

IMPROVING UNDERGRADUATE STUDENT ENGAGEMENT IN STEM THROUGH BIOINFORMATICS RESEARCH

Michele C. Morris

A DISSERTATION

**Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in
The Biotechnology Science and Engineering Program
to
The Graduate School
of
The University of Alabama in Huntsville
December 2023**

Approved by:

Dr. Neil Lamb, Research Advisor
Dr. Joseph Ng, Committee Chair
Dr. Debra Moriarity, Committee Member
Dr. Sandra Lampley, Committee Member
Dr. Sara Cline, Committee Member
Dr. Luis Cruz-Vera, Program Coordinator
Dr. Jon Hakkila, College Dean & Graduate Dean

Abstract

IMPROVING UNDERGRADUATE STUDENT ENGAGEMENT IN STEM THROUGH BIOINFORMATICS RESEARCH

Michele C. Morris

**A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy**

Biotechnology Science and Engineering Program

**The University of Alabama in Huntsville
December 2023**

A novel intervention to introduce undergraduate students to bioinformatics, provide authentic research opportunities, and support student retention in science majors and careers is presented in the Characterizing Our DNA Exceptions (CODE) project. The need to increase the number of qualified STEM graduates, particularly in the fields of bioinformatics and computational research, was the impetus for this project that leads students down the path of scientific discovery as they characterize genomic variants of uncertain significance (VUS). This study sought to examine if student participation in a CODE research project would increase their bioinformatics awareness, interest, comfort, and knowledge, as well as the psychosocial measures of science self-efficacy, identity as a scientist, and intention to persist in a STEM major or career. Using the theoretical framework of Social Cognitive Career Theory (SCCT), the author hypothesized that students would see a positive shift in these constructs. Student participants at 17 colleges and universities completed pre- and post-project questionnaires. Data analysis using paired samples t-tests showed significant positive shifts in participant awareness, comfort, and knowledge of bioinformatics concepts. Students also gained significant

improvements in their science self-efficacy, science identity, and intention to persist in a STEM major or career. The measure for participant interest ranked high on the pre-surveys and showed non-significant increases following participation in a CODE project at their institution. Facilitator surveys provided positive feedback and formative suggestions for the program. Comments in the student interviews highlighted the program's strength in building student confidence and research experience. Many previous studies related to bioinformatics education activities have shown increased student knowledge following participation, but few have examined psychosocial changes in science self-efficacy, identity as a scientist, and persistence in science. This study highlights the transformative potential of bioinformatics research projects within the CODE program. CODE presents a promising model for enhancing science identity and self-efficacy in undergraduate students and facilitating the cultivation of a diverse and skilled STEM workforce.

Acknowledgements

Sincere thanks go to the members of my dissertation advisory committee for their support and advice during my pursuit of this degree. From the University of Alabama in Huntsville, Dr. Joseph Ng, Dr. Debra Moriarity, and Dr. Sandra Lampley are excellent science advisors, and I am thankful to have them on my committee. From Athens State University, Dr. Sara Cline has strongly supported this project for many years, and I appreciate her participation in my committee.

My research advisor, Dr. Neil Lamb of HudsonAlpha Institute of Biotechnology, has been a part of this project from its inception and an invaluable asset to my progress in this degree. I am thankful and appreciative of his kindness, support, and advice through the years.

I would also like to acknowledge and thank Dr. Jeremy Prokop for the initial encouragement to begin this program and his support of our faculty, participants, and my efforts. The Characterizing Our DNA Exceptions (CODE) project was born of Dr. Prokop's educational outreach and stands on a solid foundation because of him.

Dr. Cynthia Stenger of the University of North Alabama is another strong supporter who gave me the courage to pursue this degree. Her dedication to her students is inspiring, and I am thankful to have her as a part of the CODE team.

I'd also like to acknowledge the support and investment of time and knowledge of Advisory Committee members Dr. Kevin Drace and Dr. Eric Johnson, both pilot facilitators of the CODE program.

Thank you to David Hinds, Nikki Mertz, and Yokshitha Bathula for supporting CODE and the participants. You have all provided me with advice and thoughtful words that have been much appreciated.

Thank you to all CODE Faculty Facilitators who have expanded their curricula to support these bioinformatics projects and the CODE participants for exploring genetic variants and computational biology. A special thanks to the students who volunteered for an interview about their experience - your comments were enlightening.

The CODE program and this research are funded by a National Science Foundation IUSE-EHR grant, Award #2120918, and I am grateful for that support. The Alabama Power Foundation provided support for the pilot phase of CODE. Sustained support for the program was provided by HudsonAlpha Institute for Biotechnology. The project was fully approved by WCG IRB Study Number 1313664.

This dissertation is dedicated to my husband, who has remained a constant rock throughout my educational journey. I also dedicate this dissertation to my loving family, supportive friends, and wonderful colleagues at HudsonAlpha Institute for Biotechnology. Your faith in me has carried me to the end.

Table of Contents

Abstract.....	ii
Acknowledgements	v
Table of Contents	vii
List of Figures.....	xii
List of Tables	xiii
 Chapter 1. Introduction	 1
1.1 The Need for STEM Retention.....	2
1.2 The Need for Bioinformatics at the Undergraduate Level	6
1.3 The Importance of Undergraduate Research Experiences	7
1.4 Computational Characterization of Genetic Variants.....	8
1.5 Objectives of the CODE Project and the Research Study	9
1.6 Significance of the Study.....	11
1.7 Research Questions	12
1.8 Research Hypothesis	13
1.9 Methodology Overview	14
1.10 Summary.....	15
 Chapter 2. Review of the Literature	 17
2.1 History of Bioinformatics.....	18
2.2 Bioinformatics in Academia.....	19

2.3 Inquiry-Driven Learning Experiences	20
2.4 Social Cognitive Career Theory	22
2.5 Science Self-efficacy	24
2.6 Identity as a Scientist.....	26
2.7 SCCT as a Framework for Research	27
2.8 Exploring Programs Similar to CODE	28
2.9 Bioinformatics Education Interventions	29
2.10 Improving Student Knowledge and Confidence Through Instruction in Bioinformatics	33
2.11 Studies with Assessment of Student Cognitive Areas.....	35
2.12 Summary.....	38
Chapter 3. The Program Intervention.....	41
3.1 History of CODE	42
3.2 The CODE Project Pathway	44
3.3 CODE Facilitator Training Workshop	49
3.4 Implementation by Facilitators	54
3.5 Annual CODE Student Symposium	56
3.6 Quotes from Student Symposium Attendees.....	57
3.7 Advisory and Curriculum Committee	58
3.8 Summary.....	58

Chapter 4. Methodology.....	60
4.1 Quantitative Research - Design of Student Survey Instruments	60
4.2 Institutional Review Board (IRB) Approval and Ethical Compliance	62
4.3 Participants, Survey Administration, and Responses	63
4.4 Data Preprocessing	65
4.5 Survey Constructs.....	66
4.6 Analysis Methods	71
4.7 Related Samples Comparison Analysis.....	73
4.8 Qualitative Assessment - Student Interviews	75
4.9 Facilitators Surveys for Formative Evaluation.....	77
4.10 Summary.....	79
Chapter 5. Results.....	81
5.1 Student Participant Surveys - Demographic Data	82
5.2 Student Participant Surveys – Descriptive Statistics.....	84
5.3 Student Participant Surveys – Quantitative Findings.....	88
5.4 Bioinformatics Awareness and Interest.....	91
5.5 Comfort Level with Bioinformatics	93
5.6 Bioinformatics Knowledge.....	93
5.7 Research Experience	94
5.8 Science Self-efficacy	94

