

Winter 4-10-2014

Associations between Input and Outcome Variables in an Online High School Bioinformatics Instructional Program

Douglas S. Lownsbery
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds



Part of the [Online and Distance Education Commons](#), and the [Science and Mathematics Education Commons](#)

Let us know how access to this document benefits you.

Recommended Citation

Lownsbery, Douglas S., "Associations between Input and Outcome Variables in an Online High School Bioinformatics Instructional Program" (2014). *Dissertations and Theses*. Paper 1694.
<https://doi.org/10.15760/etd.1694>

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

Associations between Input and Outcome Variables
in an Online High School Bioinformatics Instructional Program

by

Douglas S. Lownsbery

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Science
in
Teaching in General Science

Thesis Committee:
William Becker, Chair
Cary Sneider
Stephanie Wagner

Portland State University
2014

Abstract

Quantitative data from a completed year of an innovative online high school bioinformatics instructional program were analyzed as part of a descriptive research study. The online instructional program provided the opportunity for high school students to develop content understandings of molecular genetics and to use sophisticated bioinformatics tools and methodologies to conduct authentic research. Quantitative data were analyzed to identify potential associations between independent program variables including implementation setting, gender, and student educational backgrounds and dependent variables indicating success in the program including completion rates for analyzing DNA clones and performance gains from pre-to-post assessments of bioinformatics knowledge. Study results indicate that understanding associations between student educational backgrounds and level of success may be useful for structuring collaborative learning groups and enhancing scaffolding and support during the program to promote higher levels of success for participating students.

Acknowledgements

I would like to thank the following individuals for their support of this Master's thesis study. Barbara Buckley, Senior Researcher at WestEd, provided a wonderful opportunity for me to participate in the evaluation activities for the Bioinformatics: Learning by Doing Program. Cary Sneider, Associate Professor in the Center for Science Education at Portland State University, provided valuable support and encouragement throughout my thesis research activities and my entire Master's degree program.

Douglas S. Lownsbery

Bioinformatics: Learning by Doing, NSF DRL 0733255,
awarded to Rutgers, The State University of New Jersey,
Andrew K. Vershon, project director.

Table of Contents

Abstract	i
Acknowledgements	ii
List of Tables	iv
List of Figures	v
Introduction	1
Literature Review	6
Method	11
Implementation Setting and Student Demographics	22
Results	25
Discussion	38
References	44
Appendices	
A. Portland State University HSRRC Memorandum	46
B. Rutgers and PSU Data Use Agreement	47
C. Student Educational Background Survey	49
D. Student Bioinformatics Pretest/Posttest	51

List of Tables

Table 1. Participation by School Setting	22
Table 2. Student Gender by School Setting	23
Table 3. Distribution by Year in School	23
Table 4. Distribution by Prior AP or Honors Biology	24
Table 5. Clone Completion Rates for All Students (n=330)	25
Table 6. Pretest and Posttest Scores for All Students	26
Table 7. Clone Completion Rates by Gender	28
Table 8. Pretest and Posttest Scores by Gender	28
Table 9. Clone Completion Rates by Setting	29
Table 10. Pretest and Posttest Scores by Setting	30
Table 11. Clone Completion Rates by Prior Biology	32
Table 12. Pretest and Posttest Scores by Prior Biology	34
Table 13. Clone Completion Rates by Year in School	35
Table 14. Clone Completion Rates by Setting by Year in School	37

List of Figures

Figure 1. Bioinformatics Program Components and Observations for Data Collection ..	17
Figure 2. Frequency Distribution of Pretest Scores for All Students	26
Figure 3. Frequency Distribution of Posttest Scores for All Students	27

Introduction

The field of bioinformatics has created a revolution in the study of biology by redefining how biological research is carried out and how that research is impacting modern societies. The National Center for Biotechnology Information (NCBI) is a premier center in the United States for biomedical and bioinformatic research, including the databasing, accessing, and analyzing of biotechnology information. According to the NCBI, “Bioinformatics is the field of science in which biology, computer science, and information technology merge to form a single discipline. Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in the biological information generated by the scientific community. This deluge of genomic information has, in turn, led to an absolute requirement for computerized databases to store, organize, and index the data and for specialized tools to view and analyze the data.”

Bioinformatics is significantly impacting many aspects of modern societies, most particularly in the field of medicine, but also in agriculture, environmental science, and forensic science. According to the NCBI, “The rapidly emerging field of bioinformatics promises to lead to advances in understanding basic biological processes and, in turn, advances in the diagnosis, treatment, and prevention of many genetic diseases. Bioinformatics has transformed the discipline of biology from a purely lab-based science to an information science as well. Increasingly, biological studies begin with a scientist conducting vast numbers of database and Web site searches to formulate specific

hypotheses or to design large-scale experiments. The implications behind this change, for both science and medicine, are staggering.”

As part of the bioinformatics revolution, bioinformatics courses and laboratories have been incorporated into many Biology departments at colleges and universities at the undergraduate and graduate levels throughout the United States. In addition to specific courses, many institutions now offer BA and MS degrees in Bioinformatics, or Bioinformatics in combination with Molecular Biology or Computational Biology. High school Biology educators are also becoming increasingly aware of the importance of bioinformatics as a component of course offerings for students who have completed their introductory Biology courses and want to deepen their knowledge and experience in biology.

At the high school level, bioinformatics instruction and laboratory experience is highly aligned with many aspects of the Next Generation Science Standards (NGSS, 2013) including the Scientific Practices of Developing and Using Models and Analyzing and Interpreting Data; the Crosscutting Concepts of Patterns, Cause and Effect, and Structure and Function; and the Life Sciences Disciplinary Core Ideas, particularly LS1A - Structure and Function, LS3A - Inheritance of Traits, LS3B - Variation of Traits, LS4A - Evidence of Common Ancestry, and LS4B - Natural Selection.

In reference to high school instruction, bioinformatics can be viewed as incorporating three main areas of emphasis that build on understandings developed in many introductory Biology courses: a) genomics, or the study of the genes that make up an organism, b) proteomics, or the study of the function, shape, interactions, and

abundance of proteins, and c) systems biology, or the study of the role of DNA and protein interactions on the function of biological systems (Campbell and Heyer, 2003). The completed Human Genome Project (National Institutes of Health, 2003) is one prominent example of a recent scientific milestone that is frequently referenced in many high school biology classrooms. Bioinformatics played a significant role in the Human Genome Project where scientists sequenced the approximately 3 billion base pairs of DNA which make up the approximately 30,000 genes residing in the 23 pairs of chromosomes found in the nucleus of human body cells. Many other current scientific discoveries and technological advancements in medicine, agriculture, and evolutionary studies are frequently discussed in the popular media which provide relevant applications of bioinformatics to the high school curriculum.

However, there are challenges in providing bioinformatics instruction at the high school level. One significant challenge is teachers' level of understanding and expertise in domain-specific bioinformatics knowledge and in the use of specialized bioinformatics procedures and tools that were developed by scientists for scientists. Even when teachers develop the necessary understandings and expertise, the challenge still remains for structuring the learning environment and sequencing the instruction for high school students. This includes providing access to, and ensuring the usability of, the bioinformatics tools, and scaffolding the learning experience to provide sufficient supports for students' conceptual and skill development.

It was to specifically address these challenges that the Bioinformatics: Learning by Doing Program was developed by the Waksman Institute at Rutgers University. The

program was supported, in part, by a grant from the National Science Foundation (NSF) as part of the Discovery Research K-12 Program. As part of the Bioinformatics: Learning by Doing program, a unique, online teaching and learning software application was developed called DSAP (the DNA Sequence Analysis Program). The Bioinformatics: Learning by Doing title reflects the program's underlying philosophy that the best way to learn science is by doing science. As part of the program, high school students not only received scaffolded bioinformatic content instruction, they used that content knowledge, along with procedural knowledge and laboratory and online bioinformatics tools, to conduct authentic genetic research. As the culmination of their participation, each student had the opportunity to conduct authentic genetic research by analyzing a novel DNA sequence and publishing their findings in the GenBank DNA sequence database of the NCBI.

The researcher's prior association with the program had been as a member of the WestEd team serving as the NSF External Evaluator for the Bioinformatics: Learning by Doing Program conducted by the Waksman Institute at Rutgers University. The researcher participated in activities to analyze and report evaluation data for the Year 3, Year 4, and Year 5 program years. By Year 5, significant aspects of the program had been in place for multiple years and could be considered well established (e.g., the version of the DSAP software, the *Wolffia australiana* clone (duckweed) used in DNA analysis, and the protocols for the laboratory modules). At the same time, as would be expected of a complex online teaching and learning software application with distributed users, aspects of the program were also continually evolving (e.g., the sequencing of the

onsite wet-lab activities with the online DSAP activities, and the enhancement of online video tutorials and Help Modules). In addition, during this time, the program experienced significant growth, which may be considered a strong indicator of the usability, feasibility, and effectiveness of the program at the high school level. Amidst this constancy and change, certain trends in the data that were collected and analyzed by both Rutgers and WestEd began to indicate areas of significant program impact, areas that needed improvement, and areas that were, as yet, unexamined. This study addressed some of these previously unexamined areas, with the promise that the results might provide additional insight into how the program has functioned in the past, as well as valuable insight for potential enhancements to the program, and related online bioinformatics programs, in the future.

Literature Review

The following literature review includes research articles and reports that address three related topics: 1) Bioinformatics Instruction in High School; 2) Types of Knowledge in Science Achievement; and 3) Learning in a Web-Based Environment.

Wefer and Sheppard (2008) describe an analysis of high school Biology standards from 49 states and the District of Columbia for content related to bioinformatics. In analyzing the state science standards that were in place prior to the release of the NGSS (2013), the authors evaluated the degree to which bioinformatics content and related issues were incorporated, even though the term *bioinformatics* did not appear in any of the standards documents. The authors identified nine distinct areas of bioinformatics content and related issues: 1) Human Genome Project/genomics; 2) forensics; 3) evolution; 4) biological classification; 5) nucleotide variations; 6) medicine; 7) computer use; 8) agriculture/food technology; and 9) science/technology/society issues. The authors found significant gaps in the degree to which these areas were addressed across the standards. In addition, standards statements related to bioinformatics content and issues were generally ambiguous and over generalized. Based on the standards documents in place at that time, the authors conclude it would take motivation for teachers to detect the relevance of bioinformatics and initiative for teachers to incorporate bioinformatics into their biology lessons.

Wefer and Anderson (2008) conducted a case study of the ways in which 10 high school students processed information and integrated procedural and analytical thought during an instructional unit on bioinformatics that was conducted over 10 class periods.

