Genetics Education

Innovations in Teaching and Learning Genetics

Edited by Patricia J. Pukkila

Involving Undergraduates in the Annotation and Analysis of Global Gene Expression Studies: Creation of a Maize Shoot Apical Meristem Expression Database

**Brent Buckner,* Jon Beck,[†] Kate Browning,* Ashleigh Fritz,* Lisa Grantham,*,[1]
Eneda Hoxha,* Zhian Kamvar,* Ashley Lough,*,[2] Olga Nikolova,[†,3]
Patrick S. Schnable,[‡] Michael J. Scanlon[§] and Diane Janick-Buckner*,[4]**

*Division of Science, Truman State University, Kirksville, Missouri 63501, [†]Division of Math and Computer Science,
Truman State University, Kirksville, Missouri 63501, [‡]Departments of Agronomy and Genetics,
Developmental Biology and Cell Biology, Iowa State University, Ames, Iowa 50011 and
[§]Department of Plant Biology, Cornell University, Ithaca, New York 14850

ABSTRACT

Through a multi-university and interdisciplinary project we have involved undergraduate biology and computer science research students in the functional annotation of maize genes and the analysis of their microarray expression patterns. We have created a database to house the results of our functional annotation of >4400 genes identified as being differentially regulated in the maize shoot apical meristem (SAM). This database is located at http://sam.truman.edu and is now available for public use. The undergraduate students involved in constructing this unique SAM database received hands-on training in an intellectually challenging environment, which has prepared them for graduate and professional careers in biological sciences. We describe our experiences with this project as a model for effective research-based teaching of undergraduate biology and computer science students, as well as for a rich professional development experience for faculty at predominantly undergraduate institutions.

ONE essential component of the success of genomics research has been the development of the field of bioinformatics, which can be defined as the use of information technology for the collection, storage, retrieval, and analysis of genomic data. Collaborations of biologists, computer scientists, and statisticians have become more robust in recent years; current graduate students in genetics commonly receive at least some formal training in computational biology. In addition, bioinformatics graduate degrees are now being offered by several institutions (ZATZ 2002). However, it remains a challenge to involve undergraduate biology students, particularly freshmen and sophomores, in genomics and bioinformatics research. Moreover, establishing undergraduate genomics research can be particularly difficult at undergraduate institutions where collaborations between biologists and computer scientists have been slower to develop, or where there historically has not been a strong culture of research.

Many undergraduate biology programs introduce cell biology and genetics during freshman introductory courses and require additional courses in cell biology and genetics later in the curriculum (LEDBETTER and CAMPBELL 2005). Thus, the beginning biology student's view of biology is largely a cellular and molecular genetics one. Too often students are taken to the brink of understanding the networks and circuitry involved in cell function, but are unable to utilize and develop this knowledge in a research environment. Furthermore,

[1]*Present address:* Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011.

[2]*Present address:* Division of Biological Sciences, University of Missouri, Columbia, MO 65211.

[3]*Present address:* Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011.

[4]*Corresponding author:* Division of Science, Truman State University, 100 E. Normal St., Kirksville, MO 63501.   E-mail: djb@truman.edu

while undergraduates are Internet savvy, few realize that most bioinformatics tools are readily accessible and user friendly. Thus, when undergraduates do engage in genetics research, they are likely to seek out "wet-lab" experiences rather than bioinformatics or wet-lab/bioinformatics combinations (DYER and LeBLANC 2002).

With appropriate training and nurturing, properly prepared undergraduate students can make meaningful contributions to the functional annotation and analysis of microarray hybridization data. In addition, students develop a true sense of biophilia while performing this type of research. The ever-increasing publicly available genomic sequence and microarray data provide an unprecedented opportunity for undergraduate students and their mentors to contribute to advances in genomics. Furthermore, bioinformatics research is relatively inexpensive to perform and can be integrated into existing laboratory exercises.

