INTRODUCTION

As data become a dominant feature of biology (1), university curricula must adapt and prepare their students for a data-centric future (2). Many life science undergraduates do not receive bioinformatics training, and a majority of professional biologists gain bioinformatics skills without institutional training (3). This is often because scientific computing courses are designed for computer science and engineering majors (4). Beyond simple introductory courses, there is a need to develop advanced curricula which truly integrate computational and biological perspectives (reviewed in reference 5) as either life science electives or a formal bioinformatics degree (6, 7). A key design strategy is identifying authentic, biologically motivated computational courses, including hands-on exercises with real data (8).

As an emerging discipline, bioinformatics is evolving rapidly. Early coursework and textbooks for bioinformatics focused on algorithms, e.g., Rosalind (http://rosalind.info/). However, the rise of deep genomic and proteomic profiles has highlighted a need for instruction in exploratory data science and statistics. Massive public data sets offer an incredible opportunity to integrate molecular and cellular biology with computation (9), and published classroom activities exist for a variety of topics, including DNA assembly (10), metagenomics (11), bacteriophage genomics (12), and viral genomics (13). Unfortunately, current topics largely focus on genomics and not a broader genotype-to-phenotype view of biology. Additionally, many public data sets have practical barriers that complicate classroom use, such as data accessibility and package installation (14). Here, we present a cancer bioinformatics module with six exercises to teach an integrated proteogenomic view of cancer. The exercises contain explanatory text, code demonstrations, and practice problems encoded in publicly available Google Colab notebooks intended for online or traditional instruction.

PROCEDURE

Computational cancer biology module

Improvements in cancer diagnosis and treatment will come from analyzing DNA, RNA, and protein data, along with traditional clinical information (15). We present a module for teaching computational cancer biology to demonstrate how modern omics data are used in precision oncology. Each lesson explores fundamental cancer concepts and supporting molecular data (Table 1). In lesson one, students are introduced to the proteogenomic data set (16–19). As DNA mutation is the root cause of cancer, lessons two through four describe three distinct kinds of mutations and their impacts on the genome. Lessons five and six use transcriptomics and proteomics data to identify differential gene expression and activated pathways. Each lesson also teaches computational tools and analytical techniques essential for modern cancer research. The first lesson introduces pandas DataFrames and manipulation of data matrices. Subsequent lessons expand the students’ skill set with DataFrames, application programming interfaces (APIs), statistical methods, graphing, and network exploration.

Overcoming instructional barriers

The module is designed to overcome both practical and philosophical barriers to student engagement with the integration of three computational tools. First, lessons are coded in Google Colab, a live Python coding environment accessible via a Web browser (https://colab.research.google.com/). Importantly, Colab does not require any software installation and works with all operating systems, including mobile platforms. Thus, students have access to a functional computing environment on the first day of class, without requiring information technology.
(IT) support. Second, real cancer data sets are streamed via the cptac Python API (20), providing access to genomic, transcriptomic, proteomic, and clinical data. Mundane and frustrating tasks like locating files or loading data are automated, thus saving time for students to focus on cancer-relevant computation.

Third, our module uses an autograder, which contains homework answers and hints used to give students immediate feedback and provide help on integrated practice problems.

Student preparation

The module is designed for upper-division undergraduates. For biological preparation, we suggest that students have completed courses in molecular biology and genetics; for computational preparation, we suggest that students have completed courses in Python programming and introductory statistics. The first lesson, called Introduction to Cancer Datasets, is a skills check. A prepared student should be able to complete it in about 30 to 45 min. This lesson is the same as the final homework in the author’s (S.H.P) Introduction to Bioinformatics course, and students with one semester of programming would be prepared computationally. If students have difficulty with this first assignment, then it is not advised to continue with the module.

Using the module in the classroom

We created this module for those wishing to teach computational cancer biology or expand their offering of bioinformatics topics to undergraduate life science majors. The module and links to the six lessons are available at https://paynelab.github.io/biograder/bio462. Each lesson (except for Lesson 1: Introduction to Cancer Datasets) is designed to contain a week’s worth of instruction material, with approximately 6 h of computational exercises. Although each lesson is independent, the later lessons build on the computational skills from earlier lessons. Therefore, the module is intended to be used as a set.

