A CURE-based approach to teaching genomics using mitochondrial genomes

Cloe S. Pogoda1*, Kyle G. Keepers1, Jacob T. Stanley2, and Nolan C. Kane1

1Ecology and Evolutionary Biology, University of Colorado, Boulder
2Biofrontiers Institute, University of Colorado, Boulder

Abstract

There is an abundance (currently over $10^{16}$ DNA bases) of publicly available genetic sequence data and a dearth of trained genomicists to process and interpret it, necessitating more trained bioinformaticians with biological expertise. For example, thousands of data sets are deposited on NCBI's Sequence Read Archive with plans to use only part of the data generated, though much of this data could be used to address other important biological questions. Course-Based Undergraduate Research Experiences (CUREs) are growing in popularity as a way to engage undergraduates in a project-based learning experience to analyze data that could not otherwise be processed. Through CUREs, students can receive training in the most relevant and up-to-date skill sets used within the field. We present a lesson plan for a CURE centered around teaching genome annotation. This project is suitable as a four week module in an undergraduate/graduate cross-listed course and focuses on annotating streamlined organellar genomes. This module is similar to other programs, such as the Genomics Education Partnership. However, students are additionally provided with the opportunity to publish their annotated genomes to NCBI's GenBank. In addition, many students who have taken this course have gone on to pursue internships and careers using the bioinformatics skills gained.

Learning Goal(s)

Students will be able to:

• Answer the question: “What is a genome?”
• Diagram the Central Dogma of biology that information is stored in the form of DNA and gets transcribed into functional RNAs (mRNA, tRNA, rRNA) and translated in protein
• Differentiate between the organellar and nuclear genomes within a eukaryotic organism (nuclear, mitochondrial, chloroplast)
• Understand what DNA components are part of a whole genome shotgun assembly (e.g., nuclear, mitochondrial, chloroplast genomes)
• Perform background research on the biological study system, including how to concisely introduce the system to a broad audience and summarize the current state of the literature. This includes knowing how to find and read relevant scientific literature and how to properly cite references
• Assemble the relevant background research and findings into a professional scientific presentation. Students can either write a short description of the work they performed, which could be submitted for publication to a scientific journal such as Mitochondrial DNA Part B, or as a poster at a conference.

Learning Objective(s)

• Install the appropriate programs such as Putty and WinSCP.
• Navigate NCBI's website including their different BLAST programs (e.g., blastn, tblastx, blastp and blastx)
• Use command-line BLAST to identify mitochondrial contigs within a whole genome assembly
• Filter the desired sequence (using grep) and move the assembled mitochondrial genome onto your own computer (using FTP or SCP)
• Error-correct contigs (bwa mem, samtools tview), connect and circularize organellar contigs (extending from filtered reads)
• Transform assembled sequences into annotated genomes
• Orient to canonical start locations in the mitochondrial genome (cox1)
• Identify the boundaries of all coding components of the mitochondrial genome using BLAST, including: Protein coding genes (BLASTx and tBLASTx), tRNAs (proprietary programs such as tRNAscan), rRNAs (BLASTn, Chlorobox), ORFs (NCBI's ORFFinder)
• Deposit annotation onto genome repository (NCBI)
• Update CV/resume to reflect bioinformatics skills learned in this lesson
INTRODUCTION

Most published sequencing data is made available on the Sequence Read Archive (SRA) and GenBank, both of which are maintained by the United States National Institutes of Health National Center for Biotechnology Information (NIH/NCBI), as well as the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and similar repositories funded by other countries. As of January 1, 2019, the SRA contained more than 23 petabytes of raw sequence data comprised of microbial, plant and human genome sequences (SRA Overview: https://www.ncbi.nlm.nih.gov/sra/docs/), the equivalent of roughly one million human genomes. Data production is currently outpacing data storage capabilities, and with the advent of Illumina technology (1), it is expected that sequencing capabilities will continue to increase rapidly (2,3,4). While raw data is easily produced, we need more trained genomicists to process, curate, and annotate these data (5). Here we describe a lesson plan that can be taught as a final capstone project in a bioinformatics course. This lesson plan is designed to train biologists the skills needed to take these raw data and annotate full organelar genomes. These sequencing data can be obtained from research university projects or from freely available public repositories such as GenBank or the SRA. By promoting individual project ownership, this lesson aims to inspire and attract undergraduate and graduate students to future careers or research involving genomics (6).

