Big Data to the Bench: Transcriptome Analysis for Undergraduates

Carl Procko,[†][‡][*] Steven Morrison,[‡] Courtney Dunar,[‡][§] Sara Mills,[‡][§] Brianna Maldonado,[‡][§] Carlee Cockrum,[‡][§] Nathan Emmanuel Peters,[‡][§] Shao-shan Carol Huang,[||] and Joanne Chory[†][¶][*]

[†]Plant Biology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; [‡]Department of Biology, University of San Diego, San Diego, CA 92110; [||]Department of Biology, New York University, New York, NY 10003; [¶]Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92037

ABSTRACT

Next-generation sequencing (NGS)-based methods are revolutionizing biology. Their prevalence requires biologists to be increasingly knowledgeable about computational methods to manage the enormous scale of data. As such, early introduction to NGS analysis and conceptual connection to wet-lab experiments is crucial for training young scientists. However, significant challenges impede the introduction of these methods into the undergraduate classroom, including the need for specialized computer programs and knowledge of computer coding. Here, we describe a semester-long, course-based undergraduate research experience at a liberal arts college combining RNA-sequencing (RNA-seq) analysis with student-driven, wet-lab experiments to investigate plant responses to light. Students derived hypotheses based on analysis of RNA-seq data and designed follow-up studies of gene expression and plant growth. Our assessments indicate that students acquired knowledge of big data analysis and computer coding; however, earlier exposure to computational methods may be beneficial. Our course requires minimal prior knowledge of plant biology, is easy to replicate, and can be modified to a shorter, directed-inquiry module. This framework promotes exploration of the links between gene expression and phenotype using examples that are clear and tractable and improves computational skills and bioinformatics self-efficacy to prepare students for the "big data" era of modern biology.

INTRODUCTION

The Human Genome Project and the subsequent advent of so-called next-generation sequencing (NGS) technologies have been the catalysts for an explosion of genomics-related information that is having a profound effect on medicine and research. This information is often grouped with proteomics, metabolomics, and other large-scale biological data sets under the daunting moniker of "big data." This change in the scale of biological data—in addition to how data are collected and processed—has created a need for novel teaching strategies that address these new technologies and the computational methods by which they are analyzed.

Despite the necessity for improved quantitative skills as biological data become increasingly large and complex, traditional teaching methods that present biology as a distinct discipline independent of mathematics has in the past resulted in isolated groups of students, each specialized within their own field, who often have difficulties communicating across the perceived divide (Gross, 2000; Bialek and Botstein, 2004). In *Vision and Change in Undergraduate Biology Education: A Call to Action*, the American Association for the Advancement of Science (AAAS, 2011) and National Science Foundation argued that the competencies needed to teach students included 1) the ability to interpret quantitative data and 2) the ability to understand the interdisciplinary nature of science. NGS, particularly RNA sequencing (RNA-seq), by its very nature, is cross-disciplinary and a useful tool to target these key competencies.

Erin L. Dolan, *Monitoring Editor*
Submitted Oct 11, 2018; Revised Jan 28, 2019; Accepted Feb 6, 2019
CBE Life Sci Educ June 1, 2019 18:ar19
DOI:10.1187/cbe.18-08-0161

[§]These authors contributed equally to this work.
*Address correspondence to: Carl Procko (procko@salk.edu) or Joanne Chory (chory@salk.edu).

© 2019 C. Procko *et al.* CBE—Life Sciences Education © 2019 The American Society for Cell Biology. This article is distributed by The American Society for Cell Biology under license from the author(s). It is available to the public under an Attribution–Noncommercial–Share Alike 3.0 Unported Creative Commons License (http://creativecommons.org/licenses/by-nc-sa/3.0).
"ASCB®" and "The American Society for Cell Biology®" are registered trademarks of The American Society for Cell Biology.

As such, developing laboratory/classroom exercises that include NGS data analysis should provide inspiration for establishing new undergraduate courses that combine computational skills with biology.

RNA-seq is an NGS-based method by which the entire pool of transcripts within a biological sample can be catalogued and quantified. For many eukaryotic organisms, this pool consists of tens of thousands of unique transcripts. To handle this scale of data, RNA-seq analysis often uses specialized programs that require some knowledge of computer coding (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008; Trapnell *et al.*, 2012). Significant challenges to teaching RNA-seq have been noted, in large part due to the likelihood that most educators lack the necessary training themselves (Peterson *et al.*, 2015). This may be particularly true at smaller universities that do not routinely conduct RNA-seq experiments. Various teaching tutorials and workshop initiatives aim to address this shortcoming (Buonaccorsi *et al.*, 2014; Makarevitch *et al.*, 2015; Peterson *et al.*, 2015; www.rnaseqforthenextgeneration.org), and with many data sets now available through public repositories such as the Sequence Read Archive at the National Center for Biotechnology Information (Kodama *et al.*, 2011) and prices for RNA-seq continuing to fall, the means to teach this material or even perform RNA-seq experiments in the undergraduate sphere is now within the intellectual and financial grasp of many educators (Buonaccorsi *et al.*, 2014).

In addition to improving students' quantitative skills, analysis of RNA-seq data is a valuable instrument for reinforcing concepts of information flow in biological systems (AAAS, 2011). Most undergraduate students majoring in biology understand the concept that genotype determines phenotype; however, they often struggle with describing the molecular mechanisms that bridge the gap between the two (Lewis and Kattmann, 2004; Reinagel and Bray Speth, 2016). RNA-seq data provide a means to improve student understanding of how the information stored in genes translates to observable changes in phenotype via alterations in gene expression. The use of RNA-seq data for this purpose may be particularly powerful when it is combined with tangible wet-lab experimentation (Makarevitch and Martinez-Vaz, 2017).

Here, we describe the implementation of a rigorous, semester-long course-based undergraduate research experience (CURE) that combined RNA-seq big data analysis with wet-lab experimentation. These experiments tested student-derived hypotheses inspired by big data analysis. The positive effect of undergraduate research experiences on student performance and perceptions toward science has been well documented, the benefits of which also extend to historically underrepresented groups in the sciences (Kardash, 2000; Lopatto, 2007; Russell *et al.*, 2007; Ellington *et al.*, 2010; Jones *et al.*, 2010). Traditionally, however, such experiences were limited to individual student internships in research laboratories. CUREs are a means to scale this experience to the entire classroom and can yield learning gains similar to those of traditional internships (Lopatto *et al.*, 2008). CUREs are generally defined as having five major properties: the use of the scientific process, discovery, relevance to the field, collaboration, and iteration (Auchincloss *et al.*, 2014). CUREs foster student participation in the scientific process, which may include the conception and testing of original research hypotheses, and allow students to experience the failures and triumphs of the scientific endeavor. They have been argued as a means to overcome inequities in engaging a broad student population in original research (Bangera and Brownell, 2014) and can promote in students a greater sense of belonging to the larger scientific community (Shaffer *et al.*, 2014).

