# Infrastructure and Training to Bring Next-generation Sequence (NGS) Analysis Into Undergraduate Education

This three-year project will develop a sustainable infrastructure and training program to assist undergraduate faculty in integrating next-generation sequence (NGS) analysis into course-based and independent student research. Participating faculty will develop a total of 30 RNA sequence (RNA-Seq) datasets that bear on novel research problems in eukaryotic genomics. Following refinement of a biochemical and bioinformatics workflow by project staff, a Working Group retreat will be conducted at Cold Spring Harbor Laboratory in Year 1 with 10 faculty. In Years 2 and 3, regional and virtual workshops will be held for 80 faculty representing diverse institutions and regions of the country. Participants will be selected on the basis of proposals for tractable projects investigating differential gene expression and transcriptome re-sequencing. Faculty teams will learn all phases of a bioinformatics workflow to analyze their datasets, and leave the workshop with a curriculum plan to distribute data analysis among student teams.

Data analysis will use large-scale data storage, bioinformatic workflows, and high performance computing provided by the *iPlant Collaborative*, an NSF-supported cyberinfrastructure for biological research. Primary training will transition from in-person workshops to online webinars and self-paced learning via a dedicated Internet microsite, providing a sustainable method to introduce large numbers of faculty to NGS analysis. Participants will also share instructional strategies and solve analysis problems during regular video conferences. A multi-faceted evaluation program will assess: 1) impact of the training on faculty participants' knowledge, behavior, and teaching confidence, 2) faculty implementation of the project in a variety of classroom and student research settings, and 3) effects on student learning, interests, and attitudes. Use of the validated SURE survey instrument will allow comparison with other students' research in other fields and educational settings.

**Intellectual Merit** NGS methods have dramatically decreased the cost of obtaining whole genome data on eukaryotic organisms, and data storage and analysis workflows are becoming freely available online. In particular, RNA-Seq can provide novel data on gene structure and function. This project aims to help move undergraduate education into an age when students work with whole genome sequences as routinely as they work with PCR amplicons today. This project operates on the continuum of biology research and education. It recognizes that many college faculty would like to bring NGS to bear on a problem of their own interest – and invite students as co-investigators in class-based and independent projects. The program will prepare faculty to operate in a new, sequence-driven paradigm and empower them to guide students in novel genome explorations.

**Broader Impact** Free online tools have made sophisticated genome analysis available to anyone with an Internet connection. This project will further extend the egalitarian nature of genome research by providing an infrastructure for undergraduate faculty to generate and analyze their own genome-scale datasets. About 25% of faculty will be from minority-serving institutions with the objective of reaching African American, Hispanic, and Native American faculty and students. The project also provides faculty at predominately teaching institutions access to high performance computing through the NSF's Extreme Science and Engineering Discovery Environment (XSEDE). The *Green Line* of *DNA Subway* is an educational workflow specifically designed to support student analysis of RNA-Seq data sets. Advanced applications, including command line customization, are supported in the research-grade *Discovery Environment.* NGS sequencing will be done at GENEWIZ, a sequencing company that has provided sequencing for other distributed projects managed by the DNALC. This infrastructure will make it possible to broadly disseminate on-demand experiments using RNA-Seq in undergraduate settings.

CSH Cold Spring Harbor Laboratory DNA Learning Center

Funded by the National Science Foundation NSF

## I. RESULTS FROM CURRENT AND PRIOR NSF SUPPORT

**A Partnership to Develop Advanced Technology Units in Genomic Biology (DUE-9752037, 7/1998-6/2001, $599,825 David A. Micklos, PI). Intellectual Merit** This project developed and disseminated laboratories and custom bioinformatics resources for introducing genomic biology into undergraduate biology courses. These labs popularized the use of polymerase chain reaction (PCR) to examine three DNA polymorphisms: the D1S80 VNTR, an *Alu* transposon insertion, and the mitochondrial (mt) control region. The mtDNA sequencing lab demonstrates tight integration of wet (*in vitro*) and computational (*in silico*) components. Free DNA sequencing of student mtDNA amplicons was continuously provided by the DNALC after the lapse of NSF funding, until we shifted to an inexpensive service ($3.00 per sequence) provided by corporate partner GENEWIZ, Inc. The DNALC's *Sequencing Service* anticipated the broad availability of personal DNA sequencing, and predated the Genographic Project, which currently charges $200 per sample. The DNALC continues to maintain a database at the *BioServers* website where student sequences are automatically posted. The site includes bioinformatics tools to search Genbank (BLASTN), compare sequences (CLUSTALW), and construct phylogenetic trees (PHYLIP). ***Publications & Products*** Two coordinating Internet sites were developed, *Genetic Origins* and *BioServers* – with protocols, multimedia resources, and custom bioinformatics tools. The experiments developed under this project are a case study in sustainability. All were developed as standalone kits as part of a long-term collaboration between the DNALC and Carolina Biological Supply Company (CBSC). The kits remain bestsellers, and are used by more than 100,000 students per year. ***Broader Impacts*** Over 90,000 student mtDNA sequences have been uploaded to *BioServers*, including 50,000 samples sequenced free-of-charge by the DNALC. *BioServers* has 51,900 registered users and has received 1.5 million visits; in 2012 the average visit length was 14:20 minutes. At follow-up, 216 high school and junior college educators who participated in summer workshops reported increased ability to incorporate genomic biology (94%) and Internet materials (78%) into their instruction. Sixty percent had analyzed an *Alu* insertion or VNTR polymorphism, and 34% had student mtDNA sequenced by the DNALC. These teachers reported 5,843 student exposures to these labs per year.

**Developing and Testing New Laboratories in RNAi and Functional Genomics (DUE-0341510, 6/2004-12/2006, Phase I: $295,611; DUE-0717765, 9/2007-8/2010, Phase II: $444,133, David A. Micklos, PI). Intellectual Merit** With Phase I funding for this project, we developed and field-tested an experiment- and bioinformatics-rich curriculum that explores RNAi in the model eukaryotic organism *C. elegans* (nematode worm). The curriculum begins with observation of mutant phenotypes and basic worm "husbandry," progresses to simple methods to induce RNA interference (RNAi) and explore the mechanism of RNAi using PCR, and culminates with experimental methods to allow students to silence essentially any gene in the *C. elegans* genome. ***Publications & Products*** The *Silencing Genomes* Internet site (www.silencinggenomes.org), which has received over 80,000 visits, includes all experiments and reagent recipes. A free strain library includes needed bacterial and *C. elegans* strains, as well as more than 100 vectors developed by faculty participants to silence worm homologs to human genes. Three stand-alone kits derived from the program were released by CBSC in 2010, and protocols are published in the lab-text *Genome Science: A Practical and Conceptual Introduction to Molecular Genetic Analysis in Eukaryotes* (2012, CSHL Press). ***Broader Impacts*** We completed a Phase II project to disseminate the RNAi curriculum to 200 faculty at five-day workshops conducted nationwide. At 15-month follow-up the majority of participants had used the workshop materials to improve their teaching – performing labs with 3,926 students and providing seminars on RNAi to10,053 students and sharing the curriculum with 392 colleagues (392). Over 1,300 strain orders have been fulfilled and used with a reported 6,000 students.