5.9 Identity as a Scientist.....	95
5.10 Intention to Persist in STEM	95
5.11 Qualitative Data – Student Interviews.....	98
5.12 Summary of the Interview Findings	98
5.13 Facilitator Surveys for Formative Assessment.....	101
5.14 Facilitator Post-Implementation Survey Results	103
5.15 Summary.....	106
Chapter 6. Discussion and Conclusions	108
6.1 Restatement of the Problem.....	109
6.2 Research Question 1: Does Participating in a CODE Project Increase Students' Awareness, Interest In, and Knowledge of Bioinformatics?	111
6.3 Research Question 2: Does Participating in a CODE Project Increase Students' Science Self-Efficacy?	114
6.4 Research Question 3: Does Participating in a CODE Project Increase Students' Scientific Identity?	118
6.5 Research Question 4: Does Participating in a CODE Project Increase Students' Intention to Persist in STEM?	121
6.6 Student Interviews	124
6.7 Research Experience	125
6.8 Implementation of CODE Projects by Facilitators.....	126
6.9 Faculty Facilitator Outcomes	127

6.10 Benefits of a Computational Research Program	129
6.11 Summation of Significant Findings.....	130
6.12 Limitations.....	131
6.13 Future Work.....	134
6.14 Conclusions	136
References.....	138
Appendix A. CODE Facilitator Training Workshop Schedule Example.....	153
Appendix B. IRB Protocol and Certificate of Action Approvals for Study	156
Appendix C. Study Recruitment Emails.....	170
Appendix D. Student Survey Consent.....	175
Appendix E. Summative Instruments for Student Participants.....	187
Appendix F. Survey Instruments for Facilitator Participants.....	195

List of Figures

Figure 2.1 Self-efficacy and career interests.....	23
Figure 3.1 The CODE project pathway	44
Figure 3.2 Carnegie classifications of participant institutions.....	54
Figure 5.1 Pre- and post-survey means: 5-point scale items.....	89
Figure 5.2 Pre- and post-survey means: 4-point scale items.....	89

List of Tables

Table 3.1	Bioinformatics techniques utilized by CODE projects	47
Table 3.2	Demographics of CODE facilitators	52
Table 4.1	Survey construct categories and internal reliability scores	69
Table 4.2	Facilitator formative evaluation methods.....	78
Table 5.1	Demographic statistics of student respondents to the pre- and post-surveys....	82
Table 5.2	Descriptive statistics for the pre-and post-survey data for each construct.....	86
Table 5.3	Student pre-and post-survey means and differences	90
Table 5.4	Student pre-and post-surveys – Paired samples comparison analysis data.....	96
Table 5.5	Facilitator's Training Workshop survey analysis	102
Table 5.6	Facilitator post-implementation survey items and means	103
Table 6.1	Significant findings of the CODE project study	130

Chapter 1. Introduction

Many voices have joined the call to bolster this country's STEM workforce over the past few decades. It is well-acknowledged that science and technology fields need well-trained graduates to maintain continued innovation and growth. The American Association for the Advancement of Science (AASA) and the National Science Foundation (NSF) issued a call to action in the 2009 Vision and Change report¹ that sought innovation in life science education to meet societal needs. That report presented several principal findings, which included the following: biological research is increasingly interdisciplinary, the complexity of the data generated has increased the need for computational and modeling skills in students, and new technologies allow questions to be posed and answered with new approaches that were not possible before.²

Meeting our nation's growing STEM workforce needs requires educational methods that not only introduce students to computational techniques but solidify them as members of the scientific community, increasing persistence in STEM majors and careers. The Characterizing Our DNA Exceptions (CODE) project was established with these goals. This research study examines the hypothesis that participating in a CODE research project will positively impact students in ways that will support the STEM workforce.

This chapter will present a clear rationale for the study and highlight published studies addressing similar topics and how this study complements those accomplishments. The research objectives will be described, as well as the research

questions and hypotheses. Additionally, this chapter will address the significance of this study, providing support for the value of its findings. Lastly, a high-level overview of the research methodology will be presented, followed by a guide to the organization of this dissertation.

1.1 The Need for STEM Retention

A tremendous amount of energy, time, and research has been devoted to transforming biology education to meet the global needs of our workforce. The AAAS Vision and Change reports have called the science community to action and chronicled the forward strides.^{1,3,4} While progress has been made, the U.S. still lags in STEM proficiency compared to other countries.⁵ Reversing this trend is a national responsibility for several reasons. Strengthening the science workforce will benefit the national economy because scientific innovation is linked to global competitiveness.^{6,7} Additionally, science occupations tend to be high-paying and well-perceived by the public. The predicted median annual wage in 2022 for a STEM position was \$97,980 vs. \$44,670 for a non-STEM position.⁸ A 2022 report from the National Science Board found that most Americans believe there will be more opportunities for the next generation because of science and technology. The report found that the American public's confidence in science and scientists is high, with 89% of Americans believing scientists make life better for the average person through their work.⁹ Thus, boosting access to these domains may benefit individuals.¹⁰

There is still a disparity between the number of STEM graduates with adequate qualifications and the number of STEM positions that need to be filled. Science and engineering employment constituted 6.2% of all U.S. jobs in 2021. STEM occupations

have grown more rapidly than the overall workforce.^{8,11} U.S. Bureau of Labor Statistics projects that jobs in STEM fields will increase by 10.8% between 2021 – 2031 compared to a 4.9% growth rate for non-STEM occupations.⁸ Recent graduation data show that of the ~2.1 million bachelor's degrees awarded in 2020-21, only 21% were in STEM fields.¹²

In a positive turn, the number of STEM graduates has been slowly increasing. Between 2010 and 2020 the 6-year graduation rate at public institutions increased from 58% to 64%.¹³ However, specific gaps still exist. There is a growing demand for employees with bioinformatics expertise, but the workforce and talent development pipelines lag behind. The mathematical science occupational group is projected to grow the fastest among all STEM occupational groups¹⁴, yet only 5% of STEM degrees awarded in 2020-21 were in computer science fields.¹² To compete globally, especially in emerging technologies such as quantum computing and artificial intelligence (AI), our nation needs to reinvigorate the STEM education system by reducing barriers such as high cost and low access for many populations.⁵

Underrepresented STEM groups have a higher dropout rate at the undergraduate level than white students. Black and Hispanic students enroll in STEM majors at rates proportionate to white students yet leave those majors at higher rates, with a 26% dropout rate for Black STEM majors compared to only 13% for whites.¹⁵ Even though women earn approximately 57% of all bachelor's degrees, they make up only about 18% of graduates with computer science degrees.¹⁶ These statistics highlight that the U.S. is currently preparing an insufficient workforce that is low on diversity, a characteristic that is vital to innovation.^{17,18}

Undergraduate populations are key to reducing the shortage of qualified STEM workers.^{19–21} Fewer than 40% of students who enter college as a STEM major actually earn a STEM degree within six years.²² This exodus from STEM majors is expected to leave an anticipated two million jobs unfilled by 2025 due to a lack of skilled candidates.²³

Undergraduate research is a powerful tool to engage students and increase their persistence to a degree, particularly in STEM fields. The participation of undergraduate students in research endeavors has yielded significant outcomes in enhancing their inclination towards scientific pursuits, bolstering their belief in their abilities, and fostering the cultivation of their scientific identity. Each of these factors contributes to student persistence in STEM majors.^{24–26} The 2012 President's Council of Advisors on Science and Technology Report expressed the need to increase the number of STEM college graduates by one million to fill current and future demand for STEM professionals. A growing amount of research suggests that student interest in STEM and the ability to visualize themselves in a STEM career play a significant role in student retention in the STEM field.^{27–29} Increasing student persistence in STEM is key to building the workforce required to prepare the United States to be scientifically competitive in the global economy.⁵

Traditionally, bioinformatics programs and experiences have only been available at large, research-intensive universities and inaccessible to students from smaller, less-resourced institutions. However, technological advances in molecular biology - including genomics, transcriptomics, and proteomics - have led to an explosion in publicly accessible biological databases containing oceans of information. Coupled with the

emergence of user-friendly tools to query and analyze this information, students from all types of post-secondary institutions can now participate in authentic bioinformatics research.