While other published RNA-seq teaching descriptions have highlighted the possibility of original, student-led follow-up computational studies using the same or other RNA-seq data sets (Makarevitch *et al.*, 2015; Peterson *et al.*, 2015), ours is one of the first that we know of to combine RNA-seq analysis with student-led follow-up studies in a living system. This approach may better integrate the "abstract" data analysis with more traditional concepts held by the students of what biology is and how it is practiced, and it may improve student understanding of information flow from genes to phenotype. Assessment activities, student course evaluations and comments, pre- and postclass perception surveys, and a postclass assessment quiz all suggested that students acquired knowledge and confidence in programming skills and an understanding of NGS analysis. We provide herein our lesson plans and in-class tutorials to be modified as needed by other interested instructors. Depending on institution resources and time, the wet-lab component of our course can be easily modified to a shorter inquiry-based module that complements the big data RNA-seq analysis.

METHODS

Human Subjects Protocol

This project was exempted by the University of San Diego (USD) Institutional Review Board (IRB-2019-46).

Course Description and Class Enrollment

USD is a small, private, liberal arts college. In total, 5605 undergraduate students were enrolled full-time in the 2017–2018 academic year. Students majoring in biology must fulfill a research requirement. To do this, most students enroll in a semester-long Biology 490: Research Project class (4 credits). Each section enrolls eight senior students, and each covers different topics and scales in biology depending on the research interests of the instructor. Generally, three or more research sections are offered per semester that emphasize molecular, physiological, or ecological/population biology. The main objectives for all Biology 490 sections are for students to design and execute research experiments, generate original data, and present their results at an intradepartmental poster session. Later, these same data are used as the basis for a 20- to 25-minute departmental seminar that students deliver orally as part of the required Biology Capstone Seminar course. Each section of Biology 490 meets twice per week for 4 hours, with additional nonscheduled hours as needed.

Our course was offered in the Fall of 2017 as one of three different sections of Biology 490. We titled our section "Big Data in a Post-Genome World." Our section was unusual in that it was taught by two adjunct professors (C.P. and S.M.) with complementary strengths in teaching bioinformatics and plant physiology in a liberal arts college environment. Students in our section included five females and three males. These students had diverse backgrounds that impacted implementation of the course. All had already taken genetics, ecology, research methods, and introductory biology courses as prerequisites for Biology 490. However, while all of our students had some familiarity with basic molecular biology techniques, only one had previous experience with quantitative reverse transcriptase (qRT) polymerase chain reaction (PCR), two had prior experience with the coding language R from a previous biostatistics class, and none had ever undertaken an analysis of NGS data. In addition, only two students had exposure to plant physiology or botany, so exposure to this material and the relevant laboratory procedures was included in the course curriculum and taught as part of or in parallel with lecture and computational modules.

Choice of the Model System

For our course, we chose plants as the model system. The reasons for this were multiple, many of which have been highlighted by other educators (Ebert-May and Holt, 2014; Makarevitch and Martinez-Vaz, 2017). First, seeds and reagents for research in plant systems are readily available from a variety of repositories. These are supported by excellent databases of genomic information and an abundance of primary literature. Second, equipment needed to grow plants is inexpensive and accessible at most schools. While many of the experiments performed by our students took advantage of more elaborate and controlled tissue culture methods, seeds can also be germinated just as well on soil or filter paper. Third, experiments with plants avoid many ethical concerns. Fourth, when performing experiments on the fast-growing seedlings of the model plant *Arabidopsis thaliana* (*Arabidopsis*) or the oilseed/vegetable crop species *Brassica rapa* (*Brassica*), results can be generated very quickly. This is important given the time constraints of most undergraduate courses. *Arabidopsis* seeds for educators are available from the Arabidopsis Biological Resource Center at the Ohio State University (Columbus, OH), while different sources provide *Brassica* seeds. *Brassica* varieties used by our students included those already in wide circulation among educators for teaching genetics; for example, the fast-cycling Wisconsin Fast Plants from Carolina Biological and their related self-compatible varieties (https://fpsc.wisc.edu/). Finally, it has been recognized that there is a need to improve student perceptions of plant biology. Despite the fact that plants are fundamental for life on this planet as we know it, and that the National Research Council (2009) has identified the development of food plants that grow across changing environments as one of four key challenges facing the next generation of biologists, many students have a distinct disinterest in plant biology and prefer to study animals (Wandersee, 1986; Marbach-Ad, 2004). This phenomenon is known as “plant blindness” (Wandersee and Schussler, 1999). Improving student perceptions toward plants might be achieved through using plants as classroom models to learn general biological principles (Ebert-May and Holt, 2014). Indeed, while plants were our system of choice for all the reasons described, we did not necessarily consider this a “plant biology” course. To highlight the universal applications of NGS technology, we deliberately chose non-plant papers for in-class primary literature readings; were broad in our discussions of how NGS technologies are implemented; used human ethical questions as hooks to capture student interest (Loike *et al.*, 2013); and invited guest lecturers from other fields, including a research scientist and a practicing prenatal genetic counselor who advises patients on test results generated by NGS platforms.

Learning Outcomes

The main core learning outcomes were similar across all sections of Biology 490. It was expected that, after this course, students would be able to 1) design and conduct independent research projects, 2) demonstrate a command of the scientific literature associated with their research topics, 3) show mastery of techniques related to their research, and 4) articulate scientific information orally and in writing. In addition to these core learning outcomes, at the completion of our big data section, students were also expected to be able to 5) explain NGS, that is, what it is, how it is performed, and its various applications; 6) analyze an RNA-seq big data set to find differentially expressed genes and formulate hypotheses in light of the relevant primary literature; 7) demonstrate an understanding of basic plant anatomy and growth responses to environmental stimuli; and 8) demonstrate an ability to independently plan, execute, and document phenotypic and/or molecular–genetic experiments with plants to evaluate gene expression hypotheses.

Course Implementation

A summary of all 16 weeks of in-class activities and graded assessments can be found in Table 1 and Figure 1. Class periods included lectures, guest lectures, journal club readings, bioinformatics analyses, R programming tutorials, and wet-lab work, among other activities. We provide our in-class tutorials and worksheets for most of these activities in the Supplemental Material. These are intended to complement instructor-led discussions and lectures on NGS methodologies. Students were required to have access to their own laptops or institutional computers on which they installed the necessary software (see the Supplemental Material). Wireless Internet access was provided in class. A subscription to The Arabidopsis Information Resource (TAIR; www.arabidopsis.org) was provided to the class free of charge for educational purposes.

Broadly, the class was divided into two lab modules, one computational and the other a wet lab, with significant overlap between the two. Group work and discussion was strongly encouraged at all steps; however, students were expected to plan and execute their own independent experiments within any given group. Assessment activities included both individual and pair grades.