The bioinformatics unit was embedded in an elective, non-honors Genetics course. All participating students in the Genetics course had completed an introductory high school biology course that covered the appropriate prerequisite content (i.e., DNA structure, replication, DNA isolation, etc.). The bioinformatics instructional unit extensively incorporated the use of computer-based bioinformatics programs including Basic Linear Alignment Search Tool (BLAST), the Protein Data Bank (PDB), and ClustalW.

Quantitative data representing students' understanding of domain specific content and procedural skills were obtained using a quiz and comprehensive end of unit test.

Qualitative data was obtained by semi-structured interviews conducted at the end of the unit and supported by daily student journal entries. The authors conclude that effective bioinformatics instruction requires the integration of procedural knowledge and higher-order knowledge. The authors suggest that teachers should scaffold bioinformatics instruction that integrates procedural knowledge with major conceptual understandings and domain-specific factual knowledge into a coherent schema. Teachers should incorporate resources that diagnose individual student cognitive differences and organize bioinformatics instruction in a way that recognizes those cognitive differences.

Gelbart, Brill, and Yarden (2009), described a study of the impact of a web-based simulation in bioinformatics on high school biology students' understanding of genetics research. The web-based genetics research simulation was embedded at the end of a 30-hour unit on genetics. The genetics content of the simulation was based on students' knowledge received during instruction prior to beginning the simulation. The simulation guided students through five sequential, scaffolded assignments using the methods and

tools of modern geneticists, including bioinformatics, to identify a mutated gene which causes deafness in humans. The researchers claim that student participation in the research simulation promoted the generation of explanations that connect molecular mechanisms and phenotype. Participation in the research simulation also promoted understanding of one of the basic heuristics of genetic research in which a link can be formed between the normal and the mutated forms of a character at the molecular level to facilitate the identification of the gene involved in determining a certain phenotype.

Shavelson, Ruiz-Primo, Li, and Ayala (2003) define science achievement as consisting of four types of knowledge. Declarative knowledge (knowing what) includes scientific definitions and facts, mostly in the form of terms, statements, description, or data. Procedural knowledge (knowing how) includes if-then production rules or sequences of steps that can be carried out to achieve a sub-goal leading to completion of a task. Schematic knowledge (knowing why) includes principles, schemes, and mental models that can be used to interpret problems, troubleshoot systems, explain what happened, and predict changes. Strategic knowledge (knowing when, where, and how to use knowledge) includes domain-specific conditional knowledge and strategies such as planning, problem-solving, and monitoring progress toward a goal.

Weisman (2009) conducted a study during the redesign of a lecture-only, upper-level undergraduate Bioinformatics course into a web-enhanced, collaborative, virtual laboratory. Of the 43 students who completed the course, eight were juniors, 34 were seniors, one was a first-year graduate student, and 37 were Biology majors. All participating students had completed the prerequisite courses which included a two-

semester biology sequence, a one-semester genetics course, general chemistry, and college-level algebra. Traditional on-site teaching laboratories typically used in courses such as cell biology, genetics, and biochemistry often incorporate informal collaborative small groups performing an exercise. Subcomponents of the study were designed to test whether collaborative learning was practical in an online bioinformatics lab and whether student groups could provide peer support for routine questions regarding bioinformatics tools. All virtual laboratory experiments were performed using standard web-based bioinformatics tools. Students were taught to become self-sufficient in using bioinformatic tool documentation but were also encouraged to seek peer support when encountering problems. Qualitative data showed that a large majority of students found that the online collaborative learning environment benefitted their learning. In addition, the data indicated that students routinely approached their peers for help in the virtual bioinformatics lab setting.

Raes, Schellens, De Wever, and Vanderhoven (2011) conducted a large-scale study involving 347 high school students in grades 9 and 10 from 18 different classes in 10 different high schools. The study examined the impact of multiple modes of scaffolding (teacher-enhanced, technology-enhanced, and combined teacher-enhanced and technology-enhanced) on students' domain-specific knowledge and metacognitive awareness during information problem solving (IPS) in a web-based collaborative inquiry project. Students worked collaboratively in pairs during the investigation. The researchers concluded that teacher-enhanced scaffolding had the greatest effect on acquisition of domain-specific knowledge during the IPS project, particularly for students with low

prior knowledge. Technology-enhanced scaffolding was most beneficial for students' metacognitive awareness that facilitated self-regulated learning. The researchers concluded that both teacher-enhanced and technology-enhanced modes of scaffolding are needed to support a diversity of students during web-based inquiry learning.

Gelbart and Yarden (2011) analyzed the same web-based bioinformatics simulation described earlier (Gilbert, Brill, and Yarden, 2009) of high school biology students' understanding of genetics research through a different research lens. The researchers addressed the following question: What kind of support does a teacher provide during enactment of the research simulation and how does it facilitate students' ability to coordinate between declarative and procedural knowledge? According to the researchers, engaging students in authentic research practices in the bioinformatics simulation required students to use conditional knowledge. The authors' definition of conditional knowledge incorporates many similar aspects of Shavelson, Ruiz-Primo, Li, and Ayala's definition of strategic knowledge. The author's premise was that because coordination between declarative and strategic knowledge is not typical in regular school tasks, students were unlikely to carry out such coordination without guidance from the teacher. The researchers' conclude that the teacher role is essential in facilitating students' use of conditional knowledge in conducting authentic research as opposed to simply performing a set of procedures.

Method

This descriptive research study analyzed quantitative data from a completed year of the Bioinformatics: Learning by Doing online high school bioinformatics instructional program. Quantitative data were analyzed to identify associations between the input variables of implementation setting, gender, and students' educational backgrounds and the outcome variables of students' knowledge growth and level of completion of the program. For this study, knowledge growth was determined by pretest to posttest gains, and students' level of completion was determined by completion rates for analyzing the DNA sequences in a series of clones (four practice clones and one or more unknown clones).

Research Question

This study addressed the following research question.

In an online bioinformatics instructional program for high school students, are there associations between the intervention setting, gender, and students' educational backgrounds and students' level of success in the program as indicated by completion rates for clone analysis and performance gains from pretest to posttest assessments?

To address this research question, quantitative data was analyzed from the Year 5 implementation of the Bioinformatics: Learning by Doing program (2011 - 2012 academic year). The Year 5 implementation data was exported directly from the DSAP online instructional and assessment program. The DSAP data included individual student records containing intervention setting data, educational background data, clone analysis

completion rate data, and pretest and posttest performance data. These sets of data were chosen for analysis of associations because they contain the most relevant indicators of the setting, the characteristics that students bring to the learning environment (educational background and prior knowledge) and the most relevant indicators of the outcomes of the intervention (clone completion rates and knowledge gains). The intervention setting, students' educational background characteristics, and students' performance on the pretest were analyzed as independent input variables. Clone completion rates and posttest gains were analyzed as dependent outcome variables. Clone completion rates were the most direct indicator of students' progress and success in the program. Pretest scores indicated students' knowledge prior knowledge and posttest scores indicated learning gains as a result of the intervention. Where applicable, mean pretest to posttest gains were analyzed using paired t-tests to determine the statistical significance of the observed score. Given that clone completion rates were based on ordinal rather than scaled data, only descriptive statistics were used in the analysis of that indicator of success in the program.

Intervention

The Year 5 implementation of the Bioinformatics: Learning by Doing program included a total of 38 high schools and approximately 1,600 students in seven different states and the District of Columbia. The high schools tended to be located in middle-class urban and suburban neighborhoods and included both public and private schools. A majority of the schools were located in New Jersey (28 schools). For the purposes of this study, only those 17 New Jersey schools that participated in the Waksman Summer

Institute at the Rutgers campus in Piscataway, New Jersey, were considered for this study based on the premise that those schools had a relatively high degree of similarity in the training and structure of the intervention. Of these 17 schools, schools with very small class sizes or anomalously low completion rates were excluded, for a total of 13 schools and 330 students included in this study.

The intervention began with a three-week Summer Institute conducted by the Waksman Institute. One teacher (designated a Team Teacher) and two students (designated as Student Scholars) from each high school attended the Summer Institute. The Summer Institute was a three-week period of intensive wet-lab exercises, content lectures, online DSAP bioinformatics exercises, and group meetings. The Summer Institute was led by the Waksman Institute's Project Director and other Rutgers staff who had led each prior year of the Bioinformatics: Learning by Doing program (Year 1 -Year 4). During the three-week Summer Institute, the Team Teachers and Student Scholars received a fast-paced version of the instruction, practice, and research opportunities that other students received during the academic year over a period of one or two semesters. Although the Team Teachers and the Student Scholars received nearly identical lab exercises and lectures during the Summer Institute, the two groups rotated through the structured activities on separate tracks.

The researcher had the opportunity to attend a three-day period of the Summer Institute in July, 2011, and observe this important process that began the Year 5 intervention. During a debriefing session with the Team Teachers conducted by the Rutgers staff, each teacher described the successes and challenges of implementing the

program at his or her school. One comment stood out because it was consistent for all teachers and because it significantly reflected the learning environment. The Team Teachers commented that the Student Scholars functioned in many ways as their peers in providing instruction and support to other students during the academic year.

Another prominent observation at the Summer Institute was that the Student Scholars typically worked together in both formal and informal collaborative learning groups during the wet-lab exercises and the online DSAP bioinformatics exercises. The formal collaborative learning groups consisted of the pair of Student Scholars that attended from each high school. In addition to these predetermined pairings by school, the researcher observed that students continually clustered for brief periods around a student, or student pair, which had successfully completed an activity. Other students were eager to learn the strategies or insights that the successful students had learned or used. These groupings quickly formed and as quickly disbanded as students worked through the wet-lab exercises and used the online bioinformatics tools to complete the steps of DNA analysis.

In addition to the students supporting each other during the activities at the Summer Institute, Rutgers staff were continually present to guide the students to uncover answers to their own questions, and to problem-solve solutions to the problems they encountered. The researcher's observations at the Year 5 Summer Institute helped shape the research questions for this study.

During the academic year, the Bioinformatics Program was conducted in two distinct instructional settings, formal classes and informal school clubs. As would be

expected, the formal class was a scheduled period during the school day and is lead by the Team Teacher, with the Student Scholars serving as teaching assistants. Students received a class grade that is part of their official academic record. The informal clubs were typically held after scheduled classes and were supervised by the Team Teacher, with Student Scholars playing a significant role in providing instruction.