**Genomic Approaches in Biotechnology (DUE-1104236, 4/2011-3/2014, $774,809, David A. Micklos, PI). Intellectual Merit** In collaboration with community college faculty and the National Advanced Technological Education Center for Biotechnology (Bio-Link), the DNALC developed the *Genomic Approaches in Biosciences* workshop as a cost-effective means to disseminate instructional modules that provide the scientific foundation for biotechnology careers in the genome age. Over its three-year term, the program will reach 288 biotechnology faculty at week-long workshops conducted at 12 community colleges nationwide. The program focuses on four key technologies – PCR, DNA sequencing, RNAi, and bioinformatics. The first seven workshops were held in 2011 and 2012. Of 272 applicants, 155 participated – including 43% two-year college faculty, 31% four-year college or university faculty, and 26% high school faculty, with 25% underrepresented minorities. ***Publications & Products*** The lab protocols are published in the lab-text *Genome Science* (2012, CSHL Press). ***Broader Impacts*** Pre-workshop (n=145) and post-workshop (n=142) survey data revealed notable increases in both faculty knowledge and confidence. Before the workshop only 12% of participants knew "a lot" about the key genomic concepts compared with 41% afterwards. There was a marked increase in participants who felt "extremely confident" in teaching the genomics labs (14% to 36%) and bioinformatics (5% to 18%). In follow-up surveys 12–15 months later (n=70), participants reported incorporating DNA barcoding (45%), genotyping (58%), RNAi (24%), and bioinformatics (46%) into their teaching. Respondents reported student exposures to labs (9,029), bioinformatics (4,796), and biotech careers (2,605) in schools with an average of 46% students from ethnic groups underrepresented in sciences.

**Course-Based Undergraduate Research Experiences Network (CUREnet) Genomic Approaches in Biotechnology (DBI-1061874, 7/2011-6/2015, $497,556, E. Dolan, D. Micklos, and N. Trautman, PIs). Intellectual Merit** This project brings together people and programs that are creating course-based undergraduate research experiences (CUREs), which offer engage large numbers of undergraduates in using the same data and tools as scientists. In the past three years alone, NSF has funded 29 CURE projects. This project aims to address substantive issues related to CURE implementation and effectiveness – documenting current projects, investigating data ownership and quality standards, and developing tools for accessing, analyzing, and contributing data. CUREnet also aims to identify effective strategies for broadening the diversity of faculty and students involved in CUREs. ***Publications & Products*** A clearinghouse website, which will be a single "go-to" location for finding and publicizing CURE information, is under development. ***Broader Impacts*** CUREnet brings together people with relevant expertise from diverse fields and institutions to address priority issues that may limit the impact and adoption of CURE pedagogies. CUREnet fosters collaborations and publicizes existing CURE resources so that undergraduate faculty can avoid "reinventing the (instructional) wheel."

**iPlant Collaborative: Cyberinfrastructure for the Biological Sciences (EF-0735191 2/08-1/13, $50,000, 000, Steve Goff, PI). Intellectual Merit** As part of the NSF's "Cyberinfrastructure Vision for 21st Century Discovery" (www.nsf.gov/pubs/2007/nsf0728/index.jsp), a partnership between the University of Arizona, Texas Advanced Computing Center, and CSHL, the *iPlant Collaborative* is developing a national computation infrastructure to solve problems in the life sciences. Community-driven working groups guided the development of accessible cyberinfrastructure (CI) resources that include a web interface to analysis tools (*Discovery Environment*), on-demand cloud computing (*Atmosphere*), cloud-based data storage (*Data Store*), and a command-line login and web-based application programming interface (API) for high performance computing on Extreme Science and Engineering Discovery Environment (XSEDE) supercomputers.[1]

***Publications & Products*** As lead of *iPlant's* Education, Outreach, and Training (EOT) component, the DNALC developed *DNA Subway*, an educational analog to the research CI. An intuitive bioinformatics platform based on the metaphor of a subway map, *DNA Subway* makes high-level sequence analysis readily available to educators and students. The site includes workflows and shared workspaces to

construct gene models and annotate gene function in raw DNA sequence (*Red Line*) and to prospect for transposons and other repeated elements in sequenced genomes (*Yellow Line*). The *Blue Line* supports a DNA barcoding workflow that encompasses biochemical and bioinformatics (B&B) components. Students isolate DNA and amplify barcode loci, then submit the DNA for sequencing by GENEWIZ. Within 48 hours, the finished sequences are automatically uploaded to the *Blue Line*, which includes all tools needed to visualize and edit barcode sequences, search Genbank (www.ncbi.nlm.nih.gov/genbank/) for matches, align sequences, and construct phylogenetic trees. The DNA barcoding experiment, is available in three formats: the online lab notebook www.dnabarcoding101.org, in the lab-text *Genome Science* (2012, CSHL Press), and a stand-alone kit from CBSC.

The *Green Line* makes next-generation sequence (NGS) analysis of eukaryotic genomes accessible to students and researchers without computation experience. Computing power needed to analyze NGS data greatly exceeds the capabilities of ordinary computing systems, so the *Green Line* is delivered entirely through the *iPlant* Foundational API, providing easy access to high performance computing (HPC) and high capacity cloud-based storage. The *Green Line* integrates the *Tuxedo Protocol*,[2] a workflow incorporating open source components for all steps of RNA sequence (RNA-Seq) data analysis – from the processing of raw data from major sequencing platforms through to publication-quality results. First, millions of NGS reads are aligned against a reference genome. The aligned reads are then used to assemble and quantify transcripts. For comparative RNA-Seq analysis, the relative abundance of transcripts from different samples can be compared and the results visualized at any scale – from whole transcriptomes to individual genes. The *Green Line* is integrated with the *Red Line*, where assembled transcripts can provide evidence for the community (or class) annotation of sequenced genomes. Tight integration with the *iPlant* CI allows users to seamlessly "graduate" to advanced analyses.

***Broader Impacts*** The DNALC has trained researchers and science educators to use *iPlant* tools and services, reaching 1,040 participants at 68 two-day workshops conducted from 2009–12 at 27 sites nationwide. At follow-up, 80% of educators trained at *Genomics in Education* workshops had used the materials in courses including general biology (34%), genetics/genomics (30%), molecular biology (19%), biotechnology (17%), and bioinformatics (13%), reaching over 1,600 students. Educators used the curriculum for background information (59%), class resources (37%) and laboratory protocols (23%), with 21% introducing a new topic, wet lab, or bioinformatics lab. Educators also shared materials with colleagues (20%), including training them in bioinformatics (15%).

