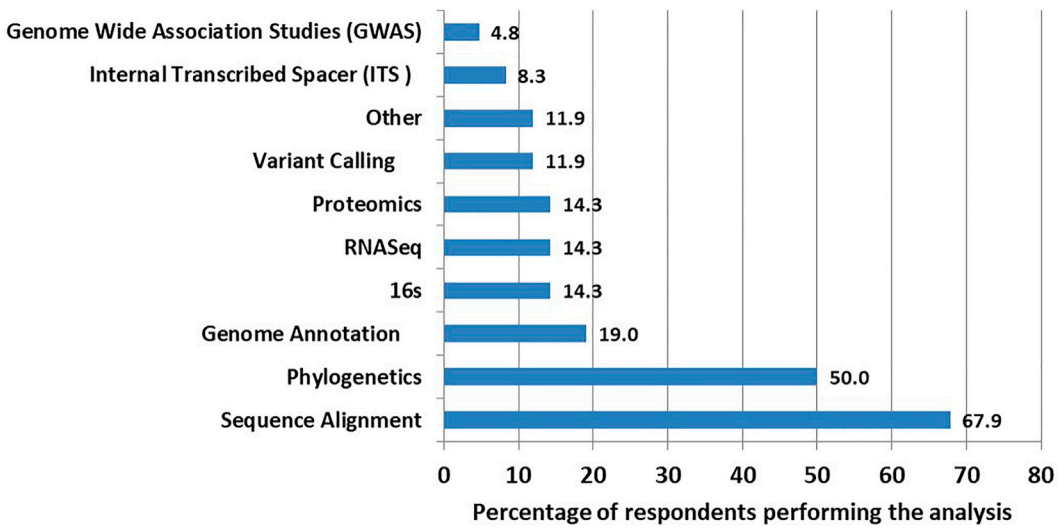


FIGURE 2 | Percentage of analysis done by the respondents (multiple responses possible *N* = 84).

seniority on a scale of 0–100, the respondents rated themselves with mean seniority of 39.1, further indicating the junior ship's perception in the area of bioinformatics practice. Only 21.4% were PhD holders; this pool of scientists can mentor the early-career counterparts. Interestingly, most of the respondents' current specialization area was mostly molecular biology. Only a few related their complete research interest in genomics and bioinformatics, suggesting that molecular biology scientists diversify their careers into bioinformatics.

This survey pointed out that the infrastructure and the human capacity to conduct bioinformatics-related research in Tanzania

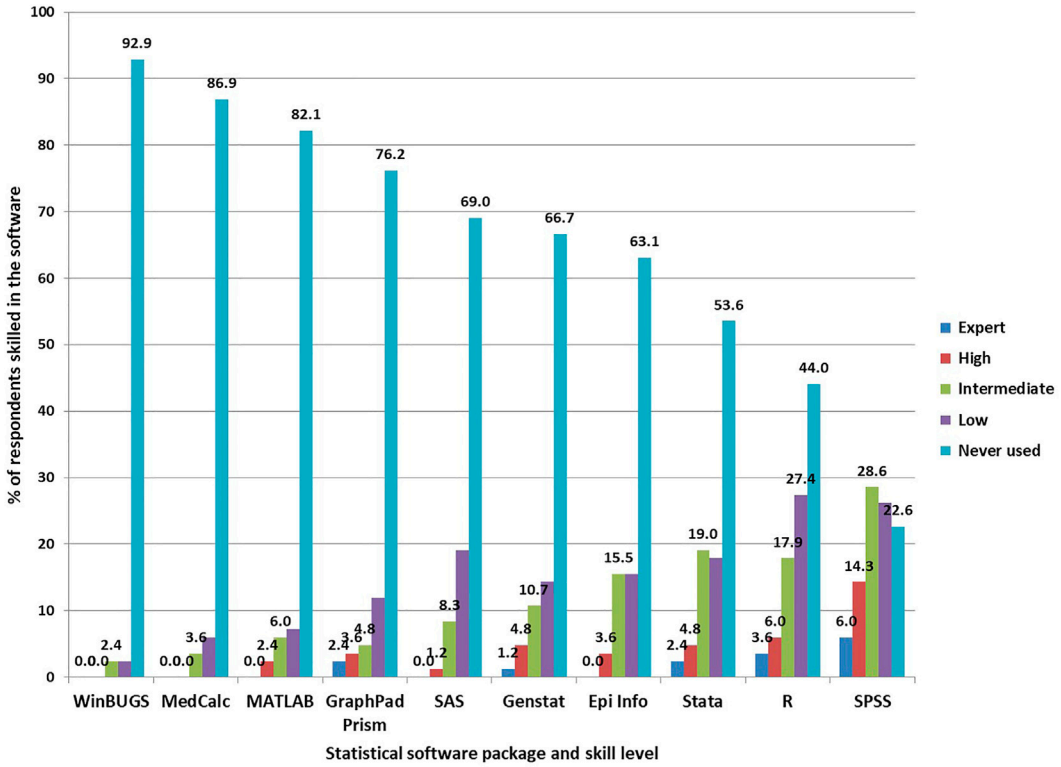


FIGURE 3 | Level of uses of standard statistical software packages.

are underdeveloped. Precisely, 96.4% of the respondents perform bioinformatics analysis using personal computers/laptops, with only about 10% having access to advanced infrastructures such as high-performance computers, cloud computing and institutional servers. Although 40.5% of respondents have access to the institutional computer server, these servers are mainly available to provide file and printing services rather than bioinformatics services.

This severely limits the capacity to conduct bioinformatics-related research. It usually involves massive datasets and requires reliable high computing capacity that personal computers cannot afford alone (Johansen Taber et al., 2014). More than 67% of the respondents use Windows operating system (OS), which does not support many genomics and bioinformatics analysis platforms, contrary to only about 14.3% who use the Linux OS that supports a broad range of bioinformatics analysis tools. However, there is a possibility that respondents using Windows use it to run bioinformatics analysis such as phylogenetics with Windows-based software. The same respondents may also use their personal Windows machines to access online-based tools such as BLAST. There are software programs that efficiently run in Windows MEGA (Kumar et al., 2016) and UGENE (Okonechnikov et al., 2012), JALVIEW (Waterhouse et al., 2009) for protein and DNA alignments. In addition, Windows 10 ships with a Windows Subsystem for Linux (WSL), which provides support to run native Linux command-line tools directly on Windows operating system (https://docs.microsoft.com/en-us/windows/wsl/faq). This has allowed running most of the bioinformatics tools directly on Windows. Nevertheless, it is still necessary to know Linux command lines to use this resource. In addition, some Linux-based packages may be hard to run in this environment.

For most respondents, the usage of standard bioinformatics analysis tools was also low; therefore, it comes as no surprise that 66.7% of the respondents had no publication related to bioinformatics. These findings align with Lyantagaye's (2013) review, which noted that the level of bioinformatics research in Tanzania was still in its infancy, lacking investment and underdeveloped infrastructure. The review noted the presence of one modern laboratory at SUA, capable of generating molecular biology and genomics data. The STM-1 SEACOM undersea fiber-optic cable was expected to increase the internet speed bandwidth (Lyantagaye, 2013). The situation is not unique to Tanzania alone. Karikari (2015) noted a low level of bioinformatics capacity in terms of personnel and infrastructure in Ghana, with frequent electrical power failures, unreliable internet connections, and lack of high-speed computing power being significant infrastructural challenges (Karikari, 2015). In Africa, three countries are responsible for a large fraction of the continent's bioinformatics output; South Africa, Kenya, and Nigeria. The existence of H3ABioNet has, to a large extent, tried to reduce this disparity by empowering other countries in Africa to participate and contribute to bioinformatics (Matovu et al., 2014; Mulder et al., 2016a).

Bioinformatics consists of multidisciplinary fields, including mathematics, computer science, statistics and others. Statistics

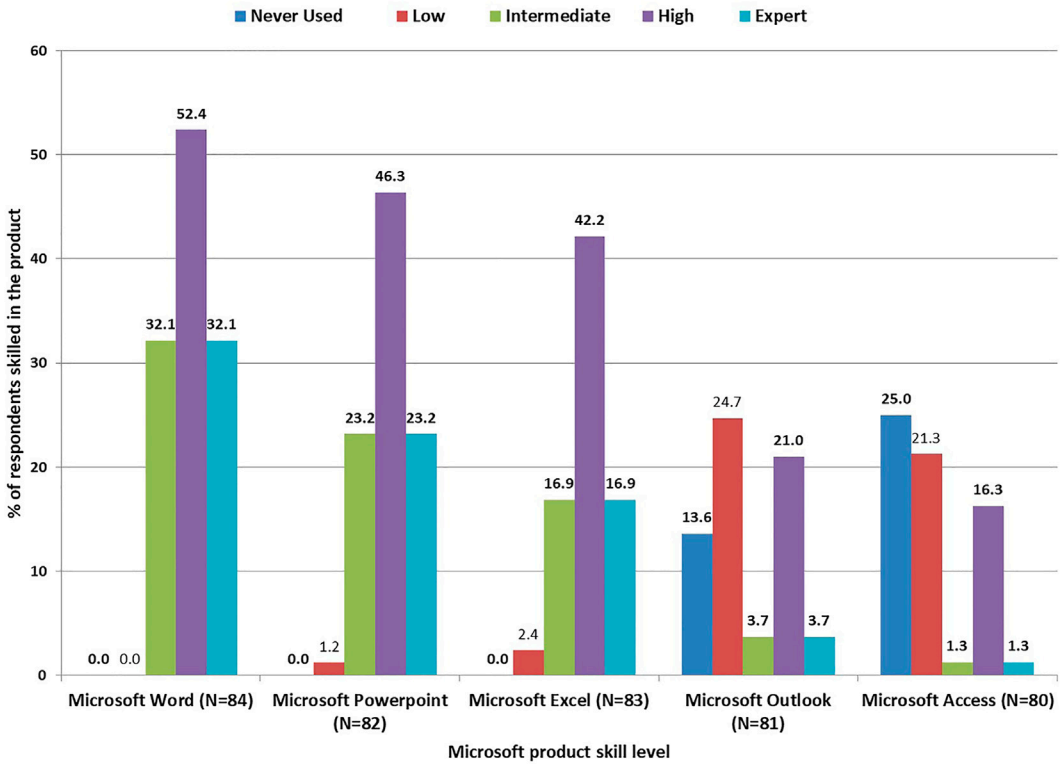


FIGURE 4 | Frequency of usage of Microsoft Office software by respondents from Tanzania.

and programming are among the disciplines that play significant roles in building reproducible methods for biological discovery and validation, especially for complex, high-dimensional data as encountered in genomics. Therefore, assessing the knowledge and level of usage of statistics and programming among the respondents was essential. We found that only a quarter of respondents reported using computer programming language and 17.9% used a database management system. The most used programming language was Python by 8 (9.5%) of the respondents and the database management systems most used were Microsoft Access and MySQL. Both Python and MySQL find wide applications in bioinformatics tools and pipelines (Pasculescu et al., 2014). However, there are a large proportion of respondents without skills in hardcore programming. Short training may help to improve the skills of these researchers. It was also evident that the knowledge and usage of different statistical packages are mainly based on IBM's SPSS package. On the one hand, many respondents are using R statistical packages. On the other hand, packages like WinBUGS and SAS are rarely used by bioinformatics researchers in Tanzania.

In bioinformatics, it is essential that computational thinking is adopted to increase the pool of hardcore programmers. This will facilitate efficient bioinformatics analyses and communication among scientists, bioinformaticians, and data analysts. To this end, short and long-term training are necessary for computer programming such as Python and R statistical package, among others. Other training should focus on database management. These efforts are essential in Tanzania and other African Scientific communities (Gurwitz et al., 2017). Nevertheless, software like Galaxy (Giardine, 2005) offers a potential advantage for non-programmers. Galaxy training can therefore be handy for biologists who undertake bioinformatics analysis.

Our respondents made high use of Microsoft Office Products, particularly Microsoft Word, Microsoft PowerPoint and Microsoft Excel. Only a few individuals made occasional use of Microsoft Access and Microsoft Outlook, again showing less advanced use of these products. These Microsoft Office Products are not essential for running bioinformatics. However, high reliance on Microsoft Office Products indicates an inclination towards using a Windows-based operating system. In addition, the use of Microsoft Access products may be a step for scientists to begin the use of large databases.

There is a wide range of bioinformatics tools and resources that respondents said they could access, with PubMed, which they use to retrieve scientific literature, which is the most popular. PubMed is widely used by the scientific community, not necessarily by the bioinformatics community. However, the responses about PubMed allow us to gauge its use in comparison with other resources that are widely used in the bioinformatics community. The other frequently used resources in this community were GenBank and some sequence alignment tools, showing good progress as users can access relevant and essential resources. Commercial products such as CLC Workbench (a QIAGEN platform for DNA, RNA and protein sequence data analysis) were limited, probably due to a shortage of funding (Smith, 2015).

More than half of the respondents reported one or more problems they face in relation to bioinformatics practice in Tanzania. The

majority of the respondents reported a lack of training and skills as a significant problem. Only a few respondents reported inadequate electrical power supply and lack of internet access as challenges. The reduced cost of internet connectivity and bandwidth improvement has helped other Africa nations improve their bioinformatics infrastructure and capacity (Mulder et al., 2016b). Tanzania has equally benefited from bandwidth improvement, which may be why few respondents cited internet connectivity as a challenge. Capacity building through training and infrastructural support for bioinformatics research remains the major challenge, as noted in other African countries (Karikari, 2015; Karikari et al., 2015; Mulder et al., 2016b; Shoko et al., 2018).

The majority of respondents reported having knowledge of at least two to three bioinformatics skills. The most commonly performed analyses were sequence alignment and phylogenetics. Other methods of analysis, such as GWAS were less commonly used. The most and the least frequent applications may require training modules for long or short-term training to allow scientists to master these critical bioinformatics skills.

In our study, most of the respondents, 40 (47.6%), reported learning bioinformatics at bachelor's degree level, followed by 27 (32.1%) who learned at the masters' training and only 18 (21.4%) during PhD training. Conferences and workshops also serve as essential sources of bioinformatics skills for some respondents (28.6%), while a small percentage (15.5%) used online resources to learn bioinformatics skills. These later may have benefitted from the opportunity provided by the H3ABioNet (Gurwitz et al., 2017) in addition to other training opportunities such as those used in other countries (Cattley and Arthur, 2007; Ding et al., 2014; Vincent et al., 2018).

It is possible that most of the surveyed Tanzanian bioinformatics researchers were either trained abroad or learned bioinformatics through postgraduate research projects. Today, no full bioinformatics or computational biology degree program exists in the country. Bioinformatics courses are part of undergraduate and postgraduate degree programs at the University of Dar es Salaam (UDSM) and Sokoine University of Agriculture (SUA). Two undergraduate courses exist at the UDSM according to the UDSM undergraduate prospectus 2018/2019. Besides, seven postgraduate courses also exist at UDSM according to the 2019/2020 postgraduate prospectus. At SUA, three undergraduate and three postgraduate courses are offered (SUA prospectus 2014/15) (Lyantagaye, 2013). Therefore, it is not surprising that most respondents, 16.7 and 14.3% in this study, are from UDSM and SUA, respectively.

There is a long way to go and an opportunity to fill the expertise gap observed in this survey. For starters, Muhimbili University of Health and Allied Sciences (MUHAS) is preparing to start a Master's of Science in Bioinformatics through collaboration with EANBiT (Eastern Africa Network for Bioinformatics Training) (http://eanbit.icipe.org/). EANBiT has developed a 2-years master's degree curriculum that has been used in training since 2017 and is expected to be adopted by MUHAS in the foreseeable future (http://eanbit.icipe.org/) (EANBiT). This will be important in establishing a critical mass of expertise in bioinformatics and computational biology in Tanzania. Eventually, it may attract grants, research projects, collaborations, and the development of infrastructure necessary to research in the field.

In terms of curriculum development and training establishment, there are examples to learn from other countries such as India and South Africa (Kulkarni-Kale et al., 2010; Mulder et al., 2016b). In the early days of bioinformatics, the discipline was not embedded in undergraduate curricula in South Africa. To address the gap, students registered for postgraduate degrees in bioinformatics in South African Universities had to start with short formal bioinformatics training before embarking on their studies. Later, the National Bioinformatics Network (NBN) developed joint courses compulsory for NBN-funded students, introducing them to a range of bioinformatics topics, programming and other technical skills (Mulder et al., 2016b). In India, similar initiatives were undertaken by the Biotechnology Information System (BTIS) under the Department of Biotechnology (DBT), Government of India (Ding et al., 2014).

Equally in Tanzania, there is also a need to develop relevant skills by extending undergraduate bioinformatics courses to other universities that offer biomedical, life and computer science courses. Students will be exposed to the field early on and potentially incite their interest. It will also prepare them with basic knowledge and skills for postgraduate research and education specializing in bioinformatics education (Bishop et al., 2015). Besides, we advocate for establishing short programs for professionals who may be constrained by time to do a full-fledged degree. This can go hand in hand with existing programs and infrastructure and collaborate with other organizations in Tanzania, Africa and worldwide. EANBiT, for example, offers a residential training course on bioinformatics for East African students and early career researchers (http://eanbit.icipe.org/content/2018-trainees). Other successful training models were in Sudan (Ahmed et al., 2020).

In the era of digital technologies, bioinformatics capacity in Tanzania could greatly benefit from online learning and has to be prioritized. It is less costly, often self-paced and accessible to many people at the same time. Online learning may be more suitable for professionals who cannot spend time in physical classes. Although a multitude of online learning platforms for bioinformatics exist, relevant organizations and institutions have a critical role in developing an appropriate curriculum and mobilizing resources to facilitate the learning process and ensure that online learning is effective. The duration of vast online courses and resources and providing guidelines to learners is also essential.

Collaborative programs with hybrid virtual-physical models have become especially attractive recently, such as the Courses such as the 3-months Introduction to Bioinformatics (IBT) course offered by H3ABioNet (https://www.h3abionet.org/training/ibt)m) (H3ABionet, 2021). The annual system that has been provided since the year 2016 attracted 364 enrolled participants hosted at 20 institutions across 10 African countries in the inaugural year (Gurwitz et al., 2017). In 2020, the course went utterly online due to physical meeting restrictions caused by the pandemic of COVID-19 but still had over 1,000 participants distributed across 40 classrooms in Africa (H3ABioNet newsletter May 2020: https://spark.adobe.com/page/

0OVCv7sPapYfa/). H3ABioNet has also hosted a 16S analysis course since 2019 in a similar manner.

Bioinformatics and computational biology research are expensive to conduct. Establishing collaborations among relevant institutions and stakeholders in Tanzania and with external partners may help develop the necessary infrastructure and conduct research. Collaboration between research institutions, academia, and civil society with similar objectives regarding bioinformatics research catalyzes the field's rapid growth. The recent establishment of the Tanzania Society of Human Genetics (TSHG) (http://tshg.or.tz/) indicates both the need and interest in furthering this critical biological sub-discipline. This will lead to the development of vital programs and improve the competitiveness of funding. In addition to joining Pan African and global networks, Tanzania needs to plan to improve and offer streamlined bioinformatics services. Initiatives of this nature have worked in other countries such as Australia (Schneider et al., 2019; Tauch and Al-Dilaimi, 2019). To build total capacity in bioinformatics, Tanzania needs to work closely with existing bioinformatics networks to strengthen its capacity through training. The H3ABioNet help desk can help African countries quickly grasp the assistance needed to get going to bioinformatics tasks (Kumuthini et al., 2019). Fostering collaboration in bioinformatics will depend on both scientist-led and Government-led initiatives.

The Government has a pivotal role to play by supporting basic infrastructure for education and training as well as for research and application. The Government also plays a crucial role in promoting human capacity building in bioinformatics and computational biology by ensuring that graduates are recognized by the government scheme and get job opportunities. The collaborative approach will help guarantee the sustainability of the initiatives, training, and infrastructure and research activities. Tanzania can emulate examples from other countries where government funding has facilitated bioinformatics (Mulder et al., 2016b; Schneider et al., 2019; Tauch and Al-Dilaimi, 2019). In South Africa, the bioinformatics leader in Africa, the very early phase of bioinformatics at the South African National Bioinformatics Institute (SANBI) on the University of the Western Cape (UWC) campus was co-funded by the Government through the South Africa's National Research Foundation (NRF) (Mulder et al., 2016b). Tanzania and other African countries need to emulate the funding models of SANBI to improve bioinformatics skills and research in their institutions.

The respondents agreed to participate in the bioinformatics network and genomics initiative in Tanzania. The bioinformatics community needs to work with the Government to support a national forum that brings together bioinformaticians and genomics practitioners to discuss common interest issues. Such a forum can already build on the existing platforms such as TGN and the TSHG to facilitate joint meetings and promote a bioinformatics agenda. Similar National platforms have been shown to help build bioinformatics capacity in South Africa, India and Australia (Kulkarni-Kale et al., 2010; Mulder et al., 2016b; Schneider et al., 2019).

CONCLUSION

In this study, we found out that the majority of the respondents engaging in bioinformatics research in Tanzania were at the early stages of their careers. Although there is a high level of interest in bioinformatics in Tanzania, a low level of skilled human resources and the lack of infrastructure pertinent to research in the field are limited. The use of bioinformatics tools for data analysis is still low, even for essential analysis tools such as BLAST (McGinnis and Madden, 2004), GenBank (Clark et al., 2016), sequence alignment software, Swiss-prot (Bairoch, 1996) and TrEMBL (Bairoch, 1996). This may be because most respondents also lacked access to basic tools and resources for bioinformatics research.

Investment in human capacity building through undergraduate and postgraduate training and encouraging and promoting digital learning may help improve the situation. Provision of infrastructure, mentorship and networking is needed to improve bioinformatics capacity in Tanzania. We recommend building strong collaborations among Tanzania institutions to promote the effective utilization of shared resources and expertise. Moreover, regional and global network partners and stakeholders may be crucial in developing infrastructure and research activities and ensuring sustainability. Support from the Government by setting the groundwork and funding basic teaching and research infrastructure is also essential to the growth and success of the field. The launch of a community of practice such as the TSHG of the TGN may help continue the Pan-African efforts to promote the use of bioinformatics for the betterment of humankind.

DATA AVAILABILITY STATEMENT

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

AUTHOR CONTRIBUTIONS

All authors have read and approved the manuscript; RS and AM designed the survey and collected the data; UM and RS performed the statistical analysis and resulted in interpretation; UM, RS, SN, LM, SLL, AM, DM and JM contributed to writing and reviewing the manuscript.

ACKNOWLEDGMENTS

We wish to thank the respondents who took the time to respond to this survey.

This article has been released as a pre-print at https://www.researchsquare.com/article/rs-112131/v1, (Sangeda et al., 2020).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.665313/full#supplementary-material

REFERENCES

Adedokun, B. O., Olopade, C. O., and Olopade, O. I. (2016). Building Local Capacity for Genomics Research in Africa: Recommendations from Analysis of Publications in Sub-saharan Africa from 2004 to 2013. *Glob. Health Action.* 9, 31026. doi:10.3402/gha.v9.31026

Ahmed, A. E., Awadallah, A. A., Tagelsir, M., Suliman, M. A., Eltigani, A., Elsafi, H., et al. (2020). Delivering Blended Bioinformatics Training in Resource-Limited Settings: a Case Study on the University of Khartoum H3ABioNet Node. *Brief. Bioinform.* 21, 719–728. doi:10.1093/bib/bbz004

Bairoch, A. (1996). The SWISS-PROT Protein Sequence Data Bank and its New Supplement TREMBL. *Nucleic Acids Res.* 24, 21–25. doi:10.1093/nar/24.1.21

Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Tastan Bishop, Ö., Adebiyi, E. F., Alzohairy, A. M., Everett, D., Ghedira, K., Ghouila, A., et al. (2015). Bioinformatics Education-Pperspectives and Challenges Out of Africa. *Brief. Bioinform.* 16, 355–364. doi:10.1093/bib/bbu022

Cattley, S., and Arthur, J. W. (2007). BioManager: the Use of a Bioinformatics Web Application as a Teaching Tool in Undergraduate Bioinformatics Training. *Brief. Bioinform.* 8, 457–465. doi:10.1093/bib/bbm039

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2016). GenBank. *Nucleic Acids Res.* 44, D67–D72. doi:10.1093/nar/gkv1276

de Martel, C., Georges, D., Bray, F., Ferlay, J., and Clifford, G. M. (2020). Global burden of Cancer Attributable to Infections in 2018: a Worldwide Incidence Analysis. *Lancet Glob. Heal.* 8, e180–e190. doi:10.1016/S2214-109X(19)30488-7

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

Ding, Y., Wang, M., He, Y., Ye, A. Y., Yang, X., Liu, F., et al. (2014). "Bioinformatics: Introduction and Methods," a Bilingual Massive Open Online Course (MOOC) as a New Example for Global Bioinformatics Education. *Plos Comput. Biol.* 10, e1003955. doi:10.1371/journal.pcbi.1003955

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

Giardine, B. (2005). Galaxy: A Platform for Interactive Large-Scale Genome Analysis. *Genome Res.* 15, 1451–1455. doi:10.1101/gr.4086505

Gurwitz, K. T., Aron, S., Panji, S., Maslamoney, S., Fernandes, P. L., Judge, D. P., et al. (2017). Designing a Course Model for Distance-Based Online Bioinformatics Training in Africa: The H3ABioNet Experience. *PLOS Comput. Biol.* 13, e1005715. doi:10.1371/journal.pcbi.1005715

H3ABionet (2021). *H3ABioNet - a Pan African Bioinformatics Network for the Human Heredity and Health in Africa (H3Africa) Consortium*. Available at: https://www.h3abionet.org/(Accessed May 4, 2021).

H3Africa (2021a). *EANBiT Eastern Africa Network for Bioinformatics Training*. Available at: http://eanbit.icipe.org/news/eastern-africa-network-%28eanbit%29-sample-%0Aarticle-1 (Accessed July 1, 2020).

H3Africa (2021b). *H3Africa Human Heredity for Health Africa*. Available at: https://h3africa.org/(Accessed July 1, 2020).

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., and Conde, J. G. (2009). Research Electronic Data Capture (REDCap)-A Metadata-Driven Methodology and Workflow Process for Providing Translational Research Informatics Support. *J. Biomed. Inform.* 42, 377–381. doi:10.1016/j.jbi.2008.08.010

Harris, P. A., Taylor, R., Minor, B. L., Elliott, V., Fernandez, M., O'Neal, L., et al. (2019). The REDCap Consortium: Building an International Community of Software Platform Partners. *J. Biomed. Inform.* 95, 103208. doi:10.1016/J.JBI.2019.103208

Hernández-de-Diego, R., de Villiers, E. P., Klingström, T., Gourlé, H., Conesa, A., Bongcam-Rudloff, E., et al. (2017). The eBioKit, a Stand-Alone Educational Platform for Bioinformatics. *PLOS Comput. Biol.* 13, e1005616. doi:10.1371/journal.pcbi.1005616

Johansen Taber, K. A., Dickinson, B. D., and Wilson, M. (2014). The Promise and Challenges of Next-Generation Genome Sequencing for Clinical Care. *JAMA Intern. Med.* 174, 275. doi:10.1001/jamainternmed.2013.12048

Karikari, T. K., Quansah, E., and Mohamed, W. M. Y. (2015). Developing Expertise in Bioinformatics for Biomedical Research in Africa. *Appl. Translational Genomics* 6, 31–34. doi:10.1016/j.atg.2015.10.002

Karikari, T. K. (2015). Bioinformatics in Africa: The Rise of Ghana? *Plos Comput. Biol.* 11, e1004308. doi:10.1371/journal.pcbi.1004308

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., et al. (2002). The Human Genome Browser at UCSC. *Genome Res.* 12, 996–1006. doi:10.1101/gr.229102

Kulkarni-Kale, U., Sawant, S., and Chavan, V. (2010). Bioinformatics Education in India. *Brief. Bioinform.* 11, 616–625. doi:10.1093/bib/bbq027

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

Kumuthini, J., Zass, L., Zass, L., Panji, S., Salifu, S. P., Kayondo, J. K., et al. (2019). The H3ABioNet Helpdesk: an Online Bioinformatics Resource, Enhancing Africa's Capacity for Genomics Research. *BMC Bioinformatics* 20, 741. doi:10.1186/s12859-019-3322-3

Li, W., Cowley, A., Uludag, M., Gur, T., McWilliam, H., Squizzato, S., et al. (2015). The EMBL-EBI Bioinformatics Web and Programmatic Tools Framework. *Nucleic Acids Res.* 43, W580–W584. doi:10.1093/nar/gkv279

Lyantagaye, S. (2013). Current Status and Future Perspectives of Bioinformatics in Tanzania. *Tanzania J. Sci.* 39, 1–11. doi:10.4314/tjs.v39i1

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

Mashima, J., Kodama, Y., Fujisawa, T., Katayama, T., Okuda, Y., Kaminuma, E., et al. (2017). DNA Data Bank of Japan. *Nucleic Acids Res.* 45, D25–D31. doi:10.1093/nar/gkw1001

Matovu, E., Bucheton, B., Chisi, J., Enyaru, J., Hertz-Fowler, C., Koffi, M., et al. (2014). Enabling the Genomic Revolution in Africa. *Sci. (80-* 344, 1346–1348. doi:10.1126/science.1251546

McGinnis, S., and Madden, T. L. (2004). BLAST: at the Core of a Powerful and Diverse Set of Sequence Analysis Tools. *Nucleic Acids Res.* 32, W20–W25. doi:10.1093/nar/gkh435

Mulder, N. J., Adebiyi, E., Alami, R., Benkahla, A., Brandful, J., Doumbia, S., et al. (2016a). H3ABioNet, a Sustainable Pan-African Bioinformatics Network for Human Heredity and Health in Africa. *Genome Res.* 26, 271–277. doi:10.1101/gr.196295.115

Mulder, N. J., Christoffels, A., de Oliveira, T., Gamieldien, J., Hazelhurst, S., Joubert, F., et al. (2016b). The Development of Computational Biology in South Africa: Successes Achieved and Lessons Learnt. *PLOS Comput. Biol.* 12, e1004395. doi:10.1371/journal.pcbi.1004395

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., et al. (2017). Development of Bioinformatics Infrastructure for Genomics Research. *gh* 12, 91. doi:10.1016/j.gheart.2017.01.005

Okonechnikov, K., Golosova, O., and Fursov, M. (2012). Unipro UGENE: a Unified Bioinformatics Toolkit. *Bioinformatics* 28, 1166–1167. doi:10.1093/bioinformatics/bts091

Ooijen, J. W. (2021). *JoinMap ® 5 Software for the Calculation of Genetic Linkage Maps in Experimental Populations of Diploid Species*. Available at: https://www.kyazma.nl/index.php/JoinMap/(Accessed May 4, 2021).

Ossom Williamson, P., and Minter, C. I. J. (2019). Exploring PubMed as a Reliable Resource for Scholarly Communications Services. *jmla* 107, 16–29. doi:10.5195/JMLA.2019.433

Pasculescu, A., Schoof, E. M., Creixell, P., Zheng, Y., Olhovsky, M., Tian, R., et al. (2014). CoreFlow: A Computational Platform for Integration, Analysis and Modeling of Complex Biological Data. *J. Proteomics* 100, 167–173. doi:10.1016/j.jprot.2014.01.023

R Development Core Team (2020). *R: A Language and Environment for Statistical Computing*. Available at: http://www.r-project.org/ (Accessed April 1, 2020).

Sangeda, R. Z., Mwakilili, A. D., Masamu, U., Nkya, S., Mwita, L. A., Massawe, D. P., et al. (2020). Baseline Evaluation of Bioinformatics Capacity in Tanzania. [Epub ahead of print]. doi:10.21203/RS.3.RS-112131/V1

Sangeda, R. Z., Mwakilili, A. D., Masamu, U., Nkya, S., Mwita, L. A., Massawe, D. P., et al. (2021). *Dataset and Supplementary Materials for Baseline Evaluation of Bioinformatics Capacity in Tanzania in 2018*. Mendeley Data doi:10.17632/t79ddvj48j.1

Schneider, M. V., Griffin, P. C., Tyagi, S., Flannery, M., Dayalan, S., Gladman, S., et al. (2019). Establishing a Distributed National Research Infrastructure

Providing Bioinformatics Support to Life Science Researchers in Australia. *Brief. Bioinform.* 20, 384–389. doi:10.1093/bib/bbx071

Shoko, R., Manasa, J., Maphosa, M., Mbanga, J., Mudziwapasi, R., Nembaware, V., et al. (2018). Strategies and Opportunities for Promoting Bioinformatics in Zimbabwe. *PLOS Comput. Biol.* 14, e1006480. doi:10.1371/journal.pcbi.1006480

Smith, D. R. (2015). Buying in to Bioinformatics: an Introduction to Commercial Sequence Analysis Software. *Brief. Bioinform.* 16, 700–709. doi:10.1093/bib/bbu030

Tauch, A., and Al-Dilaimi, A. (2019). Bioinformatics in Germany: toward a National-Level Infrastructure. *Brief. Bioinform.* 20, 370–374. doi:10.1093/bib/bbx040

Vincent, A. T., Bourbonnais, Y., Brouard, J.-S., Deveau, H., Droit, A., Gagné, S. M., et al. (2018). Implementing a Web-Based Introductory Bioinformatics Course for Non-bioinformaticians that Incorporates Practical Exercises. *Biochem. Mol. Biol. Educ.* 46, 31–38. doi:10.1002/bmb.21086

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2–a Multiple Sequence Alignment Editor and Analysis Workbench. *Bioinformatics* 25, 1189–1191. doi:10.1093/bioinformatics/btp033

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 Education

OPINION
published: 27 July 2021
doi: 10.3389/feduc.2021.710971

Challenges and Considerations for Delivering Bioinformatics Training in LMICs: Perspectives From Pan-African and Latin American Bioinformatics Networks

*Verena Ras[1], Patricia Carvajal-López[2], Piraveen Gopalasingam[2], Alice Matimba[2], Paballo Abel Chauke[1], Nicola Mulder[1], Fatma Guerfali[3], Victoria Dominguez Del Angel[4], Alejandro Reyes[5], Guilherme Oliveira[6], Javier De Las Rivas[7] and Marco Cristancho[8]**

[1]Computational Biology Division, Department of Integrative Biomedical Sciences, Institute of Infectious Disease and Molecular Medicine, CIDRI Africa Wellcome Trust Centre, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa, [2]European Molecular Biology Laboratory, European Bioinformatics Institute, Hinxton, United Kingdom, [3]Institut Pasteur de Tunis, Tunis, Tunisia, [4]INRAe, Centre de Recherche de Versailles, Institut Français de Bioinformatique, UMS3601-CNRS, Paris, France, [5]Max Planck Tandem Group in Computational Biology, Department of Biological Sciences, Universidad de Los Andes, Bogotá, Colombia, [6]Instituto Tecnologico Vale, Belem, Brazil, [7]Cancer Research Center (CiC-IBMCC), Consejo Superior de Investigaciones Científicas (CSIC) and Universidad de Salamanca (USAL), Campus Miguel de Unamuno s/n, Salamanca, Spain, [8]Vicerrectoria de Investigación y Creación, Universidad de Los Andes, Bogotá, Colombia

Keywords: training, low to middle income countries, best practices, Africa, Latin America

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Laurent Emmanuel Dardenne, National Laboratory for Scientific Computing (LNCC), Brazil
Ana Carolina Guimarães, Oswaldo Cruz Institute(FIOCRUZ), Brazil

****Correspondence:***
Marco Cristancho
ma.cristancho29@uniandes.edu.co

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *17 May 2021*
Accepted: *12 July 2021*
Published: *27 July 2021*

Citation:
Ras V, Carvajal-López P, Gopalasingam P, Matimba A, Chauke PA, Mulder N, Guerfali F, Del Angel VD, Reyes A, Oliveira G, De Las Rivas J and Cristancho M (2021) Challenges and Considerations for Delivering Bioinformatics Training in LMICs: Perspectives From Pan-African and Latin American Bioinformatics Networks. Front. Educ. 6:710971. doi: 10.3389/feduc.2021.710971

INTRODUCTION

In general, institutions and research groups based in Low to Middle Income Countries (LMICs) battle a number of challenges ranging from a shortage of infrastructure, lack of training facilities, and poor internet, to a lack of local expertise (Karikari et al., 2015; Tastan Bishop et al., 2015; Shaffer et al., 2019). A shortage of local experts, particularly within more specialised topics, still remains a key obstacle in developing bioinformatics research capacity within LMICs (Tastan Bishop et al., 2015; Mulder et al., 2018).

Over the last decade, some barriers experienced by scientists in LMICs have been lowered, in part, by increasing access to more reliable internet, facilities and infrastructure (De Las Rivas et al., 2019; Mulder and H3ABioNet, 2020). This has facilitated the analysis of increasingly complex datasets, which has led to a growing deficit in scientists able to perform these analyses. Access to high quality, affordable training resources and well-trained trainers is typically also a major challenge. Training within LMICs, similar to any context, requires a well-planned approach and often depends on a unique understanding of their specific challenges in order to provide effective training. It is thus useful to outline what we consider to be a range of best practices that may be followed when designing and delivering training in LMICs that may significantly increase a training event's chance of success and lasting impact. We focus here on short term training (workshops, hackathons or data analysis jamborees, etc.) but a number of the best practices provided can be broadly applied to degree modules–in both undergraduate and postgraduate programmes (Fatumo et al., 2014).

We have curated a set of *Best Practices for delivering Bioinformatics Training in LMICs*, with an assemblage of guidelines available as a "living document" on GitHub. The "living document" is based on our own experiences in different settings, where we aim to highlight some of the major challenges that organizers and trainers might face when delivering bioinformatics training in LMICs. These challenges have been divided into three key categories/themes i.e., 1) Planning, 2) Development and

3) Implementation, which loosely represent the three major phases of organizing and delivering a training event, and provides some potential solutions, guidelines or reusable templates.

PLANNING

Training should be tailored for relevance and application to topics of interest in the country or region where the training is taking place rather than a broad bioinformatics subject. Given the particular locations and interests of LMICs, some of the subjects that might be of great interest include biodiversity, rare and communicable diseases, tropical diseases, and native crops research. It is always possible to incorporate social aspects in technical subjects such as genomics and bioinformatics and organizers should make an effort to achieve that as a goal for all training events.

A key challenge in LMICs, even where knowledge, skills and capacity are available locally, is that there may be limited financial or institutional support. Organizers require buy-in and commitment from key stakeholders to support training development and delivery. Even a small/short course may require close involvement of the heads of institutions, government ministries and political representatives who can influence decisions that provide money and infrastructure. The involvement of public agencies will also have the benefit of contextualising the training to local and regional socio-economic environments.

Unlike in High Income Countries (HICs), well-renowned institutions, societies and even local governments might play a significant role in the support and organization of training activities in LMICs, even in such a technical subject as bioinformatics. This is very convenient for funding of the event, infrastructure, involvement of public institutions and even the inclusion of trainers who have been involved in policy making and other global treaties. For instance, training organizers and scientists can influence incorporation of relevant training in programmes aimed at building capacity for biodiversity. Equally, training on use of human data should include information about local data sharing policies and ethics.

Regionally, LMICs show variable degrees of development in bioinformatics training. For instance Brazil, Chile, Argentina, México and Colombia have relatively advanced undergraduate and postgraduate bioinformatics programs compared to other countries in Latin America, which are lagging behind in capacity for delivering training (De Las Rivas et al., 2019). Therefore it is imperative to collaborate with scientists from countries which need support for establishing training initiatives. This is particularly important to reinforce the United Nations' Sustainable Development Goals (SDGs, 2015) where inclusive and equitable education is key to reaching both gender and race equity, globally.

Having explored the above issues and once training goals have been set, a project plan is useful for effective implementation of the training. An ideal development team should map out the tasks required to effectively deliver the training on time and within a proposed budget. Training or scientific development involves leadership in the development of resources, setting the timelines and developing the training agenda. Event administration is important for overall logistical organization and processes. Technical expertise is required for bioinformatics training for ensuring infrastructure, training resources and softwares are prepared and functional. Tasks such as travel arrangements can be outsourced if needed but can be dealt with "in-house" where enough logistical expertise or guidance is available.

DEVELOPMENT

A number of training organizations provide information about their courses and often even provide access to the training materials from their courses. It is good to explore what these organizations are doing successfully and if appropriate, reach out for advice, guidance or support in the form of collaborative funding or delivery. H3ABioNet for example, administers a free to use public helpdesk (https://helpdesk.h3abionet.org/) where a variety of queries can be submitted. One could look at the models being used, content focused on and determine whether any of their approaches may work for your purposes. CABANA is also developing e-learning resources relevant for Latin American research scientists and this material will be available in Spanish (https://www.cabana.online/elearning). Some organizations and conference organizers partner with regional teams to develop workshops linked to conferences offering value addition to attendees and maximizing on the convergence of a specific target audience for scientific topics.

Securing the most appropriate instructors is important and has implications for funding and training format. When holding a regional course in LMICs, where international trainers are involved, the input of local scientists and experts should be a primary consideration. Firstly, local instructors have a sound appreciation of the target audience, they understand local research interests and needs and are comfortable with the local languages/dialects. Secondly, local scientists have the unique opportunity to strengthen and grow local capacity for bioinformatics training by collaborating and, working with guest trainers or international experts, who typically have scarce or special skills, thus increasing sustainability for training and research locally. It is important to acknowledge that representation matters, especially along gender, ethnicity, expertise, and geography. A diverse trainer team can make your course look more attractive, encouraging higher registrations.

It is vital that course organizers provide support, resources, and guidance to course developers/contributors/instructors throughout the development process. It is often useful to conceptualize your training first using a brainstorming document or perhaps a planning template like the one developed by H3ABioNet (this and other templates can be accessed and downloaded at this link).

Language can be a major barrier where trainees are not fluent in the language of the event. An option is to provide language support during the course by having instructors or additional

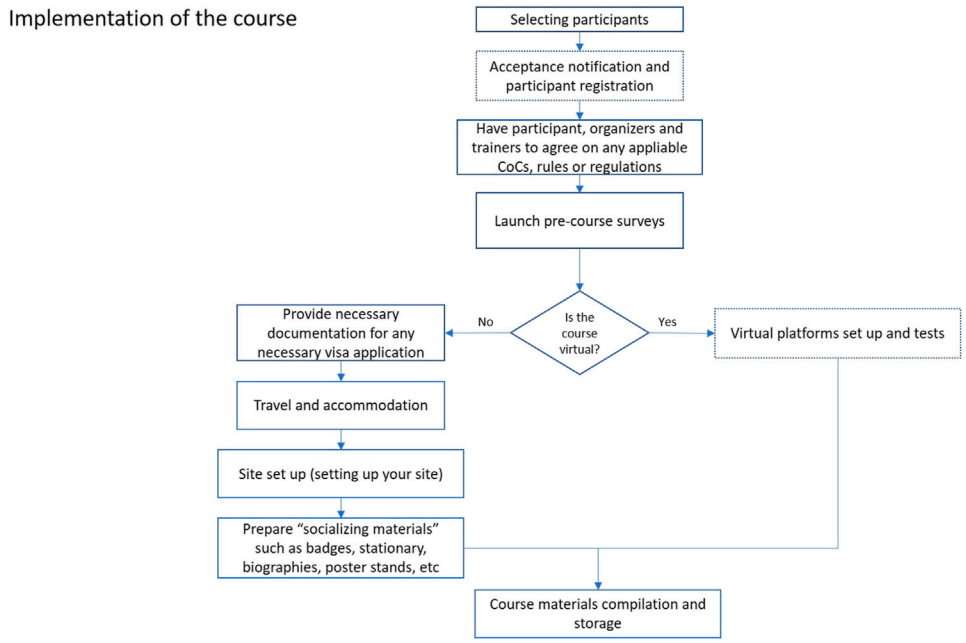


FIGURE 1 | Flowchart for implementing a bioinformatics training event. Diagram has been taken directly from the "Best Practices for delivering Bioinformatics Training in LMICs" living document and represents the set of steps advised by the contributors for implementation of a training event.

assistants who speak the local language. It is highly recommended to provide introductory courses in the local language of the trainees; introducing complex topics such as genomics and bioinformatics. It is recommended to reduce costs, to use reference materials such as articles, books and databases which have open access licenses.

IMPLEMENTATION

In **Figure 1** we describe the main steps that must be followed for implementation of a workshop/course. Selection of participants should strongly consider gender and geographical diversity, ensuring that training can reach diverse participants in the region. A key challenge for face-to-face courses for travelling participants however, is the visa application process - this step can be one of the lengthiest in the whole process of organizing a course. For trainees, a visitor visa is usually sufficient for participating in the workshop/course; for trainers some institutions require a work visa, even for short training courses. Venues for training events must preferably be chosen within close proximity to the accommodation of trainers and trainees. Transportation in most large and middle-size cities can be rather problematic and public transport might not be the best option due to issues like safety and long commute times.

A final consideration is the socio-political situation that must be taken into account when organizing any training in LMICs. In some LMICs civil unrest, political and economic instability, means that safety and infrastructure circumstances may change over short periods of time. It is advisable for organizers to insure any costs that might be reimbursable in case the training event has to be postponed or cancelled. Contingency plans should be made for running events with less infrastructural capacity, such as having pre-recorded content or non-computer-based exercises. Offline solutions like the eBioKit (http://www.ebiokit.eu/) should also be considered in regions with unreliable internet access.

CONCLUDING REMARKS

It is noted by the authors of this piece, that the advent of blended/online learning approaches holds the potential to resolve many of the points raised here. Some general guidelines for virtual bioinformatics teaching have been published given the recent Covid-19 pandemic world lockdown (Gallardo-Alba et al., 2021). H3ABioNet likewise, have developed a comprehensive "multiple-delivery-approach training model" that combines both physical face-to-face style classrooms (situated in any country) with online learning (using a Learning Management System to manage trainees/local staff), distance learning (trainers connect from anywhere in the world to participants and staff in distributed classrooms) and open educational resources (OER) (all materials developed under a creative commons [cc-by] license and publicly accessible). This model has helped H3ABioNet reach upwards of 1,000 participants annually (see Ras et al., 2021). While H3ABioNet has extended this model to deliver more advanced, data intensive training, it must be noted that although blended and online learning can result in higher participation/uptake of the training, there are still many limitations. Highly advanced/technical topics are still difficult to teach as participants or institutions may lack ongoing local

capacity or support. It is also the experience of all authors that the practical/hands-on components of highly advanced topics become increasingly difficult to teach well in an online or hybrid environment, while skills transfer between trainers, trainees and (potential) international trainers is typically greatly reduced. Thus, there currently is no complete substitute for face-to-face/hands-on workshops.

Bioinformatics training in LMICs have a lasting positive impact if done well, ranging from individual and institutional capacity development to better research, more collaboration, and increased funding for activities. Training organizers and developers should tailor content, format and style to suit the local or regional needs and environment. This can be strengthened by community building efforts during and after training events which supports building and maintaining a critical mass of skilled scientists.

A vital consideration that anyone endeavouring to deliver bioinformatics training must at all times keep in mind, is availability of materials, resources and tools post-course. Some LMICs have enhanced local infrastructure, however, many institutions and regions exist where infrastructure has not progressed. This means trainees may not have access to appropriate resources and tools after a course ends to perform real analyses. This perhaps remains one of the most challenging obstacles to overcome but designing a course with this in mind often leads to a more well-received course and may mitigate some of the infrastructural challenges. It often also fosters creativity and solution-driven thinking locally. Finally, please refer to our "Best Practices for delivering Bioinformatics Training in LMICs (link)" living document, for the perspectives that we have highlighted in this opinion article and to broaden the guidelines that we suggest you consider when delivering bioinformatics training in LMICs.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

The authors who contributed to this manuscript are funded in whole, or in part, by the Wellcome Trust (WT108749/Z/15/Z), (WT108749/Z/15/A), H3ABioNet NIH grant U24HG006941, H3ABioNet NIH grant U24HG006942, CABANA—Capacity building for bioinformatics in Latin America' (CABANA), funded by UKRI-BBSRC on behalf of the Global Challenges Research Fund (BB/P027849/1), CSIC/USAL and EMBL-EBI.

ACKNOWLEDGMENTS

The authors would like to acknowledge the many contributors of templates, case studies and experiences to the larger body of work contained on Github and who may not necessarily be authors on this particular paper. A full list of contributors to the larger project can be accessed on the Github repository (https://bioinfo-training.github.io/guidelines/).

REFERENCES

De Las Rivas, J., Bonavides-Martínez, C., and Campos-Laborie, F. J. (2019). Bioinformatics in Latin America and SoIBio Impact, a Tale of Spin-Off and Expansion Around Genomes and Protein Structures. *Brief. Bioinform.* 20 (2), 390–397. doi:10.1093/bib/bbx064

Fatumo, S., Shome, S., and Macintyre, G. (2014). Workshops: A Great Way to Enhance and Supplement a Degree. *Plos Comput. Biol.* 10 (2), e1003497. doi:10.1371/journal.pcbi.1003497

Gallardo-Alba, C., Grüning, B., and Serrano-Solano, B. (2021). A Constructivist-Based Proposal for Bioinformatics Teaching Practices during Lockdown. *Plos Comput. Biol.* 17 (5), e1008922. doi:10.1371/journal.pcbi.1008922

Karikari, T. K., Quansah, E., and Mohamed, W. M. Y. (2015). Developing Expertise in Bioinformatics for Biomedical Research in Africa. *Appl. Translational Genomics* 6, 31–34. doi:10.1016/j.atg.2015.10.002

Mulder, N., Abimiku, A. L., Adebamowo, S. N., de Vries, J., Matimba, A., Olowoyo, P., et al. (2018). H3Africa: Current Perspectives. *Pharmgenomics Pers Med.* 11, 59–66. doi:10.2147/PGPM.S141546

Mulder, N. J., and H3ABioNet (2020). H3ABioNet, Infrastructure for African Genomics Data. *Gates Open Res.* 4, 132. doi:10.21955/gatesopenres.1116670.1

Ras, V., Botha, G., Aron, S., Lennard, K., Allali, I., Claassen-Weitz, S., et al. (2021). Using a Multiple-Delivery-Mode Training Approach to Develop Local Capacity and Infrastructure for Advanced Bioinformatics in Africa. *Plos Comput. Biol.* 17 (2), e1008640. doi:10.1371/journal.pcbi.1008640

SDGs, U. N. (2015). United Nations Sustainable Development Goals. Available at: https://sdgs.un.org/goals (Accessed March 1, 2021).

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

Tastan Bishop, O., Adebiyi, E. F., Alzohairy, A. M., Everett, D., Ghedira, K., Ghouila, A., et al. (2015). Bioinformatics Education-Pperspectives and Challenges Out of Africa. *Brief. Bioinform.* 16 (2), 355–364. doi:10.1093/bib/bbu022

Conflict of Interest: The authors declare that Patricia Carvajal-López is a coordinator of the Original Strategies for Training and Educational Initiatives in Bioinformatics Research Topic. All other authors have declared no further conflicts of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

PERSPECTIVE
published: 28 July 2021
doi: 10.3389/feduc.2021.711403

Integrating Bioinformatics Tools Into Inquiry-Based Molecular Biology Laboratory Education Modules

*Carlos C. Goller[1,2†], Melissa C. Srougi[1,3†], Stefanie H. Chen[1,2†], Laura R. Schenkman[1] and Robert M. Kelly[1,4]**

[1]Biotechnology (BIT) Program, North Carolina State University, Raleigh, NC, United States, [2]Department of Biological Sciences, College of Sciences, North Carolina State University, Raleigh, NC, United States, [3]Department of Molecular Biomedical Sciences, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, United States, [4]Department of Chemical and Biomolecular Engineering, North Carolina State University, Raleigh, NC, United States

OPEN ACCESS

Edited by:
Raquel Cardoso de Melo Minardi, Minas Gerais State University, Brazil

Reviewed by:
Renato Augusto Corrêa Dos Santos, State University of Campinas, Brazil
Stephan Daetwyler, University of Texas Southwestern Medical Center, United States

***Correspondence:**
Robert M. Kelly
rmkelly@ncsu.edu

ORCID:
Melissa C. Srougi
orcid.org/0000-0003-2183-0233
Stefanie H. Chen
orcid.org/0000-0002-9550-971X
Carlos C. Goller
orcid.org/0000-0002-2013-0334
Robert M. Kelly
orcid.org/000-0002-0639-3592

†These authors have contributed equally to this work

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: 18 May 2021
Accepted: 19 July 2021
Published: 28 July 2021

Citation:
Goller CC, Srougi MC, Chen SH, Schenkman LR and Kelly RM (2021) Integrating Bioinformatics Tools Into Inquiry-Based Molecular Biology Laboratory Education Modules. Front. Educ. 6:711403. doi: 10.3389/feduc.2021.711403

The accelerating expansion of online bioinformatics tools has profoundly impacted molecular biology, with such tools becoming integral to the modern life sciences. As a result, molecular biology laboratory education must train students to leverage bioinformatics in meaningful ways to be prepared for a spectrum of careers. Institutions of higher learning can benefit from a flexible and dynamic instructional paradigm that blends up-to-date bioinformatics training with best practices in molecular biology laboratory pedagogy. At North Carolina State University, the campus-wide interdisciplinary Biotechnology (BIT) Program has developed cutting-edge, flexible, inquiry-based Molecular Biology Laboratory Education Modules (MBLEMs). MBLEMs incorporate relevant online bioinformatics tools using evidenced-based pedagogical practices and in alignment with national learning frameworks. Students in MBLEMs engage in the most recent experimental developments in modern biology (e.g., CRISPR, metagenomics) through the strategic use of bioinformatics, in combination with wet-lab experiments, to address research questions. MBLEMs are flexible educational units that provide a menu of inquiry-based laboratory exercises that can be used as complete courses or as parts of existing courses. As such, MBLEMs are designed to serve as resources for institutions ranging from community colleges to research-intensive universities, involving a diverse range of learners. Herein, we describe this new paradigm for biology laboratory education that embraces bioinformatics as a critical component of inquiry-based learning for undergraduate and graduate students representing the life sciences, the physical sciences, and engineering.

Keywords: bioinformatics, science education, molecular biotechnology, software tools, case studies

INTRODUCTION

Students benefit from inquiry-based laboratory courses (Myers and Burgess, 2003; Wallace et al., 2003; Weaver et al., 2008) and several initiatives have formed to help instructors implement these laboratory experiences in undergraduate and graduate curricula [CUR (CUR, 2021), CUREnet (CUREnet, 2021), etc.]. However, transitioning a lab course to the inquiry-based format requires an extensive commitment of time and resources. To lower the entry barrier for participation, pre-structured labs, with open research questions, could be centrally produced and distributed.

The Biotechnology Program (BIT) at North Carolina State University (NC State) (www.ncsu.edu/biotechnology) offers cutting-edge laboratory experiences for students coming from eight colleges. In addition to full-time teaching faculty, teaching postdoctoral scholars (Chen and Goller, 2019) are critical to the Program. As part of their training, teaching postdocs design, implement, and assess novel laboratory courses, based on their research expertise, ranging from metagenomics to protein sciences to organoids. These Molecular Biology Laboratory Education Modules (MBLEMs) are implemented through several iterations, and course details are subsequently published in relevant education journals. This paradigm exposes students to the latest in the biomolecular sciences, including hands-on training in cutting-edge techniques in the context of research questions, while allowing postdoctoral scholars to gain valuable teaching experience as well as develop their research agenda.

MBLEM creation involves iteration of a procedure we define as the "5D Process": Designation, Design, Development, Deployment, and Dissemination (**Supplementary Figure S1**). After a topic has been Designated (i.e. committed to), approximately 1 year is taken for Design and Development of the course, with the instructor concomitantly participating in Deploying existing MBLEMs to understand the framework and possibilities. This Design includes "backwards design" (Wiggins and McTighe, 1998) of course structure and assessments from Designated learning outcomes and Development includes piloting of the proposed experiments and assessments, often by undergraduate student researchers during the academic year or summer months. The course is then Deployed by the instructor and a graduate teaching assistant for a small cohort (typically twelve students). Depending on the results of the Deployment, additional Design and Development, followed by another round of Deployment, may be needed before Dissemination, which includes publication in an appropriate educational journal. Reasons for additional rounds of Development include activities not working as intended, or assessments revealing that scaffolding of the material was insufficient to achieve the learning objectives. Through the 5D process, MBLEMs have been created and disseminated over the years (Witherow and Carson, 2011; Srougi and Carson, 2013; Ott and Carson, 2014; Lentz et al., 2017; Chen and Goller, 2020; Goller and Ott, 2020; Samsa et al., 2020; Garcia et al., 2021), ranging in topics from protein-protein interactions, signal transduction to metagenomics. This model has proven effective as demonstrated by the successful achievement of student learning outcomes and technical skills acquisition in these courses as assessed from quantitative course data (i.e., lab reports, exams, lab notebooks, and projects) and through analysis of qualitative student survey data.

MBLEMS IN FOCUS

The BIT Program serves students, ranging from freshmen that are not STEM majors to upper-level undergraduates and graduate students seeking life science-related degrees. MBLEMs are continually being created and updated to reflect current biotechnological advances and high impact teaching practices (HIP) (White, 2018). Bioinformatics are at the core of several of our course offerings and are included to varying degrees in all MBLEMs (**Figure 1A**), providing both vertical and horizontal integration of bioinformatics throughout the entire Program. Selected MBLEMs below illustrate how bioinformatics can be incorporated in the context of different class settings at varying educational levels.

INTEGRATING BIOINFORMATICS INTO A FIRST-YEAR COURSE FOR STEM AND NON-STEM MAJORS

Course Summary: Current Topics in Biotechnology

- 16-weeks, 4-credit lecture/laboratory course (twice weekly meetings of 2 h 45 min)
- General elective in the natural sciences for first-year students
- Introduction to the science and ethics of biotechnology

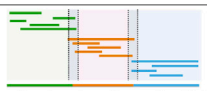
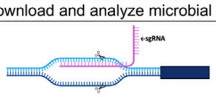
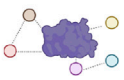
Current Topics in Biotechnology is a first-year inquiry lecture/laboratory course that provides science, technology, engineering, and mathematics (STEM) majors and non-STEM majors the opportunity to learn about biotechnology topics, ranging from biofuel production to genome editing using CRISPR-Cas9. The course has three goals: 1) "Think and Do" biotechnology, 2) Communicating scientific findings, and 3) Becoming a responsible community scientist. Learning outcomes for each of these goals are listed in **Table 1**.

It is challenging to teach bioinformatics to first-year students, especially those who are non-STEM majors. Introducing bioinformatics tools through the use of case studies is an effective way to involve all students. Different student interpretations of datasets provide a backdrop for rich discussions. Resources are available for peer-reviewed case studies on a variety of topics [e.g., CourseSource (Rosenwald et al., 2017), National Center for case Study Teaching in Science (NCCSTS, 2021)], thereby reducing the preparation required. A few examples of how we use bioinformatics in a first-year course are described below.