This study examines the CODE project, which was designed to capitalize on this sea change by building a regional network of faculty from smaller colleges and universities and training them to offer their students entry-level bioinformatics and protein modeling experiences. The CODE program incorporates skill development, active learning, and research-driven coursework. Students are brought explicitly into the world of research. They become knowledge producers, gaining insight into the biological significance of DNA variants identified from sequencing the genomes of patients with a disease or medical condition. Along the way, they gain confidence working with informatics tools and an awareness of related career pathways. This project developed a new model for bioinformatics research by partnering with HudsonAlpha Institute for Biotechnology, a non-profit genomics research and education center, and with post-secondary academic institutions, many in rural or low-income areas or serving predominantly underrepresented populations.

The CODE program addresses two issues that face undergraduate students, faculty, and the STEM workforce. First, there is a strong need to introduce students to bioinformatics tools, techniques, and careers with the goal of increasing the retention rate for STEM majors to fill the nation's workforce needs in this area. Second, smaller colleges often lack research opportunities due to insufficient funding and trained faculty. Valuable benefits can be obtained from participating in an authentic research experience at the undergraduate level.

1.2 The Need for Bioinformatics at the Undergraduate Level

The increasing output of sequence data over the past two decades has fostered the rise of countless bioinformatic databases and analysis tools. Bioinformatics is the interdisciplinary science of collecting, curating, analyzing, publishing, documenting, and archiving complex biological data, including molecular and sequence data. It has become a routine element of scientific research.¹¹ Over 20 years ago it was recognized that educators should provide biology graduates with an understanding of molecular structure and bioinformatics to keep pace with scientific advancements and this mandate still holds true today.^{30–33} The rate of discovery and the need for skill acquisition has only increased. It is becoming indispensable for biology students to master concepts related to the sequence and structure of proteins in order to develop skills that may be useful in a wide range of applications.³¹

Bioinformatics is an attractive method for students to engage in research experiences and inquiry-based learning due to the fact that it can be performed relatively cheaply with freely available data and software.^{34,35} Anyone who works with these resources knows how rapidly they can change to adapt to advancements in technology and new discoveries.³⁶ In many life science fields, particularly genetics, genomics, and biotechnology, knowledge of and education in using these resources is critical. While these skills have become indispensable for research, bioinformatics education is poorly integrated at the undergraduate level, especially at liberal arts institutions.³⁷ It is not uncommon to find a postgraduate course on bioinformatics in our nation's universities. However, it is vital that these subjects are introduced to students at the undergraduate level. A positive experience during these early years of matriculation can spark interest

and engagement for students in biology and bioinformatics and encourage their progress toward a postgraduate degree.^{38,39}

One obstacle to offering bioinformatics education at the undergraduate level is the need for faculty who feel qualified to teach these topics. A study by the Network for Integrating Bioinformatics into Life Sciences Education (NIBLSE)⁴⁰ revealed that most current professors in the life sciences did not study bioinformatics during their training due to the relative newness of the field. Faculty often lack the confidence to introduce these topics to their classes without prior training and a robust curriculum. The result is a gap in learning for many undergraduate and graduate students.³⁷

1.3 The Importance of Undergraduate Research Experiences

The expansive lab requirements and knowledge base required for DNA interpretation and variant analysis have historically limited scientific undergraduate research experiences to large universities with academic medical centers and departments staffed by federally funded research faculty. However, many undergraduates attend two- or four-year institutions with restricted access to research-intensive opportunities. In these settings, it can be difficult for faculty members to identify students with an interest in and aptitude for scientific discovery. It is also more likely that faculty members are limited in the amount of research they can conduct due to teaching responsibilities, inadequate funding, and lack of administrative support. Students from schools without access to research possibilities are disadvantaged when applying to professional or graduate programs compared to their peers at more prominent research-level universities. Overall, this reduces the number of students entering the pipeline for the STEM workforce.

Engaging in practical undergraduate research is important for encouraging students' dreams of attending graduate school or entering the STEM workforce. These experiences are particularly impactful for underrepresented minority (URM) students who may not have access to these opportunities owing to racial, ethnic, or financial inequities. These research experiences are often the most important part of a student's undergraduate education and determine their future job paths. Additionally, students with undergraduate research experience have substantially higher chances of being admitted into graduate programs and landing a job in the scientific field.⁴¹

In this study, the researcher implemented a program to address these needs at multiple colleges and universities across the southeast. CODE introduces authentic bioinformatics research to students from schools historically underrepresented in this type of undergraduate experience. This ongoing program is building a regional network of smaller post-secondary institutions and providing faculty with the training and educational resources to implement CODE.

1.4 Computational Characterization of Genetic Variants

A significant challenge in using genetic information for healthcare is that there is insufficient knowledge about how certain gene changes relate to diseases.⁴² Clinical genomics currently faces significant challenges in establishing the relevance of the majority of variants detected by sequencing research. In fact, according to Oliver *et al.*⁴³ the bulk of mutations found are present in proteins whose functions are yet unknown. Using computational analysis as a first step in characterizing these variants is a cost-effective and time-efficient method. Bioinformatics tools such as molecular modeling and molecular dynamic simulations can significantly increase the yield of information to aid

in the evaluation of pathogenicity.⁴³ Computational methods of variant analysis integrate with the high-throughput nature of genomics by providing robust predictions quickly. These methods share the goals of wet-lab experimentation but differ in their methods. Computer analyses can test a hypothesis that is difficult or impossible to test in a wet lab. The computational method of variant pathogenicity has been recognized by the American College of Medical Genetics (ACMG) as a viable method to contribute to the overall assessment of a mutation.⁴⁴ Pathogenicity prediction tools (SIFT, PolyPhen2, etc.) use various algorithms to examine such measures as evolutionary conservation and protein-level structural information to assign pathogenicity status. These programs are useful but have limitations. CODE projects utilize these predictive algorithms in conjunction with molecular modeling techniques. Molecular modeling examines variants in a three-dimensional, dynamic space allowing for a deeper understanding than viewing variants as occurring in the 2-D linear structure of nucleotides or amino acids.

During this research project, faculty, most with no experience in bioinformatics, were trained in a pipeline of techniques to characterize genetic variants using open-source databases and tools and 3-D modeling software. Taking these skills back to their schools, the facilitators guided their students in research experiences that contribute to the science community and provide wide-ranging benefits to the students. The focus of this study goes beyond the implementation of these projects and assesses the value of student and faculty participation in CODE research experiences.

1.5 Objectives of the CODE Project and the Research Study

- **Provide research opportunities for students** that increase their science self-efficacy, identity, and intent to enter or remain in a STEM major or career.