## II. PROJECT GOALS AND OBJECTIVES

The goal of this project is to create an extensible infrastructure and training program that develops faculty expertise to integrate next-generation sequencing (NGS) and analysis into undergraduate instruction. The project will focus on RNA-Seq as the most tractable analysis for faculty who are new to NGS. The project will make cost-effective use of existing infrastructure, expertise, and partnerships developed with previous NSF support. The project will also democratize access to the Extreme Science and Engineering Discovery Environment (XSEDE), providing students and faculty an "on ramp" to the national supercomputing highway.

After initial testing by project staff, a Working Group of faculty from diverse undergraduate institutions will collaborate to develop training materials, RNA-Seq datasets, and related curriculum resources during a workshop in Year 1 summer workshop at CSHL.The training curriculum will be refined and disseminated through hybrid (face-to-face and web-based) workshops in Year 2, transitioning to webinar-based workshops and a sustainable, fully online learning program in Year 3. Each teacher cohort will implement RNA-seq analysis and test curriculum materials in courses and individual projects during the ensuing academic year. Over the term of the project, faculty participants will develop 30 novel datasets to

explore genome structure and gene expression in a range of eukaryotic organisms. Intensive follow-up – including regular web conferencing, a dedicated project website, and "close support" from project staff – will help ensure successful implementation by workshop participants. A multifaceted evaluation program will gauge the range of authentic research supported by the RNA-Seq datasets and determine how participation impacts teachers, students, and institutions.

**Objectives**

*Creating Learning Materials and Strategies*

1. Refine a cost-effective, student-friendly RNA-Seq workflow, and create an extensible web infrastructure to support student research projects (Year 1).

*Developing Faculty Expertise and Sustainability*

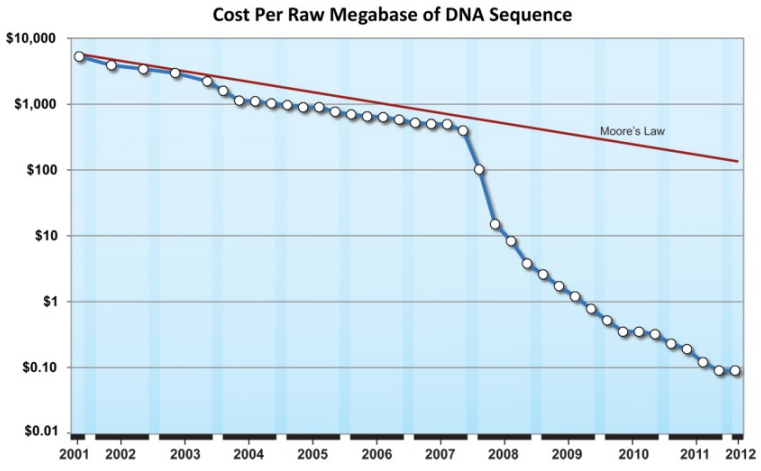
2. Convene a Faculty Working Group from diverse institutions to test the RNA-Seq workflow and develop complementary educational resources (10 faculty, Year 1).
3. Transition from in-person training (40 faculty, Year 2) to sustainable virtual training using webinars and self-paced learning (40 faculty, Year 3).

*Evaluation and STEM Research*

4. Evaluate the program to review and refine training and materials; assess teacher impact and implementation in a variety of settings; gauge student learning and attitudinal effects; and compare models of face-to-face vs. virtual training.

## III. SCIENTIFIC BASIS AND EDUCATIONAL RATIONALE

Beginning in 2005, "next-generation" sequencing methods replaced the first generation of automated Sanger DNA sequencing that was used to complete the first eukaryotic genome sequences – including *Drospohila, C. elegans, Arabidopsis*, humans, and maize. NGS methods have decreased sequencing costs 10,000-fold, to $0.10 per megabase (million basepairs), making it routine to obtain the equivalent of an entire eukaryotic genome in a single week-long experiment.[3] A literature search of the Scopus database (www.scopus.com) for "next-generation sequencing" and related terms returned over 3,000 abstracts in 2011, doubling to 7,000 abstracts in 2012. Clearly, raw DNA sequence is becoming the currency of modern biological research, and the availability of DNA and RNA-derived sequence is no longer a bottleneck to discovery. DNALC surveys of attendees at a plant biology meeting and *iPlant* training workshops suggest that 87% of researchers are currently, or soon will be, using large sequence datasets.




Easy access to genome information represents not only a quantitative, but also a qualitative shift in biological research. A decade ago, most hypotheses about molecular genetic variation were limited by the cost and difficulty of obtaining large amounts of sequence data. Now, NGS presents a virtually data-unlimited paradigm in which hypotheses are often derived from the sequence data itself. NGS technologies have "not simply changed the landscape but have placed basic, clinical and translational research scientists into a new and unfamiliar world in which entirely different types of questions can be addressed."[4]

Online databases and bioinformatics tools provide the means to explore this abundance of sequence information by predicting genes, refining gene models, identifying polymorphisms, and comparing genomes. Reduced sequencing costs and freely-available analysis resources promise to make genome analysis an egalitarian pursuit open to virtually anyone. For faculty and students alike, analyzing and comparing whole genome sequences can reveal elements of genome organization and gene structure/function that previously could only be approached as abstractions. Furthermore, the whole genome paradigm will dominate the biological landscape for students seeking future careers in biomedical and agricultural research.

***RNA-Seq and Commercial Illumina Sequencing*** We have carefully considered the experimental options that can best facilitate widespread access to NGS. Thus, the project will focus on RNA-Seq data generated on the Illumina HiSeq™ 2000 system by our commercial partner, GENEWIZ (see *Letter of Support*). RNA-Seq, or transcriptome sequencing, is based on sequencing cDNA synthesized by reverse transcriptase from extracted RNA. A cDNA library provides a detailed biological "snapshot" of the entire transcriptional activity of the eukaryotic genome, and recent studies have pushed aside the "one gene – one protein" dogma. There is now compelling evidence that 95% of multi-exon genes in the human genome are alternatively spliced, and, on average, each produces seven differently spliced mRNAs.[5] This post-transcriptional modification expands the human protein repertoire many-fold beyond the actual number of structural genes. The power of high-throughput methods including RNA-Seq was illustrated by the Encyclopedia of DNA Elements (ENCODE) project (genome.ucsc.edu/ENCODE/), which recently showed that over 80% of the human genome has a function.[6]

There is evidence to support the contention that RNA-Seq is the right entry point into NGS analysis for undergraduate faculty. According to a survey of GENEWIZ clients, transcriptome analysis was the most frequently used among ten types of NGS analyses. Two-thirds of respondents expected their high-throughput sequencing to increase over the next two years. A recent survey of *iPlant* workshop attendees and participants at a large plant biology meeting found that 87% of responders are currently using large datasets in their research, or will do so in the next 12 months. As further evidence, RNA-Seq is the training module most requested by faculty, post-doctoral fellows, and graduate students who registered for 2012 *iPlant* workshops.