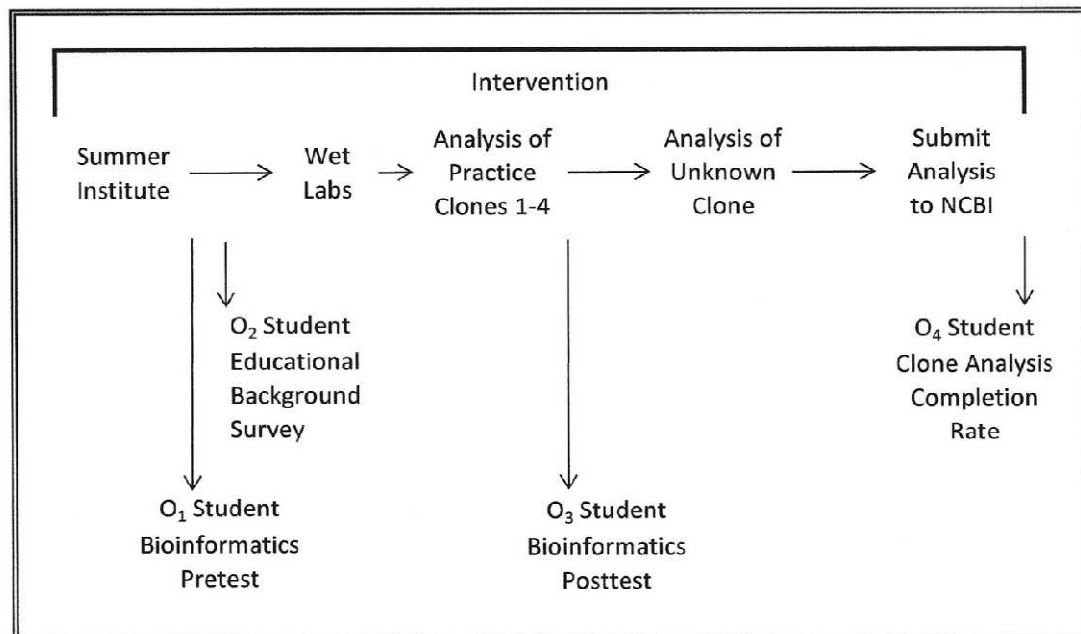
At the beginning of the academic year in both setting, students conducted a series of wet-lab exercises to construct a cDNA library of clones of the studied organism. Students then determined the size of the clones by performing plasmid DNA purification, restriction digests, Polymerase Chain Reaction (PCR), and gel electrophoresis. These wet lab exercises prepared the unknown DNA sequences that the students later analyzed as part of conducting authentic genetics research. Simultaneous to the wet lab activities, students logged into the DSAP online application and begin analyzing a series of practice clones that teach them how to use the bioinformatics tools. Students must complete the four practice clones, designated PC1 - PC4 before being allowed to analyze an unknown clone (designated UC). The practice clones were designed to scaffold students' bioinformatics knowledge and experience through a sequence of steps that progress from simple to more abstract or sophisticated.

The practice clones contained tasks and challenges that students would typically encounter when analyzing unknown clones. In the DSAP application, students were supported by selectable features including Help Modules, a list of Frequently Asked Questions (FAQs), videos, and tutorials. Additional support included online Clone Discussion modules between students, teachers, and Rutgers faculty. Rutgers faculty and

staff also provided online feedback to students' responses to embedded analysis questions associated with the practice clones. After students had completed PC1 - PC4, they could begin analyzing their unknown clone (UC). In the class setting, completion of the practice clones and an unknown clone were potentially course requirements that factored into the course grade, and consequently, affected students' academic records. In the club setting, completion of any level of clone analysis was voluntary and did not affect students' academic records.

The successful analysis of an unknown clone was considered an indication that students could transfer the knowledge they had learned working with the practice clones to the analysis of novel data during a process of authentic bioinformatics research. A student's analysis of an unknown clone was reviewed by the Rutgers faculty and returned to the student with feedback if correction was needed. Upon successful completion, the student's authentic genetic research resulting in the analysis of a previously unknown DNA sequence was published in GenBank, the DNA sequence database maintained by the NCBI, in the name of the student, the student's teacher, and the Rutgers researcher.

Figure 1 below shows the major components of the Bioinformatics: Learning by Doing Program as they relate to the sequence of data collection instruments used during



the program. The intervention begins with the Summer Institute for Team Teachers and a selected subset of Student Scholars. After the Summer Institute, components O₁ through O₄ apply to all students. For the first observation, students were administered the online Bioinformatics Knowledge Pretest (O₁) immediately after logging into DSAP. The second observation consisted of students' educational background characteristics collected during an online Survey (O₂) immediately after completing the pretest.

Figure 1. Bioinformatics Program Components and Observations for Data Collection

Legend

- O₁ Student Bioinformatics Knowledge Pretest (at the beginning of DSAP login)
- O₂ Student Educational Background Survey (immediately following the Pretest)
- O₃ Student Bioinformatics Knowledge Posttest (identical to Pretest)
- O₄ Student Clone Analysis Completion Rate (Practice Clones 1-4 and Unknown Clone)

Students then proceeded through the wet lab activities and the same sequence of analyzing four practice clones (PC1 - PC4). After students completed analysis of the four

practice clones, students were administered the online Bioinformatics Knowledge Posttest (O_3), which was identical to the pretest. Students then had the opportunity to complete the analysis of an unknown clone. Following analysis of an unknown clone (or the highest level of practice clone completion), students' Clone Analysis Completion data was gathered (O_4). Data for observations O_1 through O_4 were exported directly from the DSAP software application for analysis by the researcher.

Instruments and Observations

The Student Bioinformatics Knowledge Pretest and Posttest (O_1 and O_3) analyzed for this study contained fourteen 1-point multiple-choice items and one 4-point constructed-response item, for a total of 18 possible points. The items were developed by WestEd content experts and vetted by the Rutgers staff. Following the administration of the pretest and posttest, those items that did not match goodness-of-fit item parameters established by WestEd were eliminated from item analysis and were not included in this study. The researcher classified the individual items with reference to the Science Knowledge Framework developed by Shavelson, Ruiz-Primo, Li, and Ayala (2003). The Science Knowledge Framework was part of the initial Bioinformatics program proposal for developing assessments that addressed the types of knowledge required for using bioinformatics tools to analyze the practice clones and to conduct authentic research by analyzing an unknown clone. As an indication of its broader usage, the Science Knowledge Framework was also used as the knowledge framework for the 2009 and 2011 NAEP Science Assessments. The Science Knowledge Framework includes four types of knowledge: Declarative Knowledge, Procedural Knowledge, Schematic

Knowledge, and Strategic Knowledge. The majority of the pretest/posttest items aligned with either Declarative Knowledge (6 points) or Strategic Knowledge (9 points).

The pretest was administered online immediately after students logged into DSAP for the first time. An identical posttest was administered online typically after students had completed all four practice clones (PC1 - PC4) in the DSAP instructional sequence, but before students began analyzing an unknown clone.

The constructed-response item was a complex 4-point prompt where students were required to describe and explain their answer. For hand scoring, an initial set of 50 student responses reflecting an intended range of scores were selected and independently scored by the researcher and two other project evaluators using a detailed scoring rubric. Discrepant scores were discussed and any agreed upon decision rules were documented. For the second round of scoring, each scorer independently reviewed his or her initial scores and made any changes resulting from the team discussion and decision rules. Rater agreement after the second round of scoring was greater than 0.90. For any responses with less than total agreement after the second round, a consensus score was established. Based on the scoring discussions, decision rules, and consensus scores from round one and round two of scoring the 50 responses, the researcher independently scored the remaining student responses.

One potential weakness of the DSAP pretest and posttest data is that there were little or no external incentives for the students associated with performance on the tests. The pretest and posttest data were gathered primarily for use in program evaluation and students were aware that the results would not pertain to them directly. As such,

performance on the pretest and posttest may not reflect the same level of effort as performance if there had been stronger incentive for the students to do their best. Even with this weakness, the data is still considered relevant to this study for comparing gains between groups of students with different characteristics. Additionally, the tests did not adequately sample all four type of knowledge in the Science Knowledge Framework developed by Shavelson, Ruiz-Primo, Li, and Ayala (2003).

The Student Educational Background Survey (O₂) consisted of eight questions addressing the demographics and career interests of students. The Survey was administered online to all participants immediately after they completed the pretest. Student responses to the questions about Gender and Ethnicity were optional. All information in the Survey was self-reported by the students, and no separate verification of the reported data was performed. Survey data that was analyzed for this study included current year in high school and prior completion of Honors or AP[®] Biology.

In addition to the categories of educational background, the study also examined potential associations relative to gender and instructional setting (class or club). The gender demographic characteristic was also collected during the online Survey. The instructional setting was determined by the school the student attended.

Data for the Student Clone Analysis Completion Rate (O₄) were tracked internally by the DSAP administrative module. The clone completion rate data served as the most important measure of student progress and success in the program. The data identified the highest level of completion of practice clones for each student (PC1 - PC4) and whether or not the student completed analysis of one or more unknown clones. The analysis of an

unknown clone represented the highest level of student achievement for all students participating in the program. The analysis of an unknown clone allowed high school students to engage in the scientific enterprise by conducting authentic bioinformatic research and publishing the DNA sequences in public databases through the NCBI. Unlike the four practice clones, which were identical for all students, the unknown clones varied in difficulty across all students based on the unique DNA sequence that each student had obtained or were provided for analysis.

Implementation Setting and Student Demographics

The first level of analysis of the program begins by looking at the distributions of participants according to the independent variables of the study (program implementation, gender demographic, and student educational backgrounds). A total of 330 students began the Bioinformatics: Learning by Doing program in the 13 schools included in the study in the 2011-2012 academic year. As indicated earlier, the Bioinformatics program was implemented as either a formal class or informal club at each school—with one exception. At one school, one group of students participated as a formal class, and a different group participated as an informal club. Table 1 shows the number of schools and students participating by setting.

Table 1

Participation by School Setting

Setting	Schools	Students	Size			
			Minimum	Maximum	Median	Mean
Class	9	193	6	52	17	21
Club	5	137	17	64	19	27
Total	13*	330				

Note. One school participated with both a class and a club.

As shown in Table 2, across the 13 schools, the percentage of male and female students was identical (50%), with slightly more males than females in class settings (51% to 49%), but fewer males than females (47% to 53%) in club settings.

Table 2

Student Gender by School Setting

Setting	Male		Female		Total	
	n	%	n	%	n	%
Class	99	51	94	49	193	58
Club	65	47	72	53	137	42
Total	164	50	166	50	330	100

Table 3 and Table 4 show the distribution of students according to the student educational background variables. Across the schools, the program included students from all high school grades. Sixty-nine percent of the students who began the program were Freshman or Sophomores, and 31% were Juniors or Seniors. Thirty-nine percent of all students had not completed a prior AP Biology or Honors Biology course.

Table 3

Distribution by Year in School

Year in school	n	%
Freshman	101	31
Sophomore	127	38
Junior	66	20
Senior	36	11
Total	330	100

Table 4

Distribution by Prior AP or Honors Biology

Status	Completed prior AP or Honors Biology	
	n	%
No	129	39
Yes	198	61
Total	327	100

Note. Three student records are missing.