- **Broaden the students' awareness, interest, and skills** in bioinformatics careers and practices.
- **Create a robust community** of college faculty facilitators trained to mentor undergraduate students in bioinformatics research projects, recruiting facilitators from non-research-intensive institutions throughout the Southeast.
- **Develop a variant analysis and molecular modeling workflow** that can be easily integrated into existing and new courses at various educational institutions and used in multiple formats.
- **Provide a publicly accessible repository for student data**, particularly protein structures, molecular dynamic analyses, and findings of genetic variant impacts on those proteins.
- **Support students and facilitators** with technical assistance and online resources to ease bioinformatics hesitation.
- **Host an annual CODE Student Symposium** for students and mentors to share work and network with the science community.
- **Assess the impact of the facilitator training** using pre/post surveys in workshops and post-implementation surveys of active faculty members.
- **Assess the impact of project participation** on students through formative and summative evaluations to determine how participation shapes awareness and interest in informatics and whether CODE positively impacts student-reported measures of self-efficacy and resilience in the STEM career pipeline.

1.6 Significance of the Study

The work presented in this study significantly contributes to the scientific community by providing new opportunities for faculty and students to generate and analyze genomic data using computation tools. The findings from the study's assessment demonstrate the potential impact of using these research methods with students.

1. In the past 15 years, examples of bioinformatics-based research projects in the literature have slowly increased. However, only some have been assessed to evaluate the impact on student self-efficacy, science identity, and persistence in STEM. Evaluating these impacts will contribute to the knowledge of the usefulness (value) of these types of projects compared to primarily lab-based research.
2. The development of the CODE program has *practical implications* for improving bioinformatics education at the undergraduate level.
 - a. Educational outcomes for students are improved by the training of faculty in bioinformatics research techniques. CODE has provided 70 facilitators, many inexperienced in bioinformatic techniques, with the skills and confidence to introduce these concepts to their students. In 2009, the NSF issued a report on the state of science education in the United States and made recommendations for improvement in key areas.¹ The report emphasized the need for training to help present and future faculty develop effective approaches to undergraduate biology education. The CODE research project incorporates that recommendation and others in a pipeline of resources for faculty and students.

- b. CODE projects are often the first opportunity a student has been given to explore data-driven discovery. The NSF Vision and Change report noted growing complexity in data generated by research and that computational and modeling skills are needed to deal with the complexity. This finding led to the recommendation that undergraduates develop computational competence through experience with large databases, modeling, simulation, and computational and systems-level approaches in science research. The report emphasizes the importance of engaging students in the scientific process by recommending that they have opportunities to participate in authentic research experiences.²
- c. The unique student-friendly methodology for characterizing genomic variants of uncertain significance developed by CODE significantly contributes to the options available to faculty to incorporate bioinformatics and authentic research into their classes. Through publications and a public-facing website, the CODE project format will become a valuable addition to the STEM education resources, accessible to participants of all levels.

1.7 Research Questions

The CODE project's primary goal is to increase access to research-based learning experiences, particularly for students at small institutions with fewer undergraduate research opportunities using bioinformatics techniques. As part of this research study, the program sought to introduce a broad range of students to bioinformatics in a mentored project-based format to capture interest, build confidence, and encourage continued

participation in science education and a STEM career path, such as bioinformatics. To guide the investigation and provide a framework for data collection and analysis, the following research questions were posed by the author and assessed by quantitative and qualitative evaluations.

The current study is designed to investigate the following research questions (RQs):

- RQ1: Does participating in a CODE project increase students' awareness, interest in, and knowledge of bioinformatics?
- RQ2: Does participating in a CODE project increase students' science self-efficacy?
- RQ3: Does participating in a CODE project increase students' scientific identity?
- RQ4: Does participating in a CODE project increase students' intention to persist in STEM?

1.8 Research Hypothesis

CODE projects provide students exposure and introductory training in bioinformatics, demonstrating an alternative career path to more common goals such as health care or lab-based research. Participating in a CODE project provides students with a more interactive, visually oriented, and discovery-based learning approach to genetics. When compared to conventional lab courses, research projects like CODE give students the chance to make discoveries that are pertinent to stakeholders outside of the classroom, including working scientists, and to engage in iterative work like troubleshooting, problem-solving, and building off one another's progress in a way that is more similar to the practice of STEM.^{45,46} The author hypothesizes that participating in a data-driven bioinformatics research project such as CODE will increase student awareness, interest

in, and knowledge of computational biology, as well as their self-efficacy, science identity, and persistence in a STEM major. The null hypothesis holds that student participants will not see increases in these measures after participating in a CODE project at their school.

1.9 Methodology Overview

The CODE project began training facilitators in 2018 with a group of five pilot schools with a primary goal of introducing authentic bioinformatics research to undergraduate students using computational modeling and database analysis to characterize DNA variants identified through clinical studies. The program was designed to increase access to research-based learning experiences, particularly for students at small institutions with fewer undergraduate research opportunities. CODE provided facilitator training, protein modeling software, an informational website, and an annual student symposium for undergraduates and mentors. The program expanded across Alabama and neighboring states as additional facilitators were trained and began to implement CODE projects with their students. The assessment study described in this paper began in the fall semester of 2021.

A mix of quantitative and qualitative methods were used in this study. Before and after participating in a CODE project, surveys were given to student participants. The surveys included 44 statements, rated on Likert scales, relating to the hypothesis themes of bioinformatics awareness, interest, comfort, and knowledge, as well as science self-efficacy, science identity, and intention to persist in STEM. A grouping of eight statements was also included to establish prior research experience among the

participants. The resulting data were examined using non-parametric and parametric comparison tests to provide evidence to reject or accept the null hypothesis.

Focus interviews with student volunteers provided additional qualitative data to expand on the survey findings. These interviews were transcribed, coded for similar themes, and compared with the survey results. Additionally, faculty members were asked to complete surveys before and after the CODE Facilitator Training Workshops. These surveys also included open-ended questions to gather qualitative data to supplement the survey results. Nine facilitators completed a post-implementation survey to measure long-term changes in their confidence, project engagement, and opinions.

1.10 Summary

Participation in discovery research as an undergraduate has many positive impacts, including increased student retention in STEM majors and careers. Factors that contribute to retention can be gained through research experiences. The CODE project provides authentic research projects to student populations with fewer opportunities to engage in the traditional wet lab study due to limited resources. CODE projects utilize bioinformatics-based research, introducing students to a growing STEM field critical to our nation's future. The CODE program and the assessment study outlined in this study offer a novel contribution to the scientific community, providing a fresh avenue for genomic data exploration using computation tools. The assessment data offer insight into the research questions of how much participation in a CODE project impacts a student's awareness, interest, and knowledge of bioinformatics and increases self-efficacy, identity, and persistence in STEM.

The remainder of this dissertation will address the aspects of the author's research, implementation, and analysis. Chapter 2 will present a discussion of relevant literature and findings and the theoretical concept upon which the research project was based. A detailed description of CODE's history, the facilitators' training, student project implementations, program supports, and activities will be presented in Chapter 3, giving a clear picture of what comprises a CODE research project and how the program engages with faculty and students. Chapter 4 will describe the assessment measures, how they were selected and administered, and how data was collected and analyzed. The results from the data analyses will be presented in Chapter 5, followed by a discussion of the findings and recommendations for future work in Chapter 6.

Building upon the foundation in Chapter 1, the following chapter delves into an in-depth exploration of pertinent literature and findings underpinning this research effort. It offers insights into the theoretical framework upon which the project is grounded. Chapter 2 functions as a stepping stone for the subsequent chapters, laying the groundwork for the relevant work that guided the development of the CODE program.

