Published in final edited form as:

Yeast. 2016 February ; 33(2): 43–53. doi:10.1002/yea.3139.

Implementation and assessment of a yeast orphan gene research project; involving undergraduates in authentic research experiences and progressing our understanding of uncharacterized open reading frames

Bethany V. Bowling, Ph.D.,

Northern Kentucky University, Department of Biological Sciences

Patrick J. Schultheis, Ph.D., and

Northern Kentucky University, Department of Biological Sciences

Erin D. Strome, Ph.D.

Northern Kentucky University, Department of Biological Sciences, Highland Heights, KY 41099

Bethany V. Bowling: bowlingb2@nku.edu; Patrick J. Schultheis: schultheisp@nku.edu; Erin D. Strome: stromeel@nku.edu

Abstract

Saccharomyces cerevisiae was the first eukaryotic organism to be sequenced, however little progress has been made in recent years in furthering our understanding of all open reading frames (ORFs). From October 2012 to May 2015 the number of verified ORFs has only risen from 75.31% to 78% while the number of uncharacterized ORFs have decreased from 12.8% to 11% (representing more than 700 genes still left in this category) [<http://www.yeastgenome.org/genomesnapshot>]. Course-based research has been shown to increase student learning while providing experience with real scientific investigation; however, implementation in large, multi-section courses presents many challenges. This study sought to test the feasibility and effectiveness of incorporating authentic research into a core genetics course with multiple instructors to increase student learning and progress our understanding of uncharacterized ORFs. We generated a module-based annotation toolkit and utilized easily accessible bioinformatics tools to predict gene function for uncharacterized ORFs within the *Saccharomyces* Genome Database (SGD). Students were each assigned an uncharacterized ORF which they annotated using contemporary comparative genomics methodologies including multiple sequence alignment, conserved domain identification, signal peptide prediction and cellular localization algorithms. Student learning outcomes were measured by quizzes, project reports and presentations, as well as a post-project questionnaire. Our results indicate the authentic research experience had positive impacts on student's perception of their learning and their confidence to conduct future research. Furthermore we believe that creation of an online repository and adoption and/or adaptation of this project across multiple researchers and institutions could speed the process of gene function prediction.

Correspondence to: Erin D. Strome, stromeel@nku.edu.

Conflict of Interest Notification: These authors have no perceived conflicts of interest to report.

Keywords

bioinformatics; orphan genes; genetics; course-based research

Introduction

Sequencing of complete genomes has been ongoing for more than 30 years; with recent exponential growth in the number of these studies as new sequencing technologies emerge and costs decrease. Once the sequence has been assembled, researchers use software to predict ORFs and bioinformatics tools to determine homology and predict potential functionality. As the first eukaryotic genome to be sequenced, *S. cerevisiae* has been investigated extensively and yet there are still up to 1500 genes (23% of the genome) of unknown functions, split into categories of uncharacterized or dubious ORFs [<http://www.yeastgenome.org/genomesnapshot>]. We hypothesized we could begin to tackle this ‘orphan gene’ situation by utilizing undergraduate biology students to manually investigate ORFs. Using a systematic approach we would increase the exploration of these genes while exposing these students to novel and authentic research experiences.

Genomics and bioinformatics are advancing fields that have become an integral part of a broad spectrum of biological research. As a result, undergraduate biology majors should be able to use bioinformatics tools and have a general knowledge of genomic concepts to be prepared for the modern biology workforce. The application of bioinformatics tools to study genomics can also serve as an opportunity to integrate key principles of genetics (i.e. gene and protein sequence as related to structure and function) early in the curriculum (one of five core competencies cited in Vision and Change, [AAAS 2010]). Bioinformatics and genomics approaches also generate accessible and inexpensive research opportunities for undergraduates since many of the necessary tools are freely available online [Boyle 2010, Maloney *et al* 2010].

Science education research has consistently shown that learning current research methods and skills positively impacts students [NRC 2003a, NRC 2003b] and that inquiry-based laboratory exercises result in increased learning when compared to traditional, expository activities [Brownwell *et al* 2015, Corwin *et al* 2015, Lord *et al* 2006, Rissing *et al* 2009]. Course-based research experiences, which integrate a research project into a course, are one way to expose students to the process of science while increasing their knowledge about biological concepts [AAAS 2010, Wei *et al* 2010]. However, there can be significant challenges in involving a large number of students in authentic research, including costs and time requirements for guiding students through the process. Having multiple sections of a laboratory course, in order to accommodate the large number of students, can require multiple instructors to teach these sections and can complicate implementation. Nevertheless, involving multiple instructors in a novel instructional method presents an opportunity to propagate evidence-based teaching practices and expose more students to course-based research.