A vital component of developing an undergraduate bioinformatics research project is to establish a collaboration between biology and computer science faculty. Computer scientists are essential to the design, execution, and maintenance of a bioinformatic database. Also, collaboration between computer scientists and biologists will foster the creation of interdisciplinary courses that are desirable for students considering postgraduate study in bioinformatics (BECK *et al.* 2007). Finally, the biologists can provide a stimulating applied context to the computer science curriculum.

Herein we focus on the involvement of biology undergraduates in functional genomic analyses of the maize shoot apical meristem (SAM), a pluripotent mass of stem cells that is ultimately responsible for the development of all lateral organs in the plant shoot. Our project supports four to eight biology students and one to two computer science students annually. We offer our experiences with this project as a model for effective research-based teaching of undergraduate biology and computer science students, as well as for a rich professional development experience for faculty.

TRAINING AND RECRUITING UNDERGRADUATE RESEARCHERS

Introducing the relevant biology: Relevant biological background for an undergraduate research project can be introduced in a formal biology course. At Truman State University, the basics of SAM structure and function are covered in our freshman introductory biology course sequence. Students then complete courses in both cell biology and genetics in their sophomore year. Thus, when genetics students are involved in our bioinformatics laboratory exercise (described below), it is only necessary to review plant structural biology and cellular circuitry in the context of microarray technology. Since we generally use our research as a focal point for the investigative laboratory experiences in our advanced

courses, many students are exposed to additional and more substantive reviews of SAM biology. Finally, undergraduate research students are encouraged to enroll in a special topics course where they read secondary and tertiary literature on plant developmental biology, emphasizing the dual SAM functions during leaf initiation and stem cell maintenance. This course also emphasizes how the gene products highlighted in our readings might be expected to be differentially regulated in our SAM microarray studies. Throughout this special topics course, the microarray data are accessed, reviewed, and integrated into the class discussions.

Introduction to the relevant bioinformatics tools: Sophomore-year genetics students participate in an investigative bioinformatics analysis of microarray data. This exercise is conducted during four 3-hr labs toward the end of the semester. Students work in pairs to annotate the function of 5–10 of the upregulated genes from a SAM microarray hybridization experiment. Students manually retrieve DNA sequences from the NCBI database (http://www.ncbi.nlm.nih.gov/) and then perform BLAST searches on these sequences. BLASTN and BLASTX analyses are evaluated in an attempt to provide a functional annotation based on nucleic acid sequence identity and/or amino acid sequence identity/similarity for the EST probe on the microarray. The students also perform InterProScan searches (http://www.ebi.ac.uk/InterProScan/) on these sequences either to corroborate their BLAST results or to identify a functional domain that could be used in annotation. Finally, students use PubMed and Google Scholar to identify primary literature describing the molecular/biochemical function of the gene product and to consider how the sequence could potentially be involved in SAM function (Table 1).

Student pairs then give a 5- to 10-min oral presentation to the entire lab group. Students are encouraged to participate in the discussion of each presentation to help visualize the genetic and cellular circuitry underlying specific SAM functions. By the end of this laboratory exercise, students have received initial training in functional annotation sufficient to adequately prepare them for in-depth involvement in our research project. In addition, students are challenged to recall, apply, and expand knowledge gained in previous courses, especially cell biology. It also provides many students with their first opportunity to read the primary literature; a skill that has been incorporated into many upper-level science classes (JANICK-BUCKNER 1997; MUENCH 2000; SMITH 2001), as well as introductory science classes (PALL 2000; GILLEN *et al.* 2004; RUSSELL *et al.* 2004).

Training the annotators: Nearly all of the biology students who have joined our research team were recruited from our sophomore-level cell biology and genetics classes. We prefer to start working with students in the fall of their junior year, which allows these students to take the special topics course in plant developmental

TABLE 1

Summary of the annotation process

Annotation of accessions in the Genetics course bioinformatics exercise

Retrieve accession's sequence from NCBI (http://www.ncbi.nlm.nih.gov/).

Perform BLASTN and BLASTX of sequence (http://www.ncbi.nlm.nih.gov/BLAST/).

Evaluate BLAST results, examining *e*-values and the extensiveness of the "hits" with subject sequences.

Examine information in subject sequences' GenBank report.

