Teaching in a Time of Crisis
DOI: https://doi.org/10.1128/jmbe.v22i1.2315

JMBE Journal of Microbiology & Biology Education

Gene Expression and Data Analysis Pipeline Using Cancer BioPortal in the Classroom†

Chassidy N. Barnes[1], Blake P. Johnson[1], Stefanie W. Leacock[2], Ruben M. Ceballos[3], Lori L. Hensley[4], and Nathan S. Reyna[1]*
[1]*Department of Biology, Ouachita Baptist University, Arkadelphia, AR 71929;*
[2]*Department of Biological Sciences, University of Arkansas at Little Rock, Little Rock, AR 72204;*
[3]*Department of Biological Sciences, University of Arkansas, Fayetteville, AR 72701;*
[4]*Department of Biological Sciences, Jacksonville State University, Jacksonville, AL 36265*

At institutions with an emphasis on authentic research experiences as an integral part of the biology curriculum, COVID created a huge challenge for course instructors whose learning objectives were designed for such experiences. Moving such laboratory experiences online when remote learning became necessary has resulted in a new model for CUREs that utilizes free online databases to provide not only a novel research experience for students, but also the opportunity to engage in big data analysis. Cancer BioPortal (cBioPortal) is an open-access collective cancer research resource for storing and exploring clinical, genomic, proteomic, and transcriptomic data. cBioPortal eliminates the computational barrier of interpreting complex genomic data by providing easily understandable visualization that can be interpreted and translated into relevant biological insights. Because no prior computational knowledge is required, cBioPortal is an ideal educational tool for either in-person or distance learning environments. We developed a pedagogical approach, video tutorials, and data analysis workflows centered on using cBioPortal. Pedagogically, students develop an initial research outline that is continually updated and graded throughout the project. Progress during the project or course is assessed by a series of student presentations that are 5 to 15 minutes in length and are aimed at explaining the approach used in data acquisition, interpretation of the data, and relevance to the initial hypothesis. While cancer-specific, this analysis platform appeals to a wide range of classes and student interests. Further, the project has been successfully done both as an independent research experience and as part of a virtual class-based research project.

INTRODUCTION

The development of publicly available research databases is changing the way scientists approach bioinformatics and data analysis. As a result of COVID-19, we used these databases as a means for online instruction, allowing our students to interact remotely and conduct virtual research either independently or as part of class-based research projects. The cBioPortal for Cancer Genomics (https://www.cbioportal.org), known simply as cBioPortal, is an open-access collective cancer research resource for storing and exploring clinical, genomic, proteomic, and transcriptomic data across hundreds of cancer studies and includes thousands of individual samples (1). A graphical user interface (GUI) eliminates the barrier of having to work meticulously through large and complex genomic datasets (2). Readily interpretable visualizations provide robust data analysis leading to biological insights (1). Such insights permit the formulation of novel hypotheses by research scientists and students alike.

While cBioPortal is used extensively in cancer research (3, 4), streamlined workflows for data collection and analysis for junior researchers or those new to the database are currently unavailable. Our first observation when transitioning to remote learning was that, despite the utility of cBioPortal, the development of efficient workflows related to undergraduate-level training was needed.

Because no prior computational (e.g., programming) knowledge is required to use the database, we found cBioPortal to be an ideal educational tool for remote learning or a virtual lab research project. cBioPortal gives students access to data to analyze and interpret. To facilitate com-

*Corresponding author. Mailing address: Department of Biology, Ouachita Baptist University, 410 Ouachita St., Arkadelphia, AR 71998. Phone: 434-223-6175. Email: reynan@OBU.edu.
Received: 25 September 2020, Accepted: 17 December 2020, Published: 31 March 2021
†Supplemental materials available at http://asmscience.org/jmbe

petency in using cBioPortal, we developed a pedagogical workflow, including short video tutorials, that is amenable to remote learning. This workflow guides students through the process of accessing data from multiple published cancer studies (e.g., PanCancer Studies) and analyzing the data in the context of questions related to gene expression, gene mutations, and copy number.

PROCEDURE

Student groups of two or three are tasked with developing a hypothesis and then using cBioPortal to extract, analyze, and interpret data that either support or refute their hypothesis. Pedagogically, students develop an initial research outline that is continually updated and graded throughout the project (Fig. 1). Progress during the project or course is assessed by a series of student presentations (5 to 15 minutes in length) aimed at explaining the approach used in data acquisition, data interpretation, and relevance to the initial hypothesis. Often each presentation includes a literature review conducted outside of cBioPortal.

