AMERICAN SOCIETY FOR MICROBIOLOGY | JMBE Journal of Microbiology & Biology Education

Check for updates

| Curriculum

A cost-free CURE: using bioinformatics to identify DNA-binding factors at a specific genomic locus

Casey A. Schmidt,[1] Lauren J. Hodkinson,[2] H. Skye Comstra,[1] Samia Khan,[1] Henrik Torres,[3] Leila E. Rieder[1,2]

AUTHOR AFFILIATIONS See affiliation list on p. 12.

ABSTRACT Research experiences provide diverse benefits for undergraduates. Many academic institutions have adopted course-based undergraduate research experiences (CUREs) to improve student access to research opportunities. However, potential instructors of a CURE might still face financial or practical hurdles that prevent implementation. Bioinformatics research offers an alternative that is free, safe, compatible with remote learning, and may be more accessible for students with disabilities. Here, we describe a bioinformatics CURE that leverages publicly available datasets to discover novel proteins that target an instructor-determined genomic locus of interest. We use the free, user-friendly bioinformatics platform Galaxy to map ChIP-seq datasets to a genome, which removes the computing burden from students. Both faculty and students directly benefit from this CURE, as faculty can perform candidate screens and publish CURE results. Students gain not only basic bioinformatics knowledge, but also transferable skills, including scientific communication, database navigation, and primary literature experience. The CURE is flexible and can be expanded to analyze different types of high-throughput data or to investigate different genomic loci in any species.

KEYWORDS CURE, bioinformatics, transcription factor, DNA-binding protein, ChIP-seq, genetics, undergraduate research

Undergraduate research experiences are invaluable to students. Documented benefits include retention in STEM (1), increased confidence in research abilities (2), and inclusion of underrepresented populations (3). Yet many students struggle to find a space in laboratories already at capacity. Course-based undergraduate research experiences (CUREs) can remedy this problem, as they offer students authentic research experiences within the context of a classroom (4). Not only do CUREs involve many more undergraduates in research than the traditional "apprentice" model, but also they allow faculty (especially those with high teaching responsibilities) to make research progress. For example, the instructor of a CURE course can perform a screen (5, 6), follow-up on an interesting result from their laboratory (7), or increase the rigor and reproducibility of a research project through replication by different laboratory groups or sections.

Despite these clear benefits, there are often limitations to running bench-based CUREs. For example, large schools with high enrollment might face space and time constraints. In addition, the materials required to perform wet-laboratory experiments may be expensive and time consuming to prepare for large classes. Overall, these and other limitations can be prohibitive to implement wet-laboratory CUREs (8).

Bioinformatics CUREs can skirt these hurdles. Because laboratory space is not necessary, the class can be held in a computer laboratory, a classroom (if the students have access to personal laptops), or completely virtually. There are no costly reagents to purchase or biohazard concerns. Bioinformatics research can offer students with disabilities a less physically demanding alternative to bench-based experiments. It is also

Editor Samiksha Raut, The University of Alabama at Birmingham, Birmingham, Alabama, USA

Address correspondence to Leila E. Rieder, lrieder@emory.edu, or Casey A. Schmidt, casey.schmidt@emory.edu.

The authors declare no conflict of interest.

Received 20 July 2023
Accepted 14 September 2023
Published 24 October 2023

compatible with remote or asynchronous teaching, which became necessary during the early COVID-19 pandemic (9, 10).

Although bioinformatics research is typically performed on expensive computing clusters, we instead use Galaxy (11), which is a free, user-friendly platform that integrates many widely used bioinformatics tools. All memory-intensive computing is performed on Galaxy's servers, allowing students to simply set up commands, execute, and log off; neither sophisticated programming knowledge nor computing power is needed. Bioinformatics research is easily integrated into students' busy schedules, and each activity can typically be completed in less time than a traditional 3-hour wet laboratory. Students participating in bioinformatics CUREs report high sense of achievement and high levels of satisfaction with their projects (12). Furthermore, students can publish their discoveries, which fosters a sense of belonging to the scientific community (13, 14).

Here, we document a successful CURE that applies bioinformatics tools to discover candidate DNA-binding factors that interact with a genomic locus. Specifically, we investigated the *Drosophila melanogaster* histone gene locus, which encodes the replication-dependent histones. Because histones undergo non-canonical mRNA processing and exhibit cell cycle-dependent expression, they require a unique suite of transcription and processing factors (15). Although many of these factors are known, the complete inventory of histone gene expression regulators remains incomplete.

In this CURE, students utilize a hypothesis-based candidate approach to identify existing high-throughput datasets (specifically, ChIP-seq or similar techniques). By mapping the reads from a ChIP-seq experiment to the *Drosophila* histone gene locus and critically examining the alignment data, students determine if a transcription factor likely targets the locus, suggesting that it may contribute to histone biogenesis. Our approach functions as a primary screen to identify candidate regulatory proteins and provides opportunities for wet-laboratory follow-up undergraduate research projects (e.g., co-immunostaining for the candidate and a positive control to validate bioinformatics findings) (16).

We piloted our CURE remotely with students who were confined at home during the early COVID-19 pandemic. We then transitioned to an in-person experience during a 50-minute weekly "discussion" period attached to a sophomore-level genetics course. Over the course of a semester, each student chose at least one protein to investigate, identified appropriate datasets, mapped datasets to the *Drosophila* histone gene locus using Galaxy, and produced alignment figures. The semester culminated in a poster session, during which the students presented their findings to members of the Biology Department (i.e., faculty, staff, and students).

The CURE presented here is beneficial to all parties involved: not only did the students obtain valuable research experience and transferable skills, but also they identified new candidate factors to further investigate in our wet laboratory, allowing us to make research progress (14, 16). There are thousands of ChIP-seq datasets across multiple repositories that are available for analysis. Future students could examine other types of high-throughput datasets, such as ATAC-seq, to further probe the chromatin landscape of the histone gene locus. The bioinformatics analysis presented here can be extended to any annotated locus of interest in any organism. These seemingly endless possibilities support the sustainable implementation and adaptation of this CURE.