Results

The results of the program, including the student clone completion rates and the pretest to posttest bioinformatics knowledge gains, were analyzed in reference to the independent variables (program implementation, student gender, and student educational backgrounds) to determine potential associations. The results of the analyses are presented in sequence from the most general level (all students) to increasingly specific subsets of students. The two indicators of student success in the program (clone completion rates and bioinformatics knowledge gains) were first examined for all students, and subsequently in reference to: (a) the demographic variable of gender; (b) the implementation variable of setting; and (c) the student educational background variables of Completion of Prior Honors or AP Biology, and Year in School.

Table 5 shows the percentages of the highest level of clone completion for all students who began the program. Of the 330 students who initially logged into DSAP and completed the pretest and educational background survey, 19% did not complete any of the practice clones, 12% completed through just practice clone PC4, and 58% completed all four practice clones and analyzed at least one unknown clone.

Table 5

Clone Completion Rates for All Students (n=330)

0	PC1	PC2	PC3	PC4	Unknown clone
19%	5%	2%	4%	12%	58%

Note. 0 represents no completion of any clones. PC1, PC2, PC3, PC4, and Unknown Clone represent completion rates at each level of clone analysis.

Table 6 shows the mean pretest scores for all 330 students and the mean posttest scores for the 211 students (64%) who completed the program to that level. As indicated earlier, the posttest was typically administered after students completed PC1 - PC4, and before students were allowed to analyze an unknown clone.

Table 6

Pretest and Posttest Scores for All Students

Test	Number of students	Mean	SD
Pretest	330	9.30	4.12
Posttest	211	12.25	3.70

Note. Mean out of 18 maximum points possible

Figure 2 and Figure 3 show the distribution of the pretest and the posttest scores for all students.

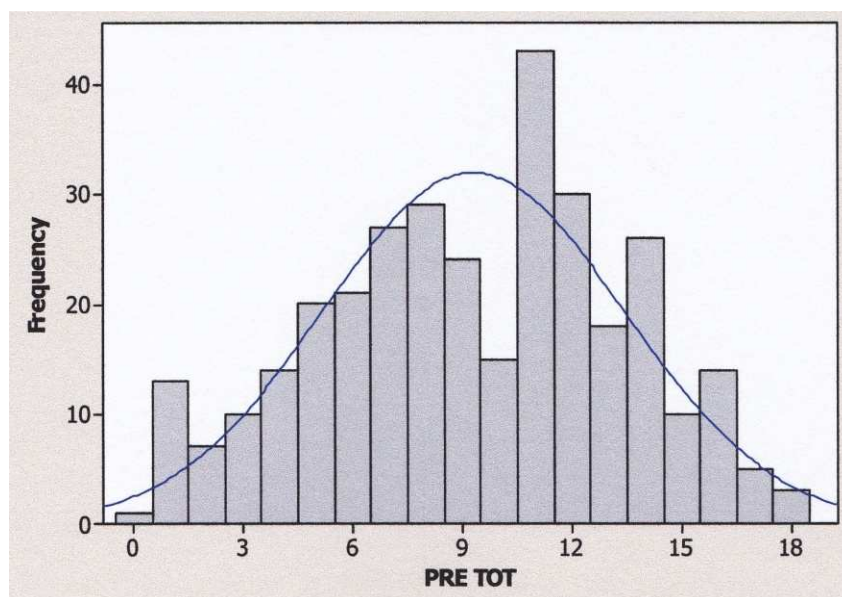


Figure 2. Frequency Distribution of Pretest Scores for All Students

Note. PRE TOT = Pretest Total Scores

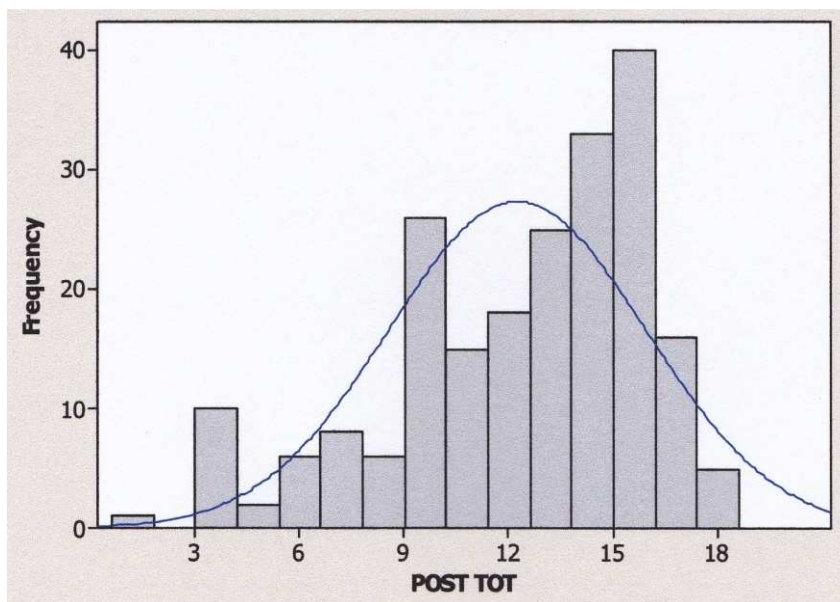


Figure 3. Frequency Distribution of Posttest Scores for All Students
Note. POST TOT = Posttest Total Scores

The second level of analysis of results looked at program success relative to the intrinsic demographic variable of gender. Table 7 shows the percentages of the highest level of clone completion for students by gender. The percentages of males and females who completed no practice clones were similar (21% for males, and 18% for females) as were the percentages who completed one or more unknown clone (57% for males, and 58% for females). Table 8 shows the pretest and posttest scores by gender with a maximum score of 18. Pretest and posttest means and mean gains ($p = 0.655$) were highly similar for males and females. Overall, the two primary indicators of success in the program, clone completion rates and knowledge gains, demonstrate minimal difference for the demographic variable of gender.

Table 7

Clone Completion Rates by Gender

Gender	0	PC1	PC2	PC3	PC4	Unknown clone
Male (n=164)	21%	5%	2%	3%	12%	57%
Female (n=166)	18%	6%	1%	5%	11%	58%

Note. 0 represents no completion of any clones. PC1, PC2, PC3, PC4, and Unknown Clone represent completion rates at each level of clone analysis.

Table 8

Pretest and Posttest Scores by Gender

Gender	Pretest mean all students	Pretest mean of posttest students	Posttest mean	Mean gain ($p = 0.655$)	Mean gain <i>SD</i>
Male					
(Pretest n=164)	9.25	10.29	12.13	1.84	4.05
(Posttest n=103)					
Female					
(Pretest n=166)	9.36	10.30	12.36	2.07	3.45
(Posttest n=108)					

Note. Mean out of 18 maximum points possible

An independent variable of the program that is extrinsic to the student is the implementation variable of school setting—either formal class or informal club. Table 9 shows the percentages of the highest level of clone completion by class or club setting.

Unlike the gender variable, where there were minimal differences between males and females for clone completion rates, there are large differences for the setting variable.

Table 9

Clone Completion Rates by Setting

Setting	0	PC1	PC2	PC3	PC4	Unknown clone
Class	5%	0%	0%	3%	13%	79%
(n = 193)	(10)	(0)	(0)	(5)	(25)	(153)
Club	39%	13%	4%	7%	9%	28%
(n = 137)	(54)	(18)	(5)	(9)	(13)	(38)

Note. 0 represents no completion of any clones. PC1, PC2, PC3, PC4, and Unknown Clone represent completion rates at each level of clone analysis.

As shown in Table 9, only 5% of the students in classes did not complete any practice clones compared with 39% of the students in clubs. Conversely, 79% of the students in classes completed all four practice clones and at least one unknown clone compared to only 28% of the students in clubs. There was minimal attrition from students in classes as they completed PC1 through PC3 (3% total) in contrast to students in clubs, where nearly one-fourth of the students (24%) left the program after completing PC1, PC2, or PC3.

The other outcome variable in the program, bioinformatics knowledge gain, was also examined for setting, class or club. The results in Table 10 comparing pretest and posttest scores by setting also reflects the high attrition rate of students in clubs that was evident in the clone completion rates in Table 9. Of the 137 students in clubs who started

the program and took the pretest, only 44 (32%) took the posttest. In contrast, of the 193 students in classes who took the pretest, 167 students (86%) took the posttest. The pretest mean of all students in classes was much higher than the pretest mean of all students in clubs (10.10 to 8.18, or an 11% difference on an 18-point test). However, the pretest means for only those students who completed both the pretest and the posttest (column Pretest Mean of Posttest Students), were highly similar for classes and clubs (10.34 to 10.11). Additionally, the posttest means and mean gains were highly similar for classes and clubs ($p = 0.783$).

Table 10

Pretest and Posttest Scores by Setting

Setting	Pretest mean all students	Pretest mean of posttest students	Posttest mean	Mean gain ($p = 0.783$)	Mean gain <i>SD</i>
Class					
(Pretest n = 193)	10.10	10.34	12.26	1.92	3.46
(Posttest n = 167)					
Club					
(Pretest n = 137)	8.18	10.11	12.21	2.09	4.72
(Posttest n = 44)					

Note. Mean out of 18 maximum points possible

Based on Table 10 showing the pretest and posttest scores by setting, the mean bioinformatics knowledge level at the beginning of the program for approximately one-third of the students in clubs (44 of 137 students, or 32% with pretest mean of 10.11) was highly similar to the pretest mean for all students in classes (10.10). Additionally, the posttest means for students in clubs and classes were highly similar (12.21 for clubs and 12.26 for classes). The implication of these results is that students who completed all four practice clones (PC1-PC4) demonstrated similar mean knowledge gains irrespective of setting. Based on Table 9 showing the clone completion rates by setting, although over one-third of the students in clubs (39%) quit the program soon after starting, 28% did reach the highest level of the program and complete one or more practice clones, and fully 79% of the students in classes analyzed one or more unknown clones. The implication of these results is that irrespective of setting, students were able to use the DSAP online software in the Bioinformatics: Learning by Doing program to conduct authentic bioinformatics research. The difference in attrition rates between classes and clubs may relate to structural factors that operated in the different settings such as participation criteria, attendance criteria, grading criteria, the role of collaborative student groups, etc., but these factors were beyond the scope of this study. The difference in attrition rates between settings may also relate to the educational backgrounds of the students who participated, i.e., what characteristics did the students bring to the program. The next section of this report examines potential associations between the independent student educational background variables and success in the program.

Table 11 compares the clone completion rates for all students based on the student educational background variable of whether or not the student had previously completed an Honors Biology or AP Biology course. As shown in Table 11, an interesting trend appears in the clone completion rates. Compared to students who had completed a prior course, a greater percentage of students who had not completed a prior course finished all the practice clones PC1 - PC 4 (73% to 67%), but a smaller percentage actually analyzed an unknown clone (55% to 60%). The most notable distinction was between the percentage that successfully finished all the practice clones but did not go on to analyze an unknown clone (18% difference for students without the prior courses, compared to 7% for students with the prior courses).