An RNA-Seq dataset is filled with novel and interpretable results. Scaffolding RNA-Seq reads onto a previously-assembled genome (or even the genome of a near relative) is much less expensive and less demanding computationally than *de novo* genome assembly. RNA-Seq provides a "one-stop" method to provide data suitable for two major types of investigations: differential gene expression and transcriptome sequencing. Differential gene expression compares transcript abundance for control *vs.* experimental conditions – such as wild-type *vs.* mutant organism, healthy *vs.* diseased tissue, differences in abiotic factors or stressors, or at two different time points. Transcriptome sequencing provides data to annotate gene structure and function, to identify alternative splicing variants, and to document protein diversity. Other investigations may analyze non-coding RNAs, pseudogenes, transposons, and micro-RNAs. The RNA-Seq dataset is also the starting point for a range of experimental follow-up and student projects, including isolation of full-length genomic clones and *in vitro* expression studies.

This proposal operates on the continuum of biological research and teaching. It acknowledges that many faculty who choose teaching as their primary focus are, in fact, looking for ways to extend genomic approaches to their own research and teaching. The National Science Education Standards and education research literature emphasize that students need to be engaged in the "process" of science—asking questions, forming hypotheses, designing experiments, collecting data, analyzing results, and drawing conclusions. There is growing evidence of the effectiveness of authentic learning,[7] especially in the area of student undergraduate research programs.[8] Using the Student Undergraduate Research Experience

(SURE) survey, David Lopatto (see *Staff Roles*) showed that undergraduate research experiences provide significantly higher gains in knowledge and understanding of science and the research process, problem solving, technical skills, data analysis, collaborative work, independent thinking, ethics, and science communication.[9] Applying RNA-Seq to a problem of their own choosing provides faculty with the opportunity to further their research while engaging students in authentic, publishable research. RNA-Seq experiments are particularly flexible, with ample data to allow multiple student projects or to engage whole classes in course-based undergraduate research experiences (CUREs).

Two decades ago, teachers came into the gene age by learning to analyze individual genes; now they need to come into the genome age by learning to identify and analyze gene networks at work within biological systems. We believe that the only way for faculty to move into a synthetic way of thinking about genomes is to own and struggle with their own NGS dataset. The rapid pace at which these new sequencing technologies have developed means that faculty, even at tier one research institutions, have had little time to bring these methods into the classroom. While publication trends leave little doubt of the impact these technologies will have on biology, NGS and computational resources are currently limited to upper-echelon researchers at comprehensive research institutions. We want to bring NGS to the cusp of "doability" for faculty at smaller institutions – those who do not have access to a DNA sequencer, a supercomputer, or an experienced bioinformatician. This project provides a sustainable infrastructure that will empower faculty to bring NGS datasets into college classroom and to invite students as co-investigators in the exploration of eukaryotic genomes.

## IV. WORK PLAN

**Objective 1: Refine a cost-effective, student-friendly RNA-Seq workflow, and create an extensible web infrastructure to support student research projects (Year 1).**

At the outset of the project, DNALC project staff will test the RNA-Seq biochemical and bioinformatics protocols to generate a novel transcriptome from a plant sample. This will involve: 1) isolating total RNA, 2) transferring samples to GENEWIZ for library preparation and sequencing, 3) uploading sequence files to the *iPlant* Data Store, and 4) bioinformatics analysis. It is critical to make use of techniques and resources that are robust and widely-available, so that our teaching model can be deployed across diverse classroom settings. To that end, the biochemical workflow will minimize specialized equipment, and the analytical workflow will be based on open-source tools and services available from the *iPlant Collaborative*. We negotiated with GENEWIZ to provide a 15–20% educational discount. This service will be available to any educator affiliated with the NSF project. We expect the cost of sequencing will continue to decline over the course of the project and will renegotiate educator prices periodically, thus providing sustainable access to sequencing and advances in technology for future participants.

***RNA Extraction*** Kits from Ambion (Life Technologies®) are tailored to animal, plant, and cultured tissues. These kits involve: 1) mechanical disruption of the cells, 2) reagents to lyse cellular structures and stabilize RNA, 3) RNA precipitation on a disposable mini-spin column, and 4) elution into a buffer. RNA samples are then quantified using either a BioAnalyzer™ (Agilent®) if available, or by electrophoresis with pre-cast denaturing gels. Acceptable RNA samples will be dried with an RNAstable® reagent, and shipped to GENEWIZ for sequencing. At GENEWIZ, the samples will be quantitated and prepped using the standard Illumina protocols.

We will support two types of RNA-Seq analysis projects, requiring distinct but overlapping approaches. Although these approaches differ slightly in experimental design, both make use of the same bioinformatics and web resources:

- ***Transcriptome sequencing*** This experiment is designed to confirm the structure and expression levels of known or predicted transcripts, as well as to identify novel transcripts from previously unknown genes and novel alternatively spliced isoforms for known genes. RNA from at least two biological replicates will be extracted to satisfy experimental best practices. After sequencing, the *Green Line* will be used to align the reads to a reference genome, merge results from samples, and analyze the transcriptome. The structure of transcripts derived from the RNA-Seq experiments can be used as evidence for gene model annotation in the *DNA Subway's Red Line*. Where existing genome annotations are available, *Cuffcompare* can be used to compare the experimentally-derived transcriptome data with the reference annotation to identify novel transcripts.

- ***Analysis of differential gene expression*** RNA from samples that differ due to mutation, developmental time point, or experimental condition is sequenced to identify genes whose expression is either increased or decreased relative to a reference sample. In cases where more than two samples are considered, pair-wise comparisons are done for all combinations. This analysis uses all components of the *Tuxedo* workflow within the *Green Line*. The identification of novel transcripts is also possible with this type of experiment.

***Web and Instructional Infrastructure*** A dedicated website will support all aspects of the project, linking participants and students across the country. This site will be modeled on an existing website created for the DNALC's *Urban Barcode Project* (www.urbanbarcodeproject.org, see *Institutional Capability*).