Intended audience

We implemented this CURE in a 200-level genetics course that contained 25 sophomores, juniors, and seniors, most of whom were biology majors. Previously, we piloted the CURE virtually with smaller groups of volunteer college students of similar demographics. We also sponsored a remote high school student, indicating that students with a wide range of experience levels can perform the research with appropriate training.

Learning time

The course had two 75-minute lecture periods and one 50-minute "discussion" period per week over a 14-week semester. Traditionally, the discussion period for this course was used for worksheets, activities, and/or literature discussions. Instead, we implemented the CURE during this time over the entire semester, which accounted for 20% of the overall course grade.

Prerequisite student knowledge

We covered all of the background information on conceptual topics, such as transcription factors and ChIP-seq, in the lecture portion of the class (see Appendix 1 for ChIP-seq resources). Therefore, the only prerequisites for the CURE were the course prerequisites (freshmen-level introductory courses for biology majors). In addition, students did not need prior bioinformatics or computer science experience; all required skills were taught in the training modules.

Learning objectives

Our overall goal was to provide students with an authentic bioinformatics research experience. Upon completion of this CURE, students will be able to:

1. Search peer-reviewed literature to identify candidate proteins that target the *Drosophila* histone gene locus.
2. Form a hypothesis about the candidate protein based on background literature.
3. Identify appropriate datasets (e.g., ChIP-seq or CUT&RUN) through literature or database searches.
4. Map datasets to the *Drosophila* histone gene locus using bioinformatics tools in Galaxy.
5. Visualize data by producing alignment figures in Integrative Genomics Viewer (IGV) software.
6. Synthesize data and conclude if the candidate targets the *Drosophila* histone gene locus.
7. Propose at least two follow-up experiments related to the candidate protein based on ChIP-seq outcome.
8. Present findings to a wider audience (i.e., peers and department) at an in-person poster session.

PROCEDURE

Materials

The following materials are required for this CURE:

- Computer and internet access
- Galaxy account (free web-based platform, www.usegalaxy.org)
- Integrative Genomics Viewer software (free downloadable software, https://software.broadinstitute.org/software/igv/)
- Learning management software such as Canvas or cloud storage program such as Google Drive or OneDrive to house files and course materials

- Customizable form software, such as Google forms, to assess weekly student progress. Alternatively, students could use a software such as Benchling, OneNote, or Google Docs as a laboratory notebook and allow instructors access to monitor progress
- Poster making software, such as PowerPoint, Google Slides, or BioRender
- Poster printing facility or online poster platform such as SpatialChat
- Optional: video production software such as Zoom, if the instructor is generating pre-recorded tutorials or the CURE is conducted remotely

Student instructions

Students received the schedule (Fig. 1) at the beginning of the semester, which we divided into four general categories:

1. Background (weeks 1–4), during which students read and discussed review (15) and research (17) articles
2. Tutorials (weeks 5–7), during which students learned how to use Galaxy and IGV through instructor-led in-person tutorials

Week	Category	Topic	Assignment
1	Background	Introductions - histone gene expression, high-throughput dataset repositories	
2		Discuss review paper (15)	Read paper (due before class)
3		Discuss research paper (17)	Read paper (due before class)
4		Histone gene expression knowns & unknowns; how to select a candidate	Fill out Google spreadsheet with your candidate
5	Tutorials	Tutorial - finding data (NCBI GEO) and Galaxy introduction	Google form with screenshot
6		Tutorial - Galaxy commands	Google form with screenshot
7		Tutorial - Galaxy outputs, IGV	Google form with screenshot
8	Work days	Work session 1	Google form with screenshot
9		Work session 2	Google form with screenshot
10		Work session 3	Google form with screenshot
11		Work session 4	Google form with screenshot
12		Poster tutorial (work session 5)	
13		Poster making session (work session 6)	Poster draft
14	Poster session	Poster session	Fill out 3 peer review forms during the poster session

FIG 1 Weekly class schedule for the CURE. We divided the 14-week semester into four categories (background, tutorials, work days, and poster session). We assessed student participation through activity logs (Google forms).

3. Work days (weeks 8–13), during which students independently carried out their bioinformatics analyses and created their poster under in-person supervision from the instructor

4. The poster session (week 14), during which students presented their work

For the background sessions, we assigned small groups a figure from the review and research papers to annotate using a presentation template (Appendix 2). During the tutorial and work day sessions, students completed a Google form at the end of class describing their efforts and progress that day (Appendix 3). At the poster session, each student presented their poster and filled out three peer review forms (Appendix 4).

Faculty instructions

Background

Our class met twice weekly for the lecture portion (a 75-minute period) and once weekly for the bioinformatics CURE portion (two sections of a 50-minute period) over 14 weeks (see Fig. 1 for the schedule). During lectures, we followed a "molecules first" rather than "Mendel first" approach (18) to introduce CURE-relevant concepts earlier. For example, concepts covered in the first weeks included transcription, transcriptional regulation, and epigenetics. Lecture topics also paid special attention to high-throughput procedures, such as ChIP-seq (see Appendix 1 for ChIP-seq teaching resources). Students learned how wet-laboratory scientists generate sequencing data, how to identify appropriate experimental controls, and the types of research questions that these techniques address. This approach synchronized the lecture and discussion sessions and provided the students with the required background knowledge for the CURE.

During the discussion period, we spent the first four weeks introducing students to *Drosophila* histone gene expression through literature discussions. Students read and discussed both a review article (15) and a research article that used a bioinformatics approach similar to that introduced in the CURE (17). For each paper, we assigned small groups a figure to annotate during class and submit to the instructor (see Appendix 2), which served as their graded assessment for the week.

In the fourth week, we shifted to candidate protein selection. Students gathered additional information on histone gene expression and DNA-binding proteins from PubMed and FlyBase (19). We gave several guiding criteria for finding a candidate factor, such as (A) proteins that interact with known histone regulators, using protein interaction databases such as STRING (https://string-db.org/) (20); (B) transcription factors that act in the early *Drosophila* embryo, which requires rapid histone biosynthesis (21); (C) DNA-binding factors implicated in cell cycle progression, as histone expression is linked to S-phase (15); and (D) dosage compensation factors, because a prominent histone gene regulator is also involved in dosage compensation (17). Students worked independently while the instructor circulated the classroom for individual *ad hoc* check-ins. Although the instructors provided guidance, candidate selection was ultimately student driven. At the end of this class period, students recorded their chosen protein on a class-wide Google spreadsheet, which served as their assessment for the week.