Table 11

Clone Completion Rates by Prior Biology

Prior Honors or AP Biology	Took pretest	Did not finish PC1-PC4	Finished PC1-PC4	Analyzed UC	Difference finished PC1-PC4 and analyzed UC
No (n = 129)	100%	27%	73%	55%	18%
Yes (n = 198)	100%	33%	67%	60%	7%

Note. Three missing records out of total of 330 participants. PC1- PC4 represent the four practice clones, and UC represents one or more unknown clones.

The implication of these results is that students who both have, and have not, completed an Honors or AP Biology course were successful in the program, but that there may be an advantage in completing what may be the most difficult component of the program, analyzing an unknown clone, for students who have completed the prior courses.

Table 12 shows the pretest and posttest scores for all students based on the student educational background variable of whether or not the student had previously completed an Honors or AP Biology course. As shown in Table 12, the same percentage of students in each group took the posttest (64%). Unlike the pretest to posttest comparisons by gender and by setting, there is suggestive, but inconclusive, evidence of a statistically significant difference ($p = 0.054$) in the mean gain between students who had, and had not, completed a prior Honors or AP Biology course. Differences in mean gains between groups (prior and no prior) on the subsets of items measuring Declarative Knowledge and Strategic Knowledge were also analyzed but did not suggest statistically significant differences ($p = 0.133$ for Declarative Knowledge, and $p = 0.472$ for Strategic Knowledge). Considered together, the results in Table 11 and Table 12 indicate that students who had completed a prior Honors or AP Biology course had greater overall success in the Bioinformatics: Learning by Doing program.

Table 12

Pretest and Posttest Scores by Prior Biology

Prior	Pretest		Posttest		Mean gain	Mean gain
Honors or AP Biology	Pretest mean	Took posttest	mean of posttest students	mean	($p = 0.054$)	<i>SD</i>
No (n = 129)	7.80	64%	8.62	11.17	2.55	3.95
Yes (n = 198)	10.35	64%	11.43	12.97	1.54	3.57

Note. Three missing records out of total of 330 participants. Mean out of 18 maximum points possible.

A second student educational background variable, the year in school in which the student participated in the program (freshman, sophomore, junior, or senior), was also analyzed relative to clone completion rates and mean knowledge gains. Table 13 compares the clone completion rates for all students based on the year in high school. Completion rates by year in high school showed marked differences between freshman and sophomores as a group compared to juniors and seniors as a group. Although completion rates for the practice clones PC1 - PC4 were relatively high for both groups, the differences are large between the two groups for analysis of one or more unknown clones. Approximately half (51%) of the freshman and sophomore group analyzed an unknown clone compared to nearly three-quarters (73%) of the junior and senior group.

Table 13

Clone Completion Rates by Year in School

Year in school	Took pretest	Did not finish PC1-PC4	Finished PC1-PC4	Analyzed UC	Difference finished PC1-PC4 and analyzed UC
Freshman/Sophomore (n = 228)	100%	14%	86%	51%	35%
Junior/Senior (n = 102)	100%	5%	95%	73%	22%

Note. PC1- PC4 represent the four practice clones, and UC represents one or more unknown clones.

Although there were large differences between the year in school groups for clone completion rates, separate analysis of the differences between years in school for pretest to posttest mean gains were not statistically significant ($p = 0.607$). Considered together, these results indicate that all grade levels of students in high school can be successful in the program, but that there may be an advantage for juniors and seniors compared to freshman and sophomores in completing what may be the most difficult component of the program, analyzing an unknown clone.

Data for a third student educational background variable, the number of years the student had participated in the Bioinformatics program, was also collected via the student questionnaire. The distribution of students across the possible responses (First, Second, Third, Fourth) to the questionnaire prompt suggested that students interpreted prior participation differently. For example, some students may have considered dropping out

of the program during a prior year as not participating because they did not finish some specific stage such as the wet labs or practice clones. Other students may have considered starting, but not completing any stage of the program, as participating. As a result, data for this background variable were not analyzed.

The last level of analysis of results in the study looked at two independent variables together relative to the outcome variable of clone completion rates. As shown earlier in Table 9, there were large differences in clone completion rates by setting, with only 28% of students in clubs completing one or more unknown clones in contrast with 79% in classes. As shown above in Table 13, there were also large differences in clone completion rates by year in school, with only 51% of the freshman and sophomore group completing one or more unknown clones in contrast with 73% of the junior and senior group.

Table 14 shows the clone completion rates by combinations of these groupings, the freshman and sophomore group by class and club, and the junior and senior group by class and club.

Table 14

Clone Completion Rates by Setting by Year in School

Year in school and setting	Took pretest	Did not finish PC1-PC4	Finished PC1-PC4	Analyzed UC	Difference finished PC1-PC4 and analyzed UC
Freshman/Sophomore					
Class (n = 126)	100%	9%	91%	75%	16%
Club (n = 102)	100%	66%	34%	23%	11%
Junior/Senior					
Class (n = 67)	100%	6%	94%	88%	6%
Club (n = 35)	100%	54%	46%	43%	3%

As shown in Table 14, the differences between groups by year in school were most extreme for clubs. In clubs, the percentage of juniors and seniors who completed one or more unknown clone (43%) was nearly twice the percentage for freshman and sophomores (23%). The difference was also evident for classes, where the percentage of juniors and seniors was 13 percentage points higher than the percentage for freshman and sophomores (88% to 75%). These results indicate that there was an advantage in both settings for juniors and seniors in analyzing an unknown clone, but that the advantage was much greater in clubs than in classes.

Discussion

This study contributes to the existing body of literature that addresses bioinformatics instruction at the high school level by analyzing selected program input variables and selected student outcome variables to identify possible associations that promote success in the online Bioinformatics: Learning by Doing program. In the online Bioinformatics program, students at all high school grade levels, and in both the traditional formal class setting and the informal club setting not currently represented in the literature, were successful at using bioinformatics tools to analyze a sequence of pre-determined practice clones and to analyze one or more unknown clones.

The Bioinformatics program incorporated multiple modes of scaffolding in a web-based information problem solving project as did the study conducted by Raes, Schellens, De Wever, and Vanderhoven (2011). The results of this Bioinformatics study support the conclusion by Raes, Schellens, De Wever, and Vanderhoven (2011) that both teacher-enhanced and technology-enhanced modes of scaffolding are needed to support a diversity of learners in a web-based inquiry learning program. Scaffolding in the Bioinformatics program included the technology-based analysis of the practice clones from simple to more sophisticated levels, and expert feedback from the Rutgers staff to the students' responses to the embedded analysis questions associated with each practice clone. An additional support feature of the Bioinformatics program was the online Clone Discussion modules that allowed for extensive, online remote communication and collaboration between students, Student Scholars, teachers, and Rutgers staff.

A particular aspect of high school bioinformatics instruction highlighted by the Bioinformatics program is the role of the Student Scholars. The two Student Scholars at

each school provided support as student-teachers in both the classes and clubs, but particularly in the informal club setting. Although clone completion rates were much higher for classes than for clubs, 28% of the students in clubs did complete the practice clone sequence and analyze one or more unknown clones. Given the less structured open-entry, open-exit, and ungraded nature of the club setting compared to the class setting, it may be that the role performed by the Student Scholars made a significant contribution to the observed level of student success.

For students in both classes and clubs, increased clone completion rates and bioinformatics knowledge gains were associated with prior completion of either an Honors Biology or AP Biology course. Additionally, increased clone completion rates were associated with grade level, or year in school. As a group, students who were juniors or seniors had much higher rates for analyzing unknown clones than students who were freshman or sophomores.

In comparing the Bioinformatics program with programs reported in the current literature, the Bioinformatics program appears to be unique in including participants at all high school grade levels who may not have completed a set of prerequisite courses. For example, in the case study of high school students conducted by Wefer and Anderson (2008), the students had previously completed an introductory high school biology course that covered the appropriate prerequisite content. In the Gelbart, Brill, and Yarden study (2009) of a web-based bioinformatics simulation embedded at the end of a 30-hour high school unit on genetics, students had completed the earlier portions of the unit covering the prerequisite genetics content. In contrast, for the schools with clubs in the

Bioinformatics program (5 of the 13 schools in the study), participation was open to all students who signed up at the beginning of the school year.

Through the unique club setting, the online Bioinformatics program provided students with access to bioinformatics instruction that might not have otherwise be available to them at their schools. By extending participation to freshman and sophomores, the Bioinformatics program provided access to a self-selected group of students interested in this rapidly emerging field of study impacting many aspects of modern society. If some of these students subsequently repeated the program as juniors or seniors, they may have developed the capacity to serve as student-teachers or to provide peer-to-peer support to other students in collaborative student learning groups as demonstrated in the Weisman study (2009) where students routinely approached their peers for help in a virtual bioinformatics laboratory setting.

Limitations of the Current Research

The data used in this study were originally collected for the purpose of general program evaluation, and therefore, the use of the data in this study is limited to those program input and outcome variables applicable to all students. For example, the data collected for general program evaluation supports comparing clone completing rates by setting by year in school. However, the data does not support additional level of analyses such as the effect of different forms of implementation of the Bioinformatics program (wet labs, lectures, online labs, etc.) within the class or club settings at different schools, since that data was not collected. Additionally, although clone completion rates can be compared, the data does not provide any information about why students did, or did not,

complete certain levels of clone analysis or information about what problems students may have encountered that influenced clone completion.

The data originally collected for program evaluation included approximately 1,600 participating students in 38 high schools in seven different states and the District of Columbia in Year 5 of the program. The data used in this study was limited to 13 schools that participated in the Waksman Summer Institute at the Rutgers campus in Piscataway, New Jersey at the beginning of the program year. Although the study data may be considered inclusive of virtually the entire subset of participants in New Jersey, the degree to which the data from the New Jersey participants is representative of the entire participating population was not examined.

Although this study used student performance on the Bioinformatics Knowledge Pretest and Posttest as an indicator of learning gains, the identical pretest/posttest had limitations in both test design and test administration. In terms of test design, the test did not adequately sample all four knowledge types in the Science Knowledge Framework developed by Shavelson, Ruiz-Primo, Li, and Ayala (2003). The test primarily addressed either Declarative Knowledge or Strategic Knowledge. As a result, the pretest to posttest comparisons provided limited information relative to one of the original goals of the Bioinformatics program — to develop student understanding of important bioinformatics and genetics concepts and science practices. Because of the limitations in test design, one of the issues highlighted in the Gelbart and Yarden study (2011) regarding students' incorporation of Strategic Knowledge in conducting authentic bioinformatics research as

opposed to simply performing tasks based on Procedural Knowledge could not be examined in this study.