Due to increasingly rapid genomic sequencing combined with the accessibility and inexpensive nature of bioinformatics analysis, potential questions for use in course-based

research are abundant. A number of projects in genomics using a bioinformatics approach have been described in the literature [Almeida *et al* 2004, Ault *et al* 2011, Beagley 2013, Boyle 2004, Chen *et al* 2005, Ditty *et al* 2010, Ditty *et al* 2013, Goodner *et al* 2003, Hatfull *et al* 2006, Hingamp *et al* 2008, Reed *et al* 2013, Shaffer *et al* 2010 and Shapiro *et al* 2013], across a variety of instructional settings and research situations. Several of the initiatives, such as Phage Hunters Advancing Genomics and Evolutionary Science (PHAGES) project, supported by the Howard Hughes Medical Institute, have incorporated bioinformatics approaches into curricula through authentic research in lower-level, general courses or even with students in high school [e.g. Hatfull *et al* 2006]. Others have utilized the Joint Genome Institute's (JGI) Integrated Microbial Genome Annotation Collaboration Toolkit (IMG-ACT) to engage students at various levels including bioinformatics and cell biology courses at different types of institutions [Beagley 2013, Ditty *et al* 2013, Reed *et al* 2013, and Shapiro *et al* 2013]. Additional programs have focused on eukaryotic genetics, such as the Genomics Education Partnership organized by the Department of Biology and Genome Center of Washington University in St. Louis [Shaffer *et al* 2010] and the Dynamic Genes Project through the Dolan DNA Learning Center (<http://www.dnalc.org/websites/dynamic.html>). These efforts have demonstrated the positive effects of integrating research into the undergraduate curriculum and the flexibility of doing so through the use of online tools.

This study sought to build on recent work by incorporating authentic research using bioinformatics tools into a core genetics course with limited resources. Our primary questions were regarding the feasibility of implementing such a project in a large, multi-section genetics course with multiple instructors and determining the type, quality, originality and consistency of data that could be generated by undergraduate students concerning uncharacterized ORFs. Also, we hypothesized that students would gain confidence in their abilities as future scientists, which could be important in stemming the tide away from STEM majors [Ketelhut 2007, Pajares *et al* 1994, and Stout *et al* 2011].

We describe the project implementation, production of materials, the impact on student learning and perceptions, potential contributions to the *S. cerevisiae* research community, and insights into successfully executing a course-based genome annotation research project. Due to the nature of the course (lower-level, high enrollment, multi-section, with multiple instructors), and the flexibility of the materials developed, this type of project may be amenable to a variety of courses and institutions.

Materials and Methods

A total of 179 students participated in the gene annotation project over four semesters, fall 2012-spring 2014, in a core required genetics course for all biology and biochemistry majors. As the course is a pre-requisite for many of the upper-level courses required for a genetics/molecular/cellular BS biology track degree, on average one-third of the students in the course were sophomores on this degree path. The remaining two-thirds of the class were composed of juniors, seniors, and a small number of post-baccalaureate students. The four-credit hour genetics course consists of three 50-minute lecture meetings and three-hours of laboratory each week. Enrollment in the course was between 30-60 students with three lab

sections of around 20 students each in fall semesters and 15-20 students across two lab sections in spring semesters. Two instructors alternated semesters in which they taught the lecture portion of the course and three instructors taught the laboratory sections (all instructors were tenured or tenure-track faculty). Laboratory instructors were given materials prior to the start of the course, participated in a brief meeting on the project concept and principles, and met on a weekly basis throughout the semester to further discuss and manage the project. Production of student materials was undertaken as part of instructors normal teaching load and did not require additional funding. Student work was computational and supported by computer and software already available within the department, again requiring no additional funds.

The project originated from the desire to incorporate bioinformatics and genomics into the undergraduate curriculum, while engaging students in the scientific process. A collaboration was established between faculty teaching the course and faculty conducting genetics research using *S. cerevisiae*, providing the opportunity for students in the genetics course to investigate orphan genes, predict putative functions, learn about ongoing research projects and impact the scientific community. Sequencing, functional and comparative genomics, as well as gene and genome structure were discussed in the lecture portion of the course, prior to introducing students to the independent gene annotation project (Table 1). Modules were generated to provide a framework to answer questions of interest to the project through a series of bioinformatics algorithms including multiple sequence alignment, conserved domain identification, signal sequence prediction and cellular localization (Table 2). The modules are available through the SGDE educational Resources website page [<http://www.yeastgenome.org/educational-resources-on-the-sgd-community-wiki>], and include introductory information for each software application based on that provided by the hosting site, as well as step-by-step directions for implementation and interpretation of data. Each semester, instructors assembled a set of orphan genes designated in SGD as encoding 'putative proteins of unknown function.' Furthermore each set shared at least one common characteristic (e.g. deletion results in sensitivity to methyl methanesulfonate), to encourage student interaction and discussion of data. Each student in one lab section was assigned one ORF from this set, and each lab section was assigned the same total gene set, allowing the research to be run in duplicate or triplicate depending on the semester.