Case Study in Metagenomics

Students are led through the case study Unique Down to Our Microbes (Lentz et al., 2017), exposing undergraduates to methodologies used to study various types of microbial life that inhabit human bellybuttons, while allowing agency in the direction of research inquiry. Students analyze community-science collected data using the open-access bioinformatics tool Phinch.org (Bik and Pitch Interactive Inc, 2014), an open-source tool for visualizing analyzing large open-source biological datasets (**Figure 1A**). With this software, students analyze a large cohort of data collected from individuals from around the world (Hulcr et al., 2012) to investigate trends in microbial biodiversity in the human belly button. The

A

BLAST	
Current Topics in Biotechnology	Phylogenetic tree analysis
Breseq	
Directed Evolution	Analyze Illumina sequencing results for mutations within a population
CLC Genomics	
High-throughput Discovery	Download and analyze microbial genomes
CRISPOR	
CRISPR	Design guide RNA
NextStrain	
Current Topics in Biotechnology *Virus Biotechnology*	SARS-CoV-2 sequence comparison SARS-CoV-2 variant analysis
Phinch	
Current Topics in Biotechnology	Visualization of large data sets in graphical formats
Pymol	
Protein Interactions	Examine protein structure and ligand interactions
SnapGene	
Manipulation of Recombinant DNA *Protein Engineering* *Yeast Metabolic Engineering*	View vector maps, design primers, analyze Sanger sequencing, simulate cloning and agarose gels Analyze Sanger sequencing Design primers, analyze sequencing

B

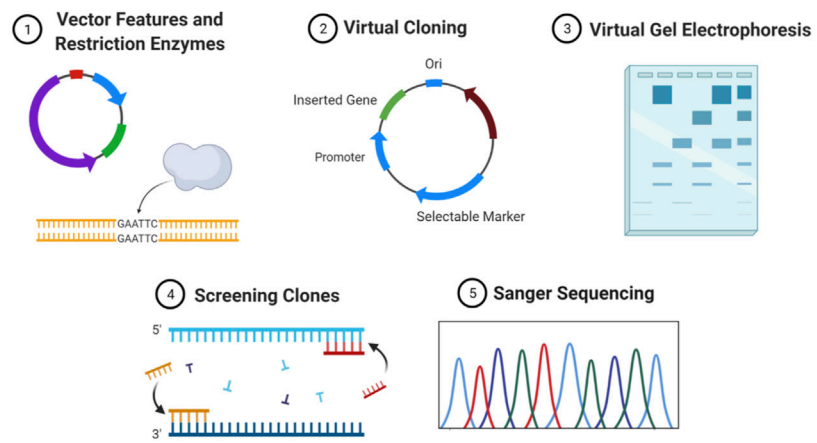


FIGURE 1 | Bioinformatics usage across MBLEMs. **(A)** Examples of bioinformatics software used in MBLEMs. **(B)** Student-driven in silico cloning projects challenge individuals to practice using bioinformatics software to solve tasks they would encounter in molecular biology. Key aspects of the project design are illustrated including: vector features and restriction enzymes, virtual cloning, virtual gel electrophoresis, screening, and Sanger sequencing. Images created with BioRender.com.

TABLE 1 | MBLEM course objectives and assessment methods.

Course objectives (COs)	Methods for assessing COs
MBLEM: Current topics in biotechnology	
By the end of this course, students will be able to	
CO 1: "Think and do" biotechnology	
• Understand the scientific concepts that underlie and experiment	o Lab activities
• Generate testable hypotheses	o Lab reports
• Create figures of scientific results	o Guided worksheets
• Interpret qualitative and quantitative experimental results	o Quizzes
CO 2: Communicate scientific findings	
• Keep detailed, concise, organized record of scientific experiments	o Lab notebook entries
• Communicate experimental results in a discipline-appropriate writing	o Lab reports
• Explain biotechnology concepts to different audiences	o Multimedia presentations
CO 3: Become a responsible community scientist	
• Identify and critique biotechnology issues relating to society or the responsible conduct of biotechnology research	o Reflection journal o Multimedia presentations
MBLEM: Manipulation of recombinant DNA	
CO 1: Design experiments to manipulate DNA	o Lab reports
• Design strategies to manipulate DNA to create new proteins	o Lab reports o Take home exams o Individual exams
• Perform a variety of techniques in the manipulation of recombinant DNA and protein expression	o Final exam o BIT 510 project Active learning activities
CO 2: Communicate scientific findings	
• Create a detailed written record of experimental procedures, results, and conclusions	o Lab notebook
• Interpret data and controls related to gene cloning, protein expression and hypothesis testing	o Lab reports
• Troubleshoot experiments that do not work	o Active learning activities
• Reflect on their own thinking and the thinking of others	
CO 3: Evaluate research questions	
• Evaluate a specific hypothesis	o Lab notebook o Lab reports o Take home exams o Final exam
CO 4: Exercise problem-solving skills in molecular biotechnology	
• Apply critical and creative thinking skills and behaviors in the process of solving problems or addressing questions	o Take home exams o Individual exams o Lab report 3 o Final exam
MBLEM: Metagenomics	
CO 1: Become a responsible community scientist	
• Demonstrate laboratory skills required of a modern-day molecular biologist in the era of next-generation sequencing. This includes keeping detailed and accurate laboratory notes (e.g., electronic records for sequence analyses) and choosing appropriate sequencing based on goals	o Critical thinking scenarios and discussion posts
CO 2: Read scientific literature	
• Read a scientific article and evaluate how bioinformatics methods were employed by the authors to explore a particular hypothesis. (From CourseSource framework)	o Article summaries and annotations o Collaborative notes o Knowledge check questions o Individual podcast explanation assignment
CO 3: Evaluate research questions	
• Given a scientific question, develop a hypothesis and define computational approaches that could be used to explore the hypothesis. (From CourseSource framework)	o Group data analyses project drafts and final submission
CO 4: Analyze experimental data	
• Use pre-existing tools to analyze a metagenomic data set to determine the set of organisms present in a metagenomic sample (e.g., 16s rRNA, greengenes, mothur, etc.). (From CourseSource framework)	o Individual podcast explanation assignment o case studies using KBase and QIIME2/DADA2
CO 5: Critically evaluate limitations of data analysis	
• Interpret data and identify limitations related to metagenomic surveys	o Individual podcast explanation assignment o Group data analyses project drafts and final submission
CO 6: (For graduate students) Explain analyses to different audiences	
• Design a critical thinking scenario and explain analyses for hypothesis testing of metagenomic data	o Video tutorial

familiarity (and somewhat whimsical nature) of the bellybutton generates genuine curiosity: How often should one wash their own bellybutton? Does frequency affect their microbial biota? Students eagerly form hypotheses and share their data analysis with the class.

Case Study in Virus Biotechnology

During the virus biotechnology unit, students use a case study to examine the spread and monitor the evolving genomes of SARS-CoV-2 in COVID-19: Where did you come from, where did you go? (Chen et al., 2020). Students use bioinformatics tools in the National Center for Biotechnology Information (NCBI) to perform sequence searches and alignments. Specifically, they learn how to interpret BLAST results (max score, query cover, E-value, and percent identity) and create phylogenetic trees of coronavirus sequence divergence using the open-source software Nextstrain (Hadfield et al., 2018). The website enables users to track real-time data of evolving pathogen populations and create interactive data visualizations. Data analyzed from NCBI BLAST and Nextstrain are used to answer important questions regarding SARS-CoV-2 spread. This case has been used widely in high school and college classrooms to engage students directly with the data underlying the COVID-19 pandemic, and can be adapted to answer emerging questions, such as the nature of the SARS-CoV-2 variants (**Figure 1A**).

INTEGRATING BIOINFORMATICS IN A DUAL-LEVEL UNDERGRADUATE, GRADUATE COURSE IN MOLECULAR BIOTECHNOLOGY

Course Summary: Manipulation of Recombinant DNA

- 16-weeks 4-credit lecture/laboratory course (lecture 1 h 50 min, lab 5 h weekly)
- Elective for upper-level undergraduate and graduate students
- Students perform a cloning project from gene to protein expression and testing

Manipulation of Recombinant DNA is a foundational lecture/lab course offered to STEM undergraduate majors and graduate students covering basic techniques in cloning, protein expression/purification, and prokaryotic and eukaryotic expression systems. Student learning objectives and assessments are described in **Table 1**. Lecture topics are directly related to the project-based laboratory sequence. For example, students learn about different types of screening methods and then perform three of those methods (i.e., restriction digestion, PCR screening, Sanger sequencing) (Garcia et al., 2021).

Bioinformatics tools in this course are first introduced through active learning assignments performed in collaborative groups, which we have found to be the most effective for student learning (Srougi et al., 2013; Tanner, 2017). Peer interactions are critical for learning key concepts and provide valuable experience in working in diverse groups, a critical skill in the biotechnology workforce. Introducing bioinformatics tools is a natural progression of the active-learning-based structure already established in the course. To extrapolate and hone in on their skills, students then individually complete a final capstone cloning project that involves bioinformatics. Since any software will involve a learning curve for students, it is imperative to develop students' familiarity with bioinformatics programs through frequent use. This ensures that students can focus critically on the project itself without the burden of mastering new software. A brief description of how bioinformatics tools are used in this course can be found below:

Molecular Cloning Visualized

Bioinformatics tools are introduced at the start of the course and interwoven throughout the curriculum. SnapGene software is the primary tool utilized to facilitate students' understanding of gene cloning by allowing them to design primers, generate the results of Gibson assembly or ligation cloning, and simulate polymerase chain reaction (PCR), restriction digestion, and agarose gel electrophoresis (**Figure 1B**). Using this software, students reinforce hands-on laboratory skills virtually by simulating the course cloning project. Graduate and honors students in the course go on to utilize SnapGene in a capstone project where they design an experimental cloning and protein expression strategy for a gene of interest related to their own interests. Due to its extensive utility in simulating experiments in a visually appealing manner before any reagents or time are wasted in the lab, students typically continue to use this software outside of the classroom in their own research.

Exposure to Complex Data Sets

Students in modern biotechnology education courses should have exposure to and practice with using complex data sets. To achieve this goal, the course incorporates the use of the La Cuadrilla case study (discussed in detail later). The case study presents a real-life scenario where villagers in La Cuadrilla, Mexico were getting sick from an unknown biological agent. Scientists from the Centers for Disease Control (CDC) collected water samples to perform a culture-independent diagnostic test with high-throughput sequencing. Students are given the raw data from this sequencing analysis and employ CLC Genomics software (Qiagen) to trim, filter, and analyze reads. CLC Genomics is an all-in-one software that enables analysis and visualization of data from all major next-generation sequencing (NGS) applications and provides an adaptable workflow for users depending upon their needs. The software does come at a fee, but discounted licenses are available for teaching purposes (see freeware options in **Supplemental Table S1**).

BIOINFORMATICS IN 8-WEEK SPECIALTY COURSES

Upon completion of the foundational course in the Manipulation of Recombinant DNA, students can choose from a menu of continuously updated 8-week, 2-credit hour specialty module courses to complete the biotechnology minor. Each of these lecture/laboratory courses focus on a particular cutting-edge aspect of biotechnology with an emphasis on hands-on laboratory skills. Current examples include: Plant Genetic

Engineering, CRISPR, High-Throughput Discovery, among others (see www.ncsu.edu/biotechnology). The selection of courses continually adapts with developments in the life sciences.

Course Summary: Metagenomics

- 8-weeks lecture/laboratory course (1 h 50 min lecture, 5 h lab weekly)
- Pre-requisite is Manipulation of Recombinant DNA or equivalent
- Students analyze metagenomic populations from various niches

Metagenomics (Goller and Ott, 2020) is an inquiry-based course that provides advanced level undergraduate and graduate students hands-on exposure to a wide-variety of methods for analyzing unique and complex microbial communities. A focus is on computational skills to evaluate interactions between microbial populations and their surroundings. The course incorporates a variety of pedagogical practices to aid student learning in the process-orientated nature of metagenomics research. To this end, the course was built around concept mapping (Ritchhart et al., 2011) and reflective writing assignments. In the laboratory, students are guided through a series of wet labs where they isolated and purified genomic DNA, then made DNA libraries for NGS sequencing. During sequencing, students are introduced to bioinformatics tools, including the CyVerse Discovery Environment (Cyverse, 2021) and QIIME (Bolyen et al., 2019), to prepare them to perform data Q/C and analyses. Through this process, students are introduced to computing resources on the cloud. During the COVID-19 pandemic, this course was successfully delivered online asynchronously; the laboratory component in this context consisted of bioinformatics exercises, including data analysis case studies and publications using high-throughput approaches to understand microbial communities. The student learning objectives and assessments for Metagenomics are detailed in **Table 1**.

Bioinformatics are at the core of Metagenomics; highlights of the bioinformatics tools and activities used in Metagenomics are featured below.

METAGENOMICS (TOOL) BOX

Interpreting microbial genomics with CLC. CLC Genomics Workbench (QIAGEN, 2021) and the Microbial Genomics Module (MGM) provide a powerful yet easy-to-use graphical interface through which students can import raw reads from an experiment they conduct in lab. Using the QIAGEN 16S/ITS panel and phased primers, sequencing libraries targeting multiple regions of the 16S ribosomal RNA gene and ITS can be prepared. CLC has a metagenomics workflow that is, included in the MGM that starts with importing and pairing reads and continues on to quality filtering and taxonomic classification using a downloaded database of choice (e.g., SILVA). CLC then displays taxonomic classification based on the sample metadata provided, allowing students to filter by sample type, re-graph by relative abundance, and run PERMANOVA analyses. This activity provides a user-friendly introduction used in the Metagenomics and Manipulation of Recombinant DNA MBLEMs.

CLC is used to engage students in analyzing next-generation sequencing datasets from a student-produced metagenomic survey. Using a case study approach (Herreid, 2011), students in the Manipulation of Recombinant DNA MBLEM learn about the use of high-throughput sequence to investigate an outbreak in a small village in rural central Mexico. The "La Cuadrilla" case study (**Supplementary Table S2**) challenges groups of students to analyze raw data and interpret it to make recommendations to the health department. A CLC metagenomics workflow that is, part of the MGM plugin of CLC enables students to work together to solve this mystery. This case study aligns with course learning objectives and introduces the use of databases, thus building on previous concepts addressed in the course. We created a second case study that addresses the concepts of genome assembly by having students assemble and explore sequencing results from a mystery yeast (**Supplementary Table S2**). While the cost of CLC can be significant, alternatives include free web-based platforms like Nephele (Weber et al., 2018), PUMAA (Mitchell et al., 2020), and KBase (Arkin et al., 2018). Nephele and KBase are important tools used in various MBLEMs to teach students about the analysis and applications of high-throughput sequencing. The importance of Nephele and KBase in MBLEMs is described below.

Nephele (Weber et al., 2018) is a web-based microbiome analysis pipeline developed by US National Institutes of Health National Institute of Allergy and Infectious Diseases. Nephele has pipelines for analysis of microbiome sequence data using popular tools, such as DADA2, QIIME2, and bioBakery. Users can upload datasets and modify workflow parameters to submit jobs for the heavy computational analyses. Results are then emailed with access to download results that include graphs, analyses, raw and filtered data, and log files. Students use Nephele to learn about the effects of different parameters and pipelines without the need of background in coding. For example, data from published microbiome studies can be reanalyzed to teach students about the significance of quality control and parameter optimization. Participants can also learn the importance of metadata for downstream analysis including hypothesis testing.

The KBase (Arkin et al., 2018) web-based platform developed by the US Department of Energy is used to analyze metagenomic datasets using sharable interactive workflow "narratives" focusing on metagenomic assembly, taxonomic identification, binning, and metabolic modeling. Students learn about "read hygiene" and compare different assemblers and the advantages and limitations of short-read taxonomic inferences and contig binning. Students employ narratives to learn key steps and explore new datasets, learning key concepts, and procedures in the process of examining datasets.

QIIME2/DADA2 (Bolyen et al., 2019) is used to expose participants to high-performance computing (HPC) and the ability to submit different jobs. Students log in and complete the QIIME Moving Pictures tutorial to practice working in the

command-line environment and understand the fundamental steps in a 16S amplicon metagenomics analysis. Emphasis is placed on the use of DADA2 and the differences between Operational Taxonomic Units (OTUs) and Amplicon Sequence Variants (ASVs) in the context of accuracy and reproducibility. Students use a common script to run tutorials and then adapt it for their own datasets.

SnapGene (Science, 2021) bioinformatics software has been integrated across several BIT MBLEMs to help students understand critical molecular biology concepts and gain experience with sequence analysis tools and approaches. This has been done in alignment with course objectives and frameworks describing bioinformatics core competencies for undergraduate education (Sayres et al., 2018; Williams et al., 2019). The flagship MBLEM Manipulation of Recombinant DNA includes activities using SnapGene, as discussed above. Additionally, SnapGene has been vertically integrated in other MBLEMs; in Yeast Metabolic Engineering SnapGene is used to analyze sequencing reactions and identify barcode sequences from promising yeast mutants producing beta-carotene. Students learn the concepts and skills using user-friendly software to apply in more advanced MBLEMs (**Figure 1B**). In addition, SnapGene integrates with the electronic lab notebook (ELN) system we use in the laboratory for several MBLEMs (LabArchives).

While SnapGene is not free, a site license can be purchased for campus-wide use; as an alternative, freeware sequence visualization, and analysis software such as Benchling and ApE are available alternatives (**Supplementary Table S1**). Besides cost, instructors should consider the accessibility features of the software and logistics required to provide access to students. For example, during the pandemic, students had to access SnapGene off-campus using full tunnel VPN, which required an additional setup on the part of the users.

SUMMARY

The BIT Program at NC State offers a dynamic set of cutting-edge courses in modern biotechnology through an innovative paradigm. A special feature of MBLEM design is the incorporation of datasets and bioinformatics tools into each course offering, based on best pedagogical practices (i.e., user-friendly software, peer support, and authentic inquiry-based projects). The examples provided encourage students to perform novel analyses on real datasets. Currently, MBLEMs are being actively employed by a wide variety of institutional partners to demonstrate that bioinformatics modules are portable and valuable in supplementing molecular biotechnology training.

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

SC, CG, MS, and RK developed the courses and exercises described herein, SC, CG, MS, LS, and RK planned, prepared, and wrote the manuscript.

FUNDING

Support for the activities described here was provided by the North Carolina State University's Provost's Office. We also acknowledge support from the National Institutes of Health Innovative Programs to Enhance Research Training (IPERT) (R25 GM130528) and the National Science Foundation Research Experience for Undergraduates (REU) (1659225).

ACKNOWLEDGMENTS

We thank the many North Carolina State University graduate and undergraduate students enrolled in the Biotechnology Program courses who helped develop and test the exercises described herein.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.711403/full#supplementary-material

REFERENCES

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* 36 (7), 566–569. doi:10.1038/nbt.4163

Bik, H. M.Pitch Interactive Inc. (2014). *Phinch: An Interactive, Exploratory Data Visualization Framework for –Omic Datasets*. bioRxiv.

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 (8), 852–857. doi:10.1038/s41587-019-0209-9

Chen, S. H., and Goller, C. C. (2020). Harnessing Single-stranded DNA Binding Protein to Explore Protein-Protein and Protein-DNA Interactions. *Biochem. Mol. Biol. Educ.* 48 (2), 181–190. doi:10.1002/bmb.21324

Chen, S. H., and Goller, C. C. (2019). Shifting Faculty Approaches to Pedagogy through Structured Teaching Postdoc Experiences. *J. Microbiol. Biol. Educ.* 20 (2). doi:10.1128/jmbe.v20i2.1789

Chen, S. H., Goller, C. C., and Srougi, M. C. (2020). *COVID-19: Where Did You Come from, where Did You Go?*. National Center for Case Study Teaching in Science.

CUR (2021). Council on Undergraduate Research. Retrieved from https://www.cur.org/ (Accessed 6 17, 2021).

CUREnet (2021). CUREnet: Course-Based Undergraduate Research Experience. Retrieved from https://serc.carleton.edu/curenet/index.html (Accessed 6 17, 2021).

Cyverse (2021). Cyverse Discovery Environment 2.0. Retrieved from https://de.cyverse.org/ (Accessed 6 17, 2021).

Garcia, C. B., Chapman, I. F., Chen, S. H., Lazear, E., Lentz, T. B., Williams, C., et al. (2021). Integrating Research into a Molecular Cloning Course to Address the Evolving Biotechnology Landscape. *Biochem. Mol. Biol. Educ.* 49 (1), 115–128. doi:10.1002/bmb.21402

Goller, C. C., and Ott, L. E. (2020). Evolution of an 8-week Upper-division Metagenomics Course: Diagramming a Learning Path from Observational to Quantitative Microbiome Analysis. *Biochem. Mol. Biol. Educ.* 48 (4), 391–403. doi:10.1002/bmb.21349

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., et al. (2018). Nextstrain: Real-Time Tracking of Pathogen Evolution. *Bioinformatics* 34 (23), 4121–4123. doi:10.1093/bioinformatics/bty407

Herreid, C. F. (2011). "Case Study Teaching," in *New Directions for Teaching & Learning*.

Hulcr, J., Latimer, A. M., Henley, J. B., Rountree, N. R., Fierer, N., Lucky, A., et al. (2012). A Jungle in There: Bacteria in Belly Buttons Are Highly Diverse, but Predictable. *PLoS One* 7 (11), e47712. doi:10.1371/journal.pone.0047712

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 (2). doi:10.1128/jmbe.v18i2.1284

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

Myers, M. J., and Burgess, A. B. (2003). Inquiry-based Laboratory Course Improves Students' Ability to Design Experiments and Interpret Data. *Adv. Physiol. Educ.* 27 (1-4), 26–33. doi:10.1152/advan.00028.2002

NCCSTS (2021). National Center for Case Study Teaching in Science. Retrieved from https://sciencecases.lib.buffalo.edu/ (Accessed 6 17, 2021).

Ott, L. E., and Carson, S. (2014). Immunological Tools: Engaging Students in the Use and Analysis of Flow Cytometry and Enzyme-Linked Immunosorbent Assay (ELISA). *Biochem. Mol. Biol. Educ.* 42 (5), 382–397. doi:10.1002/bmb.20808

QIAGEN (2021). CLC Genomics Workbench 20.0. Retrieved from https://digitalinsights.qiagen.com/.

Ritchhart, R., Church, M., and Morrison, K. (2011). *Making Thinking Visible: How to Promote Engagement, Understanding, and Independence for All Learners*. San Francisco, CA: Jossey-Bass.

Rosenwald, A. G., Pauley, M. A., Welch, L., Elgin, S. C., Wright, R., and Blum, J. (2017). The *CourseSource* Bioinformatics Learning Framework. *CBE Life Sci. Educ.* 15 (1), le2. doi:10.1187/cbe.15-10-0217

Samsa, L. A., Anderson, L., Groth, A., and Goller, C. C. (2020). A CRISPR/Cas Guide RNA Design *In Silico* Activity. *CourseSource* 7. doi:10.24918/cs.2020.46

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 (6).

Science (2021). SnapGene. Retrieved from https://support.snapgene.com/hc/en-us.

Srougi, M. C., and Carson, S. (2013). Inquiry into Chemotherapy-Induced P53 Activation in Cancer Cells as a Model for Teaching Signal Transduction. *Biochem. Mol. Biol. Educ.* 41 (6), 419–432. doi:10.1002/bmb.20741

Srougi, M. C., Miller, H. B., Witherow, D. S., and Carson, S. (2013). Assessment of a Novel Group-Centered Testing Schema in an Upper-Level Undergraduate Molecular Biotechnology Course. *Biochem. Mol. Biol. Educ.* 41 (4), 232–241. doi:10.1002/bmb.20701

Tanner, K. D. (2017). Structure Matters: Twenty-One Teaching Strategies to Promote Student Engagement and Cultivate Classroom Equity. *CBE Life Sci. Educ.* 12 (3).

Wallace, C. S., Tsoi, M. Y., Calkin, J., and Darley, M. (2003). Learning from Inquiry-Based Laboratories in Nonmajor Biology: An Interpretive Study of the Relationships Among Inquiry Experience, Epistemologies, and Conceptual Growth. *J. Res. Sci. Teach.* 40 (10), 986–1024. doi:10.1002/tea.10127

Weaver, G. C., Russell, C. B., and Wink, D. J. (2008). Inquiry-based and Research-Based Laboratory Pedagogies in Undergraduate Science. *Nat. Chem. Biol.* 4 (10), 577–580. doi:10.1038/nchembio1008-577

Weber, N., Liou, D., Dommer, J., MacMenamin, P., Quiñones, M., Misner, I., et al. (2018). Nephele: a Cloud Platform for Simplified, Standardized and Reproducible Microbiome Data Analysis. *Bioinformatics* 34 (8), 1411–1413. doi:10.1093/bioinformatics/btx617

White, A. (2018). Understanding the University and Faculty Investment in Implementing High-Impact Educational Practices. *JoSoTL* 18 (2), 118–135. doi:10.14434/josotl.v18i2.23143

Wiggins, G., and McTighe, J. (1998). *Understanding by Design*. Association for Supervision and Curriculum Development.

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 (11). doi:10.1371/journal.pone.0224288

Witherow, D. S., and Carson, S. (2011). A Laboratory-Intensive Course on the Experimental Study of Protein-Protein Interactions. *Biochem. Mol. Biol. Educ.* 39 (4), 300–308. doi:10.1002/bmb.20506

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

PERSPECTIVE
published: 09 August 2021
doi: 10.3389/feduc.2021.711512

HITS: Harnessing a Collaborative Training Network to Create Case Studies that Integrate High-Throughput, Complex Datasets into Curricula

*Sabrina D. Robertson[1], Andrea Bixler[2], Melissa R. Eslinger[3], Monica M. Gaudier-Diaz[1], Adam J. Kleinschmit[4], Pat Marsteller[5], Kate K. O'Toole[5], Usha Sankar[6] and Carlos C. Goller[7]**

[1]Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, Chapel Hill, NC, United States, [2]Department of Science and Mathematics, Clarke University, Dubuque, IA, United States, [3]Department of Chemistry and Life Science, United States Military Academy, West Point, NY, United States, [4]Department of Natural and Applied Sciences, University of Dubuque, Dubuque, IA, United States, [5]Department of Biology, Emory University, Atlanta, GA, United States, [6]Department of Biological Sciences, Fordham University, Bronx, NY, United States, [7]Department of Biological Sciences and Biotechnology Program, North Carolina State University, Bronx, NY, United States

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Jean-Karim Hériché, European Molecular Biology Laboratory Heidelberg, Germany
Martin Leonard Jones, Francis Crick Institute, United Kingdom

***Correspondence:**
Carlos C. Goller
ccgoller@ncsu.edu

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *18 May 2021*
Accepted: *20 July 2021*
Published: *09 August 2021*

Citation:
Robertson SD, Bixler A, Eslinger MR, Gaudier-Diaz MM, Kleinschmit AJ, Marsteller P, O'Toole KK, Sankar U and Goller CC (2021) HITS: Harnessing a Collaborative Training Network to Create Case Studies that Integrate High-Throughput, Complex Datasets into Curricula. Front. Educ. 6:711512. doi: 10.3389/feduc.2021.711512

As educators and researchers, we often enjoy enlivening classroom discussions by including examples of cutting-edge high-throughput (HT) technologies that propelled scientific discovery and created repositories of new information. We also call for the use of evidence-based teaching practices to engage students in ways that promote equity and learning. The complex datasets produced by HT approaches can open the doors to discovery of novel genes, drugs, and regulatory networks, so students need experience with the effective design, implementation, and analysis of HT research. Nevertheless, we miss opportunities to contextualize, define, and explain the potential and limitations of HT methods. One evidence-based approach is to engage students in realistic HT case studies. HT cases immerse students with messy data, asking them to critically consider data analysis, experimental design, ethical implications, and HT technologies.The NSF HITS (High-throughput Discovery Science and Inquiry-based Case Studies for Today's Students) Research Coordination Network in Undergraduate Biology Education seeks to improve student quantitative skills and participation in HT discovery. Researchers and instructors in the network learn about case pedagogy, HT technologies, publicly available datasets, and computational tools. Leveraging this training and interdisciplinary teamwork, HITS participants then create and implement HT cases. Our initial case collection has been used in >15 different courses at a variety of institutions engaging >600 students in HT discovery. We share here our rationale for engaging students in HT science, our HT cases, and network model to encourage other life science educators to join us and further develop and integrate HT complex datasets into curricula.

Keywords: high-throughput, collaborative network, data analyses, quantitative skills, pedagogical case studies

INTRODUCTION—A CALL TO INTEGRATE HIGH-THROUGHPUT DISCOVERY SCIENCE INTO CURRICULA

High-throughput (HT) approaches have become critical for contemporary scientific discovery. HT research often leverages the power of automation and miniaturization to make large-scale experimentation feasible and promote discovery. Advances in sequencing (McCombie et al., 2019), drug discovery (da Silva Rocha et al., 2019; Farha and Brown, 2019; Vamathevan et al., 2019; Zhu, 2020), imaging (Pegoraro and Misteli, 2017), proteomics (Tyers and Mann, 2003), machine learning and genome editing are transforming the way scientists perform research. HT approaches allow researchers to conduct multiple chemical, genetic, and phenotypic tests to identify active molecules, conditions, or genes with desired or useful features (Appleton et al., 2017). HT *omics* have enabled simultaneous examinations of hundreds or thousands of genes, proteins, and metabolites (Judes et al., 2016). Large NIH (Collins, 2010) and NSF initiatives (e.g., BREAD PHENO) also take advantage of HT methodologies (ENCODE, 2004; Regev et al., 2017; Hutter and Zenklusen, 2018; Koroshetz et al., 2018; McDonald et al., 2018). This fast-paced development and diversity of HT approaches necessitates effective training of scientists and engineers (Miller, 2014; Stephens et al., 2015; Batut et al., 2018). To better prepare students for this new HT research world, an understanding of HT approaches and dataset wrangling must be broadly integrated into curricula to empower all students to analyze and discover.

Variations in HT terminology may refer to a discipline (e.g., biology, genetics, bioinformatics), a technique (e.g., screening, sequencing), or a volume of data, but it is the data science and analysis skills that are more essential than ever for STEM students. For students to analyze and appreciate results, a general understanding of how data is validated and processed is required. Yet, undergraduate biology programs have been slow to adopt practices and curricula that enhance such quantitative skills (Feser et al., 2013; Hoffman et al., 2016). Educators have generated a multitude of reports and educational initiatives to fill this void (National Research Council, 2003; Bialek and Botstein, 2004; National Research Council, 2009). Despite the calls to action, there remains a relative dearth of educational resources that engage students with complex, HT datasets aimed at enhancing their quantitative skills (Aikens and Dolan, 2014). In part, this may reflect a lack of sufficient faculty training and a need for professional development and teaching support opportunities (AAAS, 2011). Since the design, execution, and analysis of HT data require fundamental quantitative skills, collectively this provides a ripe opportunity for faculty and students alike to expand these skills. Indeed, these skills align with the first two core competencies outlined by *Vision and Change* and the BioSkills Guide for revolutionizing undergraduate biology education (AAAS, 2011; Clemmons et al., 2020). Successful examples of undergraduate data mining and contribution to research (*i.e.,* Genomics Education Partnership (Lopatto et al., 2014) and Genome Solver programs (Rosenwald et al., 2012) are becoming more common, and this integration of authentic research has a multitude of strongly supported student benefits (Lopatto et al., 2008; Thiry et al., 2011; Jordan et al., 2014; Shaffer et al., 2014).

High-Throughput Discovery Science and Inquiry-Based Case Studies for Today's Students Network

We created the NSF-funded Research Coordination Network (RCN) known asHigh-throughput Discovery Science & Inquiry-based Case Studies for Today's Students (HITS), to develop innovative curriculum materials using a case-based approach. RCNs have had a profound scholarly impact (Porter et al., 2012) and continue to generate and disseminate resources that improve research and education. Other RCNs have focused on the development of case studies (NeuroCaseNet, Molecular case Net (Goodsell et al., 2021), OCELOTS), bioinformatics (NIBLSE) or next generation sequencing (GCAT-SEEK). Additional educational initiatives (BEDROCK, ESTEEM, NUMB3R5 COUNT!, EDDIE) aim to increase student quantitative skills in biology but are not focused on case studies or high-throughput discovery science. HITS specifically addresses a unique need for case studies that provide exposure to and practice with manipulating and analyzing high-throughput datasets and understanding HT experimental approaches.

Case-based learning has been adopted and adapted into undergraduate and graduate education (Blosser, 1988; Gallagher et al., 1995; Krynock and Robb, 1996) and has become recognized as a high impact practice that supports *Vision and Change* core competencies (Lombardi, 2007; Freeman et al., 2014). According to the *Vision and Change* 2010 report (AAAS, 2011), all students need to develop the ability to apply the process of science, use quantitative reasoning skills, use models and simulations, understand the interdisciplinary nature of scientific investigations, and be able to communicate science well and in collaboration with other disciplines. Students also need to place their science knowledge in a societal context and learn to apply scientific concepts to real world problems. Case-based instructional methods, which include the case study method, problem-based learning, and investigative case-based learning, use realistic narratives to engage students in solving problems, building analytical skills, and working cooperatively in a self-directed way (Herreid, 1994; Torp and Sage, 2002; Duch et al., 2011). The complex nature of case studies leads students to assess problems from a myriad of perspectives and those that emphasize the human dimension of an issue or controversy powerfully demonstrate the relevance of a given topic to students and generate engagement (Yadav et al., 2007). Approaches to writing and adapting case studies have been previously well documented (Herreid, 1997; Herreid, 2018; Prud'homme-Generaux et al., 2018; https://sciencecases.lib.buffalo.edu/teaching/publications/).

Datasets produced by HT approaches contain untold stories about the biological world around us. Providing "scaffolded" cases for students to understand HT methodology, dive into data, and uncover trends promotes quantitative skills, hypothesis development, and the excitement for discovery. The HITS research and educational network provides structure and

support for the development of HT case studies that promote student learning of core HT concepts and competencies. The resulting inquiry-based cases engage students in authentic research and better train future scientists for jobs in modern biology. Recently, our network shared insights on how to adapt HT cases for the classroom (Bixler et al., 2021). The HITS network also includes established non-profit and industry collaborations with researchers, institutes, and companies that produce datasets amenable for exemplary HT case studies. For example, The Allen Institute has worked with HITS to create case studies that use resources on the Allen Cell Explorer and Allen Brain Map websites to engage students in analysis of thousands of microscopy images of engineered cells and real electrophysiology data from human neurons. These datasets and equipment alone encompass several research areas and HT approaches, including sequencing, imaging, and screening. As HITS expands, additional collaborations and projects provide more datasets and resources to explore ways to connect students and courses at different institutions. By partnering researchers, educators, and case study experts to create, implement, and share HT case-based tools, our HITS network improves curricula across the nation and better equips life science students with essential science process skills and complex dataset literacy.

How High-Throughput Discovery Science and Inquiry-Based Case Studies for Today's Students Cases Are Unique

HT case studies characteristically integrate large quantitative datasets with an engaging biological narrative to directly showcase the real-world relevance of course concepts. The unique features that make HT cases stand out include the combined focus on HT experimental approaches and contemporary techniques, the use of quantitative data analysis skills and computational tools, and showcasing the interdisciplinary nature of science. To ensure our cases capture the essence of HT discovery science, we crafted a set of learning objectives to ground our HT case study development (**Supplementary Table S1**). Our HT cases also often use publicly-available large, complex datasets to facilitate reproducibility studies (OSC, 2015) as well as authentic data mining (Hand, 2007). In addition to introducing students to databases/repositories and HT experimental design, a second focus of HT cases is to immerse students in the development of analytical skills (e.g., filter, analyze, and interpret data; synthesize data and draw appropriate inferences) with large, complex datasets.

Williams et al. (2019) identified that faculty perceive barriers to integrating computational and bioinformatics competencies involving HT "big data" into curricula. HITS helps lower these barriers by offering validated case studies that can facilitate the execution of competency-based learning outcomes (e.g., analytical skills). HT cases also address the barriers of limited instructor training and lack of HT technology on many campuses. The use of publicly available datasets, instructor collaboration with experts, detailed implementation notes, and the modular and adaptable nature of HITS cases make them easy to adopt or alter to meet specific curricular learning goals. HT cases are also well suited for use in a variety of modalities including interactive face-to-face classrooms, collaborative distance learning, and the laboratory environment (Bixler et al., 2021). HT cases in the laboratory setting can also provide learning resources for HT computational biology and bioinformatics-centric experiments. Over recent decades, computational biology has shifted from being a set of tools used to support biological studies to its own discipline that often drives research programs, with physical experiments and field research playing an equal or supporting role (Yanai and Chmielnicki, 2017; Tapprich et al., 2021). This interdisciplinary (e.g., life science, computer science, mathematics, statistics) nature of HT cases promotes multidisciplinary knowledge within the classroom, providing students with practice synthesizing ideas in the context of the collaborative nature of science. Here we share several examples of successful implementation of HT cases in a broad range of institutions and courses (**Table 1**) as well as our model for HT case development.

The High-Throughput Discovery Science and Inquiry-Based Case Studies for Today's Students Network Model for High-Throughput Case Study Development

The goals of the HITS network (**Figure 1**; go.ncsu.edu/hits) are to. . .

- Organize interdisciplinary annual workshops, create a virtual community, and publish educational resources to foster HT research and its integration into the classroom.
- Create novel working groups of researchers and educators to design, assess, and disseminate much needed quantitative biology case studies based on HT discovery research.
- Build a diverse consortium of institutions committed to implementing these quantitative educational tools in biology courses and curricula across the country and world.

HITS case Fellows: As seen in **Figure 1** and **Supplementary Figure S1**, the HITS network provides networking, research, and training opportunities for a diverse group of undergraduates, graduate students, postdoctoral fellows, researchers, and educators. Participants from a myriad of institutions and backgrounds interact at meetings, bringing creative energy and varied expertise to make HT discovery science, bioinformatics tools, case study design, and large, complex datasets accessible. Key among our participants are the HITS case Fellows. Given the barriers to integrating HT data into curricula (e.g., time, bioinformatics training, case study familiarity), HITS provides structured time, accountability, training, professional development, and financial support to case fellows who commit to creating and implementing HT case studies. We support at least 10 new HITS case Fellows each year by providing their HT case study training (i.e., 2 years of stipends and the opportunity to participate in two annual HITS meetings free of cost). In turn HITS fellows collaborate in teams to build modular and adaptable HT case studies for a variety of classroom

TABLE 1 | HITS network high-throughput case study projects.

Case Title (and link), Authors, *Case Description*	Curricular and Student Impact
COVID-19: **Where Did You Come From, Where Did You Go?** Chen, S., Goller, C.C., Srougi, M.C. *Focused on the evolution and spread of SARS-CoV-2 during the pandemic, students use freely available bioinformatics tools (NCBI, BLAST, NextStrain) to track changes in SARS-CoV-2 through its genetic code*	N.C. State University (20–32 students) in either an introductory 100 level biotechnology course and a 400 level virus biotechnology course
Applying High-Throughput Analysis to Biofilms Homesley, Jr, M.L. *This case uses a hypothetical scenario in which a dental practice is contaminated from an unknown bacterial source. Students determine the origin of the bacteria and learn about a high-throughput experiment for analyzing biofilms*	This student-produced case serves as an example for other students to create and publish case studies that combine their research interests and high-throughput approaches. For example, this case inspired another student to create and publish a case study on their research
[a]**Does Organelle Shape Matter: Exploring Patterns in Cell Shape and Structure with High-throughput (HT) Imaging** Goller, C.C., Casimo, K *Using the Allen Cell Explorer, students study shape changes in the endoplasmic reticulum*	N.C. State University (10–16 students) in upper-division genomics class *every* semester since Spring 2020
[a]**Single Cell Insights into Cancer Transcriptomes: A Five-Part Single-Cell RNAseq Case Study Lesson** Samsa, L., Eslinger, M., Kleinschmit, A., Solem, A., Goller, C *Students experience how single cell RNA sequencing transcriptomics is used to discover novel ways that cells function in homeostasis and disease*	N.C. State University (10–16 students in upper-division genomics class every semester since Spring 2020) University of Dubuque (24 students/semester- Spring 2020 and 2021) Hastings College (12 students)
[b]**Seq'ing the Cure: Standard Edition and Neuroscience Editions** Miller, H.B., Robertson, S.D., Srougi, M.C. *Students analyze RNA-seq data and identify differentially expressed genes in the context of a patient who presents at the ER with a neurological disorder*	UNC Chapel Hill (>100 students across 3 semesters of a 400-level Neurotechnology course) High Point University (~10 students in Biochemistry II and 18 in Biochemistry of Gene Expression per semester)
[b]**New Tricks for Old Drugs: Using High-throughput Screening to Repurpose FDA-Approved Drugs to Combat Zika Virus** Goller, C., Chen, S., Srougi, M *Students explore a publication about Zika virus drug screen and use PubChem*	N.C. State University (10–16 students in upper-level Genomics course every semester since Spring 2020)
[c]**Pesticides in My Smoothie Bowl?** Yu, S., Weir, S *Introduces basic concepts and processes of ecological and human health risk assessment using pesticides as an example and provides students an opportunity to analyze real-world pesticide monitoring data*	Central Piedmont Community College (>30 students in Introductory Biology per semester)
[c]**What's the Difference?: Differential Methylation Within Schizophrenia** Simkin, A., Niedziela, L., Eslinger, M *Students examine the genetic methylation patterns of people with and without schizophrenia to identify potential candidate genes whose abnormal methylation may be associated with schizophrenia*	West Point (>50 students over two semesters of 300-level Genetics) Elon University (~20 students in Genetics Lab)
[c]**Mapping Human Neuron Diversity in the Search for New Therapies** Casimo, K. Robertson, S *Students collect data from the Allen Cell Types Database, a resource by the Allen Institute for Brain Science, and evaluate the properties of neurons in the temporal lobe compared to other brain regions*	UNC Chapel Hill (>200 students in 100 level introduction to neuroscience course and 400 level neurotech course)
[b]**Screening Nanobodies to Target** Chen, S., Bixler, A *Students learn techniques involved in directed evolution, particularly yeast surface display, in the context of research on SARS-CoV-2*	Clarke University (Introductory version tested with 18 intro students; an upper-level/ graduate version will also be available)
[d]**You Are What You Eat: Microbiome and Disease** Carroll C., Dihle P., S., Salger S., Vega N., Gaudier-Diaz M., Kleinschmit A., Robertson *Students explore high-throughput sequencing technologies used to study microbiota, factors to consider when designing microbiome studies, data processing and visualization of microbiome data, and societal considerations touching on ethical issues, which may lead to microbiome inequality*	University of Dubuque (48 students in 300-level Microbiology class) UNC Chapel Hill (~100 students across 2 semesters of a 400-level Neurotechnology course) UNC Pembroke (18 students in a Physiology class) Emory University (15 students in an upper-level undergrad/grad Microbiome Ecology class) West Point (>40 students in 3 sections of Genetics)

(Continued on following page)

TABLE 1 | (*Continued*) HITS network high-throughput case study projects.

Case Title (and link), Authors, *Case Description*	Curricular and Student Impact
[d]**What's in a Wetland?** Bixler, A., Eslinger, M., Holmberg, T.J., Levine, T *Introduces students to metabarcoding in the context of investigating which species live in a wetland. Students compare visual taxonomy and sequence data from the National Ecological Observatory Network either through their own analysis or by studying prepared summaries*	Clarke University (Piloted with ~18 students each of two semesters in introductory biology. Could also be used in ecology and conservation courses.)
[d]**Get on the OMNI-bus: Applied Toxicogenomics in Breast Cancer** Schockley, K., Sankar, U., Eslinger, M *Students become familiar with a primary article on response to treatment and use guided R programming with the embedded large dataset to determine target genes affected*	West Point (piloted R coding with genetics class)
[d]**Survive and Conquer: Dissecting the MEGA-plate Experiment** Carr S., Mathew S., Pruneski J., Stasulli N This case study explores the antibiotic crisis using the MEGA-plate (Microbial Evolution Growth Arena) experiment, which visually and genetically tracks resistance development in time and space, to help students understand the development and evolution of antibiotic resistance in bacteria	Hartwick College, Campbell University, Heidelberg University in Tiffin, University of New Haven (modules piloted in microbiology--both lower level for allied health and upper level for majors, bioinformatics and biology seminar courses)

[a]*Publication accepted @ CourseSource (Peer Reviewed Open Educational Resource Journal).*
[b]*Publication accepted @ NCCSTS (no cost access to the student case study documents, but requires NCCSTS membership for access to answer keys and teaching notes $25/year).*
[c]*Publication in preparation.*
[d]*Case in development, revision or assessment phase. Contact sabrinae@email.unc.edu or ccgoller@ncsu.edu to access and use our HT cases in your classroom.*

environments (**Table 1**). Our case Fellows also commit to validating the newly created HT cases through cycles of assessment and revision. Once validated, dissemination of the HT cases for broader use in the educational community is a key goal. To recruit HITS case Fellows, each year we advertise widely primarily through the QUBES platform (Donovan et al., 2015) and our collective listservs and professional networks. We leverage attendance and participation at various national meetings (e.g., NIBLSE conference 2019; PANPBL 2019; NCCSTS 2018, BIOME 2018 and 2020). Now in our fourth year we've expanded to an international network that includes >80 members, and to date, we've supported >60 HITS participants from 34 institutions at our annual conference (**Supplementary Figure S1**).

HITS Annual Conference: The HITS conference occurs annually and has taken place both in person (2018 and 2019) and virtually due to the COVID-19 pandemic (2020 and 2021). The workshop's three-day schedule has evolved over time but generally includes: 1) exposure to HT Discovery Science from research experts on day one, 2) interactive workshop sessions with case study experts and brainstorming of case ideas to facilitate team formation on day two, and 3) protected time for interdisciplinary teams to collaborate on cases on day three (Sample Schedule **Supplementary Material** 3). Key to the success of these workshop sessions is the active collaboration of HT researchers and case study development experts. HT scientists provide short talks and opportunities for case Fellows to learn about or work with large, complex HT datasets. To complement the HT training, pedagogical case study experts provide evidence-based best practices for case design and have participants work through exemplary case-based learning resources and tools.

Case Fellows participate in our HITS summer workshop twice, which allows them to first become familiar with HT research and case study development, and later gain confidence in their abilities to generate, implement, assess, and disseminate HT case studies. Given the conference format and their 2-year involvement, all fellows contribute to multiple cases. Indeed, interdisciplinary collaboration is key in the development of these HT case studies. Often, research experts familiar with the datasets or even responsible for generating them collaborate with classroom teachers to select and trim appropriate datasets as well as design case studies that can be implemented in a variety of courses. At the conference, case Fellows generate a list of HT topics which later narrows as teams are formed. Within each team tasks are distributed to lessen each instructor's burden, ensuring the HT case is integrated into each classroom. Teams then have the opportunity to present at the conference and receive feedback from HT experts, other fellows, and case design experts. This piloting with both students and HITS participants alike enables rounds of revision to refine and validate the newly created HT case for assessment and broader dissemination, which includes student materials, detailed teaching notes, and in some cases cloud-based computing resources through QUBES.

The diverse perspectives among case Fellows ensures the development of cases that can be adapted to either introductory or advanced courses in a wide variety of topics (e.g., Neuroscience, Anatomy and Physiology, Genetics, Ecology, Biotechnology), broadening the impact of our network. During our conference there is also opportunity to discuss and address barriers to implementation (e.g., teaching load, class size, heterogeneity of student background and preparation). Modular design to enable adaptability for VERY different classrooms and virtual case Fellow support throughout the 2 years beyond the conference are key. This continued virtual collaboration to facilitate development, implementation, assessment, and dissemination is supported by QUBES (Donovan et al., 2015) and modeled by other successful

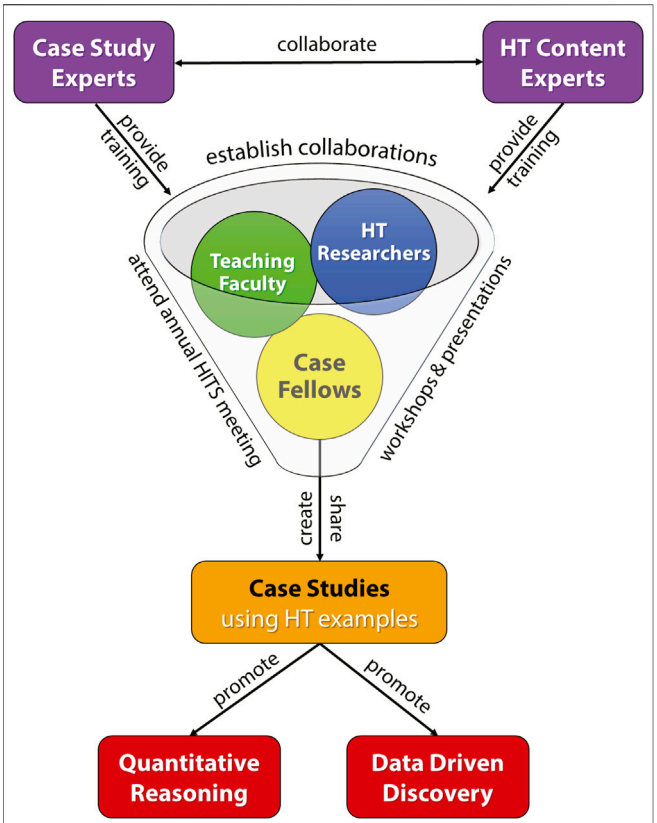


FIGURE 1 | The HITS Network Model to integrate HT discovery into curricula. Case Fellows collaborate with educators and HT Researchers at the annual HITS meeting and within a virtual support community (funnel). They receive training from case Study Experts and HT Content Experts in order to gain the skills and knowledge to develop and disseminate novel HT teaching resources that promote data driven discovery and quantitative reasoning.

networks (Ryder et al., 2020). Our HITS website allows us to track case study team progress and to feature examples of newly created HT case studies. Access to collaborative teams and the virtual HITS support network is essential to the success of each case Fellow and HITS participant. Indeed, the three-year evaluation of the program prepared by the NC State Office of Faculty Development found that 94% of prior year respondents published one or more cases since becoming a HITS case Fellow and 13% created ≥3.

HITS Steering Committee: HITS also relies on a senior leadership team composed of the PI, Co-PI and steering committee that analyzes report data to identify improvement ideas, suggests methods for network expansion, and promotes inclusivity. This team includes diverse faculty from all career levels and types of academic institutions (public and private, R1, primarily undergraduate, and minority serving institutions). The diversity of perspectives has enabled us to identify and adjust to major roadblocks in the development of the HITS network. Our steering committee has also shared their own HT research, led case study workshop sessions and interactive QUBES tutorials. Steering committee members work with individual teams of case fellows throughout each year to assist in continued virtual case study development, assessment and dissemination. The extensive networks of our steering committee team has also enabled us to coordinate with existing resources (ScienceCaseNet, QUBES, NIBLSE, BioQuest) to disseminate products and recruit interested faculty. The steering committee's exemplary, strong, and diverse leadership is essential to meeting our call to integrate HT discovery science into life science education.

Discussion

Case studies based on authentic HT data present an untapped opportunity to engage students in the process of discovery while also teaching them fundamental quantitative reasoning skills. HITS has capitalized on the synergy between diverse HT researchers and educators to create such cases through interdisciplinary collaboration and professional development training opportunities at all levels. The accomplishments of HITS case Fellows are showcased in several publications as either open educational resources or as free resources with access to supplemental documents for a nominal fee depending on the outlet (**Table 1**). These cooperative teams leverage individual expertise to develop ambitious, often modular, cases for use across the nation. Our cases have immersed hundreds of students in cutting-edge HT-experimental approaches, which are transforming how biology research is conducted, all while ensuring students develop essential data literacy and quantitative skills. Such curricular reform establishes a shared community of ideas with themes that cross multi-disciplinary thresholds and incorporate our established shareholders.

Our network has developed a long-term sustainability plan to continue to generate high-impact, HT case studies. A growing awareness of HITS and our goals and sense of community are necessary to have a lasting impact (AAAS, 2011). In particular, this requires a clear identification of stakeholders, in our case, 1) HITS case Fellows (e.g., seasoned faculty, mid-career, new trainees) 2) researchers (e.g., science, technology, bioinformatics), 3) educators (undergraduate and graduate education) 4) curriculum specialists (educators, information technologists) and 5) partner organizations (e.g., HHMI, AAAS, SABER). Synergistic approaches with these stakeholders will widen the HITS aperture to support our organizational goals. Here are five key steps in our strategic plan to further integrate HT discovery science and bioinformatics skill development into life science curricula across the nation and internationally:

1. Recruit and develop enduring teams who commit to leveraging their interdisciplinary expertise in STEM-related fields and bioinformatics to integrate HT science into undergraduate and graduate curricula.
2. Identify and address barriers to implementing HT approaches in the classroom (e.g., faculty/student reluctance, technology, accessibility, language/terminology, perceptions).
3. Continue to create, validate, and share curriculum products that train other faculty and students in the acquisition, analysis, and discovery of the HT data sciences.
4. Encourage the diversity, equity, and inclusion of students and trainees in HT discovery science through the development of cutting-edge curricula that promote STEM retention and leadership of the next generation.

5. Expand the HITS network through broad dissemination of our goals and products as well as through strategic partnerships with key life science and bioinformatics educational communities.

Our network's educational resources can be implemented in a variety of classrooms. Our HITS framework can also serve as a model for related fields with similar training bottlenecks to help create resources that address key gaps in current biology curricula. We also welcome collaboration to support other ongoing related initiatives (e.g., NIBLSE, NEUBIAS, NUMB3R5 COUNT!, EDDIE) and will continue to share products through publications, presentations, workshops, and open educational resources through QUBES. This will encourage adoption, adaptation, assessment, and creation of additional cases that use HT approaches and data to transform life science curricula and adequately equip our next generation of scientists with key quantitative and data science skills.

Permission to Reuse and Copyright

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: go.ncsu.edu/hits.

AUTHOR CONTRIBUTIONS

All authors collaborated to create the manuscript. CG and SR lead the HITS network and led the development and publication of this manuscript. AB, ME, MG-D, AK, PM, KO'T, and US are all HITS participants, and all contributed significantly to the writing and development of the manuscript including figure and table creation.

FUNDING

NSF HITS RCN network (NSF award 1730317).

ACKNOWLEDGMENTS

We appreciate the patience, energy, and wonderful ideas students provided. We the authors are members of the High-throughput Discovery Science and Inquiry-based case Studies for Today's Students (HITS) RCN network (NSF award 1730317). Our goal is to raise awareness of the use of high-throughput approaches and datasets using case study pedagogies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.711512/full#supplementary-material

Supplementary Figure 1 | Distribution of HITS network across North America. The HITS network includes 77 participants from 35 institutions from 2018 to 2020. Each circle represents a unique academic, non-profit, or government institution. Size and color represent the number of participants from each institution. Represented institutions in alphabetical order: Adams State University, Alamance Community College, Allen Institute for Science, Barton College, Campbell University, Carroll University, Central Piedmont Community College, Clarke University, Colorado State University, Davenport University, Elon University, Emory University, Fordham University, Hartwick College, Hastings College, Heidelberg University, High Point University, Meredith College, Morehouse College, National Institute of Environmental Health Sciences (NIEHS), New School, North Carolina Agricultural and Technical State University, North Carolina State University, Northwestern Connecticut Community College, Saint Augustine University, Salem College, Telus World of Science, Trinity Washington University, United States Military Academy West Point, University of Dubuque, University of New Haven, University of North Carolina at Chapel Hill, University of North Carolina at Pembroke, University of Pittsburgh, and University of Richmond.

REFERENCES

Aikens, M. L., and Dolan, E. L. (2014). Teaching Quantitative Biology: Goals, Assessments, and Resources. *MBoC* 25, 3478–3481. doi:10.1091/mbc.E14-06-1045

American Association for the Advancement of Science (2011). *Vision and Change: A Call to Action*. Washington, DC: Final report.

Appleton, E., Densmore, D., Madsen, C., and Roehner, N. (2017). Needs and Opportunities in Bio-Design Automation: Four Areas for Focus. *Curr. Opin. Chem. Biol.* 40, 111–118. doi:10.1016/j.cbpa.2017.08.005

Batut, B., Hiltemann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Blank, C., et al. (2018). Community-driven Data Analysis Training for Biology. *Cel Syst.* 6, 752–758. doi:10.1016/j.cels.2018.05.012

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

Bixler, A., Eslinger, M., Kleinschmit, A. J., Gaudier-Diaz, M. M., Sankar, U., Marsteller, P., et al. (2021). Three Steps to Adapt Case Studies for Synchronous and Asynchronous Online Learning†. *J. Microbiol. Biol. Educ.* 22, 22. doi:10.1128/jmbe.v22i1.2337

Blosser, P. E. (1988). Teaching Problem Solving--Secondary School Science ERIC/SMEAC Science Education Digest No. 2. Available at: https://eric.ed.gov/?id=ED309049 (Accessed May 14, 2021).

Clemmons, A. W., Timbrook, J., Herron, J. C., and Crowe, A. J. (2020). BioSkills Guide: Development and National Validation of a Tool for Interpreting theVision and ChangeCore Competencies. *Life. Sci. Edu.* 19, ar53. doi:10.1187/cbe.19-11-0259

Collins, F. S. (2010). Opportunities for Research and NIH. *Science* 327, 36–37. doi:10.1126/science.1185055

da Silva Rocha, S. F. L., Olanda, C. G., Fokoue, H. H., and Sant'Anna, C. M. R. (2019). Virtual Screening Techniques in Drug Discovery: Review and Recent Applications. *Cur. Top. Med. Chem.* 19, 1751–1767. doi:10.2174/1568026619666190816101948

Donovan, S., Eaton, C. D., Gower, S. T., Jenkins, K. P., LaMar, M. D., Poli, D., et al. (2015). QUBES: A Community Focused on Supporting Teaching and Learning in Quantitative Biology. *Lett. Biomathematics* 2, 46–55. doi:10.1080/23737867.2015.1049969

Duch, B., Groh, S., and Allen, D. (2011). The Power of Problem-Based Learning. Available: http://site.ebrary.com/id/11170657 (Accessed May 14, 2021) Dulles: Stylus Publishing. doi:10.2307/j.ctvt7x6h8

ENCODE (2004). The ENCODE (ENCyclopedia of DNA Elements) Project. *Science* 306, 636–640. doi:10.1126/science.1105136

Farha, M. A., and Brown, E. D. (2019). Drug Repurposing for Antimicrobial Discovery. *Nat. Microbiol.* 4, 565–577. doi:10.1038/s41564-019-0357-1

Feser, J., Vasaly, H., and Herrera, J. (2013). On the Edge of Mathematics and Biology Integration: Improving Quantitative Skills in Undergraduate Biology Education. *Life. Sci. Edu.* 12, 124–128. doi:10.1187/cbe.13-03-0057

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active Learning Increases Student Performance in Science, Engineering, and Mathematics. *Proc. Natl. Acad. Sci.* 111, 8410–8415. doi:10.1073/pnas.1319030111

Gallagher, S. A., Sher, B. T., Stepien, W. J., and Workman, D. (1995). Implementing Problem-Based Learning in Science Classrooms. *Sch. Sci. Math.* 95, 136–146. doi:10.1111/j.1949-8594.1995.tb15748.x

Goodsell, D. S., Dutta, S., Voigt, M., and Zardecki, C. (2021). Molecular Storytelling for Online Structural Biology Outreach and Education. *Struct. Dyn.* 8, 020401. doi:10.1063/4.0000077

Hand, D. J. (2007). Principles of Data Mining. *Drug Safe.* 30 (7), 621–622. doi:10.2165/00002018-200730070-00010

Herreid, C. F. (1994). Case Studies in Science--A Novel Method of Science Education. *J. Coll. Sci. Teach.* 23, 221–229. doi:10.2505/4/jcst18_048_02_34

Herreid, C. F. (2018). Exercises in Style: Is There a Best Way to Write a Case Study? *J. Coll. Sci. Teach.* 48 (2), 34–38. doi:10.2505/4/jcst18_048_02_34

Herreid, C. F. (1997). What Is a Case?. *J. Coll. Sci. Teach.* 27 (2), 92–94.

Hoffman, K., Leupen, S., Dowell, K., Kephart, K., and Leips, J. (2016). Development and Assessment of Modules to Integrate Quantitative Skills in Introductory Biology Courses. *Life. Sci. Edu.* 15, ar14. doi:10.1187/cbe.15-09-0186

Hutter, C., and Zenklusen, J. C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond its Data. *Cell* 173, 283–285. doi:10.1016/j.cell.2018.03.042

Jordan, T. C., Burnett, S. H., Carson, S., Caruso, S. M., Clase, K., DeJong, R. J., et al. (2014). A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students. *mBio* 5, e01051. doi:10.1128/mBio.01051-13

Judes, G., Rifaï, K., Daures, M., Dubois, L., Bignon, Y.-J., Penault-Llorca, F., et al. (2016). High-throughput "Omics» Technologies: New Tools for the Study of Triple-Negative Breast Cancer. *Cancer Lett.* 382, 77–85. doi:10.1016/j.canlet.2016.03.001

Koroshetz, W., Gordon, J., Adams, A., Beckel-Mitchener, A., Churchill, J., Farber, G., et al. (2018). The State of the NIH BRAIN Initiative. *J. Neurosci.* 38, 6427–6438. doi:10.1523/JNEUROSCI.3174-17.2018

Krynock, K. B., and Robb, L. (1996). Is Problem-Based Learning a Problem for Your Curriculum? *Ill. Sch. Res. Dev.* 33, 21–24.

Lombardi, M. M. (2007). Authentic Learning for the 21st century: An Overview. *Educause Learning Initiative* 1, 1–12.