Chapter 2. Review of the Literature

The landscape of science education is evolving in response to the increasing complexity of research-generated data, as highlighted in the National Science Foundation's report, *Vision and Change in Undergraduate Education, A Call to Action*.⁴⁷ This report underscored the growing importance of equipping undergraduates with computational and modeling skills to navigate this complexity effectively. Consequently, it recommended that students gain practical experience with large databases, modeling, simulation, and computational approaches in scientific research. Moreover, the report strongly emphasized engaging students in authentic research endeavors, particularly those that offer opportunities for computational learning.²

The future job market further reinforces the urgency of this paradigm shift. The U.S. Bureau of Labor Statistics⁴⁸ projects a remarkable 22.7% growth in employment for computer and information research scientists from 2022 to 2032, significantly outpacing the average growth rate across all occupations. This expanding demand for professionals skilled in computational informatics has spurred post-secondary institutions to explore effective methods of integrating computational learning to inspire and prepare students to pursue careers in this burgeoning field. However, several challenges loom on the path to achieving this goal. These challenges, including faculty expertise/training, student engagement, curriculum constraints, and student preparedness, have been recurrently identified as barriers in the scholarly literature.^{49–51}

The upcoming chapter conducts a comprehensive review of the scholarly literature pertaining to academia's response to the burgeoning field of bioinformatics. It explores the strategies employed to cultivate effective computational biology learning experiences while navigating the associated challenges. First, it provides an overview of the recent developments within the bioinformatics discipline, followed by an examination of how academia has sought to incorporate bioinformatics training into the curricula. The theoretical framework of Social Cognitive Career Theory and its role in framing this study's research plan will be described. Furthermore, the chapter highlights specific instances of bioinformatic interventions within educational settings, highlighting their evaluation methodologies (or the absence thereof) and drawing parallels to the CODE program.

2.1 History of Bioinformatics

The relatively short history of bioinformatics began in the 1960s – 1970s when scientists first started using computers to analyze biological data such as nucleotide and protein sequences. The European Molecular Biology Laboratory (EMBL)⁵² was launched in Europe in 1974, followed by GenBank in the United States in 1982.⁵³ These sequence databases laid the foundation for storing and retrieving genetic information. In the following years, the field progressed with the development of sequence analysis algorithms and tools, such as the Smith-Waterman algorithm for sequence alignment⁵⁴ and the Basic Local Alignment Search Tool (BLAST) for sequence similarity searching.⁵⁵ The demands of managing and analyzing vast amounts of genetic data generated by the Human Genome Project in the 1990s fueled the advancement of computational tools and methods. During the early years of the twenty-first century, the

field expanded into predicting and analyzing the three-dimensional structures of biological macromolecules, such as protein and nucleic acids, and using modeling techniques to understand gene function and interactions. The advent of next-generation sequencing technologies dramatically increased the amount of biological data to be analyzed. Bioinformatics tools and pipelines became essential for genomics research. This trend continues today, as bioinformatics techniques are rapidly evolving to support personalized medicine, drug discovery, and collaborative work across the globe.⁵⁶

2.2 Bioinformatics in Academia

A 1998 paper in bioinformatics education by R. Altman³⁰ recognized the need for training scientists with bioinformatics skills due to a demand for more employees with the knowledge needed to pursue research in the field. Altman focused on graduate-level training and proposed a two-year curriculum to teach graduate students the required skills. To round out this training, Altman recommended students also participate in journal clubs, study publications in the field, attend professional meetings, explore career options in both academic and industrial settings, and hone presentation skills. Finally, students should engage in an original research project.

An early advocate of undergraduate training, H. Salter argued that the fundamentals of sequence analysis should be part of every biochemist's or molecular biologist's skill set since they are a necessary tool for current macromolecular analysis in the laboratory.⁵⁷ His work highlighted several online resources and suggestions for integrating them into undergraduate curricula.

Stephan and Black called attention to the slow integration of bioinformatics and computational biology programs in universities.⁵⁸ They argued that the slow growth was

driven by the fact that the size and direction of Ph.D. programs in the life sciences are more closely tied to funding opportunities than to job market demands and that the field's interdisciplinary nature required interdepartmental collaboration that can be hindered by academic bureaucracy. Interestingly, they questioned in their paper whether bioinformatics/computational biology is really a field at all and if the interest in this type of education is “but a flash in the pan” or if the demand will continue in the future. The early 2000s saw an uptick in scholarly literature that provided curriculum support for integrating bioinformatics into the classroom.^{31,59–69} Many of these publications included assessments of the program. Most often, the assessment examined students’ knowledge gains, perceived gains in skills, or enjoyment of the implementations.^{62,70–73}

2.3 Inquiry-Driven Learning Experiences

Concurrent with the growth of bioinformatics integration into university classrooms was the effort to increase inquiry-driven learning through expanded undergraduate research experiences (URE). Apprenticeships, internships, summer research experiences, collaborative team learning, and course-based undergraduate research experiences (CUREs) have many instructional approaches and outcomes in common and facilitate this style of interactive learning.^{26,74–82} CUREs have become particularly popular because they can provide an authentic research experience for many more students than the traditional research internship.⁸³ The CURE outcomes outlined by Corwin *et al.*⁴⁶ support the benefits of these types of research experiences.⁴⁶ CUREs can be defined in many ways, and although a formal definition does not exist, it is generally agreed that these types of UREs will include these activities⁴⁶:

1. Reading and evaluating science literature

2. Selecting or designing methods
3. Collecting novel data
4. Analyzing results
5. Working collaboratively
6. Presenting results outside of class

One definition of a CURE is “a course in which students are expected to engage in science research with the aim of producing results that are of interest to a scientific community”.⁴⁶ These types of active learning experiences are common in universities today. They are built on the theory of situated learning, first postulated by Jean Lave and Etienne Wenger in the early 1990s.⁸⁴ This theory holds that students are more inclined to learn through active participation in a learning experience. One early program, the Genomics Education Partnership,^{75,85} introduced a course-based research project at diverse schools that could be tailored to meet local curriculum and student needs. This program has grown dramatically, and today is a nationwide collaboration of over 200 institutions conducting research in bioinformatics and genomics.

CODE projects are not explicitly labeled as CUREs but are often taught using similar parameters by the faculty facilitators. Most CODE projects include the six activities defined by Corwin as being CURE-specific.⁴⁶ These types of inquiry-based learning experiences, such as a CODE project, enable undergraduates to engage with the culture of scientific research by participating in the work scientists perform to produce knowledge.^{45,84,86} Participation in CUREs has been shown in numerous studies to increase students' science interest, retention, confidence, and motivation.^{83,87,88}

2.4 Social Cognitive Career Theory

In 1994, Lent, Brown, and Hackett introduced the Social Cognitive Career Theory (SCCT) as a comprehensive framework for comprehending career decision-making.⁸⁹ Grounded in Albert Bandura's social cognitive theory, SCCT explores the interplay among personal factors (*e.g.*, beliefs, attributes), environmental factors (*e.g.*, feedback, culture), and behavior.⁹⁰

At its core, SCCT focuses on the interrelationships among four primary elements: self-efficacy (belief in one's abilities), outcome expectations (anticipated consequences of behavior), interests (attraction to specific domains), and goals (intentions to pursue particular actions), as outlined by Lent and colleagues and depicted in Figure 2.1.⁸⁹ For instance, within a STEM-specific model, an individual who perceives proficiency in mathematics and foresees positive outcomes in a math-oriented career is likelier to exhibit a heightened interest in mathematics, establish goals to study math and persist in a math-centered career. Confidence and belief in favorable outcomes play pivotal roles in shaping goals and sustaining commitment to STEM fields. Hence, self-efficacy, outcome expectations, and interests influence career-related goal-setting and behavior.⁹¹