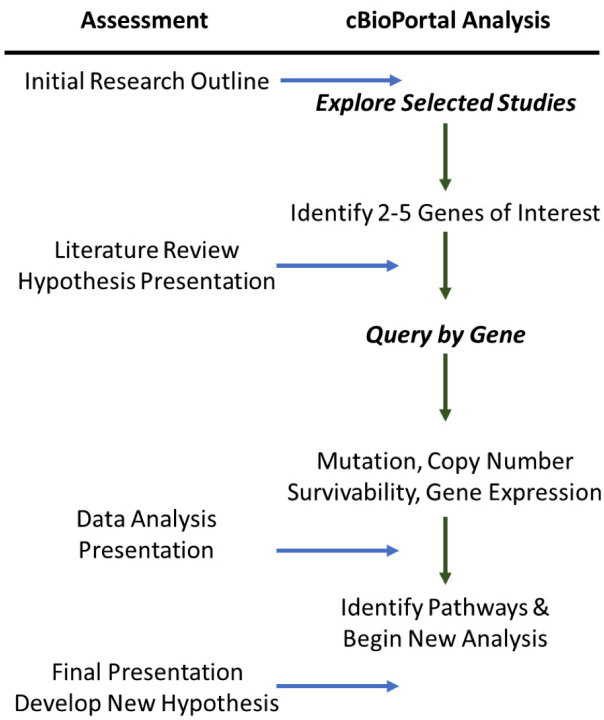

FIGURE 1. Assessment and cBioPortal analysis pipeline. Vertical arrows indicate the required analysis steps in cBioPortal. Horizontal arrows indicate assessment requirements for each step.

Two options for data acquisition are available in cBioPortal: "Query by Gene" and "Explore Selected Studies." The analysis pipeline we developed was adapted from previously published protocols (4) in which these two options were merged into one pathway, making data collection and analysis more efficient (Fig. 1). As proof of concept, students conducted an analysis of cancer-related genes to assess whether mutations and interactions among these genes lead to breast cancer liver metastasis (BCLM). Genes pertinent to breast cancer were first identified using the Explore Selected Studies pathway. These genes were verified through a review of the literature, and five candidate genes were chosen for further analysis using the Query by Gene pathway (Fig. 1).

Query by Gene

The Query by Gene function (Appendix 1) generates a series of tabs specific for a queried gene. The OncoPrint tab provides a graphical summary of the genetic alterations present in each queried gene (Appendix 2). The mutations tab allows students to view genetic alterations and determine whether the genetic alterations of their genes are mutually exclusive or co-occur (Appendices 4 and 5). The plots tab shows a box and whisker plot of mRNA expression of the queried genes and their respective copy number alterations (Appendix 5). Other tabs show data on mutations and amino acid alterations and genetic pathways.

One disadvantage of the Query by Gene pathway is that students may not know candidate genes to query when they begin their exploration. Additionally, a limited query may miss information about relevant genes. These challenges are overcome through the portal's second pathway of analysis, Explore Selected Studies.

Explore Selected Studies

The Explore Selected Studies option provides a summary containing charts and tables listing molecular data, such as expression and copy number variations, of significantly impacted genes (Appendix 5). This immense dataset also includes clinical data related to patient survival rate, patient age, and stage of cancer at the time of diagnosis. Each data set is hyperlinked, allowing for further analysis. While the Explore Selected Studies query provides clinical and genomic data, there is no clear connection as to how the data presented relate to the development and progression of different cancer types. The use of this pathway alone is insufficient to obtain a complete understanding of complex cancer development and progression pathways.

Class data analysis pipeline combines the two approaches

The students' own hypotheses directed the analysis. For example, a student identified genes that are differentially expressed in noninvasive versus invasive breast cancer and searched for these same genes in hepatocellular carcinoma–related pathways to see their relevance to liver cancer. The student hypothesized that commonalities in gene expression for both cancer types could provide insight on the survival and proliferation of breast cancer cells in the liver. The

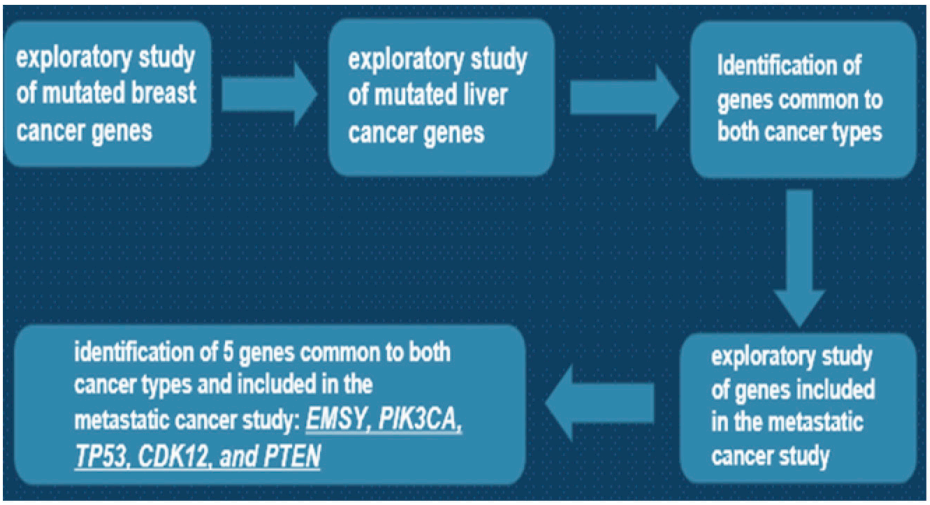

FIGURE 2. Slide from student presentation, outlining the process for the identification of target BCLM genes. The example shows the pathway for identification of target BCLM genes using the "Explore Selected Studies" pathway of analysis. Using this pathway, *EMSY*, *PIK3CA*, *TP53*, *CDK12*, and *PTEN* were identified as genes of interest entered to run a query.

analysis was then narrowed to focus on cases where breast cancer (primary diagnosis) metastasized to the liver (secondary). The student identified a novel set of target genes expressed in both breast and liver cancer. These data were presented as part of a class research presentation (Fig. 2). More examples of student-generated content are included as Appendices 2 to 9.