The multi-week project incorporated brief direction from instructors on how to use the tools and interpret data, but the majority of time was dedicated to students collecting and analyzing data. The investigation was incorporated alongside other wet-bench labs to fill time lapses such as incubation periods or gel electrophoresis times. These sessions spanned 12 of the 15 weeks of the semester. Students utilized class time to collect and analyze their data and instructors circulated between students to assist with questions. Data collection and analysis was frequently completed during class time on a module-by-module basis but students reported needing additional out of class time for project synthesis and creation of their presentations. This time was not reported to take longer than the stated expectation of 2-3 hours of outside of class time per week for this portion of the course. Students uploaded their data and analyses to a course management system, which enabled a check for plagiarism. Instructors graded these submissions and provided specific feedback throughout

the process. During the final class meetings the students reported their findings in a five-minute conference-like presentation with additional time allotted for questions. The *S. cerevisiae* research group also visited class sessions to consult with students on the interpretation of their data and attended the final presentations.

Institutional review board approval was obtained for the study (IRB 12-058). Student learning for the lab section of the course was assessed via quizzes, assignments, a midterm exam, the gene annotation data and written analysis documented in the modules, a final report and a presentation (worksheets and rubrics available in supplemental materials). The grade for the course was based on an equal distribution of scores from the midterm exam, quizzes/assignments, and the gene annotation project. The quizzes and assignments tested student understanding of molecular biology techniques such as gel electrophoresis, gene cloning, bacterial transformation, and polymerase chain reaction and one quiz specifically evaluated students' abilities to use concepts and skills related to the bioinformatics project. The midterm exam focused on molecular biology techniques discussed in the first half of the course (Table 1). For the gene project modules, points were assigned for inclusion of data, but greater weight was placed on data analysis and hypothesis development.

A post-project questionnaire (supplemental materials) on research experience and perceptions of learning was administered on the last day of class. A χ^2 test was utilized to analyze responses from students that had conducted research prior to the course to determine if there was an impact of previous experience on their perceptions of the project (Figures 1 and 3).

Results

The majority of students formed a feasible hypothesis for the functionality of the protein encoded by the uncharacterized ORF based on the data they collected. The amount of data collected, its inter-relatedness, and the level of detail of the prediction varied widely (Table 3). Students most frequently used data about protein domains (34%), genetic and physical interactions (87%), microarray expression data (68%), cellular localization predictors (70%), drug sensitivities and co-fitness interactions (47%) to generate a hypothesis. In every case they found clues that led to the potential for a testable hypothesis and starting point for functional investigation. For example, the gene *YPL034W* (described in SGD as 'Putative protein of unknown function; YPL034W is not essential gene'), was investigated and while no protein domains were identified within the Conserved Domain Database (CDD) or listed within SGD, when using PFAM directly, a weak hit (0.0093) was found for the Meiotically Up-regulated Gene (MUG113), contained within a family of fungal proteins that become up-regulated in meiosis. When analyzing gene ontology (GO) term enrichments for interactors three terms ontologies were found: reproductive sporulation (2e-02), sporulation (sensu Fungi)(2e-02) and sporulation (3e-02). A GeneMania genetic interaction map revealed a connection to *AFT2*, an iron regulated transcription factor and *YLR126C*, (at the time classified as a putative protein thought to be involved in iron homeostasis) now classified as a putative glutamine amidotransferase which may have involvement in iron and copper homeostasis. Literature searches revealed a manuscript reporting a connection of Aft1 (also involved in iron homeostasis and with a connection to Aft2) to *RIM101* involved in

sporulation pathways [Berthelet *et al* 2010]. Therefore the synthesis of data points led the student to hypothesize that *YPL034W* may have a role in meiosis connecting iron homeostasis and sporulation events.

Of the 179 students completing the genetics course, 16 failed to complete the bioinformatics assignment and their data were not included in our analysis. A comparison, using Bloom's Taxonomy, of the assessment tools indicated the quizzes, assignments, and midterm focused more heavily on lower-order cognitive skills, requiring students to demonstrate comprehension, apply knowledge, display practical skills and interpret data (Table 4). The rubrics developed for evaluating student completion of the gene project modules and the project presentation demonstrate a shift to higher-order cognitive skills, with a third of the annotation requirements assessing student application or analysis and another third focused on synthesizing data from multiple sources (Table 4).

One hundred and twenty-nine of one hundred and seventy-nine (72%) enrolled across these four semesters completed the post-project questionnaire. These students were asked about their perceived learning during the project and approximately 79% said they had learned more than from a typical lab project (Figure 1a). While 8% of the respondents felt they had learned less from the gene annotation project than from a typical lab project, the majority of students (85%) felt better prepared to take on a new problem without a known solution after doing the orphan gene project and 82% of the students confirmed that having an understanding of the capabilities of bioinformatics databases and programs would benefit them in the future (Figure 1b-c).