In terms of test administration, students had little incentive to demonstrate their best performance on either the pretest or posttest since they were aware that the results would not pertain to them directly. Although useful for making some levels of comparison, anecdotal data from the teachers and Rutgers staff, as well as some of the open-response item data, support the limitation of the data in making strong conclusions about the effect of the Bioinformatics program on students' knowledge.

Additional Research

The results and limitations of this study suggest several avenues for additional research in future iterations of the program or in other high school bioinformatics programs. One avenue is to examine more closely students' educational backgrounds, and specifically, students' prior knowledge at the beginning of the program in comparison to indicators of success at the end of the program, to determine if prerequisites are warranted for participation in either the class or the club setting to more fully meet the program's goals. This analysis might also provide insight into strategies that teachers can use to structure collaborative student learning groups that foster higher levels of success for all participating students.

Another avenue for research is to examine the role and impact of "student scholars" who receive specialized advance training, and the role of students who are repeating the program for a second or third year, on the level of success of other students in the program. This analysis may provide insight into what problems in the program

sequence can typically be addressed by teacher (including student-teacher) or peer-to-peer support, and what problems typically require expert-level support, such as from the Rutgers staff.

A final suggested avenue for research relates specifically to the club setting. The club setting provides a unique opportunity, based on the current literature, for high school students to participate in bioinformatics learning and research when a school does not have a formal class incorporating the field of study. Analysis of the features of clubs with relatively high clone completion rates, as well as the characteristics of the students within those clubs who persist through all stages of the program, may provide insight into strategies that all clubs could employ to increase overall student success.

References

Alexander, P.A., & Judy, E.J. (1988). The interaction of domain-specific and strategic knowledge in academic performance. *Review of Educational Research*, 58(4), 375-404.

Campbell, A.M., and Heyer, L. J. (2003). *Discovering genomics, proteomics, and bioinformatics*. San Francisco: Benjamin Cummings.

Cassidy, S. (2004). Learning styles: an overview of theories, models, and measures. *Educational Psychology*, 24(4), 419-444.

Gelbart, H., Brill, G., & Yarden, A. (2009). The impact of a web-based research simulation in bioinformatics on students' understanding of genetics. *Research in Science Education*, 39(5), 725-751.

Gelbart, H., & Yarden, A. (2011). Supporting learning of high-school genetics using authentic research practices: the teacher's role. *Journal of Biological Education*, 45(3), 129-135.

Quellmalz, E. S., & Kozma, R. (2003). Designing assessment of learning with technology. *Assessment in Education: Principles, Policy & Practice*, 10(3), 389-408.

Raes, A., Schellens, T., De Wever, B., & Vanderhoven, E. (2012). Scaffolding information problem solving in web-based collaborative inquiry learning. *Computers & Education*, 59(1), 82-94.

Shavelson, R., Ruiz-Primo, M., Li, M., & Ayala, C. (2003). *Evaluating new approaches to assessing learning*. Los Angeles: National Center for Research on

Evaluation, Standards, and Student Testing (CRESST), Center for the Study of Evaluation (CSE), University of California.

National Research Council. (2011). *A framework for k-12 science education: practices, crosscutting concepts, and core ideas*. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: for states, by states*. Washington, DC: National Academies Press.

Wang, K. H., Wang, T. H., Wang, W. L., & Huang, S. C. (2006). Learning styles and formative assessment strategy: enhancing student achievement in web-based learning. *Journal of Computer Assisted Learning*, 22(3), 207-217.

Wefer, S. H., & Sheppard, K. (2008). Bioinformatics in high school biology curricula: a study of state science standards. *CBE - Life Sciences Education*, 7(1), 155-162.

Wefer, S. H., & Anderson, O. (2008). Identification of students' content mastery and cognitive and affective percepts of a bioinformatics miniunit: a case study with recommendations for teacher education. *Journal of Science Teacher Education*, 19(4), 355-373.

Weisman, D. (2010). Incorporating a collaborative web-based virtual laboratory in an undergraduate bioinformatics course. *Biochemistry and Molecular Biology Education*, 38(1), 4-9.

Appendix A. Portland State University HSRRC Memorandum

**Human Subjects Research Review Committee
Research and Strategic Partnerships**

Post Office Box 751 (RSP) 503-725-3423 tel
Portland, Oregon 97207-0751 503-725-8170 fax
hsrrc@pdx.edu



Portland State University HSRRC Memorandum

To: Cary Sneider/Douglas Lownsbery

From: Todd Bodner, Chair, HSRRC 2013

Date: May 24, 2013

Re: Your HSRRC application titled, "Using Multivariate Correlational Analysis to Enhance Online High School Bioinformatics Instruction" (HSRRC Proposal #132499)

In accordance with your request, the Human Subjects Research Review Committee has reviewed your proposal for compliance with DHHS policies and regulations covering the protection of human subjects. The committee is satisfied that your provisions for protecting the rights and welfare of all subjects participating in the research are adequate, and your project is approved.

Please note the following requirements:

Changes to Protocol: Any changes in the proposed study, whether to procedures, survey instruments, consent forms or cover letters, must be outlined and submitted to the Chair of the HSRRC immediately. The proposed changes cannot be implemented before they have been reviewed and approved by the Committee.

Continuing Review: *This approval will expire May 24, 2014, one year from the approval date.* It is the investigator's responsibility to ensure that a *Continuing Review Report* (available in RSP) of the status of the project is submitted to the HSRRC approximately two months before the expiration date, and that approval of the study is kept current.

Adverse Reactions: If any adverse reactions occur as a result of this study, you are required to notify the Chair of the HSRRC immediately. If the problem is serious, approval may be withdrawn pending an investigation by the Committee.

Completion of Study: Please notify the Chair of the Human Subjects Research Review Committee (campus mail code RSP) as soon as your research has been completed. Study records, including protocols and signed consent forms for each participant, must be kept by the investigator in a secure location for three years following completion of the study.

If you have questions or concerns, please contact the HSRRC at hsrrc@pdx.edu or (503)725-2227.

Appendix B. Rutgers and PSU Data Use Agreement

RUTGERS, THE STATE UNIVERSITY OF NEW JERSEY Data Use Agreement

This **Data Use Agreement** ("DUA") is effective on the 24th day of July, 2013, ("Effective Date") by and between Rutgers, The State University of New Jersey on behalf of Andrew Vershon, Director Waksman Student Scholars Program ("Rutgers"), and Portland State University on behalf of Douglas Lownsbery ("Recipient"), located at Portland State University, Research and Strategic Partnerships, Post Office Box 751 (RSP), Portland, Oregon 97207-0751, collectively hereinafter referred to as the "Parties".

The Parties agree to the provisions of this DUA in order to address the requirements of Rutgers transferring data to Recipient and to protect the interest of both Parties.

1. **INTRODUCTION** In accordance with National Science Foundation Grant DRL - 0733255 ("Grant") and the requirement to share data generated under Grant awards with other researchers, Rutgers is willing to transfer the data developed under the Grant to Recipient under the project Bioinformatics: Learning by Doing ("Data") for the use described in Article 2 below.
2. **USE** Rutgers will transfer an electronic copy of the original Data electronically to Recipient solely for the basis to facilitate Douglas Lownsbery's Master's Thesis Research Proposal ("Thesis"). Recipient shall have the right to use the Data provided by Rutgers for this thesis.
3. **RESTRICTIONS ON USE** Recipient agrees Douglas Lownsbery will not use or further disclose the Data other than as permitted by this DUA, or as otherwise required by law or regulation. Recipient shall use appropriate safeguards to protect the Data from misuse or inappropriate disclosure and to prevent any use or disclosure other than as provided in this DUA or as otherwise required by law or regulation. Recipient shall not attempt to identify the individuals to whom the Data pertains, or attempt to contact such individuals.
4. **REPORTING** Recipient shall report annually to Rutgers the status, use and location of the Data.
5. **TERM AND TERMINATION**
 - (a) **Term.** The Term of this DUA shall be effective as of the date first written above, and shall terminate not to exceed three (3) years thereafter. At such time, Recipient shall destroy or securely delete Data from any Recipient servers and any non-network computers.
 - (b) **Termination for Cause.** Should Recipient or Douglas Lownsbery commit a material breach of this DUA, which is not cured within thirty (30) days after Recipient receives notice of such breach from Rutgers, Rutgers shall terminate the DUA immediately. Neither Recipient nor Douglas Lownsbery shall retain any copies of the Data.

6. **PUBLICATION AND ACKNOWLEDGEMENT** When Douglas Lownsbery publishes the results of his thesis, it is an NSF requirement to credit the project as an NSF award and NSF division, the project name, the university and the project director as follows:

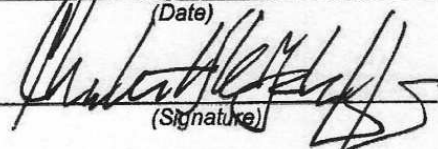
Bioinformatics: Learning by Doing, NSF DRL 0733255, awarded to Rutgers, The State University of New Jersey, Andrew K. Vershon, project director.

7. **MISUSE OF DATA** Recipient agrees to be responsible for its own negligent acts and omissions to the full extent allowed by law for the misuse or unauthorized use of Data.

RUTGERS, The State University of New Jersey


PORTLAND State University

13, Aug 2013

 (Date)


 (Signature)

8/12/13

 (Date)


 (Signature)

Charles H. Wyckoff, CCM, CFCM
Director, Industrial Agreements

 (Title)

Jennifer L. Ward
 Associate Director
 Sponsored Projects (Project Name)
 Portland State University

 (Title)

Appendix C. Student Educational Background Survey

1. In the fall semester, I will be a high school: _____ Freshman _____ Sophomore _____ Junior _____ Senior _____ Home Schooled

2. This is my: _____ First, _____ Second _____ Third _____ Fourth year participating in the WSSP project.

3. Which of the following classes have you taken? (Check all that apply)
 - AP or Honors Biology Calculus
 - Chemistry Trigonometry
 - AP or Honors Chemistry Statistics
 - Physics Other Advanced mathematics:
 - AP or Honors Physics _____
 - Other Advanced Sciences Advanced computing course _____

4. What advanced Science course(s) are you taking this year?
(Text box)

5. Are you interested in pursuing science as a college major?
 - Yes, I am interested in the Life Sciences (e.g. Biology)
 - Yes, I am interested in other types of science
 - No, I am interested in another major (Text box _____)
 - Don't know yet
 - I am not planning to attend college

6. For each statement below, circle the number that most closely corresponds to your feelings about the statement. Select a 5 if you strongly agree with the statement; select a 1 if you strongly disagree with the statement. Select 3 if you neither agree nor disagree.
 - a. It is important to know science in order to get a good job. 5 4 3 2 1
 - b. I am interested in a career in scientific research. 5 4 3 2 1
 - c. I enjoy doing laboratory exercises more than computer analysis. 5 4 3 2 1
 - d. I am interested in a career in the medical field. 5 4 3 2 1
 - e. There is little need for science in most jobs. 5 4 3 2 1
 - f. I am interested in a career that is related to the life sciences. 5 4 3 2 1
 - g. Science is useful for the problems of everyday life. 5 4 3 2 1

Your responses to these questions are optional, but appreciated.