Computer Lab

The *Arabidopsis* genome contains more than 25,000 coding genes (Arabidopsis Genome Initiative, 2000). RNA-seq can be used to assess the expression level of each of these genes simultaneously by collecting millions of sequencing reads of cDNA template generated from the RNA from an *Arabidopsis* tissue sample. The relative number of reads that match a given gene is used as a measure of transcript abundance (Nagalakshmi *et al.*, 2008; Wilhelm *et al.*, 2008). Such RNA-seq data sets are typically extremely large and present a daunting challenge to a novice student: How do you make sense of tens of millions of reads that originate from tens of thousands of genes to find biologically important changes in gene expression across different tissue samples? Students were guided through all aspects of this big data analysis using a combination of free, online cybercomputing infrastructure provided by CyVerse (previously iPlant) and local R-based coding on student or departmental

TABLE 1. Schedule of class activities

Week(s)	Class activities	Graded assessment
1	Lecture: The scientific method in the context of "big data"	
2	Lecture: Sanger sequencing; NGS technologies (Illumina); traditional gene expression analysis vs. RNA-seq; introduction to plants and plant anatomy Journal club: "What became of the Neanderthals? An introduction to NGS" (Green *et al.*, 2010) Wet lab: Plant culture techniques and growth medium preparation	
3	Lecture: Introduction to *Arabidopsis*; RNA-seq experimental design Guest lecture: NGS in the clinic and genetic counseling. Journal club: "How does gene expression change during human embryo development? Visualizing RNA-seq results" (Xue *et al.*, 2013) Computer lab: Introduction to CyVerse; aligning sequencing reads Wet lab: Preparing light chambers; growing plants to test effect of light environment.	
4 and 5	Computer lab: Introduction to R; differential gene expression analysis Wet lab: Scoring the effect of light on plant phenotype	
6	Lecture: Performing qRT-PCR Guest lecture: NGS applications in research (ChIP-Seq) Computer lab: GO analysis and making sense of the data Discussion: Formulating hypotheses and project selection	R script files
7	Peer review and finalization of research proposals	
8–12	Wet lab: Supervised independent research	Literature review and written proposal; theory exam
13	Completion of research and poster preparation	
14	Poster feedback; peer review; finalization; printing	
15	Departmental poster presentation	Poster presentation
16	No class	Gene expression analysis with R assignment; lab notebooks

laptops (Goff *et al.*, 2011; R Core Team, 2015). This approach isolated the largest data files and computational steps to remote servers and avoided many of the difficulties encountered by others in teaching the Linux command line (Peterson *et al.*, 2015). However, unlike other approaches to teach RNA-seq—notably the much-simplified Green Line of the DNA Subway developed by the DNA Learning Center (https://dnasubway.cyverse.org)—our course still guided students through all steps of data analysis and genome read alignment, with an emphasis on graphical interpretations of the data and biological meaning. Our analysis pipeline is shown in Figure 2A. Quality assessment of reads using FastQC, alignment to the genome with TopHat, and counting reads over a given gene with HTSeq were performed using the Discovery Environment in CyVerse (Trapnell *et al.*, 2009; Goff *et al.*, 2011; Anders *et al.*, 2014). The longest computational steps (read alignment and counts) were set to run as homework exercises to be ready for the next class period.

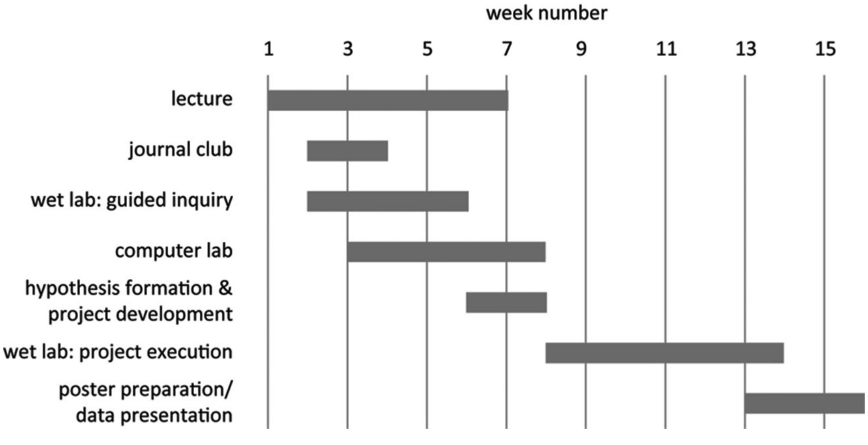

FIGURE 1. Summary of the schedule of class activities.

One of the best ways to make RNA-seq analysis feel less abstract is to visualize it. To do this, students completed an in-class tutorial in pairs using the Integrated Genome Browser to view read alignments to the genome (Freese *et al.*, 2016). This tutorial was intended to serve as a discussion platform of the *Arabidopsis* genome, gene nomenclature in *Arabidopsis*, and how to use TAIR as a resource for finding information on any given gene. For this tutorial, we directed students to genetic intervals that included genes with altered gene expression level across environmental treatments, whose functions matched phenotypic data that the students had collected (Figure 2, B and C). Implementing this particular tutorial required downloading a large representative bam alignment file and its index for both a control and an environmental treatment sample. These files were provided for students on an external USB drive. Otherwise, students needed only to download the very small count files generated by HTSeq for downstream analysis in R. Alternatively, tutorials provided in CyVerse explained how to acquire a URL for the bam alignment file for use with the Integrative Genomics Viewer (Robinson *et al.*, 2011).

edgeR was used for identification of differentially expressed genes (Robinson *et al.*, 2010). After an in-class overview of R and the R Studio environment, students completed an in-class guided R tutorial intended to cover all the basic functions needed in R: reading and writing tables,

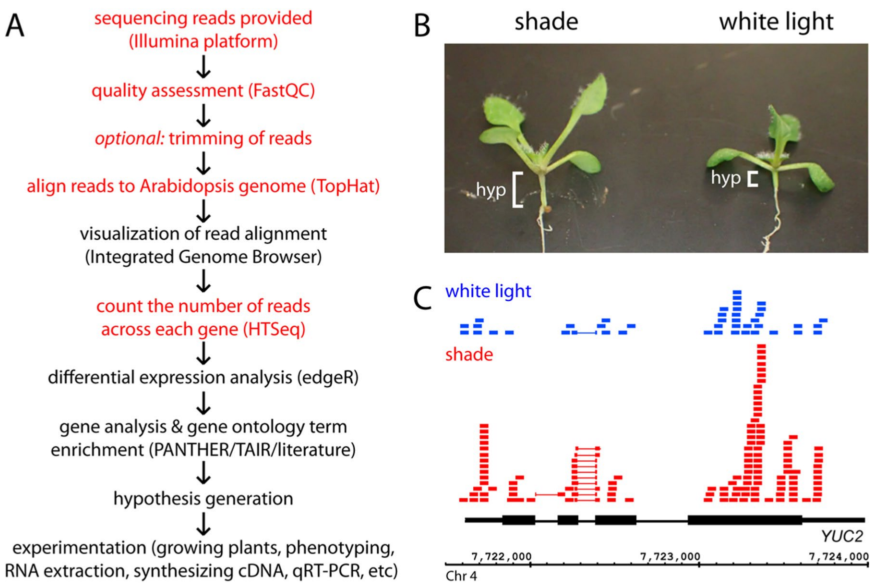
A
sequencing reads provided (Illumina platform)
quality assessment (FastQC)
optional: trimming of reads
align reads to Arabidopsis genome (TopHat)
visualization of read alignment (Integrated Genome Browser)
count the number of reads across each gene (HTSeq)
differential expression analysis (edgeR)
gene analysis & gene ontology term enrichment (PANTHER/TAIR/literature)
hypothesis generation
experimentation (growing plants, phenotyping, RNA extraction, synthesizing cDNA, qRT-PCR, etc)
B
shade
white light
hyp
hyp
C
white light
shade
YUC2
7,722,000
7,723,000
7,724,000
Chr 4