Thirty-nine percent of the respondents had participated in a research experience and 40.5% said they had conducted authentic research as part of a course (Figure 2). Forty-eight percent of students found it “exciting” and 93% found it “challenging” to investigate an authentic research question (Figure 3). In comparing responses of those students who had conducted research previously to those that had not using a χ^2 analysis, there were no statistically significant differences identified in their perceptions of the project difficulty, perceived learning, etc.

The majority of students found the project rewarding once they reached the end point and were able to use the data they had assembled to synthesize a hypothesis, in fact 33.3% of students indicated that ‘making connections’ was their favorite part of the project. Another 16% of students most enjoyed that the project was an ‘authentic’ research project and another 10% liked most that they had learned ‘a tremendous amount.’ One student stated “I really enjoyed investigating the gene's function because it helped me to become more familiar with many biological processes” while another student stated they enjoyed “finding clues on genes that have unknown functions, I like puzzles so putting all the details together was kind of exciting.”

When asked what their least favorite part of the project was, 25% of students indicated that ‘getting negative results’ was most troubling while 21% cited the large amount of time needed to collect and analyze data and another 21% cited discontent with the process of putting the pieces together. Notably only 13% of students found the lack of one correct

answer or interpretation of data as they completed the project to be disheartening. One student stated that “there were no examples of starting the thought processes to run through” in analyzing information and another thought that “nothing makes sense while you are doing it in the beginning for a while”.

Students also seemed challenged by the fact that multiple types of data were compared to support or contest the hypothesized function of the gene product. Several stated that they were frustrated that “different module information can lead to different paths.” Others appreciated the facet of “finding a correlation between results.”

Discussion

When first developing the orphan gene project, we were concerned about the feasibility of implementation. Multiple instructors teaching several laboratory sections can present challenges regarding consistency, but compounding that with the inquiry nature of the assignment was an additional strain. We found it essential to develop and use rubrics for each section of the orphan gene project in order to execute the project and meaningfully assess student learning across the different sections, thus this process could be adapted for similar projects. An analysis of the evaluation measures indicated the orphan gene project required students to demonstrate learning at higher-order cognitive levels relative to other evaluation measures used in the course (Table 4).

Students delved into running the bioinformatics analysis interfaces, learned about the algorithms underlying bioinformatics tools used to collect data, and learned how variations of algorithms could result in different outputs. They saw first hand how data output could affect the ability of the user to effectively interpret results and collect all relevant data points. The classes also came away with a better sense of what non-‘wet-lab’ science can be and the vast amount of information that has been accumulated and is open for public investigation. A greater appreciation for authentic research was seen in the students, from the professor perspectives, when they realized the time and effort it took to synthesize their results, determine what to do with conflicting data and try to narrow down the information to a prediction of functionality that could serve as a starting point for wet-lab investigation.

Students perceived their learning to overall be greater than a typical lab project (Figure 1). We also hypothesized that students would gain confidence in their abilities as future scientists and 85% of the respondents felt that it better prepared them to take on future research projects. A substantial portion of the students (79%) felt that they learned even more from this project than other lab projects (Figure 1). All instructors reported increased engagement relative to previous laboratory experiments and assignments. Future directions for the work include conducting more comprehensive evaluation of student learning, for instance in their abilities to analyze data, synthesize data from multiple sources, and connect broader biological concepts to the specific research question. To accurately assess the impact of this project on student learning gains would also require running control courses without the project to determine impacts in these areas of interest. We would also like to have students develop ways to test their hypotheses of their gene's functions. This would both bridge the wet-lab to computational gap for students and allow us to better determine

the accuracy of the students' predictions on gene function. Further we would like to determine the feasibility of working with strains from the yeast deletion collection (mutant for each gene assigned) to conduct these experiments, both analyzing cost, equipment needs and ability to conduct meaningful experiments and determine results in a large classroom setting. We believe that creation of an orphan gene project wiki page could allow collaboration between any institutions that wish to be involved while maintaining a separate source of curated student data. Inclusion of a mechanism to curate publications involving novel gene functions assigned to these ORFs would aid our ability to track benefits to the scientific community moving forward.

Challenges experienced were similar to those encountered in comparable projects: student attitudes toward learning, computer-based research not viewed as “real research”, and technical challenges with software (especially sites being down or slow loads due to traffic) (21). Not all students appreciated the experience, with approximately 8% feeling they had learned less than from a typical lab project (Figure 1), and 18% suggesting they did not see the value in knowing how to utilize bioinformatics tools (Figure 1). As many students specifically referred to the amount of data in their comments, a possible contributing factor in students' negativity about the project was the amount and diversity of data they were managing and interpreting to answer the research question. Since the experiments are conducted *in silico*, many sets of data could be collected in a relatively short amount of time. This contrasts with other assignments in the course and pre-requisite courses in which data analysis is more simplified and usually involves one, maybe two, types of data that are similar among the entire class. The students were likely not prepared for the volume of data and interpretation required. A potential future approach would be inclusion of discussion of the field of systems biology with students. In acknowledging data complexity and contradictions from the start and working with students from the perspective of a way to think of connections and systems to interpret findings we might bring in existing tools and open up viewpoints with the students.