Student video tutorials

While written protocols and how-to videos are available for cBioPortal, they are long (40 to 60 min) and difficult for undergraduate students to follow (1, 4, 5). Despite this, students in the first group to move online as a result of COVID watched these videos prior to designing their research questions. Subsequently, we addressed the need for shorter, undergraduate-level tutorials by generating three short student-friendly video tutorials that allow students to integrate both approaches more quickly: (1) *Introduction to cBioPortal* (https://youtu.be/fsmjcu3VXf0) is an overview and explains how to navigate each tab; (2) *Searching the cBioPortal Database* (https://youtu.be/6Y5G49RDyKs) is a walkthrough tutorial on running a query; and (3) *Interpreting results on cBioPortal* (https://youtu.be/WNSfh61BLg8) explains how to interpret and refine results. Videos are publicly available on the Cell Biology Education Consortium's (CBEC) YouTube channel (https://www.youtube.com/c/CellBiologyEducationConsortium). Because of the unexpected timing of COVID, we did not initially use a rubric for cBioPortal projects. Students are now required to watch these videos prior to the design of their research questions. Though no formal assessment of this assignment was undertaken, anecdotal evidence from student presentations demonstrated that students who had not watched the video were less able to communicate the significance of their projects than students who had. Additionally, students who watched the videos seemed more confident in their ability to navigate cBioPortal than those who did not.

CONCLUSION

Cancer BioPortal is a publicly accessible data analysis interface that does not require prior computational knowledge. Our pedagogical and data analysis workflows and short video tutorials make cBioPortal an ideal resource for the undergraduate classroom. This approach can be used as an independent student project or as part of a class-embedded research project. Importantly, this pedagogy transitions easily to an online learning experience while still providing an authentic research experience and is flexible enough to accommodate students moving from face-to-face instruction to online instruction and back again as COVID causes students to isolate or quarantine. We have found that exploring cancer biology appeals to many undergraduates because of its relevance to personal health and medical careers. Students encounter topics in cancer biology at all levels of biology courses. Students in genetics courses can research and compare specific genes or cancer types. Cell or developmental biology students might search for genes in pathways known to affect cell function or differentiation. Biochemistry or molecular biology students could search for mutations in known drug targets. Analyses such as these push students toward higher-level thinking by challenging them to connect concepts read in the literature and develop critical thinking skills.

SUPPLEMENTAL MATERIAL

Appendix 1: Performing a query
Appendix 2: An OncoPrint for genetic alterations of EMSY, PIK3CA, PTEN, TP53, and CDK12
Appendix 3: Cancer types summary for queried genes
Appendix 4: Plots for copy number alterations vs. mRNA expression of *EMSY*, *PIK3CA*, *PTEN*, *TP53*, and *CDK12*
Appendix 5: Mutual exclusivity for queried genes
Appendix 6: Explore Select Study view
Appendix 7: CN segments for EMSY, PIK3CA, TP53, PTEN, and CDK12
Appendix 8: Mutation charts of EMSY, PTEN, TP53, CDK12, and PIK3CA
Appendix 9: Signaling pathways for the queried genes

ACKNOWLEDGMENTS

This project was developed and supported by the Cell Biology Education Consortium (Award# 1827066), an NSF-funded Research Collaborative Network for Undergraduate Biology Education (RCN-UBE). A portion of the research was done in partnership with the NSF-REU Indigenous America to Indigenous Mekong—Adventures in Biology and Biodiversity (Award #1659858). We thank our NSF program officers for allowing us to pilot virtual undergraduate research experiences in summer of 2020. The authors have no conflicts of interest to declare.

REFERENCES

1. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E, Antipin Y, Reva B, Goldberg AP, Sander C, Schultz N. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2:401–404.
2. Buechner P, Hinderer M, Unberath P, Metzger P, Boeker M, Acker T, Haller F, Mack E, Nowak D, Paret C, Schanze D, von Bubnoff N, Wagner S, Busch H, Boerries M, Christoph J. 2020. Requirements analysis and specification for a molecular tumor board platform based on cBioPortal. Diagnostics (Basel) 10(2):93.
3. Jiao XD, Qin BD, You P, Cai J, Zang YS. 2018. The prognostic value of TP53 and its correlation with EGFR mutation in advanced non-small cell lung cancer, an analysis based on cBioPortal data base. Lung Cancer 123:70–75.
4. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E, Cerami E, Sander C, Schultz N. 2013. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 6:pl1.
5. Unberath P, Knell C, Prokosch HU, Christoph J. 2019. Developing new analysis functions for a translational research platform: extending the cBioPortal for cancer genomics. Stud Health Technol Inform 258:46–50.