Lopatto, D., Alvarez, C., Barnard, D., Chandrasekaran, C., and Chung, H.-M. (2008). Undergraduate Research. Genomics Education Partnership. *Science* 322, 1684–12685. doi:10.1126/science.1165351

Lopatto, D., Hauser, C., Jones, C. J., Paetkau, D., Chandrasekaran, V., Dunbar, D., et al. (2014). A Central Support System Can Facilitate Implementation and Sustainability of a Classroom-Based Undergraduate Research Experience (CURE) in Genomics. *Life. Sci. Eduse* 13, 711–723. doi:10.1187/cbe.13-10-0200

McCombie, W. R., McPherson, J. D., and Mardis, E. R. (2019). Next-Generation Sequencing Technologies. *Cold Spring Harb. Perspect. Med.* 9, a036798. doi:10.1101/cshperspect.a036798

McDonald, D., Hyde, E., Debelius, J. W., Morton, J. T., Gonzalez, A., Ackermann, G., et al. (2018). American Gut: An Open Platform for Citizen Science Microbiome Research, *mSystems* 3, e000031. doi:10.1128/mSystems.00031-18

Miller, S. (2014). Collaborative Approaches Needed to Close the Big Data Skills gap. *J. Organisation. Design.* 3 (1), 26–30. doi:10.7146/jod.9823

National Research Council (2009). *A New Biology for the 21st Century*. Washington DC: National Academies Press. doi:10.17226/12764

National Research Council (2003). *BIO2010: Transforming undergraduate education for future research biologists*. Washington DC: National Academies Press.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science* 349, aac4716. doi:10.1126/science.aac4716

Pegoraro, G., and Misteli, T. (2017). High-throughput Imaging for the Discovery of Cellular Mechanisms of Disease. *Trends Genet.* 33, 604–615. doi:10.1016/j.tig.2017.06.005

Porter, A. L., Garner, J., and Crowl, T. (2012). Research Coordination Networks: Evidence of the Relationship between Funded Interdisciplinary Networking and Scholarly Impact. *Bioscience* 62 (3), 282–288. doi:10.1525/bio.2012.62.3.9

Prud'homme-Generaux, A., Schiller, N. A., Wild, J. H., and Herreid, C. F. (2018). Guidelines for Producing Videos to Accompany Flipped Cases. *J. Coll. Sci. Teach.* 46 (5), 40–48. doi:10.2505/4/jcst17_046_05_40

Regev, A., Teichmann, S. A., Lander, E. S., Amit, I., Benoist, C., Birney, E., et al. (2017). The Human Cell Atlas. *eLife* 6, e270401. doi:10.7554/eLife.27041

Rosenwald, A. G., Russell, J. S., and Arora, G. (2012). The Genome Solver Website: A Virtual Space Fostering High Impact Practices for Undergraduate Biology. *J. Microbiol. Biol. Educ.* 13, 188–190. doi:10.1128/jmbe.v13i2.444

Ryder, E. F., Morgan, W. R., Sierk, M., Donovan, S. S., Robertson, S. D., Orndorf, H. C., et al. (2020). Incubators: Building Community Networks and Developing Open Educational Resources to Integrate Bioinformatics into Life Science Education. *Biochem. Mol. Biol. Educ.* 48, 381–390. doi:10.1002/bmb.21387

Shaffer, C. D., Alvarez, C. J., Bednarski, A. E., Dunbar, D., Goodman, A. L., Reinke, C., et al. (2014). A Course-Based Research Experience: How Benefits Change with Increased Investment in Instructional Time. *Life. Sci. Edu.* 13, 111–130. doi:10.1187/cbe-13-08-0152

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

Tapprich, W. E., Reichart, L., Simon, D. M., Duncan, G., McClung, W., Grandgenett, N., et al. (2021). An Instructional Definition and Assessment Rubric for Bioinformatics Instruction. *Biochem. Mol. Biol. Educ.* 49, 38–45. doi:10.1002/bmb.21361

Thiry, H., Laursen, S. L., and Hunter, A.-B. (2011). What Experiences Help Students Become Scientists?: A Comparative Study of Research and Other Sources of Personal and Professional Gains for STEM Undergraduates. *J. Higher Edu.* 82, 357–388. doi:10.1080/00221546.2011.1177720910.1353/jhe.2011.0023

Torp, L., and Sage, S. (2002). *Problems as Possibilities: Problem-Based Learning for K-16 Education*. 2nd ed. Alexandria, Va: Association for Supervision and Curriculum Development.

Tyers, M., and Mann, M. (2003). From Genomics to Proteomics. *Nature* 422, 193–197. doi:10.1038/nature01510

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., et al. (2019). Applications of Machine Learning in Drug Discovery and Development. *Nat. Rev. Drug Discov.* 18, 463–477. doi:10.1038/s41573-019-0024-5

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

Yadav, A., Lundeberg, M., DeSchryver, M., Dirkin, K., Schiller, N. A., Maier, K., et al. (2007). Teaching Science with Case Studies: A National Survey of Faculty Perceptions of the Benefits and Challenges of Using Cases. *J. Coll. Sci. Teach.* 37, 34–38.

Yanai, I., and Chmielnicki, E. (2017). Computational Biologists: Moving to the Driver's Seat. *Genome Biol.* 18, 223. doi:10.1186/s13059-017-1357-1

Zhu, H. (2020). Big Data and Artificial Intelligence Modeling for Drug Discovery. *Annu. Rev. Pharmacol. Toxicol.* 60, 573–589. doi:10.1146/annurev-pharmtox-010919-023324

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Bioinformatics

ORIGINAL RESEARCH
published: 07 September 2021
doi: 10.3389/fbinf.2021.711463

From In-Person to the Online World: Insights Into Organizing Events in Bioinformatics

*Alessandra Lima da Silva[1†], Ana Paula de Abreu[2,3], Diego Mariano[3†], Felipe Caixeta[1], Fenícia Brito Santos[1†], Fernanda Stussi D. Lage[4†], Gabriel Quintanilha-Peixoto[1†], Heron. O. Hilário[5†], Joicymara. S. Xavier[6,7†], Lucio. R. Queiroz[1†], Nayara Evelin de Toledo[1†], Raphael Tavares[1], Rodrigo Bentes Kato[1†], Roselane Gonçalves dos Santos[1†], Stellamaris Soares[1,2†], Wanessa. M. Goes[1†], Wylerson. G. Nogueira[1†], Thiago. M. Batista[8], José Miguel Ortega[1], Vasco Ariston Azevedo De Carvalho[1], Glória. Regina Franco[1], Raquel. C. de Melo-Minardi[3] and Aristóteles Góes-Neto[1]**

[1]Institute of Biological Sciences, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil, [2]Department of Clinical and Toxicological Analysis, Faculty of Pharmacy, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [3]Department of Computer Science, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Brazil, [4]Genomics for Climate Change Research Center, Universidade Estadual de Campinas (Unicamp), Campinas, Brazil, [5]Conservation Genetics Laboratory, Pontificia Universidade Catolica de Minas Gerais (PUC Minas), Belo Horizonte, Brazil, [6]Rene Rachou Institute (Fiocruz Minas), Belo Horizonte, Brazil, [7]Institute of Agricultural Sciences, Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM), Unaí, Brazil, [8]Environmental Science Training Center, Universidade Federal do Sul da Bahia, Porto Seguro

Bioinformatics is a fast-evolving research field, requiring effective educational initiatives to bring computational knowledge to Life Sciences. Since 2017, an organizing committee composed of graduate students and postdoctoral researchers from the Universidade Federal de Minas Gerais (Brazil) promotes a week-long event named Summer Course in Bioinformatics (CVBioinfo). This event aims to diffuse bioinformatic principles, news, and methods mainly focused on audiences of undergraduate students. Furthermore, as the advent of the COVID-19 global pandemic has precluded in-person events, we offered the event in online mode, using free video transmission platforms. Herein, we present and discuss the insights obtained from promoting the Online Workshop in Bioinformatics (WOB) organized in November 2020, comparing it to our experience in previous in-person editions of the same event.

Keywords: Bioinformatics, computational biology, education, Science popularization, online events

OPEN ACCESS

Edited by:
Sebastian Kmiecik, University of Warsaw, Poland

Reviewed by:
Nancy Wilkins-Diehr, University of California, San Diego, United States
Karsten Hokamp, Trinity College Dublin, Ireland

****Correspondence:***
Aristóteles Góes-Neto
arigoesneto@icb.ufmg.br

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Protein Bioinformatics, a section of the journal Frontiers in Bioinformatics

Received: *18 May 2021*
Accepted: *16 August 2021*
Published: *07 September 2021*

Citation:
da Silva AL, Abreu APde, Mariano D, Caixeta F, Santos FB, Lage FSD, Quintanilha-Peixoto G, Hilário HO, Xavier JS, Queiroz LR, de Toledo NE, Tavares R, Kato RB, dos Santos RG, Soares S, Goes WM, Nogueira WG, Batista TM, Ortega JM, De Carvalho VAA, Franco GR, Melo-Minardi RCde and Góes-Neto A (2021) From In-Person to the Online World: Insights Into Organizing Events in Bioinformatics. Front. Bioinform. 1:711463. doi: 10.3389/fbinf.2021.711463

INTRODUCTION

Education in bioinformatics is a transdisciplinary practice involving extensive knowledge in biology, statistics, mathematics, and data science (LIU et al., 2013). The demand for qualified professionals in this area emerged with the advent of big data, which in biological sciences, corresponds to topics such as DNA and RNA sequencing data, and studies of protein structure (NCBI, 1988; Illumina, 2021). Consequently, integrative educational initiatives in bioinformatics are necessary to introduce and develop the required skills. Such initiatives could also help fulfill the demands of both academia and industry concerning research progress and the creation of new tools, products, and technologies (Wilson Sayres et al., 2017; Rocha, 2021). In Brazil, disseminating the theoretical and practical knowledge in bioinformatics is mainly led by academics linked to public institutions. There are seven registered graduate programs in bioinformatics and one in computational biology, besides a technological course in bioinformatics, which make up the education and research network in this country. Moreover,

other biological sciences courses (e.g., undergraduate courses in biotechnology, biology, and graduate courses in genetics) usually include disciplines related to bioinformatics.

Graduate students play a key role in the teaching of bioinformatics. Some graduate programs and scholarships criteria require students to enroll in teacher training, which is also relevant to science popularization (CAPES, 2021). Disciplines in undergraduate and graduate courses, winter/summer schools and courses, workshops, seminars, and short courses are some of the most common event formats for science popularization in bioinformatics (Ranganathan, 2005; Welch et al., 2014; Atwood et al., 2015). In these formats, lecturers with different levels of training and expertise communicate their knowledge to heterogeneous audiences, contemplating students in universities, technical courses, professionals in the field, and even the university staff (Fuentes-Acosta et al., 2021). The covered subjects might fit into the most diverse segments of bioinformatics, ranging from basic concepts of molecular biology and computer science to Omics, MetaOmics, and Structural Biology (Gauthier et al., 2019).

During 2020, in-person events and classes were unfeasible due to the COVID-19 pandemic. Therefore, universities and students had to adapt to online classes (Mahmood, 2021). Nonetheless, many students do not possess the setup required for online classes, including a quiet room to study, reliable access to the internet, or even a compatible device (computer, tablet, mobile, and others). In addition, many students share a computer with their whole family. Interactive classes try to overcome this challenge by keeping the student concentrated on the computer screen for longer (Ibrahem and Alamro, 2020). Additionally, not seeing the audience (at least on some free streaming platforms) is another challenge for the speakers. During in-person events, speakers stand up in front of their audience and use non-verbal communication, such as reactions and facial expressions to improve their talk (Wood-Harper, 2021). On the other hand, providing an online event avoids many costs, such as travel, food, and accommodation, providing a free and globally accessible meeting that can attract a larger number of attendees (Reshef et al., 2020), despite the quality of learning, which may be inferior.

Considering this scenario, in this work, we bring the experience of the first Online Workshop in Bioinformatics (WOB) compared to our latest in-person event, the Summer Course in Bioinformatics (CVBioinfo). We prioritized engagement through quizzes and surveys during lectures, awards, and other actions to help bring the attendees together. Moreover, other activities, from posts on social media to emails with relevant topics about the online event and short breaks during the event, had a relevant impact. We also prioritized accessibility in WOB, so that all the attendees could participate and use the content without any restriction (at least 12.7 million people in Brazil have some disability or limitation) (IBGE, 2010). This report hopes to provide a model for the science popularization of online events in bioinformatics and to reflect on possible improvements through the feedback of our attendees.

MATERIALS AND METHODS

The attendees were required to fill an application form for participating in both CVBioinfo and WOB. These forms were available for approximately 1 month, comprising quantitative and qualitative sections collecting primary data (e.g., name, age, email address, and so on) and interest data from a list of topics (only for WOB), including computer science (e.g., programming languages, operational systems, and machine learning), omics sciences, and transdisciplinary topics (e.g., ecology, science outreach, and entrepreneurship). In previous CVBioinfo editions, we also required candidates to provide a resumé (generated in the Lattes platform, which is the standardized repository of curricula of Brazilian researchers) and a motivational letter, but this was not required in the last edition (nor did we levy any participation fees). Those surveys allowed not only to gather the data for this study but also guided the speakers of the audience's expertise concerning their presentation topic. In addition, we sent a feedback form to the attendees, enquiring about the overall organization of the course, its content quality and quantity, their satisfaction and willingness to indicate it for a friend and to attend future editions, their satisfaction with the course organization, and suggestions for next editions. For WOB, especially, these forms included an opinion section for the round tables and lectures. We later correlated this data using the Orange Data Mining tool (Demsar et al., 2013).

The Experience of the Summer Course in Bioinformatics in Previous Editions

The Summer Course in Bioinformatics (CVBioinfo) was initially designed to be a showcase of the Graduate Program in Bioinformatics of the Universidade Federal de Minas Gerais (PPGBioinfo-UFMG) and the research possibilities for undergraduates, and then attract new graduate students to the program. The Summer Course in Bioinformatics gradually evolved to interact with a broader audience, as a transdisciplinary bridge between attendees of diverse backgrounds and research in bioinformatics.

CVBioinfo was created in 2017 by graduate students and postdoctoral fellows of the PPGBioinfo-UFMG, supported by the professors and researchers of this graduate program and the university structure. The first edition comprised a diverse range of lectures in omics sciences and structural biology, establishing a successful model for the forthcoming years. Without a participation fee, this first edition had 600 attendees, mainly from UFMG and other institutions in the southeast of Brazil. Besides, seven practical mini-courses were offered on topics such as introduction to programming languages, data visualization, metagenomics, and molecular modeling. This event was an excellent opportunity for non-specialists to grasp the diversity of bioinformatics applications, gathering interest in this area.

In 2018, a symbolic registration fee was introduced to avoid the high evasion rate of the first edition, in which many attendees did not take part in the entire event. The fee was also a way to gain financial independence to plan the subsequent editions. This edition comprised 5 days of lectures and two different short practical courses per day. The number of attendees was now limited to 200 people, selected through a brief motivational letter. The Organizing Committee decided to limit the number of attendees to better serve the attendees and improve their

experience, in addition to adapting the event to the best infrastructure available. The number of applications was 22% higher in the following edition (2019). That year, subscriptions came from more than 70 different universities (including five institutions in other countries) and high school students. Despite this diversity, most of the attendees were undergraduate students from the state of Minas Gerais, in Biological Sciences. In contrast to the underrepresentation of women in STEM fields (Bonham et al., 2017), female participation was 60% in this edition, as computational biology is considered an interdisciplinary STEM field.

Structure of In-Person Editions

Following the aforementioned model, the order of activities was maintained flexible, aiming to maximize attendance and engagement. The mornings of the first day of the event were reserved for an opening lecture, in which the CVBioinfo history and the PPGBioinfo-UFMG structure were presented to attendees by the graduate program coordinator. From 2017–2019, the lectures occurred in the mornings while the afternoons were reserved for theoretical/practical courses. The lectures were 50 min long, with 10 min reserved for Q and A. In the 2020 edition, the courses happened during the mornings and the lectures in the afternoon to stimulate higher attendance rates. Speakers included professors, postdoctoral fellows, and graduate students of UFMG (not limited to any particular program or expertise). Although external speakers were always welcome, their invitation was dependent on the budget of that event. Most of the lectures introduced bioinformatics research applications and theory to the audience, using accessible language whenever possible while highlighting the interplay with other knowledge domains. The main guiding subjects were DNA, RNA, and proteins studies, as well as related technologies and techniques.

Some lectures, mainly in 2020, were dedicated to showing research lines open to new students and were conducted by PPGBioinfo-UFMG professors, while other lectures were reserved for the sponsors of the event to showcase their products and initiatives (always related to applied bioinformatics and biotechnology). Furthermore, some lectures were also in the roundtable format. Three to five speakers were invited to present short introductions on a common topic, followed by a discussion open to the public. The closing lecture was reserved for new researchers/professors that are part of the PPGBioinfo-UFMG alumni to provide an example of the paths traveled from the start of the formation as a bioinformatician to the many viable mature careers.

The Organizing Committee was supported by volunteers (only during event days), and they were indispensable for the proper functioning of the event. These volunteers aided in the space organization, reception of the attendees, and help on minor issues throughout the event. In 2020's in-person event, the whole organization team was composed of 29 people, divided into 14 students and post-doctoral researchers from PPGBioinfo-UFMG as the main organizers of the event, plus 15 graduate students and undergraduates volunteers and the PPGBioinfo-UFMG staff (responsible for activities directly related to the budget). The main organizers were divided into six teams with different responsibilities, such as advertisement, speakers contact, website development, registration management, content creation, and interaction with attendees, encouraging them to clarify their doubts about topics covered at the event. Additionally, short coffee breaks separated the lectures. These intervals, intended for resting, were also crucial for networking, creating opportunities for the attendees to meet each other and the lecturers to extend the discussion on the presented topics, ask spare questions, and exchange contacts. Every edition also had a dinner night before the last day to complement these breaks, where attendees, speakers, and the organizing committee interacted over drinks and regional food.

Structure of the Online Edition

A total of 20 speakers (13 from Brazilian institutions and seven from foreign institutions) were involved. The Brazilian institutions with which our speakers were associated are located in Minas Gerais (Welch et al., 2014), Bahia (NCBI, 1988), São Paulo (NCBI, 1988), and Rio de Janeiro (LIU et al., 2013), while the Brazilian speakers abroad were located in Australia (NCBI, 1988), the United States of America (NCBI, 1988), Germany (LIU et al., 2013), and England (LIU et al., 2013). For this reason, all lectures were taught in Portuguese. Eleven of these institutions are public, and eight, private. The Organizing Committee of the online edition consisted of 17 students and researchers from PPGBioinfo-UFMG (one M.Sc., 12 Ph.D. students, and four postdoctoral researchers). The organizers were divided into teams, similar to the in-person event. In addition, another team supported live lectures with live translation to Brazilian sign language (LIBRAS) from an accessibility and inclusion core at UFMG, which offers this service free of charge.

The selection of the speakers was based on the knowledge and relevance of their subject, the ability to prepare and deliver interesting short talks, which are essential in online events. Besides, our speakers were highly qualified, with 15 PhDs and 5 M.Sc. degrees, most of which are current graduate students from PPGBioinfo-UFMG or alumni. The selection of software used in this edition was focused on open-source, freemium, and low-priced tools. These tools were applied to email marketing, digital design, and the platform to manage the lectures. The lectures were made available on YouTube on live stream and later hosted in the CVBioinfo YouTube channel, with each lecture comprised of a unique video. During the event, we created an individual link for each talk (in favor of a single video containing all the lectures of that given day). After each talk, attendees were redirected to the next lecture. This was thought of to ease access to content and avoid program confusion by the attendees. Additionally, we provided interactive activities to engage and encourage the attendees to discuss the talks and ask questions to the speakers (Mariano et al., 2021).

RESULTS

In order to illustrate the differences between organizing online and in-person events, we show here a brief comparison between

TABLE 1 | Comparison of structure in online and in-person editions.

Structure	WOB (Online)	CVBioinfo (in-person)
Reach	Worldwide	Local and visitors
Travel (Speakers)	No	Yes
Travel (Attendees)	No	Yes
Dependence on internet	Yes	No
Practical courses	No	Yes
Required infrastructure	Broadcast platform	Conference rooms with projector and enough seats for attendees and speakers

FIGURE 1 | Proportion of scientific themes in bioinformatics from the schedule of lectures of CVBioinfo and WOB events. The main themes of the lectures were: **(A)** RNA and transcriptomics; **(B)** proteomics; **(C)** genomics; **(D)** data science; and **(E)** career in bioinformatics. Total attendees: WOB = 2,727; CVBioinfo = 221.

the in-person event (CVBioinfo IV) that took place in January 2020 and the online event (WOB) that happened in November 2020 (**Table 1**). To officially apply to WOB, every attendee had to fill an application form to describe their subjects of interest in sciences, be it directly related to bioinformatics or not (**Figures 1**, **2**).

The overall distribution of subjects and specific fields within bioinformatics displayed in each event shifted between editions. For example, while genomics and proteomics kept their share of the program, RNA and transcriptomics-related lectures had a higher presence on the online event than in the previous in-person version. Moreover, entrepreneurial and career-oriented presentations for bioinformatics professionals increased, especially in the schedule of WOB. Nonetheless, a considerably smaller space in the schedule was dedicated to data science (**Figure 1**). We also evaluated the topics that incite the most interest among those enrolled. The list of topics included 25 topics related to computing and biology, which were defined empirically (**Figure 2**). For this question, each student could select up to five answers. The five topics of greatest interest to event attendees were: "Genomics/ Metagenomics" (n = 1,439), "Sequencing" (n = 1,395), "Python" (n = 1,245), "R" (n = 1,014), and "Transcriptomics" (n = 931). Additionally, we correlated the most important topics of interest to identify which combination of subjects could arise the greatest interest from the attendees (**Figure 2**; **Supplementary Figure S2**). The heatmap retrieved three main groups composed of attendees with high interest in (i) computational topics (such as R, Python, Perl, and other programming languages); (ii) omics (such as genomics, transcriptomics, and proteomics); and (iii) exclusively biological themes. This information will be used, for example, to guide the choice of topics and speakers in future editions of the event. Another difference between the in-person and online events was the structure of the organizing committee. For the in-person events, the organizing committee had the support of a group of volunteers from graduate PPGBioinfo-UFMG and undergraduate students. Nevertheless, during WOB, the members of the Organizing Committee were sufficient to manage all the activities that include streaming the lectures and interact with the attendees on the chat, giving information about the event, answering questions, and collecting questions sent to the speakers.

The online event had a higher number of attendees enrolled (2,727) when compared to the in-person event (221). Most of the attendees in CVBioinfo were between 22 and 28 years old, while for WOB, at least half of attendees were between 22 and 31 years old

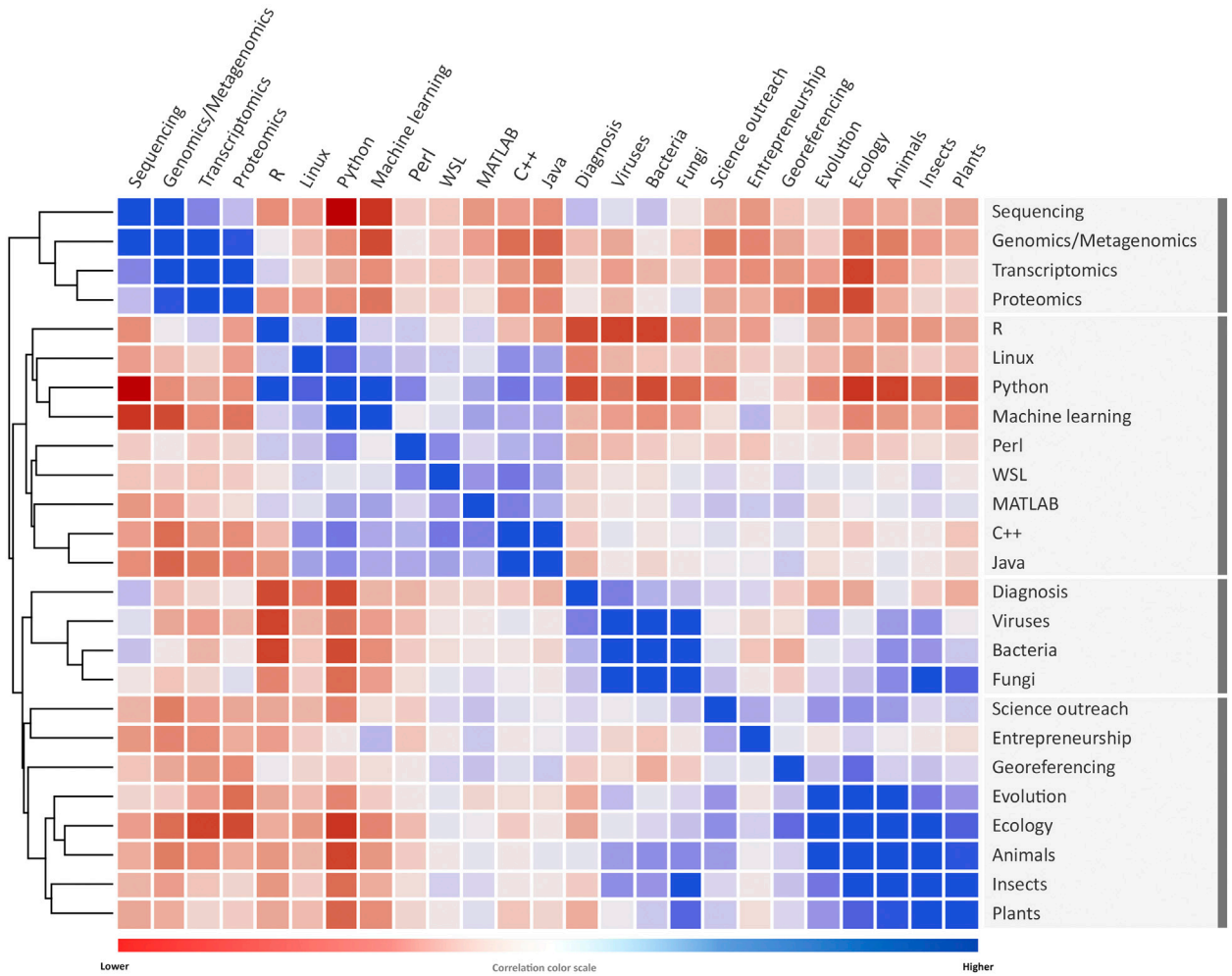


FIGURE 2 | Correlation color matrix among most interest topics of attendees enrolled in the UFMG's Online Workshop on Bioinformatics 2020 (WOB). From 2,727 students registered, we asked the question: "What are the five topics you are most interested in learning at WOB20?". Moreover, we evaluated the correlation of topic pairs. In this visualization, blue cells represent a higher correlation between topics of interest, and red indicates a lower correlation (i.e., blue suggests that students interested in topic X–column–are also interested in topic Y–row). Some clusters were highlighted in rectangles with black borders and calculated using Euclidean distance. Figure generated using Orange Data Mining (https://orangedatamining.com/).

(**Figure 3A**), representing a more diverse age group. The youngest attendee was 14 years old, while the oldest was 71 years old. Additionally, female attendees represented 52% of the CVBioinfo, against 48% of males, while in the online event, the number of female attendees increased to 59% (**Figure 3B**). The majority of the attendees were undergraduate students for both events (104 for CVBioinfo and 1,152 for WOB), and the number of attendees with bachelor degrees ranks as the second-highest for the CVBioinfo and third-highest for WOB (**Figure 4**); **Figure 5** shows the countries in Latin America while comparing the percentages of attendees in the in-person event (CVBioinfo, on the left) and the online event (WOB, on the right), displaying a broader distribution of the origins of the attendees. Nevertheless, like the in-person event, most of the attendees were from the State of Minas Gerais (MG) in WOB. Other states with a high percentage of attendees are Sao Paulo (SP) and Bahia (BA). Furthermore, the online event allowed students from other countries to attend.

We requested feedback forms from the officially subscribed attendees of both events; however, each event form has probed for different information. CVBioinfo feedback aimed to assess the background knowledge of the attendees before the event, while WOB forms screened for the attendees' interpretation of their own experience during the online edition. For the in-person setup, attendees described their skills on a scale ranging from zero to six, zero being no knowledge at all, and six related to an expert. From that group, mostly reported scores lower than four, that is, reported none or very basic level of understanding before the event. For example, 88.2% of all the feedbacks reported score lower than four in programming languages, 89.6% for GNU/Linux, 92.4% for molecular modeling, and 88.9% for bioinformatics in general. Meanwhile, only 7.7, 11.9, and 11.2% of the attendees that answered the feedback form declared to have no actual knowledge of genomics, proteomics, and transcriptomics, respectively. Both feedback

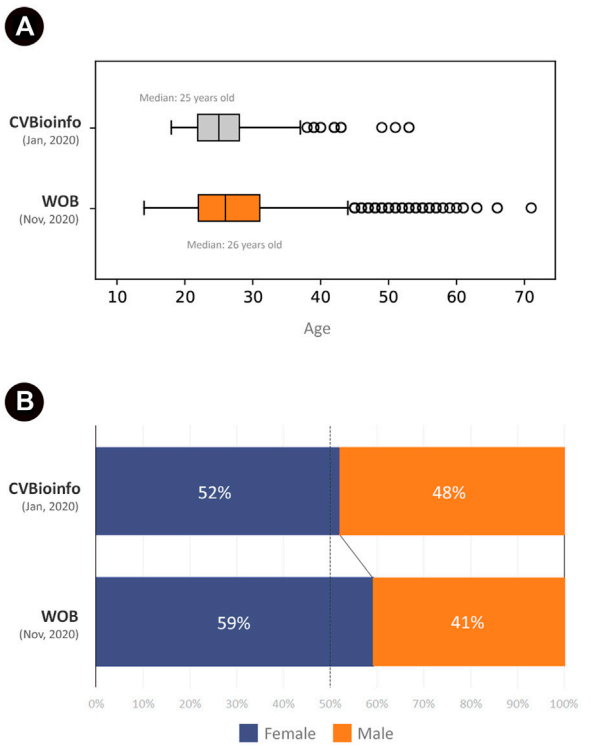


FIGURE 3 | Attendees grouped by age and sex. Total attendees: WOB 2,727; CVBioinfo 221. **(A)** Quartile (0.25, 0.5, 0.75): 22, 26, and 31 (WOB); 22, 25, and 28 (CVBioinfo). Mode: 22 (WOB) and 25 (CVBioinfo). Mean: 27.46 (WOB) and 26.44 (CVBioinfo). Median: 26 (WOB) and 25 (CVBioinfo). Max value: 71 (WOB) and 53 (CVBioinfo). Min value: 14 (WOB) and 18 (CVBioinfo). **(B)** Male/Female proportion in CVBioinfo (upper stacked bar) and WOB (lower stacked bar). Conecting line between bars indicate change in values.

tables are available in Supplementary Data (answers in Portuguese). Gathering a total of 143 replies (**Supplementary Table S1**), the answers collected by the CVBioinfo feedback form highlighted the knowledge gap previously displayed by the attendees of the in-person event.

For the online event, the focus of the feedback form was to connect with the quality of the experience for the attendees, especially considering this was the pilot trial for the online version of this event in bioinformatics. Gathering a total of 725 replies on the form (**Supplementary Table S2**), the attendees displayed high degrees of satisfaction with the adaptation of the event format for the online platforms (98.3%), scheduled program (99.0%), round conference tables (93.6%), and general organization (98%). Nonetheless, despite having been successful in its format, 42.1% of the attendees reported experiencing problems with internet connection during the Livestream, and 29.7% alleged to have difficulties remaining focused throughout the event.

To reinforce our values regarding inclusivity, besides the already mentioned LIBRAS interpreters, WOB partnered with a social project called Codigo X (available online at https://codigox.ong.br/), as a charity campaign. Codigo X is a non-governmental organization (NGO) that aims to introduce socially vulnerable girls in the technology field. As the online event had no participation fee, we encouraged the attendees to donate any value to support the activities of Codigo X. The approval feedback for both initiatives, language sign interpreters and the philanthropic campaign, was 99.2 and 97.8%, respectively.

DISCUSSION

Following a worldwide trend aimed to reimagine scientific conferences during the Sars-CoV-2 pandemic in 2020 (Hanaei et al., 2020; Jarvis et al., 2020), we compared the experiences of our online and in-person events for the popularization of Bioinformatics. Our attendee profile is slightly different between the events, with the average age of 25–26 years old, which is, in fact, our target audience: senior undergraduate students (Ristoff, 2014) probably interested in a career in bioinformatics. Female students correspond to more than half in both cases (52–59%) (IBGE, 2010). This proportion is equally displayed in our student team, where 57% of the organizers are women. Nonetheless, most lecturers were male for both CVBioinfo (69%) and WOB (62%). This reflects the changes in female access to college education in Brazil and abroad, which has increased since the 1980s to reflect the population proportion (Barros and Mourão, 2018). Furthermore, we can also correlate this proportion with the current preference for college majors in Brazil, also reviewed by Barros and Mourão (Barros and Mourão, 2018), in which men constitute the majority of the students in computational sciences, mathematics, and statistics majors. Although the data reviewed by these authors was relative to 2015–2017, approximately, female misrepresentation in some STEM fields continues to be a persistent obstacle reviewed elsewhere (Maarkert, 1996; Cheryan et al., 2017). According to (Moore, 1987), this gap seems to be present from the very beginning of the increase of female access to higher education in the 1960s and is still reflected in higher education staff throughout the late 1970s and early 1980s.

Besides a slight increase in female participation in the online event, the age range is also more diverse in this version, with more attendees over 40 years old. Although Ristoff described this cohort as a relevant part of undergraduate students, around 7.89% (Ristoff, 2014), in our experience, this cohort corresponds to graduated, employed professionals. Their participation might indicate a search for new methods and techniques already applied in their research since bioinformatics is now an indispensable part of different fields of expertise (ecology, epidemiology, pharmaceutics, and others). To contemplate this cohort, initially unexpected, we ought to bring more lectures on these topics of interest, in addition to our current panoramic perspective on bioinformatics, customized for a broader public. This category includes attendees in many different circumstances, for instance, people who have not completed their M.Sc. courses yet and some who did not attend graduate school, as well as people working in the industry, research centers, or even unemployed.

As the main positive aspect of our online event, the wide geographical distribution of the attendees was much broader than the in-person event. In CVBioinfo, we did not have more than five

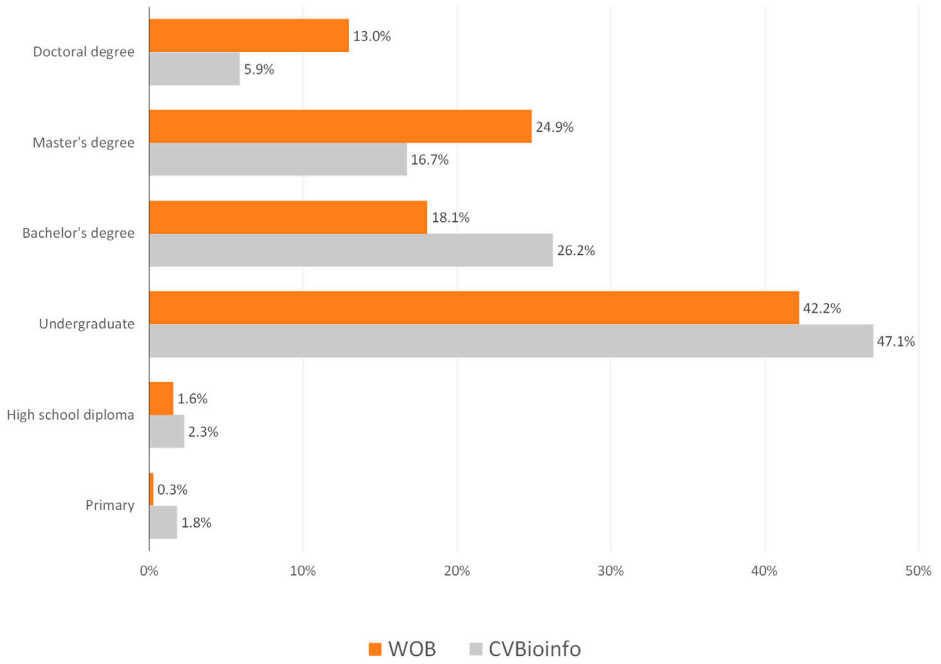


FIGURE 4 | Attendees are grouped by education levels (percentual values). For M.Sc. and Ph.D. degrees, we considered completed or in progress. Total attendees: WOB = 2,727; CVBioinfo = 221.

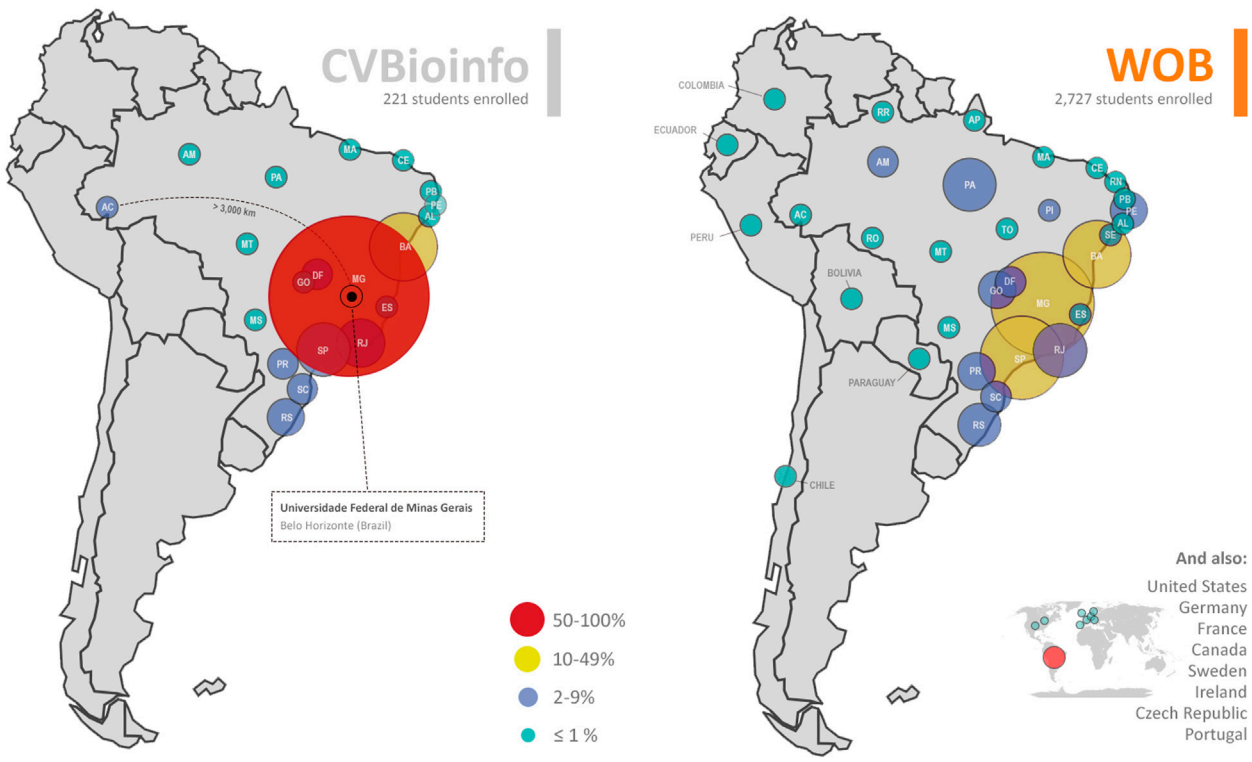


FIGURE 5 | Comparison between the percentage of registrants in the face-to-face event called CVBioinfo 2020 **(left)** and in the online event, named WOB 2020 **(right)**. CVBioinfo - Summer Course of Bioinformatics 2020 (n = 221) and WOB–Workshop Online on Bioinformatics 2020 (n = 2,727). Registrations grouped by Brazilian states are shown on the bar plot in the center. Figure generated using Python (Plotly and pandas libraries), Microsoft Excel, and Adobe Photoshop.

attendees from countries other than Brazil. On the other hand, the proportion of local attendees (from the metropolitan region of Belo Horizonte and the state of Minas Gerais as a whole) was much higher, likely due to transportation and accommodation expenses. In both events, neighboring states such as Bahia and Sao Paulo appeared with the highest number of attendees just after Minas Gerais. In contrast, the online event enabled the participation of individuals from the other Brazilian States that never attended all the previous editions of in-person events, such as from the Northern states (Amapa, Rondonia, and Roraima), from the Amazonian region, and also attendees from the states of Rio Grande do Norte, Piaul, and Sergipe, in the Northeastern region.

Besides Minas Gerais (the state where the event is situated), most of the other states displayed a significant proportional increase in participation, except for Acre, Alagoas, and Bahia. The Universidade Federal do Para (UFPA) possesses a highly active bioinformatics research group (available online at http://pctguama.org.br/), which can be correlated with their increased participation in the online event. Curiously enough, the Universidade Federal do Rio Grande do Norte (UFRN) possesses a Multidisciplinary Environment (available online at https://bioinfo.imd.ufrn.br/) and a Graduate Course in Bioinformatics. Nevertheless, it was quite surprising that the participation of this state in both the in-person and online events was still considerably low. This calls for a more collaborative effort for the scientific outreach in bioinformatics, similar to the RECOM Network (available online at http://www.recom-network.com/).

As exhibited in **Figure 2**, we can identify three major groups of interesting topics for the attendees. The first group was formed by Omics (Genomics, Transcriptomics, and Proteomics). The second group is formed by Computer Science, attendees interested in learning the GNU/Linux, programming language (R, Python, Perl, MatLab, C ++, Java), and machine learning while the third group showed greater diversity and correlation but mainly focused on Ecology (Viruses, Bacteria, Fungus, Animals, Insects, Plants, and Evolution). Curiously, cross-cutting themes such as scientific dissemination and entrepreneurship were correlated to this interest and not in a separate branch. The correlated areas displayed the need that each individual had according to their current training. Most of those enrolled were from areas focused on biological and biomedical sciences, and these students usually have experience in highly targeted research. By attending courses such as ours, which seek interdisciplinarity, the attendees can complement what they miss in their career formation and, thus, can search for complementary areas of interest.

In the in-person event (CVBioinfo), the attendees had the opportunity to participate in hands-on training. These training courses allow attendees to learn practice topics of interest to bioinformatics, complementing the topics covered in the lectures. In our experience, attendees tend to be very interested in hands-on training, and places used to be very competitive. In contrast, in this trial edition of WOB, we did not offer workshops and hands-on training, which are essential to most educators in Bioinformatics. A perspective to subsequent editions could be a model including hands-on training, short and interactive lectures as seen in (Abrahamsson and Dávila López, 2021), tools that involve the attendees as quizzes and sweepstakes, reinforcing the methods described by (Mariano et al., 2021), and activities to reach a greater regional diversity since the online event has worldwide coverage.

The increase in participation from states other than Minas Gerais resulted in more geographical diversity in our events, especially in the lectures. The student feedback was accessed through questionnaires that were sent to the attendees after the online event. Selected answers can be found in our Supplementary Data. While in the CVBioinfo event, the topics of most significant interest were related to the computing area (GNU/Linux, programming languages, and molecular modeling). This is probably due to that most of the attendees who had reported less knowledge were directly enrolled with these topics however, in WOB, the reported interest was much more spread. The topics of interest in WOB were focused on four main areas: Programming languages, Sequencing and Omics, Population and System Biology, and Machine Learning. Despite the great demand, WOB could not cover Programming language subjects and opened an excellent opportunity to improve the model hereafter. Overall, the attendees reported high satisfaction with the online event.

During all the events, there was a concern with interdisciplinarity. Thus, the lectures and courses did not focus only on computer science or biological sciences but, instead, they encompassed the intersection of these topics, so that the program was divided by theme and its applications, ranging from basic research to applications in industry. The invited professors and researchers belonged to different departments (Pharmacy, Biochemistry, Microbiology, Computer Science, Innovation, etc.), institutions (public and private), and companies (national and international). This was reflected in the satisfaction feedback with the diversity of available content. Therefore, this is a critical approach to be considered for the formation of the event schedule (Farina and Penof, 2020).

In-person CVBioinfo events had intervals between lectures, as well as dinner night aimed at promoting contact between attendees and CVBioinfo lecturers and PPGBioinfo professors. Those moments promoted interactions providing the opportunity for new partnerships to be established, as well as possibilities for attendees interested in attending PPGBioinfo as graduate students to find possible advisors. In contrast, the interaction between attendees and lecturers in the online event was limited to the live chat during the lectures and the contacts and social media the lecturers usually provided in their presentations. Nonetheless, we believe that the online event helped to promote our graduate program to a new public, which, in the case of graduate students from other institutions, might bring the possibility of scientific collaboration, for instance. To attract the attendees and overcome the loss of interaction, we promoted a gamification-based engagement strategy. This strategy consisted of grouping the attendees, based on their interest areas, into four distinct groups that competed against each other. An individual survey was sent to the attendees with

questions related to the content of the lectures that occurred during the 3 days of the event. Attendees were then able to interact with other group members through exclusive discussion rooms. There was a high number of live messages among attendees during the competition, indicating that the strategy aided an increase in engagement and interaction between attendees (Mariano et al., 2021).

CONCLUSION

Our results showed that the online event (WOB) we had offered increased access and was more inclusive when compared to our previous in-person events. Furthermore, organizing the online event was more straightforward compared to the in-person ones (CVBioinfo), especially regarding staff, since it was composed of a much smaller group of people than the in-person events, which, in turn, had larger groups with different responsibilities. Nevertheless, as the central negative aspect, our online event reduced the interactivity among participants. Moreover, problems with the internet connection (and the very access to the web) and difficulties to stay focused during the event were considered limiting factors. Therefore, we suggest that future educational events in bioinformatics should ensure that skills and information are more accessible, addressing the desires and expectations of the attendees. Thus, there is the perspective of creating thematic rooms or groups for the next events to be carried out by the Organizing Committee, aiming to increase the interaction between attendees and speakers. Most importantly, we suggest a hybrid format of in-person and online events.

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

Event organization: AS, AA, DM, FC, FS, FL, GQ-P, HH, JX, LQ, NT, RTS, RK, RGS, SS, WG, WN, TB. Data Analysis: AS, AA, DM, FC, FS, FL, GQ-P, HH, JX, LQ, NT, RK, RGS, SS, WG, WN. Document Writing: AS, AA, DM, FC, FS, FL, GQ-P, HH, JX, LRQ, NT, RK, RGS, SS, WG, WN. Document Review: AS, AA, DM, FC, FS, FL, GQ-P, HH, JX, LQ, NT, RTS, RK, RGS, SS, WG, WN, TB, JO, VA, GF, RM-M, AG-N.

FUNDING

An early edition of the event (the 2018 edition) was funded by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES) under grant number 88887.138234/2017-00.

ACKNOWLEDGMENTS

We would like to thank all people involved in organizing both the events cited here, past and present. We would also like to thank the coordinators of PPGBioinfo-UFMG, who have supported and encouraged us into organizing these events; the administration staff Sheila Magalhaes, Tiago Silva, Marcia Natalia Ramalho, and Fernanda Magalhaes. Thank you to all the speakers for their time and effort to make these events happen, and equally to the Núcleo de Acessibilidade e Inclusão of UFMG which provided Brazilian Sign Language interpreters for the online edition. Finally, we thank CAPES, CNPq, and FAPEMIG for funding the M.Sc. and Ph.D. scholarships of our team members, and to UFMG for providing the spaces used for the in-person events.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2021.711463/full#supplementary-material

REFERENCES

Abrahamsson, S., and Dávila López, M. (2021). Comparison of Online Learning Designs during the COVID-19 Pandemic within Bioinformatics Courses in Higher Education. *Bioinformatics* 37 (Suppl. ment_1), i9–15. doi:10.1093/bioinformatics/btab304

Atwood, T. K., Bongcam-Rudloff, E., Brazas, M. E., Corpas, M., Gaudet, P., Lewitter, F., et al. (2015). "GOBLET: The Global Organisation for Bioinformatics Learning, Education and Training,". Editor L Welch (San Francisco, CA: PLOS Comput Biol [Internet), 11, e1004143. Available from: https://dx.plos.org/10.1371/journal.pcbi.1004143. doi:10.1371/journal.pcbi.1004143

Barros, S. C. da. V., and Mourão, L. (2018). *PANORAMA DA PARTICIPAÇÃO FEMININA NA EDUCAÇÃO SUPERIOR, NO MERCADO DE TRABALHO E NA SOCIEDADE*. Recife, PE: Psicol Soc [Internet], 30. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-71822018000100214&lng=pt&tlng=pt

Bonham, K. S., and Stefan, M. I. (2017). in *Women Are Underrepresented in Computational Biology: An Analysis of the Scholarly Literature in Biology, Computer Science and Computational Biology*. Editor C. T Bergstrom (San Francisco, CA: PLOS Comput Biol [Internet]), 13, e1005134. Available from: https://dx.plos.org/10.1371/journal.pcbi.1005134. doi:10.1371/journal.pcbi.1005134

CAPES (2021). Plataforma Sucupira Brasilia, DF: Coordenacao de Aperfeicoamento de Pessoal de Nivel Superior (CAPES). Available at: https://sucupira.capes.gov.br/sucupira/.

Cheryan, S., Ziegler, S. A., Montoya, A. K., and Jiang, L. (2017). Why Are Some STEM fields More Gender Balanced Than Others? *Psychol. Bull.* 143 (1), 1–35. doi:10.1037/bul0000052

Demsar, J., Curk, T., Erjavec, A., Gorup, C., Hocevar, T., Milutinovic, M., et al. (2013). Orange: Data Mining Toolbox in Python. *J. Mach Learn. Res.* 14, 2349–2353.

Farina, M. C., and Penof, D. G. (2020). *AÇÕES DE INTERDISCIPLINARIDADE NA EDUCAÇÃO SUPERIOR: UMA AVALIAÇÃO COM BASE NA ANÁLISE DE REDES SOCIAIS*, 36. Sao Caetano do Sul, SP: Gestão Reg [Internet].

Available from: https://seer.uscs.edu.br/index.php/revista_gestao/article/view/5706. doi:10.13037/gr.vol36n107.5706

Fuentes-Acosta, M. A., Mulia-Rodríguez, J., and Osorio-González, D. (2021). Use of Bioinformatics Technologies and Databases to Teach Analysis of Genetic Sequences to Undergraduate Students in Physics, Biotechnology, and Biology: The Specific Case of the SARS-CoV-2 Spike Protein. *Ce* 12 (01), 193–202. doi:10.4236/ce.2021.121014

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A Brief History of Bioinformatics. *Brief Bioinform* 20 (6), 1981–1996. doi:10.1093/bib/bby063

Hanaei, S., Takian, A., Majdzadeh, R., Maboloc, C. R., Grossmann, I., Gomes, O., et al. (2020). *Emerging Standards and the Hybrid Model for Organizing Scientific Events during and after the COVID-19 Pandemic*. Cambridge, UK: Disaster Med Public Health Prep [Internet], 1–17. Available from: https://www.cambridge.org/core/product/identifier/S1935789320004061/type/journal_article

IBGE (2010). CENSO Brasilia, DF: Instituto Brasileiro de Geografia e Estatistica (IBGE). Available at: http://www.censo2010.ibge.gov.br/. Accessed August 23, 2021.

Ibrahem, U. M., and Alamro, A. R. (2020). Effects of Infographics on Developing Computer Knowledge, Skills and Achievement Motivation Among Hail University Students. *Int. J. Instr.* 14 (1), 907–926. doi:10.29333/iji.2021.14154a

Illumina (2021). NovaSeq 6000 Sequencing System Guide San Diego, CA: Illumina. Available at: https://support.illumina.com/downloads/novaseq-6000-system-guide-1000000019358.html. Accessed August 23, 2021.

Jarvis, T., Weiman, S., and Johnson, D. (2020). Reimagining Scientific Conferences during the Pandemic and beyond. *Sci. Adv.* 6 (38), eabe5815, 2020 . Available from: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.abe5815. doi:10.1126/sciadv.abe5815

Liu, W., Sun, Z., Gong, E., and Xie, H. (2013). The Comparison of Introductory and Advanced Bioinformatics Courses. *J. Electr. Electron. Educ.* 6.

Maarkert, L. R. (1996). Gender Related to Success in Science and Technology. *J. Technol. Stud.* 22 (2), 21–29. doi:10.21061/jots.v22i2.a.4

Mahmood, S. (2021). Instructional Strategies for Online Teaching in COVID -19 Pandemic. *Hum. Behav Emerg Tech* 3 (1), 199–203. doi:10.1002/hbe2.218

Mariano, D., Nogueira, W. G., Goes, W. M., dos Santos, R. G., Kato, R. B., Toledo, N., et al. (2021). Uma estratégia para engajamento de participantes de eventos online. *BIOINFO - Rev. Bras Bioinformática e Biol. Comput.* doi:10.51780/978-6-599-275326-19

Moore, K. M. (1987). Women's Access and Opportunity in Higher Education: toward the Twenty-first century. *Comp. Educ.* 23 (1), 23–34. doi:10.1080/0305006870230104

NCBI (1988). National Center for Biotechnology Information (NCBI) Bethesda, MD: National Library of Medicine (US). Available at: https://www.ncbi.nlm.nih.gov/. Accessed April 06, 2017.

Ranganathan, S. (2005). Bioinformatics Education-Pperspectives and Challenges. *Plos Comput. Biol.* 1 (6), e52–7. doi:10.1371/journal.pcbi.0010052

Reshef, O., Aharonovich, I., Armani, A. M., Gigan, S., Grange, R., Kats, M. A., et al. (2020). How to Organize an Online Conference. *Nat. Rev. Mater.* 5 (4), 1–4. Available from: http://www.nature.com/articles/s41578-020-0194-0. doi:10.1038/s41578-020-0194-0

Ristoff, D. (2014). O novo perfil Do campus brasileiro: uma análise Do perfil socioeconômico Do estudante de graduação. *Avaliação (Campinas)* 19 (3), 723–747. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1414-40772014000300010&lng=pt&tlng=pt. doi:10.1590/s1414-40772014000300010

Rocha, M. (2021). "Organizing Committee," in *BOD X - 10th Bioinformatics Open Days*. Braga, PT: Universidade do Minho, 1–62.

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 (3), e1003496, 2014 . Available from: https://dx.plos.org/10.1371/journal.pcbi.1003496. doi:10.1371/journal.pcbi.1003496

Wilson Sayres, M. A., Hauser, C., Sierk, M., Robic, S., Rosenwald, A. G., Smith, T. M., et al. (2017). Bioinformatics Core Competencies for Undergraduate Life Sciences Education. *bioRxiv* 13 (6), 1–20. doi:10.1371/journal.pone.0196878

Wood-Harper, T. (2021). "Emerging EdTechs amidst the COVID-19 Pandemic," in *Fostering Communication and Learning with Underutilized Technologies in Higher Education* (Hershey, USA: IGI Global), 93–107. doi:10.4018/978-1-7998-4846-2.ch007

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

ORIGINAL RESEARCH
published: 23 September 2021
doi: 10.3389/feduc.2021.725702

The Development of a Sustainable Bioinformatics Training Environment Within the H3Africa Bioinformatics Network (H3ABioNet)

Shaun Aron[1]†, Paballo Abel Chauke[2†], Verena Ras[2†], Sumir Panji[2†], Katherine Johnston[2] and Nicola Mulder[2] on behalf of the H3ABioNet Training and Education Work Package*

[1]Sydney Brenner Institute for Molecular Bioscience, University of the Witwatersrand, Johannesburg, South Africa, [2]Computational Biology Division, Institute of Infectious Disease and Molecular Medicine, CIDRI Africa Wellcome Trust Centre, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

OPEN ACCESS

Edited by:
Hugo Verli,
Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Peter Van Heusden,
University of the Western Cape, South Africa
Erik Bongcam-Rudloff,
Swedish University of Agricultural Sciences, Sweden

***Correspondence:**
Shaun Aron
shaun.aron@wits.ac.za

†These authors have contributed equally to this work

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: 15 June 2021
Accepted: 27 August 2021
Published: 23 September 2021

Citation:
Aron S, Chauke PA, Ras V, Panji S, Johnston K and Mulder N (2021) The Development of a Sustainable Bioinformatics Training Environment Within the H3Africa Bioinformatics Network (H3ABioNet). Front. Educ. 6:725702. doi: 10.3389/feduc.2021.725702

Bioinformatics training programs have been developed independently around the world based on the perceived needs of the local and global academic communities. The field of bioinformatics is complicated by the need to train audiences from diverse backgrounds in a variety of topics to various levels of competencies. While there have been several attempts to develop standardised approaches to provide bioinformatics training globally, the challenges encountered in resource limited settings hinder the adaptation of these global approaches. H3ABioNet, a Pan-African Bioinformatics Network with 27 nodes in 16 African countries, has realised that there is no single simple solution to this challenge and has rather, over the years, evolved and adapted training approaches to create a sustainable training environment, with several components that allow for the successful dissemination of bioinformatics knowledge to diverse audiences. This has been achieved through the implementation of a combination of training modalities and sharing of high quality training material and experiences. The results highlight the success of implementing this multi-pronged approach to training, to reach audiences from different backgrounds and provide training in a variety of different areas of expertise. While face-to-face training was initially required and successful, the mixed-model teaching approach allowed for an increased reach, providing training in advanced analysis topics to reach large audiences across the continent with minimal teaching resources. The transition to hackathons provided an environment to allow the progression of skills, once basic skills had been developed, together with the development of real-world solutions to bioinformatics problems. Ensuring our training materials are FAIR, and through synergistic collaborations with global training partners, the reach of our training materials extends beyond H3ABioNet. Coupled with the opportunity to develop additional career building soft skills, such as scientific communication, H3ABioNet has created a flexible, sustainable and high quality bioinformatics training environment that has successfully been implemented to train several highly skilled African bioinformaticians on the continent.

Keywords: LMIC, training modalities, Africa, bioinformatics training, professional development

INTRODUCTION

Bioinformatics training and education poses many challenges globally for several reasons; it is not a well-established subject that can easily be incorporated into traditional university curricular, it includes elements from multiple disciplines, application of bioinformatics requires practical skills, and the field is constantly and rapidly changing in response to the development of new data generating technologies and novel algorithms. There are also numerous different training audiences that need to acquire skills and knowledge, from wet-lab and other life science bioinformatics users to bioinformatics analysts and bioinformatics engineers, each of which require different levels of competencies (Mulder et al., 2018). Some scientists are even suggesting a nomenclature change from bioinformatics as it does not capture what the field entails (Bourne, 2021). The debate about labels is not within the scope of this paper as we present H3ABioNet's experiences in developing a holistic and diverse, multi-pronged training program in bioinformatics within an African environment. Bioinformatics skills are particularly lacking in low resourced settings, including many African countries. Ten years ago there were only a handful of institutions in Africa with bioinformatics infrastructure and skills, mostly in the middle income countries (Tastan Bishop et al., 2015). Therefore the continent was ill-equipped for a genomics revolution and to analyze and manage the ever-increasing availability of biological data. Though high-throughput data generation equipment is still limited on the continent, it is essential that data generated off samples derived from the continent be analysed by African scientists.

The Human Heredity and Health in Africa (H3Africa) Initiative is funded by the Wellcome Trust and National Institutes of Health from late 2012 to investigate the genetic and environmental basis for human diseases (Rotimi et al., 2014). With the promise of generating genomic data for thousands of African individuals, there was a need to develop bioinformatics capacity in Africa to manage, transfer, store, analyze and interpret the data in the continent. H3ABioNet, the Pan-African bioinformatics network for H3Africa was established in 2012 to develop the computing infrastructure, tools and skills required for this (Mulder et al., 2016; Mulder et al., 2017b). H3ABioNet developed a multi-faceted approach to address the vast bioinformatics training needs across diverse audiences in multiple African countries, each with different levels of infrastructure and expertise. To establish a sustainable bioinformatics infrastructure, it was necessary to build the underlying support mechanisms by training systems administrators in the effective use of computing infrastructures for bioinformatics applications, and local trainers and teaching assistants on how to run and support training interventions. In parallel, H3ABioNet focused on developing fundamental bioinformatics and computational skills in bioinformatics analysts and engineers, and in building introductory followed by specialized intermediate bioinformatics skills in the bioinformatics users. H3ABioNet also developed guidelines for establishing new bioinformatics degree programs at resource limited institutions and has supported the development of such programs. Four years ago, the Fogarty International Center introduced a dedicated funding call through H3Africa for the establishment of bioinformatics postgraduate degree programs, enabling several new degree programs to be established in three African countries that had previously lacked these programs (Shaffer et al., 2019). H3ABioNet has continued to support the Fogarty postgraduate training program through providing trainer support and hosting a joint postgraduate seminar series to provide advice on student projects. H3ABioNet has subsequently focused on short term training interventions that enable professional development in specific topics outside of degree programs. This has been achieved using different modalities from face-to-face courses at various levels, hackathons, webinars and internships to mixed-model online teaching.