While approximately half the respondents felt it was exciting to investigate an authentic question (Figure 3), the lack of established answers contributed to student anxiety about the assignment. Many students had the urge to check to see if they were “right,” and instructors were challenged with encouraging independence while also reassuring them of their analyses. Incorporating formal mechanisms for peer feedback throughout the gene annotation process in future iterations may help with students' confidence in their analyses. We were surprised to find that students who had participated in research did not report differences in their perceptions of the project (such as difficulty, perceived learning, etc.).

The project culminated in a brief presentation, approximately five-minutes, by each student in which they described their hypothesis and supporting data regarding the role of the gene product in the cell. The class session was conducted much like a research colloquium and resulted in considerable exchange between the presenters and their audience. A portion of the presentations were attended by the researchers investigating *S. cerevisiae* genetics, which contributed to the atmosphere of collaboration. The public presentation not only motivated students to have a solid understanding of their gene annotation, but based on comments, it also helped them process data into meaningful conclusions. Final submission

of a Yeast Gene Report occurred after the presentation, allowing students to reflect on the class discussions and submit modifications if necessary.

Finally, student analyses have the potential to contribute to the genomic research being conducted on the *S. cerevisiae* genome. However this will necessitate a compilation of all researched genes and a distribution mechanism to allow access to all the collected data for each individual 'putative protein'. As a result, a greater understanding of this organism could be provided from student-collected data.

Due to the nature of the course (lower-level, high enrollment, multi-section, with multiple instructors), and flexibility of the materials, this type of project may be amenable to a variety of courses and institutions. We recommend that others interested in using such a model revise the provided modules (SGD Educational Resources page) on a semester basis, as updates to applications and databases occur rather frequently. Also, to make the research more approachable for less experienced students, it would be useful to embed interpretation information in the module instructions and provide examples rather than relying on students to sift through the websites. In addition, we will include formal peer feedback to potentially help students overcome the lack of confidence in their analyses, hoping to provide all biology and biochemistry majors at our university with a productive research experience early in their undergraduate careers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank the students of BIO349: Genetics at Northern Kentucky University for their willingness to share their thoughts about the project. We would also like to thank Jeffrey Steller for his help in set-up of the project and troubleshooting along the way. Further we thank the Educational Resources group at SGD for publishing our modules and making them available as a starting point to the community at large. Research reported in this publication was partially supported by an Institutional Development Award (IDeA) from the National Institute of General Medical Sciences of the National Institutes of Health under grant number 5P20GM103436-13.

References

- AAAS. [accessed 02 Jun. 2012] Vision and change in undergraduate biology education: a call to action. A report of the American Association for the Advancement of Science. 2010. <http://visionandchange.org/>
- Almeida CA, Tardiff DF, De Luca JP. An introductory bioinformatics exercise to reinforce gene structure and expression and analyze the relationship between gene and protein sequences. *Biochem and Mol Biol Ed.* 2004; 32:239–45.
- Ault J, Renfro B, White A. Using a molecular-genetic approach to investigate bacterial physiology in a continuous, research-based, semester-long laboratory for undergraduates. *J Microbiol and Biol Ed.* 2011; 12 <http://jmbe.asm.org/index.php/jmbe/article/view/326>.
- Beagley CT. Genome annotation in a community college cell biology lab. *Biochem and Mol Biol Ed.* 2013; 41(1):44–49.
- Berthelet S, Usher J, Shulist K, Hamza A, Maltez N, Johnston A, Fong Y, Harris LJ, Baetz K. Functional genomics analysis of the *Saccharomyces cerevisiae* Iron Responsive Transcription Factor Aft1 Reveals Iron-Independent Functions. *Genetics.* 2010; 185(3):1111–1128. [PubMed: 20439772]