Tutorials

We followed background and brainstorming sessions with 3 weeks bioinformatics tutorials, during which we led students through analysis and visualization of example data using Galaxy (11) and IGV (22). Pre-recorded tutorials were also posted on our learning management site (Canvas) for students to reference outside class and contained the same information as what was presented in class. In the tutorials, we used ChIP-seq data from the background primary research article (17) to ensure that their results matched the published figures. See Fig. 2 for an overview of the tools we used in Galaxy,

Galaxy tool	Description	Input	Output
Faster Download and Extract Reads in FASTQ	Extracts sequencing reads (.fastq files) from an SRA import folder	SRA import folder	.fastq file(s)
FastQC	Quality control of the sequencing reads	.fastq file(s)	(1) "webpage" readout (2) "raw data" readout
Bowtie2	Aligns sequencing reads to genome (either built-in genome or user-provided genome)	(1) .fastq file(s) (2) Normalized .fasta genome file (only if using a user-provided genome)	.bam file (ChIP-seq reads mapped to user-specified genome)
bamCoverage	Converts .bam files to .bigwig files, which are better for visualization	.bam file	.bigwig file
bamCompare	Normalizes experimental conditions to input (if available)	(1) Input .bam file (2) Experimental .bam file	.bigwig file

FIG 2 Summary of tools used in Galaxy. Each tool can be found by using the search function in Galaxy (see Appendix 5 for galaxy tutorial).

and Appendix 5 for the Galaxy workflow tutorial. Due to computing demands on the Galaxy servers, some tools can take several hours to complete. During any downtime, students continued their background research on candidate proteins in preparation for designing their poster. We consulted with each student individually during class time to provide guidance, and students could also come to office hours for additional help.

Work days

The next six discussion periods functioned as work sessions for students to carry out their bioinformatics analyses. Because the majority of high-throughput sequencing experiments funded by the National Institutes of Health are deposited into public databases such as the NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/), many students identified ChIP-seq datasets for their candidate protein(s) by directly searching this database. Others located GEO accession numbers within primary literature. We also directed students to additional databases such as modENCODE (23), which contains ChIP-seq datasets for various transcription factors from numerous tissues and developmental timepoints in model organisms. In several cases, students formed strong hypotheses, but ChIP-seq data did not yet exist for their candidate protein. We instructed these students to choose another factor, and this did not bias their grade. Although some students went through this selection process several times, all found a unique factor to investigate. Once each student located usable data, they carried out the analysis pipeline in Galaxy (Fig. 2) and subsequently generated alignment figures using IGV (Appendix 5). Several students had time to investigate multiple (often related) candidates based on the conclusions from their first hypothesis.

We dedicated two of the work sessions to poster design. We presented resources for crafting posters (e.g., https://www.posternerd.com/tutorials) and shared our assessment rubric (Appendix 6). Students submitted a draft of the poster in week 13 (Fig. 1), for which we provided written feedback and allowed students to revise before printing.

Poster session

We held the poster session on the last day of the discussion period and invited members of the Biology Department to attend. See Fig. 3 for an example student poster. We divided the students into two groups: while the first group presented their posters, each student in the second group filled out three peer review forms (Appendix 4), which served as a graded assessment. The students switched roles halfway through the session.