Course-Based Undergraduate Research Experiences (CUREs) encourage class engagement, provide a sense of ownership over a project, increase material retention, and provide a long-term persistence of interest in science (6,7,8,9). In addition, CUREs improve student ability to interpret and analyze data (10) and provide an alternative to the historical “apprenticeship” model (11), which is only feasible on a small scale and in appropriately equipped research institutions (12). Many CUREs have been successfully implemented as undergraduate laboratories and courses (e.g. 7,10,13).

A prime example of a successful network CURE is the Genomics Education Partnership (GEP), which is a consortium started in 2006 to expose undergraduate students to hands-on scientific research (14). The collective focuses on annotating short regions of the Drosophila genome mined from publicly available genomic data sets (14). As of 2016, there were more than 100 faculty members involved in the consortium. These faculty members are free to design their own curricula and most integrate the GEP project into existing genetic and molecular biology courses (14). The benefit of these types of programs is that they are flexible and can be implemented in many different teaching environments making them more inclusive for students that plan on pursuing other careers, such as health care (15).

This lesson plan teaches complete organelar genome annotation in detail, but also promotes student authorship in the form of publicly available resources that the students create and submit by the end of the module. In keeping with the Vision and Change from the AAAS/NSF (16,17) we have organized our learning goals for this course around the mantra that “the biology we teach should reflect the biology that we practice” (17). Therefore, the learning experience focuses on process rather than rote memorization. The students learn to use multiple genome analysis and annotation programs, including Sequin (18), and web BLAST (19), to identify and locate the features encoded within genome the genome they are given to annotate. In addition, the students learn fundamental answers to questions about biology, such as “What is a genome?” Ultimately, the course gives the students the experience of doing their own scientific investigations and presenting their own novel findings in publications and public data archives. Accordingly, the goal of this lesson plan is for each individual student to submit a fully annotated mitochondrial genome to NCBI (GenBank), which will become a publicly available genomic resource.

If all lesson material is completed, each student leaves with a first authored genomic resource publication that can be added to their CVs. This class is designed to teach all the necessary skills for fully annotating organelar genomes of various species (even non-model organisms) and is broadly applicable to many different disciplines. There have been calls for professionals (medical personnel, lawyers, business people, etc.) to be better prepared and informed about genomic data specifically, and how it is produced (15,20,21), as well as more general calls for increased abilities to handle big datasets of any type (22). Therefore, even if these students do not directly pursue a career in genomics, they will have a better understanding of how these enormous data sets are produced and analyzed, and how the scientific method is used to annotate a genome.

Intended Audience

This lesson plan is meant to be incorporated into an undergraduate/graduate level bioinformatics course as either a major component or final project. It has been successfully taught to up to 30 students (20 undergraduate and 10 graduate students) at a time. However, we recommend first time teachers limit the enrollment to 20 students. This lesson requires that the students have a personal laptop to access the class projects, however access to a computer lab containing either Mac or Linux operating systems and in which programs are able to be installed is a suitable alternative. Each student will complete their own assembly and annotation. It is appropriate for a wide range of university majors including; biology, chemistry, applied mathematics, computer science, and pre-medical students. We expect students taking this lesson to have some knowledge of the Central Dogma of biology as well as experience navigating a command-line interface and using web-based BLAST. We have provided links to tutorials for each of these tools. This capstone project has been successfully taught to both undergraduate and graduate students in an evolutionary biology department. These students have minimal computer skills; however, some have been exposed to the computer language R. In addition, we have had undergraduate students in molecular biology, pre-med, physics and computer science successfully complete this course project.

Required Learning Time

We have found that four weeks is an appropriate duration for this module to be completed by instructors adopting this lesson for the first time (see lesson plan timeline: Table 1). Variations of this lesson plan have been tested over six semesters of EBIOS Genomics Course-4460/5460 taught at the University of Colorado, Boulder. In addition to class time
and recitation, students on average report spending 6.8 hours per week to complete this project.

**Prerequisite Student Knowledge**

This lesson plan focuses on teaching the basic skills necessary for organellar genome annotation and we suggest that students will need to have basic command-line skills for optimal performance in this project. Basic command-line and Linux skills are necessary to succeed and can be found in several modules (http://linuxcommand.org/fc3_learning_the_shell.php). Additional skills needed to succeed include familiarity with command-line BLAST (a beginner's guide can be found here: https://www.ncbi.nlm.nih.gov/books/NBK279680/) and web-based BLAST (https://www.youtube.com/watch?v=gKRD6t7-I42M), as well as a basic understanding of how DNA is sequenced (https://en.wikipedia.org/wiki/DNA_sequencing) and assembled into contigs or reads (https://en.wikipedia.org/wiki/Sequence_assembly). A comprehension of the Central Dogma will enhance the students' understanding of the genes they are annotating and how they are processed into functional cellular components, such as proteins or functional RNAs (https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology).