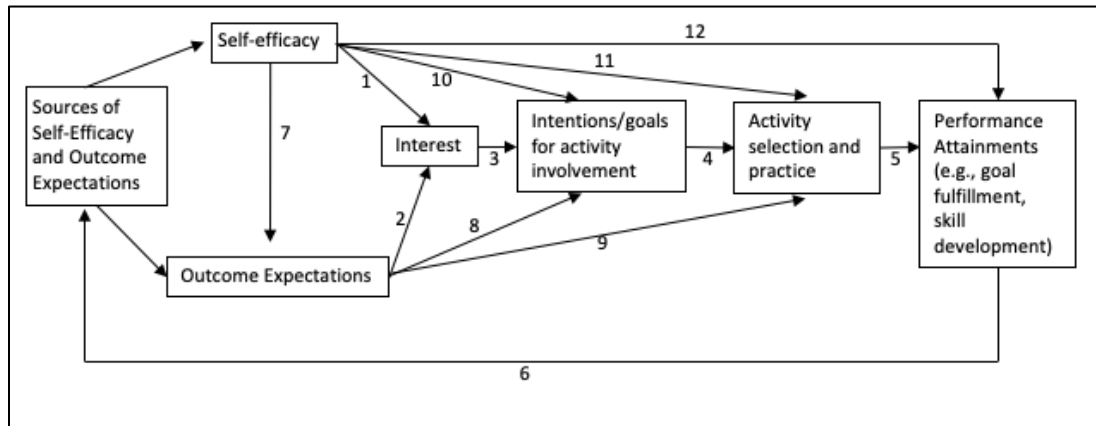


Figure 2.1 Self-efficacy and career interests. Model of how career interests develop over time as proposed by Lent *et al.* 1994.⁸⁹

Although the four social cognitive factors form the core of SCCT, they are not isolated entities. Individuals integrate social feedback from their environment (*e.g.*, peers, mentors) as they develop self-perceptions. Researchers have thus examined how additional constructs influence these core factors and, subsequently, career choices. Lent and colleagues proposed that personal characteristics (*e.g.*, gender, race/ethnicity, emotional affect, socioeconomic status), background factors (*e.g.*, family expectations, role model exposure), and educational experiences shape these cognitive factors and mediate their influence on career decisions.^{89,92} Additionally, external factors, such as support systems (*e.g.*, social networks) and barriers (*e.g.*, discrimination), influence career outcomes. While no single environmental element determines outcomes independently, each can exert a substantial impact. For instance, individuals with higher socioeconomic status, affording greater access to learning experiences and career-support resources, may experience enhanced career choices and persistence. Nevertheless, those

with ample financial resources do not consistently achieve their desired professional objectives, whereas individuals with few have succeeded.⁹¹

Social cognitive career theory describes the social cognitive process by which individuals gain interests and make career-based decisions. This framework was used as a guide to structure the student survey and interview questions for this research project. SCCT has been used for over 35 years to predict success in STEM careers, particularly for minority populations.^{93,94} This theoretical model can be used to elucidate the importance of UREs for students in the science arena. Of particular interest are impacts from participation in a CODE project on science self-efficacy, identity as a scientist, and the student's intention to remain in STEM, three of the primary elements of the SCCT model. Over three hundred undergraduates, most of whom were members of underrepresented racial and ethnic groups, were surveyed in a study by Chemers.²⁴ The findings showed that self-efficacy for scientific research mediates the relationship between UREs and intent to persist in STEM professions. Additionally, the results showed that science identity played an important role in mediating the relationship between self-efficacy in scientific research and the intention to remain in a STEM field. Research experiences, such as CODE, may increase a student's sense of self-efficacy for science research, strengthening science identity and encouraging commitment to a science career.²⁴

2.5 Science Self-efficacy

Self-efficacy expectations are a crucial early input into the career decision-making process. Albert Bandura pioneered the study of self-efficacy and defined it as the levels of confidence one has in one's ability to perform an action, achieve a goal, or attain a

specific performance outcome.⁹⁵⁻⁹⁷ Bandura argued that a person's belief in their ability to complete a task strongly influences the effort they will direct toward that goal.⁹⁵ Thus, if one believes they can be successful, they will expend higher levels of effort, engagement, and persistence to achieve their goal.⁹⁸

Across the breadth of his work, Bandura postulated four sources of efficacy expectations: mastery experiences, vicarious experiences, social persuasion, and physiological and emotional states.^{95-97,99} Mastery experiences are the most influential source of efficacy. They involve completing tasks and challenges successfully. In science, this could mean successfully completing a lab experiment or learning to navigate a database. Vicarious experiences can enhance self-efficacy, especially in science, when a student observes others, such as a peer, successfully completing a task and achieving a boost in confidence that they could do the same. Social persuasion plays a role in increasing self-efficacy when others, such as a teacher or mentor, praise one's abilities and progress and encourage continued work. Lastly, physiological and emotional states can play a key role in one's feelings of self-efficacy. Recognizing and managing stress and other emotional states related to a task can enhance self-efficacy. Both anxiety and excitement can add to the sense of mastery or incompetence based on one's perception of success in a task. Self-efficacy is enhanced if success can be attributed to internal or controllable causes, such as ability or effort, rather than to external factors, such as luck or the intervention of others.^{100,101} Engagement and self-efficacy are intrinsically linked. When students are interested in courses or activities, they may be more motivated to develop their skills, increasing self-efficacy. Conversely, a sense of self-efficacy can

encourage students to explore subject matter more thoroughly, fostering greater engagement.¹⁰²

2.6 Identity as a Scientist

In 1968, Erik Erikson defined identity as an individual's cohesive perception of oneself that remains consistent across time and in many situations.^{103,104} The construct 'identity as a scientist' was described and examined by Robnett *et al.*²⁵ in their longitudinal study to understand the associations between research experiences, science self-efficacy, and the extent undergraduates view 'scientist' as a core component of their identity. The concept is aligned with prior work by others,^{103,105} which emphasized the importance of forming an identity that ties one's sense of self with one's academic pursuits. The development of an identity as a scientist has been associated with favorable outcomes, such as expected and actual persistence in the science pipeline.^{24,106–108} Implementing interventions that foster the growth of a scientific identity among students is essential to increasing long-term outcomes such as persistence in a STEM field. Research has shown that students who identify as members of the scientific community are more likely to persist in science careers than students who do not adopt a professional science identity.^{109,110}

A scientific identity can be fostered through authentic research experiences.¹¹¹ These experiences affect a student's sense of belonging, recognition, interest, and performance in science, as well as their scientific competence. The integration of students into a community of experienced scientists during a research experience likely fosters their identity as scientists.⁸⁴ One study found that after participating in an intensive summer research experience, students reported learning to think and work like scientists,

and their faculty mentors confirmed these opinions.¹¹² Studies examining the benefits of research experience in the context of academic outreach programs have yielded comparable results. For instance, a study by Nagda *et al.*¹¹³ found that student-faculty research partnerships integrated underrepresented ethnic minority students into the university's culture, resulting in increased retention.

2.7 SCCT as a Framework for Research

This study examined variables most likely to lead to retention in STEM, such as increased science identity, improved confidence, and self-efficacy beliefs, through participation in active learning and real-world research experiences in an interdisciplinary mix of courses at 17 community colleges, liberal arts colleges, and universities. Social cognitive career theory (SCCT) expounds upon the social cognitive mechanisms through which individuals develop their interests and reach decisions about their careers. This study employed the SCCT framework as a guiding structure to investigate the following research inquiries regarding student involvement in a CODE project:

1. Impact on awareness, interest, and knowledge: What were the consequences of participating in the CODE program on undergraduates' awareness, interest, and comprehension of bioinformatics?
2. Influence on science self-efficacy and identity: What were the effects of participating in the CODE program on undergraduates' perceptions of their self-efficacy in science, their sense of scientific identity, and their inclination to pursue STEM disciplines?