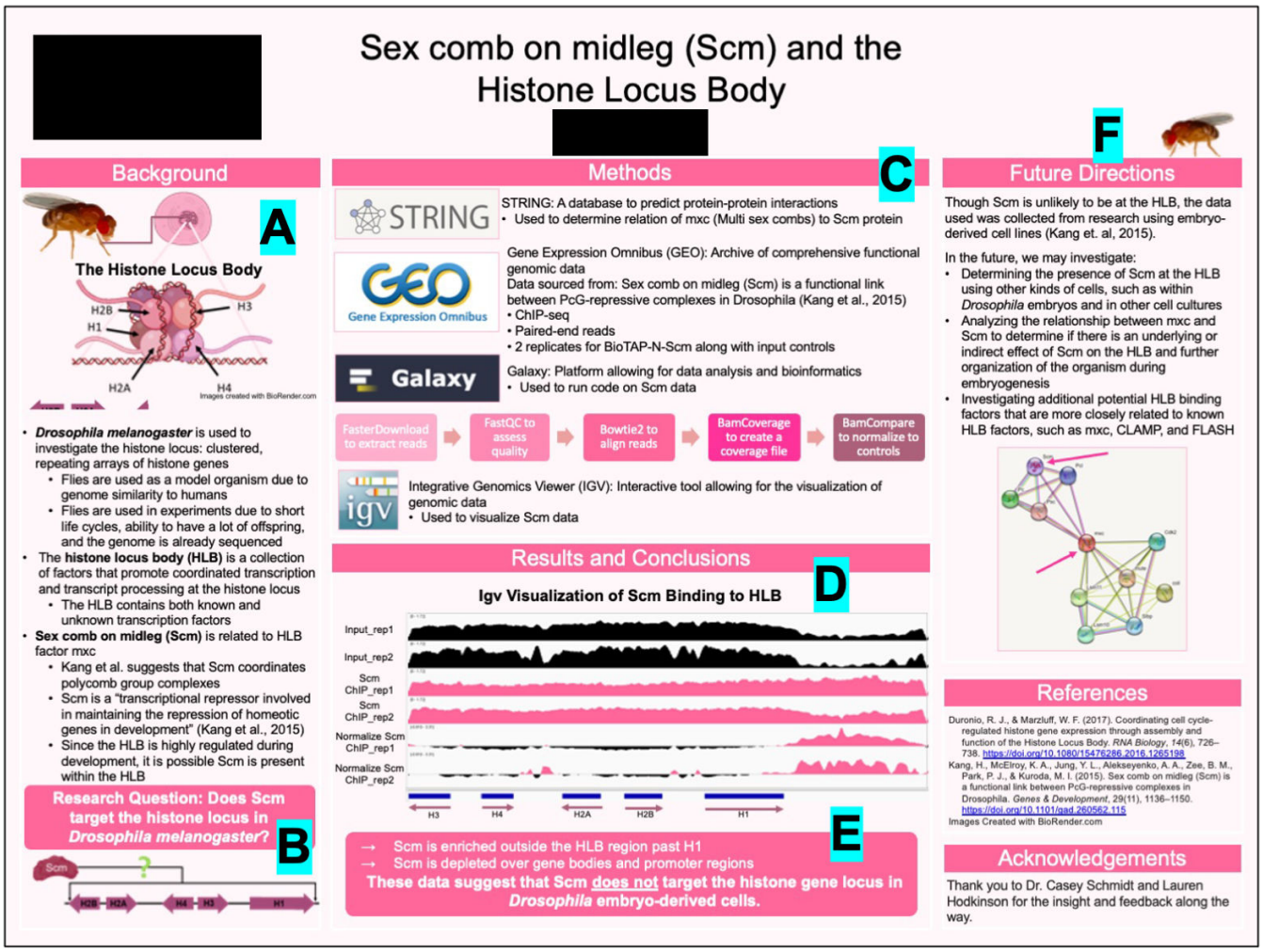

FIG 3 Example student poster. (A) The background section contains an image created in BioRender, information on the specific factor the student investigated, and information on the *Drosophila* histone gene locus. (B) The research question accurately summarizes the project. (C) The methods section lists the type of data analyzed (ChIP-seq, paired-end reads, two replicates), the programs/databases used, and the specific tools in Galaxy. (D) In the results section, the student re-labeled the tracks to descriptive titles and increased the default font size. The IGV tracks are also color-coded. (E) The conclusion section summarizes the data while considering limitations (i.e., the cell type used for the ChIP experiment). (F) At least two future experiments are proposed.

Notes and recommendations

Instructors may wish to have a set of "backup" datasets for students who have difficulty locating appropriate data. Students can map data from ChIP-seq variation techniques, such as ChIP-nexus (24), CUT&RUN (25), and CUT&Tag (26) using the same bioinformatics analysis as ChIP-seq. However, ChIP-chip (chromatin immunoprecipitation followed by microarray) datasets cannot be used with our pipeline (Fig. 2) because microarrays utilize different analyses. Unfortunately, some datasets do not contain appropriate controls. For example, we routinely find ChIP-seq datasets that do not include an input or control immunoprecipitation (e.g., IgG) condition, which are important to normalize or compare to the experimental ChIP data. The lack of normalization can sometimes lead to misleading or false-positive results, wherein small local peaks appear as positive signal (16). Although there is no way to rectify the lack of controls, it allows for important discussions with students on what conclusions one can draw from their datasets.

Suggestions for determining student learning

Student posters were the primary mode of assessment for our CURE (worth 25% of the discussion grade, plus 15% for the poster peer review assignment). The remaining 60% of the discussion grade was based on participation in the research, assessed through student-reported activity logs. It is sometimes difficult to assess inquiry-based research, and the bioinformatics component added additional hurdles for some students. For example, there may not exist appropriate datasets for a student's selected candidate, Galaxy may perform slowly, or a dataset from a large study may contain many variables (e.g., environmental conditions, mutant genotypes, treatments, or tissue types) such that students struggle to determine which samples are relevant (see Appendix 5). Therefore, we emphasized progress and effort over results and did not penalize students for things