Students will need to be able to access a secure shell (SSH) client, such as PuTTY (https://www.putty.org/) for Windows machines. Mac and Linux machines should come already equipped with default SSH abilities. Students using Windows machines should also download a file transfer program (FTP), such as WinSCP (https://winscp.net/) or Filezilla (https://filezilla-project.org/) to facilitate easy movement of files to and from their computers to a centralized Linux server. Mac and Linux users can use command-line secure copy (SCP, a guide to using SCP can be found here: https://linuxize.com/post/how-to-use-scp-command-to-securely-transfer-files/) if desired. Students should also be able to use their web browser of choice to navigate to some of the websites that contain tools for genome annotation such as GenBank's BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi), OGDraw (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html) and Chlorobox GeSeq (https://chlorobox.mpimp-golm.mpg.de/geseq.html).

**Prerequisite Teacher Knowledge**

The instructor must have working knowledge of command-line navigation and file management as well as programs useful for parsing large files, such as awk (https://www.tutorialspoint.com/awk/), grep (https://open.sourceforgeu.com/2012/06/beginners-guide-grn-grep-basics/) and sed (https://www.tutorialspoint.com/sed/). If the instructor is rusty, they can be reviewed at: http://linuxcommand.org/index.php.

Familiarity with file formats such as .fasta, .fastq, .bam/.sam as well as bioinformatics programs such as bwa mem (user's guide: http://bio-bwa.sourceforge.net/; 23), samtools tview (user's guide: http://samtools.github.io/tview.shtml; 24, 25), bcftools call (user's guide: https://samtools.github.io/bcftools/bcftools.html; 26), BLASTn (19), Chlorobox (27,28) and NCBI's Sequin (18) is required to prepare the data for student annotation. The program of whole genome assembly. The choice is to take raw genomic reads and assemble them into as few, long contigs as possible. These contigs should ideally relate to the number of chromosomes an organism contains, but expectations in this endeavor are typically that there are imperfections built into the process of Whole Genome Shotgun (WGS) de novo assembly and that a first pass won't achieve a perfect representation of the number of chromosomes. However, in contrast organellar genomes are small enough and in high enough copy number to where a researcher can have reasonable expectations of completely assembling the entire genome of the organelle. Moreover, organellar genomes tend to have a common set of a relatively small number of features that those annotating may know to look for, as opposed to nuclear genomes which contain a much larger, variable set of features. This makes organellar genomes an ideal locus for complete genome annotation (29,30,31,32) in the four-week lesson plan presented here.

The programs chosen here (samtools tview (24,25), bcftools call (26), BLASTn (19), Chlorobox GeSeq (28) and NCBI's Sequin (18)) provide an excellent toolbox for an instructor to guide students through their annotation. samtools tview (24,25), bcftools call (26) were chosen as they are simple to implement on the command-line and are appropriate for viewing the de novo assembled organellar contigs to check for assembly errors. BLASTn (19), Chlorobox GeSeq (28) and NCBI's Sequin (18) are all required for different stages of annotation. While NCBI's Sequin (18) is no longer actively supported, NCBI still accepts files prepared by Sequin. Moreover, it provides the student an ability to double-check and compare the results produced from using Chlorobox GeSeq (28), which is a newer tool and may not be able to correctly annotate genomes originating from species other than model organisms. The ability to hand-annotate each gene using NCBI's Sequin (18) is important even if the student does not ultimately use Sequin to submit their final annotation.

**SCIENTIFIC TEACHING THEMES**

**Active learning**

Students are provided with a tutorial that guides them through the steps of genome annotation, which is meant to complement the lecture portion of the bioinformatics/genetics course. Each student annotates their own genome, although class activities can be performed in groups. The tutorial contains section reviews and questions to evoke thought during the reading/learning process. Several class periods during the four week time table should be used as workshops, in which students can work on their annotations and potentially help each other during the troubleshooting process. In previous iterations of this lesson plan students have reported this being very beneficial. If students have problems during class relating to their annotations, instructors can use those as teaching moments to clarify and reiterate new concepts.