Success and Challenges With Traditional Face to Face Training

At the outset of the establishment of H3ABioNet, there was an initial need to develop a critical mass of skilled bioinformatics experts within Africa. While small pockets of skilled bioinformaticians did exist at some of the H3ABioNet nodes who had previously invested in bioinformatics skills development, the majority of the continent had little or no bioinformatics expertise. To tackle this initial lack of skills, a number of face-to-face training events were developed and run at the various nodes within the network across the African continent. Initially, international trainers were sourced where necessary to run earlier courses with the goal of developing local training expertise in specific data analysis topics. Some of these earlier training events took the form of a longer Postgraduate Introductory Bioinformatics workshop which covered different modules over 4–5 weeks to train bioinformaticians, and Train-the-Trainer events focused on developing regional training capacity. This formed the foundation to subsequently develop and host week-long workshops focusing on specialised bioinformatics analysis topics such as metagenomics, microbiome, genome-wide association studies (GWAS), data analytics and next generation sequencing (NGS). While one of the major focus areas of H3ABioNet was to develop bioinformatics capacity across the continent, it was soon realised that the design, planning and running of face-to-face workshops was not a sustainable long-term approach. Apart from the limited reach and immensely high costs associated with running face-to-face workshops, the challenges associated with working in an African context soon came to light.

While some of these challenges are specific to an African context, some may be encountered in other low-to-middle income countries (LMICs). The first challenge was internet access. While good internet connectivity is taken for granted as a necessity in developed countries, limited and slow internet access is a stark reality in most of Africa. Some universities have access to good local internet infrastructure, but issues with bandwidth and download speed restrictions are exacerbated by unexpected power outages and limited backup resources. H3ABioNet addressed this in some of the training courses by

running all practical analyses using the eBiokit, which is a standalone Mac-based hardware resource that contains all the software and databases required for analyses, without the need for an internet connection to the outside world (Hernández-de-Diego et al., 2017). The second challenge, as already alluded to, is an unstable electricity supply, with planned and unplanned rolling blackouts a daily occurrence in several African countries. Slightly less common, but still observed were political unrest, visa bureaucracy, excessive costs for airline flights within Africa and endemic disease outbreaks. The disruptive nature of removing participants from their home institutes for extended periods of time coupled with participants not having access to local infrastructure to work on upon returning to their home institution, meant the face-to-face mode of training was not a long-term sustainable approach. This presented H3ABioNet with an opportunity to develop a new, more sustainable approach to build bioinformatics capacity across the African continent. Until the COVID-19 pandemic, these modes of training ran in parallel as the advantages of personal interaction, networking opportunities, community building and other benefits of face-to-face training are still important to exploit when feasible.

Need to Develop and Combine Alternative Modes of Training

Having built up a critical mass within the network, there was an opportunity to explore additional forms of training modalities. These training modalities and associated resources were borne out of necessity at the various stages of the development of the network and highlights the need for a flexible approach to developing capacity in any field. To complement the face-to-face training, H3ABioNet used training modalities that are targeted to different audiences with specific desired outcomes. Internships provide an opportunity for a researcher to be embedded in a host laboratory and work alongside experts to achieve a particular goal, usually to develop skills for analysis of their data. While the H3ABioNet internship program has been successful, it is a costly endeavour and is very limited in terms of reaching a wider audience. Therefore, H3ABioNet embarked on a new challenge to develop knowledge and skills in a much larger audience using an adaptation of traditional online courses. In our mixed-model approach, we developed courses running over a longer time period in an online format, but to multiple physical classrooms with local support. These have been introduced at both the introductory/overview and intermediate/specialized course levels and have been instrumental in addressing the enormous demand for accessible bioinformatics training across the continent.

The H3ABioNet training environment also includes longer term building of communities from the training courses, further career professional development in scientific communication and grant writing, and increased accessibility to our training resources by making them Findable, Accessible, Interoperable and Reusable (FAIR) (Garcia et al., 2020). Through a combination of all of these aspects of the training program, H3ABioNet has developed a sustainable training environment that has addressed many of the skills gaps on the continent with demonstrable impact on the ability of African scientists to analyse their own data. In this paper we describe how H3ABioNet has developed and expanded its traditional training approaches to create a training environment suited to an African audience and aimed at developing an African bioinformatics community through the establishment of suitable training models and approaches. We further describe the components of this training environment, which has been developed to address multiple audiences and transfer multiple competencies to trainees based on their selected career paths. Finally, we present some results evaluating the impact and accessibility of our training program.

METHODS - DEVELOPING THE H3ABIONET TRAINING ENVIRONMENT

As previously mentioned, H3ABioNet uses a multi-pronged approach to training in both its target audiences and training modalities. The latter includes face-to-face, mixed-model online training and hackathons to enable the transfer of skills in various bioinformatics subject areas, systems administration and additional career development topics. We offer certificates of participation and acknowledgement for all of the workshops we host, and we have found that this encourages active participation and engagement through the value of adding these achievements to participants' CVs for job and fellowship applications.

Our course development uses competencies to drive the curriculum design and we strive to make the resulting training resources as accessible as possible through the curation of training materials. Below we provide further details about each of the modalities and the audiences and competencies that we aim to address with each one.

Switching to a Mixed-Model Training Approach

The need for basic bioinformatics skills in Africa is growing rapidly and H3ABioNet was faced with the challenge of finding a modality to train students and scientists "en masse" in bioinformatics topics, an endeavour that soon proved to be far more difficult than initially anticipated. Many trainees across Africa do not have access to powerful machines or stable internet/electricity, nor do they have access to expensive software or costly training materials. Additionally, many institutes attempting to establish bioinformatics programs do not have much experience in more advanced topics, and often lack experience with the organisational aspect of running these types of courses. Sourcing local trainers with the required expertise to teach specialized topics also proved difficult and costly, particularly where bioinformatics uptake within the region has already been low. This often results in data science courses, particularly within more specialized fields like bioinformatics, becoming very costly and highly competitive, reducing access to bioinformatics training significantly. It quickly became clear that any traditional approach to training would not suffice, and that a unique model would need to be implemented to meet this demand for training in Africa.

Online learning coupled with distance learning seemed the most cost-effective approach, but tackling other challenges like receiving adequate technical and teaching support at remote regions across Africa (at minimal cost) were more complex. It is well known that online courses that fail to provide additional teaching support, and even at times those that do, experience large drop-out and failure rates (Onah et al., 2014; Bawa, 2016). To combat this, the establishment of a local hosting classroom with some minimum infrastructural requirements such as a training room with computers that could connect to the internet and a projector to play lectures, would allow trainees the opportunity to learn (mostly) uninterrupted. An added human capital requirement of 1) someone to coordinate the classroom, 2) onsite support in the form of teaching assistant/s (TA) familiar with the content and, 3) a systems administrator (SA) for technical assistance ensures trainees receive the physical support required throughout the course. Any classroom across Africa that could meet these criteria would be eligible to join the course and become a local hosting site/classroom, once thoroughly vetted.

Since online and distance learning were to be used, course organisers could partner with international subject experts situated across Africa and abroad, to develop and teach course content. The first course to be developed using this mixed-model approach was H3ABioNet's flagship Introduction to Bioinformatics Training Course (IBT), piloted in 2016 (Gurwitz et al., 2017). Since the training was introductory, trainers could develop content around predominantly online tools ensuring infrastructural requirements remained low and participants could repeat at least some analyses once the course concluded. All content was designed to ensure participants on the course developed appropriate competencies (as described by the International Society for Computational Biology's (ISCB) core bioinformatics competencies) (Welch et al., 2014).

Participation in the course for all participants and staff is completely free of charge and voluntary to ensure the course reaches the widest possible audience. This has allowed not only thousands of participants to be trained over the last 5 years in introductory bioinformatics skills, but has also provided staff with the support required to implement this training locally. The detailed course model is described by Gurwitz et al. (2017).

Training an increasing number of scientists in introductory bioinformatics skills meant the need and demand for more specialized/advanced courses, implemented using a similar model, was also growing. The first course established using a similar mixed-model approach was the African Genomic Medicine Training Initiative (AGMT), targeted at nurses across Africa and first run in 2017 (Nembaware and Mulder, 2019). The program uses a similar mixed-model approach, but with an added requirement of a collaborative research project toward the end of the program. The AGMT course has been accredited and forms part of ongoing professional development for nurses. More importantly, AGMT allows for a wider uptake of genomic medicine within primary health care as nurses are trained in a range of relevant topics.

H3ABioNet then piloted an intermediate level microbiome 16 S rRNA data analysis course (IntBT) in 2019 which saw local classrooms across Africa now needing stronger local technical expertise to access and maintain advanced software to perform more complicated and data intensive analyses, typically on a computing cluster environment using containers (Ras et al., 2021). The course allowed local staff the opportunity to pull and maintain containers while being supported by an experienced core team. These containers were then accessed by their local participants to perform in depth analyses of microbiome data. The use of containers had the added benefit of all software and tools remaining available for further analyses once the course ended and allowed local staff the opportunity to receive some systems administration training and support. Ongoing support was also made available via the H3ABioNet Helpdesk to ensure staff and participants could continue to receive support with local project analyses once the course ended (Kumuthini et al., 2019).

Feedback is collected at various time points (before, during and after) throughout the training to continually monitor the trainee's learning experience as well as the experiences of the staff to allow organisers to continually improve the model from year to year. The pilot phases of both the AGMT and the 16 S rRNA data analysis courses were successful (see (Nembaware and Mulder, 2019) and (Ras et al., 2021)) and demonstrates the flexibility and potential of the model for teaching more advanced and specialised topics. The success of the model with more advanced topics led to a partnership with Wellcome Connecting Science in the United Kingdom (UK) and the development of a Next Generation Sequencing Course which was run across 31 classrooms in Africa in 2021 with ~400 participants registered across them; an exciting "next step" in making high quality bioinformatics training more accessible to African scientists.

When the coronavirus pandemic struck during 2020, the flexibility of the mixed-model approach was tested even further as a result of lockdowns across most African countries. The pandemic meant nearly all face-to-face classrooms could not run, leaving the organisers with the difficult task of ensuring effective support was still provided from local staff teams, while also ensuring that the participants still gained the classroom experience without access to an actual classroom and the ability to meet physically. These courses also typically ran across many months with biweekly contact sessions which ended in a live, online Question & Answer session (Q&A) with the module trainer. While previously trainees were situated within a physical classroom, meaning only one online connection was required per classroom, each participant now needed to connect individually. For IBT, this meant over 1000 simultaneous connections, which quickly became a limiting factor as most online conferencing platforms at the time could not handle the large groups with a standard license, and the large number of connections quickly meant a bandwidth too large for trainees to maintain. While materials are made available via a Learning Management Site, Vula (Learning Management System used at the University of Cape Town which is based on the Sakai (https://www.sakailms.org/) platform), where participants may access the content at any time, the live Q&A was frequently not a component that could be shifted/modified. In some instances, trainees also struggled with internet accessibility/stability, which

often disrupted their access to the LMS, Vula and online sessions. Similarly, while Vula allowed structured forums to be made available to participants and staff, this was not enough to effectively support trainees who were now typically very isolated. The organisers thus worked alongside classrooms and found many creative solutions such as: the creation of breakout rooms on Zoom for more of a classroom feel which were accessible during the stipulated contact session times (in the case of IntBT); staggering sign on times to the online platforms with rolling 30 min slots for Q&As with the trainer (in the case of IBT); having some local classrooms create Whatsapp groups/ independent Zoom rooms (used in both IBT & IntBT) and; even at times, using social media to stay in touch. In regions where trainees really struggled with internet accessibility, some classrooms opted to provide materials on removable storage drives or set up contact points where a head TA would meet with participants to transfer materials. H3ABioNet has also made use of Virtual Machines to ensure all trainees have access to a standardised system for more advanced courses. Both the IBT and IntBT courses still ran successfully in 2020 despite the various challenges, illustrating the effectiveness of the mixed-model approach to teaching.

The model has been used successfully for many years and while it holds a great deal of potential for training across many different topics/domains, it does not come without its share of challenges. It is beyond the scope of this paper to provide a detailed account of these challenges, however, a few are briefly discussed here (in addition to the COVID-19 related challenges discussed previously). As mentioned, a major benefit within the IBT course has been the use of predominantly online tools–this has allowed anyone with access to a browser and internet the ability to join the course. A major disadvantage of using mostly online tools, however, is that course materials need to be constantly updated as online resources tend to change and be updated fairly regularly. Course convenors are thus required to regularly engage with course trainers throughout the year to ensure materials are frequently updated and remain current for every iteration of the course. This is in contrast to the 16 S rRNA course (IntBT), where software and tools used within the course are packaged within containers and run locally. Since this course makes use of few online tools, materials mainly require updating when there have been significant updates or changes to the containers, software or workflows used within the course. It also requires course convenors to work closely with trainers as well as software developers, and to remain abreast of major changes to software and tools used within the field. Another key challenge within the current model and one that is faced by numerous training providers globally, is that all course content has historically been developed and provided in English. While this has resulted in a great uptake of materials within English-speaking African countries, it has meant little to no uptake within Francophile/French speaking countries. Accessibility to the course materials for those with other impairments i.e. hearing, visual are also major obstacles that need to be considered if the model is to reach as many people as possible (material accessibility is discussed in more detail later). Finally, the model constitutes a large amount of logistical and administrative overhead, mostly undertaken by (typically) a single course convenor who manages and coordinates the various components of the model. Convenors need to have advanced coordination skills and need to have the ability to deal with, and address a wide range of issues while balancing many competing priorities—often managing cohorts and staff teams in the region of 500 to greater than 1000 people at any given time.

Applied Continuous Skills Development for Scientists, Engineers and Systems Administrators Through Hackathons

Training workshops, in their various formats, provide a controlled environment for the learning and use of different types of bioinformatics resources and tools which is beneficial to audiences being newly introduced to bioinformatics. Part of bioinformatics capacity development involves continuous learning and the practical application of new methods, tools, technologies and leveraging multi-disciplinary skills in a collaborative environment to solve various problems and produce outputs. Science hackathons provide a collaborative environment for scientists to work together over a short period of time fully concentrating on solving problems that enables concomitant skills transfer and learning (Groen and Calderhead, 2015; Aboab et al., 2016; Lyndon et al., 2018). To ensure new skills acquisition and continuous development through multi-disciplinary knowledge sharing via peer learning within the network, H3ABioNet has organized and held a variety of hackathons for audiences with diverse skill sets ranging from bioinformaticians, computer scientists, systems administrators, and statisticians to biologists (Ahmed et al., 2018; Ghouila et al., 2018; Fadlelmola et al., 2021). Each of the H3ABioNet hackathons have their specific target audiences that encourage multi-disciplinarity and are designed to achieve predetermined goals related to specific ongoing H3ABioNet projects that continue post-hackathon.

To address the issues of reproducibility and portability of executing bioinformatics software stacks in heterogeneous computing environments while keeping up with advances in technology, H3ABioNet organized a workflows and cloud computing hackathon (Ahmed et al., 2018; Baichoo et al., 2018). Introductory training on workflow languages (Common Workflow Language and Nextflow) and containerization of bioinformatics software applications for portability were provided to H3ABioNet personnel, including bioinformaticians, software developers and systems administrators (Baichoo et al., 2018; Ahmed et al., 2019). The goal of the H3ABioNet workflows and cloud computing hackathon was to create four containerized workflows that are portable for Genome Wide Associations Studies, imputation, variant calling from NGS data, and the processing and analysis of 16 S rRNA microbiome data. Specific workflow development teams consisted of bioinformaticians and software developers to develop the workflows, and a systems administrator for software containerization (Ahmed et al., 2018). Some participants of the hackathon have gone on to participate in

other Nextflow community organized hackathons outside of Africa, and have provided training on these topics in subsequent workshops. Skills gained from this hackathon are currently being applied to various H3ABioNet projects and have led to the development of workflows and containers used to provide training for advanced topics such as the 16 S rRNA data analysis course (IntBT) (Ras et al., 2021). The GWAS workflow and imputation service was used to provide training and analysis of data to H3Africa consortium members in a bring your own data (BYOD) face-to-face workshop.

Another goal-oriented hackathon, the DREAM of Malaria hackathon hosted by H3ABioNet and other organizations, brought together scientists at various career stages encompassing data generators, bioinformaticians, statisticians and data modellers to investigate various methods for predicting dihydroartemisinin (DHA) sensitivity of *P. falciparum* isolates using genome wide expression profiles (Ghouila et al., 2018). The multi-disciplinary composition of the various hackathon teams enabled more in-depth discussions and knowledge transfer via peer learning between the hackathon participants within and outside of their teams, while testing various methods in preparation for the Malaria DREAM Challenge (Davis et al., 2019). Data modellers gained a much better understanding and appreciation of why biological data can inherently be noisy, while data generators gained more knowledge of statistical modelling and analysis methods for their data (Ghouila et al., 2018). Apart from peer learning and knowledge transfer, hackathons are a great way of improving communication, fostering long term cohesion and collaboration between scientists and engineers while working towards common goals, as in the case of the H3ABioNet hackathon for the development of a genomic medicine and microbiome portal for African genomics data (Radouani et al., 2020; Fadlelmola et al., 2021; Hamdi et al., 2021).

With the onset of the COVID-19 pandemic, H3ABioNet switched from a planned face-to-face hackathon to a virtual hackathon format utilizing Zoom and breakout rooms for a project aimed at creating an African genome reference graph. The virtual hackathon organization and format was quite different from the traditional face-to-face hackathons held by H3ABioNet. An advantage of having a virtual hackathon format that ran over 2 weeks for 4 to 5 hours a day was that more people within H3ABioNet could participate. The average cost per participant for a face-to-face hackathon (~USD 1,500) imposes a limit on the number of individuals to between 25–30 that can participate. Depending on the nature of the goals for a specific H3ABioNet hackathon, a predetermined selection criteria that places emphasis on existing pre-requisite skills is applied for a number of reasons. Some applicants mistakenly interpret a hackathon to primarily be a training workshop, rather than a goal-oriented event. Participation requires a high level of engagement and contribution that draws upon pre-existing knowledge at the scientist, engineer or systems administrator level, while rapidly learning and applying specialized skills. It can also be disheartening for a participant to attend a specific hackathon for which they are not equipped to make any meaningful contributions to at that particular point in time.

The prerequisite skills required to participate in some of the sub-projects for the virtual African genome graph hackathon included familiarity with the use of the version control system Github, graph building tools, Nextflow workflow language, NGS sequence formats and alignment tools and using High Performance Computing (HPC) environments. A number of applicants had some very basic skills or understanding of using Github, Nextflow and little knowledge about how to containerize software. As there was no financial constraint on participant numbers for the hackathon in its virtual format and to be as inclusive as possible, virtual training open to all network members was provided on using version control and Github. Introductory training on Nextflow and software containerization was also provided by participants who attended the previous H3ABioNet workflows and cloud computing hackathon. A specific sub-project within the African genome reference graph hackathon to containerize software used by the other sub-projects was aimed at systems administrators within H3ABioNet. A surprising outcome was the number of applicants who were not systems administrators that signed up specifically to be part of the containerization sub-project, rather than the other sub-projects using various bioinformatics tools. Further enquiries revealed that some of the applicants were bioinformatics students and scientists within H3ABioNet wanting to make their software and tools available in the form of containers. The virtual format of the hackathon provided an interactive environment for bioinformatics scientists to learn how to containerize software from systems administrators, and for interested users to work with experienced bioinformatics engineers to run novel tools in different HPCs while applying the skills learned during the pre-hackathon workshops. Overall, H3ABioNet has made effective use of the hackathon format to provide paths for continuous applied skills development and peer learning for African bioinformatics scientists, engineers and systems administrators as well as advanced users in multi-disciplinary, collaborative environments within Africa.

Internships as a Means to Bridge the Skills Divide

While training workshops and hackathons can impart and strengthen an individual's knowledge and skills base, they might not necessarily be a suitable bridge for a bioinformatics user to gain scientist level expertise in a specialized bioinformatics topic. To address this, H3ABioNet has implemented an internship program since its inception. Internship programs are a form of training that are essentially supervised work experiences over a dedicated period of time and have been found to be impactful, and in some cases essential for budding African scientists at postgraduate level (Mlotshwa et al., 2017). Recognizing that specialized bioinformatics skills are needed to conduct in-depth analysis of genomics data from a myriad of studies, H3ABioNet's internship program was developed specifically for the H3Africa and H3ABioNet consortium, where students and staff members from an H3Africa data generating project can apply to spend time at an H3ABioNet Node and learn to work with, analyse and interpret their data.

H3ABioNet staff and students can apply to spend dedicated time at another H3ABioNet Node or an external institution with established expertise for a specific computational analysis of data that they seek to better understand.

An important requirement for the H3ABioNet internship is that the individual applying for an internship should have a dataset which they will analyse for the duration of the internship, as this will enable them to gain practical in-depth expertise with this type of analysis. Additionally, the H3ABioNet internship program does not cover the costs of data generation, as H3ABioNet is a funded resources project with clear deliverables for capacity development and not research data generation. A proposed analysis plan, expected outcomes and motivation letters from the applicant's current supervisor and hosting institution and supervisor, as well as a plan to disseminate the skills they have learned at their home institution to ensure capacity development, are mandatory with the application. Additional information such as courses attended, presentations and publications are also required to determine whether an applicant has the foundational skills and knowledge to fruitfully undertake the proposed internship. While the internship program was directly affected by the COVID-19 travel restrictions, it proved a highly valuable and successful component of the H3ABioNet training endeavours and has resumed where possible. Further details on the internship program have been highlighted in the training impact section.

Developing Additional Career-Building Skills

Most early-career research scientists in Africa face continent specific and debilitating challenges which include poor access to libraries and online resources such as journals, lack of funds, inadequate support and lack of good research infrastructure and tools (Nchinda, 2002; Mccullough, 2010). It is from this perspective that the H3ABioNet training and educational initiatives are not only based on the technical or scientific knowledge transfer, but also focus on multiple, holistic, interdisciplinary, practical and diverse soft skills that scientists need for their scientific studies and careers, especially in a challenged continent like Africa. H3ABioNet focuses on experiential learning (learning by doing, which is hands-on in scope) that allows participants to benefit significantly from our offerings. Similar to Margaret Hostetter (Hostetter, 2002), the ideology that career development will enable our diverse, previously and currently disadvantaged scientists and especially bioinformaticians to become highly motivated and intensively trained scientists ready for their different careers across Africa is strongly espoused. We regularly offer face-to-face and online workshops (especially since the onset of COVID-19, all of our training has been virtual), webinars, conferences, career development, internships, mentorship and colloquiums all geared towards developing the soft skills of members and affiliates, be they professionals or students. Most of the soft skills training is offered in collaboration with a diverse group of scientists and trainers from across Africa working in different organizations such as H3Africa and the African Academy of Sciences (AAS).

H3ABioNet provides short term and long term soft skills training depending on the context and temporality, for instance the H3ABioNet UCT CBIO Node contributes to an annual, all year postgraduate development program geared towards assisting students to learn how to be better academic writers and presenters. The shorter format training includes face-to-face workshops that are provided during H3Africa consortium meetings to H3Africa fellows, before the onset of the "new normal" imposed by COVID-19 restrictions. To quote from King (King, 2013), it is necessary to develop creative and sustainable ways to help "early career research scientists ascend the professional ladder." Regardless of the limitations inherent in online learning and training such as the lack of human contact, Brancaccio-Taras et al. have proven that effective scientific training can be offered to fellows with impactful results even on virtual platforms such as webinars (Brancaccio-Taras et al., 2016).

A recently run soft skills H3ABioNet workshop was the Scientific Communication Workshop for H3Africa fellows, staff and scientists, offered at the 17th Consortium Meeting held in April 2021. The main purpose of the workshop was to train attendees on different scientific communication platforms, styles and methods for specific needs. The workshop was facilitated by diverse volunteer trainers from across H3ABioNet nodes based in different African countries and the US. Similar to most of the training H3ABioNet provides, trainers conducted the training free of charge in order to offer necessary workshops to participants at no cost. The Scientific Communication Workshop had multiple learning outcomes aimed at addressing the often overlooked skills required in progressing in an academic environment. These included learning the importance of effective communication and writing, good practices for designing eye-catching posters, developing elevator pitches for sharing of research, valuable tips for providing constructive feedback, advice on the value of timely thesis preparation and referencing software, and guidance on how to effectively critique and summarise a research article.

Building Bioinformatics Communities Through Online Training and Support

The H3ABioNet training network spans many African countries (currently more than 16 countries). It has, and, continues to grow in large part due to our mixed-model training approach and blended learning courses. The Learning Management System used within the mixed-model courses, Vula, allows structured forums and a real time chatting tool to be made available to trainees which quickly enables participants from all over the world to connect in real time. Participants from across Africa and across all classrooms use these forums to connect and interact while joining one of our training courses. Since our training is often open to anyone with the prerequisite skills to apply, these courses double as a mechanism to connect local bioinformatics students and/or enthusiasts, to staff and peers at their local classrooms, institutions or regions. This, coupled with the

skills gained as part of various training events allows bioinformatics communities to begin forming, often at remote regions where bioinformatics uptake may not be as high. Many trainees return to act as staff i.e. TAs or SAs, or may even start up additional classrooms in future iterations of courses—further developing local capacity for bioinformatics teaching and training. Trainees are also encouraged to remain in touch with their local classmates and staff for ongoing training or further opportunities to participate in webinars, events and conferences. More recently we have attempted to implement "Open Learning Circles" (https://h3abionet.github.io/LearningCircles/), a coordinated "Mozilla open classrooms style" space where trainees that have attended any of our mixed-model courses could continue to come together, share new content, work on ongoing data challenges or simply continue to connect. H3ABioNet has also created Slack workspaces where participants from various training programmes could continue to connect, share resources and work on coding problems together, well beyond a course concluding. The creation of mailing lists and various social media accounts also ensures trainees are kept up to date on opportunities and allows for ongoing communication between H3ABioNet and these trainees.

One of the skills which is essential for future academics, but seldom included in traditional educational programs is training skills for trainers. There are several train-the-trainer programs developed for different contexts, but few specific to bioinformatics. The European Bioinformatics Institute (EBI) has run bioinformatics train-the-trainer courses for specific bioinformatics applications, and Wellcome Connecting Science has recently developed a train-the-trainer Massive Open Online Course. H3ABioNet has been developing trainers using different approaches. Several H3ABioNet members attended the EBI NGS train-the-trainer course and went on to assist at a Wellcome Connecting Science NGS course and are now trainers on the previously mentioned mixed-model NGS course. A group of African Carpentries (https://carpentries.org/) trainers is currently being established through working with the Carpentries community to host a Carpentries Instructor training course for H3ABioNet members. H3ABioNet has also developed a training guide that describes the processes followed in planning, designing and running courses, including relevant templates for the different steps (accessible at https://doi.org/10.25375/uct.14337806.v1). This is a valuable resource that can be used by others who wish to set up and run any type of training course, with a focus on bioinformatics resources. More informally, students who have participated in H3ABioNet courses are often given an opportunity to be teaching assistants on the next iteration of the course. Working alongside experienced trainers, they gain practical experience in providing training. The establishment of a strong training network has allowed a budding bioinformatics training community to begin forming in Africa which is now gaining traction across other continents too.

H3ABioNet members have also contributed to global efforts to develop training resources for bioinformatics trainers, such as competency frameworks, guidelines for course and curriculum development and a trainer portal, hosted by GOBLET (Global Organization for Bioinformatics Learning, Education and Training: https://www.mygoblet.org/training-portal/trainer-resources/). H3ABioNet initiated and hosted the first Bioinformatics Education Summit in Cape Town in 2019, which brought together bioinformatics trainers and educators from around the world to develop trainer resources. This community of trainers meets monthly and has since organized 2 further summits. Through these and other activities, such as the ISCB Education Committee and COSI (community of special interest), H3ABioNet has developed long term partnerships with key global bioinformatics trainers and training organizations, with whom we have co-organized several training events and are currently working with the global bioinformatics trainer community to develop an online train-the-trainer course which will be run using our mixed-model training approach.

Addressing Gender Inclusion in Data Science

According to Stads and Beintema (Stads and Beintema, 2006) "the science world appears to be greatly affected with gender barriers that disadvantage female scientists in their career development." To address this gender disparity and inequality, H3ABioNet intentionally organizes spaces and training that targets African women in science. The focus on training African women is based on research and statistics that show that there are many barriers that African female scientists face, such as lack of funding, the work and family balance as well as patriarchal gatekeeping (Sonnert, 1999; Prozesky and Mouton, 2019). A few H3ABioNet female scientists are WiDS (Women in Data Science) ambassadors, and in 2021 we organized the inaugural WiDS Africa event which was hosted virtually and intended to engage women in data science across Africa. This was also an important collaborative space where women were encouraged by other women on ways to navigate the very exclusionary science fraternity. At the conference there was an echoing of the need for more women in science communities which has sparked future collaborations necessary for this endeavour within H3ABioNet (Chauke, 2021).

Curation of Training Materials and Improving Accessibility

Running several training events generates a wide range of training materials, from lecture slides to codebooks. H3ABioNet has thus embarked on a rather large effort to make our training materials (and supporting tools such as containers and workflows) more accessible, starting with the implementation of FAIR principles (Garcia et al., 2020) across our training materials and training pages available on our website. In order to make our training materials FAIR i.e. Findable, Accessible, Interoperable and Re-usable, H3ABioNet is ensuring our webpages reflect the most relevant information to ensure optimum findability and accessibility of our materials. The use of tags and keywords across H3ABioNet webpages to improve Search Engine Optimisation (SEO) along with submitting materials to a public repository to assign a permanent identifier, improve

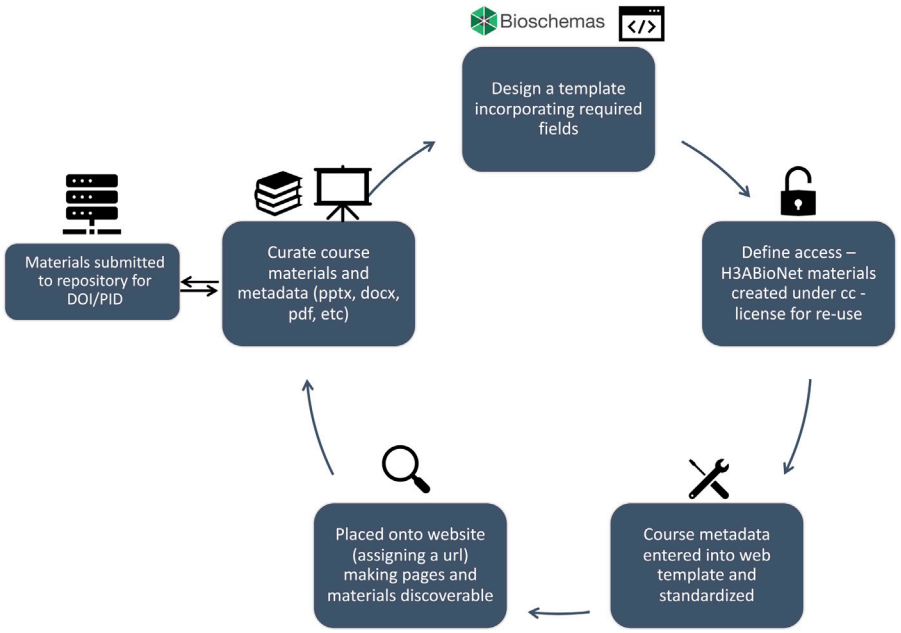


FIGURE 1 | Graphical depiction of the process being used by H3ABioNet to make training materials FAIR.

data provenance and discoverability, and are two ways in which we are improving FAIR compliance across our materials. We have also recently begun the process of implementing bioschemas (https://bioschemas.org/) across our training webpages (**Figure 1**). Bioschemas allow for a standard schema markup to be applied across our pages and materials, making them more discoverable to web scrapers and training databases and repositories like GOBLET (https://www.mygoblet.org/) and TeSS (https://tess.elixir-europe.org/), thus increasing their findability. These tasks are currently underway with the aim to make all H3ABioNet training materials FAIR and freely accessible before the end of 2021 (many materials can already be accessed, freely and openly, via our website and course pages). H3ABioNet is also improving accessibility and reusability of materials by transcribing lectures and videos to add subtitles. The aim is to transcribe all materials into English in the first instance after which transcriptions can then be more easily translated to other common languages spoken in Africa like Arabic and French. Since most materials within H3ABioNet are generally released under a creative commons license, our materials are also free for re-use and distribution by anyone who needs them.

RESULTS AND DISCUSSION ON THE REACH AND IMPACT OF THE H3ABIONET TRAINING ENVIRONMENT

The different approaches discussed above were developed to address specific training needs and to adapt to local and global challenges. The H3ABioNet training environment, summarized in **Figure 2**, includes not only the training interventions for imparting bioinformatics skills, but also career development opportunities, networking and community building, and access to training materials and trainer guidelines and resources. This is to ensure our training has a longer term impact with continued support. For all our training interventions we usually conduct pre- and post-training surveys and evaluations to determine what the participants learned, what they want to learn in future, and their impression of the event. The results of each training event's evaluations and feedback are discussed in post-course meetings and results are sent to trainers or guest lecturers to assess training successes, materials or implementations that may need to be improved. Monitoring, Evaluation and Learning (MEL) are imperative to all our training endeavours at H3ABioNet as we believe that it is necessary to constantly take stock and learn while iteratively providing alternative and additional training solutions. In the section below, we present selected results sourced from some of our experiences, in addition to feedback from a biannual long term evaluation survey sent out to each participant of an H3ABioNet training event. We discuss some of the results in terms of the impact of the various training modalities H3ABioNet has employed on the acquisition of knowledge and skills and the progression of participants' careers.

Training Audiences

Using a multi-pronged approach, H3ABioNet has developed a comprehensive bioinformatics training program for a diverse audience. To date (May 2021), 4,466 unique trainees have attended one or more H3ABioNet training course. Courses are tailored to the audience by designing the curriculum based on the competencies required by participants, taking into account their background and prerequisite skills. A summary of attendee backgrounds is presented in **Figure 3**. Since not all our

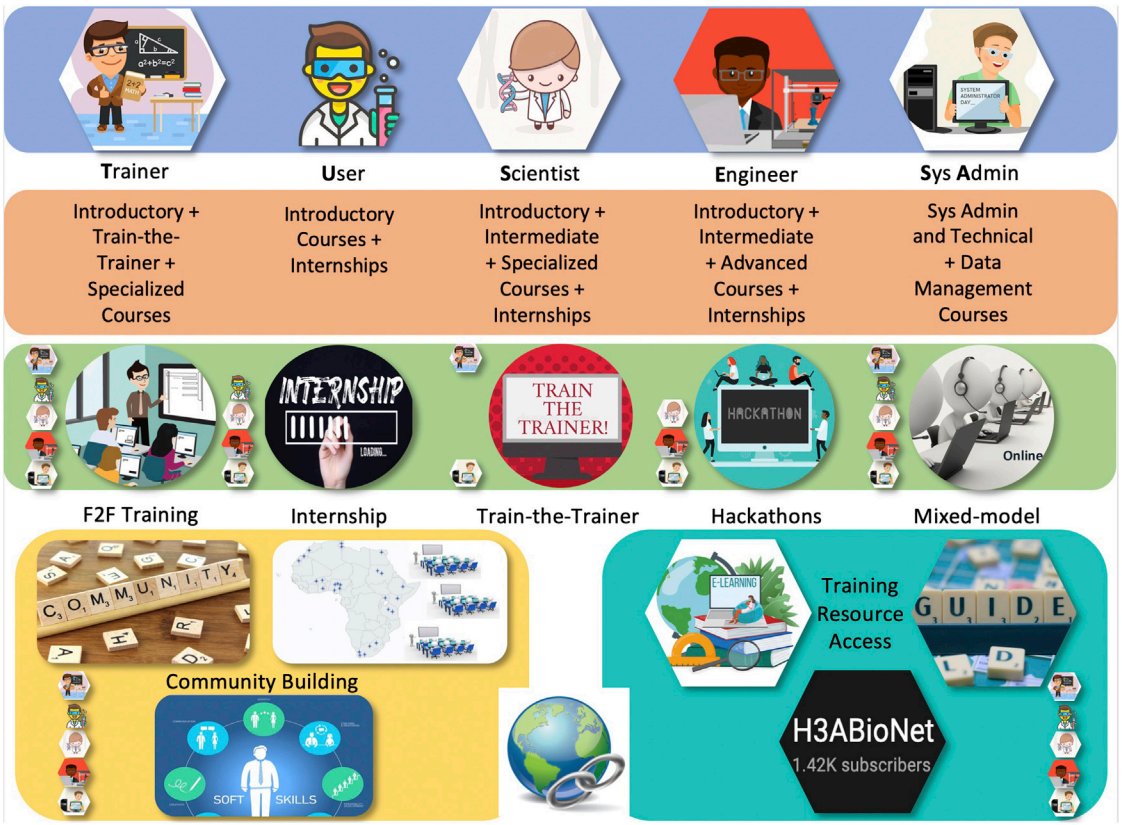


FIGURE 2 | A diagrammatic representation of the H3ABioNet training environment. The blue bar (first tier) represents the different audiences (personas) identified as necessary to develop within the network. The image of each persona is used to represent the choice of training approach/resource in the figure. Courses were designed and presented at varying levels of difficulty (red bar—second tier) to allow for progressive growth of individuals based on their desired backgrounds and planned career paths. Similarly, the complementary modes of teaching (green bar—third tier) presented the opportunity for individuals to explore and learn in unique environments that fostered the support and opportunities required to develop specific competencies and personas. The fourth tier (yellow block) of the training environment provides individuals with the opportunity to learn as part of a larger community as opposed to individual silos and further develop the complementary skills required to excel in building a successful academic career. The final component (teal block) highlights our efforts to make our training resources accessible beyond H3ABioNet to ensure that the efforts to continue training in bioinformatics are sustainable and continue to be utilised by the broader bioinformatics community.

application forms captured information on the background of participants, this data was derived from those participants who completed the long term evaluation survey, but we believe this is generally representative of the broader participant pool. **Figure 3** shows that the majority of trainees are from a life sciences background, predominantly at postgraduate level, and consider themselves to be bioinformatics users. This reflects the large IBT audiences, as these courses are primarily aimed at life scientists who need to use some bioinformatics tools for their research.

In terms of geographical distribution, the initial courses were limited to locations that fulfilled the requirements to host face-to-face workshops. The transition to a mixed-model training approach resulted in an increased reach with several classrooms being able to be hosted for a single course. The interest in the mixed-model approach training courses such as IBT and IntBT has grown each year, with new countries enrolled during each iteration of the course. While we have not yet reached each country on the continent, we have provided training to attendees in a substantial number of African countries as shown in **Figure 4**. Attendees from outside the continent, particularly from the USA and UK, were generally attendees of one of the hackathons. The mixed model courses, in addition to increasing the geographical reach, have resulted in a large increase in the annual number of attendees at our training events (**Figure 5**). Even at the height of the pandemic (2020), we trained over 1200 people.

Training Topics and Course Attendance

The training events have covered a wide variety of topics at introductory, intermediate and advanced levels (**Supplementary Table S1**). The selection of training topics has primarily been driven by the perceived analysis needs within the H3Africa consortium as well as the need to develop highly skilled bioinformaticians within H3ABioNet. In addition to the annual Introduction to Bioinformatics course, which covers 6 different basic bioinformatics modules, the most common topics covered at varying levels of complexity are NGS and GWAS data analysis. While around 70% (3024) of the trainees attended one of our Introduction to Bioinformatics courses (these attract the largest audience per course because of the scalable training

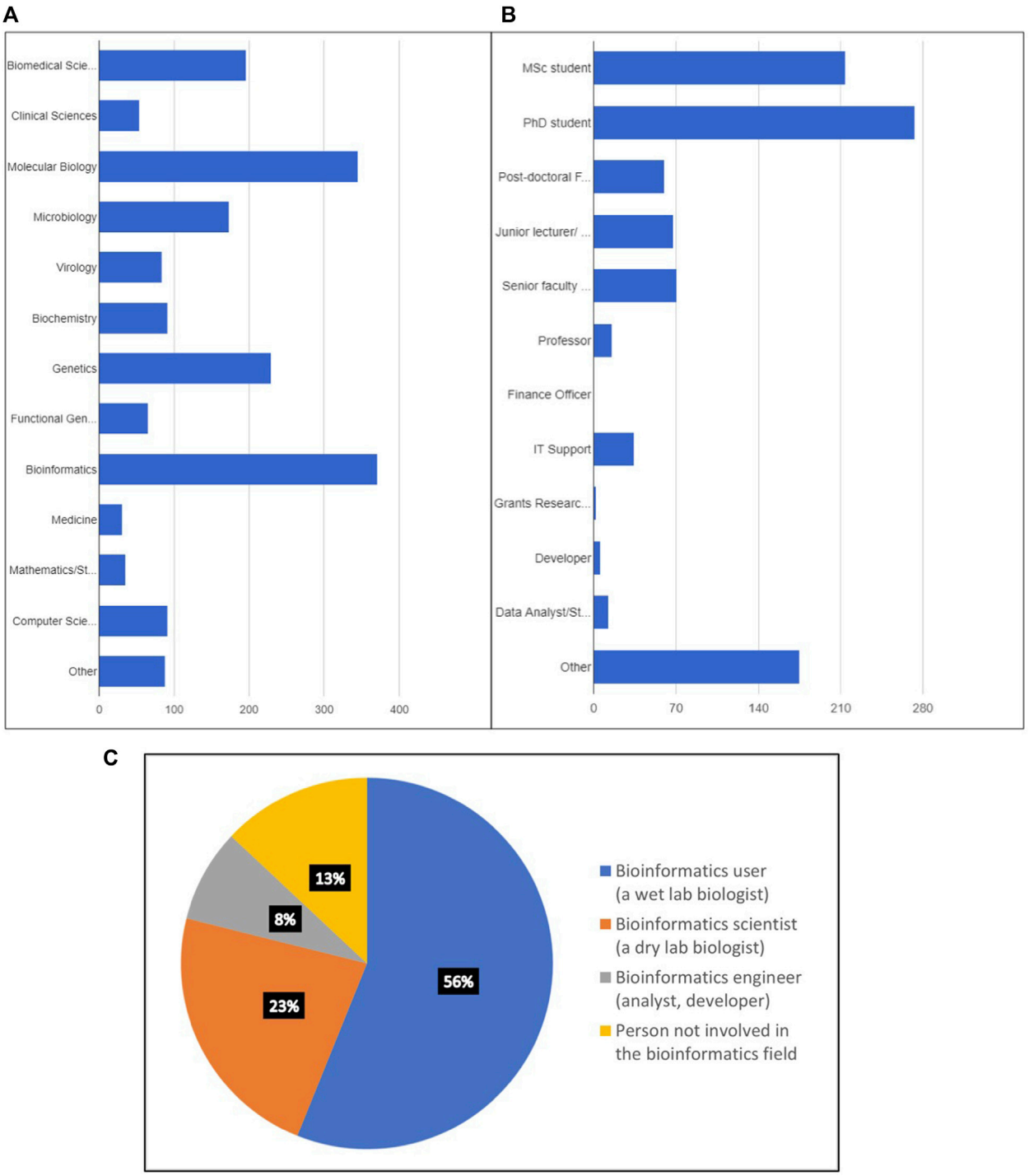


FIGURE 3 | Shows the breadth of different audiences we have trained in terms of **(A)** the academic background of the trainees (Academic Backgrounds: Biomedical Sciences, Clinical Sciences, Molecular Biology, Microbiology, Virology, Biochemistry, Genetics, Functional Genomics, Bioinformatics, Medicine, Mathematics/Statistics, Computer Science, Other), **(B)** their current role or highest educational level (Educational Levels: MSc student, PhD student, Post-doctoral fellow, Junior Lecturer/Lecturer, Senior Faculty Member, Professor, Finance Officer, IT Support, Grants Researcher, Developer, Data Analyst/Statistician, Other), and **(C)** how they perceive their role in the context of bioinformatics.

model), several trainees have attended multiple courses (over 400 people have attended 2 courses and nearly 100 have attended 3 courses, see **Figure 6**) to gain general bioinformatics knowledge, more specialized bioinformatics skills, and career-building skills such as grants management and scientific communication. Many people who attended an IBT course did not go on to do

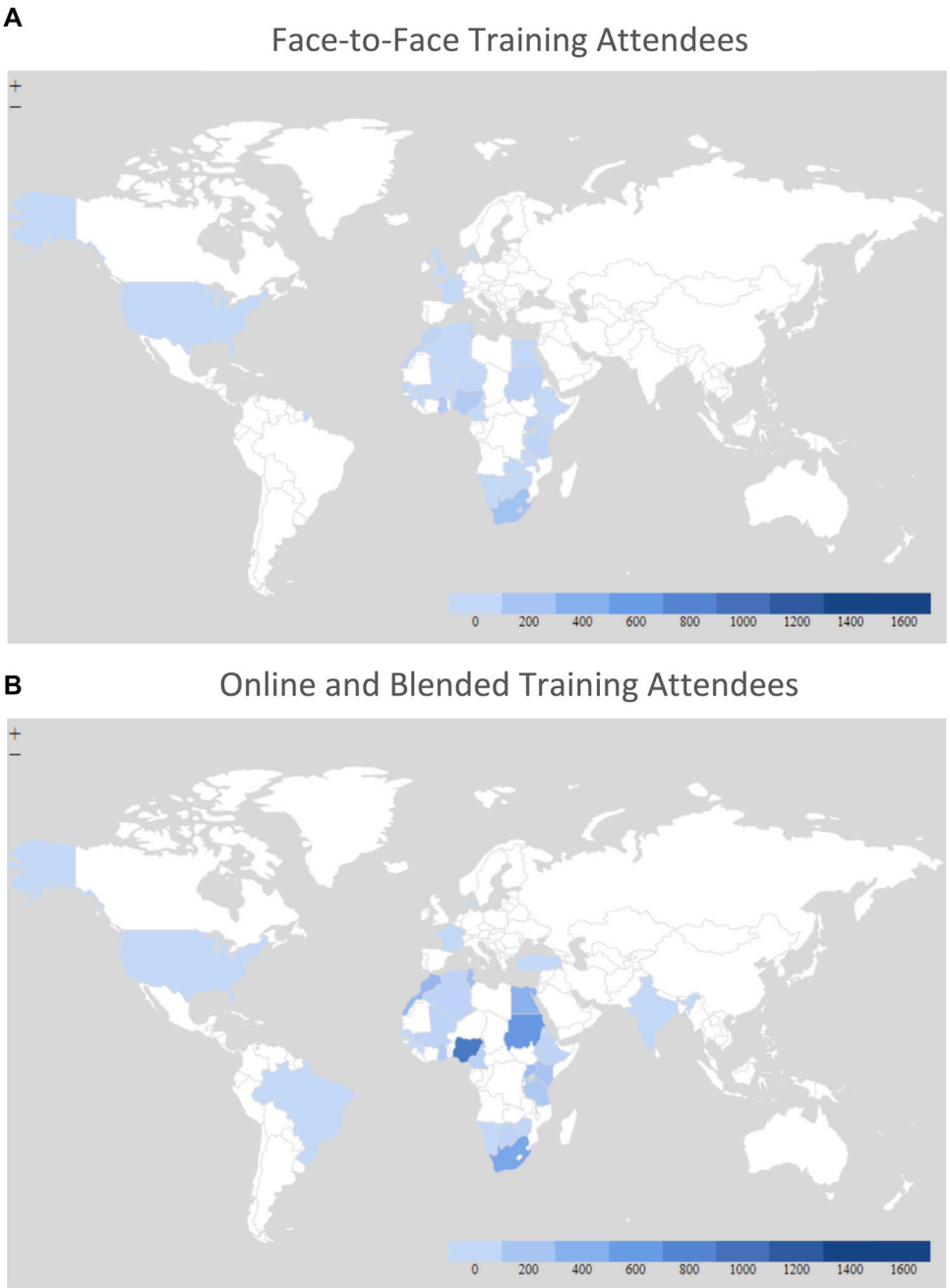


FIGURE 4 | Map showing the distribution of the country of residence for all individuals who have attended any H3ABioNet training event. **(A)** is for face-to-face training events, while **(B)** is for online and blended training events. The darker the shade of blue, the higher the number of individuals. ***Note:*** A number of individual's country of residence was not collected especially in early training events; as a result, in **(A)** 437 individuals (34%) are not represented on the map. In **(B)** 977 individuals (19.5%) are not represented on the map.

further training, though around 200 trainees have also done one of the IntBT courses. It is worth noting that apart from the IBT courses, to cope with the demand, our events are generally limited to H3Africa or H3ABioNet members, so most IBT attendees were not eligible to attend the specialized courses. Nevertheless, the most common pairwise combination of courses attended by H3ABioNet/H3Africa members was an IBT and the 16 S rRNA IntBT (**Figure 7A**), followed by an IBT and a career development course. When grouped by category (**Figure 7B**), the same most popular combinations were found, suggesting that more than 200 of our trainees have gained foundational bioinformatics skills, then specialized by increasing their expertise in a particular data analysis area, and supplemented these with scientific or

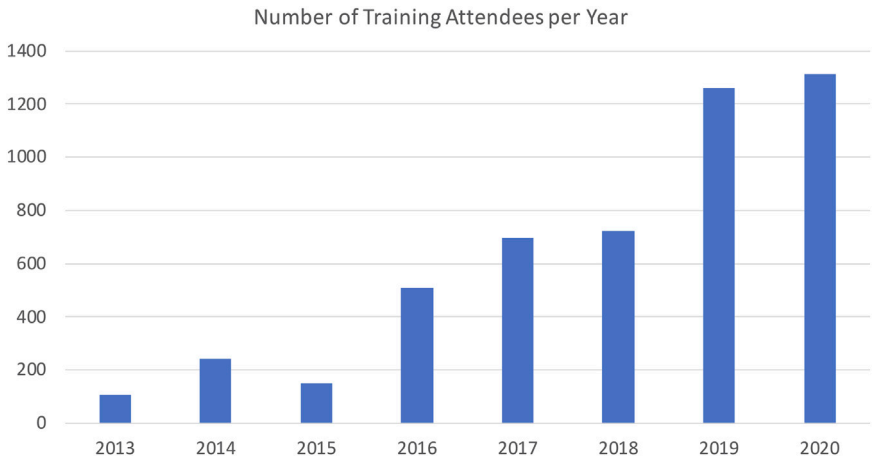


FIGURE 5 | The number of individuals attending H3ABioNet training events per year since the network's inception in 2013. The dramatic increase in 2016 was due to the inclusion of the mixed-model training approach.

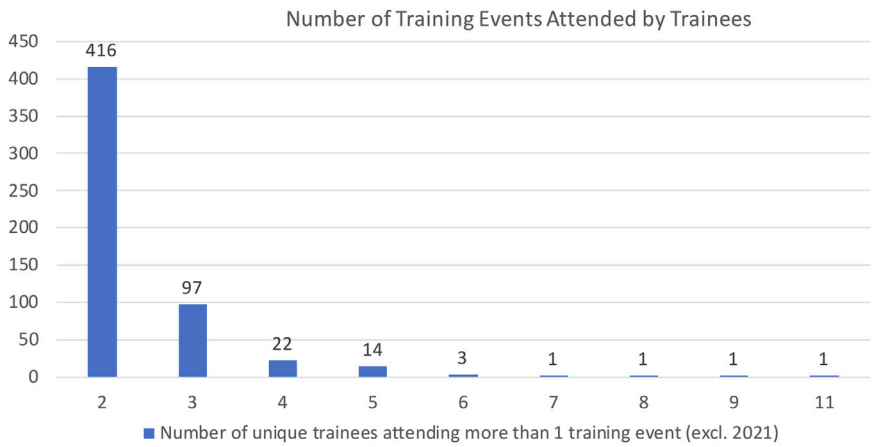


FIGURE 6 | Number of H3ABioNet training events attended by trainees. Most trainees attended just 1 training event (most often IBT, 3777 individuals), however, several went on to attend additional H3ABioNet training events.

grant writing skills. These combinations would serve to build researchers from novice bioinformatics users to more well-rounded academics able to analyse their own data. Many of those who attended face-to-face, intermediate or advanced courses did not also attend an IBT course, presumably because they already had the basic bioinformatics knowledge they required.

Training Modalities

As described earlier, training has been delivered using a variety of approaches and modalities. While there is no overall preference for a single mode of training we did observe that individuals from different backgrounds preferred different modes of training. Based on the experience of running the different modes of training, **Table 1** highlights these preferences as well as some of the advantages and challenges encountered when implementing the different training approaches.

Training Impact

Different training approaches are likely to have a different impact, which, in turn, will be affected by the trainee's background and skills coming into the training intervention as well as their expectations and their engagement during the training. For individuals, internships have enormous potential to provide long term impact on their research and careers, due to the hands-on, personal nature of an internship at an expert facility, and the opportunity to develop partnerships and future collaborations. H3ABioNet has awarded 20 internships to H3Africa and H3ABioNet consortium members that mainly comprise of postgraduate students and staff members who wished to develop expertise and analyse data in topics ranging from structural modelling, GWAS, human variant calling, large scale data transfers to metabolic modelling, pathogen informatics, human population genetics and microbiome studies. Of the 20 H3ABioNet internships

A

	Introductory Bioinformatics	Metagenomics	Professional development	Genomics	Technical/Software	GWAS	Data Management	Education	Intermediate Bioinformatics	NGS	SysAdmin	Advanced Bioinformatics	Research skills
Introductory Bioinformatics	3382	207	32	36	19	17	6	10	28	18	2	15	2
Metagenomics	207	492	8	16	12	9	2	5	6	9	0	1	1
Professional development	32	8	93	0	3	13	8	7	4	4	2	2	0
Genomics	36	16	0	335	5	1	2	3	0	3	0	0	2
Technical/Software	19	12	3	5	78	6	5	4	2	4	2	0	1
GWAS	17	9	13	1	6	105	12	10	10	6	2	1	2
Data Management	6	2	8	2	5	12	81	12	2	5	3	0	1
Education	10	5	7	3	4	10	12	68	1	6	2	1	4
Intermediate Bioinformatics	28	6	4	0	2	10	2	1	62	3	2	21	0
NGS	18	9	4	3	4	6	5	6	3	68	4	0	1
SysAdmin	2	0	2	0	2	2	3	2	2	4	39	0	2
Advanced Bioinformatics	15	1	2	0	0	1	0	1	21	0	0	31	0
Research skills	2	1	0	2	1	2	1	4	0	1	2	0	22

B

	Bioinformatics general - Introductory	Bioinformatics general - Intermediate	Bioinformatics general - Advanced	Career Development	Specialized - Introductory	Specialized - Intermediate	Specialized - Advanced	Specialized - Sys Admin
Bioinformatics general - Introductory	3382	28	15	39	71	247	0	2
Bioinformatics general - Intermediate	28	62	21	4	5	20	0	2
Bioinformatics general - Advanced	15	21	31	2	0	2	0	0
Career Development	39	4	2	172	26	80	0	5
Specialized - Introductory	71	5	0	26	507	85	0	6
Specialized - Intermediate	226	15	2	32	50	619	0	4
Specialized - Advanced	0	0	0	0	0	0	0	0
Specialized - Sys Admin	2	2	0	5	6	0	0	39

FIGURE 7 | Heatmaps showing the total number of trainees who attended combinations of training events (i.e. more than 1 training), the darker blues represent higher figures and therefore the most popular training combinations. **(A)** Shows which other training topics were commonly attended by the same trainee. **(B)** Also shows which other trainings were commonly attended by the same trainee but grouped by training category. Refer to **Supplementary Table S1** for further details on the training events under each training category and topic.

TABLE 1 | A description of the various training modalities preferred by the diverse audience types, highlighting the advantages and disadvantages associated with each modality.

Pros	Cons
Online Training	
Preferred by: Developers and Junior Lecturers	
Cost efficient	Challenge to foster sense of community
Reach a large audience	Challenge to form collaborations
Easy to share material	Licensing of materials could limit accessibility
Face-to-Face Workshops/Internships	
Preferred by: MSc students, Doctoral and Post-doctoral students + IT Professionals + Professors + Senior Professionals	
Dedicated time	Costly/unpredictable logistics
Close interaction	Limited audience reached
Lots of support	Trainees return to environments where they do not have access to software or tools
Hackathons	
Preferred by: Developers, Junior Faculty, Data Analysts/Statisticians	
Defined aims and output	Selection of participants crucial
Develop practical skills	Base knowledge level required
Cross-disciplinary	Limited audience
Mixed-Model Learning	
Preferred by: MSc/PhD Students, Data Analysts/Statisticians and Professors	
Cost efficient	Huge administrative overhead on organisers
Reaches a wide audience	Difficult to deal with local challenges remotely
Easily adaptable	Regular updating of training material is challenging for courses with multiple modules and trainers, or content which changes frequently
Develops local expertise	

awarded, 18 have been between H3ABioNet Nodes with 5 of the internships taking place at the H3ABioNet USA partner nodes, while 13 internships were within H3ABioNet African Nodes, and 2 of the internships have been to external institutions in Europe. The H3ABioNet internships have resulted in work being accomplished that has contributed to a number of publications by the interns, with many of the interns successfully completing their degree programs where applicable, or moving to new positions. In addition, selected internships have resulted in participants gaining strong applied skills in specific bioinformatics areas, creating the foundational knowledge required to contribute to several ongoing larger projects within the H3Africa consortium (Mulder et al., 2017a; Jongeneel et al., 2017; Azarian et al., 2018; Hamda et al., 2018; Ahmed et al., 2019; Choudhury et al., 2020; Sengupta et al., 2021). Due to the COVID-19 pandemic, the H3ABioNet internship program temporarily halted with internship applications having to be deferred to when international air travel was possible between different countries. The most recent H3ABioNet internship award was approved in April 2020, but due to travel restrictions the internship only commenced in March 2021.

The fact that internships are easily defined with agreed upon goals means they have an increasingly important role in education and bridging the gap between a bioinformatics user and a bioinformatics scientist. Internships are arguably one of the best ways for individuals to acquire practical skills and prepare for their careers as they provide many benefits such as learning how different scientific groups work, accessing compute resources, setting up environments for their analysis, and working with dedicated bioinformatics scientists to produce outputs. Internships also provide intangible benefits that include being exposed to new environments, cultures and modes of working, while being immersed in a new environment, they enable networking and provide career-related experiences while facilitating skills transfer (Scott, 1992; Beard, 1998; Cook et al., 2004). These have benefited a number of H3ABioNet internship awardees (Scott, 1992; Beard, 1998; Cook et al., 2004).

Though face-to-face courses and hackathons are ideal for fostering interactions, these were not always possible when trying to reach a wide audience and in negotiating the global pandemic. Nevertheless, whatever the modality used, our training interventions had impact beyond just transferring skills. When trainees were surveyed, nearly 90% (1142 of 1289 responses) indicated that they had shared their new knowledge or the training materials with at least 3 other people, thus extending the reach of our training. While some passed on their knowledge through informal discussions or practical demonstrations, a few did this through hosting workshops. Additionally, approximately 33% (421 of 1289 responses) said the training led to, or facilitated a publication mostly by improving their knowledge and understanding of bioinformatics, through exposure to new ideas or topics, or due to improved data analysis skills. About 30% (413 of 1289 responses) indicated that the training facilitated submission of their thesis, for the same reasons as above. Interestingly, more than 45% (602 of

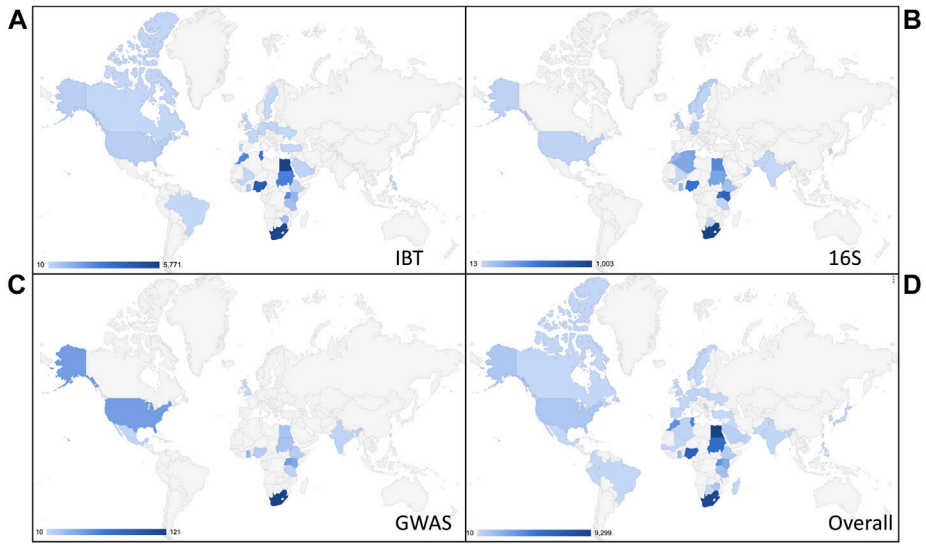


FIGURE 8 | Geographical locations (countries) where H3ABioNet YouTube playlists are being accessed from for the **(A)** Introduction to Bioinformatics Training (IBT) playlist, **(B)** 16 S rRNA data analysis Intermediate Bioinformatics Training Course (IntBT) playlist, **(C)** Genome Wide Association Study (GWAS) playlist and **(D)** H3ABioNet channel views overall. Maps do not depict views from countries with a total number of views <10 and represent the number of overall playlist views (clicks). Individual video views and analytics will thus differ. Colour scales on the maps represent the actual number of playlist clicks/views.