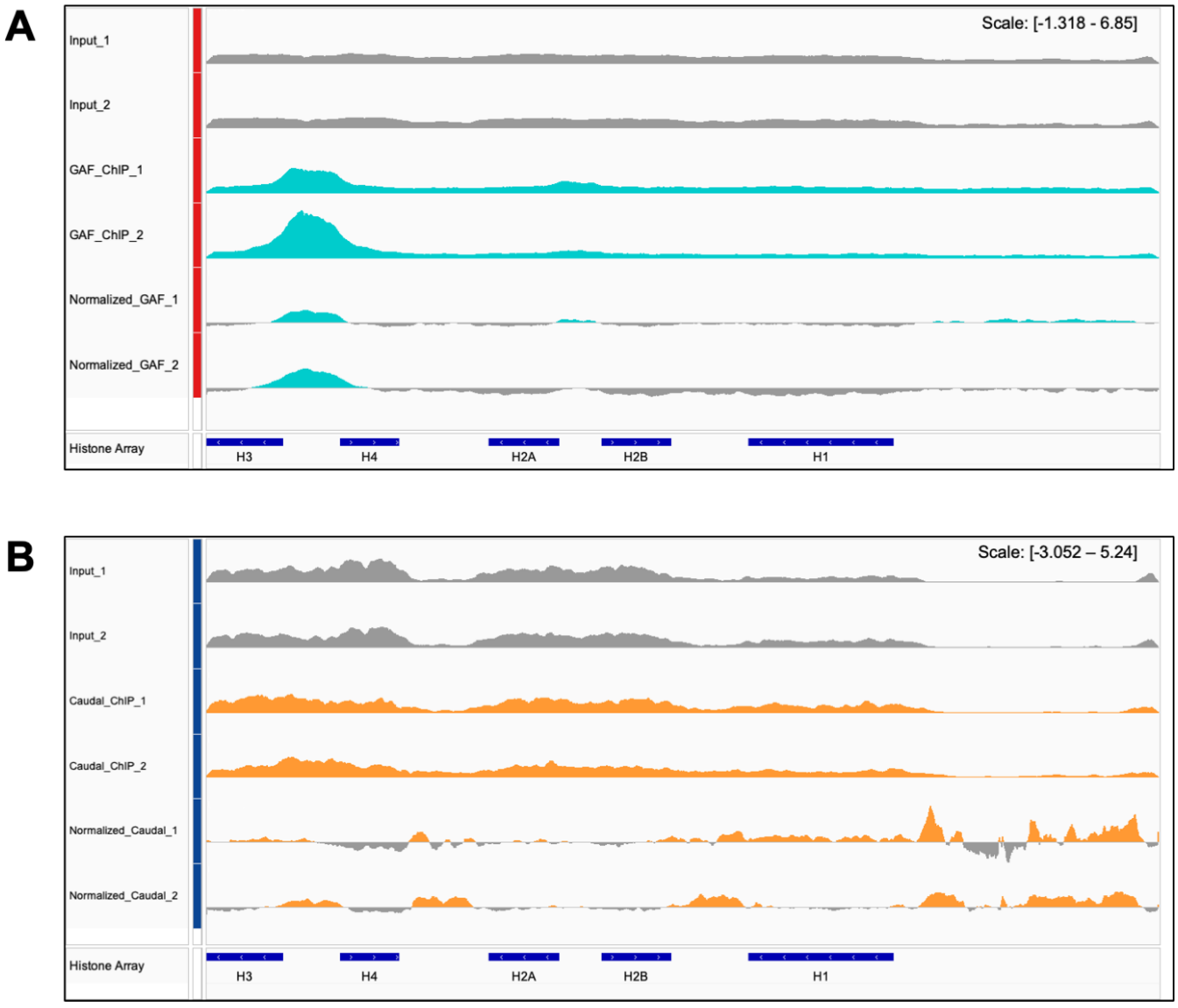

FIG 4 Sample data. (A) ChIP-seq alignment of GAGA Factor (GAF) in stage 3 *Drosophila* embryos (teal; input, gray). The figure shows two replicates from the same study. There is a clear peak between the *H3* and *H4* genes, suggesting that GAF localizes to this region. This finding was surprising, given that GAF does not target the histone gene locus in cultured S2 cells or by immunofluorescence in early embryos (17). Data from Ref. (28), GEO accession no. GSE152770. (B) ChIP-seq alignment of Caudal in 0- to 4-h embryos (orange; input, gray). The figure shows two replicates from the same study. Although there is a signal upstream of *H1* in the normalized panels, the peaks are not reflected in the ChIP panels, suggesting that they are not true signal. Thus, there is no clear enrichment of Caudal at the histone gene locus. Data from reference (23), GEO accession no. GSE20000.

out of their control. At the end of each discussion session, the students filled out a Google form describing the activities they performed that day. These forms included space to upload a screenshot of Galaxy or IGV (Appendix 3). Through the Google forms, we assessed participation and monitored progress so that we could intervene if necessary. Specifically, we ensured that students had found a dataset for their candidate by week 9 (work session 2; see Fig. 1) and provided guidance if they had not.

An additional approach to determining student learning is to include formative assessments throughout the semester. For example, groups of students might complete a worksheet such as the Figure Facts template (27) that walks through a figure from a primary research article, which could be a graded formative assessment. Students could also gain presentation experience by sharing a research article that includes the dataset that they plan to analyze. The instructor may choose to have students self-report their activities in graded laboratory notebooks. These assessments offer additional opportunities for instructor feedback but may be impractical in a larger class.

Sample data

We present example candidates (Fig. 4). First, we identified a dataset for GAGA Factor (GAF) (28) and mapped ChIP-seq reads to the *Drosophila* histone locus. We classify GAF as a "positive" candidate due to the strong, broad peak between the *H3* and *H4* genes (Fig. 4A). This result suggests that GAF targets this region of the histone locus and is a good candidate for wet-laboratory follow-up experiments. Second, we mapped ChIP-seq data for the transcription factor Caudal (23) but did not observe meaningful signal (Fig. 4B). Although the normalized panels appear to have signal, the peaks are not reflected in the ChIP panels, suggesting that they are an artifact of normalization and thus not a true signal. Other students also observed this phenomenon (Fig. 3). We classify Caudal as a "negative" candidate. The results from these and other CURE iterations are suitable for publication (14, 16).

Safety issues

Because this activity does not involve a traditional laboratory setup, we do not foresee any safety issues.

DISCUSSION

Field testing

We began this bioinformatics project as a strategy to engage our junior laboratory members in remote work during the early COVID-19 pandemic. During the fall of 2020, undergraduates at our institution were not permitted to work in research buildings. Instead, our undergraduate laboratory researchers collectively learned basic bioinformatics skills. Four students each chose a protein to study, identified datasets, mapped data to the histone gene array, and presented their findings to the larger laboratory group. After this first pilot, we recruited nine naive undergraduates from our institution to remotely study the chromatin landscape of the *Drosophila* histone gene locus in the spring of 2021. For this iteration, students chose a histone post-translational modification and mapped ChIP-seq data from the modENCODE project (23). The students presented their findings to a wider audience via a virtual poster session. Three students from this group joined our wet laboratory when we returned to in-person instruction and carried out independent projects.

Our laboratory also sponsored a remote high school student that continued the bioinformatics project during the summer of 2021. This student investigated several early *Drosophila* embryo patterning factors (Fig. 4), providing our wet laboratory with candidates for follow-up studies. Most recently, we implemented the project as an in-person CURE in a 200-level genetics course with 25 students.

The class size will likely contribute to the effectiveness of this CURE. Our weekly discussion period was split into two 50-minute sections, with 14 students in one and

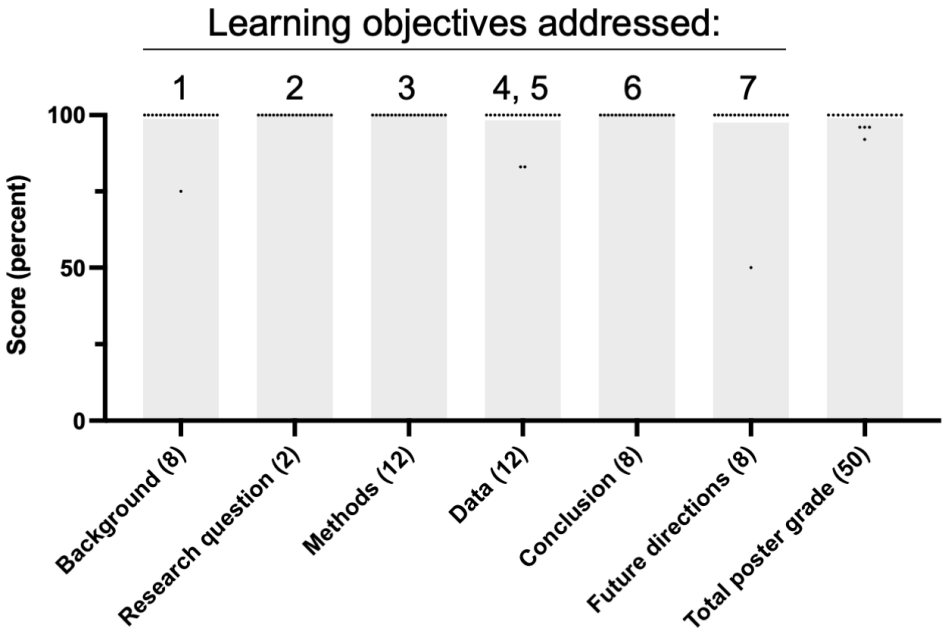

FIG 5 The primary form of summative assessment for this CURE was the students' posters. The bar graph represents the score (as a percent) for individual poster sections and the entire poster. Each dot (black) represents an individual student's score. Each bar (gray) represents the average of the dots. The point value of each poster section is listed in parentheses. Learning objectives addressed by each poster section are listed above the bars. Data were obtained were consenting students.