Perform InterProScan on sequence (http://www.ebi.ac.uk/InterProScan/).

Collect information on the molecular/biochemical function of the protein that the gene encodes:
    at PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi).
    at Google Scholar (http://scholar.google.com).

Annotation of accessions in database

Examine results of BLASTN and BLASTX results for original sequence.
    Evaluate *e*-values and the extensiveness of the hits with subject sequences.
    Examine information in subject sequences' GenBank report.

Run the sequence through Repeatmasker to detect presence of repetitive DNA (http://magi.plantgenomics.iastate.edu/repeatmasker.html).

Use BLASTN to identify a longer maize sequence (*i.e.*, a maize EST contig or predicted gene) at the MAGI or MADI database (http://magi.plantgenomics.iastate.edu/ or http://schnablelab.plantgenomics.iastate.edu:8080/madi/, respectively).

Perform BLASTX on longer sequence (http://www.ncbi.nlm.nih.gov/BLAST/).

Perform InterProScan on longer sequence (http://www.ebi.ac.uk/InterProScan/).
    Gather information on protein functional domains (http://www.ebi.ac.uk/InterProScan/ and http://pfam.janelia.org/).

Run the longer sequence through Repeatmasker to detect the presence of repetitive DNA (http://magi.plantgenomics.iastate.edu/repeatmasker.html).
    If repetitive DNA is present, perform BLASTN of the Cereal Repeats database (http://magi.plantgenomics.iastate.edu/blast/blast_r.html).

If no hits on BLASTX are obtained and sequence is not repetitive, perform search for microRNA (http://microrna.sanger.ac.uk/sequences/search.shtml).

Collect information on the molecular/biochemical function of the protein that the gene encodes:
    PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi)
    Google Scholar (http://scholar.google.com)
    The Arabidopsis Information Resource (http://www.arabidopsis.org)
    Saccharomyces Genome Database (http://www.yeastgenome.org/)

Identify EC number of enzymes.
    Collect information on enzyme function and which biological pathway(s) enzyme is involved (http://www.genome.jp/kegg/pathway.html).

Include GO numbers if possible.
    Check GenBank reports, Gramene, and InterPro. (http://www.gramene.org/plant_ontology/index.html#ontology and http://www.ebi.ac.uk/InterProScan/).

biology and to remain at Truman State University the following summer to perform research full time and to continue to work and make presentations during their senior year.

During the initial training period, it is valuable for the mentor to walk students through one or more complete annotations of an accession or EST. This process serves as a foundation for all of the student's initial development and provides the student with a comfortable set of annotation guidelines (Table 1). In addition, it allows the mentor to reinforce the "big picture" concept to young students, who often lack this perspective. One of the first hurdles for the students to overcome is gaining confidence in differentiating biologically relevant and significant information from irrelevant noise. While it may be comforting to the students to establish a BLAST *e*-value cutoff, it is more important to illustrate the value of multiple lines of evidence to distinguish relevant data. We train our research team to retrieve a longer

sequence [usually a *maize EST contig* (MEC) (http://magi.plantgenomics.iastate.edu/)] and to perform BLASTX, InterProScan, and Repeatmasker searches on it, which often corroborate the conclusions drawn from the initial BLASTX results. It is also very important that students understand that it is not currently possible to annotate all genes using database-mining tools.

For many undergraduate institutions, providing students meaningful exposure to microarray technology may require collaborations with laboratories at research level I institutions. The data that our students annotate were generated by our collaborators at Iowa State University, Cornell University, and Cold Spring Harbor Laboratory. These collaborations provide an opportunity for our students to visit research level I institutions where they are exposed to state-of-the-art instrumentation and interact with the laboratory personnel who generate the SAM genomics data. During these visits, undergraduate students make informal presentations

describing their annotations and analyses. Typically, these trips invigorate the undergraduate researchers, who gain a greater understanding of professional life at a research level I institution and of their own contributions to the collaborative project.