1289 responses) said the training led to a new collaboration, demonstrating the opportunity that training events provide for networking and building relationships with peers.

Accessibility of Training Materials

Although the H3ABioNet training materials have only recently been curated and some of the lectures and videos have been on YouTube for just a couple of years, all materials have been viewed to varying extents after the training event. For one of our GWAS courses, we ran a series of webinars providing theoretical content, followed by a week-long face-to-face hands-on workshop to ensure maximum use of the time together to focus on the practical components of the analysis. These webinar lectures have each been viewed over 1100 times from multiple countries both within and outside of Africa (**Figure 8C**). The IBT video series has between 100 and 3000 views per country on YouTube (**Figure 8A**) and draws a large number of viewers from outside of Africa, with our channel now having well over 100 000 individual video views overall. Despite videos only being available since 2018, they have drawn interest from the global community with a large number of views from the United States, Canada and India (**Figures 8A–D**). In addition, the training pages on our website (www.h3abionet.org/training) have been accessed several thousands of times. Our flagship IBT course pages alone have been accessed between 795 and 15,324 times between 2016 and 2020. Similarly our GWAS training pages have been accessed >1500 times. As we will be one of the first organizations to make our training pages Bioschemas compliant, all H3ABioNet training material should become even more easily findable and accessible. These results demonstrate the immediate impact and growth in uptake of materials by making materials more openly accessible and FAIR.

CONCLUSION

Here we have presented the H3ABioNet training environment which aims to provide a holistic training experience. We target bioinformatics users for introductory training and bioinformatics users, scientists and engineers for more specialized training. A parallel effort to develop local trainers and build grant writing and scientific communication skills ensures sustainability and improved prospects for career progression. While the COVID-19 pandemic has impacted our ability to take advantage of the personal interactions offered by face-to-face events, we were able to reasonably easily adapt to virtual training and continue intensive training throughout the lockdowns. In addition to transferring skills and knowledge, our training environment works towards making training materials FAIR and more accessible to those who are not native English speakers, and enables us to share our experience in training through our training guide and templates. Through the creation of a sustainable training environment, H3ABioNet has provided the foundation for further development of bioinformatics capacity across the continent beyond the lifetime of the network. Though funding for H3ABioNet will end soon, the systems and processes are well documented, and materials are available for the network of trainers that have been developed to pick up and continue at minimal cost. Already, the infrastructure is being leveraged in a separately funded project to roll out training for pathogen surveillance. The enormous demand for bioinformatics training in Africa will hopefully ensure that the legacy continues. While there are many successful training programs world-wide, H3ABioNet

has overcome many of the challenges of working in resource-limited settings and provided a multidisciplinary training program that has reached a very wide audience both directly and indirectly.

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.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Centre for Higher Education Development and the Faculty of Health Sciences Human Research Ethics Committee at the University of Cape Town. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SA, PC, SP, VR, and NM contributed to the conception and design of the study and wrote the first draft of the manuscript. SA, PC, VR, SP, and NM contributed towards the development and running of the various training endeavours of H3ABioNet. KJ analysed the training evaluation results and generated the related figures for the manuscript. All authors contributed to the manuscript and read and approved the final submitted version.

FUNDING

H3ABioNet is supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number U24HG006941. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ACKNOWLEDGMENTS

We would like to express our gratitude to the members of the H3ABioNet Education and Training Work Package who have contributed to much of the work discussed in this paper. We would also like to thank all the trainers who have contributed towards the development and presentation of the H3ABioNet training events as well as all the course participants. Finally we would like to thank Wisdom Akuguru for assisting in extracting and compiling the statistics for the H3ABioNet YouTube account.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.725702/full#supplementary-material

Supplementary Table S1 | | List of H3ABioNet training events by topic and category.

REFERENCES

Aboab, J., Celi, L. A., Charlton, P., Feng, M., Ghassemi, M., Marshall, D. C., et al. (2016). A "Datathon" Model to Support Cross-Disciplinary Collaboration. *Sci. Transl Med.* 8, 333ps8. doi:10.1126/scitranslmed.aad9072

Ahmed, A. E., Heldenbrand, J., Asmann, Y., Fadlelmola, F. M., Katz, D. S., Kendig, K., et al. (2019). Managing Genomic Variant Calling Workflows With Swift/T. *Plos one.* 14, e0211608. doi:10.1371/journal.pone.0211608

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. *AAS Open Res.* 1, 9. doi:10.12688/aasopenres.12847.1

Azarian, T., Mitchell, P. K., Georgieva, M., Thompson, C. M., Ghouila, A., Pollard, A. J., et al. (2018). Global Emergence and Population Dynamics of Divergent Serotype 3 CC180 Pneumococci. *Plos Pathog.* 14, e1007438. doi:10.1371/journal.ppat.1007438

Baichoo, S., Souilmi, Y., Panji, S., Botha, G., Meintjes, A., Hazelhurst, S., et al. (2018). Developing Reproducible Bioinformatics Analysis Workflows for Heterogeneous Computing Environments to Support African Genomics. *BMC Bioinformatics.* 19, 457–513. doi:10.1186/s12859-018-2446-1

Bawa, P. (2016). Retention in Online Courses: Exploring Issues and Solutions—A Literature Review. *SAGE Open.* 6, 1–11. doi:10.1177/2158244015621777

Beard, D. F. (1998). The Status of Internships/Cooperative Education Experiences in Accounting Education. *J. Account. Education.* 16, 507–516. doi:10.1016/s0748-5751(98)00021-9

Ben Hamda, C., Sangeda, R., Mwita, L., Meintjes, A., Nkya, S., Panji, S., et al. (2018). A Common Molecular Signature of Patients With Sickle Cell Disease Revealed by Microarray Meta-Analysis and a Genome-Wide Association Study. *Plos one.* 13, e0199461. doi:10.1371/journal.pone.0199461

Bourne, P. E. (2021). Is "Bioinformatics" Dead? *Plos Biol.* 19, e3001165. doi:10.1371/JOURNAL.PBIO.3001165

Brancaccio-Taras, L., Gull, K. A., and Ratti, C. (2016). The Science Teaching Fellows Program: A Model for Online Faculty Development of Early Career Scientists Interested in Teaching. *J. Microbiol. Biol. Education.* 17 (3), 333–338. doi:10.1128/jmbe.v17i3.1243

Chauke, P. (2021). *Data Science Needs Input from Women*. Johannesburg: Mail and Guardian. Available at: https://mg.co.za/opinion/2021-04-11-data-science-needs-input-from-women/.

Choudhury, A., Aron, S., Botigué, L. R., Sengupta, D., Botha, G., Bensellak, T., et al. (2020). High-Depth African Genomes Inform Human Migration and Health. *Nature.* 586, 741–748. doi:10.1038/s41586-020-2859-7

Cook, S. J., Parker, R. S., and Pettijohn, C. E. (2004). The Perceptions of Interns: A Longitudinal Case Study. *J. Education Business.* 79, 179–185. Available at: https://www.proquest.com/scholarly-journals/perceptions-interns-longitudinal-case-study/docview/202821430/se-2?accountid=15083.

Davis, S., Button-Simons, K., Bensellak, T., Ahsen, E. M., Checkley, L., Foster, G. J., et al. (2019). Leveraging Crowdsourcing to Accelerate Global Health Solutions. *Nat. Biotechnol.* 37, 848–850. doi:10.1038/s41587-019-0180-5

Fadlelmola, F. M., Ghedira, K., Hamdi, Y., Hanachi, M., Radouani, F., Allali, I., et al. (2021). H3ABioNet Genomic Medicine and Microbiome Data Portals

Hackathon Proceedings. *Database: J. Biol. databases curation.* 2021, 1–13. doi:10.1093/database/baab016

Garcia, L., Batut, B., Burke, M. L., Kuzak, M., Psomopoulos, F., Arcila, R., et al. (2020). Ten Simple Rules for Making Training Materials FAIR. *Plos Comput. Biol.* 16, e1007854. doi:10.1371/journal.pcbi.1007854

Ghouila, A., Siwo, G. H., Entfellner, J. D., Panji, S., Button-Simons, K. A., Davis, S. Z., et al. (2018). Hackathons as a Means of Accelerating Scientific Discoveries and Knowledge Transfer. *Genome Res.* 28, 759–765. doi:10.1101/gr.228460.117

Groen, D., and Calderhead, B. (2015). Science Hackathons for Developing Interdisciplinary Research and Collaborations. *eLife.* 4, e09944. doi:10.7554/eLife.09944

Gurwitz, K. T., Aron, S., Panji, S., Maslamoney, S., Fernandes, P. L., Judge, D. P., et al. (2017). Designing a Course Model for Distance-Based Online Bioinformatics Training in Africa: The H3ABioNet Experience. *Plos Comput. Biol.* 13, e1005715. doi:10.1371/journal.pcbi.1005715

Hamdi, Y., Zass, L., Othman, H., Radouani, F., Allali, I., Hanachi, M., et al. (2021). Human OMICs and Computational Biology Research in Africa: Current Challenges and Prospects. *OMICS.* 25, 213–233. doi:10.1089/omi.2021.0004

Hernández-de-Diego, R., de Villiers, E. P., Klingström, T., Gourlé, H., Conesa, A., and Bongcam-Rudloff, E. (2017). The eBioKit, a Stand-Alone Educational Platform for Bioinformatics. *Plos Comput. Biol.* 13, e1005616–14. doi:10.1371/journal.pcbi.1005616

Hostetter, M. K. (2002). Career Development for Physician-Scientists: The Model of the Pediatric Scientist Development Program. *J. Pediatr.* 140, 143–144. doi:10.1067/mpd.2002.121584

Jongeneel, C. V., Achinike-Oduaran, O., Adebiyi, E., Adebiyi, M., Adeyemi, S., Akanle, B., et al. (2017). Assessing Computational Genomics Skills: Our Experience in the H3ABioNet African Bioinformatics Network. *Plos Comput. Biol.* 13, e1005419–10. doi:10.1371/journal.pcbi.1005419

King, L. (2013). Helping Early Career Research Scientists Ascend the Professional Ladder. *Trends Biochem. Sci.* 38, 373–375. doi:10.1016/j.tibs.2013.06.001

Kumuthini, J., Zass, L., Panji, S., Salifu, S. P., Kayondo, J. K., Nembaware, V., et al. (2019). The H3ABioNet Helpdesk: an Online Bioinformatics Resource, Enhancing Africa's Capacity for Genomics Research. *BMC Bioinformatics.* 20, 741–747. doi:10.1186/s12859-019-3322-3

Lyndon, M. P., Cassidy, M. P., Celi, L. A., Hendrik, L., Kim, Y. J., Gomez, N., et al. (2018). Hacking Hackathons: Preparing the Next Generation for the Multidisciplinary World of Healthcare Technology. *Int. J. Med. Inform.* 112, 1–5. doi:10.1016/j.ijmedinf.2017.12.020

Mccullough, H. (2010). Using Personal Development Planning for Career Development With Research Scientists in Sub-saharan Africa PhD thesis, University of Liverpool, 4–61. doi:10.17037/PUBS.03894557

Mlotshwa, B. C., Mwesigwa, S., Mboowa, G., Williams, L., Retshabile, G., Kekitiinwa, A., et al. (2017). The Collaborative African Genomics Network Training Program: A Trainee Perspective on Training the Next Generation of African Scientists. *Genet. Med.* 19, 826–833. doi:10.1038/gim.2016.177

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

Mulder, N., Adebamowo, C. A., Adebamowo, S. N., Adebayo, O., Adeleye, O., Alibi, M., et al. (2017a). Genomic Research Data Generation, Analysis and Sharing - Challenges in the African Setting. *Data Sci. J.* 16, 49. doi:10.5334/dsj-2017-049

Mulder, N. J., Adebiyi, E., Adebiyi, M., Adeyemi, S., Ahmed, A., Ahmed, R., et al. (2017b). Development of Bioinformatics Infrastructure for Genomics Research. *Glob. Heart.* 12, 91–98. doi:10.1016/j.gheart.2017.01.005

Mulder, N. J., Adebiyi, E., Alami, R., Benkahla, A., Brandful, J., Doumbia, S., et al. (2016). H3ABioNet, a Sustainable Pan-African Bioinformatics Network for Human Heredity and Health in Africa. *Genome Res.* 26, 271–277. doi:10.1101/gr.196295.115

Nchinda, T. C. (2002). Research Capacity Strengthening in the South. *Soc. Sci. Med.* 54, 1699–1711. doi:10.1016/s0277-9536(01)00338-0

Nembaware, V., Mulder, N., and Mulder, N. (2019). The African Genomic Medicine Training Initiative (AGMT): Showcasing a Community and Framework Driven Genomic Medicine Training for Nurses in Africa. *Front. Genet.* 10, 1209. doi:10.3389/fgene.2019.01209

Onah, D. F. O., Sinclair, J., and Boyatt, R. (2014). "Dropout Rates of Massive Open Online Courses : Behavioural Patterns MOOC Dropout and Completion: Existing Evaluations," in Proceedings of the 6th International Conference on Education and New Learning Technologies EDULEARN14, 1–10. doi:10.13140/RG.2.1.2402.0009

Prozesky, H., and Mouton, J. (2019). A Gender Perspective on Career Challenges Experienced by African Scientists. *S. Afr. J. Sci.* 115, 40–44. doi:10.17159/sajs.2019/5515

Radouani, F., Zass, L., Hamdi, Y., Rocha, J. D., Sallam, R., Abdelhak, S., et al. (2020). A Review of Clinical Pharmacogenetics Studies in African Populations. *Per Med.* 17, 155–170. doi:10.2217/pme-2019-0110

Ras, V., Botha, G., Aron, S., Lennard, K., Allali, I., Claassen-Weitz, S., et al. (2021). Using a Multiple-Delivery-Mode Training Approach to Develop Local Capacity and Infrastructure for Advanced Bioinformatics in Africa. *Plos Comput. Biol.* 17, e1008640. doi:10.1371/JOURNAL.PCBI.1008640

Rotimi, C., Rotimi, C., Abayomi, A., Abimiku, A., Adabayeri, V. M., Adebamowo, C., et al. (2014). Research Capacity. Enabling the Genomic Revolution in Africa. *Science* 344, 1346–1348. doi:10.1126/science.1251546

Scott, J. (1992). Scotland. *Adoption & Fostering* 16, 59. doi:10.1177/030857599201600412

Sengupta, D., Choudhury, A., Fortes-Lima, C., Aron, S., Whitelaw, G., Bostoen, K., et al. (2021). Genetic Substructure and Complex Demographic History of South African Bantu Speakers. *Nat. Commun.* 12, 2080. doi:10.1038/s41467-021-22207-y

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, 331–413. doi:10.3389/fgene.2019.00331

Sonnert, G. (1999). Women in Science and Engineering: Advances, Challenges, and Solutions. *Ann. NY Acad. Sci.* 869, 34–57. doi:10.1111/j.1749-6632.1999.tb08353.x

Stads, G.-J., and Beintema, N. M. (2006). *Women Scientists in Sub-saharan African Agricultural R&D. Brief Prepared for the USAID Meeting on Women in Science: Meeting the ChallengeLessons for Agricultural Sciences in Africa*. Washington, D.C: The Agricultural Science and Technology Indicators (ASTI). June 21.

Tastan Bishop, Ö., Adebiyi, E. F., Alzohairy, A. M., Everett, D., Ghedira, K., Ghouila, A., et al. (2015). Bioinformatics Education-Pperspectives and Challenges Out of Africa. *Brief Bioinform.* 16, 355–364. doi:10.1093/bib/bbu022

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

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

BRIEF RESEARCH REPORT
published: 30 September 2021
doi: 10.3389/feduc.2021.725591

Transdisciplinary Approach for Bioinformatics Education in Southern Brazil

Marcio Dorn[1,2], *Rodrigo Ligabue-Braun*[3,4]* *and Hugo Verli*[2,5]

[1]*Structural Bioinformatics and Computational Biology Lab, Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil,* [2]*Graduate Program in Cellular and Molecular Biology, Center of Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil,* [3]*Department of Pharmacosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, Brazil,* [4]*Graduate Program in Biosciences, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, Brazil,* [5]*Structural Bioinformatics Group, Institute of Biosciences, Department of Molecular Biology and Biotechnology, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil*

The development and application of bioinformatics has been growing steadily, but its learning and training has been lagging. We have approached this problem through a bi-annual event, called EGB (Escola Gaúcha de Bioinformática), dedicated to undergraduate and graduate students (mainly from biology, biomedicine, chemistry, physics, and computer sciences), as well as professionals, to mingle and be presented to bioinformatics from sequence, structure, and computational standpoints simultaneously. The interactive environment provided by EGB allows for participants mingling, independently from their training background, fostering collaborative learning and experience exchange. Both lecturers and students are encouraged to collaborate and communicate, with no formal acknowledgement of "status differentiation".

Keywords: bioinformatics, Brazil, education, Latin America, outreach, South America

OPEN ACCESS

Edited by:
Chi-Cheng Chang,
National Taiwan Normal University, Taiwan

Reviewed by:
Jan Egger,
Graz University of Technology, Austria
Jose R. Valverde,
Centro Nacional de Biotecnología, Consejo Superior de Investigaciones Científicas (CSIC), Spain
Alexander Miguel Monzon,
University of Padua, Italy

***Correspondence:**
Rodrigo Ligabue-Braun
rodrigolb@ufcspa.edu.br

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *12 July 2021*
Accepted: *20 September 2021*
Published: *30 September 2021*

Citation:
Dorn M, Ligabue-Braun R and Verli H (2021) Transdisciplinary Approach for Bioinformatics Education in Southern Brazil. Front. Educ. 6:725591. doi: 10.3389/feduc.2021.725591

INTRODUCTION

Bioinformatics can be defined as simply as the "application of tools of computation and analysis to the capture and interpretation of biological data" (Bayat, 2002). Nonetheless, such definition is not as clear cut as it seems, with some authors being much more detailed in their description, e.g., "Bioinformatics is conceptualizing biology in terms of macromolecules (in the sense of physical chemistry) and then 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" (Luscombe et al., 2001). More recently, a definition of bioinformatics has been proposed as "an interdisciplinary field that is concerned with the development and application of algorithms that analyze biological data to investigate the structure and function of biological polymers and their relationships to living systems" (Tapprich et al., 2021). Contrasting with its array of definitions, there is no contending that bioinformatics has been steadily growing, both in its use and as a research field, at least since the early 2000s (Hodcroft et al., 2021; Wilson Sayres et al., 2018; Brusic, 2007; Perez-Iratxeta et al., 2007). This lack of agreement reflects the quick paced development of an evolving discipline whose target is not well defined yet. This lack of a settled focus is one of the major problems in teaching Bioinformatics: despite bioinformatics' half-century history (Gauthier et al., 2019), its learning and teaching seem to be trailing behind its observed growth and use (Hack and Kendall, 2005; Wilson Sayres et al., 2018). This trend is observed especially in undergraduate courses focused in biological and health sciences

(Madlung, 2018), but computational science courses also lack intersection with biological applications and implications (Atwood et al., 2019). Thus, while students of life sciences lack formal training in data management and programming, computer science students have little to no contact with biological data and its intrinsic variability. This poses the immediate problem: how to address the shortcomings in bioinformatics teaching and learning when dealing with such diverse set of actor backgrounds?Considering such problem, we herein present the results from a bioinformatics learning model based on: 1) a space for undergraduate and graduate students interactions, among each other and with professionals; 2) integration of participants originated from different backgrounds; 3) exposition of the participants, simultaneously, to bioinformatics from sequence, structure and computational standpoints, through 4) both theoretical lectures and hands-on courses, from introductory and advanced, methodological and applied standpoints. Encompassing three editions so far of the Escola Gaúcha de Bioinformática (EGB), or "Southernmost Brazilian School of Bioinformatics" in a free translation, such model has been applied with variations in Brazil since the 1980s through multiple fields, such as physics, chemistry, molecular simulation and others, and has been able to offer a generational impact on the formation of highly qualified researchers.

EGB: ESCOLA GAÚCHA DE BIOINFORMÁTICA

Background

Different strategies to consolidate bioinformatics as part of the main curriculum of different courses, especially in the biological sciences, have been proposed through the years. These include the definition of its core competencies (Welch et al., 2014), its defining elements (Tapprich et al., 2021), the inspection of successful teaching cases in the US and United Kingdom (Hack and Kendall, 2005), the need to go beyond traditional classroom short courses (Atwood et al., 2019), and the analysis of a dedicated learning module of bioinformatics for biology students (Madlung, 2018). The frequently mentioned transdisciplinary nature of the field, however, remains observational rather than practical in most of these propositions, with disciplines borrowing techniques from each other, but without forming a single, major area of knowledge. Thus, in order to show explicitly the combination of technical fields that make up this fast-paced field, we decided that, during the EGB, the transdisciplinarity should be presented as a living experience, bringing together participants from the multiple backgrounds that encompass the definition of bioinformatics.

The quick development of Bioinformatics is also often uneven spatially, with some regions developing faster than others, and often each group specializing in a very limited scope of the field of temporary interest for their research, neglecting other possibilities that might – if considered – improve their capacities. For example, as a research field, bioinformatics has been concentrated in the Southeast Region of Brazil, especially in the states of São Paulo and Minas Gerais (Bicudo, 2016). Even with government initiatives to spread research facilities all over the country (from 2007 onwards), such as BioMicro nucleation effort (CAPES, 2008) and a major Computational Biology financial laid call (CAPES, 2013), other regions remained underrepresented (Bicudo, 2016). This also results in numerous groups working in specialized subfields, with little knowledge of other areas, which further hampers development of usefully complete Bioinformatics curricula. By taking advantage of the budding bioinformatics groups in the South Region of Brazil, particularly those in the southernmost state of Rio Grande do Sul, and connecting them with professional from all over Brazil and Latin America, EGB presents itself as a networking environment for the supervised development of abilities required to navigate the field. It is noteworthy that AB3C (Brazilian Association for Bioinformatics and Computational Biology) *via* its annual X-meeting event, provides training and networking opportunities. Likewise, the Brazilian ISCB Regional Student Group has been connecting bioinformatics students regionally. Nonetheless, these initiatives target subjects already working with bioinformatics, not specifically beginner, non-experienced users. The EGB works both as a scientific event (such as a symposium) and as a teaching-learning environment (like a "school"). This model has been successfully applied to more specific study areas in Brazil, for instance in the Brazilian School of Electronic Structure (since 1987) (dos Santos et al., 2003; EBEE, 2018) and in the School for Molecular Modeling in Biological Systems (since 2002) (EMMSB, 2021). Participation can include research paper submission and poster presentation, but these are not mandatory activities. The integration of all participants in an experience-sharing environment with provocative ideas is the main goal of EGB. A regular day at the "school" would include three to four lectures in the morning, followed by a lunch break, and flash courses in the afternoon, with both morning and afternoon periods having coffee break intermissions. These intermissions are particularly relevant for allowing participants to mingle in a friendly, casual environment. The school duration is proposed to be of 5 days (Monday to Friday), taking place during the Brazilian college winter break (in July). From the practical standpoint, the school requires a large auditorium for lectures, open halls for intermission activities and small individual computer labs for flash courses. Participants have the option to validate their participation as a credit in partner undergraduate and graduate programs. Thus far, three editions of EGB were organized (in 2015, 2017, and 2019), taking place in the Informatics Institute of the Federal University of Rio Grande do Sul, in Porto Alegre, Brazil. Supporting lab spaces were also provided by the Center of Biotechnology and the Institute of Biosciences at the same university. Financial management and enrollment assistance was provided the Brazilian Genetics Society (SBG).

Conceptually, the transdisciplinary characteristic of bioinformatics is approached by a simultaneous exposition of the students to lectures from introductory to advanced graduate levels, and from methodological to applied standpoints. A particular care is taken to include, in the same morning, lecturers from multiple fields in bioinformatics, as those

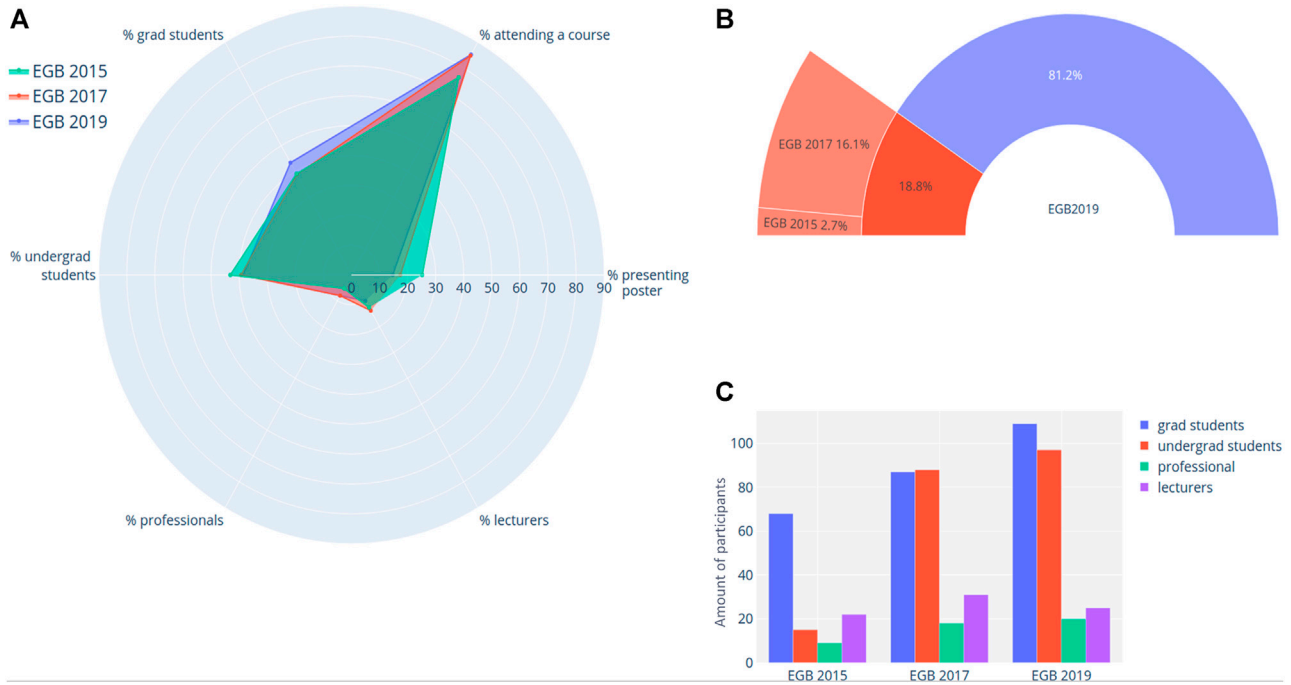


FIGURE 1 | EGB participant demographics. **(A)** Radar plot highlighting shifts in participant profiles; **(B)** ratio of participants returning from previous EGB editions; **(C)** Total participant amount by year, indicating growth in participation at each edition of the event.

working with sequence- or structure-based approaches, from biological to computer science backgrounds. Consequently, for example, some undergraduate biology student will be familiarized with coding while a computer science graduate student will be presented to protein biochemistry and molecular biology. During the intermissions, a more experienced student could help a newcomer to a particular area of bioinformatics to better understand previously discussed concepts, while the lecturer could both deepen some methodological details or explain in an even more basic level some aspect for participants. Finally, in the afternoon, the aspects from bioinformatics discussed in the morning lectures will be sedimented in hands-on courses on computer labs, also from the basic to more advanced levels. While it may be particularly challenging for some students to have a first contact to advanced coding, protein structure modeling or genome annotation using this approach, it is simultaneously a highly contextualized introduction to such field supported by the student familiarity with some of the other fields discussed during the event. In addition, these interactions have the benefit of encouraging and fostering inter-group cooperation.

Participant Demographics

The target audience for EGB was conceived to be wide, from undergraduate to graduate students and professionals. Hence, since its first edition in 2015, similar amounts of undergraduate and graduate students have participated in the event (**Figure 1**), simultaneously to a steady increase in the interest for the school, at an approximate 25% rate per edition, ranging from 174 participants in 2015 to 251 participants in 2019. We have observed a slight increase in graduate students as event editions progressed (**Figures 1A,C**). The same tendency was observed for the interest in focused flash courses (**Figure 1A**). More data is required to interpret these trends, but we hypothesize that graduate courses (especially those in the biological areas) are being insufficient in providing their students with the bioinformatics training they need or aspire. Since it is not directly enforced, the varied audience emerges as reflection of the public interest in the subject. Admissions are processed in first-come, first-served basis, with the only selection criteria in place being the preference for new participants (in lieu of returning ones). So far, no exclusion has occurred in the admissions process.

Around 90% of the participants were Brazilians, with three quarters of them being from the Rio Grande do Sul State. These participants listed traditional University cities as their origin, with a concentration in the capital city of Porto Alegre and its surroundings. Non-Brazilian participants were Latin-American, coming from Argentina, Bolivia, Chile, Colombia, Mexico, Peru, and Uruguay (**Figure 2**). Direct involvement of international participants remains dependent on additional funding for travel expenses, something that is yet to be secured. The origin institutions of the participants (universities, research centers, etc) were also listed, forming a total of 42 different affiliations. At the end of each event, all participants were invited to provide anonymous evaluations on multiple aspects of the school. This qualitative data has been used to adjust and improve subsequent editions of EGB. These observations are reported in **Section 3**. Participant demographic questionnaire and obtained results are

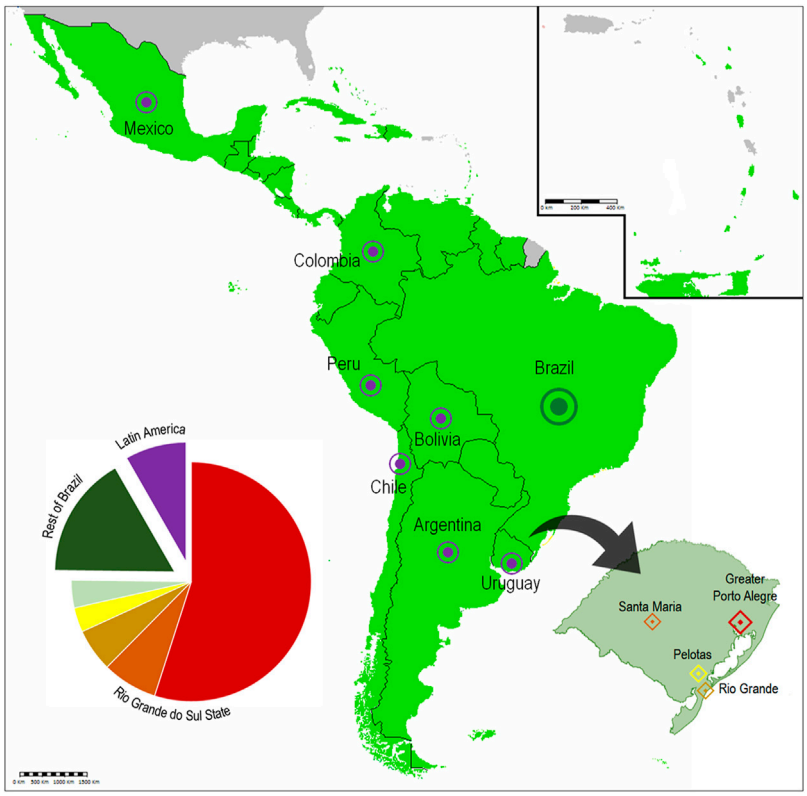


FIGURE 2 | EGB participant coverage. The pie chart is color-coded as depicted in the map and reflects relative proportion of participants from each location. The State of Rio Grande do Sul is highlighted, pinpointing University cities that most contributed participants to the event. (Incorporates "Map for Community of Latin American and Caribbean States", licensed under the Creative Commons Attribution-Share Alike 4.0 International, by Roblespepe).

presented in the Supplementary Material. The full questionnaire was not applied in the first edition of the event.

Lectures and Flash Courses

Following the general concept of "mixing and matching", the lectures were planned to expand the participants' perspectives on the field. Accordingly, during each morning talks from specialists were given to the students covering multiple bioinformatics aspects, under three main standpoints: sequence or structure-based methods, biological applications and questions, and computational approaches to solve them. The main goal here is that the students get familiarized with multiple perspectives of bioinformatics, which are reinforced during the event. Thus, a lecture cycle could begin with observations on ways bioinformatics assisted zoologists to solve how tigers got their stripes (based, for example, on genomic data), followed by drug discovery against a specific pharmacological target (supported by docking, for instance), and ending with examples of computer clusters being used for crop enhancement via synthetic biology. Themes covered by lectures included (to cite a few) agricultural enhancements via genetic manipulation, big data in health and ecology, chemoinformatics, evolution, forensic applications of bioinformatics, massive parallel processing for bioapplications, molecular docking, multiscale molecular modeling and simulation, protein engineering, and vaccinology. Themes were changed at each event, and while one theme was given in a basic level in one edition, in other edition it could be offered in an intermediate level. Such approach accommodates students from multiple backgrounds and stimulate the return of participants for a new addition of the event with, accordingly, a continuous learning on the field. A similar approach was used for the flash courses, that include mandatorily some level of hands-on activity, from physical modeling of protein structures to viral genome assembly. The main themes covered by flash courses included introductory biochemistry, genome assembly and annotation, molecular dynamics, Python programming, structural biology, systems biology, and transcriptome data processing. All flash courses have allotted time at their initial session for familiarization with computer use and (when needed) with nongraphic interfaces. The ratio of 1 assistant for every five flash course participants is preconized. One day of the event was reserved to lectures from researchers in industry, emphasizing commercial applications and opportunities for the students. Also, these entrepreneurs reserved part of the day to mix with the participants. Course instructors and lecturers comprised Brazilian and Latin American experts from various fields pertaining to bioinformatics. These included professors, researchers, product managers, and industry representatives. Each lecture had an expected length of 50 min, followed by around 10 minutes of questioning, while flash courses lasted

around 15 hours in total (3 h s/day), with participants opting for one course to follow along the 5 days. The lectures are of general attendance, while participants are invited to choose one flash course to follow through. Lecturers were selected based on suggestions from the Organizing and Scientific Committees and approved upon members' voting. Training assistants for flash courses were selected among volunteers with a track record of work with the computational tools. Lectures were given mostly in Portuguese and Spanish.

OBSERVATIONS AND PERSPECTIVES

The qualitative evaluation by the participants of the event was highly positive, with courses and lectures being considered "excellent" or "good" by more than 90% of the participants. Likewise, the scientific program and the venue were deemed excellent, and about three quarters of the participants expressed interest in participating in future editions of EGB. This interest is confirmed by the amount of returning participants at each new edition (**Figure 1B**). From the open-ended questionnaire, some aspects were brought up and have helped shape a more inclusive program in bioinformatics. Such aspects encompass the requirement for more physically accessible auditoria for mobility challenged individuals, the option for sign language interpreters and commitment to gender-balanced lineup of lecturers and instructors. Considering the need for training in bioinformatics and the observed success of the EGB strategy, impacting directly around 200 people per edition, plans are underway for expanding its geographical coverage. Initial arrangements are in progress for providing simultaneous instances of the event in different regions/countries, with live webcasting of lectures and local offerings of flash courses. Nonetheless, with the lengthening of the COVID-19 pandemic and its ongoing effects (Zoumpourlis et al., 2020), these plans were halted. Considering the social mingling that has been considered the highlight of EGB, by providing ways of connecting people with different backgrounds with a common interest in bioinformatics, and the additional need for supervised activities in the flash courses, the total conversion of the event to distance activities was deemed inadequate by the organizing committee. Despite still lacking proper assessment, information volunteered by former participants indicate that the interaction opportunities provided by EGB have fostered collaborations among researchers and students from different institutions, allowing for multiple groups joining forces for topical research projects. Such collaborations were also stimulated with the bridging of academia and different economic sectors, with students familiarizing themselves with the entrepreneurial environment. Undergrad students were also able to socialize with graduate students, sparkling interest in pursuing further academic career pathways. Still, the enrolled students came majorly from life sciences backgrounds, with computer science students representing a minor portion of the attending public. Even if the social component of the event may not apply to all participants, the knowledge gathered can supplement the current curricula by providing up-do-date lectures and high-level discussions. We hope these strategies may aid other endeavors for consolidating bioinformatics as a perennial knowledge in students and professionals alike.

In conclusion, the opportunity for interaction between people from biological and computational science backgrounds in a casual environment emerges as the main advantage of the model for bioinformatics training, teaching, and learning presented in the EGB school events.

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

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

Previous editions of EGB were funded by grants from FAPERGS (570-2551/15-6 and 17/2551-0000373-0), CAPES (23038.001974/2015-73, 88881.123406/2016-01, and 88881.290880/2018-01) and Asociación de Universidades Grupo Montevideo (AUGM). Support was also provided by the Universidade Federal do Rio Grande do Sul (UFRGS) via its Institute of Informatics, Institute of Biosciences, and Center of Biotechnology.

ACKNOWLEDGMENTS

The authors are extremely grateful to all students who worked in the organization of previous editions of EGB, and to all lecturers and course tutors for accepting the challenge. The authors acknowledge Diego Bonatto for his involvement in the first edition of the event, and the Brazilian Genetics Society (SBG) for providing instrumental support for the second and third editions of the event.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.725591/full#supplementary-material

REFERENCES

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 Bioinform* 20, 398–404. doi:10.1093/bib/bbx100

Bayat, A. (2002). Science, Medicine, and the Future: Bioinformatics. *BMJ* 324, 1018–1022. doi:10.1136/bmj.324.7344.1018

Bicudo, E. (2016). Genomics Politics through Space and Time: The Case of Bioinformatics in Brazil. *Public health genomics* 19, 81–92. doi:10.1159/000443472

Brusic, V. (2007). The Growth of Bioinformatics. *Brief Bioinform* 8, 69–70. doi:10.1093/bib/bbm008

CAPES (2008). Doutorados em bioinformática e microeletrônica (BioMicro), Available at: https://www.gov.br/capes/pt-br/acesso-a-informacao/acoes-e-programas/bolsas/programas-estrategicos/outras-informacoes/programas-encerrados-estrategicos/biomicro (Accessed June 7, 2021).

CAPES (2013). Edital Biologia Computacional–CAPES Nº051/2013, Available at: https://www.gov.br/capes/pt-br/centrais-de-conteudo/edital-051-2013-biologiacomputacional-pdf (Accessed June 7, 2021).

Dos Santos, H. F., Coura, P. Z., Dantas, S. O., and Barone, P. M. V. B. (2003). *Escola Brasileira de Estrutura Eletrônica*. Juiz de Fora: Editora Livraria da Física, 324.

EBEE (2018). XVI Escola Brasileira de Estrutura Eletrônica, Available at: https://sites.google.com/site/xviebee/home (Accessed June 7, 2021).

EMMSB (2021). 10a Escola de Modelagem Molecular em Sistemas Biológicos, Available at: http://www.emmsb.lncc.br/(Accessed June 7, 2021).

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A Brief History of Bioinformatics. *Brief Bioinform1996* 20, 1981–1996. doi:10.1093/bib/bby063

Hack, C., and Kendall, G. (2005). Bioinformatics: Current Practice and Future Challenges for Life Science Education. *Biochem. Mol. Biol. Educ.* 33, 82–85. doi:10.1002/bmb.2005.494033022424

Hodcroft, E. B., De Maio, N., Lanfear, R., MacCannell, D. R., Minh, B. Q., Schmidt, H. A., et al. (2021). Want to Track Pandemic Variants Faster? Fix the Bioinformatics Bottleneck. *Nature* 591, 30–33. doi:10.1038/d41586-021-00525-x

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

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

Perez-Iratxeta, C., Andrade-Navarro, M. A., and Wren, J. D. (2007). Evolving Research Trends in Bioinformatics. *Brief Bioinform* 8, 88–95. doi:10.1093/bib/bbl035

Tapprich, W. E., Reichart, L., Simon, D. M., Duncan, G., McClung, W., Grandgenett, N., et al. (2021). An Instructional Definition and Assessment Rubric for Bioinformatics Instruction. *Biochem. Mol. Biol. Educ.* 49, 38–45. doi:10.1002/bmb.21361

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

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

Zoumpourlis, V., Pearson, W., Ryder, E. F., Tosado-Acevedo, R., Tapprich, W., Tobin, T. C., et al. (2020). The COVID-19 Pandemic as a Scientific and Social challenge in the 21st century. *Mol. Med. Rep.* 22, 3035–3048. doi:10.3892/mmr.2020.11393

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Bioinformatics

PERSPECTIVE
published: 11 October 2021
doi: 10.3389/fbinf.2021.727066

U-Hack Med Gap Year—A Virtual Undergraduate Internship Program in Computer-Assisted Healthcare and Biomedical Research

*Stephan Daetwyler†, Hanieh Mazloom-Farsibaf†, Gaudenz Danuser and Rebekah Craig**

Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, Dallas, TX, United States

The COVID-19 healthcare crisis dramatically changed educational opportunities for undergraduate students. To overcome the lack of exposure to lab research and provide an alternative to cancelled classes and online lectures, the Lyda Hill Department of Bioinformatics at UT Southwestern Medical Center established an innovative, fully remote and paid "U-Hack Med Gap Year" internship program. At the core of the internship program were dedicated biomedical research projects spanning nine months in fields as diverse as computational microscopy, bioimage analysis, genome sequence analysis and establishment of a surgical skill analysis platform. To complement the project work, a biweekly Gap Year lab meeting was devised with opportunities to develop important skills in presenting, data sharing and analysis of new research. Despite a challenging year, all selected students completed the full internship period and over 30% will continue their project remotely after the end of the program.

Keywords: bioinformatics, science education, undergraduate research, virtual experience, peer-mentoring

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Rachel Sparks, King's College London, United Kingdom
Ulrik Günther, Helmholtz Association of German Research Centers (HZ), Germany

****Correspondence:***
Rebekah Craig
Rebekah.Craig@UTSouthwestern.edu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Computational BioImaging, a section of the journal Frontiers in Bioinformatics

Received: *18 June 2021*
Accepted: *23 September 2021*
Published: *11 October 2021*

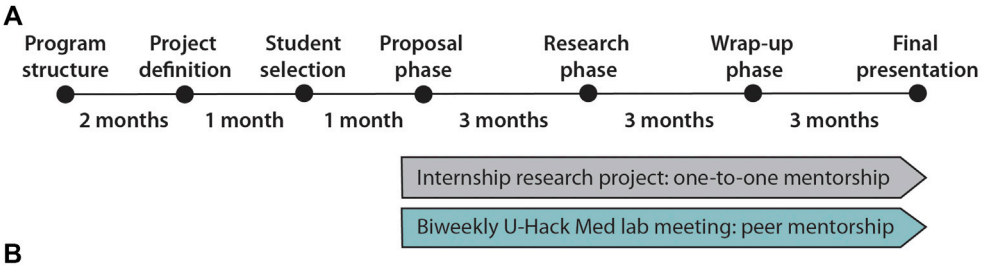
Citation:
Daetwyler S, Mazloom-Farsibaf H, Danuser G and Craig R (2021) U-Hack Med Gap Year—A Virtual Undergraduate Internship Program in Computer-Assisted Healthcare and Biomedical Research. Front. Bioinform. 1:727066. doi: 10.3389/fbinf.2021.727066

INTRODUCTION

The COVID-19 health crisis has heavily impacted traditional learning environments (Slavin and Nathan, 2020; UNESCO 2020). Many universities have cancelled or shifted their courses to a virtual learning space, depriving students of essential elements of the college experience (Abbasi et al., 2020; Kogan et al., 2020; Sidpra et al., 2020). This includes laboratory rotations, hands-on-experiments, face-to-face social interactions and important opportunities to learn together and discuss assignments (Daniel 2020; Kogan et al., 2020). Consequently, in the 2020–2021 academic year many students decided to take a "gap year" and were looking for new opportunities to be involved in active research.

The Lyda Hill Department of Bioinformatics at UT Southwestern Medical Center (UT Southwestern) was also affected in educational programming for undergraduates. Traditionally, every year UT Southwestern offers an international hackathon for students interested in application of computational approaches to solve "real-world" biomedical problems (www.u-hackmed.org). These events take place in-person, across 3 days during which teams of students, researchers and physicians intensely endeavor to bring new computational approaches to clinical and biomedical research and utilize the analytic power of the BioHPC, UT Southwestern's high-performance computing resource. The hackathon format attracts students from community colleges to Ivy League schools and is particularly successful, as students work on clearly pre-defined, stimulating research questions with faculty across UT

A



B

Selected learning goals	Methods for Achieving Learning Goals
Technical skills in computational biomedicine, computational bioimaging, and computer science	Internship project with mentor
Learn about cutting-edge research	Presentation of mentors and peers, journal club presentation, being a full lab member of mentor's lab
Present research data in an understandable manner	3 presentations (proposal, mid-term and final presentation)
Interpret scientific data and summarize it	Present a scientific paper in a journal club
Exchange ideas and work in a team	Collaborative efforts to present a paper, mentor-mentee team work
Learn about different career paths	Career development sessions and panel discussions
Networking with peers beyond science	Online happy hours and board-game nights

FIGURE 1 | Project milestones and learning goals. **(A)** Schematic of important milestones of the U-Hack Gap Year internship program **(B)** Summary of the major learning goals for the undergraduate students.

Southwestern. Unfortunately, during the COVID-19 health crisis, travel and laboratory restrictions led to the cancellation of this successful format.

To overcome this lack of exposure for undergraduate students to pressing questions in healthcare and computer-assisted biomedical research, a new format, the U-Hack Med Gap Year program was conceived. To comply with local and state regulations, the internship was defined as a completely remote, virtual experience that provided new opportunities for industrious students searching for meaningful research projects during this time of crisis. At the core of our U-Hack Med Gap Year internships were clearly defined projects that were attractive to undergraduate students, useful for mentors to advance their research program, and discrete in scope so that substantial project completion could be achieved within 6–9 months. To build a sense of community among the intern cohort and mentors, the experience was enhanced with a biweekly virtual U-Hack Med lab meeting. In the next sections, we describe administrative best practices for setting up a virtual internship program, the project and student selection process, the internship program structure with an initial proposal phase, a core research phase, and a wrap-up phase with a student symposium, before describing in detail the surveyed outcome of the program (**Figure 1**).

Administrative Structure for Remote Internship Program

Offering innovative opportunities such as a fully remote, paid internship may require overcoming administrative barriers as a first step. Many institutions of higher education have not traditionally allowed fully remote work due to university telecommuting policies, tax-withholding issues, and other state regulations. The COVID-19 crisis presented a window of opportunity for advocacy toward greater openness and administrative restructuring with regard to remote work. Prior to launching the U-Hack Med Gap Year applications, facilitators coordinated with several administrative offices including graduate school, payroll, accounting, international affairs and human resources temporary employment recruitment. From these meetings, emerged the following recommendations for how research departments in the United States can administratively structure this type of program.

State-to-state registration: To employ remote workers in another state, employers must register with the department of state in that jurisdiction and these requirements vary by state. It is important to determine what states the institution is registered, or is willing to register in, prior to selection of student applicants.

Job classification: A custom job code, reserved for undergraduate students, was used for this opportunity. The job code allowed the intern to remain in "student" status with compensation in the form of a monthly stipend versus being considered a "temporary staff" employee paid an hourly wage. This distinction is important as stipends support educational and training opportunities and thus do not represent compensation for work performed. Depending on the job code used, stipends therefore may not be subject to the federal and state payroll taxes that are assessed for salaries and wages. This is an important factor to be considered, together

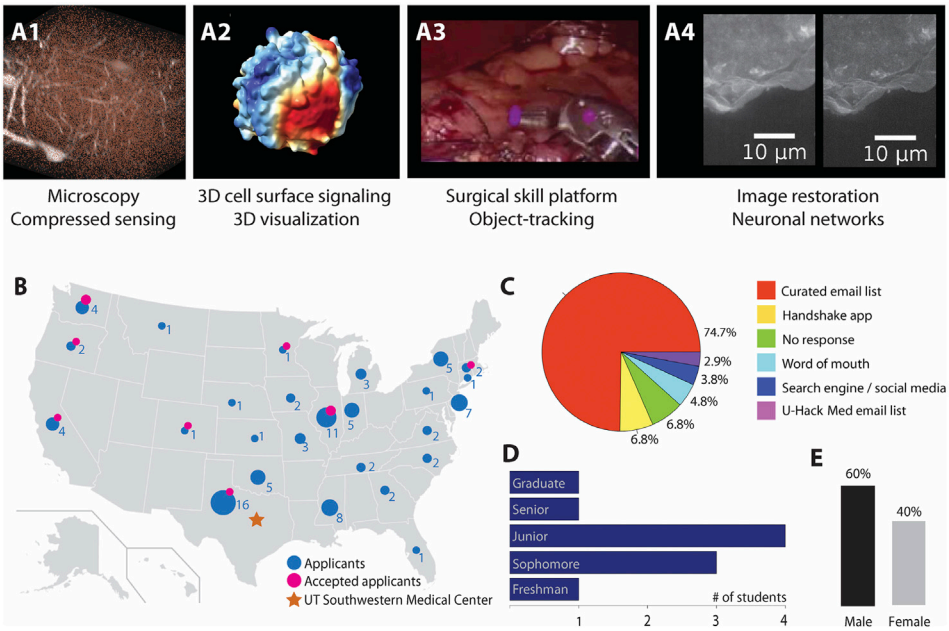


FIGURE 2 | Internship projects and student characteristics. **(A1-A4)** The internship projects spanned topics as diverse as microscopy, 3D visualization, surgical skill platform analysis and image restoration. **(A1)** A student applied compressed sensing algorithms (Rani, Dhok, and Deshmukh 2018) to reconstruct an original image of mouse brain neurons (white) from sparsely sampled points (brown) in a 3D volume obtained by simulating a random-access microscope. **(A2)** A student analyzed molecular signals on a 3D cell surface of a melanoma cell (Driscoll et al., 2019) and visualized the resulting data. **(A3)** A student established a full pipeline to perform automated analysis to score the performance of different surgeons on a novel surgical skill platform (Battaglia et al., 2021). **(A4)** A student explored artificial neuronal networks such as CARE (Weigert et al., 2018), to obtain higher resolution images (right) from raw data (left). **(B)** Applications (blue, circle size scaled for number of applications) for the U-Hack Gap Year were received from all over the US. Consequently, participating students (pink, circle size scaled for number of interns) were working remotely from all over the US, while the mentors were working at UT Southwestern in Dallas (orange). **(C)** Source of information that made participant aware of the U-Hack Med Internship program included advertisement through our curated college email list (74.7%), Handshake app (6.8%), no response (6.8%), word of mouth/ friend (4.8%), search engine and social media (3.8%), U-Hack Med former participant mailing list (2.9%). **(D)** The knowledge background of the participating students was highly diverse ranging from freshman to graduate student knowledge. **(E)** The internship participants were six male and 4 female students.

with the states an employer is registered in, to determine the geographic region from which remote student interns may be recruited.

Student stipend: To support living expenses during the course of the internship, every accepted student received a stipend of $1,600 dollars per month. To incentivize research labs to participate in the program, the Department subsidized 50% of the stipend cost for each student, while the sponsoring lab was responsible for providing the remaining 50% of the funding.

Remote employee onboarding: It is important that a system is established for onboarding of a fully remote employee. UT Southwestern requires in-person "check-in" for all employees during their first 3 days of employment. During this appointment, personal documents are verified, and an I-9 employment eligibility verification form is completed by the human resources representative. UT Southwestern partnered with a third-party vendor at a nominal fee per student. Each student intern was then able to complete the check-in with this authorized representative at a location within driving distance of their place of residence. Another aspect of remote employee onboarding made possible in response to the COVID-19 crisis was the institution's mandatory new employee orientation. This typically is required as an in-person group meeting on the first day of employment. However, the human resources department quickly adapted to the need for social distancing by moving all new employee orientation sessions to a virtual platform.

International student interns: The international affairs office of each institution has some measure of discretion in interpretation of federal regulations. In the UT Southwestern context, completion of the I-9 employment eligibility verification by a third-party vendor for international students was not allowed. Since travel and on-campus activity was also severely limited by the COVID-19 crisis, it was determined that students in a temporary visa status could not be admitted to the program. This was made clear in the application announcement.

Project Selection and Scope

To capture the breadth of research within the Lyda Hill Department of Bioinformatics, all faculty were invited to draft project proposals and present their ideas in a planning meeting. Ideal projects were limited in scope so that they could be completed in 6–9 months with a student time commitment of approximately 30 h per week. Since the interns would be fully remote, projects also had to be feasible without any specialized equipment. In total, nine faculty members proposed eleven projects and half of the faculty integrated trainees and senior researchers from their labs as project leaders and mentors. All eleven projects complied with the requirements and were selected

TABLE 1 | Overview about selected projects. Graduate Student (GS), Postdoctoral Fellow (PF), Instructor (I), Principal Investigator (PI). * Biweekly lab meeting coordinator.

Mentor(s)	Project title
J. Zhou (PI)	Making Deep Learning Models Interactive and Self-Interpreting
X. Wang (PF), G. Danuser (PI)	Building a Deep Learning-Based Shape Selector
H. Mazloom-Farsibaf* (PF), M. Driscoll (I), G. Danuser (PI)	Computer Graphics of Cancer Cells
B.J. Chang (I), R. Fiolka (PI)	Deep Learning Deconvolution
S. Daetwyler (PF), R. Fiolka (PI)	Really Smart Microscopes for Cancer Cell Biology
J. Lee (PI)	Deciphering Visual Evaluation on Reconstructive Surgery Outcomes Using an Eye-tracking Platform and Machine Learning Techniques
A. Jamieson (PI)	Developing an Automated Surgical Skill Analysis Platform
S. Nguyen (PF), A. Treacher (GS), A. Montillo (PI)	Image Segmentation, Deep Learning Architectures & Predictive Modeling to Advance Neuroscience
D. Beshnova (PF), Bo Li (PI)	Interpretable Deep Learning for Cancer-Associated T-Cell Receptors
D. Kim (PI)	3D Visualization of An Entire Cell
S. Rajaram (PI)	Deep Learning for Histopathology and Broad Utility Biomedical Tools

for publication on the U-Hack Med website (www.u-hackmed.org/2020gapyear). The projects were as diverse as visualization of 3D data, image processing with deconvolution, genome sequencing with deep-learning networks, and developing an automated surgical skill analysis platform (**Figure 2A**; **Table 1**).

Student Recruitment

To recruit students, we leveraged a curated list of administrator contacts in STEM departments at colleges and universities across the US. This list was first established for the U-Hack Med Hackathon, with the help of high school interns who personally contacted hundreds of university administrators to introduce the U-Hack Med program and invite their partnership in promoting the event. For the U-Hack Med Gap Year promotion, university career services or advising office contacts were added. While initially a labor-intensive task, this list is now updated annually and has proven effective in widely promoting the U-Hack Med events. Through this list, a promotion email for the 2020 Gap Year program was sent to 808 administrators at 132 research-intensive undergraduate universities. Besides this list, the UT Southwestern and U-Hack Med websites, social media networks, email to former U-Hack Med hackathon participants, and the student recruitment platform *Handshake* were used to advertise. From the 133 student applicants, we learned that emails directly to university STEM departments and career offices were the most effective form of promotion, with nearly 75% of applicants indicating this communication as the way they learned about the program (**Figures 2B,C**).

To apply, undergraduate students were required to submit a resume, transcript and personal statement. In the personal statement, applicants were asked to describe their educational and career goals, circumstances that led to interest in a gap year internship, and their motivation for a specific U-Hack Med Gap Year project. In the resume, students were further asked to include links to code samples and completed or in-process computational or research projects.

Each mentor was asked to review applications and rank their preferred interns with consideration for the intern's project preferences. In a group meeting of all mentors, possible matches were tallied and from an initial 133 applicants, 26 students were selected for interviews. Group interviews were arranged with all potential intern-mentor pairs to determine the best fit of intern skills and interest with project requirements. They also served to further narrow the pool of applicants. Final assignments were made in a subsequent mentor meeting. For one of the projects, there was only one suitable student applicant, but the selection process helped to determine that another project was a better fit for that student's interests and enthusiasm; thus, the faculty member chose to withdraw the project from consideration. A final ten matched students were then invited to participate.

The selection process led to a student population with diverse knowledge backgrounds (**Figure 2D**). Most students were junior classification (40%), followed by sophomore (30%), while freshmen, senior and recent graduates each represented 10% of the intern cohort. Moreover, the interns were located all over the US and had a gender distribution of 40% female and 60% male (**Figure 2E**). Of note, this percentage of female students is significantly higher than the percentage of women in computer science [19% on Bachelor's level, (NSF 2017)].

Key Features of the U-Hack Med Gap Year Program

At the heart of the U-Hack Med Gap Year program was a dedicated research project. To achieve project goals, students received one-to-one mentoring and participated as full lab members in the mentor's lab. This included virtually attending all lab meetings. In several cases, graduate students and postdoctoral fellows were paired with interns (**Table 1**) to achieve defined research goals as a mentee-mentor team, providing an excellent professional growth opportunity for both students and mentors. In these cases, an experienced faculty member further supervised the mentee-mentor team.

As many projects were computationally expensive, UT Southwestern provided all students access to its Biomedical High-Performance Computing cluster, the BioHPC, via remote VPN login. The BioHPC provides state-of-the art service with access to over 30 PB of data storage capacity and 28,000 CPU cores (https://portal.biohpc.swmed.edu/content/). Currently, the simultaneous allocation limit for an individual user (and thus the

interns) is set at 4 GPU nodes with high-end Nvidia GPU modules (Nvidia Tesla K20/K40, P4, P40, P100, V100, A100) and 64 "light-weight" CPUs nodes with each 32 GB memory, or alternatively 16 large memory nodes with up to 384 GB memory. To further facilitate computational analysis, the BioHPC is equipped with remote virtual desktops to access the cluster and licensed software needed for the projects. As part of the onboarding process, all students received training to access and use the compute cluster for their analysis.

Besides the research project, the biweekly U-Hack Med lab meeting was a crucial element of the U-Hack Med Gap Year internship program. It provided a platform to achieve important learning goals (**Figure 1B**). As students came with diverse academic backgrounds (**Figure 2D**), the lab meeting included high-level training in presentation, interpreting research data, structuring data and data sharing. Specific technical training was the responsibility of the mentors. A postdoctoral fellow was chosen to design and lead the biweekly program and student symposium, with the input and mentoring of two faculty members, providing additional trainee professional enrichment and growth.

The structure of the U-Hack Med lab meetings supported their internship experience. First, to give students time to become fully acquainted with their project, all mentors presented their work to highlight ongoing cutting-edge research at UT Southwestern. Mentor talks were complemented with toolbox presentations on how to give a presentation. This phase was concluded with a first "proposal" presentation by the students. This proposal presentation was intended for students to concisely articulate their ideas and understanding of the project and receive feedback. Afterwards, students worked on their projects and showcased their progress in a mid-term presentation. Alongside these, the students teamed up in pairs to choose a new research paper and present it in a journal club format. The journal club presentation was a central element in the training effort to enhance the students' skills in extracting the main messages of a paper. During the wrap-up phase, two career sessions were included to increase students' awareness of job opportunities for biomedical scientists. One session was dedicated to industry careers with an invited speaker. The second session was a panel discussion devoted to academic careers with two faculty and two trainees from the Lyda Hill Department of Bioinformatics at UT Southwestern. Finally, the U-Hack Med lab meetings culminated in a symposium with a presentation by each student and attendance from the broader Department and institutional education leaders.