11 in the other. This small size allowed us to grant individual attention to each student. Because several of our students ran into difficulties finding appropriate ChIP-seq datasets for their chosen candidate factor, we found that this one-on-one time was necessary to ensure the success of all students, and we recommend a ratio of one instructor to no more than 15 students. If individual conversations are not feasible, the instructor could employ additional experienced teaching assistants to consult with the students or students could operate in small groups.

Evidence of student learning

We primarily evaluated the CURE learning objectives through the student posters, which served as a summative assessment (Fig. 5). Learning objectives 1–7 were reflected in the poster rubric (Appendix 6). The posters were worth 50 points in total. We gave all students ungraded feedback on their poster before the final submission by providing written comments. Student grades for the poster ranged between 92% and 100%. Most deductions were related to data presentation, as we instructed students to change the default labels and font size in the IGV plots (Fig. 5; Appendix 6).

We also documented student learning in CURE-related exam questions, which at least 80% of students answered correctly (Fig. 6). For example, we asked what experiment a student would perform to determine the genomic localization of a hypothetical new histone variant protein. This question, which we classify in the "Apply" level of Bloom's taxonomy (29), required students to recall that histones are DNA-binding proteins and to differentiate between types of experiments (Fig. 6A). In addition, we asked students to draw the results of a ChIP-seq experiment if the researcher forgot to add the primary antibody (Fig. 6B). We classify this question in the "Analyze" level of Bloom's taxonomy because it addresses the role of different reagents in an experiment. Collectively, these results demonstrate that our students displayed higher-order reasoning on CURE-related topics in their exams.

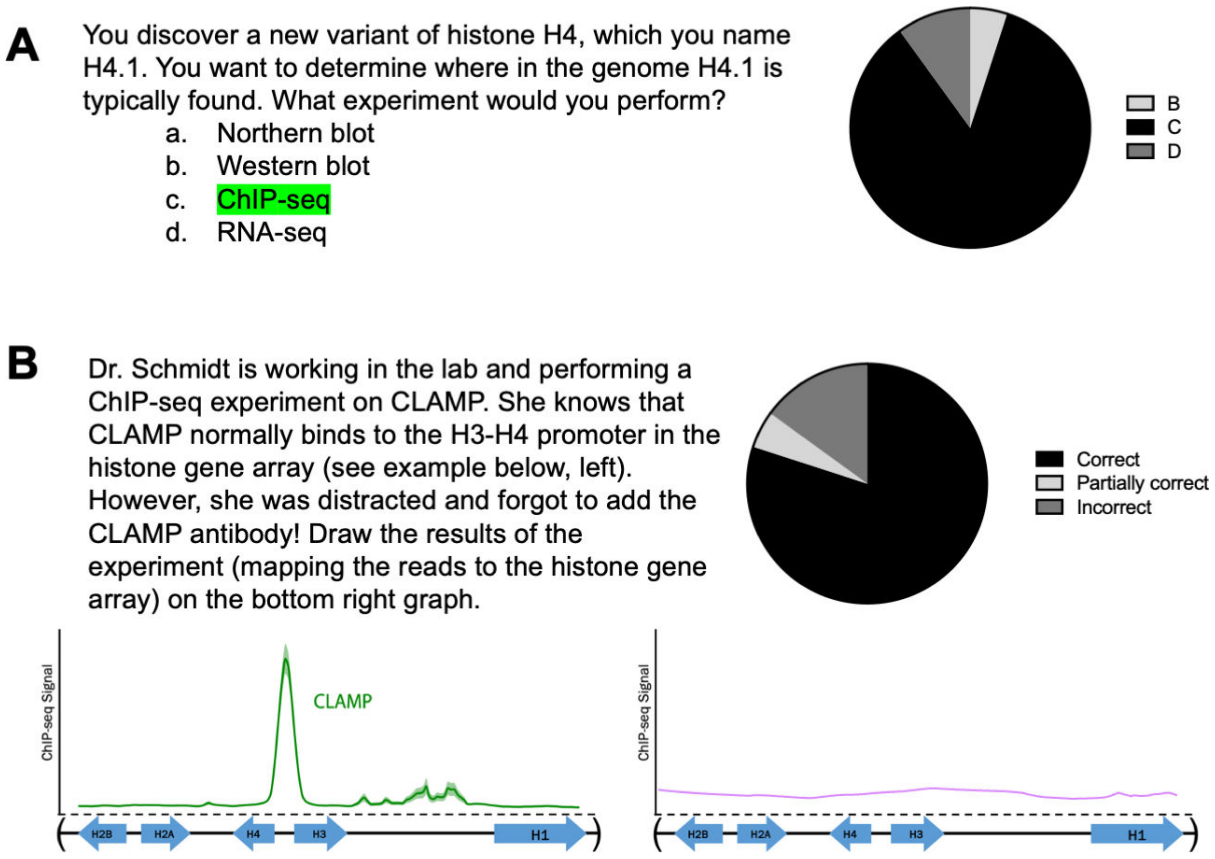

FIG 6 CURE-related exam questions. (A) A multiple choice exam question was answered correctly by 85% of students; the correct answer is highlighted in green. (B) An open-ended exam question was answered correctly by 80% of students; a correct answer is drawn on the right panel by the instructor in purple. CLAMP ChIP-seq data are from reference 17. Data were obtained from consenting students. (A portion of panel B is reproduced from reference 17 with permission of the publisher; copyright Cold Spring Harbor Laboratory Press.)