Moving from annotation to analysis and presentation: Students can become so driven to annotate that they view it as a goal unto itself, especially when working with very large genomic data sets. To offset this tendency, the mentor must provide opportunities for the students to analyze the annotated data set with respect to SAM function. Undergraduate students often prefer to become resident experts on ESTs placed into specific functional categories, such as transcription factors or signal transduction components. Encouraging opportunities for group interaction among the students can be synergistic and may provide new insights into genetic interactions during SAM function. Thus, it is valuable to designate specific times for student discussions of the annotated data.

The ultimate goal of scientific research is its presentation to the scientific community. Altogether, our students have made seven scholarly presentations at local, regional, and international professional conferences in the last 2 years, as well as several additional seminar presentations. These have been either individual oral presentations or posters authored and presented by student pairs. Presentations have included an overview of database content, analysis of particular aspects of the microarray hybridization data, as well as some of the "wet-lab" work that we have performed to validate and extend our understanding of tissue-specific expression of select SAM genes. At Truman State University, students may opt to use their research as a senior honors project, which requires the preparation of a thesis as well as an oral presentation to the biology discipline. Other students have received course credit for their research; grades are assigned on the basis of productivity as well as the quality of poster or oral presentation.

THE PRODUCT: THE TRUMAN STATE UNIVERSITY SAM DATABASE

Creating and maintaining the database platform for use by biologists: Making meaningful use of a large data set of differentially regulated ESTs, and adding value to that data set in the form of functional annotations and analyses discussed here, requires a robust data storage and manipulation platform. The phases of the project include receiving and loading the initial set of raw microarray expression data, the presentation of these data to the annotators in useful and understandable formats, automated retrieval of auxiliary and associated information such as BLAST searches, review of the students' annotations by the faculty mentors, real-time generation of statistical analyses of the annotated results, and presentation of the results to the scientific community. The system needs to provide these capabilities securely to authenticated researchers at multiple locations.

Our system, a MySQL relational database-driven system with a web front end programmed in PERL running on a dedicated Linux platform, was developed by the computer science students and faculty on the team working in close collaboration with the biologists. The computer scientists began by carefully observing the manual annotation process. The computer scientists also met frequently with the biologists to develop a deeper understanding of the goals of the research project. The main system was built as a series of prototypes of increasing complexity over a period of ∼8 months. Each prototype was used and tested by the biologists, with their positive and negative feedback incorporated into the subsequent prototype. Through this prototyping refinement process, a coherent system design was developed. The current system comprises ∼6500 lines of PERL code and 22 database tables containing ∼1.5 gigabytes of data. Maintenance of the system is an ongoing process. The design continues to evolve as new genetics experiments are conceived, new sets of data are acquired, and new ways of viewing and analyzing the data are developed.

The Truman State University SAM database: The Truman State University SAM database (http://sam. truman.edu) houses the results of functional annotations of >4400 maize ESTs that have been identified as being differentially regulated in three separate microarray hybridization experiments. These experiments were conducted at Iowa State University (Kazuhiro Ohtsu and Patrick Schnable), the University of Georgia (Xiaolan Zhang), and Cornell University (Michael Scanlon). Specifically, these experiments were aimed at identifying genes that are differentially regulated in (1) the maize SAM compared to the above-ground parts of the whole seedling, (2) the L1 and L2 histological layers of the SAM, and (3) the SAM of the leaf developmental mutant narrow sheath (SCANLON *et al.* 2000) compared to the nonmutant maize SAM. Methods used in tissue preparation, RNA isolation, amplification, and labeling, as well as microarray hybridization for these studies are found at http://maize-meristems.plantgenomics.iastate.edu/ resources/protocols/. Three different SAM microarray gene chips [SAM 1.0 (GPL2557), SAM 2.0 (GPL2572), and SAM 3.0 (GPL3538)] containing >38,000 ESTs were used for the hybridization experiments and were manufactured at Iowa State University (http://www. plantgenomics.iastate.edu/maizechip/).