7. Gender ___ Female ___ Male

8. Ethnicity

- ___ American Indian or Alaska Native
- ___ Asian
- ___ Black or African-American
- ___ Native Hawaiian or Pacific Islander
- ___ White
- ___ Other

Appendix D. Student Bioinformatics Original Pretest/Posttest

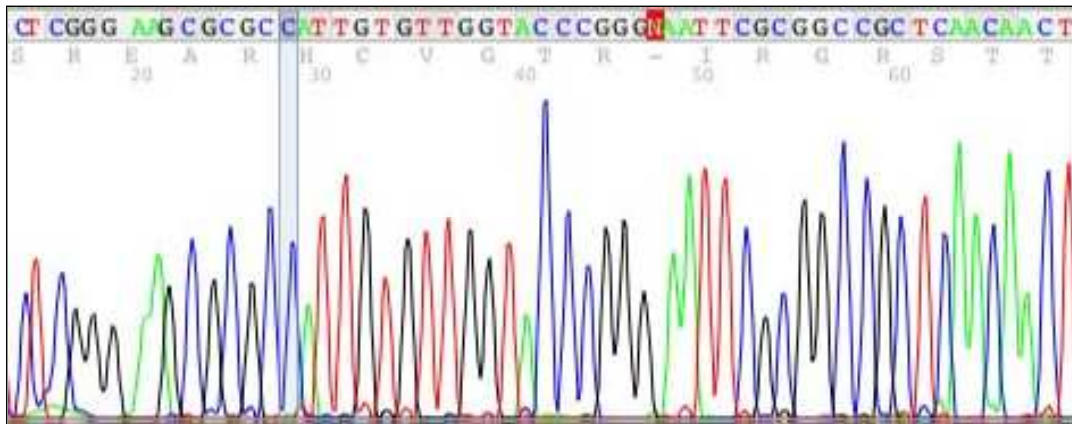
As part of the registration process for DSAP we are asking students to take a survey before you begin the analysis of your DNA sequences. We will also ask you to take the survey a second time before we submit your DNA sequences for publication on the NCBI databases. The following questions will give researchers a chance to see how much you know about genomics and genetics before and after using the DSAP program.

At the end of the survey there are some questions that will help researchers gather information about DSAP users. Your responses will help improve future versions of DSAP.

The WSSP instructors and your teachers will not know who completed which assessment! All students are number coded.

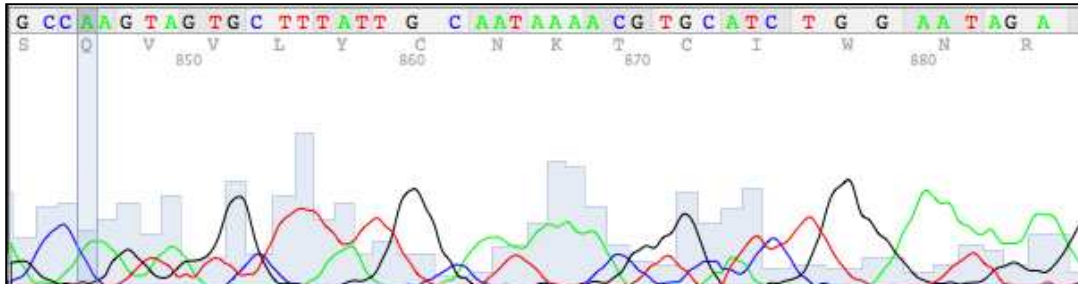
Please take this survey without looking up the answers on the internet, or in a textbook or lecture notes. Please limit yourself to 60 minutes. You will sign an electronic pledge on this at the end of the survey.

Thank you for participating in this survey.



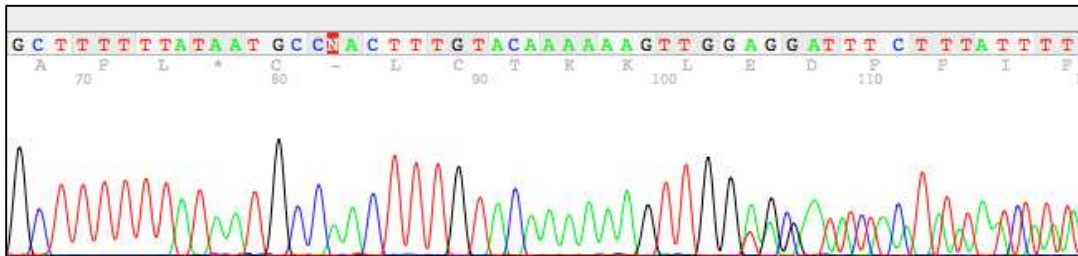
Q1. The DNA sequence for a cDNA clone is derived from a waveform generated by a DNA sequencing instrument. In the colored waveform shown above, the different peaks represent:

- A) the protein sequence.
- B) fluorescent labeled DNA pieces of different sizes.
- C) the different sizes of each actual base.
- D) probabilities for each base based on typical DNA sequences.



Q2. The first step in analyzing a DNA sequence is to determine if it is readable by looking at the waveform of the sequence. Which of the following is the best reason why you would consider the above waveform to be unreadable?

- A) The waveform is unreadable unless you have special software to read it.
- B) The letters in the top sequence have spaces that are too large.
- C) The waveform peaks are not distinct and overlap each other too much.
- D) The numbers under the letter sequence are too large.



Q3. Which statement best identifies the last readable base in the DNA sequence from the waveform pictured above?

- A) C82, because it is the last base in the sequence before N83.
- B) A98, because it is the last base in a sequence of consecutive A bases.
- C) G103, because it is the last clear peak in the waveform with no overlaps.
- D) T112, because it is the last clear peak in the waveform with no overlaps.

Q4. If the number of bases in a DNA molecule were counted or compared or calculated, you would find:

- A) $A = C$ and $G = T$.
- B) $A = G$ and $C = T$.
- C) $T = A$ and $C = G$.
- D) no two bases would be equal in number.
- E) that all bases are equal in number.

Q5. In the analysis of a cDNA insert you may be asked to convert the sequence to its “reverse complement” form. This is due to the antiparallel nature of the DNA strands, which means that:

- A) the twisting nature of DNA creates nonparallel strands.
- B) the 5' to 3' direction of one strand runs counter to the 5' to 3' direction of the other strand.
- C) base pairings create unequal spacing between the two DNA strands.
- D) one strand is positively charged and the other is negatively charged.
- E) one strand contains only purines and the other contains only pyrimidines.

The screenshot shows a software interface with a left sidebar and a main content area. The sidebar is titled 'Clone: 1AH3.07' and contains a 'Clone Editing' section with options: Welcome, Crop Ends - SP, Crop Ends - XP (highlighted), SP/XP Overlap, and Edit Sequence. Below this are sections for 'BLAST' and 'Analysis & Submission'. The main content area is titled 'Crop Ends - XP Sequence' and contains the following text:

In this step you will examine a waveform and determine the start and end of the cDNA insert.

4 Remove (crop) the sequence after the last base (question 3). Remove (crop) the sequence before the start base (question 2). Use the waveform software to convert this cropped XP sequence to its reverse complement form.

5 Copy the cropped reverse complement XP sequence here:

```
ATAATTGATGGTTTTACTTGCTTATACAAGAGTAACATTGATTTGTTTTCTATGTGCTAGGCAATG
TCAACGAGAATGAGCTGCTTCTGTCAAACGTTTTGAACGTGTATACGAATCCATCTGCCTAATTT
TGACAAAGGATGGAAAAAAGGAATCTCATAGACAACCTTGATGCTATAATGCTAGTCATAGACG
AAATTGTTGATAAAGGAATAATTCTTGAAGCCGATCCAGAAGCTATTGCCATAAGGTATCAATAAG
```

Number of characters: 562 Cursor is on base: A001; range marked:

At the bottom of the main content area are two buttons: '<< Back' and 'Next >>'.

Q6. DSAP asks you to compare two DNA sequences from the same clone to see if there is an overlap. The first sequence (the SP sequence) is entered as is. The second sequence (the XP sequence) is derived from the complementary strand of DNA from the same region. Looking at step #4 above, why do you need to use the waveform software to convert the XP sequence to its reverse complement form before entering it in step #5 to compare it to the SP sequence?

- A) Complementary strands of DNA use different base pair languages, so the XP strand needs to be converted by the waveform software before it can be compared to the SP strand.
- B) The XP strand of DNA is considered “junk” DNA and needs to be converted by the waveform software before it is compared to the SP strand.
- C) The SP and XP sequences have different open reading frames, so the XP strand needs to be converted before it can be compared to the SP sequence.
- D) Complementary strands of DNA are anti-parallel, so the XP sequence needs to be reversed and converted to complementary base pairs by the waveform software before looking for overlaps with the SP sequence.

Q7. The information contained in a strand of DNA may determine the sequence of:

- A) amino acids in a protein.
- B) sugars in a polysaccharide.
- C) amino acids in a fat.
- D) fatty acids in a protein.
- E) fatty acids in a fat.

Q8. The BLASTN search program allows you to evaluate how similar a DNA sequence from one organism is to DNA sequences found in other organisms by comparing e-values. Small e-values usually suggest that sequence similarities are due to shared ancestry. Which process best explains a small e-value for two organisms that are **not** very closely related?

- A) Polymorphism
- B) Convergent evolution
- C) Polyandry
- D) Crossing over

Clone: 1AH3.07

► Clone Editing

▼ BLAST

BLASTn

BLASTx

Define ORF

BLASTp

BLASTx vs. BLASTp

► Analysis & Submission

BLASTn

In this step you will perform a BLASTn search of two DNA databases for matches with the edited sequence.

1 Open the [BLASTn](#) search page and run a search with the edited sequence against the [nr/nt database](#) Retrieve sequence

2 List the best matches from three [different](#) organisms:

Accession #	Definition	Organism	E Value
XM_001944025	Acyrtosiphon pisum	pea aphid	3e-37
XM_001599118	AGAP007963-PA (Aga	Armigeres subaibatus	7e-12
XM_623377	Drosophila grimshawi	Drosophila grimshawi	1e-09

3 Open the [BLASTn](#) search page and run a search with the edited sequence against the [est database](#)

« Back
Next »

Q9. Your instructor is convinced that your sequence is very similar to sequences from the pea aphid, Armigeres, and fruit fly, Drosophila, that were found in the above BLASTn search. Based on the e-values, would you agree or disagree with your instructor's position?