As shown in the table, the website will support a variety of educational activities and materials, The electronic resources will evolve over the course of the project and incorporate materials that are further described under Objectives 2 and 3. The website will provide practical information and insights to support classroom implementation and student research projects: 1) video screencasts explaining RNA-Seq experiments, RNA extraction, and the bioinformatics workflow; 2) downloadable protocols and multimedia resources; 3) teaching resources developed by project staff and workshop participants.

| Year | 1 | 2 | 3 | Post-Grant |
|---|---|---|---|---|
| *iPlant* analysis tools: *DNA Subway, Discovery Environment, Atmosphere* | ✓ | ✓ | ✓ | ✓ |
| Website protocols, how-to videos, project summaries and lesson plans | ✓ | ✓ | ✓ | ✓ |
| Tele/video introduction and RNA isolation | ✓ | ✓ | ✓ | |
| RNA-Seq datasets | ✓ | ✓ | ✓ | ✓ |
| Tele/video follow-up and implementation | ✓ | ✓ | ✓ | |
| Instructional workshops | ✓ | ✓ | | |
| Instructional webinars | | | ✓ | ✓ |
| Online self-paced course | | | ✓ | ✓ |

A *Google Maps* utility will show the locations of RNA-Seq projects with links to project summaries, relevant metadata on the RNA-seq datasets, and interviews with faculty and student participants. An interface to project datasets will allow faculty participants to distribute analyses to different student groups. Two thirds of faculty participants (Years 1-3) will make use of datasets produced by others and the datasets will be freely available to students and teachers nationwide at the close of the project.

The project will make use of two types of synchronous electronic instruction, in which faculty participants receive the same content in the same time frame. Webinars will be "live" online productions, with formal content "pushed" to participants, followed by an interactive question and answer period. Web conferences will be informal sessions with participants connecting by Internet or telephone. In each case, the presenter and participants can share presentations on their computer desktops. Although the *iPlant Collaborative* has a license for the Webex system, we will also explore the use of Skype, which is expanding its capabilities under Microsoft. Webinars and web conferences will be professionally produced and distributed from the DNALC's *Landeau Multimedia Studio* (see *Institutional Capabilities).*

**Objective 2: Convene a Faculty Working Group from diverse institutions to test the RNA-Seq workflow and develop complementary educational resources (10 faculty, Year 1).**

Ten faculty members will join a project Working Group, using the participant recruitment and proposal selection criteria described below. A ten-day retreat at CSHL in summer 2013 will introduce RNA-Seq analysis to Working Group members and develop curriculum to support student use of the RNA-Seq workflow. Participants will be housed on the main CSHL campus, providing an insider's view of a world-class research institution and providing opportunities to interact with CSHL staff, visiting scientists, and other participants in concurrent Postgraduate Training Courses. Many ideas and collaborations have gotten started during informal discussions at the CSHL cafeteria or after hours in the bar! We expect that one faculty participant will be from the local area and can commute to the retreat.

***Participant Recruitment, Proposal Evaluation, and Pre-Training Preparation*** The project will be advertised annually, using faculty mailing lists maintained by the DNALC (~4,000), *iPlant Collaborative* (~4,000), Bio-Link National ATE Center for Biotechnology (~1,000), and CUREnet. We have access to a large number of highly motivated DNALC workshop alumni, and the workshop will also be announced in a feature on the DNALC homepage, which receives two million visits annually. To reach our goal of 25% underrepresented minorities, we will use mailing lists and postings at the Society for the Advancement of Chicanos and Native Americans in Science (SACNAS), Historically Black Colleges and Universities (HBCU), Tribal Colleges and Universities (TCU), and the Black Collegian Online.

Applicants will complete a formatted proposal including: 1) importance and relevance of the proposed dataset, 2) management plan about how data will be shared/partitioned between student teams and made available to collaborating institutions, 3) description of course and/or research contexts in which students will be involved, and 4) numbers of students involved and duration of exposures. The application will also include information on personal attributes (academic preparation and research, classes taught, students mentored, and computational expertise) and institutional attributes (NGS capability, rural or non-urban setting, and percentages of disadvantaged and minority students). A support letter from the department head or division director must demonstrate an institutional commitment to incorporate NGS analysis in appropriate courses and to directly analyze the RNA-Seq dataset during the upcoming academic year. In addition to supporting high-quality proposals, participant selection will be balanced to represent diverse types of institutions (two- *vs.* four-year, research *vs.* liberal arts, etc.) and domains of expertise (organism/system specific knowledge *vs.* computation/informatics).

Based upon review of proposals received, four Working Group members will be "funded" to develop their proposed transcriptome sequencing projects, and two will be "funded" to develop their differential gene expression projects. Four Working Group members, who may have limited laboratory resources or expertise, will join one of the "funded" projects or work with the transcriptome dataset generated during workflow testing at CSHL. Pre-retreat web conferences will introduce Working Group members to the project and to each other, and will assist them with their RNA extractions (using kits provided by the project). RNA will be shipped to GENEWIZ for sequencing, and the RNA-Seq datasets will be available at the start of the retreat.

***The Training Experience*** The retreat is designed around individual learning cycle.[10] Goals and theoretical background will be introduced in concept seminars followed by hands-on bioinformatics sessions. As shown in the agenda below, the retreat will include conceptual seminars (S), research facility tours (T), hands-on bioinformatics (B), curriculum planning (C), and online materials development (D).

| Day | Morning | Afternoon |
|---|---|---|
| **1** | Welcome/Introductions/Course Objectives | S: GENEWIZ and DNA sequencing |

| | | |
|---|---|---|
| | S: NGS and Data-driven Biology<br>S: Presentation of "funded" faculty experiments | T: CSHL Core Sequencing Facility<br>T: CSHL and Modern Biology |
| **2** | S: Introduction to *iPlant*<br>B: Using *DNA Subway* | B: *Tuxedo* package for RNA-Seq<br>S: Data interpretation and validity testing |
| **3** | B: Analysis of faculty datasets | B: Analysis of faculty datasets |
| **4** | B: Advanced use of the *iPlant Discovery Environment/ Atmosphere* | B: Analysis of faculty-generated data |
| **5** | B: Analysis of faculty datasets<br>S: Case Studies of NGS Research | C: Curriculum assignments<br>C: Break-out group assignments |
| **6** | C: Break-out groups on course-based projects | C: First presentation of lesson plans |
| **7** | C: Planning and developing course materials | C: Breakout groups work on course outlines |
| **8** | C: Final presentation of course outlines | S: Workshop and project evaluation plan |
| **9** | D: Materials development & videography | D: Materials development & videography |
| **10** | D: Materials development & videography | Evaluation of the Working Group retreat: Participant feedback |

Bioinformatics sessions will introduce the concept of a "power desktop" in which a participant's personal computer becomes a portal for high performance computing (HPC) and high capacity, cloud-based data storage. Instruction will be based on the *Green Line* of *DNA Subway* (described in *Results from Current and Prior NSF Support*), which provides an intuitive, and virtually foolproof, user interface for all phases of RNA-seq analysis. However, participants will also learn how to use "expert" configurations and command-line capabilities of the *Discovery Environment*, which are appropriate for computationally minded students.

We *do not* intend that faculty participants will fully analyze their datasets during the retreat, but, rather, that they will return to their institutions with a "work in progress." The retreat will arm them with targeted bioinformatics skills and plans to explore the datasets for course-based and/or individual student projects. The aphorism "garbage in, garbage out" is especially germane as large-scale datasets and sophisticated bioinformatics resources become widely available. Thus, the retreat will also emphasize that critical evaluation of RNA-Seq results is a prerequisite to drawing valid biological conclusions.