Users of the SAM database select "GENEVA home" on the home page to enter the database, which displays the "summary" page of the results for SAM chip 1 in the SAM *vs.* whole-seedling experiment (Figure 1, top). It is then possible to select one of the three SAM chips and display the genes on the chip that were found to be differentially regulated for a particular tissue comparison. Thirty accessions are displayed on each summary page, with each line including brief information about

FIGURE 1.—View of a summary page. From the summary page, users of the database can select which SAM chip they want to view, search the database by entering keywords or accession number into the boxes in the search bar, or browse the entries in each experiment by viewing subsequent summary pages.

the accession [GenBank accession number, gene name, functional category, Gene Ontology (GO) molecular function, fold change and *P*-value], as well as links to BLAST results (Figure 1, bottom). Additional information stored on the details page is accessible by clicking on the "details" link (see Figure 1, bottom). We plan to expand the database to include the complete annotations from additional hybridization experiments being conducted in our collaborators' labs.

The process of functionally annotating these differentially regulated ESTs is similar to that carried out in our teaching lab described above, but more extensive and rigorously applied (Table 1). For each significantly up- or downregulated EST, BLASTN and BLASTX searches were performed on NCBI's nonredundant database. Since many of the EST sequences are short and could not be annotated from these initial BLAST searches, a longer, corresponding genomic sequence or EST contig [*i.e.*, a *m*aize *a*ssembled *g*enomic *i*sland (MAGI) or a MEC, respectively (http://magi.plantgenomics.iastate. edu/)] was identified and evaluated with BLASTN,

BLASTX, and InterProScan. When evaluating BLAST results, no *e*-value cutoff was automatically applied; results were evaluated in relation to a series of criteria including EST length, percentage identity and similarity to the subject sequence, and, if possible, the presence of functional domains. In addition, ESTs were routinely evaluated for repetitive sequences using the Repeat-masker function at the MAGI database. If repeat DNA was found in the EST, the accession's sequence was used to query the Cereal Repeats database (http://magi. plantgenomics.iastate.edu/) to identify to what repetitive DNA it was most similar. After each differentially regulated gene was annotated, it was placed into a functional category. Assigning a gene to 1 of 26 functional categories was done following an evaluation of scientific literature identified using PubMed (http://www.ncbi. nlm.nih.gov/entrez/query.fcgi) and Google Scholar, as well as additional information found in databases including the Kyoto Encyclopedia of Genes and Genomes (http://www.genome.jp/kegg/), the European Bioinformatics Institute (http://www.ebi.ac.uk/), the Expert Protein

Analysis System (http://us.expasy.org/), The Arabidopsis Information Resource (http://www.arabidopsis.org/), and others.

Each accession in the Truman SAM database has an individual "details" page, which contains the gene's functional information, as well as fold change and *P*-value information for differential expression (see "details" pages at http://sam.truman.edu). Gene name(s), enzyme commission (EC) number (when appropriate), and GO numbers (when possible) are entered manually by researchers following evaluation of BLASTN, BLASTX, and InterProScan searches. The details page includes links to the GenBank report for the accession, the accession's sequence, as well as BLASTN and BLASTX results. A BLASTN of the maize genomic sequences at the MAGI database is generated when the "MAGI blast" link is selected. The notes section of the details page allows the inclusion of detailed functional information collected at various databases (indicated above), as well as abstracts and other information from journal articles. Links that are entered into this section of the details page allow users of the database direct access to relevant functional information at other locations on the Internet.

Users of the database can readily identify information of interest to them by entering terms into the boxes in the "search bar" at the top of the summary page (Figure 1). The database is searchable by gene name, accession, and keywords in the notes section. In addition, the entries in the summary page can be limited to one of the functional categories or GO molecular functions. Finally, through use of a companion website, the *MicroArray Data Interface* (MADI; http://schnablelab.plantgenomics.iastate.edu:8080/madi/), accessions of interest on the SAM chips and in the Truman SAM database can be identified by performing BLAST searches of the sequences printed on all three SAM chips. These searches identify similar sequences by accession number and spot location on the chips. The accession number can then be entered into the accession search box at the Truman SAM database to access the functional information for that sequence.