Lessons are delivered through Google Colab notebooks and interweave descriptive text and software code. Notebooks begin with a section describing the topic and relevant literature, which could be used in classroom lectures or preparatory reading. The explanatory text motivates students with the description of a cancer biology problem and then proceeds with demonstrative software examples that teach students core computing concepts. These are followed by practice problems where students write their own code. Each practice problem is graded with the autograder, and students can ask for hints on how to approach the question. As with any course, it is advised that a teaching assistant with a

| Lesson topic               | Biological learning outcomes                                                                 | Computational learning outcomes                                                                 |
|----------------------------|------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| 1. Introduction to Cancer Datasets | • Locate proteogenomic cancer datasets from real tumor samples                                | • Manipulate Pandas DataFrame                                                                     |
|                            |                                                                                                | • Select relevant clinical cancer data                                                            |
| 2. Missense Mutation       | • Identify frequent mutations in cancer cohorts                                             | • Access UniProt knowledgebase using API                                                           |
|                            | • Compare and contrast different cancer types                                               | • Parse JSON and integrate cancer data with UniProt                                              |
|                            | • Assess functional impact of DNA mutation                                                  |                                                                                                |
| 3. Truncation Mutation     | • Classify mutation impact on primary sequence, domain structure, and protein function     | • Create new DataFrame elements with \_apply() function                                           |
|                            |                                                                                                | • Access UniProt knowledgebase using API                                                           |
|                            |                                                                                                | • Import/export data from Colabs                                                                  |
| 4. Copy Number Variation   | • Identify copy number variation (CNV) changes in genes                                     | • Plot CNV events                                                                                  |
|                            | • Distinguish between focal and arm level CNV events                                        | • Integrate gene location and CNV data                                                            |
|                            | • Compare frequent CNVs in multiple cancer types                                           | • Create new DataFrame elements with \_apply() function                                           |
| 5. Transcriptomics         | • Identify differential expression of transcripts between tumor and normal samples         | • Perform pathway enrichment analysis                                                             |
|                            | • Interpret a gene list as a functional set of pathways                                     | • Create boxplot visualization                                                                    |
|                            | • Justify choice of enrichment methods based on gene list characteristics                   | • Perform unpaired t test with multiple hypothesis correction                                      |
| 6. Proteomics              | • Evaluate utility of protein coexpression networks based on agreement with known protein interaction networks | • Perform correlation analysis                                                                  |
|                            |                                                                                                | • Create network visualization                                                                     |
|                            |                                                                                                | • Access UniProt knowledgebase using API                                                           |
bioinformatics background be available to guide students through the rigorous lessons. Instructors interested in the answer key may contact the authors.

**Student feedback**

Student evaluation of the lessons and module contained both supportive and critical feedback. In general, students were enthusiastic about the authentic integration of computing and cancer biology. A significant positive of the lessons was the links to articles that prompted a deeper exploration of biological concepts and a reference for algorithm use or syntax. The most common challenge was learning to work with data matrices; for students who learned programming through the object-oriented lens, computational exploration of DataFrames was an adjustment. Student feedback on specific exercises prompted substantial revision in wording, hints, and primer examples.

**Safety issues**

This module is purely computational and does not include any laboratory components. The data used in the modules are publicly available and do not include any personally identifiable information.

**Conclusion**

Cancer bioinformatics is a growing area, and proteogenomic analyses of large cancer cohorts are rapidly changing cancer diagnosis and treatment. Many of these rich data sets are publicly available, enabling the creation of engaging and impactful curricula. We created a module that introduces undergraduate students to proteogenomic cancer data and teaches bioinformatics skills in data analysis and interpretation. By utilizing a unique combination of technologies, the module avoids common obstacles to large-scale and authentic bioinformatics exercises in a classroom setting.

**ACKNOWLEDGMENTS**

We thank members of the Payne laboratory and global collaborators who tested these lessons and provided feedback.

This work was supported by a National Cancer Institute (NCI) CPTAC award (grant number U24 CA210972) and by the Simmons Center for Cancer Research.

We declare no conflicts of interest.