Curriculum planning sessions will encourage faculty to develop lesson plans and support materials for a variety of undergraduate courses and independent student projects. With the assistance of Working Group members, these materials, as well as "how-to" videos and participant testimonial interviews, will add a new dimension of educator insight to the evolving project website and online training materials for regional and virtual workshops conducted in Years 2 and 3.

***Participant Follow-up Support*** In the academic year following the retreat, Working Group members will share experiences and insights during monthly web conferences. Importantly, these sessions will allow us to support implementation of the RNA-Seq workflow in a variety of biological and computational settings – from targeted exposures in introductory biology courses to distributed projects in mid-level courses and sustained independent research and computational investigations in upper-level courses. DNALC staff will provide "close computational support," via telephone and web, to help faculty trouble-shoot technical and data manipulation problems. We will also help faculty implement check-point tests to insure data validity. In addition to desktop sharing, follow-up support will make use of *iPlant Atmosphere*, which will allow one or more participants to work together with the same dataset and genome viewer. Using customized virtual desktops with access to multiple processors, *Atmosphere* cloud computing allows users to visualize and analyze RNA-Seq data that is too complex to view on normal browsers.

**Objective 3: Transition from in-person training (40 faculty, Year 2) to sustainable virtual training using webinars and self-paced learning (40 faculty, Year 3).**

During Year 2, we will conduct two 5-day summer workshops to introduce RNA-Seq analysis to a broad spectrum of faculty. The workshops will be sited at the institutions of two Working Group members, who will serve as local organizers and liaisons. Situating workshops at institutions in relatively large metropolitan areas will allow 25% of participants to commute, with 75% receiving housing support. Twenty participants will be recruited, selected, and prepared for each workshop as described above. Each workshop will support 4 transcriptome and 2 differential gene expression projects, with 14 faculty joining a "funded" project or working with a dataset developed in Year 1. Experience from the Working Group retreat and follow-up will allow us to structure an efficient experience – including conceptual seminars (S), hands-on bioinformatics (B), and curriculum planning (C) as outlined below. Participants will receive academic year follow-up support as described above.

| Day | Morning | Afternoon |
|---|---|---|
| **1** | Welcome/Introductions/Course Objectives<br>S: NGS and Data-driven Biology | S: Presentation of "funded" faculty experiments |
| **2** | S: Introduction to *iPlant*<br>B: Using *DNA Subway* | B: *Tuxedo* package for RNA-Seq<br>S: Data interpretation and validity testing |
| **3** | B: Analysis of faculty datasets | B: Analysis of faculty datasets |
| **4** | B: Advanced use of the *iPlant Discovery Environment/ Atmosphere* | B: Analysis of faculty datasets |
| **5** | C: Curriculum and lesson planning | C: Curriculum and lesson planning |

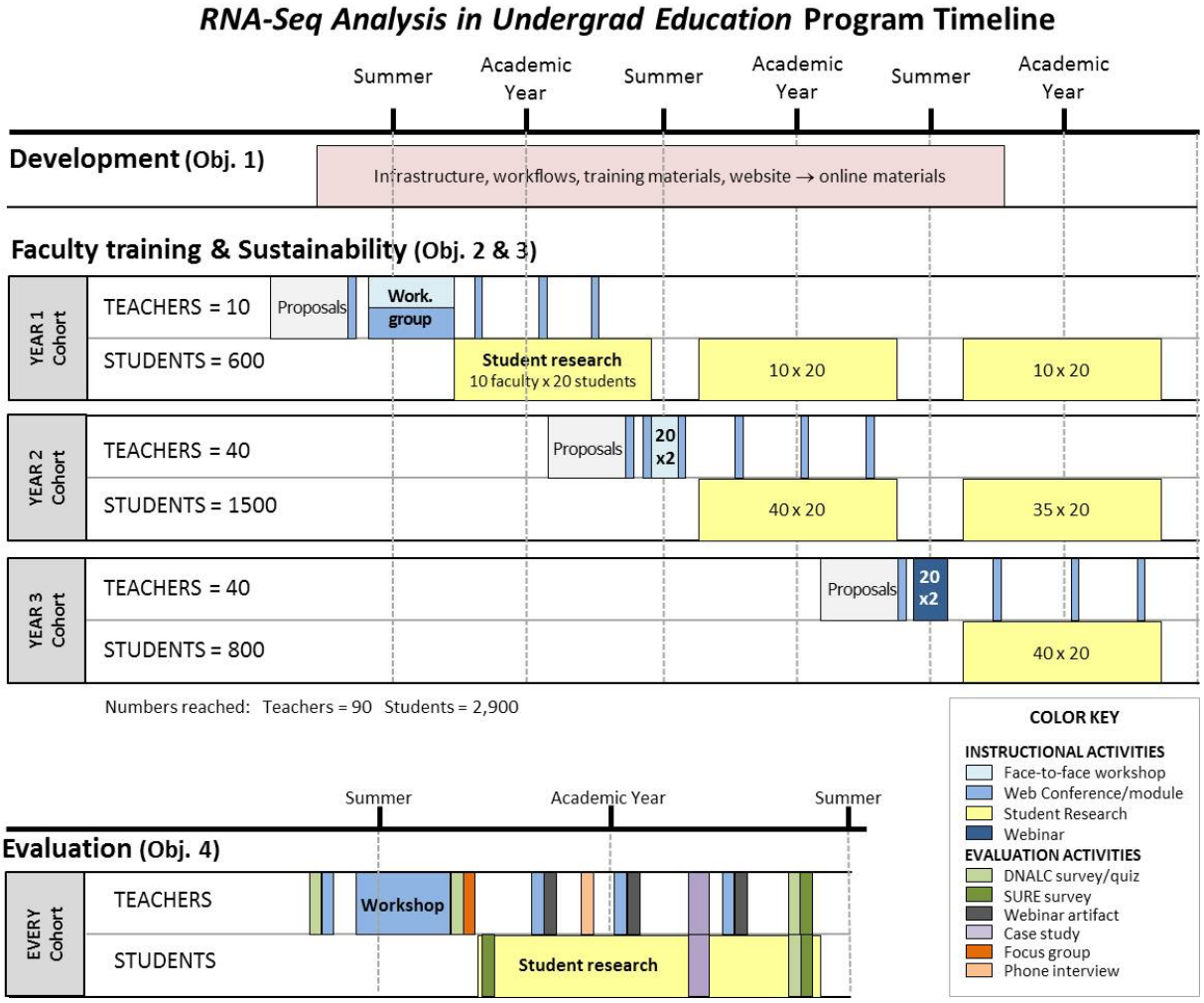
We will videotape sessions at the second iteration of the workshop, by which time we expect that presentation materials will be highly evolved. These videos will guide development of core webinars that will anchor instruction in Year 3. Twenty faculty will be selected according to the established procedure to participate in each of two webinar-based courses – each supporting 4 transcriptome and 2 differential gene expression projects. The webinar-based course will follow the basic flow of the Year 2 workshop. Following introductory webinars, participants will work independent on their datasets – with web conferencing used for questions and wrap-up. We will test asynchronous learning materials with a subset of Year 3 participants – replacing pre- and post-training web conferences with tutorials and video recordings of frequently asked questions.