Students who are introduced to a challenging field, such as bioinformatics, and have an experience that increases their self-efficacy related to that topic will increase their

interest and obtain positive expected outcomes related to that field.¹¹⁴ As theorized by SCCT, increased self-efficacy, interest, and outcome expectations are key drivers of persistence.¹¹⁵ The anticipated outcomes of this study align with the theoretical underpinnings of SCCT. It is hypothesized that students participating in a CODE project will develop more positive attitudes towards bioinformatics and STEM and feel more efficacious in pursuing a STEM career. Overall, these feelings will encourage students to be more likely to seek out paths to facilitate their career goals, be more likely to persist in a STEM major and consider a career path in computational biology as a viable option for them.

2.8 Exploring Programs Similar to CODE

Participating in undergraduate research offers a myriad of advantages. Extensive scholarly literature has consistently provided evidence supporting the positive effects of undergraduate research on science self-efficacy, science identity, and the intention to pursue a career in science. However, most of these investigations have mainly concentrated on laboratory-based encounters. Fewer studies have focused on computational programs and often have only examined improvements in student learning and, occasionally, confidence levels. This study's primary objective was to investigate the impact of an introductory-level bioinformatics project emphasizing discovery-based learning on students. The study aimed to assess shifts in student attitudes towards bioinformatics and their self-perceived levels of science self-efficacy and identification as a scientist—two critical constructs known to positively influence students' commitment to STEM majors and careers, as well as their intent to persist in STEM disciplines.

Much of the scholarly literature on bioinformatics education has predominantly focused on describing modules or instructional frameworks for introducing and teaching key concepts. Some of these papers have included assessments of knowledge acquisition and student engagement related to these activities. As the significance of integrating bioinformatics into undergraduate education has become increasingly apparent, evaluating these programs has expanded to encompass psychosocial factors such as confidence, self-efficacy, and science identity, particularly concerning STEM field persistence. The examination of the impact of bioinformatics research experiences is a growing area of inquiry, and this project contributes to a deeper understanding of the distinctive aspects and potential value of programs that feature bioinformatics research experiences.

The subsequent section will review literature related to programs that share common elements with the CODE program. Each of these studies has played a pivotal role in advancing the field of bioinformatics education at the undergraduate level, providing valuable insights that have contributed to the development and refinement of CODE.

2.9 Bioinformatics Education Interventions

In the 2000s, several institutions sought to teach genomics using sequencing data to integrate informatics with biology.^{61,116,117} Many modules and courses focused on web-based bioinformatic tools and online data and included analyses of the yeast genome, HIV mutations, phylogenetic studies of 16S rRNA sequences, and comparative genomics activities.^{118–120} Protein structure was addressed by Centeno *et al.*³¹ who described a scheme for teaching structural biology using computational approaches in 2003, and

Bednarski *et al.*⁶² in 2005, who designed an inquiry-based lab around how protein structure and function impact genetic diseases.

An early effort to introduce bioinformatics analysis to undergraduates was described in the 2008 study by Drew & Triplett.⁶⁶ This pilot course featured whole genome sequencing of a bacterial species, with the objectives of teaching students about genomics and providing an avenue for original research. Unlike many previous studies, Drew & Triplett included an assessment of their program and reported positive outcomes, including improved knowledge, laboratory skills, and understanding of sequencing technologies among students. The study emphasized the benefits of integrating advanced technology into undergraduate education. No assessment of shifts in student cognitive outcomes was measured in this study.

A program that employed a similar protein characterization method as the CODE pathway was highlighted in the paper by Badotti *et al.*³⁸ This work described a bioinformatics educational approach using comparative modeling of proteins with a class of 33 undergraduate, graduate, and postgraduate students in Brazil. The authors discussed a pedagogical method that aimed to make bioinformatics tools more accessible and interesting to students. The project employed comparative modeling techniques to predict the 3D structures of selected proteins, providing students with practical experience in using bioinformatics tools. The authors emphasized the importance of involving students in hands-on activities and problem-solving exercises to enhance their understanding of bioinformatics concepts. The benefits of this approach were described as increased student engagement, improved comprehension of complex bioinformatics topics, and the development of valuable computational and analytical skills. The students were evaluated

on a seminar presentation graded by their peers and the professor. The only assessment of this program consisted of a ten-question questionnaire, the first four regarding the students' previous knowledge and the last six regarding their learning. Fifteen students were randomly selected from the class to complete the form anonymously. The results indicated that students were able to improve their theoretical knowledge and practical skills in comparative modeling during the course, although no formal statistical analysis was performed. This hands-on educational method is similar to the techniques used by CODE projects. Unlike the CODE study, the assessment did not address the student attitude constructs and was presented to a small audience that included mainly graduate-level students. Additional work related to this program has not been published.

The Genome Solver Project (GSP), described in the 2019 paper by Mathur *et al.*⁵¹ has many similarities to CODE. Much like CODE, a primary feature of the initiative was faculty training to improve proficiency in teaching bioinformatics concepts and the use of analysis tools. GSP also designed a curriculum emphasizing problem-solving and hands-on data analysis using real-world research questions and tools commonly used in genomics and bioinformatics. The project focused on comparative microbial genomics for undergraduates and was commonly structured as a CURE. The program no longer provides training workshops, but online curricula materials are available. The activities provided by the GSP were designed to introduce bioinformatics tools and techniques but were not focused on generating new scientific findings, one of the key aspects of a CODE project. Faculty workshop participants (n=277) were surveyed to measure the effectiveness of faculty training. The results concluded that the bioinformatics training workshops effectively encouraged faculty to engage in bioinformatics instruction.

Students at five schools (n=640) were given pre- and post-participation surveys to measure knowledge gains, which showed increased student performance. No assessments were conducted to measure impacts on students' attitudes pre- and post-participation.

A very successful initiative to integrate bioinformatics techniques and research into undergraduate classrooms is the Genomics Education Partnership (GEP).^{85,121–125} This active program trains faculty to lead their undergraduate students in the annotation of eukaryotic genes. The program has integrated genomics research into over 200 institutions and provided thousands of students with authentic hands-on research experiences. The GEP offers a variety of research projects that students can participate in, allowing them to contribute to ongoing genomics research while learning key laboratory techniques and data analysis skills.