**Assessment**

We designed an informal pre- and post-survey consisting of 20 questions (supporting file S1: informal Pre-/Post- Survey). This survey was specifically developed in the context of a larger genomics course in which this lesson plan was taught. Therefore, many of the questions are specific to skills that are taught during the bioinformatics section of the larger course and the organism studied during the semester (i.e., lichens). This survey can be modified to include questions not only from the learning module, but also for the course in which it is being taught. Specifically, questions 13-15 and 18-20 address and assess gains made in this lesson.
In addition, the completion, submission and acceptance of the mitochondrial annotation is the main way to assess student engagement and performance during the course. Before students submit their fully annotated organellar genomes to NCBI, the instructor should manually check each annotation. These should include:

1. Viewing each student’s assembly in samtools tview to ensure no assembly errors are present and the genome is correctly circularized (end of Chapter 1, 2-3 minutes/student).
2. Completing a checklist of expected features including common organellar protein coding genes, rRNAs and tRNAs (see page 19 of supporting file S2: Annotation handbook, end of Chapter 2, 2-3 minutes/student).
3. Checking that there are no remaining error-level or reject-level warnings in NCBI’s Sequin (end of Chapter 3, 5-6 minutes/student).

**Inclusive teaching**

The active learning that is required for each student to successfully complete their organellar genome annotation promotes investment from the students and can help students from different backgrounds and with different skill sets to work together and engage in this challenging activity. As there is no strict pre-requisite for this class, students with many different educational backgrounds (i.e., undergraduate, graduate, traditional, non-traditional, biological science majors, computer science majors) work together to solve any in-class problems, which many students have verbally reported as being useful and having increased their learning and enjoyment of the lesson plan. This class also mixes graduate and undergraduate students, which can give the students an opportunity to mingle with people at different educational levels and may also give the graduate students an opportunity to take on a leadership role, which can promote a collegial learning environment.

**LESSON PLAN**

**Pre-Lesson Preparation**

Genome assemblies must be obtained and curated prior to the start of the lesson. Assemblies may be available from within the instructor’s department or university, but WGS assemblies that remain to be annotated are also available from NCBI. A centralized server is useful for storing all the class sequence data and installing the required programs (i.e., command-line BLAST, bwa mem and tview). Alternatively, a cloud-based service such as CyVerse or Google Drive could be used to store the assemblies and would be well within the typical course budget of a university class. Required programs can just as easily be run on a laptop computer equipped with a command-line as on a server.

We suggest that the instructor spend some time prior to the beginning of the four-week lesson to identify and download unannotated assemblies that they are interested in. If the instructor is using their own data, the curation of these genomes should only take 30-60 minutes. However, if the instructor is mining libraries/genomes from NCBI or other sequence archives, downloading and vetting enough unique mitochondrial genomes for each student to have their own to annotate can take several hours (4-6) for a class of approximately 20 people, depending on instructor familiarity with the process. We have successfully used NCBI to browse by taxonomy and search for archived sequences. An example taxon whose mitochondrial genome remains to be assembled and annotated, as of the time of writing, is Tuber melanosporum (black truffle) found on NCBI’s taxonomy database by searching the species name, filtering by type “genome”, and then clicking on the number that appears next to the taxon name after filtering the search. This takes you to the page for the draft assembly, which contains a download link to the .fasta formatted genome containing the contigs that we will be using in the main lesson. Instructors should also download raw .fastq sequences that will be used to error correct the assembled sequences. A search of Tuber melanosporum in NCBI’s SRA database yields over a dozen WGS libraries, any of which are suitable for the error-correction process so long as the final taxon ID in the annotated genome is attributed to the .fasta sequences rather than the de novo assembled .fasta sequence already downloaded. This is because error-correction of an assembly with a different individual’s sequences reveals a small number of differences (polymorphisms) that represent actual genetic differences between the reference genome and the genotype being assembled. By converting these polymorphisms in the original .fasta sequence you convert the genome from the original reference to that of the genome being assembled, which yields a unique genome that awaits annotation by the student. Instructors need to download the desired .fastq and .fasta files onto a centralized Linux server/student computer (using fastq-dump: [https://ncbi.github.io/sra-tools/fastq-dump.html](https://ncbi.github.io/sra-tools/fastq-dump.html) and will need to install bwa mem ([https://github.com/lh3/bwa](https://github.com/lh3/bwa)) and command-line BLAST ([https://www.ncbi.nlm.nih.gov/books/NBK279671](https://www.ncbi.nlm.nih.gov/books/NBK279671)). Once the desired sequences are downloaded, we suggest that the instructor check to make sure the mitochondrial genomes assembled successfully and are in no more than 2-3 separate contigs that will need to be connected by the students. If the mitochondrion assembled into more than 2-3 contigs, we suggest that the instructor choose a different assembly, as connecting more than a few is more labor intensive than we expect either instructor or students to take on. Once the desired species’ mitochondrial sequences are downloaded the instructor will need to put together an inventory of proteins, tRNAs and rRNAs that should be present in the mitochondrial genomes (a basic inventory is provided on page 19 of supporting file S2: Annotation Handbook). Using an already published resource, the instructor can click on the “send to” button in the upper right hand corner of the web page and directly download a .fasta sequence of all the coding features which will be used to make the BLAST database used during the four week lesson plan. An example suitable reference containing homologous features (rRNAs, tRNAs and protein coding genes) used for annotating Tuber melanosporum, can be found within the same subdivision Pezizomycotina. One possible such genome appropriate for this purpose is Usnea ceratina (accession number: NC_035940). The phylogenetic similarity of the reference genome to be used depends on the species type being annotated. Although closely related species make better references, due to the slowly evolving nature of the locus, distantly related species will often suffice (see our above example of subdivision level divergence). The script used to perform reference guided error-correction is provided in the supporting file S2: Annotation Handbook.
As with all scientific endeavors it is good practice to obtain permission to use data generated by other groups as well as appropriately thank and cite the individuals who originally sequenced it. We suggest that the instructor contact the data generators directly via email to alert them that you are going to be using their data for your class project. During the annotation step, it is important to cite the data generators for their work in the author's list.