CONCLUDING THOUGHTS

Through a multi-university and interdisciplinary collaboration, students at an undergraduate institution have been involved in bioinformatics research. The undergraduate students have made a significant contribution to the functional annotation and analysis of microarray data. They have applied their valuable general cellular and genetic knowledge and insights to the analysis of these data and have developed specific expertise and knowledge comparable to that of a first- or second-year graduate student. Importantly, the students found this research experience to be intellectually challenging and rewarding, as well as a good preparation for their next career step.

It is possible to train and maintain a productive undergraduate research team that can sustain a year-round research program carrying out microarray annotation and analysis. In 2 years we have engaged 19 research students in this project; 10 students have been involved during both the summer research months and the academic year. In addition, since we incorporate this research project into several of our introductory and advanced course laboratory experiences, ~60 students/year are involved in this research project. The value of this experience is evidenced by the fact that all of our research students who have graduated have gone on to postgraduate studies, the majority in Ph.D. or M.S. programs. Several of these students are attending graduate school at our collaborators' institutions while two are now students in one of our collaborator's laboratories.

It is worth noting that the undergraduate research mentors on this project received their graduate and postdoctoral training prior to the advent of genomics and bioinformatics. It is often quite difficult for a faculty member at an undergraduate institution to acquire cutting-edge expertise in their field of study. This is particularly true for a person who is at or beyond "mid-career." Our collaboration with research level I institutions not only has provided us access to state of the art technology, but also has made us valued members of a highly motivated, knowledgeable, and productive collaboration. The faculty mentors' learning experiences have greatly enhanced the quality of our classes, one of the collateral benefits of conducting research at undergraduate institutions described by numerous authors (Bender 2000; Schultz 2001; Malachowski 2004). This research project has provided outstanding opportunities for faculty development, which have had significant educational and scholarly rewards.

LITERATURE CITED

Beck, J., B. Buckner, O. Nikolova and D. Janick-Buckner, 2007   Using interdisciplinary bioinformatics undergraduate research to recruit and retain computer science students. SIGCSE Bull. **39:** 358–361.

Bender, C., 2000   Advance science: mentor an undergraduate. Pharmacologist **42:** 141–145.

Dyer, B. D., and M. D. LeBlanc, 2002   Meeting report: incorporating genomics research into undergraduate curricula. Cell Biol. Educ. **1:** 101–104.

Gillen, C. M., J. Vaughan and B. R. Lye, 2004   An online tutorial for helping nonscience majors read primary research literature in biology. Adv. Physiol. Educ. **28:** 95–99.

Janick-Buckner, D., 1997   Getting undergraduates to critically read and discuss primary literature: an approach used in an advanced cell biology course. J. Coll. Sci. Teach. **27:** 29–32.

Ledbetter, M. L., and A. M. Campbell, 2005   Arguments favoring a survey as the first course of majors. Cell Biol. Educ. **4:** 123–137.

Malachowski, M. R., 2004 The importance of placing students first in designing research programs at predominantly undergraduate institutions. Council Undergrad. Res. Q. **25:** 106–108.

Muench, S. B., 2000 Choosing primary literature in biology to achieve specific educational goals. J. Coll. Sci. Teach. **29:** 255–260.

Pall, M. L., 2000 The value of scientific peer-reviewed literature in a general education science course. American Biology Teacher **62:** 256–258.

Russell, J. S., L. Martin, D. Curtin, S. Penhale and N. A. Trueblood, 2004 Non-science majors gain valuable insight studying clinical trials literature: an evidence-based medicine library assignment. Adv. Physiol. Educ. **28:** 188–194.

Scanlon, M. J., K. D. Chen and C. C. McKnight, 2000 The *narrow sheath* duplicate genes: sectors of dual aneuploidy reveal ancestrally conserved gene functions during maize leaf development. Genetics **155:** 1379–1389.

Schultz, J. R., 2001 The transformational process of mentoring. Council Undergrad. Res. Q. **22:** 72–73.

Smith, G. R., 2001 Guided literature explorations. J. Coll. Sci. Teach. **30:** 465–469.

Zatz, M. M., 2002 Bioinformatics training in the USA. Brief. Bioinformatics **3:** 353–360.

Communicating editor: P. J. Pukkila