- A) Yes, because all three e-values are less than 1.
- B) Yes, because only very similar sequences are returned by a BLASTn search.
- C) No, because plants and animals cannot have similar DNA sequences.
- D) No, because 3e-37 is a much smaller number than 1e-09, so they are not that similar.

Q10. Scientists often compare DNA sequences from different organisms.

a) What can scientists learn from highly conserved regions of DNA?

(Text box for answer)

b) What can scientists learn from differences between the DNA sequences?

(Text box for answer)

Q11. Under what conditions would you use nucleotide sequences rather than protein sequences (or amino acid) to prepare phylogenetic trees (a scheme that shows the evolutionary relationships of organisms)?

(Text box for answer)

Q12. Arrange the following steps involved in synthesis of a protein in the correct order.

- i. A complementary RNA copy of DNA is made.
- ii. The DNA double helix unwinds.
- iii. mRNA binds to ribosomes.
- iv. The amino acids of two adjacent tRNAs form a peptide bond.
- v. mRNA leaves the nucleus.
- vi. An anticodon of tRNA recognizes an mRNA codon.

A) i, ii, iii, v, vi, iv

B) ii, i, iii, v, iv, vi

C) ii, i, iii, iv, vi, v

D) iv, v, ii, i, vi, iii

E) ii, i, v, iii, vi, iv

Q13. Using the DSAP program you will determine if your DNA sequence codes for a protein. Suppose a gene has the base sequence AUGGCCAGAAC. The polypeptide encoded by this gene has how many amino acids?

A) 2

B) 4

C) 6

D) 12

E) 36

Q14. When a BLASTP protein database search is performed using the amino acid sequence for a clone, a number of potential matches are generated from different organisms.

If the triplet CCC codes for the amino acid proline in bacteria, then in plants CCC should code for:

- A) leucine.
- B) valine.
- C) cystine.
- D) phenylalanine.
- E) proline.

Q15. One of the steps in analyzing a clone involves an examination of the potential amino acid sequence that would be generated from the cDNA during protein synthesis. In the early steps of protein synthesis, the production of mRNA from a gene sequence is called:

- A) translation.
- B) transformation.
- C) transcription.
- D) activation.
- E) replication.

Clone: IAH3.07

▶ Clone Editing

▼ BLAST

BLASTx

BLASTx

Define ORF

BLASTp

BLASTx vs. BLASTp

▶ Analysis & Submission

BLASTx vs. BLASTp

In this step you will compare the BLASTx and BLASTp search results to determine if you chose the correct ORF in the Toolbox.

1 Look at the BLASTx and BLASTp results that you have entered:

BLASTx	
Accession #	E Value
XP_001944060	
XP_001599168	
XP_001656371	

BLASTp	
Accession #	E Value
XP_001599168	
XP_001944060	
XP_316555	

Do the proteins found in the BLASTp search match the proteins found in the BLASTx search?

Compare the E-Values for each protein found by the BLASTx and the BLASTp searches. Are the E-Values similar?

Q16. In step #1 above you are asked to compare the results of a BLASTx search, which searches a protein database using a translated nucleotide query, and a BLASTp search, which searches a protein database using a protein query. Why might these two search tools produce different results if sequences generated from the same clone are used?

- A) The two searches use two different databases, which will automatically produce different results.
- B) Nucleotide sequences can produce variable match results because of different reading frames while a protein sequence can only be read one way.
- C) Translated nucleotide queries will contain many mutations that the protein queries will not.
- D) Nucleotides and proteins are completely different molecules and will produce different results.

Q17. Using the table, determine what amino acid sequence will be generated, based on the following mRNA codon sequence?

5'AUG-UCU-UCG-UUA-UCC-UUG3'

- A) met-arg-glu-arg-glu-arg
- B) met-glu-arg-arg-gln-leu
- C) met-ser-leu-ser-leu-ser
- D) met-ser-ser-leu-ser-leu
- E) met-leu-phe-arg-glu-glu

		Second Letter				
		U	C	A	G	
1st letter	U	UUU Phe UUC UUA Leu UUG	UCU Ser UCC UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G
	C	CUU Leu CUC CUA CUG	CCU Pro CCC CCA CCG	CAU His CAC Gln CAA CAG	CGU Arg CGC CGA CGG	U C A G
	A	AUU Ile AUC AUA Met AUG	ACU Thr ACC ACA ACG	AAU Asn AAC Lys AAA AAG	AGU Ser AGC Arg AGA AGG	U C A G
	G	GUU Val GUC GUA GUG	GCU Ala GCC GCA GCG	GAU Asp GAC Glu GAA GAG	GGU GGC Gly GGA GGG	U C A G

Reading Frame 1
Total codons = 265. ORF #14 (in blue) is longest with 50 amino acids.
Select a run of amino acids to see the corresponding DNA sequence.

IADSRILGI*ILIKIVHLNGNLGSHQCTLSKE*PFLIMMATDSLPTM
MRMDSQLLKNKKILRRSYFLKQAEPKRLS*LMVLLAYTRVTLICFSM
C*AMSTRMSCFCQTF*TVYTNPSA*FCTRMWKKGIS*TTLML*C*S*TK
LLIKE*FLKPIQKLLPIRYQ*GVKISQ*ESSLYLRFKQKNSLNGPCL
SDFEDLHSIVCLGLIFPNTFFSFLHEYE*NS*IKFTLLDLEDDLGLN
LIGTGLFEIHSFIIYSLL*IKLC

Reading Frame 2
Total codons = 265. ORF #27 (in blue) is longest with 32 amino acids.
Select a run of amino acids to see the corresponding DNA sequence.

*QTRGF*EFEF**K*YISMVI*GAINVHCQRNSHS***WQPTLCQIL*CGWTL
NF*RTKKF*EEVIF*NKQSQFGNYHN*WFYLLIQE*H*FVFLCARQCQRE*AA
SVKRFELCIRIHLNFAQCGGKESHROP*CYNASHRRNC**RNNS*SRSRSY
CP*GINKE*RYPNRAVCISGSSNSERTA*MVPA*VILKIYIPLFVWV*YFQI
HFFPFCCMMNKIHELNSRFWTLRMIWV*T*SELDYSKFILLFIHYCK*NF-

Reading Frame 3
Total codons = 265. ORF #2 (in blue) is longest with 174 amino acids.
Select a run of amino acids to see the corresponding DNA sequence.

SRLEDFRNLNFDKNSTSQW*SREPSMYTVKGIALDNDGNRLFAYYD
ADGLSTFKEQKNFEKLFKTSRANSEIIIDGFTCLYKSNIDLFFYV
LGNVNEHELLSNVLCVYESICLILHKDVEKRNLIDNLDAIMLVIDE
IVDKGII LEADPEAIAHKVSI RSEDIPIGEQSVSQVLQTAKEQLKWSL
LK*F*RFTFCHLFGFDISKYIFFLSVA*I*IKFMN*IHAFGP*G*FGFE
LNRNWIIRNSFFYYLFIIVNKTL-

Q18. The Toolbox program determines the protein sequence translated from three different reading frames of a DNA sequence. Based on the Toolbox results shown on the right, which reading frame most likely codes for a protein? (* indicate stop codons)

- A) Reading Frame 1
- B) Reading Frame 2
- C) Reading Frame 3
- D) Reading Frame 1 and Reading Frame 3 are both equally likely.
- E) None of the reading frames is likely to produce a protein.

Clone: 1AH3.07

▶ Clone Editing

▼ BLAST

BLASTn

BLASTx

Define ORF

BLASTp

BLASTx vs. BLASTp

▶ Analysis & Submission

Define ORF

In this step you will determine if there is an open reading frame (ORF) in your edited DNA sequence

1 Paste the edited DNA sequence into the [Toolbox](#), where it can be translated in each of the reading frames. Retrieve sequence

2 Which reading frame is most likely to code for a protein? +1 +2 +3 None

3 Paste the most likely protein sequence or a portion of a protein sequence here:

SREPSMYTVKGI AILDNDGNRLFAKYDADGLSTFKEQKNFEKLFKTSRANSEIIIDGFTCLYKSNIDLFF
 YVLGNVNEHELLSNVLCVYESTCLLHKDVEKRNLDNLDAIMLVIDEIVDKGIILEADPEAIAHKVSRSEDI
 PIGEQSVSQVLQTAKEQLKW SLLK

Number of characters: 174 Cursor is on amino acid: S901; range marked:

« Back
Next »

Q19. The Toolbox program in DSAP allows you to analyze a DNA sequence to determine each of the potential open reading frames, or ORFs. In step #2 above you are asked choose one of three likely reading frames, +1, +2, or +3, or none of them. Why does DSAP only offer three ORFs to choose from for any given sequence that is read in one direction?

- A) Because the three ORFs refer to DNA, mRNA, and protein sequences, respectively.
- B) Because DNA and RNA only share three bases: A, G, and C.
- C) Because codons contain three bases, so after the first three ORFs they start to repeat.
- D) Because DNA contains three strands, each with its own ORF.

Q20. You have found two individuals (Mut1, Mut2) carrying mutations in the wild type DNA sequence shown below. DNA from both individuals has been sequenced. In one mutant DNA sequence there is a single nucleotide change in which a T base has been substituted by a C (underlined base in the Mut1 sequence). In the second mutant DNA sequence there is a deletion of the same base (underlined position in the Mut2 sequence).

		Second Letter					
		U	C	A	G		
1st letter	U	UUU Phe UUC UUA Leu UUG	UCU Ser UCC UCA UCG	UAU Tyr UAC UAA Stop UAG Stop	UGU Cys UGC UGA Stop UGG Trp	U C A G	
	C	CUU CUC Leu CUA CUG	CCU Pro CCC CCA CCG	CAU His CAC CAA Gln CAG	CGU CGC Arg CGA CGG	U C A G	
	A	AUU AUC Ile AUA AUG Met	ACU Thr ACC ACA ACG	AAU Asn AAC AAA Lys AAG	AGU Ser AGC AGA Arg AGG	U C A G	
	G	GUU GUC Val GUA GUG	GCU Ala GCC GCA GCG	GAU Asp GAC GAA Glu GAG	GGU GGC Gly GGA GGG	U C A G	
						3rd letter	

Which mutation is more likely to have a larger effect on the activity of the protein?

Wild type 5'- CGA ATT CAT CTG ATA TTG ATA AAA ATG AGT - 3'

A) Mut1 5'- CGA ATT CAC CTG ATA TTG ATA AAA ATG AGT -3'

B) Mut2 5'- CGA ATT CA_ CTG ATA TTG ATA AAA ATG AGT -3'

Explain your answer.

(Text box for answer)

PLEDGED:

I worked on this assessment independently. I did not use any texts or other resources, and spent no more than 60 minutes on it.

Signed: _____

School: _____

Thank you for participating in this survey! Your responses will help researchers make the DSAP more effective and meaningful for students to use. Your participation is greatly appreciated.