The internship experience was further augmented by optional community building events. These included two "Virtual Happy Hour" sessions, where internship students and mentors met in a relaxed and friendly atmosphere to celebrate holidays such as the New Year. In addition, two virtual board game nights were organized with group games such as Pictionary and Code Names.

Outcome

Despite the numerous challenges during an unprecedented international crisis, all ten selected students completed the program. Students were successful in developing computer vision pipelines for performing automated analysis on raw video data from actual surgical procedures, deep learning-based object tracking, establishing of containers for high-performance cluster computation, image reconstructions from sparse signals, deconvolution of imaging data, restoration of low-resolution microscopy images to high-resolution images at a tenfold speed increase compared to conventional methods, building of a novel application of artificial intelligence to clinical image segmentation, and molecular signal analysis on 3D cell surfaces. Three of the ten students have already decided to continue their research project. At least eight of the ten projects have publications in planning or close to submission.

To ascertain areas in which the U-Hack Med Gap Year program was successful and could improve, we developed anonymous online surveys completed by all ten students and eight mentors (Supplementary Notes 1 and 2). As illustrated in **Supplementary Figures S1, S5**, students found the application process user-friendly (80 ± 13%). Mentors (79 ± 5%) and students (90 ± 8%) agreed that the matching process was also satisfactory. Students felt somewhat well prepared by their academic program for the internship (76 ± 10%), while mentors expressed a lower confidence in the academic preparation of interns (63 ± 18%). Students and mentors shared in open-ended comments that a better and longer training was needed at the beginning of the internship for using high-performance cluster computing resources.

During the biweekly lab meetings (**Supplementary Figures S2, S5**), students gained the most benefit from preparing and presenting their research in the final student symposium (83 ± 17%), a career session on prospects in industry (78 ± 11%), learning about the other projects (77 ± 13%), and the panel discussion on careers in academia (70 ± 13%). Moreover, mentors appreciated the biweekly lab meetings and felt it brought good benefit to the students (84 ± 14%). Open-ended comments from both students and mentors indicated opportunity for improvement in the amount of constructive feedback offered by other interns and mentors during biweekly meetings, and in the use of these meetings to develop problem-solving skills.

Overall, students perceived communication with the mentors via virtual platforms (**Supplementary Figure S3**) as "just right" and felt communication was frequent and easy enough (92 ± 8%). Mentors agreed that communication was frequent and easy in a completely virtual environment (81 ± 13%). Approximately half of the mentors (46 ± 9%) and students (40 ± 11%) felt that the project period was long enough for the students to make satisfactory progress. However, mentors observed strong gains (84 ± 10%) in students' skills and knowledge over the course of the program (**Supplementary Figure S5**), and all mentors reported that publications will result from the project, with the intern as co-author. In overall experience (**Supplementary Figure S4**), students regarded the U-Hack Med Gap Year as very satisfactory and beneficial to their future academic studies (93 ± 9%), and all students would recommend this type of virtual internship and the Department of Bioinformatics as a place for such an internship to others.

All ten students reported that the internship had a definite positive impact on their career plans, and half of the students

indicated that applying theoretical concepts to a "real-world" problem was the most satisfying aspect of their internship. For three students the experience solidified their desire to pursue a graduate degree, five indicated a desire to pursue computational biology and/or bioinformatics as a focus of study, and one felt reaffirmed in their decision to pursue a career in industry. Students found that the experience helped to buffer the negative impacts of the pandemic on their career motivation, with one commenting: "I was really struggling to be inspired before I started this internship because of the pandemic and feeling like I was not doing anything, and this really helped to revitalize my love of academia and research." Another commented, "This internship was incredible both in a professional and personal context. It gave me confidence that I was employable and also that I am good at what I do."

In open-ended comments, mentors reported that the internship was a very enriching and successful experience, and that they improved their skills in student research mentoring. In comparison with previous undergraduate student mentoring experiences, one mentor felt the length of the internship and the stipend students received resulted in both the mentor and the mentee being "more disciplined and responsive," and asserted that "as a result, mentors and mentees both gain a more solid experience."

DISCUSSION

The program described herein has been a unique opportunity for high-performing students to gain insights into the practices of cutting-edge research in computational biology and medical informatics. The students and mentors expressed their enthusiasm in outcomes surveys and all participating students would recommend this virtual internship and our Department to others. In the future, when COVID-19 restrictions are lifted to allow the in-person U-Hack Med hackathon format again, both events could be organized in tandem, potentially offering hackathon winners a paid virtual internship and/or offering virtual interns the opportunity to serve as team leads for future hackathon projects. The hackathon thereby will enable many students to gain brief insight into computational biomedical research, while the virtual internship will provide an opportunity for exceptional students to work in-depth, longer-term on a dedicated research project in a one-to-one mentorship setting.

For future computational biology internships, we recognized the importance of an early, in-depth training in accessing and using the high-performance computing cluster resources. Such training could be organized, together with other fundamental computing/programming concepts, as "boot-camp" style training that spans the entire first week, instead of a short introduction. A group boot-camp at the beginning of the program may also help build a sense of community among all interns in the cohort. Additionally, more social networking and community building during the internship period was desired by interns, especially as the COVID-19 pandemic had restricted other social interactions.

From the outcomes surveys, we learned that the accompanying biweekly lab meetings provided interns an important framework for professional skills development and reporting on research progress milestones. The meetings can be improved in future programs by stimulating more frequent discussion and constructive feedback from mentors. Furthermore, through exposure to different projects and increased mentor engagement in the meetings, students will be better equipped to make informed choices on academic and career paths, for example, when choosing a suitable topic and a good laboratory for a successful doctoral thesis project.

In addition to the students, mentors benefitted tremendously from the experience. The Department leadership allowed complete freedom to mentors to realize, together with the mentee, a clearly defined research goal. Moreover, the mentors often were trainees (graduate students or postdoctoral fellows), who had fewer prior experiences with research mentoring of an undergraduate student. For these mentors, the U-Hack Med Gap Year program was a supportive environment to grow their mentoring skills, with engaged faculty to advise and guide their mentoring as needed.

Together, in their 9-month U-Hack Med Gap Year internship, students achieved progress toward their research goals and mastered new technical skills, often with little previous knowledge of their subject matter. Fundamental to this process were close mentoring, being part of a cohort, and having a regularly scheduled, shared U-Hack Med Gap Year lab meeting. Overall, the internship was successful in contributing to the students' and mentors' academic and career advancement.

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

SD and HM-F wrote the original draft of the manuscript, SD, HM-F, GD and RC edited and revised the manuscript, GD and RC were responsible for supervision, project administration and funding acquisition.

FUNDING

This work was supported by the Lyda Hill Philanthropies.

ACKNOWLEDGMENTS

We would like to thank all the other mentors of the U-Hack Med Gap Year Internship Program: Albert Montillo, Bo Li, Son

Nguyen, Alex Treacher, Daria Beshnova, Bo-Jui Chang, Reto Fiolka, Xinxin Wang, Meghan Driscoll, Jian Zhou, Andrew Jamieson, Jeon Lee, Daehwan Kim, and the administrative staff at the Lyda Hill Department of Bioinformatics: Neha Sinha. We also thank Bo-Jui Chang, Andrew Jamieson, and Erik Welf for help in generating the project overview figure. In addition, this research was supported in part by the computational resources provided by the BioHPC supercomputing facility located in the Lyda Hill Department of Bioinformatics, UT Southwestern Medical Center, TX. URL: https://portal.biohpc.swmed.edu.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2021.727066/full#supplementary-material

REFERENCES

Abbasi, M. S., Ahmed, N., Sajjad, B., Alshahrani, A., Saeed, S., Sarfaraz, S., et al. (2020). E-learning Perception and Satisfaction Among Health Sciences Students amid the COVID-19 Pandemic. *Work* 67 (3), 549–556. doi:10.3233/WOR-203308

Battaglia, E., Boehm, J., Zheng, Y., Jamieson, A. R., Gahan, J., and Majewicz Fey, A., (2021). Rethinking Autonomous Surgery: Focusing on Enhancement over Autonomy. *Eur. Urol. Focus* 7, 696–705. doi:10.1016/j.euf.2021.06.009

Daniel, S. J. (2020). Education and the COVID-19 Pandemic. *Prospects* 49, 91–96. doi:10.1007/s11125-020-09464-3

Driscoll, M. K., Welf, E. S., Jamieson, A. R., Jamieson, K. M., Dean, T., Fiolka, R., et al. (2019). Robust and Automated Detection of Subcellular Morphological Motifs in 3D Microscopy Images. *Nat. Methods* 16 (10), 1037–1044. doi:10.1038/s41592-019-0539-z

Kogan, M., Klein, S. E., Hannon, C. P., and Nolte, M. T. (2020). Orthopaedic Education during the COVID-19 Pandemic. *J. Am. Acad. Orthop. Surg.* 28 (11), e456–64. doi:10.5435/JAAOS-D-20-00292

NSF (2017). Higher Education in Science and Engineering. S&E Degrees by Sex. 2017 Available at: https://ncsesnsfgov/pubs/nsb20197/demographic-attributes-of-s-e-degree-recipients#figureCtr714 (Accessed June 12, 2021).

Rani, M., Dhok, S. B., and Deshmukh, R. B. (2018). A Systematic Review of Compressive Sensing: Concepts, Implementations and Applications. *IEEE Access* 6, 4875–4894. doi:10.1109/ACCESS.2018.2793851

Sidpra, J., Gaier, C., Reddy, N., Kumar, N., Mirsky, D., and Mankad, K. (2020). Sustaining Education in the Age of COVID-19: A Survey of Synchronous Web-Based Platforms. *Quant Imaging Med. Surg.* 10 (7), 1422–1427. doi:10.21037/qims-20-714

Slavin, R. E., and Nathan, S. (2020). The US Educational Response to the COVID-19 Pandemic. *Best Evid. Chin. Edu* 5 (2), 617–633. doi:10.15354/bece.20.or027

UNESCO (2020). Education: From Disruption to Recovery. Available at: https://en.unesco.org/covid19/educationresponse.

Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., et al. (2018). Content-Aware Image Restoration: Pushing the Limits of Fluorescence Microscopy. *Nat. Methods* 15 (12), 1090–1097. doi:10.1038/s41592-018-0216-7

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.

The reviewer UG declared a past co-authorship with one of the authors SD to the handling editor.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

PERSPECTIVE
published: 14 October 2021
doi: 10.3389/feduc.2021.711618

The Bioinformatics Virtual Coordination Network: An Open-Source and Interactive Learning Environment

Benjamin J. Tully[1], Joy Buongiorno[2], Ashley B. Cohen[3], Jacob A. Cram[4], Arkadiy I. Garber[5], Sarah K. Hu[6], Arianna I. Krinos[7], Philip T. Leftwich[8], Alexis J. Marshall[9], Ella T. Sieradzki[10], Daan R. Speth[11], Elizabeth A Suter[12], Christopher B. Trivedi[13], Luis E. Valentin-Alvarado[14] and Jake L. Weissman[15] and on behalf of BVCN Instructor Consortium*

[1]Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA, United States, [2]Division of Natural Sciences, Maryville College, Maryville, TN, United States, [3]School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, NY, United States, [4]Horn Point Laboratory, University of Maryland Center for Environmental Science, Cambridge, MD, United States, [5]School of Life Sciences, Arizona State University, Tempe, AZ, United States, [6]Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, United States, [7]MIT-WHOI Joint Program in Oceanography/Applied Ocean Science and Engineering, Cambridge, Woods Hole, MA, United States, [8]School of Biological Sciences, University of East Anglia, Norwich, United Kingdom, [9]Thermophile Research Unit, Te Aka Mātuatua - School of Science, University of Waikato, Hamilton, New Zealand, [10]Environmental Science, Policy and Management Department, University of California, Berkeley, CA, United States, [11]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, United States, [12]Biology, Chemistry and Environmental Studies Department, Center for Environmental Research and Coastal Oceans Monitoring (CERCOM), Molloy College, Rockville Centre, NY, United States, [13]Interface Geochemistry, GFZ German Research Centre for Geosciences, Helmholtz Centre Potsdam, Potsdam, Germany, [14]Plant and Microbial Biology Department, University of California, Berkeley, CA, United States, [15]Department of Biological Sciences - Marine and Environmental Biology, University of Southern California, Los Angeles, CA, United States

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Aristóteles Góes-Neto, Federal University of Minas Gerais, Brazil
Miguel Rocha, University of Minho, Portugal

***Correspondence:**
Benjamin J. Tully
tully.bj@gmail.com

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *18 May 2021*
Accepted: *17 September 2021*
Published: *14 October 2021*

Citation:
Tully BJ, Buongiorno J, Cohen AB, Cram JA, Garber AI, Hu SK, Krinos AI, Leftwich PT, Marshall AJ, Sieradzki ET, Speth DR, Suter EA, Trivedi CB, Valentin-Alvarado LE and Weissman JL (2021) The Bioinformatics Virtual Coordination Network: An Open-Source and Interactive Learning Environment. Front. Educ. 6:711618. doi: 10.3389/feduc.2021.711618

Lockdowns and "stay-at-home" orders, starting in March 2020, shuttered bench and field dependent research across the world as a consequence of the global COVID-19 pandemic. The pandemic continues to have an impact on research progress and career development, especially for graduate students and early career researchers, as strict social distance limitations stifle ongoing research and impede in-person educational programs. The goal of the Bioinformatics Virtual Coordination Network (BVCN) was to reduce some of these impacts by helping research biologists learn new skills and initiate computational projects as alternative ways to carry out their research. The BVCN was founded in April 2020, at the peak of initial shutdowns, by an international group of early-career microbiology researchers with expertise in bioinformatics and computational biology. The BVCN instructors identified several foundational bioinformatic topics and organized hands-on tutorials through cloud-based platforms that had minimal hardware requirements (in order to maximize accessibility) such as RStudio Cloud and MyBinder. The major topics included the Unix terminal interface, R and Python programming languages, amplicon analysis, metagenomics, functional protein annotation, transcriptome analysis, network science, and population genetics and comparative genomics. The BVCN was structured as an open-access resource with a central hub providing access to all lesson content and hands-on tutorials (https://biovcnet.github.io/). As laboratories reopened and participants returned to previous commitments, the BVCN

evolved: while the platform continues to enable "a la carte" lessons for learning computational skills, new and ongoing collaborative projects were initiated among instructors and participants, including a virtual, open-access bioinformatics conference in June 2021. In this manuscript we discuss the history, successes, and challenges of the BVCN initiative, highlighting how the lessons learned and strategies implemented may be applicable to the development and planning of future courses, workshops, and training programs.

Keywords: bioinformatics, educational initiative, computational biology, microbiology, open source, free educational resource

INTRODUCTION

As governments across the world began issuing various lockdowns, "stay-at-home" orders, and quarantines to curtail the spread of the coronavirus, academic labs that were considered nonessential were shuttered. It was unclear how long these closures would last, but initial estimates assumed at least 6–8 weeks. However, social distancing restrictions and other mitigation protocols ultimately minimized laboratory activities in many parts of the world for far longer. We developed the Bioinformatics Virtual Coordination Network (BVCN) due to the concerns raised on social media in mid-March 2020 about the impact that lab closures would have on the progress of graduate students and early-career researchers. This network is a collective of international early career microbiology researchers; with experience and expertise in bioinformatics and computational biology. Our initial mission was to provide an outlet for bench and field researchers to learn computational methods to pursue alternative research during laboratory closures or analyze data generated after returning to normal research activities. At the start of a global pandemic, accomplishing this goal required a flexible approach that could quickly ramp up to meet the needs of an international audience.

The BVCN joins a growing number of bioinformatics training resources (e.g., Wibberg et al., 2019; https://datacarpentry.org/) that have been prescribed as a necessity to teach life science researchers the skills required for large-scale data analysis (Attwood et al., 2017; Barone et al., 2017; Batut et al., 2018; Williams et al., 2019). The BVCN has a microbiology centric approach, with many of the core concepts and skills applicable to wider life science data analysis (Welch et al., 2014; Wilson Sayres et al., 2018). In the early phases of developing the BVCN, content distribution was envisioned as live events, combining microlectures on specific topics with hands-on demonstrations that could function as standalone lessons or part of an extended series on a broad topic. We developed a platform that built on the successes of previous short-form training courses and workshops (DIBSI: http://ivory.idyll.org/dibsi/toc.html, ECOGEO: doi. org/10.17504/protocols.io.fjjbkkn, STAMPS: https://mblstamps.github.io/, etc.) and then adapted these resources to accommodate the needs of a group of instructors and learners spread throughout the world. What evolved was a decentralized platform that has persisted beyond the end of lesson development, which coincided with the relaxing of lockdown orders and a return to in-person laboratory research.

We provide details on the critical aspects of what made the BVCN a success, along with the challenges we encountered and reflections on how to mitigate those challenges. The BVCN leveraged the varied expertise of its instructors to use multiple online tools and platforms to design material and learning environments tailored to a specific topic. As one of the successes of the BVCN model, this provided experience to learners on how to use computational resources that mirrored those required for genuine research. These lessons are freely available and mutable through a public GitHub repository (https://biovcnet.github.io/). The BVCN made early commitments to establishing an inclusive environment with explicit goals to lower the entry barriers for researchers to participate in computational biology. Now that formal lesson developments have ended, we have a library that allows interested learners to pursue self-led, "point of need" instruction of microbial bioinformatics and computational biology when suitable for their particular research needs and questions.

Implementation of the Education Program

In March 2020, the BVCN lead, Dr. Benjamin Tully, put out an announcement on Twitter to gauge interest in teaching and learning bioinformatics during the pandemic. Within a few days, the announcement had gathered interest from more than 50 computational biology educators and several hundred participants. Following an introductory virtual meeting among interested instructors that discussed possible avenues for distributing lessons and the breadth of topics to include, we created a BVCN Slack workspace (Teckchandani, 2018; https://slack.com/). As instructors, we self-assigned into topics and chose one person to act as a coordinator for each topic. Based on the breadth of bioinformatics research disciplines represented by the instructors, we chose to include lessons addressing Unix, R programming, amplicons, metagenomics, functional annotation, transcriptomics, network analysis, population genetics and comparative genomics, and Python programming (**Table 1**). With an aim to promote reproducible science and good open data practice, we also created the Reproducibility Challenge (see below) to encourage learners to reproduce bioinformatic results from published research. We organized communication amongst the instructors through Slack and Zoom (https://zoom.us/), with collaborative editing on Google Docs and Sheets within

TABLE 1 | List of BVCN topics with how material was delivered, and examples of core concepts reviewed (https://github.com/biovcnet/biovcnet.github.io/wiki).

Topic	Modes of delivery	Example of tools in tutorials	No. of lessons	No. of associated YouTube videos	Slack channel membership
R	RStudio cloud; R Markdown	Tidyverse; ggplot2; phyloseq	9	14	303
Python	Binder	Pandas	9	11	206
Unix	Binder	Conda	6	6	179
Network Science	RStudio Notebook	—	8	8	116
Amplicons	RStudio cloud; Cyverse	Qiime2; DADA2; phyloseq; mothur; vegan	7	8	191
Metagenomics	Binder	FastQC; MultiQC; bbtools; Kraken2; sourmash; CheckM; MetaBat; BinSanity; DASTool	7	15	310
Functional Annotation	Binder	Prodigal; GeneMark; BLAST; DIAMOND; HMMER; FeGenie; BlastKOALA; antiSMASH	8	17	224
Transcriptomics	Jupyter notebooks; Binder	htseq-count; Trinity	4	6	180
Population genetics and comparative genomics	Binder	PAML; PGLS	3	4	131

Tool citations. Tidyverse (https://www.tidyverse.org/), ggplot2 (Wickham, 2016), phyloseq (McMurdie and Holmes, 2013), Pandas (McKinney, 2010), Conda (https://conda.io), mothur (Schloss et al., 2009), vegan (https://github.com/vegandevs/vegan), FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), MultiQC (Ewels et al., 2016), Kraken2, sourmash (Titus Brown and Irber, 2016), CheckM (Parks et al., 2015), MetaBat (Kang et al., 2015), BinSanity (Graham et al., 2017), DASTool (Sieber et al., 2018), Prodigal (Hyatt et al., 2012), GeneMark (Besemer et al., 2001), BLAST (Altschul et al., 1997), DIAMOND (Buchfink et al., 2014), HMMER (Finn et al., 2011), FeGenie (Garber et al., 2020), BlastKOALA (Kanehisa et al., 2016), antiSMASH (Blin et al., 2019), htseq-count (https://htseq.readthedocs.io), Trinity (Grabherr et al., 2011), PAML (Yang, 2007)

a shared Google Drive (https://www.google.com/drive/), accessible to all instructors. We selected GitHub (https://github.com/) as a central repository for collaboratively creating and editing tutorial material and YouTube (https://youtube.com/) for sharing recorded microlectures. Instructors leading each topic had the flexibility and independence to plan lessons in the formats most relevant to the strengths and weaknesses of the specific topic and/or lesson. While topics functioned semi-independently, weekly meetings between all instructors and coordination with the Project Lead and the Coordinating Committee helped foster the sharing of resources and creative approaches.

We started creating lessons in April continuing through August 2020, and developed lessons for each topic in various formats, with a majority scheduled and planned as synchronous learning events, either as a lesson, short-format tutorial, or live Q and A discussion. At its peak, all topics produced about one lesson per week. We recorded all of our lessons and made these available on YouTube (https://www.youtube.com/channel/UC5qVqcvUPfgPQWOhBaR_Low), along with datasets, code, and any other necessary materials. When applicable, short-format tutorials were executed in shareable computing environments (see below: platforms and tools), with the explicit goal of enabling learners to complete tutorial material without needing to modify their local computing environments, while also providing a genuine experience equivalent to what would be required for the participants' own research. All of BVCN was implemented with a forward-looking approach; our materials are available indefinitely and can be accessed and modified by others. The BVCN Slack workspace, organized in channels based on topic, continues to be a community resource where users can post broad or specific questions regarding their research and have conversations between learners and instructors.

Platforms and Tools Incorporated Into the Lesson Plans

Each topic of the BVCN has common features, as well as unique teaching platforms and systems determined by the set of instructors for that topic. We decided in the early planning meetings that every course should contain a live component, when possible, for learners interested in a synchronous approach to the lessons, and all lectures and tutorials should be recorded for asynchronous viewing at the convenience of the learners. Components needed to follow along asynchronously, such as interactive code and links to video lessons, were collated by lesson and organized through the wiki page (https://github.com/biovcnet/biovcnet.github.io/wiki). Lecture components tended to be conceptual in nature, providing essential background to the tools that we used in the tutorials. Tutorials provided an example of an applied approach using standard toolsets for accomplishing a specific computational task. For example, Lesson 4 of the metagenomics topic includes a lecture, recorded live, on the principles of read mapping and some of the main tools and algorithms used in this practice. We followed this lecture with two short-format tutorials for the read mapping tools Bowtie2 (Langmead and Salzberg, 2012) and BBmap (Bushnell et al., 2017). We delivered synchronous content as online meetings on Zoom. Meetings varied between formal microlectures and less formal Q and A sessions. Synchronous meeting planning, posting of asynchronous content, discussions about lesson content, and direct responses to learner questions were all hosted through Slack.

Each topic had a different set of instructors and within these topics the instructors chose the specific delivery platforms for the interactive course content. Two main strategies emerged (**Table 1**). Most topics packaged content using interactive Binder (https://mybinder.org/) environments which included installed dependencies, tools, and data, which gave learners access to an online Unix terminal interface that could be used to perform small-scale tasks on real data. The major advantage of using Binder environments is the ease with which tutorials can be launched regardless of a student's local computing resources and avoiding any software incompatibility, as Binder operates in the cloud using a web browser. However, this flexibility comes with the tradeoff that Binder environments are limited in the size of the data that can be stored and the processing power that can be applied. Other topics, like the R programming and amplicon channels, used the RStudio Cloud platform which offers similar functionality to Binder but is optimized to reflect an R coding environment. Budgetary restrictions put an upper limit on the amount of content we could release using the RStudio Cloud (https://rstudio.cloud/) platform (e.g., capacity for number of participants, storage, RStudio Projects, etc.). Other topics (e.g., network science) encouraged students to download lesson content from the BVCN GitHub and perform local analysis, which had the advantage of no explicit limits on the number of lessons, but required learners to possess some proficiency in setting up local computing environments before they could interact with the content.

Establishment of Diversity and Equity Statement

As a community-led project, we aspire to make the BVCN a safe and open workspace where bioinformatics can be taught and learnt by a diverse audience. Our founding document, the BVCN Code of Conduct (https://biovcnet.github.io/code-of-conduct/), describes our commitment to diversity and inclusion. Building off of the codes of conduct of multiple other open-source communities, the document's strength is that it details how the BVCN will provide accessible content, what the community expects of its members, and provides detailed examples as to what the community defines as acceptable and unacceptable behaviors. This code of conduct drove how we worked within the BVCN, enhancing the virtual learning experience of a wider global audience. We actively worked towards decreasing entry barriers into bioinformatics, including knowledge access, support, and representation. We see this as especially important for reaching and providing access to an international community of early career researchers. For many, the access to resources, such as textbooks, workshops, training courses, research experience, and computational infrastructure, can prevent interested scientists from pursuing bioinformatics.

Our primary avenue for promoting the BVCN was through professional networks on Twitter and word of mouth. Access to the growing archive of lessons and short-format tutorials, along with the Slack workspace detailing active lesson deployment, helped to facilitate access to formal support, information exchange, and collaborations. This approach aimed to lower the entry barriers between awareness (i.e., promotion) and the use of BVCN content as a resource (i.e., ease of access to archived and ongoing support), which is identified as a key guiding principle for bioinformatics education (Williams et al., 2010). The high degree of connectivity between instructors and learners laid a strong foundation for establishing mentee-mentor relationships. The combined experiences of our instructors, who represent diversity in gender, ethnicity, career stage, and academic background, provided opportunities for learners to observe representation within the field. Gender-balance at the instructor level (~59% of instructors identify as women) offered the potential for same-gender mentoring, which has been shown to increase belonging, self-efficacy, and retention in STEM and engineering fields (Dennehy and Dasgupta, 2017).

Despite these efforts, we have identified areas of current and future improvement within our approach to diversity and inclusion. One aspect that we feel could have a significant impact on learner engagement would be to increase efforts to develop formal mentor-mentee and/or peer-to-peer partnerships with a targeted effort to enhance the recruitment of historically excluded groups. In order to have the most impact, and truly expand on the BVCN's mission of increasing representation of historically excluded groups in bioinformatics, we need to continue to work to enhance the visibility and accessibility of our resources and support network. We are committed to reducing entry barriers through continuous improvements in participant recruitment and retention through purposeful global outreach (see below: ongoing collaborations) to raise awareness about the free resources and support supplied through the BVCN.

Evolution of Approaches as Topics Emerge

As instructors, learners, and topics harmonized over our various platforms, custom lesson plans were crafted. One common challenge learners expressed was how to select the "correct" tool or pipeline for an analysis and then incorporate it into their own research. Many topics provided opinions about "best practices" and highlighted multiple approaches that allowed learners to pick and choose how they approached complex topics. We regularly emphasized that there is often more than one way to approach a problem without sacrificing the quality of the results. To demonstrate this more thoroughly, we addressed the question "Which tool(s) should one use to perform amplicon analysis" by developing lessons that showcased the most commonly used environments for the analysis of 16 S/18 S rRNA gene or intergenic spacer region (ITS) amplicons. This series of lessons kicked off with a live Q and A session where several instructors discussed the methods by which we conduct amplicon sequence analysis (https://youtu.be/egkCswqQMWM). The goal of this "fireside chat" was to demonstrate that our individual bioinformatic practices vary and shed light on current trends within the interest. We then designed parallel lessons using an identical initial dataset.

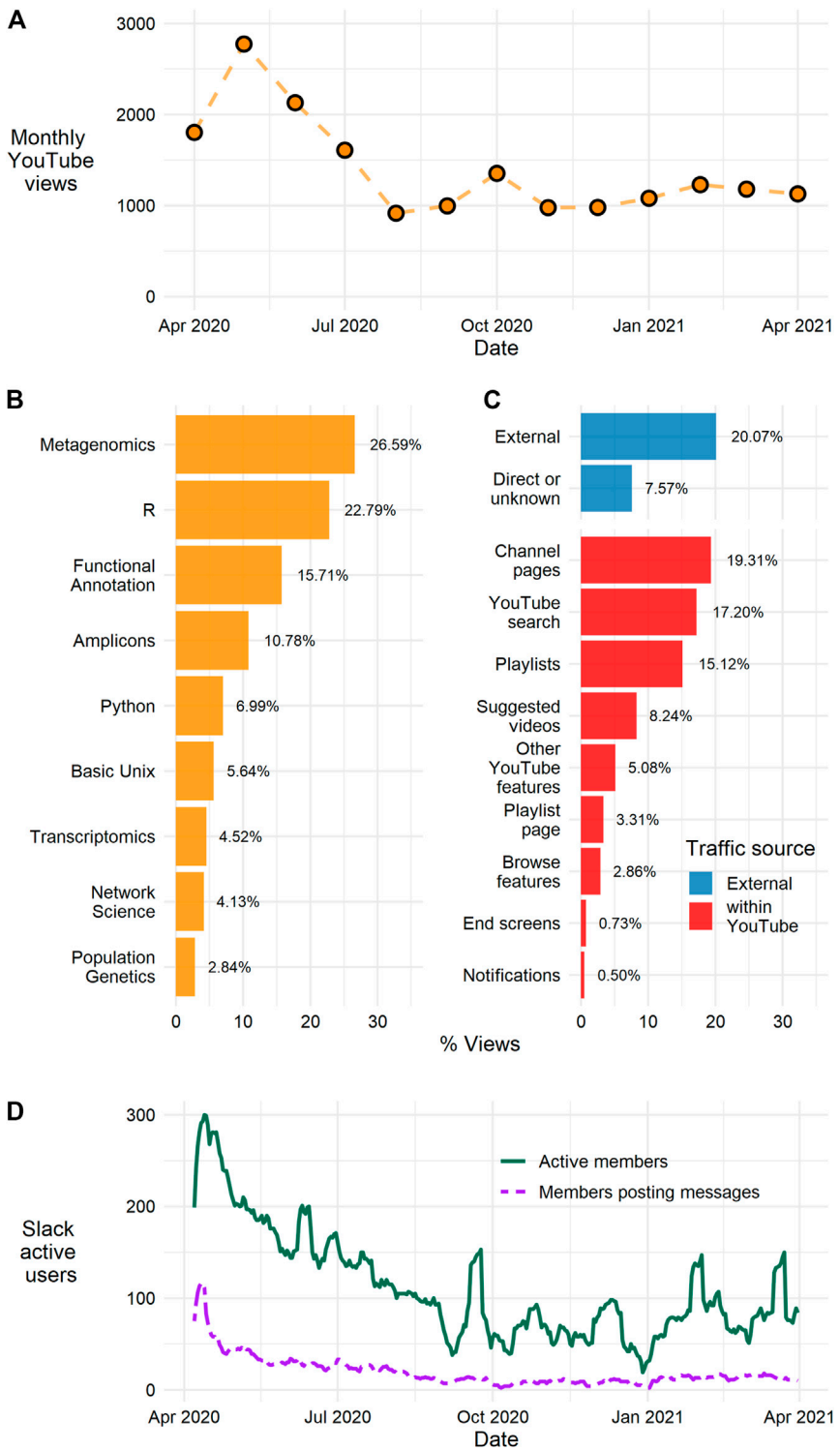


FIGURE 1 | Overview of the number of views of BVCN YouTube videos and Slack membership. **(A)** Monthly viewing figures for the BVCN YouTube channel; the total accumulated views for the channel is 16,325 views (accessed 29 April 2021), views per month peaked in May 2020 at 2774, and viewing numbers have fallen but stabilized at an average of over 1,000 views per month from August 2020 to present. **(B)** The percentage of total views for the BVCN channel split by subject playlists; the Metagenomics and R playlists are the first and second most viewed playlists accounting for almost half of the total views for the BVCN channel. **(C)** The percentage of total views for the BVCN channel split by "traffic source" top bars (blue) indicate views came from sources external to YouTube, bottom bars (red) indicate views that originate within the YouTube website. The top four bars (External, Direct or unknown, Channel pages, YouTube search) are views that originate from "searches with intent" viewers actively searching for BVCN videos ?these account for 64% of total views. **(D)** The seven-day rolling average for Slack user activity; active members are those who used the Slack workspace in the last seven days (solid green line), and the number of users that have posted messages to the workspace (dashed purple line). Slack use peaked in April 2020 and stabilized at between 50 and 100 active users per month from September 2020.

These data were processed using the tutorials from Happy Belly Bioinformatics (Lee, 2019), which implements the DADA2 pipeline (Callahan et al., 2016) in R (R Core Team. R, 2020) and the QIIME2 pipeline (Bolyen et al., 2019) on the cloud-computing infrastructure platform CyVerse (Merchant et al., 2016).

Discussions highlighting variations in approaches typically occur at conferences or in more casual settings in the workplace; without these interactions, learners working through their first datasets may feel isolated. Conversations about alternative approaches, which occurred during recorded lessons, on Slack, or during scheduled office hours, were a critical part in building a community of learning within the BVCN. Even after lesson development ended and the BVCN shifted to an asynchronous learning format, the Slack workspace continues to serve as a valuable resource for learners who may have recently completed the coursework and are pursuing follow-up questions regarding how to apply tools to their own datasets.

Asynchronous Audience and Ongoing Membership

As might be expected for an initiative started in response to a global event that caused the forced shutdown of laboratory research, the highest level of audience engagement occurred in the initial phases of government-ordered shutdowns. Using the metric of YouTube channel views, it is clear that the BVCN continues to reach an audience interested in accessing our archive of video and written tutorial materials. We collected viewership data from YouTube (accessed April 29, 2021) for the BVCN channel which hosts 95 videos (lectures and tutorials) and has 562 subscribers. The videos had a total of 16,325 cumulative views and over 1,700 cumulative watch hours. BVCN content was developed and deployed almost entirely between April 6 and August 4, 2020, approximately during the height of laboratory lockdowns. Monthly viewings peaked at 2,774 views during May 2020, shortly after we started, and dropped through August 2020. Since the formal end of BVCN lesson development, we have sustained approximately 1,000 views per month (**Figure 1A**). By cumulative views, metagenomics, R, functional annotation, amplicons, and Unix were the top five viewed subject playlists (**Figure 1B**).

A four-week window was selected from 15 February to March 14, 2021 to assess what type of users were accessing the video content. During this period, 488 unique viewers and 48 returning viewers visited the channel. YouTube provides information on where "traffic sources" originate from, quantifying how viewers reached videos on the BVCN channel (e.g., external links, visiting the channel page, YouTube search, etc.). About a fifth of traffic to YouTube originated from external links, likely learners directed from the BVCN wiki or social media posts. Roughly three-quarters of total views originated from within YouTube (**Figure 1C**), but around 64% of total traffic (internal and external to YouTube) came from "searches with intent" (i.e., users searching or accessing BVCN content directly; **Figure 1C**).

The BVCN Slack workspace remains active and provides a central location for 630 members of the community to exchange ideas, share resources, and troubleshoot problems (**Figure 1D**). Activity within Slack mirrors that of the views tracked by YouTube. Weekly active members spiked in the first month following the start of lockdowns, but steadily declined through September 2020. Since that time, the number of weekly members has stabilized to between 50 and 100 participants who are engaging with the workspace either through reading public posts, participating in private channels, or direct messaging. Data collected by Slack from a recent 30-days window (April 5 to May 3, 2020) indicates that most of this participation (47%) is based on users accessing messages in public channels and not governed strictly by direct messages (37%) or private channels (16%).

Reproducibility Challenge

In collaboration with Drs. Harriet Alexander and Maria Pachiadaki, instructors for a graduate-level course in bioinformatics at the Woods Hole Oceanographic Institution (WHOI), we developed a research task for students to enhance their hands-on bioinformatics skills. Building off of the existing knowledge base of the WHOI course, students would select a bioinformatics analysis in a publication and attempt to recreate the outputs as presented by the authors. The goal was to encourage learners to consider what would be required to create reproducible research and was accomplished by providing intermediate learners a platform to perform analysis on an existing dataset with the goal of reproducing published results. Additionally, we sought to foster growth within the learner community and provide learners with a network for peer-to-peer mentorship while working on a formalized collaborative project. We created an accompanying tutorial website (https://alexanderlabwhoi.github.io/BVCNReproducibility/intro.html) explaining some of the fundamental stages of a bioinformatics project, such as data curation/management, tool installation, and reproducible coding platforms, and provided two example datasets and exemplar code which learners could use to reproduce published research results. Learners and instructors were matched with other interested partners through an online form and learners were provided an example timeline to complete the research task. Ultimately, few learners participated in the project based on the provided timeline. While one project was successful and resulted in establishing a network of researchers working on a collaborative project, many projects suffered from scheduling challenges and time constraints. As with many facets of the global pandemic, it appeared that participants became overwhelmed with responsibilities and were more comfortable with using the materials for asynchronous learning.

DISCUSSION

Simultaneous Building of Learning Resources and a Community

The impromptu, grassroots nature of the BVCN came with a number of challenges, which were compounded by the COVID-19 pandemic. We attempted to simultaneously develop a novel learning resource while building a community–two tasks with an immense amount of

complexity even under normal conditions. To this end, we saw several areas for which the BVCN has accomplished the goals laid out in early planning phases:

1) *Provide a platform that encourages the dissemination of knowledge from experienced bioinformaticians and computational biologists to those new in the field.* This pedagogical approach leveraged our collective research experiences to help guide learners with hesitancy about using bioinformatic techniques towards a place where they felt comfortable exploring these concepts. We aggregated a number of introductory resources (e.g., the Happy Belly Bioinformatics Unix tutorial (Lee, 2019)) and paired them with lessons that advanced in complexity, allowing learners to see the progression from introduction to practical use.
2) *Implementation of a community educational resource that has previously been limited in accessibility due to numerous constraints.* While there are initiatives of learning these types of skills, many of them occur through short course formats (Attwood et al., 2017) that may have restrictions related to attendance, cost, or location. We took advantage of the limitations imposed by the pandemic by providing a resource that learners can access at their own pace, with sole limitations being access to an internet connection and personal computer, while still interacting with the instructors and their colleagues. The pedagogy of the lessons is also designed with an active learning approach in mind. Conceptual lectures are paired with tutorial content which allow learners to actively explore bioinformatic tools and techniques which tend to have better learning outcomes (Markant et al., 2016). Broadly, our goal with the BVCN was to assist in bridging life science researchers without experience in bioinformatics/computational biology towards an understanding of the tasks involved in this type of research. The persistent nature of the BVCN platform can assist in achieving this goal for years to come.
3) *Establishing an international community, oriented around microbiology, bioinformatics, and computational biology.* In many instances, forming national and international collaborations amongst early career researchers occur at meetings and conferences. As with many things, the pandemic has prevented these events, which may ultimately have an impact on these types of relationships in the long term. With the BVCN, we have built an international community, predominantly of early career researchers, that has the capacity to persist beyond the months of active lesson development and act as a meeting place for bench/wet-lab and/or field scientists to interface with bioinformaticians and computational biologists. Further, the instructors and learners of the BVCN have initiated multiple projects, activities, and collaborations that would not have occurred had it not been for the efforts of the platform (see below: Emergent Collaborative Projects).

Continued Challenges to Building an Online Learning Initiative

Some of the challenges we experienced appear to be specific to the pandemic and reflected in other similar experiences, but there are others which we could have effectively addressed during the initiation of the BVCN had the Project Lead, Coordinating Committee, and instructors been aware of such issues arising in other online initiatives. Future bioinformatics training initiatives may benefit from considering issues that we encountered highlighted below:

1) *Rapidly declining attendance/interest in organized live sessions for topics after the first several weeks.* As supported by viewership on YouTube and active member data on Slack, many learners shifted away from active participation and towards an asynchronous learning approach. While videos from all sections continue to accrue views, participation in the interactive Zoom sections tapered off after the initial pandemic lockdowns (June/July 2020). While additional live sessions mirroring standard "office hours" were established, these sessions also saw a rapid decline in attendance. Successes like the merged amplicon and R lesson (described above) resulted from active feedback from learners, but the decrease in instructor-learner interactions had impacts on our ability to develop customized BVCN lessons. Some small groups of learners and instructors have continued to meet through 2021 (e.g., weekly network science office hour). Without a direct incentive for attending lessons consistently, learners adjusted to a mode of knowledge sharing suitable for their personal research needs. This was likely driven by the ability of BVCN lessons to provide "point-of-need" training on specific concepts and tools, more suitable for researchers as they actively perform analyses. As noted before, many existing bioinformatics workshops have solved this by selecting a limited cohort of students to be fully immersed in the learning experience but inherently restrict the number of participants.
2) *No system or time to iteratively improve lessons or strategies.* The speed at which the BVCN was deployed was in direct response to the emotions and fears spun out of the early phases of the pandemic. As such, we prioritized the consistent production of new lessons. The trade-offs to this approach were clear. While it provided resources for learners immediately, it meant that our limited time was focused on producing additional content, not necessarily on returning to improve, refine, or iterate previous content. An alternative could have been to form our instructor base for the BVCN and spend several months producing and refining lesson content prior to release, though the downside would have been the trajectory of the pandemic itself, where most BVCN instructors and learners saw their availability and priorities shift back towards their teaching and/or research commitments, as conditions in the pandemic stabilized (see below: collaborations with KBase).

3) *Difficulty assessing the degree to which lessons, stylistic choices, use of computational infrastructure, etc. assisted learners in achieving their goals.* This challenge likely is a result of the initial infrastructure we established. Efforts were made from the onset to make all elements open access and reduce the "cost" of participation within the BVCN, which we saw as sufficient to encourage ongoing participation. However, without formal checkpoints or feedback routes, it was difficult to assess the needs of learners. During the peak of activity, it would have been useful if we had implemented post-lesson assessments that could have tracked sentiments and/or encouraged learners to engage in the future direction of lessons.

Emergent Collaborative Projects

One of our continuing successes from the BVCN has been the number of collaborative projects that have been initiated as a direct result of the interaction between instructors and learners. One example of such a collaboration emerged from the network science topic. In that group, discussions between instructors and learners led to the conclusion that the current "best practice" for network analysis did not address many common statistical considerations of time-series analysis. One learner, in particular, with support from two of our instructors, took the lead on developing a novel approach that is robust to common statistical artifacts and that works with unevenly spaced time-series data. Another collaborative project initiated between the instructors from WHOI and the University of Southern California has seen the development of multiple complementary approaches for the large-scale analysis of eukaryotic metagenome-assembled genomes. Additionally, the relationships created between instructors has led to additional opportunities for cross-pollination between fields. For example, an instructor-led seminar series for the Center for Dark Energy Biosphere Investigations tapped BVCN instructors as presenters for a seminar about novel and upcoming bioinformatic approaches.

One of the current ongoing projects is to continue the educational legacy of the BVCN. As the development of new content in the BVCN channels slowed, we looked to new and alternative ways our community could facilitate bioinformatics learning more broadly. We identified a gap in bioinformatics education: while tools and some workflows may have good documentation, a holistic overview of entire bioinformatics projects, including the decisions as to which tools to use and how to combine them, was missing. To address this education gap, we envisioned a virtual conference focused on open-science methods, where featured speakers outlined their bioinformatics pipelines and provided open data and code alongside their talks. In this format, the speakers would be asked to put special emphasis on the many difficult decisions and trade-offs that go into shaping and executing a project. In addition to filling a gap in the bioinformatics education infrastructure, we identified during the early exploratory stage of conference planning that a virtual conference offered opportunities to make the event more accessible to a wide audience in ways a traditional conference never could. As part of these continuing efforts, the BVCN will welcome over 200 attendees from around the world to "A BVCN Training Conference: Holistic Bioinformatic Approaches used in Microbiome Research" in June 2021 supported by the Code for Science and Society Event Fund (https://eventfund.codeforscience.org/).

The BVCN represents one of many educational initiatives started during the COVID-19 pandemic. Early in the development of the BVCN, we collaborated with the educational directive led by Drs. Ellen Dow and Elisha Wood-Charlson at the Department of Energy Systems Biology knowledgebase (KBase). Multiple conversations were held to determine how the BVCN and KBase could support each other's educational initiatives. The BVCN concentrated on short-format tutorials that utilized command line or coding examples, while KBase built educational material that used their cloud-based, graphical user interface (GUIs). BVCN content was directed at early careers researchers and the KBase educational directive, supported by tenured and tenure-track faculty, developed lesson plans for undergraduate microbiology majors (Dow et al., 2021). These lessons form a complement to those created by the BVCN and reflect the strength of formal support through a funded sponsor and long-term planning for lesson development and maintenance.

CONCLUDING REMARKS

The popularity of the BVCN makes it clear that there is a need for the bioinformatic training of life scientists, including microbiologists, at all career stages. With the BVCN, we have laid the groundwork for filling one of the most persistent gaps in bioinformatic training, moving beyond introductory content towards intermediate levels of expertise. We provided a grassroots response to the needs of the community at a time of heightened global uncertainty, and we hope that future educational initiatives, either direct descendants of the BVCN or inspired by, may be able to satisfy the continued need for these resources and take the lessons and approaches applied in this initiative to achieve further success.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. Data on YouTube and Slack usage, along with code used to produce figures, can be found at https://github.com/biovcnet/BVCN_stats

AUTHOR CONTRIBUTIONS

BT, JB, AC, JC, AG, SH, AK, PL, AM, ES, DS, ES, CT, LV-A, and JL wrote and edited the article. BT founded the Bioinformatics Virtual Coordination Network. The BVCN Instructor Consortium consists of Michael D. Lee, Harriet Alexander, R. Eric Collins, Maria Pachiadaki, Adelaide Rhodes, Wayne Decatur, who along with the authors of the article, contributed lessons and/or expertise during the Bioinformatics Virtual Coordination Network.

FUNDING

BT was supported by the Center for Dark Energy Biosphere Investigations (OCE-0939654). The BVCN Training Conference was funded in full by a grant from Code for Science and Society, made possible by grant number GBMF8449 from the Gordon and Betty Moore Foundation. AM was supported by Smart Ideas award (UOWX1602) from the New Zealand Ministry of Business, Innovation and Employment and the Rutherford Foundation Royal Society Te Aparangi Postdoctoral Fellowship (20-UOW-006). ETS. was supported by the U.S. Department of Energy, Office of Biological and Environmental Research, Genomic Science Program, Scientific Focus Area award SCW1632 to Jennifer Pett-Ridge and award DE-SC0014079 to Mary K. Firestone. CT acknowledges financial support from the German Helmholtz Recruiting Initiative (award number: I-044–16-01) and from the European Research Council Synergy Grant ("Deep Purple" grant # 856416) awarded to Liane G Benning. AK was supported by the U.S. Department of Energy Computational Science Graduate Fellowship (DE-SC0020347). DS was supported by the Netherlands Organisation for Scientific Research, Rubicon award 019.153LW.039 and the US Department of Energy, Office of Science, Office of Biological and Environmental Research under award number DE-SC0016469 to Victoria J. Orphan. JL was supported by a postdoctoral fellowship in marine microbial ecology from Simons Foundation Award 653212. The Center for Dark Energy Biosphere Investigations (OCE-0939564) supported the participation of SH through a C-DEBI Postdoctoral Fellowship.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, 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

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

Barone, L., Williams, J., and Micklos, D. (2017). Unmet Needs for Analyzing Biological Big Data: A Survey of 704 NSF Principal Investigators. *Plos Comput. Biol.* 13, e1005755. doi:10.1371/journal.pcbi.1005755

Batut, B., Hiltemann, S., Bagnacani, A., Baker, D., Bhardwaj, V., Blank, C., et al. (2018). Community-Driven Data Analysis Training for Biology. *Cell Syst* 6, 752–e1. doi:10.1016/j.cels.2018.05.012

Besemer, J., Lomsadze, A., and Borodovsky, M. (2001). GeneMarkS: a Self-Training Method for Prediction of Gene Starts in Microbial Genomes. Implications for Finding Sequence Motifs in Regulatory Regions. *Nucleic Acids Res.* 29, 2607–2618. doi:10.1093/nar/29.12.2607

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S. Y., et al. (2019). antiSMASH 5.0: Updates to the Secondary Metabolite Genome Mining Pipeline. *Nucleic Acids Res.* 47, W81–W87. doi:10.1093/nar/gkz310

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

Buchfink, B., Xie, C., and Huson, D. H. (2014). Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* 12, 59–60. doi:10.1038/nmeth.3176

Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge - Accurate Paired Shotgun Read Merging via Overlap. *PLoS ONE* 12, e0185056. doi:10.1371/journal.pone.0185056

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J., and Holmes, S. P. (2016). DADA2: High-Resolution Sample Inference from Illumina Amplicon Data. *Nat. Methods* 13, 581–583. doi:10.1038/nmeth.3869

Dennehy, T. C., and Dasgupta, N. (2017). Female Peer Mentors Early in College Increase Women's Positive Academic Experiences and Retention in Engineering. *Proc. Natl. Acad. Sci. U S A.* 114, 5964–5969. doi:10.1073/pnas.1613117114

Dow, E. G., Wood-Charlson, E. M., Biller, S. J., Paustian, T., Schirmer, C. S., Sheik, W., et al. (2021). Bioinformatic teaching resources - for educators, by educators - using KBase, a free, user-friendly, open source platform. *Front. Educ.* doi:10.3389/feduc.2021.711535

Ewels, P., Magnusson, M., Lundin, S., and Käller, M. (2016). MultiQC: Summarize Analysis Results for Multiple Tools and Samples in a Single Report. *Bioinformatics* 32, 3047–3048. doi:10.1093/bioinformatics/btw354

Finn, R. D., Clements, J., and Eddy, S. R. (2011). HMMER Web Server: Interactive Sequence Similarity Searching. *Nucleic Acids Res.* 39, W29–W37. doi:10.1093/nar/gkr367

Garber, A. I., Nealson, K. H., Okamoto, A., McAllister, S. M., Chan, C. S., Barco, R. A., et al. (2020). FeGenie: A Comprehensive Tool for the Identification of Iron Genes and Iron Gene Neighborhoods in Genome and Metagenome Assemblies. *Front. Microbiol.* 11, 37. doi:10.3389/fmicb.2020.00037

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., et al. (2011). Full-length Transcriptome Assembly from RNA-Seq Data without a Reference Genome. *Nat. Biotechnol.* 29, 644–652. doi:10.1038/nbt.1883

Graham, E. D., Heidelberg, J. F., Tully, B. J., and Tully, B. J. (2017). BinSanity: Unsupervised Clustering of Environmental Microbial Assemblies Using Coverage and Affinity Propagation. *PeerJ* 5, e3035–19. doi:10.7717/peerj.3035

Hyatt, D., LoCascio, P. F., Hauser, L. J., and Uberbacher, E. C. (2012). Gene and Translation Initiation Site Prediction in Metagenomic Sequences. *Bioinformatics* 28, 2223–2230. doi:10.1093/bioinformatics/bts429

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J. Mol. Biol.* 428, 726–731. doi:10.1016/j.jmb.2015.11.006

Kang, D. D., Froula, J., Egan, R., and Wang, Z. (2015). MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities. *PeerJ* 3, e1165–15. doi:10.7717/peerj.1165

Langmead, B., and Salzberg, S. L. (2012). Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods* 9, 357–359. doi:10.1038/nmeth.1923

Lee, M. (2019). Happy Belly Bioinformatics: an Open-Source Resource Dedicated to Helping Biologists Utilize Bioinformatics. *Jose* 2, 53. doi:10.21105/jose.00053

Markant, D. B., Ruggeri, A., Gureckis, T. M., and Xu, F. (2016). Enhanced Memory as a Common Effect of Active Learning. *Mind, Brain Educ.* 10, 142–152. doi:10.1111/mbe.12117

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proc. 9th Python Sci. Conf.*, 56–61. doi:10.25080/Majora-92bf1922-00a

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

Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., et al. (2016). The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *Plos Biol.* 14, e1002342. doi:10.1371/journal.pbio.1002342

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015). CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes. *Genome Res.* 25, 1043–1055. doi:10.1101/gr.186072.114

R Core Team. R (2020). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. Available at: https://www.R-project.org/.

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., et al. (2009). Introducing Mothur: Open-Source, Platform-independent, Community-Supported Software for Describing and Comparing Microbial Communities. *Appl. Environ. Microbiol.* 75, 7537–7541. doi:10.1128/AEM.01541-09

Sieber, C. M. K., Probst, A. J., Sharrar, A., Thomas, B. C., Hess, M., Tringe, S. G., et al. (2018). Recovery of Genomes from Metagenomes via a Dereplication, Aggregation and Scoring Strategy. *Nat. Microbiol.* 3, 836–843. doi:10.1038/s41564-018-0171-1

Teckchandani, A. (2018). Slack: A Unified Communications Platform to Improve Team CollaborationSlack: A Unified Communications Platform to Improve Team Collaboration, Available at https://slack.Com/. *Amle* 17, 226–228. doi:10.5465/amle.2018.0061

Titus Brown, C., and Irber, L. (2016). Sourmash: a Library for MinHash Sketching of DNA. *JOSS* 1, 27–31. doi:10.21105/joss.00027

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

Wibberg, D., Batut, B., Belmann, P., Blom, J., Glöckner, F. O., Grüning, B., et al. (2019). The de.NBI/ELIXIR-DE training platform - Bioinformatics training in Germany and across Europe within ELIXIR. *F1000Res* 8, 1877. doi:10.12688/f1000research.20244.1

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer-Verlag.

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

Williams, J. M., Mangan, M. E., Perreault-Micale, C., Lathe, S., Sirohi, N., and Lathe, W. C. (2010). OpenHelix: Bioinformatics Education outside of a Different Box. *Brief Bioinform* 11, 598–609. doi:10.1093/bib/bbq026

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

Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol. Biol. Evol.* 24, 1586–1591. doi:10.1093/molbev/msm088

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Education

ORIGINAL RESEARCH
published: 29 October 2021
doi: 10.3389/feduc.2021.711535

Bioinformatic Teaching Resources – For Educators, by Educators – Using KBase, a Free, User-Friendly, Open Source Platform

Ellen G. Dow[1], Elisha M. Wood-Charlson[1]*, Steven J. Biller[2], Timothy Paustian[3], Aaron Schirmer[4], Cody S. Sheik[5], Jason M. Whitham[6], Rose Krebs[7], Carlos C. Goller[7,8], Benjamin Allen[9], Zachary Crockett[9] and Adam P. Arkin[1,10]*

[1]Environmental Genomics and Systems Biology Division, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA, United States, [2]Department of Biological Sciences, Wellesley College, Wellesley, MA, United States, [3]Department of Bacteriology, University of Wisconsin Madison, Madison, WI, United States, [4]Department of Biology, Northeastern Illinois University, Chicago, IL, United States, [5]Department of Biology and the Large Lakes Observatory, University of Minnesota Duluth, Duluth, MN, United States, [6]Department of Plant and Microbial Biology, North Carolina State University, Raleigh, NC, United States, [7]Department of Biological Sciences, North Carolina State University, Raleigh, NC, United States, [8]Biotechnology Program (BIT), North Carolina State University, Raleigh, NC, United States, [9]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, United States, [10]Department of Bioengineering, University of California, Berkeley, Berkeley, CA, United States

OPEN ACCESS

Edited by:
Hugo Verli, Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Dhananjaya Pratap Singh, National Bureau of Agriculturally Important Microorganisms (ICAR), India
Manuel Corpas, Cambridge Precision Medicine, United Kingdom

***Correspondence:**
Ellen G. Dow
egdow@lbl.gov
Elisha M. Wood-Charlson
elishawc@lbl.gov

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *18 May 2021*
Accepted: *02 September 2021*
Published: *29 October 2021*

Citation:
Dow EG, Wood-Charlson EM, Biller SJ, Paustian T, Schirmer A, Sheik CS, Whitham JM, Krebs R, Goller CC, Allen B, Crockett Z and Arkin AP (2021) Bioinformatic Teaching Resources – For Educators, by Educators – Using KBase, a Free, User-Friendly, Open Source Platform. Front. Educ. 6:711535. doi: 10.3389/feduc.2021.711535

Over the past year, biology educators and staff at the U.S. Department of Energy Systems Biology Knowledgebase (KBase) initiated a collaborative effort to develop a curriculum for bioinformatics education. KBase is a free web-based platform where anyone can conduct sophisticated and reproducible bioinformatic analyses *via* a graphical user interface. Here, we demonstrate the utility of KBase as a platform for bioinformatics education, and present a set of modular, adaptable, and customizable instructional units for teaching concepts in Genomics, Metagenomics, Pangenomics, and Phylogenetics. Each module contains teaching resources, publicly available data, analysis tools, and Markdown capability, enabling instructors to modify the lesson as appropriate for their specific course. We present initial student survey data on the effectiveness of using KBase for teaching bioinformatic concepts, provide an example case study, and detail the utility of the platform from an instructor's perspective. Even as in-person teaching returns, KBase will continue to work with instructors, supporting the development of new active learning curriculum modules. For anyone utilizing the platform, the growing KBase Educators Organization provides an educators network, accompanied by community-sourced guidelines, instructional templates, and peer support, for instructors wishing to use KBase within a classroom at any educational level–whether virtual or in-person.

Keywords: computational biology, bioinformatics, data science, STEM education and learning, undergraduate education

INTRODUCTION

Modern biology is becoming more reliant on "big data" to answer a range of questions. The ability to generate or re-analyze large data sets is quickly becoming a mainstay in many areas, from cellular biology to ecosystem ecology. This is especially true for the field of genomics and molecular biology, where large data sets are now commonplace. With the rapid growth and free online availability of biological data derived from DNA sequencing technologies, the need for skilled researchers to analyze these data is growing as well (Kodama et al., 2012; Koboldt et al., 2013). Topics in bioinformatics have therefore become a key feature of biology curriculum development in recent years (Maloney et al., 2010; Machluf et al., 2017). When the 2020 global pandemic drastically increased the need for virtual, learn-from-home coursework, bioinformatics also provided an attractive option for hands-on experience outside of a traditional wet lab. However, bioinformatics and the techniques used to analyze data advance quickly, and it can be difficult to incorporate the most cutting edge resources into the classroom.

One of the core challenges of teaching data analysis courses is that many students may not have access to a computer outside of their school's computer labs. During the pandemic, many schools did what they could to provide computer access for students at home, but school-loaned laptops are likely unable to support many of the bioinformatics software programs necessary to run analyses locally. For those students with more sophisticated computers at home, instructors still have the challenge of helping them troubleshoot integration of bioinformatics software across a variety of operating systems (e.g., Windows, macOS, Chrome OS, or Unix-based systems) (Cummings and Temple, 2010). In addition, many bioinformatics programs or pipelines are coded in languages such as Java or R, or rely heavily on Perl (Perl Programming Language, RRID:SCR_018313), Python (Python Programming Language, RRID:SCR_008394), Ruby or Julia programming languages that require an understanding of command-line interfaces to operate (Ken Arnold et al., 2000; Wall et al., 2000; Flanagan and Matsumoto, 2008; Van Rossum and Drake, 2009; Bezanson et al., 2017; R Core Team, 2021). These tools and resources advance quickly, with new versions becoming available regularly, requiring updates or deprecation, depending on whether a different tool is adopted by the community as the new standard. Finally, some of the most powerful bioinformatic programs require vast amounts of memory or processing power, and larger data sets often require more computing power than is available on a personal computer. High-performance computing (HPC) clusters are often available at large research centers, which may offer these resources to students, but are relatively uncommon at teaching colleges or high schools. While cloud resources, such as Azure through Microsoft or AWS through Amazon, are more readily available, the monetary costs and time and resources to set up the workflows may still impede adoption by schools and institutions (Raj et al., 2020).

To date, acquiring a depth of knowledge in bioinformatics has largely been contingent on prerequisite computational knowledge about scripting, programming, data typology, and database management, alongside statistics and of course biology (Wilson Sayres et al., 2018). Although fundamental to research development work in bioinformatics, students and faculty are often missing one or more of these skill sets. From the lack of training and access to adequate computing resources for faculty, to the constantly changing software ecosystem, these problems can seem insurmountable for teaching and create significant challenges to the effective instruction of practical data science knowledge (Cummings and Temple, 2010; Williams et al., 2019).