FIGURE 2. Student analysis of gene expression and phenotype of shade-treated *Arabidopsis* seedlings. (A) Flowchart of RNA-seq analysis and follow-up experimentation performed by students. Steps in red were completed using the Discovery Environment in CyVerse. (B) Student photograph showing the phenotype of 10-day-old *Arabidopsis* seedlings grown in white light or shade (5 days white light + 5 days low R:FR light). Note increased elongation growth of the hypocotyl (hyp) in shade. (C) Student worksheets were used to correlate the hypocotyl phenotype with increased expression of genes involved with auxin growth hormone signaling. Shown is cDNA sequencing reads from white-light-treated and shade-treated *Arabidopsis* seedlings, aligned over the *YUC2* gene (boxes, exons). *YUC2* codes for an enzyme involved with auxin biosynthesis (Mashiguchi *et al.*, 2011). Students noted the increased number of reads in the shade-treated plants.

column and row arithmetic, statistical testing, and graphing and indexing (see the Supplemental Material and the accompanying data spreadsheet available through our class materials on CyVerse: http://datacommons.cyverse.org/browse/iplant/home/shared/USD_teaching_materials). As many of our students were interested in ecology, for this tutorial we used a data set describing the patterns of movement of local ocean wildlife, with personal instructor stories and images (including a shark-eating octopus!) to aid student engagement and friendly discourse with the instructors. From here, students worked in pairs through a series of tutorials to find differentially regulated genes (see the Supplemental Material). Students were allowed to work through each tutorial at their own pace, but were not given the next tutorial until all students were ready. This ensured that no particular group got too far ahead and that in-class questions and the resulting discussions were relevant to all students. At the completion of the tutorials, students submitted their R scripts for assessment. As a guide to instructors, we provide in our class materials on CyVerse an R markdown and html file for use with our instructor-led example data set.

Wet Lab

Students in our class analyzed gene expression in plants in response to environmental change. Accompanying this analysis, all students were tasked early during the semester with growing *Arabidopsis* plants and measuring basic phenotypic responses to the environmental condition from which the RNA-seq data were generated. This approach is similar to that reported by other educators using maize (Makarevitch *et al.*, 2015). For example, we simulated shade by reducing the ratio of red (R) to far-red (FR) light (see the Supplemental Material). Shade treatment results in large phenotypic effects on *Arabidopsis* and related plant species that are easily scored by students using basic instrumentation (Figure 2B). Our protocols for growing plants in tissue culture or on soil are available in the Supplemental Material. FR light was provided using fixed-wavelength LED bulbs (LumiGrow ECC-FR or PAR-source PowerPAR Far Red LED bulbs).

By the end of the RNA-seq analysis, students had generated lists of differentially regulated genes in *Arabidopsis* in response to the same environmental treatment, made various graphs to show these changes, performed a gene ontology (GO) enrichment analysis using PANTHER to look for pathways that may be overrepresented among the differentially expressed genes (Mi *et al.*, 2013), and investigated the function of interesting genes using TAIR. In-class tutorials guided students to look for likely causal links between gene expression changes and phenotypes that could be explored further. On the basis of these observations, students were then free to choose their own independent follow-up studies using plants. This was in part to fulfill department objectives that students taking Biology 490 generate original data. In addition, we thought it was important that students choose their own original hypothesis to test in planta on the basis that some autonomous control over the direction of their projects would lead to the positive attributes associated with project ownership (Hanauer *et al.*, 2012). With such large data sets from which to draw inspiration, and as a consequence of this approach, student projects were diverse and depended on individual student interests. Most involved finding homologous gene sequences for interesting genes in a plant species other than *Arabidopsis* using online databases. For example, five of eight students chose to test a hypothesis using the vegetable and oilseed crop species *B. rapa* (e.g., see Example Project 1 in the Supplemental Material). This may reflect that students find crop species more immediately relevant than *Arabidopsis*. Students then designed qRT-PCR experiments to detect gene transcripts in the novel species. Our protocols for RNA extraction, cDNA synthesis, primer design, and qRT-PCR measurements of gene expression are provided in the Supplemental Material. As an aid to other educators, we also provide therein tutorials for the study of plant responses to shade and two examples of individual student projects to demonstrate the quality and scale of the projects students completed during the semester, and the methods used.

RNA-Seq Data

Students analyzed one of two different RNA-seq data sets reporting changes in gene expression in *Arabidopsis* to two different environmental stimuli. These data sets were generated on the Illumina platform (stranded mRNA libraries, single-end sequencing, 50 base reads). One of these data sets was of 5-day-old *Arabidopsis* seedlings following a 4-hour low-R:FR light treatment (Procko *et al.*, 2016) and is the focus of this paper. The other data set is currently unpublished, but was similar, in that it was generated from *Arabidopsis* seedlings responding to a light treatment, specifically, high light. To avoid confusion over batch effects, students analyzed experimental and control duplicates from RNA samples harvested on the same day only. For instructor demonstrations, a third RNA-Seq data set reporting changes in gene expression in 3-day-old *Arabidopsis* seedlings in response to 1 hour of 10 μM abscisic acid (ABA) treatment was used (Song *et al.*, 2016). ABA is a plant hormone involved with mediating responses to drought and other abiotic stressors. Sequencing read files can be found at the Gene Expression Omnibus (accession numbers GSE79881 and GSE80568). For ease of use by other instructors, we have also made the relevant files for ABA and low-R:FR shade responses publicly available on CyVerse (http://datacommons.cyverse.org/browse/iplant/home/shared/USD_teaching_materials).

Plant Strains

Arabidopsis strains used in our class were Col-0 (wild type), *phyB-9* mutant, and *gi* mutant (*Salk_092757*). *Brassica rapa* strains used were R-o-18 (yellow sarson); FPsc, a rapid-cycling self-compatible variety; and the FPsc *phyB* mutant *ein194* (https://fpsc.wisc.edu).

Assessment

A challenge of bioinformatics courses is the design of quality assessment tools, in part due to the disparate methods of bioinformatics and the rapid progress in the field (Campbell and Nehm, 2013; Magana *et al.*, 2014). We designed formative assessment activities to fulfill departmental requirements that students taking Biology 490 undertake and present an independent research project. Additional activities tested cognitive and psychomotor skills specific to our section of the course, namely RNA-seq analysis and computer coding. We did not administer a final cumulative exam, but rather assessed our learning objectives by student submission of R files, a short midterm theory exam focused on NGS and how these big data sets are used, a final R assignment, a written literature review and project proposal, a poster presentation, lab performance, and a lab manual documenting experimental design and implementation (Table 2).

There are few published assessment tools to evaluate student learning of RNA-seq theory and its analysis (Makarevitch *et al.*, 2015). To assess NGS theory, we administered an in-class midterm exam (see the Supplemental Material). Questions were written by the instructor (C.P.) most knowledgeable in the field and actively using NGS methods in their research and were edited by S.M. and others to improve student readability. The exam was not intended to be difficult, and no student took longer than 1 hour to complete it. Rather, in addition to the recall and description of facts, we tested higher-order cognitive and critical-thinking skills by asking our students to apply their knowledge to previously unseen sets of mock data and to evaluate and make predictions based on these data (Bloom *et al.*, 1956; Crowe *et al.*, 2008). Our small class size made it feasible to administer this exam using mostly short-answer, constructed response–style questions, which are better suited to assessing student creativity and critical thinking than multiple choice (Martinez, 1999). In addition, the inclusion of constructed-response questions on exams has been demonstrated to correlate with study behaviors that are more cognitively active and in which the interrelationships of facts are emphasized (Stanger-Hall, 2012). Students were prepared for these exams and the style of questions by in-class discussions and examinations of selected journal club papers in which we broke down figures within papers to understand how RNA-seq data are presented and what hypotheses they can test.