The best preparation for this lesson is to read through the lesson timeline (Table 1) and the three chapters of the annotation handbook (see supporting file S2: Annotation Handbook).

In-Class Lecture Script for Week 1
The organellar genome annotation handbook chapter 1 (supporting file S2: Annotation Handbook) should be assigned to the students as required reading before the first class. The instructor can begin the first class period by introducing the differences between the nuclear and organellar genomes and letting the students know where to obtain their species sequence data (ideally stored on a central server). After these logistical issues are sorted the students can use command-line BLAST to identify the organellar contig(s) of interest.

During the second class period of the week the instructor should lead the class through the process of completing/polishing the genome from the assembled contig(s), which will usually involve ensuring that the organellar genome is circularized (in cases where applicable). Often the genome in question will have assembled in more than one contig so the instructor should guide the students through the process of connecting assembled contigs into one complete piece. The skills needed to perform this operation are covered in detail in chapter 1 of the annotation handbook (supporting file S2: Annotation Handbook).

In-class Lecture Script for Week 2
The second week will focus on finishing the preparation of the organellar genome sequence. During the first class period the instructor should lead the class through error correction and reorientation of their genomes to a canonical starting location (i.e., in the case of mitochondrial genomes, either coxI or the D-loop; in chloroplasts, the beginning of the long single-copy region [LSC]). This will conclude the portion of the lesson plan conducted using a Linux server.

During the second class period, students should be introduced to Chlorobox and the suite of genome annotation and visualization tools available (i.e., GeSeq, OGDraw, and GB2Sequin). Students should import their circularized, error corrected and reoriented .faa files into GeSeq. Once GeSeq has run, students should only spend a maximum of 40 minutes confirming the features that were identified by the program, deleting any notes that were inserted and deleting duplicate features. The students should double check the location of all 20+ putative transfer RNAs using specialized tRNA-locating programs (tRNAscan). Finally, the students will need to convert their GenBank formatted file in the appropriate format for Sequin to read using GB2Sequin also provide by Chlorobox. This Sequin formatted file should be saved as a plain text file for use during the next week.

In-class Lecture Script for Week 3
During the beginning of the first class period of the week students will need to download Sequin from NCBI. The instructor can then lead the class through reading an existing submission and opening the file saved during the previous class period. Form information such as the author, sequence information, and sample details must all be entered at this stage.

During the second class period students will add in any remaining protein-coding gene features into their organellar genome. GeSeq likely only found some of the protein-coding genes and probably did not estimate the correct boundaries, especially if the organism being annotated is lacking in published genomic resources. The instructor should demonstrate the proper method of using BLASTx, tBLASTx, and SmartBLAST to identify missing protein genes and update the boundaries of the genes identified by GeSeq. It is helpful to keep an inventory of the gene features students are expected to find, along with the expected length ranges of each feature. Filling out this inventory ensures the accuracy and completeness of the genome annotations.