Possible modifications

Because bioinformatics research does not require a wet-laboratory setup, this CURE can be implemented remotely and/or asynchronously. We held our CURE pilots synchronously over Zoom during the early COVID-19 pandemic and used the platform SpatialChat (https://spatial.chat) to hold a virtual poster session. In addition, instructors can adapt this CURE to study any genomic locus of interest (e.g., an enhancer region that might attract regulatory factors) in any species with an annotated genome. The workflow is particularly suitable for repetitive regions (such as the histone or ribosomal gene arrays) because these regions are often excluded from genome-wide analyses in prior publications. Galaxy contains many built-in genomes, but instructors can also provide a custom genome. We used a custom genome that contains a single copy of the histone gene array (30) because the sequences of the ~100 array copies are nearly identical in the *Drosophila melanogaster* genome (31). This approach amplifies the ChIP-seq signal (Fig. 4) (17). Furthermore, this CURE can be used to map other types of high-throughput data. For example, students could examine chromatin landscape data, such as ATAC-seq or FAIRE-seq, and compare to histone modification ChIP-seq datasets that correlate with different chromatin states at a particular locus (32).

An exciting follow-up to the bioinformatics CURE is to confirm positive candidates with wet-laboratory experiments. *Drosophila melanogaster* is a particularly useful model organism for these follow-up studies due to the wealth of available mutant and RNAi lines in public stock centers, as well as established protocols for staining tissues. There are also numerous custom antibodies that researchers can request from individual

laboratories or purchase from stock centers such as the Developmental Studies Hybridoma Bank (https://dshb.biology.uiowa.edu/). These wet-laboratory experiments can provide a platform for future studies; for example, testing histone gene expression in the absence of a validated protein that targets the histone gene locus (17).

Summary

The data generated from this CURE will ultimately add to the growing body of knowledge regarding transcription factor targeting of genomic loci. In addition, the CURE provides students with an authentic research experience, especially in situations where in-person wet-laboratory research is not feasible. Students also gain transferable skills that are important for STEM education, including (A) reading and interpreting primary literature, (B) forming hypotheses based on prior research, (C) navigating complex databases, (D) drawing conclusions from data, and (E) proposing future studies. Furthermore, students interested in continuing bioinformatics research will require less training because they have learned basic bioinformatics techniques. The skills gained during this CURE are crucial to both research science and critical thinking.

ACKNOWLEDGMENTS

We are grateful to the Emory University students who participated in our first (Dabin Cho, Gregory Kimmerer, Mary Wang, and Mellisa Xie) and second (Eric Albanese, Yono Bulis, Edgar Hsieh, Shaariq Khan, Andre Mijacika, Sean Parker, Rohan Ramdeholl, Annalise Weber, and Kelly Yoon) remote pilots, as well as the students who participated in the in-person CURE. This study was determined by the Emory University Institutional Review Board to be exempt from further review (STUDY00005976), and grade data in Fig. 4 and 5 were obtained only from consenting students. We also thank Dr. Karen Resendes, Dr. Kelsey Gray, Dr. Jennifer Gresham, Dr. Ethan Rundell, and Dr. Michaelyn Hartmann for critically reviewing this manuscript.

This work was supported by K12GM00068 to CAS and HSC; F32GM140778 to CAS; T32GM00008490 and F31HD105452 to LJH; and R00HD092625 and R35GM142724 to LER.

AUTHOR AFFILIATIONS

[1]Department of Biology, Emory University, Atlanta, Georgia, USA
[2]Graduate Program in Genetics and Molecular Biology, Emory University, Atlanta, Georgia, USA
[3]Choate Rosemary Hall, Wallingford, Connecticut, USA

AUTHOR ORCIDs

Casey A. Schmidt http://orcid.org/0000-0002-4678-5523
Leila E. Rieder http://orcid.org/0000-0001-9851-0145

AUTHOR CONTRIBUTIONS

Casey A. Schmidt, Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review and editing | Lauren J. Hodkinson, Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Writing – review and editing | H. Skye Comstra, Conceptualization, Funding acquisition, Methodology, Project administration, Software, Supervision, Validation, Writing – review and editing | Samia Khan, Investigation, Visualization | Henrik Torres, Investigation, Visualization | Leila E. Rieder, Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – review and editing

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Appendices (jmbe00120-23-s0001.pdf). Appendices 1–4 and 6 (Appendix 5 is separate file).

Appendix 5 (jmbe00120-23-s0002.pdf). Tutorial for using Galaxy and IGV.