- Boyle JA. Bioinformatics in undergraduate education: Practical examples. *Biochem and Mol Biol Ed.* 2004; 32(4):236–8.
- Boyle M. “Shovel-ready” sequences as a stimulus for the next generation of life scientists. *J Microbiol and Biol Ed.* 2010; 11 <http://jmbe.asm.org/index.php/jmbe/article/view/133>.
- Brownwell SE, Hekmat-Scafe DS, Singla V, Seawell PC, Imam JFC, Eddy SL, Stearns T, Cyert MS. A High-Enrollment Course-Based Undergraduate Research Experience Improves Student Conceptions of Scientific Thinking and Ability to Interpret Data. *CBE-Life Sciences Education.* 2015; 14:1–14.
- Chen J, Call GB, Beyer E, Bui C, Cespedes A, Chan A, Chan J, Chan S, Chhabra A, Dang P, et al. Discovery-based science education: Functional genomic dissection in *drosophila* by undergraduate researchers. *PLoS Biol.* 2005; 3:e59. [PubMed: 15719063]
- Corwin LA, Graham MJ, Dolan EL. Modeling Course-Based Undergraduate Research Experiences: An Agenda for Future Research and Evaluation. *CBE-Life Sciences Education.* 2015; 14:1–13.
- Ditty JL, Kvaal CA, Goodner B, Freyermuth SK, Bailey C, Britton RA, Gordon SG, Heinhorst S, Reed K, Xu Z, et al. Incorporating genomics and bioinformatics across the life sciences curriculum. *PLoS Biol.* 2010; 8:e1000448. [PubMed: 20711478]
- Ditty JL, Williams KM, Keller MM, Chen GY, Liu X, Parales RE. Integrating grant-funded research into the undergraduate biology curriculum using IMG-ACT. *Biochem and Mol Biol Ed.* 2013; 41(1):16–23.
- Goodner BW, Wheeler CA, Hall PJ, Slater SC. Massively parallel undergraduates for bacterial genome finishing. *ASM News.* 2003; 12:584–585.
- Hatfull GF, Pedulla ML, Jacobs-Sera D, Cichon PM, Foley A, et al. Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *PLoS Genet.* 2006; 2(6):e92.10.1371/journal.pgen.0020092 [PubMed: 16789831]
- Hingamp P, Brochier C, Talla E, Gautheret D, Thieffry D, Herrmann C. Metagenome annotation using a distributed grid of undergraduate students. *PLoS Biol.* 2008; 6:e296. [PubMed: 19067492]
- Ketelhut DJ. The Impact of Student Self-efficacy on Scientific Inquiry Skills: An Exploratory Investigation in River City, a Multi-user Virtual Environment. *Journal of Science Education and Technology.* 2007; 16(1):99–111.
- Kloser MJ, Brownell SE, Chiariello NR, Fukami T. Integrating teaching and research in undergraduate biology laboratory education. *PLoS Biol.* 2011; 9:e1001174. [PubMed: 22110400]
- Kushner D. DNA microarrays in the undergraduate microbiology lab: Experimentation and handling large datasets in as few as six weeks. *J Microbiol and Biol Ed.* 2009; 8(1)
- Lord T, Orkwiszewski T. Moving from didactic to inquiry-based instruction in a science laboratory. *Amer Biol Teach.* 2006; 68:342–5.
- Maloney M, Parker J, LeBlanc M, Woodard CT, Glackin M, Hanrahan M. Bioinformatics and the undergraduate curriculum. *CBE Life Sci Educ.* 2010; 9:172–4. [PubMed: 20810947]
- National Research Council. Bio2010: transforming undergraduate education for future research biologists. National Academies Press; Washington, DC: 2003.
- National Research Council. Improving undergraduate instruction in science, technology, engineering, and mathematics: Report of a workshop. Washington, DC: National Academy Press; 2003.
- Pajares F, Miller DM. Role of self-efficacy and self-concept beliefs in mathematical problem solving: A path analysis. *Journal of Educational Psychology.* 1994; 86(2):193–203.
- Reed KE, Richardson JM. Using microbial genome annotation as a foundation for collaborative student research. *Biochem and Mol Biol Ed.* 2013; 41(1):34–43.
- Rissing SW, Cogan JG. Can an inquiry approach improve college student learning in a teaching laboratory? *CBE Life Sci Educ.* 2009; 8:55–61. [PubMed: 19255136]
- Shaffer CD, Alvarez C, Bailey C, Barnard D, Bhalla S, Chandrasekaran C, Chandrasekaran V, Chung H, Dorer DR, Du C, et al. The genomics education partnership: Successful integration of research into laboratory classes at a diverse group of undergraduate institutions. *CBE Life Sci Educ.* 2010; 9:55–69. [PubMed: 20194808]
- Shapiro C, Ayon C, Moberg-Parker J, Levis-Fitzgerald M, Sanders ER. Strategies for using peer-assisted learning effectively in an undergraduate bioinformatics course. *Biochem and Mol Biol Ed.* 2013; 41(1):24–33.

- Stout JG, Dasgupta N, Hunsinger M, McManus MA. STEMing the tide: Using ingroup experts to inoculate women's self-concept in science, technology, engineering, and mathematics (STEM). *Journal of Personality and Social Psychology*. 2011; 100(2):255–270. [PubMed: 21142376]
- Wei CA, Woodin T. Undergraduate research experiences in biology: Alternatives to the apprenticeship model. *CBE-Life Sci Educ*. 2011; 10:123–131. [PubMed: 21633057]
- Ketelhut DJ. The Impact of Student Self-efficacy on Scientific Inquiry Skills: An Exploratory Investigation in River City, a Multi-user Virtual Environment. *Journal of Science Education and Technology*. 2007; 16(1):99–111.

a. In doing the gene annotation project, I learned _____ a typical lab project.

b. I feel better prepared to take on investigating a problem without a known solution or outcome after doing this project.

c. I think having an understanding of the capabilities of bioinformatics databases and programs will benefit me in the future.