We strongly argue that the only way to assess student proficiency in RNA-seq analysis using the coding language R is to do it: to write, run, execute, and debug code to achieve a particular goal. To this end, not only did we require students to submit their R script files from the in-class tutorials, but we also administered a final take-home big data coding assignment in lieu of a cumulative exam or final written report (see the Supplemental Material). This assignment required students to download a list of gene expression values and to manipulate the data to draw graphs and form conclusions. While we believe that coding is an active endeavor and that discussion with classmates is something to be encouraged, to ensure some independence on the assignment, we had each student work with one of five different gene lists, with each list representing thousands of genes from

TABLE 2. Learning outcomes and associated assessment tools

Learning outcome	Graded assessment
1. Design and conduct an independent research project.	Literature review and written proposal; participation and lab performance; lab notebook
2. Demonstrate a command of the scientific literature associated with research topic.	Literature review and written proposal
3. Show mastery of techniques related to research.	Participation and lab performance; lab notebook
4. Articulate scientific information orally and in writing.	Literature review and written proposal; poster presentation
5. Explain NGS: what it is, how it is performed, and its various applications.	Techniques and theory exam
6. Analyze an RNA-seq data set to find differentially expressed genes and formulate hypotheses in light of the relevant primary literature.	R script and bioinformatics analyses; gene expression with R assignment
7. Demonstrate an understanding of basic plant anatomy and growth responses to environmental stimuli.	Literature review and written proposal; poster presentation
8. Demonstrate an ability to independently plan, execute, and document phenotypic and/or molecular–genetic experiments with plants to evaluate gene expression hypotheses.	Participation and lab performance; lab notebook

one of the five *Arabidopsis* chromosomes. Our focus was to award completion of the activity, and, as such, all students started with a grade of 100% for assignment submission, with percentage points deducted for inaccuracies based on severity. "Severity" was a measure with some subjectivity, as it was impossible for us to predict all the ways a student might go wrong. For example, a simple error wherein a student erroneously counted a header row of a table as a gene row lost only 1 percentage point, while an error that confused values of gene expression (in the form of read counts per million, or cpm) with fold change was deemed a greater conceptual mistake and penalized more.

Assignment essays in the form of a literature review are also useful tools to assess higher-order cognitive-thinking skills and are tools that can encourage students to employ deep-learning strategies that try to integrate facts and course components into a higher understanding (Scouller, 1998). We challenged students to review and evaluate a field of literature pertinent to their proposed experiments and used this tool as a means to evaluate higher-order cognitive skills such as synthesis and evaluation (Bloom *et al.*, 1956; Crowe *et al.*, 2008). A departmental grading rubric was provided to highlight our desire for students to synthesize the literature to produce relevant, interesting scientific "stories" that looked for gaps in the field and justified their proposed wet-lab experiments (see the Supplemental Material). In addition to this activity, a further opportunity to evaluate their own work was then provided in the form of a culminating, open departmental poster session that included presentations by students from other Biology 490 sections in addition to our own. Grading for posters followed a standard departmental rubric (see the Supplemental Material). To avoid subjective bias in grading the literature review and poster presentation, these items were rated by both class instructors and agreement was reached as to where a student's effort fell within the grading rubric. A consensus through deliberation was reached in the case of instructor disagreement. Our small class size also allowed for students to challenge grades with reason for re-evaluation by one or both instructors.

Depending on the research environment, lab notebooks take many different forms and increasingly include collections of digital data as well as handwritten notes. Thus, the only requirement for the lab notebooks was that students needed to provide descriptions of their experiments so that an educated observer would be able to replicate all necessary components and a legible record of their results in conjunction with the necessary digital files. A grade of "excellence" indicated the lab notebook fulfilled these requirements, while a "satisfactory" grade or lower reflected that a student might have omitted certain details of the experiment or descriptions of the data might only have been interpretable by the student. These grades and "lab performance" (admittedly a rather subjective measure of commitment to task) were also determined through discussion by both instructors to minimize bias and improve reliability and consistency. While we reported feedback to students on all aspects of their grades and were vocal in our expectations at the start of the semester and throughout, future iterations of the course could be improved by providing a written rubric of our expectations for lab performance in particular. However, some subjectivity cannot be eliminated.

In addition to our graded activities, we also sought to assess the learning of our students ($n = 8$) by comparing them against those of other Biology 490 sections or students taking research internships ($n = 21$). Owing to in-class time constraints, this consisted of using a brief, multiple-choice quiz consisting of 10 questions that mostly assessed factual recall (see the Supplemental Material). This quiz was executed 3 months after course completion by sampling graduating seniors taking the Biology Capstone Seminar course and did not form a part of any student's grade. Quiz questions covered the scientific method—which we argued all students should have learned in lower-level classes and as part of their particular research experience—as well as questions that might be more specific to our particular section of Biology 490–Research Project: gene expression, genome evolution, and RNA-seq theory. Questions were largely taken or modified from other reported assessment quizzes (Couch *et al.*, 2015; Makarevitch *et al.*, 2015).

Finally, we were also interested in monitoring student perceptions and bioinformatics self-efficacy. This we did using anonymous pre-and postclass questionnaires (see the Supplemental Material) and student write-in comments on class evaluations.

RESULTS

Students came into the course with minimal experience in coding, and none in big data analysis (see *Methods*). Multiple forms of assessment were used to track student proficiency in these topics and the associated wet-lab experimentation, including a written report, poster presentation, and submission of R script files and lab notebooks (see *Methods*). All students successfully completed the R coding tutorials and assignment, and all scored grades of "excellence" for their culminating poster presentations, which demonstrated their ability to integrate the RNA-seq data with their in planta experiments in poster format and sufficiently communicate this material to others (Table 3).

As described earlier, we also executed a short, 10-question factual-recall postclass quiz covering various aspects of our course (see *Methods*, Supplemental Table 1, and the Supplemental Material), in which we compared our students with seniors who were enrolled in other Biology 490 sections or, in lieu of this, had completed research internships. We reasoned that our comparison group would control for general knowledge of a USD senior student majoring in biology. Initially, however, we were surprised to see that our students performed only marginally better on this quiz than the control group (74 ± 5% vs. 65 ± 3% [mean ± SE]; not significant, Wilcoxon-Mann-Whitney test). This might reflect the limitations of our small sample size (due to the fact that USD is a small university with small

TABLE 3. Student scores

Activity	% of overall grade	Mean score[a]
Participation and lab performance	20	88.6
Literature review and written proposal	20	85.6
Poster presentation	20	95.4
R script and bioinformatics analyses	10	95.4
Lab notebook	10	87.5
Gene expression with R assignment	10	96.7
Techniques and theory exam	10	88.2

[a]Mean score represents the mean across all students in the class, as a value out of 100: >90 represents a letter grade of "A" (excellence), 80–90 represents a "B" (satisfactory), and 70–80 represents a "C."