REFERENCES

1. Eagan MK, Hurtado S, Chang MJ, Garcia GA, Herrera FA, Garibay JC. 2013. Making a difference in science education: the impact of undergraduate research programs. Am Educ Res J 50:683–713. https://doi.org/10.3102/0002831213482038
2. Szteinberg GA, Weaver GC. 2013. Participants' reflections two and three years after an introductory chemistry course-embedded research experience. Chem Educ Res Pract 14:23–35. https://doi.org/10.1039/C2RP20115A
3. Bangera G, Brownell SE. 2014. Course-based undergraduate research experiences can make scientific research more inclusive. CBE Life Sci Educ 13:602–606. https://doi.org/10.1187/cbe.14-06-0099
4. Auchincloss LC, Laursen SL, Branchaw JL, Eagan K, Graham M, Hanauer DI, Lawrie G, McLinn CM, Pelaez N, Rowland S, Towns M, Trautmann NM, Varma-Nelson P, Weston TJ, Dolan EL. 2014. Assessment of course-based undergraduate research experiences: a meeting report. CBE Life Sci Educ 13:29–40. https://doi.org/10.1187/cbe.14-01-0004
5. Evans CJ, Olson JM, Mondal BC, Kandimalla P, Abbasi A, Abdusamad MM, Acosta O, Ainsworth JA, Akram HM, Albert RB, et al. 2021. A functional genomics screen identifying blood cell development genes in *Drosophila* by undergraduates participating in a course-based research experience. G3 (Bethesda) 11:jkaa028. https://doi.org/10.1093/g3journal/jkaa028
6. Olson JM, Evans CJ, Ngo KT, Kim HJ, Nguyen JD, Gurley KGH, Ta T, Patel V, Han L, Truong-N KT, et al. 2019. Expression-based cell lineage analysis in *Drosophila* through a course-based research experience for early undergraduates. G3 (Bethesda) 9:3791–3800. https://doi.org/10.1534/g3.119.400541
7. Delventhal R, Steinhauer J. 2020. A course-based undergraduate research experience examining neurodegeneration in *Drosophila melanogaster* teaches students to think, communicate, and perform like scientists. PLoS One 15:e0230912. https://doi.org/10.1371/journal.pone.0230912
8. Genné-Bacon EA, Wilks J, Bascom-Slack C. 2020. Uncovering factors influencing instructors' decision process when considering implementation of a course-based research experience. CBE Life Sci Educ 19:ar13. https://doi.org/10.1187/cbe.19-10-0208
9. Fernandes PA, Passos Ó, Ramos MJ. 2022. Necessity is the mother of invention: a remote molecular bioinformatics practical course in the COVID-19 era. J Chem Educ 99:2147–2153. https://doi.org/10.1021/acs.jchemed.1c01195
10. Anderson N, Wilch M. 2021. Online instruction – bioinformatics lesson for a COVID-19 vaccine. Am Biol Teach 83:464–471. https://doi.org/10.1525/abt.2021.83.7.464
11. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451–1455. https://doi.org/10.1101/gr.4086505
12. Kirkpatrick C, Schuchardt A, Baltz D, Cotner S. 2019. Computer-based and bench-based undergraduate research experiences produce similar attitudinal outcomes. CBE Life Sci Educ 18:ar10. https://doi.org/10.1187/cbe.18-07-0112
13. Turner AN, Challa AK, Cooper KM. 2021. Student perceptions of authoring a publication stemming from a course-based undergraduate research experience (CURE). CBE Life Sci Educ 20:ar46. https://doi.org/10.1187/cbe.21-02-0051
14. Hodkinson LJ, Smith C, Comstra HS, Ajani BA, Albanese EH, Arsalan K, Daisson AP, Forrest KB, Fox EH, GueretteMR, et al. 2023. A bioinformatics screen reveals hox and chromatin remodeling factors at the *Drosophila* histone locus. bioRxiv. https://doi.org/10.1101/2023.01.06.523008
15. Duronio RJ, Marzluff WF. 2017. Coordinating cell cycle-regulated histone gene expression through assembly and function of the histone locus body. RNA Biol 14:726–738. https://doi.org/10.1080/15476286.2016.1265198
16. Xie M, Comstra S, Schmidt C, Hodkinson L, Rieder LE. 2022. Max is likely not at the *Drosophila* histone locus. bioRxiv. https://doi.org/10.1101/2022.09.11.507040
17. Rieder LE, Koreski KP, Boltz KA, Kuzu G, Urban JA, Bowman SK, Zeidman A, Jordan WT, Tolstorukov MY, Marzluff WF, Duronio RJ, Larschan EN. 2017. Histone locus regulation by the *Drosophila* dosage compensation adaptor protein CLAMP. Genes Dev 31:1494–1508. https://doi.org/10.1101/gad.300855.117
18. Deutch CE. 2018. Mendel or molecules first: what is the best approach for teaching general genetics?. Am Biol Teach 80:264–269. https://doi.org/10.1525/abt.2018.80.4.264
19. Jenkins VK, Larkin A, Thurmond J. 2022. Using FlyBase: a database of *Drosophila* genes and genetics, p 1–34. In Dahmann C (ed), *Drosophila*: methods and protocols. Springer US, New York, NY. https://doi.org/10.1007/978-1-0716-2541-5
20. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, Doncheva NT, Legeay M, Fang T, Bork P, Jensen LJ, von Mering C. 2021. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acids Res 49:10800. https://doi.org/10.1093/nar/gkab835
21. Horard B, Loppin B. 2015. Histone storage and deposition in the early *Drosophila* embryo. Chromosoma 124:163–175. https://doi.org/10.1007/s00412-014-0504-7
22. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26. https://doi.org/10.1038/nbt.1754
23. modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797. https://doi.org/10.1126/science.1198374
24. He Q, Johnston J, Zeitlinger J. 2015. ChIP-nexus enables improved detection of *in vivo* transcription factor binding footprints. Nat Biotechnol 33:395–401. https://doi.org/10.1038/nbt.3121
25. Skene PJ, Henikoff S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. Elife 6:e21856. https://doi.org/10.7554/eLife.21856
26. Kaya-Okur HS, Wu SJ, Codomo CA, Pledger ES, Bryson TD, Henikoff JG, Ahmad K, Henikoff S. 2019. CUT&tag for efficient epigenomic profiling of small samples and single cells. Nat Commun 10:1930. https://doi.org/10.1038/s41467-019-09982-5
27. Round JE, Campbell AM. 2013. Figure facts: encouraging undergraduates to take a data-centered approach to reading primary literature. CBE Life Sci Educ 12:39–46. https://doi.org/10.1187/cbe.11-07-0057
28. Gaskill MM, Gibson TJ, Larson ED, Harrison MM. 2021. GAF is essential for zygotic genome activation and chromatin accessibility in the early *Drosophila* embryo. Elife 10:e66668. https://doi.org/10.7554/eLife.66668
29. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR, eds. 1956. Taxonomy of educational objectives Handbook I: The cognitive domain. Longman, New York.

30. McKay DJ, Klusza S, Penke TJR, Meers MP, Curry KP, McDaniel SL, Malek PY, Cooper SW, Tatomer DC, Lieb JD, Strahl BD, Duronio RJ, Matera AG. 2015. Interrogating the function of metazoan histones using engineered gene clusters. Dev Cell 32:373–386. https://doi.org/10.1016/j.devcel.2014.12.025
31. Bongartz P, Schloissnig S. 2019. Deep repeat resolution—the assembly of the *Drosophila* histone complex. Nucleic Acids Res 47:e18. https://doi.org/10.1093/nar/gky1194
32. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, Brugman W, de Castro IJ, Kerkhoven RM, Bussemaker HJ, van Steensel B. 2010. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell 143:212–224. https://doi.org/10.1016/j.cell.2010.09.009