For students and educators looking to improve their command of the command line, the Bioinformatics Virtual Coordination Network provides peer-support and resources (BVCN, https://biovcnet.github.io), or they can turn to training portals like the Global Organisation for Bioinformatics Learning (Corpas et al., 2015). For everyone else, online or graphic user interfaces (GUI) platforms [i.e., KBase, Nephele (Nephele, RRID:SCR_016595], Galaxy (Galaxy, RRID:SCR_006281), PATRIC (Pathosystems Resource Integration Center, RRID:SCR_004154), CLC Genomics (CLC Genomics Workbench, RRID:SCR_011853), or Geneious (Geneious, RRID:SCR_010519) may address some of these challenges by facilitating greater access to bioinformatic tools and data (Afgan et al., 2018; Arkin et al., 2018; Weber et al., 2018; Davis et al., 2019; Kearse et al., 2012; QIAGEN CLC Workbench, 2013). While some of these services are provided for free, several require the purchase of a software license, which can be expensive for faculty and students already struggling with the costs of education. Thus, as the community seeks to bring modern bioinformatic approaches to students in the classroom, the community should leverage resources that address many of the barriers to entry and strive to empower students.

The U.S. Department of Energy (DOE) Systems Biology Knowledgebase (KBase) is a free knowledge creation and discovery environment designed for both biologists and bioinformaticians (Arkin et al., 2018). KBase integrates a variety of data and analysis tools, from DOE and other public services, into an easy-to-use platform that employs scalable computing infrastructure to perform sophisticated systems biology analyses. KBase uses a dynamic GUI based on the Jupyter Notebooks framework (Kluyver et al., 2016) to provide documentable, shareable, and reproducible analysis workflows called Narratives. This integration allows for hosting code and programs outside of a traditional command-line interface. Data brought into KBase are transformed into a unique data model, with specifications for each type of biological data (reads, genomes, metabolic models, etc.), structured around the FAIR–findable, accessible, interoperable, and reusable–data principles (Wilkinson et al., 2016) to facilitate discovery, access, and interoperability throughout the system. Data from different sources or file formats can be used interchangeably and downloaded in standard formats, and data workflows can be directly copied to reproduce the data provenance. All the data, tools, commentary, and code used for analysis are stored as KBase Narratives on the KBase servers, for easy, persistent access with minimal requirements for the user (more information at: https://docs.kbase.us/getting-started/browsers).

While many biologists have taken advantage of KBase for academic and industry-related research, there is a growing cohort

of educators using KBase in the classroom to demonstrate fundamental content knowledge for bioinformatics. Narratives are well-suited to use in educational settings as they can be shared with other individuals, groups, or made public. Educators can leverage Narrative features to re-run workflows, modify a workflow with new data or tools, extend workflows to incorporate new analyses, and add contextual information ranging from simple text to embedded videos and images, using Markdown or HyperText Markup Language (HTML). The visual display of provenance for data objects allows instructors to see connections between data, which can be used to supplement lectures describing the workflows. For example, due to the immediate relevance of the SARS-CoV-2 pandemic, one instructor created a Narrative to show students how to obtain sequencing data from public repositories, trim the reads, assess quality, and align to a reference genome (Pasqualoni, 2020).

Over the past year, KBase staff have worked with educators to identify how the platform could support their classroom goals. The need for tools and resources that enable virtual, interactive bioinformatics courses was a common request from each instructor. Therefore, as a community, we developed a modular curriculum framework with a variety of resources to empower students (and educators) to grow their bioinformatics content knowledge across a range of instructional levels, including high school, undergraduate, and graduate courses.

MATERIALS AND METHODS

Creating FAIR, Reproducible, and Shareable Community Resources

The KBase Narrative platform was used to modularize several complete and complex bioinformatic workflows (Dow et al., 2021). Each larger concept was divided into smaller parts to ensure that students could understand the scalability of the content while retaining the ability to answer biological questions. Working groups of instructors and the KBase team developed example teaching workflows by assembling a set of Narratives for a variety of bioinformatics concepts: genome assembly, annotation, and analysis; metagenomics; phylogenetics; and metabolic pathway analysis. The Narratives include a range of the tools available within KBase that are wrapped as applications, referred to as KBase Apps. Each workflow was centered around creating engaging content that met core curriculum concepts using backward design to address learning goals within individual modules that connected to form the complete concept (McTighe and Wiggins, 1998). Alongside FAIR data principles, these Narratives can be used directly as teaching tools, be modified by individual instructors to fit their student population and class focus, or simply serve as inspiration.

The resources developed over the summer were used in a pilot program during Fall 2020 to gather and implement instructor and student feedback, and to develop supporting guidelines that included best practices for educators (https://doi.org/10.25982/1668075). Program resources, including modularized workflows, guiding documentation, the KBase Educators Organization, and KBase User Slack workspace for support and networking were made available in December of 2020 for all educators interested in using KBase and joining the community (see KBase for Educators Webinar: https://youtu.be/K9FxPc_2jzI).

Assessing Resources During the Pilot

To evaluate the use of KBase in the classroom, a subset of the KBase Educators Working Group developed a post-course student survey (**Supplementary Materials**). The survey was a student self-assessment that focused on how effective the piloted KBase learning modules were at achieving common learning objectives, the approachability of the KBase platform, and changes in student perceptions and self-confidence in the field of computational biology after completing the modules. The survey used a Likert scale to quantify student responses through self-assessment on awareness of content and growth post-course, confidence around concepts and using tools, and using KBase. The survey study was reviewed and approved by the Human Subjects Committee Institutional Review Board at Lawrence Berkeley National Lab (LBNL) (339NR001-2AP22).

The survey data was used to create frequency distributions of student responses using histograms to gauge student reception for introducing KBase into curriculum and how students measured their own self-efficacy. Likert scale data was converted to numerical equivalents to measure distribution of the data and run Wilcoxon signed rank test with continuity correction between student responses using R v3.1 (R Project for Statistical Computing, RRID:SCR_001905) (R Core Team, 2021).

Members of the KBase Educators Working Group also informally assessed how effective KBase was at addressing the challenges of teaching data science in a virtual classroom mentioned in the introduction, ranging from totally solved (5) to not solved at all (1).

A Case Study: KBase Utilization in the NC State Metagenomics Course

An example of tailoring these resources to a course was provided by the Biotechnology Program (BIT) 477/577 Metagenomics class at North Carolina State University (NC State), which adapted the Metagenomics modules workflow along with additional analyses using Phylogenetics and Metabolic Modeling tools. The goal of these activities was to provide hands-on experience with computational tools to study relationships between microbial communities and ecosystems.

Fall 2020 was the first semester BIT 477/577 students used KBase. Students also had weekly topic modules with graded activities in Nephele (Nephele, RRID:SCR_016595) (Weber et al., 2018) and QIIME2 (QIIME, RRID:SCR_008249) (Caporaso et al., 2010; Bolyen et al., 2019), as well as social annotation and collaborative note-taking assignments with Hypothes.is (Goller et al., 2021), video lectures, interviews of bioinformaticians, creation of podcasts, and a group data analysis project. The course was taught asynchronously online due to the pandemic.

Two KBase Narratives were created for this course, Gold (https://doi.org/10.25982/67335.259/1773074) and Silver

TABLE 1 | Module Learning Objectives (MO) and instructional materials (resources linked within text).

Learning Objectives	Instructional Materials
MO 2.1. Perform quality control of read libraries	1. Background reading and discussion, "Purifying the Impure: Sequencing Metagenomes and Metatranscriptomes from Complex Animal-associated Samples." (Lim et al., 2014)(MO 2.1–2.6)
MO 2.2. Predict taxonomic population structure of environmental shotgun reads	2. Module 2. Video Overview: What are MAGs? (3:54 min) (MO 2.1, 2.2)
MO 2.3. Identify high-quality bins to extract and annotate	3. KBase assignment videos (MO 2.1–2.6)
MO 2.4. Annotate genes of extracted MAGs	a. Silver Narrative Part 1 (6:06 min) and Part 2 (10:39 min)
	b. Gold Narrative (9:42 min)
MO 2.5. Place MAGs into a species tree	4. KBase Silver Narrative (MO 2.1–2.6)
MO 2.6. Build metabolic models	5. KBase Gold Narrative (MO 2.1–2.5)

(https://doi.org/10.25982/68579.143/1766297), based on frameworks from the KBase Educators Organization and adapted for specific research and course objectives. The Gold Narrative focused on genome assembly and evaluation, through an exploration of available high-throughput metagenomic sequencing data sets, to identify *Delftia acidovorans* sequences (a bacterium capable of precipitating liquid gold into nanoparticles). In the Silver Narrative, students binned a high-quality metagenome-assembled genome (MAG) from a Silvergrass hybrid soil metagenome, determined its near relatives with a phylogenomic tree, and generated metabolic models to identify the media components required to isolate this potentially important microbiome member. This provided students with the opportunity to learn skills often used by applied scientists leveraging metagenomics. The concepts of MAGs, binning, and metabolic modeling were new to the course. Therefore, a week-long module was designed to introduce key definitions and procedures through videos, a reading assignment, and slides before engaging students in KBase activities. The KBase module addressed the learning objectives listed in **Table 1**.

Prior to the beginning of the course, students were sent information about the study and the approved informed consent form. Thirteen of the 14 students agreed to participate in the study, including five undergraduates and eight graduate students from various programs. The study was approved by the NC State Institutional Review Board (IRB, # 20309). Pre- and post-course quizzes were used to assess the students' familiarity with various general concepts and ideas in bioinformatics prior to using the Narratives. The results of the pre- and post-course knowledge surveys were analyzed using Fisher's exact test and *t*-test to compare averages and variance. An open response section afforded students the opportunity to provide anonymous feedback on aspects of their experience, such as where they had the most difficulty, providing a basis for determining where future improvements to lessons could be most effectively targeted. Quiz questions are included in the **Supplementary Data Sheet 1**.

Building a Community of Practice to Address Challenges

The KBase Educators program was initiated to gather information from educators teaching bioinformatics on how KBase could support the community of users with the pandemic shifting classrooms from in-person to online. KBase advertised an open call to instructors to hold a discussion of their needs and speculate around how KBase, as a research platform, could be applied to course curriculum. In late Spring 2020, a kick-off community meeting was held for instructors that expressed the shared need for tools and resources that enable interactive bioinformatic experiences and transferable analytical skills. From the initial conversation, about a dozen educators began the KBase Educators Working Group to develop workflows that connect across basic bioinformatic concepts. Working Group members with classes in the Fall of 2020 participated in the pilot program, which prompted the creation of best practices and supporting guidelines for the community as it grows. Alongside guidelines, KBase implemented infrastructure to support members and promote collaboration and networking within the broader KBase Educator community. Many of the Working Group members co-authored this manuscript.

RESULTS

Resources to Support Educators Using KBase

The KBase teaching Narratives are organized as a series of interconnected workflows, illustrated by the Educators Concept Overview (https://doi.org/10.25982/1668075). This Narrative provides a centralized landing page for the modules and allows educators and students to choose the analysis pipelines they wish to perform. It also contains links to resources and suggestions for teaching fundamental concepts, and is easily extensible so educators can collect and share example data with others in the community or expand existing analyses. The overview Narrative contains a flow diagram that shows the connections between concepts and modules with links to each module (**Figure 1**). The concepts include broad bioinformatics topics: Genomics, Metagenomics, Phylogenetics, Pangenomes, and Metabolic Modeling (**Table 2**, with complete overview in **Supplementary Table S1**). For each concept there are one or more modules to demonstrate the roadmap of analysis within that concept. From there, educators can copy any of the module Narratives and modify the learning concepts and example questions, populate the Narrative with new data related to a topic of interest to the class, or add text to create an explanatory Narrative.

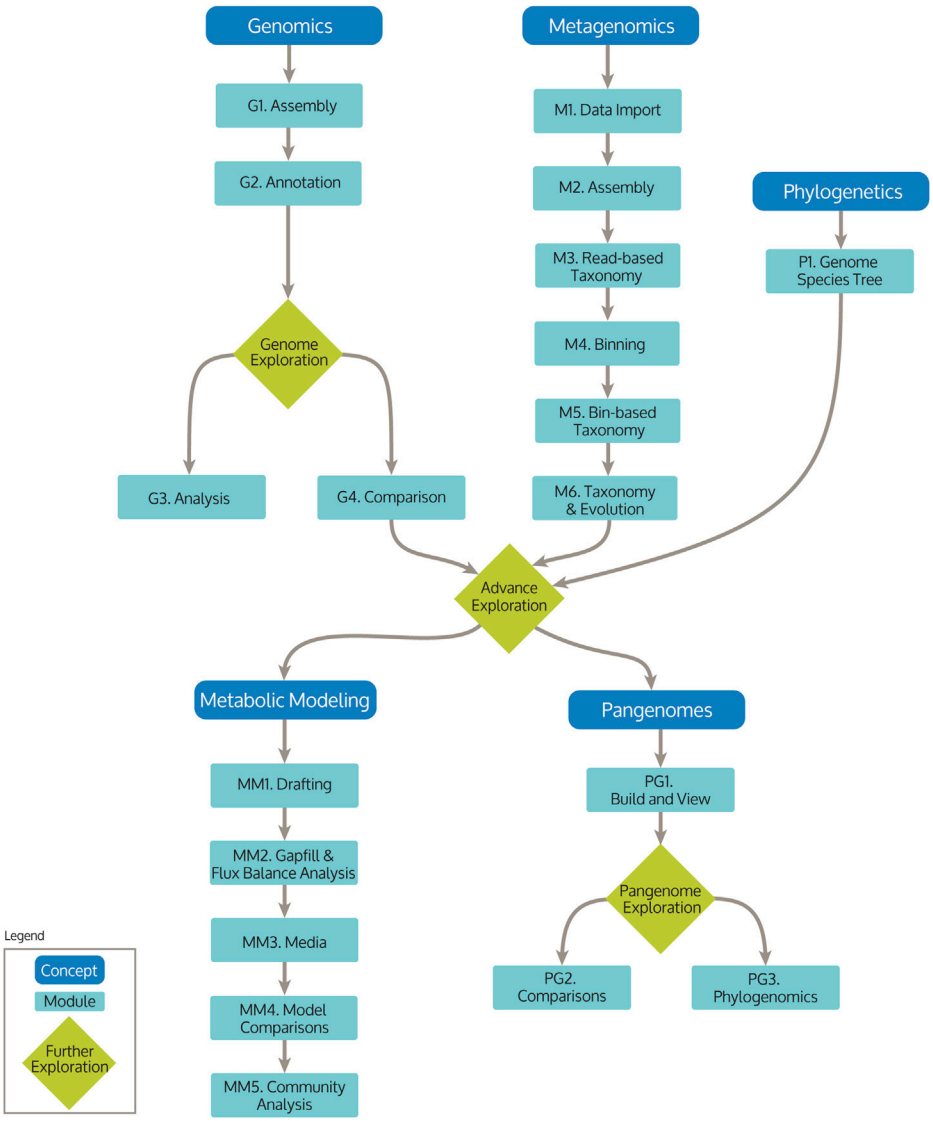


FIGURE 1 | KBase Educators Concept Workflows and their modules, including Genomics, Metagenomics, Phylogenetics, Pangenomes, and Metabolic Modeling.

Narratives are connected to each other through HTML or Markdown hyperlinks (see resources linked in Dow et al., 2021) so that related workflows can be separated into modular components. By breaking complex analysis pipelines into smaller, distinct units, this structure allows students to focus on understanding the inputs, outputs, and rationale behind each step in the process. This also enables educators to more easily pick and choose the aspects most relevant to their particular course learning goals, or to fit analyses into particular timeframes, while ensuring students understand how each step connects to the larger concept and analytical goals.

The primary tutorial Narratives are designed for instructors and contain teaching notes that include explanations and alternatives or suggestions. For each tutorial Narrative there is a paired student version, which maintains the framework of the original Narrative in nascent App workflows without data products or results. Questions are embedded within the Narratives for students to answer as they explore the workflow and data, prompting them to perform the analysis on their own. Narratives can be shared with data and analysis results preloaded, inviting students to focus on interpreting results, or empty of data and analyses, allowing students to explore data and workflow options on their own. Student versions may also provide additional resources, such as links to literature, data sources, and guides to using the tools. To create new student Narratives, educators simply copy an existing Narrative and edit the data and Apps included within the Narrative. Because of this feature, educators can place students at the beginning of a particular workflow or at any point within it, and have the ability to combine workflows for students to follow longer analyses.

TABLE 2 | Example Concept Workflow in KBase using the Narrative on Metagenomics, including module topics and corresponding Learning Objectives.

Concept	Module	Learning Objectives
Metagenomics	**M1.** Data Import and QC	**M1.1** Evaluate read quality based on FastQC reports
		M1.2 Perform read trimming with Trimmomatic
		M1.3 Explain in your own words how adapter contamination can
		M1.3a affect read quality
		M1.3b be addressed with Trimmomatic
	M2. Assembly	**M2.1** Define assembly and contig
		M2.2 Compare and contrast the output of different assemblers
	M3. Read-based Taxonomy	**M3.1** Understand the mechanisms behind taxonomic profiling
		M3.2 Compare the outputs of two applications
	M4. Binning	**M4.1** Understand how binning works to group contigs of metagenomic assemblies
	M5. Bin-based Taxonomy	**M5.1** Explain another method of how to measure taxonomic diversity through binned assemblies
	M6. Taxonomy and Evolution	**M6.1** Understand how MAGs generated in a metagenomic study can be placed into a phylogenetic tree
		M6.2 Understand how ANI can be used to determine relatedness
		M6.3 Understand how to estimate relative abundance of a MAG within a metagenomic sample
		M6.4 Learn how to use apps and tools within KBase to build upon foundational concepts and promote exploration

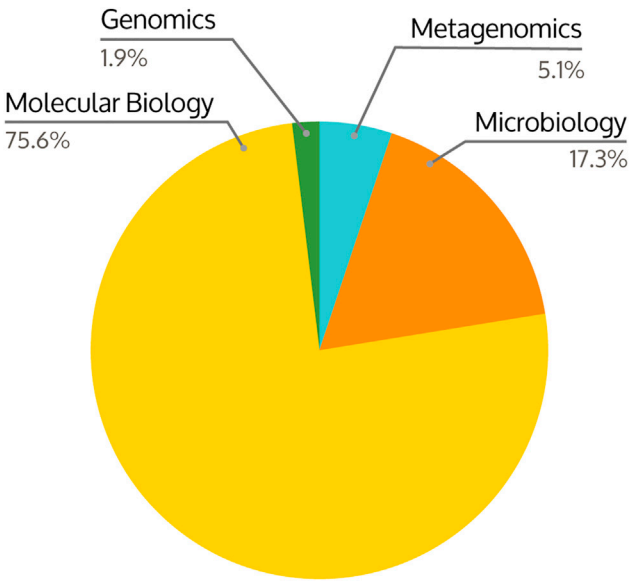


FIGURE 2 | Distribution of undergraduate course subjects using KBase within the curriculum across 156 responses. The courses included Genomics, Metagenomics, Microbiology, and Molecular Biology.

Assessment of Resources During the KBase Educators Pilot (Fall 2020-Winter/Spring 2021)

Instructors from the Working Group implemented KBase resources in a variety of settings and educational contexts during the 2020–2021 academic year. Contributors represented about a dozen institutions, with six institutions piloting KBase in the classroom in 2020–2021. For example, students at Boca Raton Community High School (Boca Raton, FL), Northeastern Illinois University (Chicago, IL), and graduate students participating in the 2020 Multiscale Microbial Dynamics Modeling course (hosted by the Environmental Molecular Sciences Laboratory (EMSL) and Pacific Northwest National Laboratory (PNNL) Subsurface Biogeochemical Research group) all simultaneously used KBase to reach different learning objectives and educational goals. At Boca Raton Community High School, students used KBase as a "one-stop shop" to explore genomic data. Undergraduate Genomic and Proteomics students at Northeastern Illinois University (Chicago, IL) used KBase as part of a flexible set of semester-long assignments focused on using real-world data to learn about genome assembly, annotation, and comparison. Graduate students and early career researchers participating in the 2020 Multiscale Microbial Dynamics Modeling course (https://www.kbase.us/multiscale-microbial-dynamics-modeling/) used newly implemented Apps, such as DRAM (Shaffer et al., 2020), in KBase for metagenomic data processing and analysis. This range of contexts and approaches highlights the scalability, accessibility, and versatility of the KBase platform.

Several undergraduate courses participated in a student survey to assess student perception of learning with KBase. So far, the survey has received 156 responses (excluding one response that only provided the course name) from students in courses covering Genomics, Metagenomics, Microbiology, and Molecular Biology (**Figure 2**).

Students were asked to assess their own experience and knowledge after the course. Coming into the course, 87% of students reported that the concepts covered were "extremely or moderately new" to them (**Figure 3A**). After the course, familiarity with bioinformatics concepts and workflows improved for around 96% of responses, with the majority of students responding that they now felt "somewhat or moderately familiar" with the concepts covered (**Figure 3B**). The change in student perception of course knowledge from pre to post-course was significant (**Figures 3A,B**; Wilcoxon test $p < 0.001$). At the end of their courses, approximately 39% of students felt they were able to describe the steps of a bioinformatics workflow, 34% were uncertain, and 27% did not feel that they were able. However, students' confidence in their ability to use bioinformatics analysis and tools to answer research questions improved in 68% of responses (**Figure 3C**). Approximately 90% of students described at least some confidence in their ability to continue to use KBase to analyze data and find answers (**Figure 4A**), with the majority of students agreeing that overall

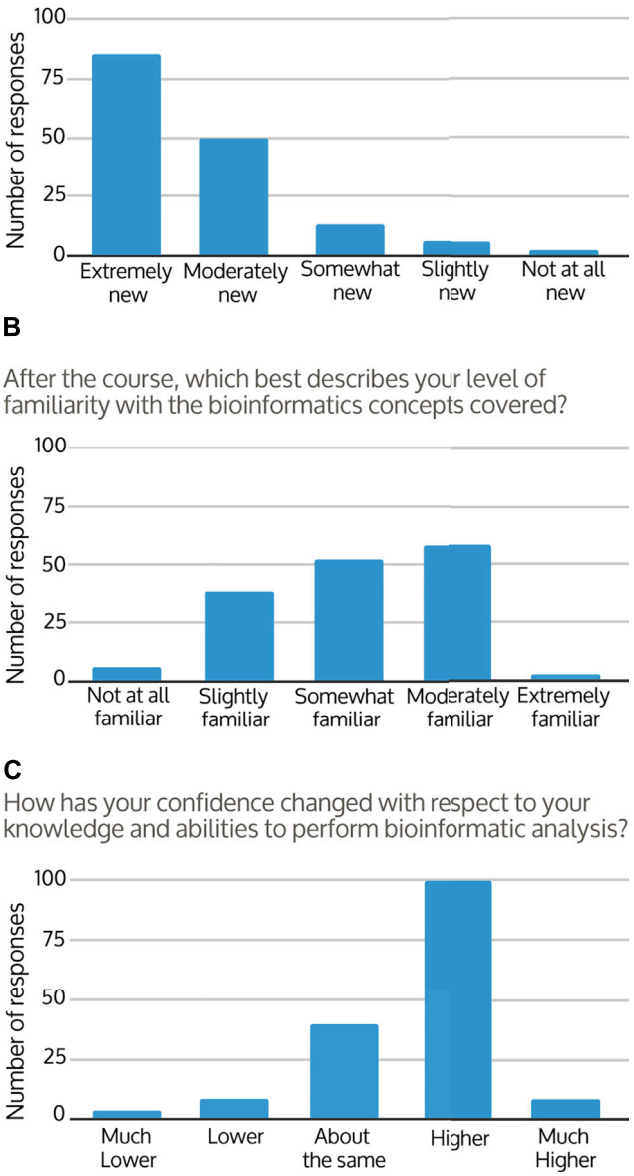


FIGURE 3 | Frequency distributions of student responses to survey questions comparing **(A)** pre-course to **(B)** post-course knowledge and **(C)** self-assessed confidence. There were 155–157 responses per question as each question was optional for students to answer.

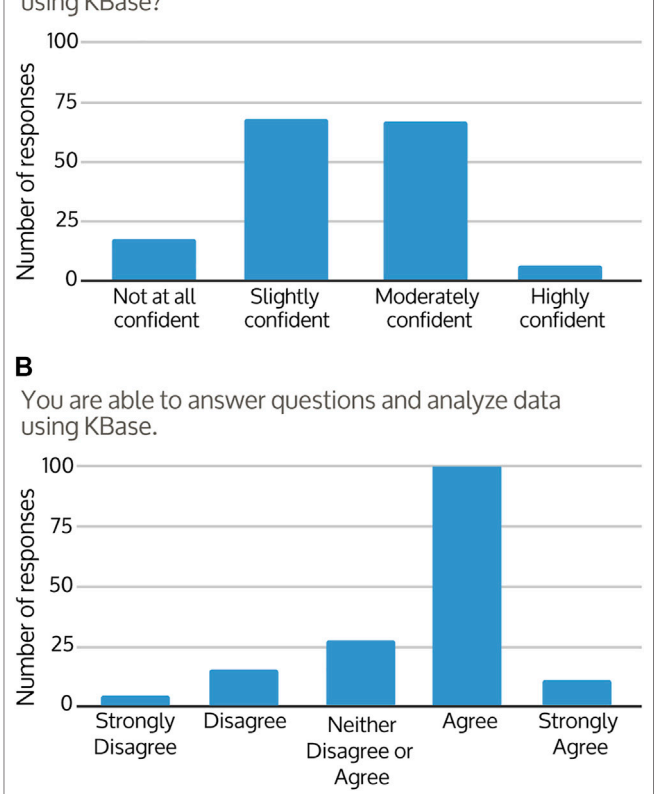


FIGURE 4 | Frequency distributions of 156 student responses to their **(A)** confidence and **(B)** ability to use KBase post-course.

they were able to use KBase effectively (**Figure 4B**). Student responses describing their change in confidence reflected a similar pattern to their perceived ability to use KBase (**Figure 3C**, **Figure 4B**; Wilcoxon test $p = 0.649$). Overall, approximately 30% of the students reported that KBase was easy to use, 41% had no strong opinion, and 30% had some difficulty using KBase.

From an Educator's perspective ($n = 6$, working group members), KBase was able to address many of the challenges facing data science in a virtual classroom for undergraduate courses. During the pilot, the platform and available resources were mostly able to address challenges (5 = totally solved, 1 = not solved at all) related to: working across different operating systems (4.3), access to and accessibility of consistent bioinformatic tools and resources (4.8), access to computational resources (4.5), and providing bioinformatics options for students without coding experience (5). Success at granting students access to computers was variable and likely dependent on pre-existing circumstances. Specifically, the education-based co-authors appreciated the ability to easily share Narratives with students; the option to gradually reduce the scaffolding as student knowledge of the platform, data, and analyses increased; and how the platform allowed students to focus on why they were learning bioinformatic concepts, instead of spending time on setup and maintenance of the tools and resources.

Results of North Carolina State University Biotechnology Course Case Study

At the beginning of the metagenomics course, students ($n = 13$) had different knowledge gaps related to the instructional concepts. For example, **Figure 5A** shows students 2, 7, and 11 each had three incorrect answers in the pre-course quiz, but only one of those

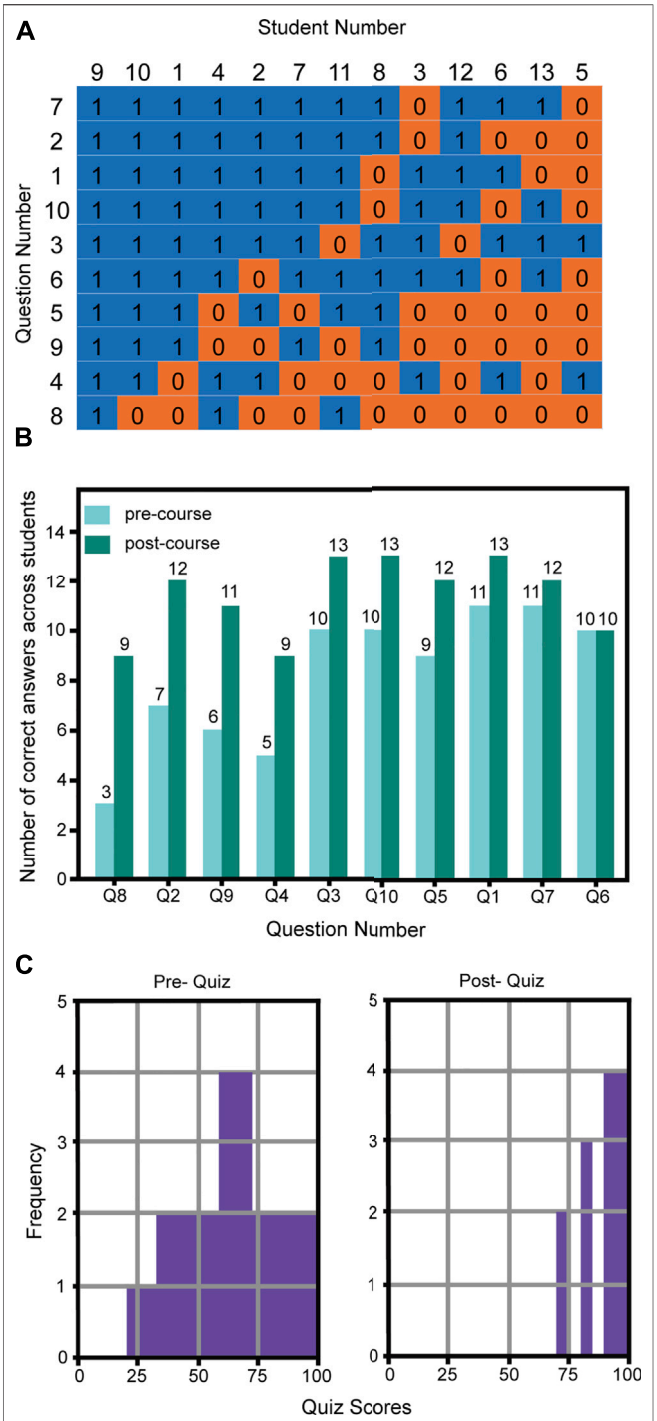


FIGURE 5 | Pre- and post-course knowledge survey of 13 students in NC State's Biotechnology (BIT) 477/577 Fall 2020 course. **(A)** Pre-quiz scores arranged by students and questions to highlight diversity in pre-knowledge. Responses are sorted to visually show pattern and clustering of correct and incorrect responses. 1 = correct, 0 = incorrect responses, n = 13 students. **(B)** Correct and incorrect pre- and post-course quiz questions. Questions can be found in the Supplemental Materials. Questions were sorted to show greatest improvement in correct answers. **(C)** Pre- and post-course quiz grade distributions. Student scores transitioned from nearly normal with a wide variance in the pre-quiz to highly skewed left with low variance in post-quiz.

overlapped. Likewise, students 3 and 8 had five and four incorrect answers, respectively, and only one overlapped. Knowledge gaps were mostly filled by the end of the course based on comparison of pre- and post-course question scores (**Figure 5B**). The highest number of students that answered correctly on any one question was 11 out of 13 on the pre-test. In contrast, three questions were 100% correctly answered in the post-test. Correct answers increased by as many as six students in the post-test for some questions. On a per question basis, there was a higher frequency of pre/post-quiz questions incorrectly then correctly answered (36) compared to incorrect/incorrect (12) and correct/incorrect (4) answers. A significant imbalance of these ratios (Fisher's test, p-value = 0.0024) supports student learning during the course. Furthermore, statistical differences in pre and post-course quiz averages and variance (**Figure 5C**, t-test, p-value = 0.0004, F-test, p-value = 0.0091) indicated significant improvement in students' understanding of the concepts. While sample size was limited during this pilot study, student responses and feedback were encouraging.

At the end of the course, students completed their own projects by applying data analysis methods to another data set. Students chose between KBase or QIIME2 (run on a local HPC cluster) to complete their final project. Of the 13 students, seven chose to complete their project through KBase, some actually expressing an aversion to using the command-line and terminal interfaces during office hours. Some students explained that they would have chosen the KBase option for their data analysis project if they had managed their time better. The KBase assignment would take more time to complete because it was more open-ended, and involved analyzing a set of data in a novel way, whereas the QIIME assignment had specific questions to answer and a prepared script that needed minimal additions from the student to work properly.

Open-ended student responses indicated new knowledge and understanding of computational pipelines for metagenomics analysis (**Table 3**). Overall, testimonials were telling about the impact of KBase. Six out of 13 students used the word "comfort" in describing their familiarity with KBase and other tools taught in the course (**Table 3**).

Crowd-Sourced Suggestions and Guidelines for Utilizing KBase Narratives in the Classroom

To facilitate the use of KBase Narratives, the Working Group developed a set of guidelines and resources associated with the Educators organization. An educator's "Standard Operating Procedures" document (SOP) includes crowd-sourced suggestions for best practices and detailed logistics for using KBase Narratives in the classroom setting. This guide includes suggestions for organizing classes within KBase, guidelines for registering student accounts, and other key details. Guidance is also provided on estimated time requirements for running different modules with students at different educational levels to help educators plan their initial rollout. These represent "living documents" which can be continually improved with feedback from the community of users.

A number of open resources are provided that educators can adopt directly to their curriculum, reducing the barriers associated with incorporating new material into a class.

TABLE 3 | Select student responses to the question "List and describe your familiarity with computational pipelines for metagenomic analyses."

Pre-course	Post-course
I have used Galaxy and other genome browsers	All of the ones used in the course (KBase, Nephele, HPC) would be comfortable using on my own with my own data set.
I took BIT 495 (RNA World) with Dr. [edit] in the beginning of this semester, so we just used a Shape-MaP computational pipeline to study reactivities and secondary structure of RNA molecules. Other than that, though, it's still pretty new ground for me	I became much more comfortable with KBase during the data analysis project, and began to better understand using the HPC and Nephele during the course
I am not familiar with using pipelines for metagenomic analyses and I chose this course because I am really interested to learn new computational tools. I am only familiar with analyzing genome sequences from pure isolates (assembly, annotation)	I feel comfortable with using KBase and Nephele. I need to work more on my own on analysis with QIIME2
Maybe a little experience with MEGA from 6 + years ago, but I can't remember them	I'm mildly familiar with KBase and Nephele now
I've used DADA2 in R and danced around using qiime but haven't actually done it	KBase (comfortable), Nephele (comfortable), QIIME/HPC (not comfortable), R packages (comfortable)
No experience	Nephele- basic understanding, KBase- intermediate understanding, HPC- basic understanding
I am not at all familiar with computational pipelines	This class has helped me learn about KBase, Nephele, and HPC which I was completely unfamiliar with before

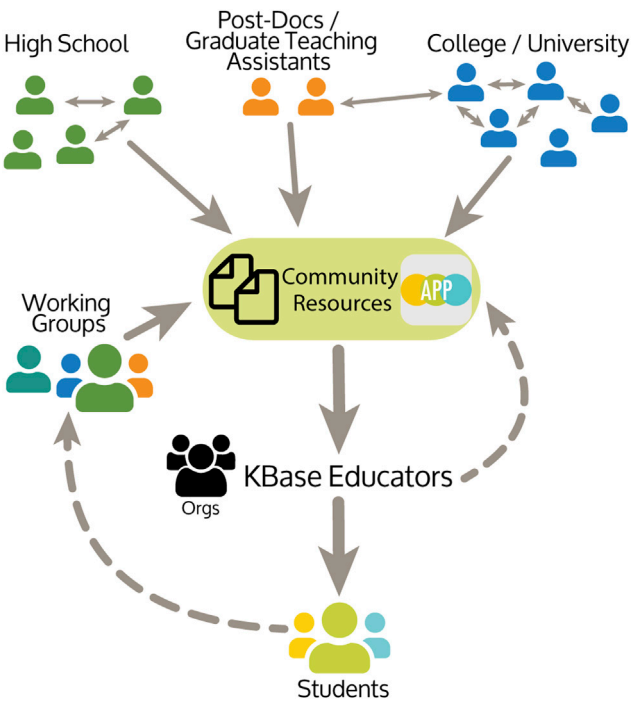


FIGURE 6 | KBase Educators Community: The community includes instructors that teach at High School and college levels. Resources are developed by the community, including teaching Narratives using bioinformatics tools wrapped as KBase Apps and supporting documentation. Working groups have developed workflows across common topics. From student and community feedback, resources including teaching Narratives, are improved and added to the KBase Educators Organization.

Examples include links to videos and specific step-by-step instructions, which can be provided to students to support their use of KBase. Guidance within the SOP describes common problems encountered in creating, modifying, and using Narratives including links to the KBase documentation site (http://docs.kbase.us). Some Narratives have associated introductory slides which can be used directly or customized by an instructor to introduce the concepts and tools utilized. Examples of student assessment questions, along with model answers, are also available. Educators interested in sharing additional or modified materials associated with the use of these Narratives in different classroom settings are encouraged to contribute these within the KBase Educators Organization, described below.

Providing Support, Networking, and Collaboration *via* Slack

The KBase Educators Organization is a community of educators, representing diverse types of educational institutions and levels, who are interested in using KBase to teach bioinformatics concepts (**Figure 6**). To provide a common location where educational Narratives could be freely shared, the "Organization" framework within KBase was used to create a central homepage for the KBase Educators Organization. This Educators Organization represents a space where conceptual tutorial Narratives, student Narratives, or other resources can be shared in a moderated fashion. The Educators Organization homepage (https://narrative.kbase.us/#/orgs/kbase-educators) has open resources from educator guides to using the Narrative and Markdown language to instructional Narratives that can be accessed by anyone, but contributions to more detailed resources come from members of the community.

KBase Organizations can also be created for individual classes. Educators can provide both informative and evaluative Narratives to their own organization, which students can copy, and then add their assigned projects to the organization or share with their instructor. Access to these organizations can be restricted, or not, as any instructor chooses.

The KBase team also supports educators and scientists to create working groups focused on specific topics. For example, the KBase Educators Working Group developed the existing resources (**Figure 6**). A subset of the Working Group recently

developed a workflow for Metabolic Modeling (**Supplementary Table S1**), and other members have an interest in working on RNAseq. These collaborations are invaluable at providing ideas, exploring questions that arise during the implementation, and novel ways to use KBase Apps and Narratives.

Communication between instructors and KBase staff is a central feature of the KBase Educators Organization. To facilitate these interactions, KBase hosts a Slack workspace open to KBase users which includes a specific channel for education purposes. This Slack workspace is monitored by KBase staff and represents an important avenue for users to obtain technical support with any aspect of the KBase website, Narrative questions, or issues with specific tools. Importantly, it also serves as a venue for educators to seek advice from other educators and to serve as a meeting space for individuals with shared interests, facilitating collaborative development of new teaching modules.

DISCUSSION

KBase provides a free, streamlined entry point for students (and educators) to explore and perform sophisticated bioinformatics analyses (Dow, et al., 2021). The KBase Narrative interface allows instructors to break down complex real-world data sets and create analyses that are approachable and easily digested by students. From the students' perspective, their familiarity with the concepts and confidence increased from the start of the courses surveyed. Most students began the course with little to no knowledge of the bioinformatics concepts covered. By the end, students gained confidence and familiarity with bioinformatics workflows overall, and in their ability to do bioinformatics analysis within KBase. By lowering many of the barriers students face when getting started in bioinformatics, it is possible to successfully increase students' confidence in bioinformatics within a single course offering. Overall, students using KBase appreciated the workflow architecture and visualizations available within the Narrative. They especially appreciated being able to complete an entire sequencing analysis process in a relatively short amount of time. Beyond the ease of using a streamlined GUI (as compared to command line), students were able to perform research using the same tools as world-class researchers in a framework that aligns the learning process with applicable experience to empower students to get excited about their own education. The web-based nature of KBase also reduced the "barriers to entry" for instructors by minimizing the time spent on setup and maintenance of resources.

KBase provides a standardized framework for the integration of tools, resulting in a modular, yet interoperable and easy-to-follow interface that connects analyses and data to results and visualizations. KBase, and the network of community developers that contribute to the platform, continually strive to update software and Apps, incorporate new resources as they are developed, and ensure computational requirements are met for each App. The KBase servers include over 6,500 total cores, which run over 660,000 central processing unit hours per month, and over 400 terabytes of storage (and growing). Students only need a computer that can run a browser and an internet connection to access the resources behind KBase. This democratization of hardware enables a greater diversity of students to explore data science and bioinformatics, to run computationally demanding programs, and to open up much deeper analysis questions than previously possible.

In the case study at NC State, KBase was introduced as a course module component, as well as an option for the students' final data analysis project. One key difference with using KBase for the course was that the Narratives enabled text and pictures to be incorporated within the analysis workflow. Students were introduced to Narratives through instructor-created videos, and then asked to copy the example Narratives for their own exploration. Embedded challenge questions allowed students to investigate publicly available data and encouraged them to be creative, while still providing scaffolding and structure via tutorial Narratives. Future course adaptations and modifications could expand the assignment by including similar individual or group challenges that require students to develop their own Narrative workflows to demonstrate an appropriate application and sequence of tools.

KBase was a favored tool for case study students. Where tools were mentioned in the post-course open-response questions, students most frequently referenced KBase by name, often first within lists, and articulated their comfort using the platform for bioinformatics analyses. Students indicated more confidence with using KBase over other tools used in the course during office hours. Notably, KBase was chosen for final projects more frequently than the HPC option by students with less coding experience.

The pivot towards confidence and comfort when using data science tools was an encouraging discovery. The perception of coding or using a computer for data analysis outside of basic spreadsheets may elicit feelings of inadequacy in the students (Cline and Prokop, 2019). These deep-seated feelings could translate into the students disregarding the assignment or not fully understanding the greater concepts of the exercise, so it is important to provide an accessible and inclusive framework that focuses on knowledge generation (Wright et al., 2020).

The interactive nature of KBase Narratives reduces the cognitive load placed on students learning bioinformatics, which allows students to focus on learning biology. Therefore, the emphasis shifts to biological concepts and the process of biological data analysis using a GUI, that provides exposure to bioinformatic tools without requiring students to become proficient with coding or command line tools. Students can run high-memory, computationally intensive applications on KBase that would not be possible on their own computers, or even in computer labs at many institutions. Access to institutional HPC resources that can handle applications requiring higher computational power may also be restricted to prevent large numbers of inexperienced users from degrading the performance for more senior researchers.

KBase is also versatile. The Narrative workflow format, where data and Apps can be easily added or removed, allows KBase to be implemented in a variety of educational environments ranging

from the high school general biology classroom to advanced graduate courses in bioinformatics. For example, each App has detailed explanations describing its purpose and links to research articles appropriate for advanced courses that need to explore the design and logic behind each analysis. In addition, students can view and analyze publicly available data sets, or they can conduct *de novo* analyses using novel data sets. In support of novel analyses, KBase is scalable–both in the size or number of data sets, as well as the functional areas available for instruction. This provides faculty members with the ability to incorporate a single platform at multiple levels and in multiple contexts within their curriculum. This flexibility and scalability provides the opportunity for a variety of both formative and summative assessments in the classroom and laboratory. Taken together, feedback from the pilot was extremely positive and demonstrated that KBase can be a powerful educational tool that can be easily applied to a variety of educational contexts.

Joining the Educators Organization and Community

The KBase Educators program is an on-going effort supported by the KBase team. All biological sciences instructors, with any experience level in bioinformatics, are welcome to join. Joining is easy and includes: 1) access to the KBase Educators Organization, which has a plethora of resources and pre-built teaching Narratives, and 2) an invite to the KBase Users Slack workspace, which has a dedicated channel for educators to converse, ask questions of the community, or get asynchronous support from KBase staff. To join, create a KBase user account at www.kbase.us, then submit a "request to join" to the KBase Educators Org (https://narrative.kbase.us/#org/kbase-educators). Send a follow up email to engage@kbase.us to introduce yourself. Please include what institution you are affiliated with and what courses you are hoping to teach using KBase. The KBase team will approve your join request and send you an invite to the Slack group.

One of the major benefits of joining the KBase Educators program is the community that has rallied around virtual learning using a GUI. The existing resources are free for use under a Creative Commons by 4.0 license, with attribution by citing this manuscript and (Arkin et al., 2018). As your instructional needs evolve and diversify, the teaching Narratives can be easily adapted and expanded. All work in KBase is backed by free large-scale computing resources and access to the KBase Help Desk (https://kbase-jira.atlassian.net/). KBase also supports the FAIR data principles (Wilkinson et al., 2016) by providing a mechanism to publish reproducible research. Static Narratives are snapshots of the current Narrative analysis that are made visible to anyone on the internet (including search engines) without a KBase account. If your class is working on a publication, it is also possible to request a DOI for a static Narrative(s) for inclusion in the manuscript's data availability statement and reference section.

Finally, for those of you that are feeling a bit more adventurous and want to also teach your students how to code, the Bioinformatics Virtual Coordination Network (BVCN, https://biovcnet.github.io) is another great pandemic-inspired program led by Dr. Ben Tully (Tully et al., 2021).

DATA AVAILABILITY STATEMENT

The datasets analyzed for the teaching portions of this study can be found through KBase and the publication Narrative [https://doi.org/10.25982/90997.49/1783189]. The study records and raw datasets comprising student survey data are not available to persons outside of the study team under stipulations of the survey study review and approval (LBNL HSC 339NR001-2AP22; NC State IRB, # 20309).

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Human Subjects Committee Institutional Review Board at Lawrence Berkeley National Lab (LBNL) (339NR001-2AP22) for the student survey and the NC State Institutional Review Board (IRB, # 20309) for the case study. The participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

ED and EWC conceptualized the project. ED supervised the project. ED, EWC, SB, CG, AS, and CS contributed to developing the student survey. ED contributed to the methodology and performed the statistical analysis for the student survey. RK, JW, and CG contributed to the case study methodology and wrote the case study sections. JW performed statistical analysis for the case study. ED, EWC, BA, ZC, SB, JW, CG, TP, and CS wrote the first draft of the manuscript. ED, EWC, SB, CG, AS, CS, and AA contributed to the manuscript revisions. All authors read and approved the final manuscript.

FUNDING

KBase and its related work is supported as part of the Genomic Sciences Program DOE Systems Biology Knowledgebase (KBase) funded by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research under Award Numbers DE-AC02-05CH11231, DE-AC02-06CH11357, DE-AC05-00OR22725, and DE-AC02-98CH10886.

ACKNOWLEDGMENTS

We recognize fellow Educator Working Group members Jon Benskin, Jennifer Chase, Kathleen Morrow, and Mike Shaffer for developing these resources and using them within their courses. We thank fellow Educator Org members for taking part in initial discussions and lending their insights, especially Bill

Andreopolos, Valerie de Crecy, Adam Deutschbauer, JP Dundore-Arias, Sharon Greenblum, Igor Grigoriev, Ana Juarez-Vazquez, Mark Martin, Victoria Orphan, Ben Tully, and Chris Vaglio. We thank members of the KBase team (https://www.kbase.us/team/) for their involvement and support, especially Dylan Chivian, Paramvir Dehal, Meghan Drake, Janaka Edirisinghe, José Faria, Chris Henry, Sean Jungbluth, Miriam Land, Erik Pearson, Bill Riehl, Boris Sadkhin, and Pamela Weisenhorn. Thank you to the Pacific Northwest National Lab and Environmental Molecular Sciences Laboratory User Facility for using KBase during the 2020 Multiscale Microbial Dynamics Modeling course, led by Tim Scheibe and Nancy Hess.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.711535/full#supplementary-material

REFERENCES

Afgan, E., Baker, D., Batut, B., van den Beek, M., Bouvier, D., Cech, M., et al. (2018). The Galaxy Platform for Accessible, Reproducible and Collaborative Biomedical Analyses: 2018 Update. *Nucleic Acids Res.* 46, W537–W544. doi:10.1093/nar/gky379

Arkin, A. P., Cottingham, R. W., Henry, C. S., Harris, N. L., Stevens, R. L., Maslov, S., et al. (2018). KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* 36, 566–569. doi:10.1038/nbt.4163

Arnold, Ken., Gosling, J., and Holmes, D. (2000). *The Java Programming Language*. 75 Arlington Street, Suite 300 Boston, MA United States: Addison-Wesley Longman Publishing Company, Inc.

Bezanson, J., Edelman, A., Karpinski, S., and Shah, V. B. (2017). Julia: A Fresh Approach to Numerical Computing. *SIAM Rev.* 59, 65–98. doi:10.1137/141000671

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

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

Cline, S. G., and Prokop, J. W. (2019). *Framework, Barriers, and Proposed Solutions for Engaging Students in Bioinformatics Research*. Red Hook, NY: CSREA Press, 6.

Corpas, M., Jimenez, R. C., Bongcam-Rudloff, E., Budd, A., Brazas, M. D., Fernandes, P. L., et al. (2015). The GOBLET Training portal: a Global Repository of Bioinformatics Training Materials, Courses and Trainers. *Bioinformatics* 31, 140–142. doi:10.1093/bioinformatics/btu601

Cummings, M. P., and Temple, G. G. (2010). Broader Incorporation of Bioinformatics in Education: Opportunities and Challenges. *Brief. Bioinform.* 11, 537–543. doi:10.1093/bib/bbq058

Davis, J. J., Wattam, A. R., Aziz, R. K., Brettin, T., Butler, R., Butler, R. M., et al. (2019). The PATRIC Bioinformatics Resource Center: Expanding Data and Analysis Capabilities. *Nucleic Acids Res.*. doi:10.1093/nar/gkz943

Dow, E., Wood-Charlson, E., Biller, S., Paustian, T., Schimer, A., Sheik, C., et al. (2021). Data from: Bioinformatic Teaching Resources – for Educators, by Educators – Using KBase, a Free, User-Friendly, Open Source Platform. *KBase*. doi:10.25982/90997.49/17831890

Flanagan, D., and Matsumoto, Y. (2008). The Ruby Programming Language. *O'Reilly*.

Goller, C. C., Vandegrift, M., Cross, W., and Smyth, D. S. (2021). Sharing Notes Is Encouraged: Annotating and Cocreating with Hypothes.Is and Google Docs †. *J. Microbiol. Biol. Educ.* 22. doi:10.1128/jmbe.v22i1.2135

KBase for Educators Webinar (2020). YouTube. Available at: https://youtu.be/K9FxPc_2jzI (Accessed May 18, 2021).

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., et al. (2012). Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data. *Bioinformatics* 28, 1647–1649. doi:10.1093/bioinformatics/bts199

Kluyver, T., Ragan-Kelley, B., Pérez, F., Bussonnier, M., Frederic, J., Hamrick, J., et al. (2016). Jupyter Notebooks—A Publishing Format for Reproducible Computational Workflows. In Positioning and Power in Academic Publishing: Players, Agents and Agendas Amsterdam: IOS Press, 87–90.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The Next-Generation Sequencing Revolution and its Impact on Genomics. *Cell* 155, 27–38. doi:10.1016/j.cell.2013.09.006

Kodama, Y., Shumway, M., and Leinonen, R.on behalf of the International Nucleotide Sequence Database Collaboration (2012). The Sequence Read Archive: Explosive Growth of Sequencing Data. *Nucleic Acids Res.* 40, D54–D56. doi:10.1093/nar/gkr854

Kluyver, T., Ragan-Kelley, B., Pérez, F., Bussonnier, M., Frederic, J., Hamrick, J., et al. (2016). "Jupyter Notebooks–A Publishing Format for Reproducible Computational Workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas* (IOS Press), 87–90.

Lim, Y. W., Haynes, M., Furlan, M., Robertson, C. E., Harris, J. K., and Rohwer, F. (2014). Purifying the Impure: Sequencing Metagenomes and Metatranscriptomes from Complex Animal-Associated Samples. *J. Vis. Exp.* 52117. doi:10.3791/52117

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

Maloney, M., Parker, J., LeBlanc, M., Woodard, C. T., Glackin, M., and Hanrahan, M. (2010). Bioinformatics and the Undergraduate Curriculum Essay. *CBE Life Sci. Educ.* 9, 172–174. doi:10.1187/cbe.10-03-0038

McTighe, J., and Wiggins, G. (1998). *"Backward Design," in Understanding by Design*. Alexandria, VA: Association for Supervision and Curriculum Development, 13–34.

Pasqualoni, S. (2020). 2019 Novel Coronavirus (COVID-19). *KBase*. doi:10.2172/1602724

QIAGEN CLC Workbench (2013). QIAGEN Bioinformatics Is Now QIAGEN Digital Insights. Available at: https://digitalinsights.qiagen.com/.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: https://www.R-project.org.

Raj, R. K., Romanowski, C. J., Impagliazzo, J., Aly, S. G., Becker, B. A., Chen, J., et al. (2020). High Performance Computing Education. In Proceedings of the Working Group Reports on Innovation and Technology in Computer Science Education New York, NY: Trondheim Norway: ACM, 51–74. doi:10.1145/3437800.3439203

Shaffer, M., Borton, M. A., McGivern, B. B., Zayed, A. A., La Rosa, S. L., Solden, L. M., et al. (2020). DRAM for Distilling Microbial Metabolism to Automate the Curation of Microbiome Function. *Nucleic Acids Res.* 48, 8883–8900. doi:10.1093/nar/gkaa621

Tully, B. J., Buongiorno, J., Cohen, A. B., Cram, J. A., Garber, A. I., Hu, S. K., et al. (2021). The Bioinformatics Virtual Coordination Network: An open-source and interactive learning environment. *Front. Educ.* doi:10.3389/feduc.2021.711618

Van Rossum, G., and Drake, F. L. (2009). *Python 3 Reference Manual. 100 Enterprise Way*. CA: Suite A200 Scotts ValleyCreateSpace.

Wall, L., Christiansen, T., and Orwant, J. (2000). *The Perl Programming Language*. Sebastopol, CA: O'Reilly Media, Inc.

Weber, N., Liou, D., Dommer, J., MacMenamin, P., Quiñones, M., Misner, I., et al. (2018). Nephele: a Cloud Platform for Simplified, Standardized and Reproducible Microbiome Data Analysis. *Bioinformatics* 34, 1411–1413. doi:10.1093/bioinformatics/btx617

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. 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. 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:10.1371/journal.pone.0196878

Wright, A. M., Schwartz, R. S., Oaks, J. R., Newman, C. E., and Flanagan, S. P. (2020). The Why, when, and How of Computing in Biology Classrooms, version 2; peer review: 2 approved. *F1000Res* 8, 1854. doi:10.12688/f1000research.20873.2

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.

NOMENCLATURE

Resource Identification Initiative

Perl Programming Language, RRID:SCR_018313

Python Programming Language, RRID:SCR_008394

Nephele, RRID:SCR_016595

Galaxy RRID:SCR_006281

PATRIC – Pathosystems Resource Integration Center, RRID:SCR_004154

CLC Genomics, Workbench, RRID:SCR_011853

Geneious, RRID:SCR_010519

R Project for Statistical Computing, RRID:SCR_001905

QIIME2–QIIME, RRID:SCR_008249

frontiers in Education

CURRICULUM, INSTRUCTION, AND PEDAGOGY
published: 17 November 2021
doi: 10.3389/feduc.2021.726930

Bioinformatics on the Road: Taking Training to Students and Researchers Beyond State Capitals

*Marcus Braga[1], Fabrício Araujo[2,3,4], Edian Franco[2,5,6], Kenny Pinheiro[2], Jakelyne Silva[1], Denner Maués[2], Sebastiao Neto[2], Lucas Pompeu[2], Luis Guimaraes[2], Adriana Carneiro[2], Igor Hamoy[3] and Rommel Ramos[2]**

[1]Paragominas Campus, Federal Rural University of Amazônia, Paragominas, Brazil, [2]Laboratory of Biological Engineering, Federal University of Pará, Belém, Brazil, [3]Laboratory of Applied Genetics, Federal University of Amazônia, Belém, Brazil, [4]Castanhal Campus, Federal University of Pará, Castanhal, Brazil, [5]Instituto Tecnológico de Santo Domingo (INTEC), Santo Domingo, República Dominicana, [6]Instituto de Innovación en Biotecnología e Industria (IIBI), Santo Domingo, República Dominicana

In Brazil, training capable bioinformaticians is done, mostly, in graduate programs, sometimes with experiences during the undergraduate period. However, this formation tends to be inefficient in attracting students to the area and mainly in attracting professionals to support research projects in research groups. To solve these issues, participation in short courses is important for training students and professionals in the usage of tools for specific areas that use bioinformatics, as well as in ways to develop solutions tailored to the local needs of academic institutions or research groups. In this aim, the project "Bioinformática na Estrada" (Bioinformatics on the Road) proposed improving bioinformaticians' skills in undergraduate and graduate courses, primarily in the countryside of the State of Pará, in the Amazon region of Brazil. The project scope is practical courses focused on the areas of interest of the place where the courses are occurring to train and encourage students and researchers to work in this field, reducing the existing gap due to the lack of qualified bioinformatics professionals. Theoretical and practical workshops took place, such as Introduction to Bioinformatics, Computer Science Basics, Applications of Computational Intelligence applied to Bioinformatics and Biotechnology, Computational Tools for Bioinformatics, Soil Genomics and Research Perspectives and Horizons in the Amazon Region. In the end, 444 undergraduate and graduate students from higher education institutions in the state of Pará and other Brazilian states attended the events of the Bioinformatics on the Road project.

Keywords: education, bioinformatic, computer science, computational biology, training

OPEN ACCESS

Edited by:
Hugo Verli,
Federal University of Rio Grande do Sul, Brazil

Reviewed by:
Rodrigo Ligabue-Braun,
Federal University of Health Sciences of Porto Alegre, Brazil
Ana Ligia Scott,
Federal University of ABC, Brazil

****Correspondence:***
Rommel Ramos
rommelramos@ufpa.br

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *17 June 2021*
Accepted: *21 October 2021*
Published: *17 November 2021*

Citation:
Braga M, Araujo F, Franco E, Pinheiro K, Silva J, Maués D, Neto S, Pompeu L, Guimaraes L, Carneiro A, Hamoy I and Ramos R (2021) Bioinformatics on the Road: Taking Training to Students and Researchers Beyond State Capitals. Front. Educ. 6:726930. doi: 10.3389/feduc.2021.726930

INTRODUCTION

Since the first studies focused on manipulating biological sequences, bioinformatics has played an essential role in the stages of data analysis. In this context, alignment methods for comparing two or more sequences have become popular in implementing software such as Fasta (Pearson and Lipman, 1988) and Blast (Altschul et al., 1990). Since their early versions, these programs have been widely used by the scientific community, regardless of the level of knowledge in bioinformatics, even with other improved software available in both accuracy and execution time, such as Diamond (Buchfink et al., 2015).

However, after the next-generation sequencing platforms (NGS) release in mid-2005 (Schuster, 2008), there has been a significant increase in complete genome sequencing projects. The demand for new bioinformatics solutions capable of handling the data offers fast and accurate results. Despite the challenges imposed by NGS technologies, such as increasing throughput and reducing the size of reads, there are currently platforms capable of generating reads larger than 4 Mb (Fujimoto et al., 2021).

Additionally, the launch of benchtop sequencing platforms such as 454 GS Junior (GSJ), Ion Torrent Personal Genome Machine (PGM) - Life Technologies (Carlsbad, CA), Illumina MiSeq (Jünemann et al., 2013), the evolution of sequencing platforms, the availability of large volumes of biological data and the arising of new analyzes that can be performed, showed that research groups that operate mainly outside the large capitals of Brazil are still not prepared to take full advantage of these technologies because of lack of skilled human resources.

The training of research groups formed by students from different areas and levels: undergraduate, master's, doctoral, and postdoc, has become essential to give small research groups from the countryside, which have significant questions and minimal structure, a chance to participate in the genomic era, improve their research and data analysis, which certainly favors better training of students and national research. The demand for bioinformatics professionals is not exclusive to researchers who work outside the large capitals.

Thus, some professors from the Federal University of Pará and the Federal Rural University of Amazônia decided to act in the training of human resources in bioinformatics, students and researchers, through extension projects in cities from the countryside in order to increase the number of professionals trained in bioinformatics and thus help them to be independent in specific analyzes of their areas.

As bioinformatics is a useful tool for several research areas, the training demands tend to have customized specificities, which in traditional courses are not met. Furthermore, the lack of knowledge of different analyzes and sometimes regarding the possibilities limits the research groups. For this reason, the training has always been designed considering the types of research of the groups/institutions covered in those cities, which is a great innovation in terms of capacity building in bioinformatics, with hands-on activities since the wet lab to dry lab.