Figure 1.

a-c. Opinions of respondents regarding their perceived learning after participating in the gene annotation research, preparation for future research, and value in learning bioinformatic tools.

Figure 2. Percentage of respondents (N=129) that had participated in a research experience and percentage that had conducted an authentic research project as part of a course.

Figure 3. Percentage of respondents (N=129) to questions pertaining to their perception of the orphan gene project. Note that students could choose more than one response to this question.

Table 1
Outline of topics and exercises conducted in genetics course

Week	Lecture Topics	Lab Exercises
1	Introduction to Genetics	Micropipetting and culture techniques
2	DNA: The chemical nature of the gene	DNA restriction and electrophoresis
3	Chromosome structure and transposable elements	Transformation and introduction to DNA sequence analysis
4	DNA replication and recombination	Construction of a recombinant plasmid
5	Transcription	Construction of a recombinant plasmid (continued)
6	RNA molecules and RNA processing	Construction of a recombinant plasmid (continued)
7	The genetic code and translation	Polymerase chain reaction
8	Control of gene expression in prokaryotes and eukaryotes	Electrophoretic Analysis, Constructing a Knockout cassette
9	Control of gene expression in prokaryotes and eukaryotes	Spring Break
10	Control of gene expression in prokaryotes and eukaryotes	Midterm Exam
11	Sequencing, Biotechnology and Genomics	Electrophoretic Analysis
12	Chromosomes and cellular reproduction	Knockout Strain Creation
13	Basic principles of heredity	Knockout Strain Creation and Phenotypic Analysis
14	Sex determination and sex-linked characteristics	Knockout Strain Creation and Phenotypic Analysis
15	Pedigree analysis, applications and genetic testing	Data compilation and Orphan Gene Presentations

Table 2
Gene annotation modules and student investigations

Module	General exercise	Specific
1	Introduction to <i>Saccharomyces cerevisiae</i>	Basic Information Module <ul style="list-style-type: none"> DNA Sequence Protein Sequence Sequence-based Similarity Data Module <ul style="list-style-type: none"> BLAST CDD (Conserved Domain Database)
2	Genetic and Physical Interactions and Expression Data	Genetic and Physical Interactions <ul style="list-style-type: none"> Gene Mania Expression Data Analysis <ul style="list-style-type: none"> SPELL (Serial Pattern of Expression Levels Locator)
3	Structure-Based Evidence and Multiple Sequence Alignment	Structure-based Evidence Module <ul style="list-style-type: none"> TIGRFam (protein family database search) Pfam (search of alternative protein family database) PDB (Protein Data Bank structure search) SGD Protein Tab SUPERFAMILY SMART (Simple Modular Architecture Research Tool) GENE3D PANTHER (Protein ANalysis THrough Evolutionary Relationships) Multiple Sequence Analysis <ul style="list-style-type: none"> T-COFFEE (Tree-based Consistency Objective Function For alignment Evaluation) Web Logo
4	Cellular Localization Data	Cellular Localization Data <ul style="list-style-type: none"> TMHMM (Trans Membrane helices Hidden Markov Models) SignalP Philius TargetP NucPred YPL (Yeast Protein Localization database)
5	Gene Deletion Phenotypes	Gene Deletion Phenotypes <ul style="list-style-type: none"> PROPHECY (PROfiling of PHENotypic Characteristics of Yeast) Yeast Fitness Database SGD Phenotype Tab