TABLE 4. Mean pre- and postclass scores for bioinformatics self-efficacy

Item	Preclass[a]	Postclass[a]	*p* value[b]
I am comfortable with statistical analyses in biology.	2.625	3.875	0.029
I am knowledgeable in computer programming.	2.25	3.375	0.026
I can use bioinformatics tools to answer biological questions.	2.5	4.125	0.00093
Molecular genetics and genomics excite me!	3.875	4.3125	0.19
I am likely to use R in the future for data analysis.	NA[c]	3.0625	NA[c]

[a] Students anonymously self-reported levels of agreement to each item statement. The response format was: strongly disagree (coded 1), disagree (2), neutral (3), agree (4), and strongly agree (5). Mean scores are shown.

[b] Significance was determined by Wilcoxon-Mann-Whitney test.

[c] NA, not applicable. This question was asked only at the completion of the class when all students had acquired some knowledge of R coding.

classes) and the limited number of questions in our postclass quiz due to time constraints. However, we noticed that our students performed surprisingly poorly on the few questions devoted to the scientific method (questions 1–3), in which we had expected all students who had undertaken a research experience to perform at least equally well. This might indicate that other sections of Biology 490 emphasized these skills and hypothesis-driven science more than we did, partly on our assumption that students already had a strong knowledge of the scientific method and the definition of a hypothesis before taking our class. This might be rectified in future iterations of this course by emphasizing the role of RNA-seq as a hypothesis generator within the traditional context of the scientific method. On the remaining seven questions pertaining to molecular genetics, genome evolution, and molecular techniques, which are ideas all relevant to either RNA-seq theory or the accomplishment of our students' research projects, our students significantly outperformed others (82.1 ± 0.06% vs. 63.9 ± 0.04% [mean ± SEM], $p < 0.05$, Wilcoxon-Mann-Whitney test). The question that showed the greatest differential between the two groups was the only question pertaining directly to RNA-seq (question 10), of which seven of our eight students answered the question correctly versus only seven of 21 students from the control group ($p < 0.05$, Fisher's exact test).

On pre- and postclass perception surveys designed to test changes in self-efficacy, students self-reported increased confidence using coding and bioinformatics tools to address biological questions (Table 4). Students also showed a slight but nonsignificant increase in their positive feelings toward molecular genetics and genomics, which might reflect already high student interest in these topics before completing the class and/or our small *n* value. While these self-efficacy surveys were encouraging, they hide some student frustration. Student write-in comments on class evaluations ranged from extremely positive ("Best course at USD"!) to some feeling "overwhelmed" by R. We reviewed student comments and identified those that appeared most positive and most negative, and we present these in Table 5 to show the range of feelings students expressed about the course. In addition, three of eight students felt that the midterm exam was unnecessary in the context of the number of other assessment activities, and two reported that the "workload (homework, papers, exams, etc.)" was "too demanding." Five students felt that the workload was "about right," and one student wavered in-between. By contrast, 14 of 15 students queried from other Biology 490 sections felt that the workload of their section was "about right," and only one judged it "too demanding" ($p = 0.10$, Fisher's exact test). These thoughts might be reflected in some negative feelings toward the time spent on wet-lab work in our section, albeit balanced by other students who "enjoyed" the novelty of the research experience (Table 5).

DISCUSSION

A challenge for biology departments—particularly at small undergraduate institutions like USD—is to teach students quantitative big data biology that reflects the current state of the life sciences. Our anecdotal experience suggests that students are often happy to discuss the ethical issues surrounding this new era of genetic information, yet have minimal understanding of the technologies that have brought about this revolution and how they are implemented. Here, we describe the design and implementation of a CURE that teaches big data NGS concepts and analysis using RNA-seq, from which students integrate this analysis with in planta observations and hypothesis testing. Student data were presented to the department in the form of a poster session and later capstone seminar, and some student experiments will continue to be followed in a research laboratory setting.

Our approach to teaching RNA-seq lies somewhere between the Green Line of DNA Subway (https://dnasubway.cyverse.org), which was developed to ease educators into introducing RNA-seq to the classroom, and the more advanced tutorials developed by the Genome Consortium for Active Teaching using Next-Generation Sequencing (GCAT-SEEK; Peterson *et al.*, 2015). The Green Line is the easiest approach, but, while useful for educators facing steep time constraints, limits user control over the data and confines the output to a series of predetermined graphs and spreadsheets, pushing much of the analysis into a "black box" (Makarevitch and Martinez-Vaz,

TABLE 5. Examples of postclass student comments

Class component	Student comments
General/computer lab	Positive: "Best course at USD; learned the most." Negative: "We went through R too fast and it was overwhelming."
Wet lab	Positive: "I enjoyed the research that we did and the fact that it is the first time being done at USD." Negative: "Research requirement … requires too much class time as well as time out of class."

2017). Furthermore, learning to manipulate large sets of biological data using coding languages such as R is a valuable learning objective unfulfilled by the Green Line. By contrast, tutorials on the other end of the spectrum provided by GCAT-SEEK require Linux command line skills in addition to R, and significant teacher investment in class preparation. By using the Discovery Environment through CyVerse in combination with R, we have deliberately avoided the complexities of running analyses through the Linux command line, while still emphasizing the need to learn some coding for big data analysis.

Other published approaches to teaching RNA-seq have encouraged students to computationally test preformed hypotheses using the data sets analyzed in class or to test follow-up hypotheses computationally with the same or other data collections (Makarevitch *et al.*, 2015; Peterson *et al.*, 2015). Ours is one of the first that we know of to combine computational analysis with student-driven wet-lab experimentation. While this approach is intended to target many of the positive attributes associated with CUREs, it requires significant teacher investment in both the design of RNA-seq analysis tutorials and instructional support for the wet-lab component. It is hoped that the availability of class materials such as ours will improve instructor confidence in tackling large computational analyses in the classroom. Future implementations of the course may assess instructor confidence to test this. However, this does not absolve the challenge of leading students through technically challenging molecular protocols, such as RNA extraction and quantitative PCR. Our course was taught by two adjunct professors, one with expertise in RNA-seq analysis and the other with greater experience teaching laboratory classes. This melding of expertise no doubt contributed to the first-round success of this course; however, we note that most larger institutions will have higher student-to-teacher ratios, and this course might be best adapted to a more streamlined inquiry-driven class with predetermined experiments in these situations. In addition, we sometimes had difficulties dividing instructor time between eight students working on eight different projects, and one student noted there were "days where time was not utilized effectively." While our approach was taken in part to fulfill departmental requirements that each student have his or her own original data set to present, our opinion is that the class could be improved with students working in larger groups on the same project, with more instructor time devoted to each group. This would also improve issues with scheduling student access to shared equipment, while all the positive aspects of CUREs and student hypothesis development and project "ownership" are still provided.

Our assessment tools and survey suggest that students acquired knowledge in programming and increased bioinformatics self-efficacy. Self-efficacy—the judgment one has about one's own capabilities toward a particular task (Bandura, 1977)—is an important indicator of student success. Students who report higher self-efficacy show greater optimism and are more persistent at tasks. These traits correlate with the higher academic success of these students and retention in their chosen academic disciplines (Lent *et al.*, 1986, 2001; Lau and Roeser, 2002). A major source of increased self-efficacy is the accomplishment of similar tasks in the past, or so-called mastery experiences (Usher and Pajares, 2008). It is likely that the increased scores of self-efficacy reported by our students in our pre- and postclass perception surveys (Table 4) were due to their completion of the class assignments and coding tutorials. We hope that this increased self-efficacy will translate to a greater confidence in our students to tackle similar and new bioinformatics-related problems in the future. However, student postclass comments also showed some frustration with R; specifically, three of eight students described R coding as "overwhelming." A broad spectrum of student comments toward learning R has also been reported by others teaching RNA-seq (Makarevitch *et al.*, 2015; Peterson *et al.*, 2015). We propose that, to mitigate these challenges, biology students should ideally be exposed to these skills earlier in their undergraduate education through similar or other bioinformatics/biostatistics courses. A consistent application of these skills during undergraduate education should in principle better prepare senior students for courses like ours and the new reality of the postgraduate research enterprise.