CAPACITATION EVENTS

In the last 4 years, five events were held in satellite cities, where the institutions offered the minimum physical structure, with computing and biology laboratories, auditoriums, local support staff, and well-established demands that can be answered by the team working in training. Mainly, the courses include theoretical and practical activities to allow participants to test their knowledge.

The project activities started in 2017, with the Computational Biology Applied to Agribusiness Meeting. The target audience were undergraduate students from the Federal Rural University of Amazônia (UFRA), from agricultural science courses, such as Agronomy, Animal Science and Forestry Engineering, and the newly created Information Systems course, in the city of Paragominas, in the southeast of State. This city is one of the largest agribusiness centers in Pará and received a green seal due to good management practices and exploitation of the environment. Thus, it is common to observe research in the region to evaluate the soil recovery time, management possibilities, including chemical and molecular analyses, where it is possible to collaborate with bioinformatics fully. This was the only course with exclusively theoretical content carried out by the team as the students were still at the beginning of their courses.

In 2019, the team expanded contact with researchers from the UFRA (Paragominas) who worked with soil analysis. Thus, it was possible to plan a special event where students had the opportunity to experience, in a biology laboratory, how a microorganism is isolated from soil samples. After, in courses at the computer lab, they had access to techniques on how to assemble prokaryote genomes and, finally, their annotation. Interestingly, the participants demonstrated that they already had different backgrounds, both biological and computational, which was important in exchanging experiences.

As part of the commitment to the dissemination of science and the training of new professionals, the project financed by the Coordination for the Improvement of Higher Education Personnel (CAPES) called PROCAD Amazônia, which involves collaboration between the Federal University of Pará (UFPA), Federal University of Minas Gerais (UFMG) and UFRA, organized the event entitled Training in Bioinformatics - Belém Stage, which aimed to immerse and train undergraduate and graduate students in bioinformatics through basic and advanced courses that addressed the standard topics of the area. The courses were given from January 13th to 17th, 2020, at the Federal University of Pará, in Belém, capital of the state of Pará.

The training was taught by graduate students who are part of the PROCAD project, under the supervision of the project's professors and researchers. The content was aimed at developing the theoretical-practical skills necessary to learn of the main bioinformatics tools and the analysis of biological data in genomics and transcriptomics to meet local demands. The training was structured with 20% theoretical and 80% practical content.

The training was split into five modules from 4 to 8 h in duration, among which we can mention: Introduction to the Linux Environment (4 h), Introduction to Programming with Python (4 h), Genome Assembly (8 h), Analysis of Transcriptomic Data: RNA-Seq (8 h) and Machine Learning Techniques: Clustering (8 h).

The Introduction to Linux Environment course had the following contents: introduction to the shell, basic Linux commands, and terminal pipeline development. Introduction to Python course had as content the installation of the python environment, main algorithms, reading biological data files, analysis methods, and data processing. The Genome Assembly course addressed topics such as data pre-processing, genome assembly algorithms, assembly of prokaryote and eukaryotic

genomes and analysis, and evaluation of assembly results. Analysis of Transcriptomic Data course contained topics such as obtaining RNA-seq data, algorithms and methods for analyzing gene expression data, differential expression analysis, and technique for data evaluation. Finally, the Machine Learning Technique course focused on clustering techniques and covered topics such as data pre-processing, clustering algorithms and their uses, metrics for evaluating results and techniques for presenting results.

The training was carried out at the Computer Laboratory of the Faculty of Biotechnology of UFPA, which has 24 computers (Intel® Core™ i5-4590S Processor, 4GB, HD 500GB, 23" LED monitor), where hands-on activities of the courses were taken.

For the selection of the participants, a web platform was developed where they could register. This made it possible for undergraduate and graduate students from different universities (public and private) to participate in the training. Ninety applications were received, from which 39 students were selected due to limited computer equipment. For the final selection of the participants, the following criteria were established: *1*) be signed into an undergraduate or graduate program related to the areas of biotechnology and bioinformatics; *2*) have basic computer knowledge; *3*) be conducting research or graduate work in the areas of bioinformatics.

Of the 39 selected students, 60% were graduate students researching in the areas of bioinformatics and 40% were graduate students. Of the total, 80% were students from public higher education institutions and 20% from private institutions. To receive the certificate, students must have participated in at least 90% of the training hours.

As a result, 95% of the participants obtained the participation certificate for having completed the minimum hours and performed all the activities required in training. Due to the great demand for training, the organizing team decided to carry out other introductory and specific training courses, which, as a result of the pandemic caused by COVID-19, were carried out virtually. This allowed the development of activities on a national and international level. In addition to the students enrolled in the event, seven instructors participated in the courses, and three research professors coordinated the training.

Online Events

During the COVID-19 pandemic, between June and August 2020, the project team was asked to conduct training in bioinformatics, especially for students from the research group Nucleus for Research in Applied Computing, at UFRA. These students would start to work in research in bioinformatics. The objective was to train critical and scientific thinking in Computer Science undergraduate students who were starting their research in bioinformatics and, in the future, to encourage their participation in graduate programs. The training program was divided into two modules and addressed topics such as Introduction to Bioinformatics, NGS Sequencing, Biological Sequences Alignment, and Genome Assembly. The training was carried out through the Google Meet platform and was theoretical and practical.

In order to meet the demands of the computing area, in 2020, a course was held to present Artificial Intelligence techniques in the Machine Learning approach and its possible applications in bioinformatics. The Introduction to Machine Learning course was offered with theoretical and practical aspects.

The course was designed over 3-month, between October and December of 2020, carried out, 100% virtually, through the Google Meet platform. The course was held in four modules: in the first unit, called Introduction to Computational Intelligence, an overview of the area was given; the second unit was Machine Learning Fundamentals and addressed types of learning, classic machine learning problems and machine learning algorithms; the third unit dealt with Artificial Neural Networks, giving an overview of the model; and the fourth unit addressed Deep Machine Learning and Deep Neural Networks, presenting models such as Convolutional Neural Networks, Recurrent Neural Networks, Autoencoders, Generative Adversarial Networks, Attention Mechanisms, and others.

50% of the training was carried out with practical activities for the construction of intelligent models and applications in various areas of science, including bioinformatics. Frameworks like Spyder (<https://www.spyder-ide.org/>), Colab (<https://colab.research.google.com/>), Jupyter (<https://jupyter.org/>) and Orange (<https://orangedatamining.com/>) were used for the hands-on activities. The course had 80 participants, 60 of which were linked to various higher education institutions in the state of Pará and 20 linked to institutions in five other states in Brazil. Among the participants, there were undergraduate and graduate students and five professors. This initiative served to introduce intelligent techniques that can be used in bioinformatics research in their respective application scenarios.

Capacitation in Numbers

Training in bioinformatics and computational biology, through the "Bioinformatics on the Road" project, reached dozens of students in recent years, as shown in **Table 1**.

International Actions

Despite not being part of the "Bioinformatic on the road" project, international initiatives were encouraged for collaborators. In April 2019, the course "Bacterial Resistome: from wet laboratory to computational biology" was organized at the Technological Institute of Santo Domingo (INTEC). Fifteen students from different Dominican universities participated in this training, with the primary objective of showing the process for the study of antibiotic-resistant bacteria, from sample collection, DNA extraction, and, finally, the process of analysis and data interpretation through the use of bioinformatics tools. The training had 30 h, divided into 20% theoretical hours and 80% practical hours.

Also, in April 2019, the lecture "Computational Biology" was organized at INTEC, which 30 students from public and private universities attended. In this lecture, the main themes related to the processing, treatment, analysis, and interpretation of genomic data through several bioinformatics tools.

In May 2020, the lecture entitled "Rights and Challenges of Bioinformatics and Computational Biology in the Dominican

TABLE 1 | List of events held in the context of the "Bioinformatics on the Road" project with information on the place where it took place, year and total number of students and teachers.

Evento	City	Year	Students	Professors
Computational Biology Applied to Agribusiness Meeting	Paragominas	2017	10	6
Bioinformatics and Biotechnology Workshops	Paragominas	2019	300	8
Training in Bioinformatics - Belém Stage	Belém	2020	39	3
Training in Bioinformatics - NPCA	Online	2020	15	4
Introduction to Machine Learning: Theoretical and Practical Aspects	Online	2020	80	5

Republic" was given as one of the main activities of the week celebrating the 54th anniversary of the Faculty of Sciences of the Universidad Autónoma de Santo Domingo (UASD).

In July 2020, the project team participated in the "International Symposium for Research and Scientific Solutions in Times of Crisis, COVID-19 and Beyond: Food Safety, Health, Education, Environment and Economy," organized by the Ministry of Higher Education Science and Technology of the Dominican Republic (MESCYT). On occasion, the lecture: "Bioinformatics and computational biology as new ferments to face a crisis in the Dominican Republic" was given.

Lessons Learned

As a basic premise for all events, we lack of qualified personnel to work with bioinformatics and one training only cannot change it. For this reason, with decentralized training, which goes to the origin of the demands, we establish connections with local researchers and also with other collaborators from outside the State, in order to make the environment more collaborative, continuing the growth of production scientific research that is dependent on bioinformatics methods.

Some students who attended the events presented their projects, and participated in courses, also had the opportunity to keep in touch with graduate professors from UFPA and UFRA. As a result, they are already attending graduate programs. Research groups in the state capital selected some students to work in scientific initiation due to their knowledge in computing and already have a glimpse of computational biology and bioinformatics.

Graduate students were also present, especially in Belém do Pará, whose central theme was genomics. Thus, they could first contact bioinformatics and even use this knowledge in their thesis and dissertations.

The lesson that attracted the most attention regarding the Bioinformatics on the Road project is that it has collaborated with the growth of research groups that today already generate independent scientific productions with their local collaborators. Thus, we consider that the events were positive for presenting scientific thinking in the context of bioinformatics, but that it is being used in several areas to establish collaborations and increase local scientific production.

Next Steps

The COVID-19 pandemic showed us that distance training is not only possible but essential, as it opens up frontiers that were previously an obstacle, such as geographic and financial. However, in the northern region of Brazil, access to the internet from the countryside cities to carry out synchronous events is not trivial. Thus, our next steps will be the release of the project's channel on the YouTube platform to disseminate the content of online training, including exercises and answers on platforms that allow us to share the content, using active methodologies applied to remote learning.

CONCLUSION

Bioinformatics remains a promising area but with little training initiatives compared to the demand, which is much more evident outside the state capitals, which is one of the factors that justifies the tremendous demand for courses where there were many places available.

On the other hand, the offering of courses based on the interaction between bioinformatics and research themes studied by local researchers becomes more attractive to students, who end up realizing that the challenge is "only" to study bioinformatics, a way to achieve their goals and not a new research area, which leads to an increase in the number of interested people and subscribers.

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 authors.

AUTHOR CONTRIBUTIONS

MB and RR designed and organize the steps of the project. MB, FA, EF, KP, JS, DM, SN, LP, LG, AC, IH, and RR were lecturers. MB, FA, EF, KP, JS, DM, SN, LP, LG, and AC were practical activities tutor.

FUNDING

The project 88887.200562/2018-00 grant from CAPES. The authors would like to thank all those who contributed and collaborated directly or indirectly to the realization of this

project since its inception. Federal University of Pará Dean of Extension and Federal Rural University of Amazon Dean of Extension for the support. Federal University of Pará Dean of research (PROPESP/UFPA) for the financial support to the manuscript production.

REFERENCES

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

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and Sensitive Protein Alignment Using DIAMOND. *Nat. Methods* 12 (1), 59–60. doi:10.1038/nmeth.3176

Fujimoto, A., Wong, J. H., Yoshii, Y., Akiyama, S., Tanaka, A., Yagi, H., et al. (2021). Whole-genome Sequencing with Long Reads Reveals Complex Structure and Origin of Structural Variation in Human Genetic Variations and Somatic Mutations in Cancer. *Genome Med.* 13, 65. doi:10.1186/s13073-021-00883-1

Jünemann, S., Sedlazeck, F. J., Prior, K., Albersmeier, A., John, U., Kalinowski, J., et al. (2013). Updating Benchtop Sequencing Performance Comparison. *Nat. Biotechnol.* 31, 294–296. doi:10.1038/nbt.2522

Pearson, W. R., and Lipman, D. J. (1988). Improved Tools for Biological Sequence Comparison. *Proc. Natl. Acad. Sci. U S A.* 85 (8), 2444–2448. doi:10.1073/pnas.85.8.2444

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/feduc.2021.726930/full#supplementary-material

Schuster, S. C. (2008). Next-generation Sequencing Transforms Today's Biology. *Nat. Methods* 5 (1), 16–18. doi:10.1038/nmeth115619

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 18 May 2022
doi: 10.3389/feduc.2022.727019

OnlineBioinfo: Leveraging the Teaching of Programming Skills to Life Science Students Through Learning Analytics

Raquel C. de Melo-Minardi[1]*, *Eduardo C. de Melo*[2] and *Luana L. Bastos*[3]

[1] Laboratory of Bioinformatics and Systems, Department of Computer Science, Institute of Exact Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, [2] Take Blip, Belo Horizonte, Brazil, [3] Bioinformatics Graduate Program, Department Biochemistry and Immunology, Institute of Biological Sciences, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Online learning has grown in recent years and has become popular with Massive Open Online Courses (MOOCs). The advent of the pandemic has undoubtedly made more teachers and students experience the online learning experience. Distance learning is going to grow even more in the coming years. In this article, we present our computational thinking and programming course focused on life science students. We introduce our approach for analyzing how students interact with didactic resources regarding their probability of completing the course. We discussed several insights this strategy brought us and how we can leverage the teaching of programming skills to life science students through learning analytics. We suggest that machine learning techniques will be increasingly essential for better monitoring and supporting students and for online courses improvements.

Keywords: bioinformatics, computer programming, online education, Python, learning analytics, artificial intelligence, neural networks, distance education

OPEN ACCESS

Edited by:
Chi-Cheng Chang,
National Taiwan Normal University, Taiwan

Reviewed by:
Candido Cabo,
The City University of New York, United States
Diego Onna,
University of Buenos Aires, Argentina

***Correspondence:**
Raquel C. de Melo-Minardi
raquelcm@dcc.ufmg.br

Specialty section:
This article was submitted to STEM Education, a section of the journal Frontiers in Education

Received: *28 July 2021*
Accepted: *12 April 2022*
Published: *18 May 2022*

Citation:
de Melo-Minardi RC, de Melo EC and Bastos LL (2022) OnlineBioinfo: Leveraging the Teaching of Programming Skills to Life Science Students Through Learning Analytics. Front. Educ. 7:727019. doi: 10.3389/feduc.2022.727019

1. INTRODUCTION

The intensification of the use of information and communication technologies (ICT) boosted the growth of the distance education (DE) modality in higher education. This growth was accelerated due to the COVID-19 pandemic, culminating in important reformulations in teaching practice (Dhawan, 2020).

The DE is intentionally designed for autonomous learning, with a variety of resources, such as study guides, digital books, video lessons, forums, and virtual assessments. It is based on the theoretical framework of pedagogy defined as the science whose object of study is education, as well as the teaching and learning process; the andragogy that deals with the study of adult learning; and heutagogy, an extension of andragogy, which deals with the student's independence about how and what he wants to learn (Agonács and Matos, 2020). It can be practiced in synchronous and asynchronous models. The synchronous learning environment is where students have access to classes in real time. In the case of asynchronous environments, the content can be offered through recorded lessons (Dhawan, 2020).

Among the advantages associated with distance learning is accessibility, enabling knowledge to reach more remote areas and giving the student greater freedom in planning the time for carrying out activities. Among the aspects of negative impact is less interaction with colleagues and teachers,

in addition to aspects related to performance evaluation and feedback, which need to be adapted to the online medium for greater effectiveness (Higashi et al., 2017).

In 2017, we implemented a distance university extension course[1] (de Melo-Minardi and Bastos, 2021) aimed at teaching computer programming to students and graduates in life sciences. Teaching computational thinking and a programming language for life science students involves some challenges. Among these, the great heterogeneity of training and level of knowledge in programming logic stands out. In the first half of 2019, we opened enrollment for the first class of the distance university extension course, with 101 enrolled for the first class. Currently, we have trained more than 1,000 students in more than 40 different undergraduate courses. The demand and success in training these students show an opportunity to offer strategic and diversified content to students in areas traditionally not covered by certain disciplines.

The course teaches computational thinking, uses Python programming language, and also teaches more in-depth concepts, such as algorithm complexity analysis techniques, and ends with classic Bioinformatics algorithms for sequence alignment.

The pandemic forced teachers around the world to quickly adapt to teaching content to hybrid teaching. This seems to be a point of no return. Education will never be the same, from now on we can take advantage of the best in face-to-face and distance learning through more hybrid strategies that make the best use of technology for teaching and assessment. With the huge amount of data that online learning platforms collect and generate, teaching methods can be improved through learning analytics.

With a large amount of data that online teaching platforms collect and generate, teaching methods can be improved through *learning analytics*. We profited from the opportunity and the increasing number of students in our course to implement these strategies and further improve our lessons, didactic resources, and pedagogic trails. It was possible due to data science methods.

We analyzed a class consisting of 245 students. Of these, 101 completed the course, which gives a 41% completion rate. According to Jordan (2015), MOOC completion rates or the percentage of enrolled students who complete the course vary from 0.7 to 52.1%, with a median of 12.6%. Our course has a significant completion rate, but we want to increase it, to comprehend better the factors that lead to student dropout, to support students in completing the course, and continually improve instructional design and course material.

We collected data from *Moodle*[2] virtual learning environment and from *Google Forms*[3] surveys. We gathered information from the students before, during, and after the conclusion of the course. We used machine learning models to comprehend the course content's use and its relation to course completion. The strategy proved to be promising in the evaluation of course resources, proposed activities, and instructional design and predinctig student's drop out.

[1]http://onlinebioinfo.dcc.ufmg.br/cursos
[2]https://moodle.org/
[3]http://forms.google.com/

2. LEARNING ANALYTICS

Keats and Schmidt (2007), Lengel (2013), and Gerstein (2014) used the term "Education 3.0" to denote a new students generation, more digital. We now teach a cohort of students who grew up in a digital world where learning occurs anywhere, anytime, mediated by ICT.

Learning analytics (Siemens and Baker, 2012) is an emerging field and aims at using data related to students to build better educational material and strategies. A particular trend is to trace the profiles of the students, collecting data about their interactions with online activities to provide reliable information about the learning results achieved (Johnson and Palmer, 2015).

Among the valuable data for this purpose, we cite pass and disapproval data; data for accessing educational resources such as texts, images, and videos; data from participation in learning activities such as quizzes, open activities, discussion forums; performance data on learning and assessment activities; social interaction data (relationship with study colleagues and professors, for example); satisfaction survey data.

Through this data-based comprehension, one can build better andragogical proposals, train students to have a pro-active rule in the learning process, identify the potential of course abandonment (Kampff, 2009), and evaluate the aspects that affect course completion. This educational data science can help provide immediate feedback and adjusts to the educational contents and activities.

Considering the possible dimensions of analysis, the potential, and the importance of adopting learning analytics methods to know the teaching situation, the related factors, and consequently, guide the implementation of improvements in educational systems.

In this article, we focus our attention on three significant dimensions for data analysis in educational environments, including the segmentation of students based on the patterns of access to course resources; the identification of at-risk of abandonment students, and the evaluation of the use and perception of each material/activity/task.

3. THE COURSE

We evaluate our online course that aims to introduce programming logic, teach the Python programming language, introduce algorithm complexity, and present classical bioinformatics.

The content of the course is as follows:

- **Module 0 - Welcome**: reception of the students, explaining the functioning of the course, and preparing the computational environment for the course.
- **Module 1 - Introduction**: definitions and fundamentals of bioinformatics, computational biology, computer science, algorithms, problems, data structures, programs, programming language, compiler, computer components and their relationship with programming, the complexity of algorithms.

- **Module 2 - Programming**: Python language, Python in bioinformatics, essential Python syntax variables, variable types (sets, tuples, lists, dictionaries), arithmetic operators, string comparators, logical operators, conditional structures, defined repetition structures and undefined, loop control, input, and output, formatted printing, code modularization (subroutines and modules), and regular expressions. It is a practical module with several practical exercises to hand on.
- **Module 3 - Algorithm complexity analysis**: algorithm complexity functions, best case analyses, average and worst case analyses, optimal algorithms, asymptotic behavior of complexity functions, asymptotic domination, O notation, complexity classes, several examples involving, among others, search algorithms. This module is largely made up of theoretical content, exercises are provided at the end of each class to fix the content learned. Submitting these activities is optional.
- **Module 4 - Algorithms for bioinformatics**: paradigm concept in computing, dynamic programming, token game example, tourist problem in Manhattan, distance metrics between sequences (Hamming and Levenshtein), maximum common subsequence problem, Needleman-Wunsch algorithm, Smith-Waterman algorithm, scoring schemes, and substitution matrices, peer-to-peer alignments, multiple alignments, global alignments, local alignments, and heuristics. This module also has some optional challenges.
- **Bonus module**: structural bioinformatics bonus module. It has no exercises or practical activities and is entirely optional.

he target audience comprises undergraduate or graduate students in biological sciences and related fields with little or no programming knowledge. We frequently receive as well students who graduated in computer science-related areas. They aim to learn Python, be introduced to bioinformatics problems, and review some computational fundamentals. Students holding more advanced knowledge of Python programming will benefit from the second half of the course, in which we cover more advanced topics related to algorithm complexity and classical algorithms in bioinformatics. This course can help prepare them to enter the graduate course in bioinformatics and computational biology, contributing to their acquisition of solid computing skills.

The course resources consist of the following:

1. **4 digital books**: in pdf format, containing theory and challenges (mostly solved) for practical exercises in logical reasoning and programming, totaling 100 pages.
2. **Video recorded classes**: 35 classes (approximately 7 h).
3. **Slides**: presentations used in classes will be made available in pdf format.
4. **Review and programming tasks**: quizzes for review of taught concepts and lists of programming exercises.
5. **Google Colab Notebooks**: solved and commented programming exercises.
6. **Live classes**: meetings to clarify doubts through video conferences with the students.

The course lasts for up to 90 days and takes about 40 h to attend classes, read the material, solve course exercises, and complete the course.

3.1. Student Evaluation

Student assessment considers attending asynchronous classes, reading materials, review exercises, and practical programming exercises. To be considered a complete student, the student must attend 75% of classes, attend final classes, and complete 60% of the submitted exercises. We collect the data describing the use of the resources by the students from the Moodle platform. The data set consists of yes/no values for each pair of student-resource. Review exercises contain closed questions of various types: multiple-choice, association, and filling in gaps, among others. The programming exercises are practical and have to be solved using the Python programming language. Proposed solutions in a *Google Colab Notebook*[4] and a video explaining the solution step-by-step accompanies each list of programming exercises. The correction of the exercises is done by the students themselves through correction classes and workbooks of solved activities. The analysis was carried out with data from 245 students from different courses, most of them coming from the biological sciences course, about 39.1% (see **Figure 1**). The data used consists of a table formed by the course resources such as class and delivery of activities with values of yes for attended and delivered and no for unattended and not delivered and the course completion section with values of yes and no. Data were obtained from the Moodle platform. To obtain the reports used in the analysis, the function "course management" and then "view participation report" was used. We chose "all course activities," throughout the course period, filtered only by students and the "view" action. Then, we download a spreadsheet in xls format and perform all the analysis using Orange Data Mining.

4. MATERIALS AND METHODS

To obtain the reports used in the analysis, we used the Moodle function "course management" and then "view participation report." We chose "all course activities," throughout the course period, filtered only by students and by the "see" action. Each line of the data set is a student, each column is a course resource, and the domain of the features is in *Yes*, *No* domain. We then downloaded an xls format sheet and performed all the analysis using *Orange Data Mining*[5] software.

4.1. Visualizing Students According to Their Profiles of Material Accession

First of all, we needed to have a visual grasp of the whole set of students according to completion or abandonment of the course. We used Multidimensional scaling (MDS) Carroll and Arabie (1998) which is a technique that finds (in this case) a 2D projection of instances, reproducing their distances as well as possible. As input, the technique needs a matrix of distances.

[4]https://colab.research.google.com/
[5]https://orangedatamining.com/

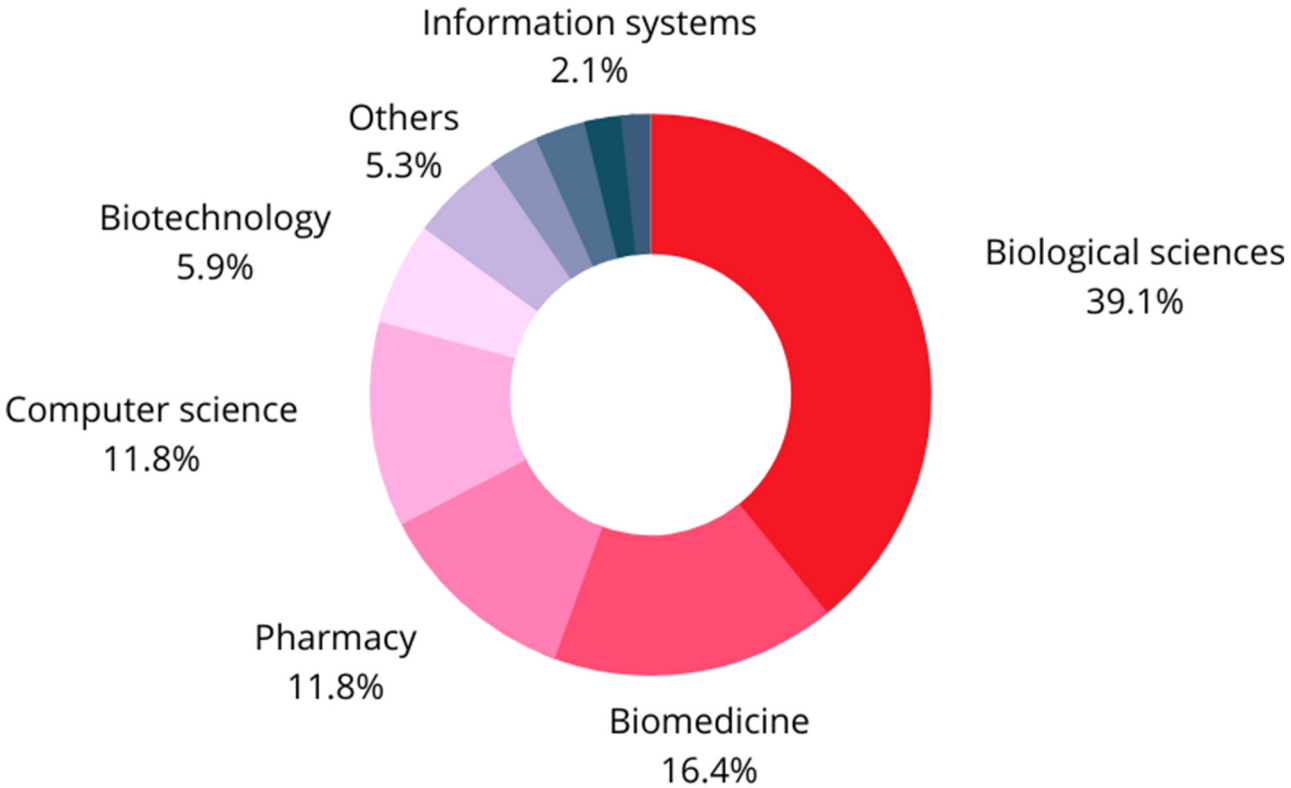


FIGURE 1 | Student graduation areas. 2% of the students have completed the high school level and are not yet enrolled in a graduate course, 28% are undergraduates, 27.5% are graduates, 27.5% are masters, and 15% are Ph.D.s.

As our attributes are categorical having values yes/no according to whether the student had accessed the material or not, the distances are zero for equal values and one for different values. The MDS algorithm iteratively moves the points around in a simulation of a physical-based model. When two points are too close/too far to each other, it applies a force pushing them apart/together.

We visualized the set of students in a 2D cartesian plan and colored each one according to completion (red) or abandonment (blue). It is presented in **Figure 2** which will be discussed later in the Results session.

4.2. Identification of At-Risk of Abandonment Students

To identify whether a particular student would be at risk of dropping out of the course, we built a binary classification model. We evaluated through cross-validation how accurately the model can get the label right (conclusion = *yes* or *no*). For the classification task, we tried several algorithms: K-Nearest Neighbors (KNN) (Peterson, 2009), Decision tree (Chang and Pavlidis, 1977), Random Forest (Breiman, 2001), Gradient Boosting (Friedman, 2002), Support Vector Machine (SVM) (Noble, 2006), and Logistic Regression.

All the models achieved accuracy greater than 0.8 with the complete data set, but neural networks outperformed, achieving 1.0. For this reason, we present and discuss only these results. We used a neural network with 100 neurons in the hidden layers, ReLu activation function, Adam optimizer, and a maximum number of interactions of 200. These were all default values, and we did not make any further assessment of the impact of varying these choices.

4.3. Evaluation of the Use and Perception of Course Material

To assess how closely each material is related to the target class (which indicates whether the student has completed the course or not), we created a ranking of attributes (resources) in classification using various indices. This ranking is based on a metric that scores attributes according to their correlation with the class, based on internal scorers (information gain, gain ratio, gini index, χ^2, reliefF, and FCBF). Although the final ranking is calculated based on several indices, we present in the tables the information gain and the gain ratio due to space restrictions as they also discriminated the more informative attributes for classification. Information gain (Kent, 1983) is a classical machine learning score and measures the reduction in entropy by splitting a dataset. It is calculated by comparing the entropy of the dataset before and after this split. The concept of entropy comes from information theory and is the purity of a dataset or how balanced the distribution of classes happens to

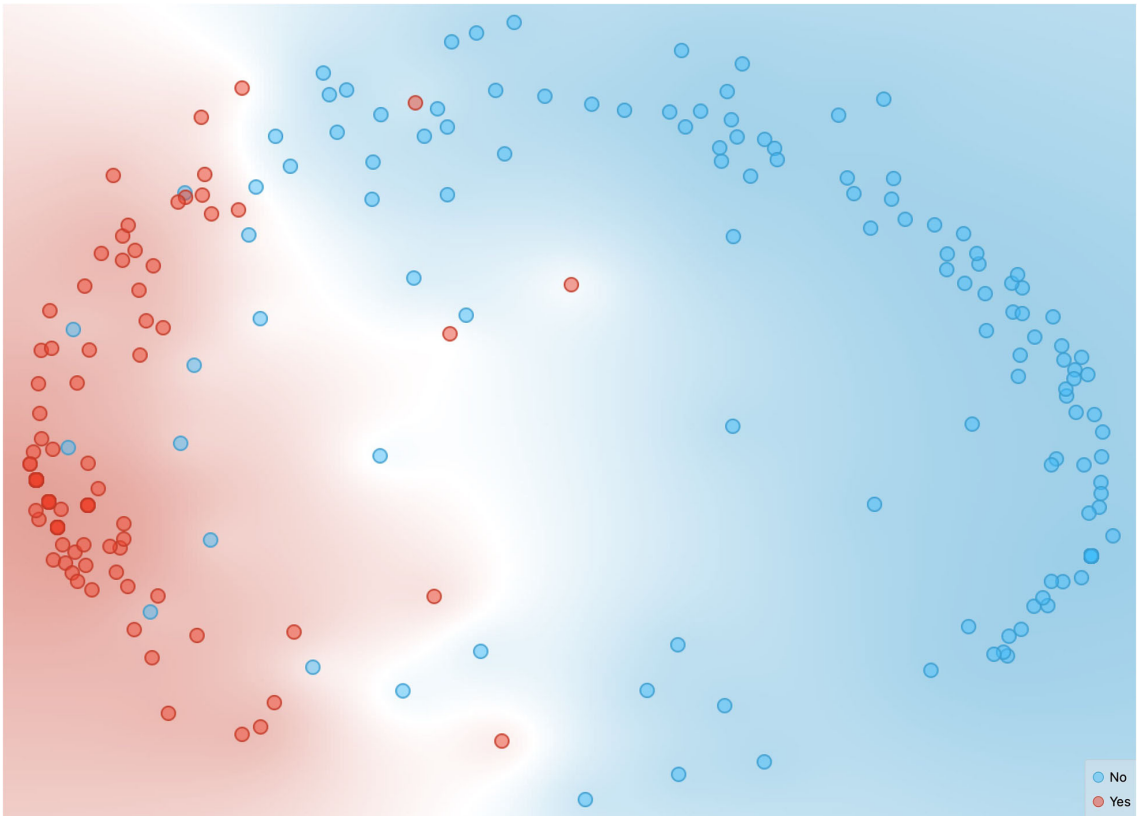


FIGURE 2 | Multidimensional scaling (MDS) visualization of the students. In red, we show the students who complete the course, and blue represents the ones that dropped out.

be. An entropy of 0 (minimum) means that all the instances are of the same class, and an entropy of 1 (maximum) means that the classes are equally represented in the group of instances.

5. RESULTS AND DISCUSSION

We will present the results of the analysis carried out with data from a single class of students who took the course from August 2020 onwards. We had 245 students. Of these, 101 completed the course, which gives a 41% completion rate. According to Jordan (2015), completion rates of MOOCs vary from 0.7 to 52.1%, with a median of 12.6%. Although we have a relatively high completion rate, we want to better understand the use of the course resources, the main difficulties, and the improvements that can be implemented. Thus, the objectives of our analysis were student monitoring, support and satisfaction, and instructional design and planning.

A central question we wanted to answer was whether it would be possible to predict that a particular student would complete the https://www.overleaf.com/project/623b8acc04b7506347793b4a course by observing his/her pattern of resources study. We would also like to understand the pattern of dropping out of the course concerning the module/lesson/task done. Are there more difficult lessons or tasks that contribute to student dropout and whose content could be improved?

In **Figure 2**, we show the set of 245 students colored by red (completed the course) and blue (dropped out). We plotted them according to the distance between their profiles of study of course resources. It is possible to discriminate between the ones that conclude or not conclude the course indicating they have different behaviors.

Next, we wanted to see if a machine learning algorithm was able to differentiate the students who would finish those who would give up. A neural network hit 100% of the instances. We did the same with resources from each module of the course and the results are presented in **Table 1**. The same metrics for module 0 obtained by a varied set of classifiers are in **Table 2**.

We can see that the last module allows us to predict without error if a student completes the course, which is obvious. However, surprisingly, it is also possible to predict the completion with a relatively high accuracy through the course's first module (Module 0 - Welcome, 0.71). This prediction accuracy naturally increases as more resources are studied and more modules are concluded.

We observed that module 0 was efficient to predict the dropout of students from the course even using different classification algorithms. Accuracy values equal to 0.71 were obtained for Neural Network, 0.70 KNN, Decision tree 0.73,

TABLE 1 | Neural network metrics for the prediction of course completion.

Resources	AUC	CA	F1	Prec.	Rec.	TP	FP	TN	FN
Module 0 - Welcome	0.71	0.71	0.71	0.72	0.71	63.2%	36.8%	78.7%	21.3%
Module 1 - Introduction	0.75	0.69	0.68	0.69	0.69	66.4%	33.6%	70.2%	29.8%
Module 2 - Programming	0.81	0.77	0.77	0.77	0.77	73.0%	27.0%	79.8%	20.2%
Module 3 - Complexity	0.90	0.89	0.90	0.91	0.90	82.9%	17.1%	96.3%	3.7%
Module 4 - Algorithms	1.00	1.00	1.00	1.00	1.00	100.0%	0.0%	100.0%	0.0%
Bonus	0.78	0.85	0.85	0.85	0.85	84.8%	15.2%	84.7%	15.3%
Practical exercises	0.78	0.79	0.79	0.79	0.79	74.4%	25.6%	81.8%	18.2%
Review + Practical	0.77	0.78	0.78	0.78	0.78	73.9%	26.2%	80.5%	19.5%
Task Loop + Input and output	0.87	0.79	0.79	0.80	0.80	72.4%	27.6%	85.2%	14.8%

AUC, area under ROC curve; CA, classification accuracy; F1, weighted harmonic mean of precision and recall; Prec., precision; Rec., recall; TP, true positive rate; FP, false positive rate; TN, true negative rate; FN, false negative rate.

TABLE 2 | Comparison of metrics of a varied set of classifiers.

Model	AUC	CA	F1	Prec.	Rec.
Neural network	0.71	0.71	0.71	0.72	0.71
KNN	0.70	0.70	0.69	0.69	0.69
Decision tree	0.73	0.74	0.73	0.74	0.74
Random forest	0.72	0.72	0.71	0.72	0.72
Gradient boosting	0.72	0.73	0.72	0.73	0.73
Support vector machine	0.793	0.74	0.73	0.75	0.74
Logistic regression	0.76	0.67	0.67	0.67	0.67
Naive bayes	0.787	0.64	0.63	0.72	0.64

TABLE 3 | Top 10 course resources correlated with course completion.

# Rank	Content	Info gain	Gain ratio
1	Lesson 47 - Multiple alignment algorithms	0.978	1.000
2	Lesson 46 - Smith-Waterman algorithm	0.816	0.831
3	Lesson 45 - Types of sequence alignment algorithms	0.746	0.751
4	Slides 47 - Multiple alignment algorithms	0.734	0.741
5	Lesson 43 - Needleman-Wunsch algorithm	0.732	0.736
6	Lesson 44 - Needleman-Wunsch algorithm implementation	0.720	0.726
7	Lesson 42 - Sequence distance measures	0.681	0.683
8	Slides 46 - Smith-Waterman algorithm	0.680	0.683
9	Slides 44 - Needleman-Wunsch algorithm implementation	0.644	0.645
10	Slides 45 - Types of sequence alignment algorithms	0.644	0.645

Random Forest 0.72, Gradient Boosting 0.72, Support Vector Machine 0.79, Logistic Regression 0.76, and Naive Bayes 0.78. The suggested hypothesis is that module 0 contains the introductory information of the course, in addition to tutorials for the preparation of the environment, it is imagined that when the first instructions are not met, students have more difficulties in performing programming activities, which can lead to abandonment. In addition, the results obtained may simply reflect a behavioral characteristic of the students, since they start the course performing tasks and attending classes, they are more likely to complete the course than those who do not start fulfilling the proposed activities.

We analyzed whether the use of review exercises and programming exercises themselves were sufficient to indicate a greater chance of completing the course. We have noticed that review exercises are less explanatory while programming exercises can help predict with a bit higher accuracy. We evaluated each exercise list and remarked on the great importance of content referring to loop structures. This topic is fundamental in programming logic. Students who complete this list of programming exercises and the following (about input/output) are more likely to complete the course (precision and recall of 0.8, AUC of 0.87, and an F1-measure of 0.79).

We built a ranking of resources by their correlation to course92s completion by students who accessed them. The most correlated resources are at the end of the course naturally (**Table 3**). It is noteworthy that recorded video lessons are always best ranked than respective slides. A curious fact is that all review exercises are at the very bottom of the ranking, after every extra and support only resources (**Table 4**).

A curious fact is that all review exercises are at the very bottom of the ranking, after every extra and support only resources. This surprises us because we assumed that review exercises were very useful in online courses. They are important to give the student immediate feedback on what he/she has learned in each lesson, helping to absorb the content, and are a guide in the need to re-study certain topics. We have to investigate further the use students make of review quizzes and understand how we can use them more effectively.

Table 5 shows that all four review exercise lists were widely accessed, and the average grades are reasonable. Consequently, we have to investigate with next classes if and how this type of exercise is contributing to learning. They are very theoretical and aim to fix concepts and diagnose problems in acquiring some knowledge related to a set of particular lessons.

TABLE 4 | The bottom 10 course resources correlated with course completion.

# Rank	Content	Info gain	Gain ratio
236	Extra - OnlineBioinfo YouTube chanel	0.119	0.120
237	Poll best days and times for live classes	0.118	0.148
238	Inaugural class	0.118	0.154
239	Lesson 2 - Preparing the environment for the course	0.118	0.154
240	Pre-course form - Multiple alignment algorithms	0.114	0.150
241	Schedule	0.071	0.138
242	Review exercises - Lessons 4–6	0.071	0.076
243	Review exercises - Lessons 10–16	0.179	0.165
244	Review exercises - Lessons 17–19	0.142	0.134
245	Review exercises - Lessons 7–9	0.111	0.110

TABLE 5 | A number of attempts and average grades in the review exercises.

Task	# attempts	Percentage students	Average grade
Review exercises - Lessons 4-6	149	60.1%	72.8
Review exercises - Lessons 7-9	262	106.9%	82.6
Review exercises - Lessons 10-16	191	77.9%	85.4
Review exercises - Lessons 17-19	189	77.1%	80.8

TABLE 6 | A number of attempts in the programming exercises.

Task	# submissions
Task submission - Basic syntax	185
Task submission - Strings	163
Task submission - Tuples	159
Task submission - Lists	150
Task submission - Sets	147
Task submission - Dictionaries	144
Task submission - Operators	114
Task submission - Conditionals	109
Task submission - Loops	66

Regarding the programming exercises, the first that appears in the ranking (54th position) is about loop structures, with the info gain and gain ratio meager at values of 0.336 and 0.337, respectively. **Table 6** shows the number of students who have turned in each of the programming exercise lists. As expected, the number of work submissions decreases throughout the course, indicating more difficulty for some students in more complex exercises. There is a significant drop in the number of deliveries of the list of loops. From our point of view, it is the heaviest in terms of the programming logic domain. Despite the low predictive capacity and the low number of deliveries of the latest exercise lists, we note that there is higher access to the solutions for these exercises (**Table 7**).

In **Figure 3**, we illustrate the distributions of attributes referring to the delivery of the programming exercises. Blue represents the students who abandoned the course, and red represents the students, who concluded the course. The left bars

TABLE 7 | Solutions to the programming exercises.

# Rank	Content	Info gain	Gain ratio
51	Solution notebook - Lists	0.396	0.414
60	Solution notebook - Strings	0.387	0.399
61	Solution notebook - Dictionaries	0.387	0.410
62	Solution notebook - Operators	0.387	0.410
63	Solution notebook - Tuples	0.386	0.400
65	Solution - Lists	0.385	0.404
66	Solution notebook - Basic syntax	0.377	0.388
68	Solution - Sets	0.376	0.401
69	Solution notebook - Sets	0.376	0.401
70	Solution - Strings	0.376	0.388
76	Solution - Basic syntax	0.386	0.377
77	Solution - Dictionaries	0.366	0.392
78	Solution - Operators	0.366	0.392
83	Solution notebook - Conditionals	0.357	0.383
89	Solution notebooks - Loops	0.346	0.370
90	Solution - Tuples	0.346	0.358
93	Solution - Conditionals	0.339	0.371
95	Solution - Loops	0.338	0.366

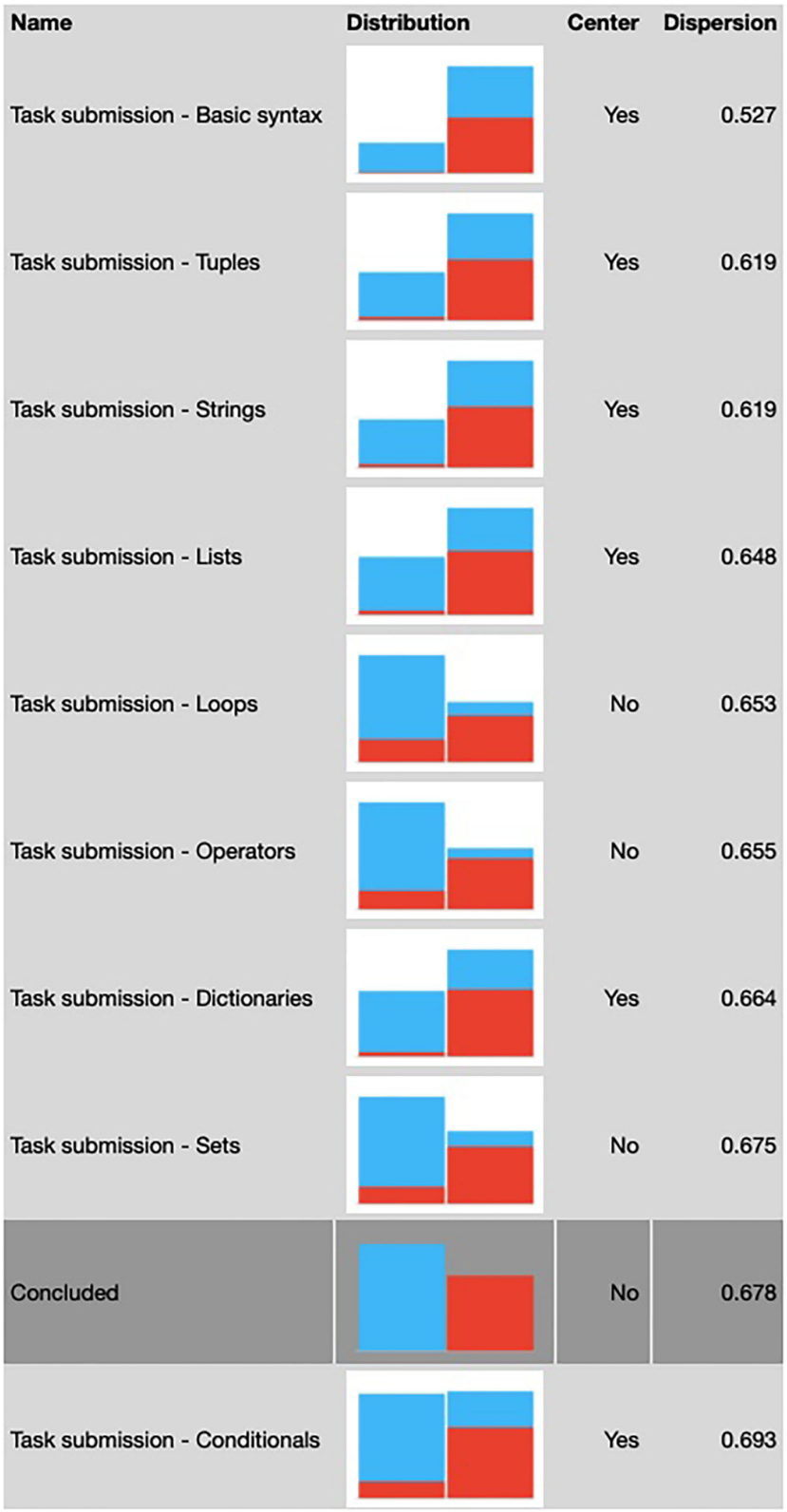
indicate those who did not hand in the task, and the right ones indicate those who did. The column "central" shows the mode of that attribute and the "dispersion", the entropy of the distribution. We can notice that most of the students who completed the course handed most of the tasks (big red fraction on the right bar). The tasks that were considered the most challenging were about loops, operators, and sets. This is a very useful analysis to guide the instructor to improve course material and design.

The programming exercises are critical tasks in the course since computer programming is essentially a practical activity. Thus, we would like all students who completed the course to be fully capable of performing all programming exercises. **Table 6** shows us that 66 students submitted the last list of practical exercises among the 101 who completed the course. This means they attended the classes until the end but did not submit all the tasks.

Hence, we consider that only 65% of graduates completed all the course exercises. Consequently, if we were to consider graduates only those who successfully submitted all activities, we would have a completion rate of 27%. Thus, this analysis is critical to raise the quality of the course, to produce more support teaching resources, and offer better follow-up and support to students.

As a result in next classes, we created surveys, using Google Forms, to assess the students' perception of the difficulty of each programming exercise. For each question, we use a Likert scale between 1 and 5, from easiest to the most difficult. As a result, we obtained the following (**Table 8**) perception of students from the two classes after the class that was initially analyzed.

We confirmed that students perceive the last three lists of programming exercises (operators, conditional structures, and loops) as the most difficult. Therefore, we evaluate each of the



Name	Distribution	Center	Dispersion
Task submission - Basic syntax		Yes	0.527
Task submission - Tuples		Yes	0.619
Task submission - Strings		Yes	0.619
Task submission - Lists		Yes	0.648
Task submission - Loops		No	0.653
Task submission - Operators		No	0.655
Task submission - Dictionaries		Yes	0.664
Task submission - Sets		No	0.675
Concluded		No	0.678
Task submission - Conditionals		Yes	0.693

FIGURE 3 | Task submission statistics related to course completion. Each line is a specific task list. The line in a darker shade of gray shows the proportion of students who conclude the course (red) and dropped out (blue). The left bar is composed of students who did not hand the task and the right is composed of the ones who did it.

TABLE 8 | The average level of difficulty perceived by the students in the programming exercises.

Programming list	Perceived difficulty	# Responses
Basic syntax	1	152
Strings	1	111
Tuples	1	109
Lists	2	95
Sets	2	90
Dictionaries	1	88
Operators	3	78
Conditionals	3	77
Loops	4	51

We used a Likert scale between 1 and 5, from easiest to most difficult.

exercises separately and use this information to identify the most critical difficulties and propose new exercises in the following classes. We now also have live lessons for these topics.

We did this having in mind the concept of *flow* proposed by Csikszentmihalyi et al. (2014). According to them, it is ideal for learning that the tasks offered for the students respect a balance between the level of difficulty and the student's knowledge at the point. Tasks far below this level can be tedious, and tasks far above the acquired skills are also discouraging. The professor's challenge involves keeping the difficulties of the tasks in an equilibrium between difficulty and acquired capacity, the flow region.

This challenge becomes more defiant when we have heterogeneous students in distance learning. We have a very limited perception of engagement and student performance. Accordingly, an exciting direction for learning analytics and machine learning techniques in education is developing educational paths and the recommendation of resources suited to each student's particular profile and level. In this way, we could recommend exercises that are more and more challenging according to the evolution of each student. We could still offer more fundamental or reinforcement resources for those with more elementary difficulties on specific contents.

6. PERSPECTIVES

The computer programming online course for life science students has been offered since 2019 and has already graduated more than 600 students. They originated from more than 40 different undergraduate courses. It is a vast and heterogeneous audience. We ask students to voluntarily evaluate the course through a survey implemented through the Google Forms tool. Among the 496 graduates in the first four classes, 124 evaluated the course. More than 95% gave grades of 7 or more out of 10 on the question "Would you recommend this course to a friend?" It shows that, in general, the course was considered helpful by the graduates. Approximately 25% of the enrolled got to know the course through the indication of a colleague.

Although we have been monitoring student satisfaction from the beginning, this is the first time we have collected data on their access to course content. We have gained essential insights through learning analytics for course improvement and the teaching-learning process.

For future work, we intend to collect interaction data from the forums, analyze temporal data for each student (order and number of times they use each resource), and better assess the students' performance in the programming exercises. We also want to accumulate historical series from several classes, making possible broader inferences about the results. We would like to evaluate if there are groups or profiles of students that can be treated collectively in terms of necessities or styles of learning. It can guide future decisions about learning paths and guide the elaboration of new resources making it more assertive.

Another natural direction is the use of automatic code correction tools. For instance, Pereira et al. (2020) developed *CodeBench*, a tool that allows data collection of student programming at the level of keystrokes, number of code submissions, and grades. They also have carried out a fine-grained analysis of effective/ineffective behaviors regarding learning programming. We intend to evaluate how those platforms that integrate programming IDEs and programming data collection could be integrated into our course (Berland et al., 2013; Fu et al., 2017; Grover et al., 2017). We also intend to study how the mistakes made can be used to assist in the correction of codes and the recommendation of didactic resources (Fernandez-Medina et al., 2013).

It may also be interesting to have a more detailed analysis of the codes produced by the students, as done by Blikstein (2011). He evaluated the programming style of the students and developed metrics related to compilation frequency, code size, code evolution pattern, and frequency of correct/incorrect compilations that could assess the students and the course itself. In another work (Blikstein et al., 2014), the same authors stated that learning to program is very personal, and there are multiple ways to build expertise in programming. We want to investigate the personal aspects and modes of learning in the context of our course in deep.

So our central perspective is to collect and integrate more data on the use of the resources and evaluate more machine learning techniques. We want to perform other tasks we did not go further in this study, such as student cohort analysis (through clustering), identifying sequences of students' behaviors, and identifying rules that indicate risk of abandonment, among others.

7. CONCLUDING REMARKS

In the present study, we present our online course content to teach introductory computer science concepts to life science students, mainly intending to give them skills to start in bioinformatics.

This course has been offered since 2019, having trained more than 600 people from more than 40 different undergraduate courses. The course has received good feedback from students on course evaluation surveys, and more than 95% of graduates who evaluated the course gave grades higher than 7 out of 10.

This study accounts for how we use learning analytics techniques to scrutinize student access data to course resources.

We evaluate their access to recorded lessons, lesson slides, digital books, review exercises, programming exercises, and the proposed solutions for them.

For the class evaluated in this study, we had a completion rate of 41% if we consider attending to 75% of the classes and delivering 60% of the practical exercises. A total of 27% of the students delivered all the programming exercises. It is a relatively high completion rate when compared to the average completion of online courses. According to Jordan (2015), MOOC completion rates vary from 0.7 to 52.1%, with a median of 12.6%.

This study showed us that it was possible to predict student dropout risk with considerable accuracy by evaluating even the course's first welcome and introduction modules. Students who tend to complete the course already show a different behavior from those who do not. In addition, we were able to notice that students access recorded video lessons much more than written material (slides and books). We also notice that there are a greater number of visualizations for solutions of programming exercises than submissions of solved exercises. In addition, there are also fewer views to guided solutions for exercises than to the written solutions (in Google Colab Notebooks). We believe that both patterns are related to "better use of time." We note that performing the review exercises in quiz format is less significant for predicting the course's completion. We believe that students consider that the practical activity of programming will give them more gain than theoretical exercises. These hypotheses confirm the point initially made about a new generation of students strongly motivated by what they want to learn and willing to pursue the desired skills, going straight to the point in a very pragmatic way.

We observed that only 65% of graduates delivered all activities and identified three main contents that caused more difficulties. To confirm this hypothesis, we applied a questionnaire to survey the difficulties perceived by the students in each of the exercises. The exercises more related to programming logic (operators, conditionals, and loops) were perceived as more difficult. It led us to propose new sets of exercises related to these topics. Our goal is to maintain a good balance between the challenges offered to students and the level of skills acquired so far. It is a great challenge, and in future works, we also intend to use machine learning techniques that support us in improving instructional design and providing more personalized learning trails.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because Brazilian general law of individual data protection. Requests to access the datasets should be directed to raquelcm@dcc.ufmg.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

AUTHOR CONTRIBUTIONS

RM-M conceived the instructional project of the online course, prepared, and taught the recorded and live classes, conceived and executed the analysis, wrote the article. EM conceived the study, performed the analysis, and revised the manuscript. LB prepared exercises for the course, tutored students in the class whose data were analyzed, conceived and performed the analysis, and revised the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior–Brasil (CAPES)–Grant 51/2013 – 23038.004007/2014-82; Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). The funding body did not play any roles in the design of the study, collection, analysis, or interpretation of data, or in writing the manuscript.

ACKNOWLEDGMENTS

We would like to thank Luciana Carvalho for her great support with Moodle from the Department of Computer Science at the Federal University of Minas Gerais.

REFERENCES

Agonács, N., and Matos, J. F. (2020). Os cursos on-line abertos e massivos (mooc) como ambientes heutagógicos heutagógicos. *Revista Brasileira de Estudos Pedagógicos* 101, 17–35. doi: 10.24109/2176-6681.rbep.101i257.4329

Berland, M., Martin, T., Benton, T., Petrick Smith, C., and Davis, D. (2013). Using learning analytics to understand the learning pathways of novice programmers. *J. Learn. Sci.* 22, 564–599. doi: 10.1080/10508406.2013.836655

Blikstein, P. (2011). "Using learning analytics to assess students' behavior in open-ended programming tasks," in *Proceedings of the 1st International Conference on Learning Analytics and Knowledge*, 110–116.

Blikstein, P., Worsley, M., Piech, C., Sahami, M., Cooper, S., and Koller, D. (2014). Programming pluralism: using learning analytics to detect patterns in the learning of computer programming. *J. Learn. Sci.* 23, 561–599. doi: 10.1080/10508406.2014.954750

Breiman, L. (2001). Random forests. *Mach. Learn.* 45, 5–32. doi: 10.1023/A:1010933404324

Carroll, J. D., and Arabie, P. (1998). "Multidimensional scaling," in *Measurement, Judgment and Decision Making*, 179–250.

Chang, R. L., and Pavlidis, T. (1977). Fuzzy decision tree algorithms. *IEEE Trans. Syst. Man Cybern.* 7, 28–35. doi: 10.1109/TSMC.1977.4309586

Csikszentmihalyi, M., Abuhamdeh, S., and Nakamura, J. (2014). "Flow," in *Flow and the Foundations of Positive Psychology* (Springer), 227–238.

de Melo-Minardi, R. C., and Bastos, L. L. (2021). Expandindo as paredes da sala de aula: aprendizados com o ensino a distância e ensino remoto emergencial. *Revista da Universidade Federal de Minas Gerais* 28, 106–125. doi: 10.35699/2316-770X.2021.29089

Dhawan, S. (2020). Online learning: a panacea in the time of covid-19 crisis. *J. Educ. Technol. Syst.* 49, 5–22. doi: 10.1177/0047239520934018

Fernandez-Medina, C., Pérez-Pérez, J. R., Álvarez-García, V. M., and Paule-Ruiz, M. d. P. (2013). "Assistance in computer programming learning using educational data mining and learning analytics," in *Proceedings of the 18th ACM Conference on Innovation and Technology in Computer Science Education*, 237–242.

Friedman, J. H. (2002). Stochastic gradient boosting. *Comput. Stat. Data Anal.* 38, 367–378. doi: 10.1016/S0167-9473(01)00065-2

Fu, X., Shimada, A., Ogata, H., Taniguchi, Y., and Suehiro, D. (2017). "Real-time learning analytics for c programming language courses," in *Proceedings of the Seventh International Learning Analytics and Knowledge Conference*, 280–288.

Gerstein, J. (2014). Moving from education 1.0 through education 2.0 towards education 3.0.

Grover, S., Basu, S., Bienkowski, M., Eagle, M., Diana, N., and Stamper, J. (2017). A framework for using hypothesis-driven approaches to support data-driven learning analytics in measuring computational thinking in block-based programming environments. *ACM Trans. Comput. Educ.* 17, 1–25. doi: 10.1145/3105910

Higashi, R. M., Schunn, C. D., and Flot, J. B. (2017). Different underlying motivations and abilities predict student versus teacher persistence in an online course. *Educ. Technol. Res. Dev.* 65, 1471–1493. doi: 10.1007/s11423-017-9528-z

Johnson, D., and Palmer, C. C. (2015). Comparing student assessments and perceptions of online and face-to-face versions of an introductory linguistics course. *Online Learn.* 19, n2. doi: 10.24059/olj.v19i2.449

Jordan, K. (2015). Massive open online course completion rates revisited: Assessment, length and attrition. *Int. Rev. Res. Open Distribut. Learn.* 16, 341–358. doi: 10.19173/irrodl.v16i3.2112

Kampff, A. J. C. (2009). Mineração de dados educacionais para geração de alertas em ambientes virtuais de aprendizagem como apoio à prática docente.

Keats, D., and Schmidt, J. P. (2007). The genesis and emergence of education 3.0 in higher education and its potential for africa. *First Monday* 12, 3–5. doi: 10.5210/fm.v12i3.1625

Kent, J. T. (1983). Information gain and a general measure of correlation. *Biometrika* 70, 163–173. doi: 10.1093/biomet/70.1.163

Lengel, J. G. (2013). *Education 3.0: Seven Steps to Better Schools*. Teachers College Press.

Noble, W. S. (2006). What is a support vector machine? *Nat. Biotechnol.* 24, 1565–1567. doi: 10.1038/nbt1206-1565

Pereira, F. D., Oliveira, E. H., Oliveira, D. B., Cristea, A. I., Carvalho, L. S., Fonseca, S. C., et al. (2020). Using learning analytics in the amazonas: understanding students' behaviour in introductory programming. *Br. J. Educ. Technol.* 51, 955–972. doi: 10.1111/bjet.12953

Peterson, L. E. (2009). K-nearest neighbor. *Scholarpedia* 4, 1883. doi: 10.4249/scholarpedia.1883

Siemens, G., and Baker, R. S. d. (2012). "Learning analytics and educational data mining: towards communication and collaboration," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge*, 252–254.

Conflict of Interest: EM was employed by Take Blip.

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.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

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