Table 3
Sample of student collected data and prediction of function

Gene Name	Source	Data Collected
YBR220C	CDD	<ul style="list-style-type: none"> Hit#1: PTR2 dipeptide/tripeptidepermease (COG3104) Hit#2: AmpG muropeptide transporter (PRK11010)
	GeneMania	<ul style="list-style-type: none"> Shared Protein Domain: <i>FEN2</i> (plasma membrane symporter) Genetic Interactor: <i>AMS4</i> (NPC component)
	SPELL & Gene Ontology	<ul style="list-style-type: none"> Co-expression: <i>VRG4</i> (Golgi GDP-mannose transporter) YBR220C is consistently co-expressed with genes that play roles in glycosylation and/or have relationships to the ER and ribosomes. GO Biological Process: protein amino acid acetylation
	TIGRFAM & PFAM	<ul style="list-style-type: none"> TIGRFAM#1: AmpG-related permease (TIGR00901) TIGRFAM#2: sider_RhtX_FptX siderophore transporter (TIGR02718)
	Philius, SignalP, & NucPred	<ul style="list-style-type: none"> PFAM#1: Acetyl-coenzyme A transporter 1 (PF13000) Philius: transmembrane protein SignalP: No signal peptide NucPred: 0.22
	PROPHECY & Yeast Fitness Database	<ul style="list-style-type: none"> PROPHECY: Null has growth resistance to diamide YFD: null has sensitivity to lovastatin (inhibitor of HMG CoA reductase; rate limiting step for cholesterol biosynthesis) Co-fitness interactions: SDS24 (involved in cell separation) and HMX1 (ER localized hemeoxygenase)
	Student Prediction of Function	Transmembrane protein with a transporter/permease related domain. This protein could serve as a transporter of acetyl co-enzyme A into the cell or ER/Golgi to be used in lipid biosynthesis.
YCL002C	BLAST	<ul style="list-style-type: none"> Hit#1: <i>C. briggsae</i> Cystinosin (transmembrane protein involved in excretion of cysteine out of lysosomes) Hit#2: <i>S. kudriavzevii</i> IFO 1802 (hydrogen ion transmembrane transporter)
	GeneMania	<ul style="list-style-type: none"> Physical Interactions: <i>OST6</i>, <i>ARV1</i>, <i>FRT1</i> (all ER localized)
	TIGRFAM	<ul style="list-style-type: none"> TIGRFAM#1: PQ loop repeat proteins
	Philius, SignalP, TargetP & NucPred	<ul style="list-style-type: none"> Philius: 7-pass transmembrane protein SignalP: No signal peptide NucPred: 0.05 TargetP: secretory pathway score 0.919 RC class 1
	PROPHECY & Yeast Fitness Database	<ul style="list-style-type: none"> PROPHECY: Null has growth sensitivity in DTT YFD: Drug sensitivity: null has growth inhibition in tryptanthrin (inhibitor of transmembrane proteins.) Co-fitness interactors: (osmotic and isotonic regulation)

Gene Name	Source	Data Collected
YLR152C	Student Prediction of Function	Seven-pass transmembrane protein with transporter activity localized to the ER. Cells without YCL002C show increased sensitivity to oxidative and reductive stress.
	CDD	<ul style="list-style-type: none"> Hit#1: Predicted permeases COG# COG0679
	GeneMania	<ul style="list-style-type: none"> Co-expression: MAL31, GAL3, GCY1, YAK1 (sugar transport, catabolism, or sensing)
	SPELL & Gene Ontology	<ul style="list-style-type: none"> YLR152C shares expression patterns with genes that encode proteins that function in membrane transportation.
	TIGRFAM & PFAM	<ul style="list-style-type: none"> PFAM#1: Membrane transport protein (PF03547)
	Philius, SignalP, & NucPred	<ul style="list-style-type: none"> Philius: 9-pass transmembrane protein SignalP: No signal peptide NucPred: 0.42
YOR365C	PROPHECY & Yeast Fitness Database	<ul style="list-style-type: none"> Co-fitness interactions: AQY1 (transport across membranes), YDL180W (vacuole localized), YAR028W (integral membrane protein), YOL092W (vacuolar membrane transport)
	Student Prediction of Function	Vacuole membrane protein involved in lipid homeostasis through the transport of lipids and sugar-modified lipids.
	BLAST & CDD	<ul style="list-style-type: none"> BLAST: <i>S. pombe</i>: TRP-like ion channel pkd2; CDD: Hit#1 PG-PI_TP COG# cd00917
	GeneMania	<ul style="list-style-type: none"> Physical Interaction: <i>BRE1</i> (E3 ubiquitin ligase) Genetic Interaction: <i>SMA1</i>, <i>HMS1</i>, <i>ABM1</i>, <i>ATH1</i> (assembly of prospore, hyperfilamentous growth, microtubule organization, digestion of trehalose) Co-expression: <i>ABM1</i>, <i>IME4</i>, <i>SPO75</i>, <i>SPO77</i>, <i>HOP2</i> (meiosis-specific proteins, spore wall formation, entry into meiosis, microtubule organization),
	SPELL & Gene Ontology	<ul style="list-style-type: none"> YOR365C share expression patterns with genes involved in spore wall formation and meiosis.
	TIGRFAM & PFAM	<ul style="list-style-type: none"> PFAM#1: ML-Like Domain (PF14558) PFAM#2: Transient receptor potential ion channel (PF06011)
	Philius, SignalP, TargetP & NucPred	<ul style="list-style-type: none"> Philius: 9-pass transmembrane protein SignalP: Signal peptide and cleavage site predicted with score of 79% TargetP: secretory peptide score 0.844 (out of 1) NucPred: score of 0.38 (out of 1)
	PROPHECY & Yeast Fitness Database	<ul style="list-style-type: none"> YFD: Drugs: null has increased sensitivity to dyclonine (decreases membrane permeability)
	Student Prediction of Function	Protein involved in spore wall synthesis

Table 4
Bloom Taxonomy levels of Assessment Tools

Assessment	Knowledge/Comprehension	Application/Analysis	Synthesis/Evaluation
Quizzes and Assignments	48%	37%	15%
Midterm Exam	76%	24%	0%
Orphan Gene Project	34%	34%	32%