In their follow-up studies, many of our students used qRT-PCR methods to follow a gene of interest in *Arabidopsis* or another plant species. This technique allows the measurement of transcript abundance of a chosen gene across biological samples using real-time PCR (Bustin, 2000). Despite the importance of real-time PCR to estimate DNA concentrations in modern academic and industrial molecular biology laboratories, very few papers describe the successful implementation of qRT-PCR modules in the classroom (Hancock *et al.*, 2010; Makarevitch and Martinez-Vaz, 2017). This may in part be due to experimental complexity, inhibitory costs, or lack of access to the necessary equipment at many undergraduate institutions. Indeed, while most students in our class had previously performed a PCR, only one had prior experience with quantitative PCR through a research internship. Here, we used the double-stranded DNA binding dye SYBR Green and the ΔΔCt approach to estimate the fold change in transcript abundance for a student's chosen gene (Hancock *et al.*, 2010). The time that students needed to treat plants, extract RNA, generate cDNA, and perform qRT-PCR was approximately 3 to 4 weeks, which is similar to the time frame described by Hancock and colleagues (2010) in their teaching module using primer pairs predetermined by the instructor. However, our students needed some additional time to design and test novel primer pairs and repeat experiments when initial results were ambiguous. In addition, while we discussed the theory behind normalizing transcript abundance to a constitutively expressed housekeeping gene, in the interest of time, we used BioRad's CFX Manager software to calculate the fold change in gene expression rather than generating ΔΔCt values manually.

Our course is easy to adapt to guided inquiry, which will likely be more appropriate at institutions with larger student-to-teacher ratios or who do not have access to real time PCR thermocyclers or the financial resources for some of the experiments discussed herein. Plant responses to shade are an excellent platform for guided inquiry: *Arabidopsis* and *Brassica* seedlings are easy to grow in a short period of time, and many schools already have the requisite skills and materials; the R/FR light environment is easy to manipulate; and the hypocotyl elongation phenotype is easy for students to measure and conceptually place in an ecological context (Figure 2B). Furthermore, shade induces obvious changes in auxin biosynthesis and auxin target-gene expression in the RNA-seq data, which can explain the observed growth phenotype (Figure 2C). If instructors so choose, students can then test the hypothesis that increased auxin levels are causative for the increased growth in shade by manipulating the auxin pathway with application of exogenous auxin transport inhibitors or comparing *Arabidopsis* wild-type seedlings to any of many auxin-deficient mutants available from the Arabidopsis Biological Resource Center (https://abrc.osu.edu; Tao *et al.*, 2008). While we find shade responses in plants particularly compelling for these reasons, our tutorials and approach to teaching RNA-seq can be applied to many other published data sets (or instructor-generated data) that use model organisms suitable for classroom instruction. For example, single-celled organisms such as yeast and bacteria are easy for students to grow and manipulate, and gene expression changes might be matched to growth curves or other phenotypes.

We present here a CURE that combines quantitative assessment of NGS data with wet-lab experimentation that tests student-generated hypotheses. We provide our RNA-seq analysis tutorials and make suggestions for how the course can be adapted by other institutions depending on their needs and constraints. Our assessment tools suggest that students gained experience in R-based coding and showed subjective and objective improvements in bioinformatics skills and knowledge of molecular genetics. Introducing biology students to coding even earlier in their undergraduate education would likely further improve student performance and satisfaction in quantitative biology courses and better prepare students for the "big data" era of biology.

ACKNOWLEDGMENTS

We thank Lisa Baird and Joseph Provost for discussion; Kate Boersma and the USD Department of Biology for sharing grading rubrics and materials; Adam Seluzicki, Laura Krenik, and Tyler McElroy for comments on the article; Jianyan Huang, Liang Song, Scott Woody, and Björn Willige for reagents; and Matthew Spees for contributing example data and comments. We are also grateful for extensive comments from Erin Dolan and anonymous reviewers during the review process. This material is based on work supported by National Institutes of Health (NIH) awards 1F32GM101876 (C.P.) and 5R35GM122604 (J.C.), the Howard Hughes Medical Institute (J.C.), and the National Science Foundation under award numbers DBI-0735191 and DBI-1265383 (www.cyverse.org).

REFERENCES

American Association for the Advancement of Science. (2011). *Vision and change in undergraduate biology education: A call to action*. Washington, DC.

Anders, S., Pyl, P. T., & Huber, W. (2014). HTSeq—A Python framework to work with high-throughput sequencing data. *Bioinformatics*, *31*, 166–169.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana. Nature*, *408*, 796–815.

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., ... Dolan, E. L. (2014). Assessment of course-based undergraduate research experiences: A meeting report. *CBE—Life Sciences Education*, *13*, 29–40.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215.

Bangera, G., & Brownell, S. E. (2014). Course-based undergraduate research experiences can make scientific research more inclusive. *CBE—Life Sciences Education*, *13*, 602–606.

Bialek, W., & Botstein, D. (2004). Introductory science and mathematics education for 21st-century biologists. *Science*, *303*, 788–790.

Bloom, B. S., Krathwohl, D. R., & Masia, B. B. (1956). *Taxonomy of educational objectives: The classification of educational goals*. New York: David McKay.

Buonaccorsi, V., Peterson, M., Lamendella, G., Newman, J., Trun, N., Tobin, T., ... Roberts, W. (2014). Vision and Change through the Genome Consortium for Active Teaching Using Next-Generation Sequencing (GCAT-SEEK). *CBE—Life Sciences Education*, *13*, 1–2.

Bustin, S. (2000). Absolute quantification of mRNA using real-time reverse transcription polymerase chain reaction assays. *Journal of Molecular Endocrinology*, *25*, 169–193.

Campbell, C. E., & Nehm, R. H. (2013). A critical analysis of assessment quality in genomics and bioinformatics education research. *CBE—Life Sciences Education*, *12*, 530–541.

Couch, B. A., Wood, W. B., & Knight, J. K. (2015). The Molecular Biology Capstone Assessment: A concept assessment for upper-division molecular biology students. *CBE—Life Sciences Education*, *14*, ar10.

Crowe, A., Dirks, C., & Wenderoth, M. P. (2008). Biology in Bloom: Implementing Bloom's taxonomy to enhance student learning in biology. *CBE—Life Sciences Education*, *7*, 368–381.

Ebert-May, D., & Holt, E. (2014). Seeing the forest and the trees: Research on plant science teaching and learning. *CBE—Life Sciences Education*, *13*, 361–362.

Ellington, R., Wachira, J., & Nkwanta, A. (2010). RNA secondary structure prediction by using discrete mathematics: An interdisciplinary research experience for undergraduate students. *CBE—Life Sciences Education*, *9*, 348–356.