In-class Lecture Script for Week 4
During the first class period students should visualize their annotation using the organellar genome drawer, OGDraw (https://chlorobox.mpimp-golm.mpg.de/OGDraw.html). This visualization helps students identify long gaps lacking annotated features. These gaps may contain open reading frames containing hypothetical proteins and parasitic elements that the students might have missed in the inventory or representing novel proteins that the students should add.

The second period of this lesson will have the students address any errors and warnings in their annotations, which are provided using Sequin’s “Validate” tool. Students should perform a final OGDraw to make sure that there are no remaining gaps or errors in their annotation. Once all possible errors have been eliminated and the instructors has performed quality control, the students should submit their annotation directly to GenBank by emailing them the Sequin formatted file.

Once the students receive confirmation that their submission was accepted, they receive an accession number and a link to the public record of their newly published genomic resource. This lesson concludes by instructing students on how to update their CV/resume to include this genetic resource publication, as well as the marketable skills they obtain in this lesson. Some students have been motivated to continue working on this project after the semester/course has ended. These students wrote up the analyses and prepared a manuscript for later submission to an appropriate journal (e.g., Mitochondrial DNA Part B Resources; 29, 30).

TEACHING DISCUSSION
Using the informal survey developed to measure learning gains we find that students largely met the learning goals of the lesson plan. We found large increases in personal interest in genomics, as well as their confidence in manipulating large data sets (students reported gains up to ~50%). Students...
responses to the free-response question “What is a genome?” were variable in their quality, but overall demonstrated a higher quality in the post-survey responses. An example, pre-survey response to the question “What is a genome?”: “I believe a genome is the mapping of one's genetic information into code to understand the DNA of that individual.” An example, post-survey response: “All of the DNA present in the nucleus, ribosomes and mitochondrion of an organism's cells.” By performing the annotation of different organellar genome components (rRNAs, tRNAs and protein coding genes) the students gain firsthand knowledge of the Central Dogma of biology. Many students report an increased understanding of how DNA is processed into RNA and proteins. This lesson focuses on organellar genomes and requires that the students differentiate between the mitochondrial/chloroplast DNA and the nuclear DNA present in WGS assemblies, this helps to cement where in the cell genetic information is stored and in what form. The students are challenged to discover more about their organism and present their findings as either a scientific poster or in the form of a peer reviewed manuscript, which gives more ownership over the project. This lesson plan places much emphasis on student involvement and engagement and they must take on an entire genome annotation on their own. We have seen that students generally report very positive feelings about their learning gains and project completion by the end of the module. This course has been taught several times and we have made improvements based upon student feedback. The most significant improvement has been the development of the annotation handbook (supporting file S2: Annotation Handbook). Students of this course may add sections to their resumes/CVs, such as technical computer skills, published genomic resources, and peer-reviewed journal publications that elevate their quality as applicants to trade, graduate school, or other professional programs. This course is adaptable to many different types of university environments, is useful for students in multiple different fields of study and has sufficient supplementary material to be easily implemented by interested faculty. We believe this integrative lesson plan that yields tangible scientific resources and valuable, marketable skills exemplifies the principles of a CURE-based learning approach.

Possible Modifications

We suggest adding additional course validation tools, such as The Project Ownership Survey (POS; 8) to assess the student’s feelings and experience about the course or the Classroom Undergraduate Research Experience (CURE) survey, which has been successfully implemented by the Genomics Education Partnership (14). These will help to expand confidence in this modules’ success in promoting project ownership and the resulting knowledge gains from participating in this lesson.

This lesson can also be adapted to teach faculty the process of genome annotation, which they can then go onto use in their own research projects. The three chapters of our annotation handbook (supporting file S2: Annotation Handbook) provide enough information that an individual with a basic understanding of genome sequencing and data manipulation can easily gain the skills necessary to complete a full annotation of an organellar genome.

SUPPORTING MATERIALS

• S1. CURE-based Genomics - Informal Pre- and Post-Survey
• S2. CURE-based Genomics - Annotation Handbook

ACKNOWLEDGMENTS

This work was funded in part by a teaching grant from the University of Colorado, Boulder. The lab work portion was supported by a grant from the National Science Foundation’s Dimensions of Biodiversity Program (award #1542639 [University of Colorado] and award #1432629 [New York Botanical Garden]).