Based upon experience and formative evaluation from the webinar-based training, the project website will complete its evolution as the central portal of an integrated training solution for asynchronous, self-paced learning. The website will provide an easy-to-use interface for users to navigate tutorials and video instructional materials, save personal notes, and track personal progress. The online course will integrate instructional and testimonial videos recorded over the term of the project. Parallel video and narrative quick-start guides will briefly introduce the RNA-Seq workflow, while more extensive tutorials will provide users detailed guidance on how to use the tools with their own data. The website will also benefit from an extensive bank of faculty-developed lesson plans and instructional use cases – available in narrative and video formats.

**Objective 4: Evaluate the program to review and refine training and materials; assess teacher impact and implementation in a variety of settings; gauge student learning and attitudinal effects; and compare models of face-to-face vs. virtual training.**

A multi-faceted evaluation program will build on several large-scale evaluations of DNALC web- and experiment-based programs.[11,12] The program will employ qualitative and quantitative methods to a)

monitor the evolving program, b) assess impact on faculty, c) assess impact on students, and d) compare the effectiveness of face-to-face vs. virtual training. Group feedback will take advantage of occasions when participants are gathered at the retreat or a workshop, supplemented by remote data collection tools, such as online surveys and telephone/web-based interviews. An external evaluator, David Lopatto, of Grinnell College, will conduct a summative evaluation using the Student Undergraduate Research Experience (SURE) survey, developed to measure the effects of student research on students and faculty. The study design is summarized below in the context of the program objectives.




***a) Monitor the evolving program*** Formative evaluation will review and refine faculty training experiences and workshop materials. The Year 1 Working Group retreat and Year 2 faculty workshops will provide quantitative and qualitative data on the usability and effectiveness of the project website, RNA-seq workflow, and instructional materials. Faculty workshops will incorporate a feedback protocol at the end of each day to reflect on accomplishments and obstacles encountered/overcome and surveys will assess workshop satisfaction, highlights, and challenges. The internal evaluator (Dr. Nisselle) will attend the Working Group retreat to observe faculty interactions, and conduct informal one-on-one interviews and a group discussion addressing overall satisfaction, areas of workshop strength and areas in need of improvement, technical problems, and anticipated implementation of the RNA-Seq materials. Focus groups will also be held at the Year 2 workshops, and via teleconference in Year 3. We will also collect program data relating to number of faculty trained, RNA sequences, curricula implemented

(number of projects and student exposures), use of online materials, barriers to implementation, etc.

***b) Assess impact on faculty*** A set of longitudinal surveys will provide both formative and summative data. We will track faculty attitudes and classroom behavior at three time points, using methods developed at the DNALC over two decades (see *Results from Current and Prior NSF Support*). The faculty surveys document changes in knowledge, lab and bioinformatics expertise, teaching confidence and behavior, as well as implementation – courses taught, numbers of students reached, problems encountered, and areas requiring further support. Respondents will be added to cumulative data from over 7,000 faculty respondents to previous DNALC surveys. To allow for comparison with our existing data sets, identical or equivalent questions will be asked whenever possible. Standard question sets include teaching experience, attitude and confidence, workshop experience, knowledge items tailored to content (NGS, RNA-Seq), and anticipated vs. actual implementation and dissemination of materials. The DNALC surveys will be supplemented with the external evaluation of faculty experiences of undergraduate research using the SURE-III survey for faculty mentors. Qualitative data will include focus groups (Years 1 and 2 and via teleconference in Year 3), and transcripts of all web conferences and webinars.

***c) Assess impact on students*** A mixed methods summative evaluation will assess student learning, interests, attitudes, and career choices relating to the undergraduate research experience. To assess student learning, at the completion of the unit students will complete a standard quiz developed by the DNALC to assess their understanding of key NGS, RNA-Seq, and bioinformatics concepts. The SURE-III survey will be externally administered before and after the research experience. The SURE-III survey was developed for college settings, with data from over 3,000 students and faculty. Students will complete the survey at the beginning of the project (for baseline data) and again as a hurdle requirement for completing a unit to reflect on their experience and compare it with other research subjects. We will also conduct classroom case studies in the second half of each academic year with one trained faculty per cohort per year (a total of 6 case studies) to gauge how working with NGS datasets affects student attitudes towards science and scientific research and the formation of key biological concepts. These case studies will involve in-class ethnographic observations, teacher interviews, and student focus groups (6–8 participants). Case studies will include a total of six teachers and classes and 36–48 student focus group participants.

***d) Compare face-to-face vs. virtual training*** We will use all data from all three aims above to conduct STEM research about these different models of faculty training, comparing impact (to both trained faculty and their students), sustainability, and practical applications. In particular we will focus on whether there are any differences in faculty knowledge and behavioral changes, curriculum implementation, and ultimately student impacts, and, if so, how best to leverage positive elements of both training models.

***Data collection, storage and analysis*** All surveys will be administered using validated survey methods that have produced high response rates in previous surveys of workshop participants.[13] Surveys administered by the DNALC will be collected using Survey Monkey (www.surveymonkey.com). The SURE surveys will be completed online via the Grinnell College portal (www.grinnell.edu/academic/csla/assessment/sure). This equates to potentially over 6,000 completed surveys – 2,900 x 2 student surveys and 90 x 3 teacher surveys. Student submissions to public browsers and databases, use of the Internet website, and publications arising from the program will also be tracked as part of the summative evaluation. All electronic data will be collected using *Survey Monkey* and *Google Analytics* software, and via auditing public browsers and databases. Descriptive and inferential analyses and multiple logistic regressions will be used to highlight demographic, attitudinal, and practical contributors to teaching practice; modified grounded theory will be used to analyze qualitative data for themes. Both quantitative and qualitative data will be used to respond to the four evaluation aims, especially, to compare the experiences and impact of face-to-face vs. virtual training.

***Dissemination of research findings*** Training materials, experiences, and evaluation insights from this project will help faculty bring other student research projects to scale, and inform STEM education. We will provide updates on the project website and at relevant educational, scientific, and professional meetings, e.g. the National Association for Research in Science Teaching, which generate refereed publications. We will also submit at least one publication focusing on educational development and research to a journal such as *The American Biology Teacher* or *Journal of Research in Science Teaching*.