Freese, N. H., Norris, D. C., & Loraine, A. E. (2016). Integrated genome browser: Visual analytics platform for genomics. *Bioinformatics*, *32*, 2089–2095.

Goff, S. A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A. E., Gessler, D., ... Stanzione, D. (2011). The iPlant Collaborative: Cyberinfrastructure for plant biology. *Frontiers in Plant Science*, *2*, 34.

Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U., Kircher, M., ... Paabo, S. (2010). A draft sequence of the Neandertal genome. *Science*, *328*, 710–722.

Gross, L. J. (2000). Education for a biocomplex future. *Science*, *288*, 807–807.

Hanauer, D. I., Frederick, J., Fotinakes, B., & Strobel, S. A. (2012). Linguistic analysis of project ownership for undergraduate research experiences. *CBE—Life Sciences Education*, *11*, 378–385.

Hancock, D., Funnell, A., Jack, B., & Johnston, J. (2010). Introducing undergraduate students to real-time PCR. *Biochemistry and Molecular Biology Education*, *38*, 309–316.

Jones, M. T., Barlow, A. E. L., & Villarejo, M. (2010). Importance of undergraduate research for minority persistence and achievement in biology. *Journal of Higher Education*, *81*, 82–115.

Kardash, C. M. (2000). Evaluation of undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *Journal of Educational Psychology*, *92*, 191–201.

Kodama, Y., Shumway, M., & Leinonen, R. (2011). The sequence read archive: Explosive growth of sequencing data. *Nucleic Acids Research*, *40*, D54–D56.

Lau, S., & Roeser, R. W. (2002). Cognitive abilities and motivational processes in high school students' situational engagement and achievement in science. *Educational Assessment*, *8*, 139–162.

Lent, R. W., Brown, S. D., Brenner, B., Chopra, S. B., Davis, T., Talleyrand, R., & Suthakaran, V. (2001). The role of contextual supports and barriers in the choice of math/science educational options: A test of social cognitive hypotheses. *Journal of Counseling Psychology*, *48*, 474–483.

Lent, R. W., Brown, S. D., & Larkin, K. C. (1986). Self-efficacy in the prediction of academic performance and perceived career options. *Journal of Counseling Psychology*, *33*, 265–269.

Lewis, J., & Kattmann, U. (2004). Traits, genes, particles and information: Re-visiting students' understandings of genetics. *International Journal of Science Education*, *26*, 195–206.

Loike, J. D., Rush, B. S., Schweber, A., & Fischbach, R. L. (2013). Lessons learned from undergraduate students in designing a science-based course in bioethics. *CBE—Life Sciences Education*, *12*, 701–710.

Lopatto, D. (2007). Undergraduate research experiences support science career decisions and active learning. *CBE—Life Sciences Education*, *6*, 297–306.

Lopatto, D., Alvarez, C., Barnard, D., Chandrasekaran, C., Chung, H. M., Du, C., … Elgin, S. C. (2008). Undergraduate research. Genomics Education Partnership. *Science*, *322*, 684–685.

Magana, A. J., Taleyarkhan, M., Alvarado, D. R., Kane, M., Springer, J., & Clase, K. (2014). A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research. *CBE—Life Sciences Education*, *13*, 607–623.

Makarevitch, I., Frechette, C., & Wiatros, N. (2015). Authentic research experience and "big data" analysis in the classroom: Maize response to abiotic stress. *CBE—Life Sciences Education*, *14*(3), ar27.

Makarevitch, I., & Martinez-Vaz, B. (2017). Killing two birds with one stone: Model plant systems as a tool to teach the fundamental concepts of gene expression while analyzing biological data. *Biochimica et Biophysica Acta*, *1860*, 166–173.

Marbach-Ad, G. (2004). Expectations and difficulties of first-year biology students. *Journal of College Science Teaching*, *33*, 18–23.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207–218.

Mashiguchi, K., Tanaka, K., Sakai, T., Sugawara, S., Kawaide, H., Natsume, M., … Kasahara, H. (2011). The main auxin biosynthesis pathway in *Arabidopsis*. *Proceedings of the National Academy of Sciences USA*, *108*, 18512–18517.

Mi, H., Muruganujan, A., Casagrande, J. T., & Thomas, P. D. (2013). Large-scale gene function analysis with the PANTHER classification system. *Nature Protocols*, *8*, 1551–1566.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, *320*, 1344–1349.

National Research Council. (2009). *A new biology for the 21st century*. Washington, DC: National Academies Press.

Peterson, M. P., Malloy, J. T., Marden, J. H., & Buonaccorsi, V. P. (2015). Teaching RNAseq at undergraduate institutions: A tutorial and R package from the Genome Consortium for Active Teaching. CourseSource. doi: https://doi.org/10.24918/cs.2015.14

Procko, C., Burko, Y., Jaillais, Y., Ljung, K., Long, J. A., & Chory, J. (2016). The epidermis coordinates auxin-induced stem growth in response to shade. *Genes & Development*, *30*, 1529–1541.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved October 17, 2015, from www.R-project.org/

Reinagel, A., & Bray Speth, E. (2016). Beyond the central dogma: Model-based learning of how genes determine phenotypes. *CBE—Life Sciences Education*, *15*, ar4.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*, 24–26.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140.

Russell, S. H., Hancock, M. P., & McCullough, J. (2007). The pipeline. Benefits of undergraduate research experiences. *Science*, *316*, 548–549.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, *35*, 453–472.

Shaffer, C. D., Alvarez, C. J., Bednarski, A. E., Dunbar, D., Goodman, A. L., Reinke, C., … Elgin, S. C. R. (2014). A course-based research experience: How benefits change with increased investment in instructional time. *CBE—Life Sciences Education*, *13*, 111–130.

Song, L., Huang, S. C., Wise, A., Castanon, R., Nery, J. R., Chen, H., & Ecker, J. R. (2016). A transcription factor hierarchy defines an environmental stress response network. *Science*, *354*, aag1550.

Stanger-Hall, K. F. (2012). Multiple-choice exams: An obstacle for higher-level thinking in introductory science classes. *CBE—Life Sciences Education*, *11*, 294–306.

Tao, Y., Ferrer, J. L., Ljung, K., Pojer, F., Hong, F., Long, J. A., … Chory, J. (2008). Rapid synthesis of auxin via a new tryptophan-dependent pathway is required for shade avoidance in plants. *Cell*, *133*, 164–176.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics*, *25*, 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., … Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*, 562–578.

Usher, E. L., & Pajares, F. (2008). Sources of self-efficacy in school: Critical review of the literature and future directions. *Review of Educational Research*, *78*, 751–796.

Wandersee, J. H. (1986). Plants or animals—Which do junior high school students prefer to study? *Journal of Research in Science Teaching*, *23*, 415–426.

Wandersee, J. H., & Schussler, E. E. (1999). Preventing plant blindness. *American Biology Teacher*, *61*, 82–86.

Wilhelm, B. T., Marguerat, S., Watt, S., Schubert, F., Wood, V., Goodhead, I., … Bahler, J. (2008). Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, *453*, 1239–1243.

Xue, Z., Huang, K., Cai, C., Cai, L., Jiang, C. Y., Feng, Y., … Fan, G. (2013). Genetic programs in human and mouse early embryos revealed by single-cell RNA sequencing. *Nature*, *500*, 593–597.