REFERENCES

1. Van Dijk, EL, Auger, H, Jaszczyzyn, Y, & Thermes, C. 2014. Ten years of next-generation sequencing technology. Trends in genetics, 30(9), 428-426.
2. Leinonen, R, Sugawara, H, Shumway, M, & International Nucleotide Sequence Database Collaboration. 2010. The sequence read archive. Nucleic acids Res, 39(suppl_1):D19-D21.
3. Regalado, A. 2014. EmTech: Illumina says 228,000 human genomes will be sequenced this year. Technology Review, 24. Sciences Educ, 14(2):ar21.
4. Stephens, ZD, Lee, SY, Faghi, F, Campbell, RH, Zhai, C, Efron, MJ, Iyer, R, Schatz, MC, Sinha, S, and Robinson, GE. 2015. “Big data: astronomical or genomical?” PLoS biol, 13(7): e1002195.
5. Carnevale, AP, Smith, N, & Strohl, J. 2013. Recovery: Projections of jobs and education requirements through 2020. Washington, DC: Georgetown Public Policy Institute.
6. Hanauer, DI, Frederick, J, Fotinales, B, & Strobel, SA. 2012. Linguistic analysis of project ownership for undergraduate research experiences. CBE Life Sci Educ, 11(4):378-385.
7. Jordan, TC, Burnett, SH, Carson, S, Caruso, SM, Clase, K, Dejong, RJ, Dennehy Jj, Denver DR, Dunbar D, Elgin SC, Findley, AM. 2014. A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. MBio, 5(3):e01051-13.
8. Hanauer, DI, & Dolan, EL. 2014. The Project Ownership Survey: measuring differences in scientific inquiry experiences. CBE Life Sci Educ, 13(1):149-158.
9. Corwin, LA, Graham, MJ, & Dolan, EL. 2015. Modeling course-based undergraduate research experiences: an agenda for future research and evaluation. CBE Life Sci Educ, 14(1):es1.
10. Brownell, SE, Hekmat-Scafe, DS, Singla, V, Sewell, PC, Imam, JFC, Eddy, SL, Stearns, T, & Cyert, MS. 2015. A high-enrollment course-based undergraduate research experience improves student conceptions of scientific thinking and ability to interpret data. CBE Life Sci Educ, 14(2):ar21.
11. Wei, CA, & Woodin, T. 2013. Undergraduate research experiences in biology: alternatives to the apprenticeship model. CBE Life Sci Educ, 10(2):123-131.
12. Linn, MC, Palmer, E, Baranger, A, Gerard, E, & Stone, E. 2015. Undergraduate research experiences: impacts and opportunities. Science, 347(6222):1263757.
13. Shafer, CD, Alvarez, C, Bailey, C, Barnard, D, Bhatta, S, Chandrasekaran, C, Chung, HM, Dorer, DR, Du, C & Eckdahl, TT. 2010. The Genomics Education Partnership: successful integration of research into laboratory classes at a diverse group of undergraduate institutions. CBE Life Sci Educ, 9(1):55-69.
14. Elgin, SC, Hauser, C, Holzen, TM, Jones, C, Kleinschmit, A, Leatherman, J, & Partnership, TGE. 2017. The GEP: Crowd-Sourcing Big Data Analysis with Undergraduates. Trends Genet, 33(2):81-85.
15. Feero, WG, & Green, ED. 2011. Genomics education for health care professionals in the 21st century. Jama, 306(9):898-990.
16. Woodin, Smith, D, & Allen, D. 2009. Transforming undergraduate biology education for all students: an action plan for the twenty-first century. CBE Life Sci Educ, 8(4):271-273.
17. Ledbetter, MBL. 2012. Vision and Change in Undergraduate Biology Education: A Call to Action Presentation to Faculty for Undergraduate
A CURE-based approach to teaching genomics using mitochondrial genomes.