## V. INSTITUTIONAL CAPABILITY

***Cold Spring Harbor Laboratory (CSHL)*** is ranked #1 worldwide in citation impact in molecular genetics and genomics research. A private, non-profit, basic research and educational institution, over 400 scientists conduct groundbreaking research at CSHL in cancer, neurobiology, plant genetics, and bioinformatics. Established 122 years ago, CSHL is recognized internationally for its educational activities, including international scientific meetings and courses, Watson School of Biological Sciences doctoral program, CSHL Press, Banbury Conference Center, and Hazen Genome Research Center.

The ***DNA Learning Center (DNALC)*** is an operating unit of CSHL, extending its traditional research and postgraduate education mission to the college, pre-college, and public levels. Founded in 1988, the DNALC is the world's first science center devoted entirely to genetics education, and the largest provider of student lab instruction in molecular genetics, operating six teaching laboratories in Cold Spring Harbor, Lake Success, and Manhattan. The DNALC popularized useful methods for delivering genetics laboratory instruction to large numbers of teachers and students – including equipment-sharing consortia, mobile vans to carry instructional labs to remote sites, and laboratory field trips. Over 280,000 precollege students have conducted hands-on experiments at the DNALC; an additional 145,000 have received intensive instruction from DNALC staff at their schools. The *DNA Science* laboratory curriculum has sold 90,000 copies and catalyzed bringing hands-on bacterial molecular genetics experiments into classrooms. *Genome Science* now updates lab and bioinformatics explorations of eukaryotes. More than 6,500 precollege and college faculty received intensive training in lab and computer technology at DNALC workshops conducted in 48 states and 12 foreign countries. Twenty-five kits, developed with Carolina Biological Supply Company, are used by 200,000 students annually. The DNALC is also one of the largest providers of multimedia learning materials for biology education, publishing 22 content and bioinformatics sites with major funding from federal and private foundation sources. These sites, along with a *YouTube* Channel, and smartphone apps received over 7.2 million visitors in 2012. The scope and value of these educational tools was recognized by the journal *Science*, which honored the DNALC websites with the 2012 *Science* Prize for Online Resources in Education.[12]

In addition to the mtDNA sequencing program discussed in Results from Current and Prior NSF Support, the DNALC recently implemented another distributed DNA sequencing project for students. The *Urban Barcode Project* (UBP) is a science competition aimed at supporting independent, open-ended investigations using DNA barcoding led by New York City high school students. Seventy-five teams, including 218 students from 31 high schools, completed research projects within five thematic categories: 1) wildlife in parks and public spaces; 2) biodiversity of traded products and possible commerce of endangered species; 3) food mislabeling; 4) public health and diversity of disease vectors; and 5) presence of exotics and invasive species. Twenty-six percent of students come from underrepresented minorities (i.e. Hispanic or African American) and 38 teachers and mentors, from 29 institutions, participated as advisors. More than 2,500 single read sequences have been obtained, from approximately 1,000 samples (15 samples per team on average). During the 18-month grant period, we developed an entire infrastructure that can be replicated elsewhere in the world as outreach for the *International Barcode of Life Project*. This includes simplified protocols to amplify DNA barcodes from a broad diversity of plant, animal, and fungi species, and experiment support through *Open Lab* sessions at several venues and equipment footlockers shipped to schools. An agreement with GENEWIZ provides high-quality DNA sequencing at low cost and rapid turn-around (48 hours). The *Blue Line* of *DNA Subway* provides a simplified bioinformatics workflow for DNA barcode analysis. The project website supports all phases of the project, including management and tracking of student projects.

## STAFF PROJECT ROLES

**David Micklos** will oversee all aspects of the project, working with senior staff to ensure timely workflow and accomplishment of project objectives. With dual education in biology and journalism, Mr. Micklos has 30 years' experience administering grants from federal and private sources. He is founder and executive director of the DNALC. His laboratory texts and kits have helped to popularize DNA lab instruction – including *Genome Science: A Practical and Conceptual Introduction to Molecular Genetic Analysis in Eurkaryotes* (2012), co-authored with Bruce Nash. He was a member of the National Research Council study, *The Role of Scientists in the Professional Development of Teachers*.

**Sheldon McKay** will be responsible for the bioinformatics workflow deployment and testing, as well as preparing and instructing workshops and webinars. Dr. McKay has bachelor and master degrees in genetics, a Ph.D. in evolutionary genetics and training in bioinformatics. Dr. McKay developed the Green Line of *DNA Subway*, which is the bioinformatics basis of this proposal. He has extensive experience with comparative and functional genomics analysis and related bioinformatics, including NGS workflows. For more than a decade, he has been active in collaborative, open source scientific software development and is a leading member of the Generic Model Organism Database (GMOD) project.

**Bruce Nash** will be responsible for developing technical aspects of the demonstration workflow, selecting participants, and workshop/webinar instruction. Dr. Nash has a bachelor's degree in genetics and Ph.D. in molecular genetics. His expertise in RNAi and the RNA world are relevant to this project. He has been instrumental in developing, field-testing, and refining DNALC curricula, teaching high school and college students and instructors, and supervising students doing independent research projects.

**Amy Nisselle** will oversee all aspects of website and online materials production, as well as program evaluation, including study design, participant management, data collection, analysis, and report writing. With a Ph.D. in genetics education and evaluation, she has worked and published in medical genetics, multimedia science education, and evaluation across academic, health, education and corporate sectors internationally. She has extensive practical experience with both quantitative and qualitative research methods and is the evaluator for the NSF-funded *Genomic Approaches in Biosciences* (DUE1104236).

**Eun-Sook Jeong**, DNALC multimedia designer, will be responsible for designing and implementing the project website, based upon the design for the *Urban Barcode Project.* She will videotape and produce audio/video podcasts that explain RNA-Seq experiments and bioinformatics workflows. Ms. Jeong is the lead designer for the *DNA Subway* and she will help develop interfaces to RNA-Seq datasets.

**Young Koung Lee**, CSHL post-doctoral fellow in the Ware Lab, has expertise in NGS, particularly RNA-Seq biochemistry and analysis. Dr. Lee will advise on RNA isolation and RNA-Seq analysis. She has undergraduate and masters degrees in plant applied science and molecular breeding, and a PhD in plant molecular biology. She has experience in microarray analysis and RNA-Seq analysis and comparative genomics. Dr. Lee also works on *iPlant*, validating *iPlant* tools and services through comparison with independently-derived analyses.

**David Lopatto** is the Samuel R. and Marie-Louise Rosenthal Professor of Natural Science and Mathematics at Grinnell College. With undergraduate, masters, and doctoral degrees in psychology, Dr. Lopatto's research interests include assessment of undergraduate science learning, including classroom and interdisciplinary science assessment in terms of students' personal and professional growth. He has published extensively in this area, and has evaluated similar programs, such as the HHMI-funded Genomics Education Partnership and the SEA PHAGES program. He is an external evaluator for the

program using his Survey of Undergraduate Research (SURE-III) instrument to collect and analyze pre- and post-experience data for students and faculty.