**REFERENCES**

1. Barone L, Williams J, Micklos D. 2017. Unmet needs for analyzing biological big data: a survey of 704 NSF principal investigators. PLoS Comput Biol 13:e1005755. https://doi.org/10.1371/journal.pcbi.1005755.
2. Wilson Sayres MA, Hauser C, Sierk M, Robic S, Rosenwald AG, Smith TM, Triplett EW, Williams JJ, Dinsdale E, Morgan WR, Burnette JM, Donovan SS, Drew JC, Elgin SCR, Fowlks ER, Galindo-Gonzalez S, Goodman AL, Grandgenett NF, Goller CC, Jungck JR, Newman JD, Pearson W, Ryder EF, Tosado-Acevedo R, Tapprich W, Tobin TC, Toro-Martinez A, Welch LR, Wright R, Barone L, Ebenbach D, McWilliams M, Olney KC, Pauley MA. 2018. Bioinformatics core competencies for undergraduate life sciences education. PLoS One 13:e0196878. https://doi.org/10.1371/journal.pone.0196878.
3. Atwood TK, Blackford S, Brazas MD, Davies A, Schneider MV. 2019. A global perspective on evolving bioinformatics and data science training needs. Brief Bioinform 20:398–404. https://doi.org/10.1093/bib/bbx100.
4. Qin H. 2009. Teaching computational thinking through bioinformatics to biology students. ACM SIGCSE Bull 41:188–191. https://doi.org/10.1145/1539024.1508932.
5. Magana AJ, Taleyarkhan M, Alvarado DR, Kane M, Springer J, Clase K. 2014. A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research. CBE Life Sci Educ 13:607–623. https://doi.org/10.1187/cbe.13-10-0193.
6. Altman RB. 1998. A curriculum for bioinformatics: the time is ripe. Bioinformatics 14:549–550. https://doi.org/10.1093/bioinformatics/14.7.549.
7. Welch L, Lewitter F, Schwartz R, Brooksbank C, Radijovic P, Gaeta B, Schneider MV. 2014. Bioinformatics curriculum guidelines: toward a definition of core competencies. PLoS Comput Biol 10:e1003496. https://doi.org/10.1371/journal.pcbi.1003496.
8. Maloney M, Parker J, LeBlanc M, Woodard CT, Glackin M, Hanrahan M. 2010. Bioinformatics and the undergraduate curriculum. CBE Life Sci Educ 9:172–174. https://doi.org/10.1187/cbe.10-03-0038.
9. Hankey W, Zhangi N, Crow MM, Dow WH, Kratz A, Robinson AM, Robinson MR, Segarra VA. 2020. Using the Cancer Genome Atlas as an inquiry tool in the undergraduate classroom. Front Genet 11:573992. https://doi.org/10.3389/fgene.2020.573992.
10. Jensen PA. 2017. Hands-on assembly of DNA sequencing reads as a gateway to bioinformatics. J Microbiol Biol Educ 18:18.2.34. https://doi.org/10.1128/jmbe.v18i2.1295.
11. Ditty JL, Kvaal CA, Goodner B, Freyermuth SK, Bailey C, Britton RA, Gordon SG, Heinhorst S, Reed K, Xu Z, Sanders-Lorenz ER, Axen S, Kim E, Johns M, Scott K, Kerfeld CA. 2010. Incorporating genomics and bioinformatics across the life sciences curriculum. PLoS Biol 8:e1000448. https://doi.org/10.1371/journal.pbio.1000448.
12. Jordan TC, Burnett SH, Carson S, Caruso SM, Clase K, Dejong RJ, Denney JJ, Denver DR, Dunbar D, Elgin SCR, Findley AM, Gissendanner CR, Golebiowska UP, Guild N, Hartzog GA, Grillo WH, Hollowell GP, Hughes LE, Johnson A, King RA, Lewis LO, Li W, Rosenzweig F, Rubin MR, Saha MS, Sandoz J, Shaffer CD, Taylor B, Temple L, Vazquez E, Ware VC, Barker LP, Bradley KW, Jacobs-Sera D, Pope WH, Russell DA,
Cresawn SG, Lopatto D, Bailey CP, Hatfull GF. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio 5: e01051-13. https://doi.org/10.1128/mBio.01051-13.

13. Steel JJ. 2021. Genome analysis of SARS-CoV-2 case study: an undergraduate online learning activity to introduce bioinformatics, BLAST, and the power of genome databases. J Microbiol Biol Educ 22:221.1.16. https://doi.org/10.1128/jmbe.v22i1.2245.

14. Learned K, Durbin A, Currie R, Kephart ET, Beale HC, Sanders LM, Pfeil J, Goldstein TC, Salama SR, Haussler D, Vaske OM, Bjork IM. 2019. Barriers to accessing public cancer genomic data. Sci Data 6:98. https://doi.org/10.1038/s41597-019-0096-4.

15. Rodriguez H, Zenklusen JC, Staudt LM, Doroshow JH, Lowy DR. 2021. The next horizon in precision oncology: proteogenomics to inform cancer diagnosis and treatment. Cell 184:1661–1670. https://doi.org/10.1016/j.cell.2021.02.055.

16. Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, da Veiga Leprevost F, Reva B, Lih T-SM, Chang H-Y, Ma W, Huang C, Ricketts CJ, Chen L, Krek A, Li Y, Rykunov D, Li QK, Chen LS, Ozbek U, Vasaikar S, Wu Y, Yoo S, Chowdhury S, Wyczalkowski MA, Ji J, Schnaubelt M, Kong A, Sethuraman S, Avtonomov DM, Ao M, Colaprico A, Cao S, Cho K-C, Kalayci S, Ma S, Liu W, Ruggles K, Calinawan A, Gümüş ZH, Geiszler D, Kawaler E, Teo GC, Wen B, Zhang Y, Keegan S, Li K, Chen F, Edwards N, Pierorazio PM, Clinical Proteomic Tumor Analysis Consortium, et al. 2019. Integrated proteomic characterization of clear cell renal cell carcinoma. Cell 179:964–983.e31. https://doi.org/10.1016/j.cell.2019.10.007.

17. Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, Petralia F, Li Y, Lih T-SM, Reva B, Krek A, Ji J, Song X, Liu W, Hong R, Yao L, Blumenberg L, Savage SR, Wendt MC, Wen B, Li K, Tang LC, MacMullan MA, Avanessian SC, Kane MH, Newton CJ, Cornwell M, Kothadia RB, Ma W, Yoo S, Mannan R, Vats P, Kumar-Sinha C, Kawaler EA, Omelchenko T, Colaprico A, Geffen Y, Maruva YE, da Veiga Leprevost F, Wiznerowicz M, Gümüş ZH, Veluswamy RR, Hostetter G, Heiman DJ, Wyczalkowski MA, Hiltke T, Mesri M, Kinsinger CR, Boja ES, Omenn GS, Chinnaiyan AM, Rodriguez H, Li QK, Clinical Proteomic Tumor Analysis Consortium, et al. 2020. Proteogenomic characterization reveals therapeutic vulnerabilities in lung adenocarcinoma. Cell 182:200–225.e35. https://doi.org/10.1016/j.cell.2020.06.013.

18. Dou Y, Kawaler EA, Cui Zhou D, Gritsenko MA, Huang C, Blumenberg L, Karpova A, Petyuk VA, Savage SR, Satpathy S, Liu W, Wu Y, Tsai C-F, Wen B, Li Z, Cao S, Moon J, Shi Z, Cornell M, Wyczalkowski MA, Chu RK, Vasaikar S, Zhou H, Gao Q, Moore RJ, Li K, Sethuraman S, Monroe ME, Zhao R, Heiman D, Krug K, Clauser K, Kothadia R, Maruva Y, Pico AR, Oliphant AE, Hoskins EL, Pugh SL, Beecroft SJL, Adams DW, Jarman JC, Kong A, Chang H-Y, Reva B, Liao Y, Rykunov D, Colaprico A, Chen XS, Czekanski A, Jedyryka M, Clinical Proteomic Tumor Analysis Consortium, et al. 2020. Proteogenomic characterization of endometrial carcinoma. Cell 180:729–748.e26. https://doi.org/10.1016/j.cell.2020.01.026.

19. Wang L-B, Karpova A, Gritsenko MA, Kyle JE, Cao S, Li Y, Rykunov D, Colaprico A, Rothstein JH, Hong R, Stathias V, Cornell M, Petralia F, Wu Y, Reva B, Krug K, Pugliese P, Kawaler E, Olsen LK, Liang W-W, Song X, Dou Y, Wendt MC, Caravan W, Liu W, Cui Zhou D, Ji J, Tsai C-F, Petyuk VA, Moon J, Ma W, Chu RK, Weitz KK, Moore RJ, Monroe ME, Zhao R, Yang X, Yoo S, Krek A, Demopoulos A, Zhu H, Wyczalkowski MA, McMichael JF, Henderson BL, Lindgren CM, Boekweg H, Lu S, Baral J, Yao L, Stratton KG, Clinical Proteomic Tumor Analysis Consortium, et al. 2021. Proteogenomic and metabolomic characterization of human glioblastoma. Cancer Cell 39:509–528.e20. https://doi.org/10.1016/j.ccell.2021.01.006.

20. Lindgren CM, Adams DW, Kimball B, Boekweg H, Tayler S, Pugh SL, Payne SH. 2021. Simplified and unified access to cancer proteogenomic data. J Proteome Res 20:1902–1910. https://doi.org/10.1021/acs.jproteome.0c00919.