18. Benson, DA, Karsch-Mizrachi, I, Lipman, DJ, Ostell, J, Rapp, BA, & Wheeler, DL. 2000. GenBank. Nucleic acids Res, 28(1), 15-18.
19. Altschul, SF, Gish, W, Miller, W, Myers, EW, & Lipman, DJ. 1990. Basic local alignment search tool. J. Mol. Biol., 215(3), 403-410.
20. Korf, BR. 2011. Genetics and genomics education: the next generation. Genet Med, 13(3):201-202.
21. Chen, LS, & Goodson, P. 2013 Genomics education training needs of US health educators: a (qualitative) pilot study. Health promot pract, 14(1):44-52.
22. Siemens, G, & Long, P. 2011. Penetrating the fog: Analytics in learning and education. EDUCAUSE review, 46(5), 30.
23. Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:1303.3997.
24. Li, H, Handsaker, B, Wysoker, A, Fennell, T, Ruan, J, Homer, N, Marth, G, Abecasis, G, & Durbin, R. 2009. The sequence alignment/map format and SAMtools. Bioinform, 25(16):2078-2079.
25. Li, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinform, 27(21):2987-2993.
26. Narasimhan, V, Danecek, P, Scally, Xue, Tyler-Smith, C, & Durbin, R. 2016. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics, 32(11):1749-1751.
27. Greiner, S, Lehwark, P, & Bock, R. 2019. OrganellarGenomeDRAW (OGDRAW) version 1.3. 1: expanded toolkit for the graphical visualization of organellar genomes. bioRxiv, 545009.
28. Funk, ER, Adams, AN, Spotten, SM, Van Hove, RA, Whittington, KT, Keepers, KG, Pogoda, CS... & Kane, NC. 2018. The complete mitochondrial genomes of five lichenized fungi in the genus Usnea (Ascomycota: Parmeliaceae). Mitochondrial DNA B Resour., 3(1), 305-308.
29. Brigham, LM, Allende, LM, Shipley, B., Boyd, KC, Higgins, TJ, Kelly, N., ... & Tripp, EA. 2018. Genomic insights into the mitochondria of 11 eastern North American species of Cladonia. Mitochondrial DNA B Resour., 3(2), 508-512.
30. Pogoda, CS, Keepers, KG, Lendemer, JC, Kane, NC, & Tripp, EA. 2018. Reductions in Complexity of Mitochondrial Genomes in Lichen-Forming Fungi Shed Light on Genome Architecture of Obligate Symbioses. Mol Ecol., 27(5), 1155-1169.
31. Pogoda, CS, Keepers, KG, Nadiadi, AY, Bailey, DW, Lendemer, JC, Tripp, EA, & Kane, NC. 2019. Genome streamlining via complete loss of introns has occurred multiple times in lichenized fungal mitochondria. Ecol. Evol.
A CURE-based approach to teaching genomics using mitochondrial genomes.

Table 1. CURE-based Genomics - Teaching Timeline

| Description                                                  | Time* | Notes                                           |
|--------------------------------------------------------------|-------|-------------------------------------------------|
| **Week 1**                                                   |       |                                                 |
| Day 1: Identifying the organellar contig.                    | 75 minutes |                                           |
| Day 2:                                                       |       |                                                 |
| • Connecting contigs.                                        | 60 minutes | Chapter 1 of S2. Annotation Handbook             |
| • Circularizing the genome.                                  | 15 minutes |                                                 |
| **Week 2**                                                   |       |                                                 |
| Instructor should review chapter 2 prior to class.           | |                                                 |
| Day 1:                                                       |       |                                                 |
| • Error correction.                                          | 40 minutes | Chapter 1 of S2. Annotation Handbook             |
| • Reorienting the genome to cox1.                            | 35 minutes | Chapter 2 of S2. Annotation Handbook             |
| Day 2:                                                       |       |                                                 |
| • Chlorobox GeSeq.                                           | 40 minutes |                                                 |
| • And confirm tRNAs.                                        | 20 minutes |                                                 |
| • Convert from GenBank to Sequin format.                     | 15 minutes |                                                 |
| **Week 3**                                                   |       |                                                 |
| Instructor should review how to use Chlorobox GeSeq, and GB2Sequin and chapter 3 prior to class. | |                                                 |
| Day 1:                                                       |       |                                                 |
| • Download Sequin.                                           | 20 minutes | Chapter 3 of S2. Annotation Handbook             |
| • Open file in Sequin that was generated in Chlorobox GeSeq and converted using GB2Sequin. | 10 minutes |                                                 |
| • Add in meta data for submission.                           | 45 minutes |                                                 |
| Day 2:                                                       |       |                                                 |
| • Adding protein coding genes (BLASTx and iBLASTx).          | 75 minutes |                                                 |
| **Week 4**                                                   |       |                                                 |
| Instructor should review how to use OGDraw and ORFFinder prior to class. | |                                                 |
| Day 1:                                                       |       |                                                 |
| • Use OGDraw to find gaps in annotated features.             | 20 minutes | Chapter 3 of S2. Annotation Handbook             |
| • Add parasitic elements and hypothetical proteins.          | 55 minutes |                                                 |
| Day 2:                                                       |       |                                                 |
| • Error elimination.                                         | 40 minutes |                                                 |
| • Final OGDraw.                                              | 15 minutes |                                                 |
| • Submitting annotation to NCBI.                            | 20 minutes |                                                 |

*Each class is expected to take 75 minutes.*