Much like CODE, the GEP provides faculty training and involves a wide range of institutions, including small colleges, liberal arts universities, and larger research universities. The program has expanded since its inception in 2006 and offers several different projects for students to contribute to, including research related to drosophila and parasitoid wasps. A CURE version of the program was assessed to evaluate its effectiveness in achieving educational objectives using surveys, interviews, and analysis of student work.¹²³ The study reported positive outcomes, including increased student engagement, improved understanding of genomics concepts, and the development of research skills. An assessment of the attitudes of students engaged in a GEP CURE was described in the 2020 paper by Lopatto *et al.*¹²⁶ The assessment used scales to measure epistemic beliefs about work and science and interest in science. Their findings showed that students who entered the class with more positive attitudes toward science showed

greater learning gains and self-reported benefits. The GEP program is a strong model for CODE in terms of its robust curriculum, faculty network, and student engagement model. Several studies have shown that GEP students experience increased engagement, improved understanding of genomics concepts, and the development of research skills. However, the assessment of impacts on science self-efficacy, identity, and intention to persist in science have not been examined in this program

2.10 Improving Student Knowledge and Confidence Through Instruction in Bioinformatics

A 2007 study that integrated the study of bioinformatics into the undergraduate life science curricula was described by Howard *et al.*⁶⁸ While the authors did not seek to examine shifts in psychosocial measures, they did use a program of three different assessment tools (student self-assessment of learning, content exam, and faculty survey) to evaluate the integration of the computational modules, tools, and resources into existing courses at the university. The paper reported positive outcomes from the assessment, indicating that students who participated in bioinformatics activities demonstrated an improved understanding of computational biology concepts and data analysis skills. Additionally, they expressed increased interest and confidence in using bioinformatics tools. Some of the survey questions for the CODE project were drawn from this previous work because the questions most closely evaluated the type of bioinformatics practices conducted during a CODE project. In particular, the questions in this paper included statements about protein domains, the three-dimensional structure of a protein, and genetic mutations responsible for disease.

Similarly, researchers Wightman and Hark explored the integration of bioinformatics into multiple levels of undergraduate biology courses and its influence on

students' mathematical skills and confidence in their 2012 paper.⁷² The program implemented a survey of student attitudes and a direct assessment of student performance on specific bioinformatics and mathematical skills. The survey encompassed 14 questions, prompting students to self-assess their confidence levels and comprehension of the math and bioinformatics skills. Subsequently, it included ten multiple-choice questions designed to directly evaluate their knowledge in both domains. The surveys were administered at three points in the semester (pre-, mid-, and post-learning). The results showed increases in student knowledge and confidence in performing bioinformatics activities. A key conclusion of the authors was that introducing students to bioinformatics concepts early in their careers is effective and results in significant learning gains. This study was primarily focused on assessing the impact of integrating bioinformatics on students' mathematical skills and did not include additional assessment measures or faculty training in bioinformatics techniques.

Another study that describes the design and implementation of an undergraduate module designed to teach applied bioinformatics to biology students is presented in a 2018 paper by Madlung.⁴⁹ The author integrated bioinformatics activities into multiple courses with topics that ranged from NCBI reference databases to command-line training for RNA sequence analysis. This work highlighted positive student outcomes, including improved knowledge and skills, and emphasized the importance of adapting teaching methods to accommodate diverse student backgrounds. The assessment found that students gained confidence in their ability to master the bioinformatics tools. The module underscored the interdisciplinary nature of bioinformatics and its significance in modern biological research. Madlung also highlighted the importance of preparing educators to

teach computational skills. Again, this assessment did not examine the factors the CODE assessment is focused on. Still, the program similarities make it a valuable resource for improving the CODE curriculum modules accessible to novice faculty members.

2.11 Studies with Assessment of Student Cognitive Areas

A recent study assessed several psychosocial correlates of academic and STEM success in a population of entering freshmen at a short co-curricular STEM Academy.¹²⁷ The authors were particularly interested in determining if a weeklong intervention could impact areas related to STEM retention, such as science identity, self-efficacy, sense of belonging to the university, career expectancies, and the intention to remain in a STEM major). Using pre-and post-surveys, the researchers found the participants significantly increased their sense of science identity and sense of belonging to STEM and the university. The authors suggest that these impacts are predictive of increased STEM retention and show that the student cohort showed 98% first-year retention and 92% STEM major retention and plan follow-up studies to assess the longitudinal impacts of this intervention. This study is of note because similar to CODE, it assessed students with an interest in STEM. This study examined a very short intervention of only one week. While some CODE facilitators integrate the project into a genetics or molecular biology class for 2-4 weeks, many other facilitators implement CODE as a semester-long course or independent study. A longer-term study looked at participants in an 8–10-week science or engineering internship or a year-long research study program.¹²⁸ Their research showed that science support experiences, such as mentoring, and integration into the scientific community, drive the formation of science self-efficacy and identity. While neither of these studies included bioinformatics education, it is encouraging to see the

positive results on students' self-efficacy and science identity development from co-curricular STEM activities and support.

Cognitive traits that are known to support interest in STEM careers were the focus of a study of high students who engaged in a bioinformatics class.¹⁰² The program developed two units focusing on genetic testing and research to teach bioinformatics concepts. Extensive teacher training and diverse resources were provided to support instructors in integrating the materials into the classroom. Introductory unit students (n = 289) showed significant gains in awareness, relevance, and self-efficacy but not engagement, while advanced unit students (n = 41) improved in all four cognitive areas. A key focus of the program was the use of bioinformatics to increase student interest in STEM careers, particularly among students who may have yet to consider a career in STEM. Like the philosophy of CODE, the authors suggest that teacher professional development is essential for successful implementation. The assessment results in this program, also similar to the CODE findings, showed gains in student awareness and self-efficacy, which point to the potential advantages of introducing bioinformatics education at the high school level for promoting STEM careers.

Faculty were trained to integrate bioinformatics techniques to analyze phage genomes in the successful program outlined in the 2014 paper *A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students*.⁷⁴ The Science Education Alliance Phage Hunters Advancing Genomics and Evolutionary Science (SEA-PHAGES)¹²⁹ course is ongoing and aims to increase undergraduate interest and retention in the biological sciences through authentic research experiences. The program incorporates lab work with bioinformatics analysis of

microbial species found in soil samples, with an emphasis on the publication of student findings. An assessment of the program included questions from the Survey of Undergraduate Research Experience (SURE) survey¹³⁰ and the Classroom Undergraduate Research Experience Survey (CURE).¹³¹ The surveys sought to measure the self-perceptions of learning gains, motivation and attitude, and career aspirations of the SEA-PHAGES course participants. The results were compared to surveys of students in a summer research experience and students in a traditional science course and revealed that SEA-PHAGES students had improved learning gains and increased scientific self-efficacy. Positive influences on student retention were also attributed to program participation. This nationwide program expertly models methods of faculty training, building a faculty network, and student engagement in undergraduate research that are also found in the CODE program.

The paper titled *Gains in Scientific Identity, Scientific Self-Efficacy, and Career Intent Distinguish Upper-Level CUREs from Traditional Experiences in the Classroom* by Newell *et al.*¹³² published in 2022 explored the distinctions between upper-level course-based undergraduate research experiences (CUREs) and traditional classroom-based experiences in terms of their impact on students' scientific identity, scientific self-efficacy, and career intentions. These students participated in authentic research projects, similar to CODE participants. However, the focus of the projects was not bioinformatics but zoology. The study authors assessed how participation in this CURE affected students' scientific identity, self-efficacy, and intentions to pursue careers in scientific fields. They compared the results to those in traditional classroom-based settings. The paper reports that students in the research-focused CUREs (n=182) demonstrated higher

gains in scientific identity, self-efficacy, and career intentions than their peers in non-CURE settings. The study suggests that UREs can strengthen the relationships among these variables and can be used to maximize the likelihood of students progressing through the STEM pipeline. Like the CODE assessment, this study contributes to the literature by quantifying the gains associated with undergraduate research participation. The study utilized the Tripartite Integration Model of Social Influence (TIMSI) survey instrument, measuring students' scientific self-efficacy, scientific identity, value orientation, mentorship, and career intent. The TIMSI uses a lens of social influence to analyze STEM students' persistence and reveal influencing agents that affect persistence and motivation in academic environments.^{20,133} This measure was more extensive than the survey instrument used for CODE but contained similar elements. The inclusion of a control group was a valuable addition to this study.

This endeavor follows the example of other bioinformatics-focused undergraduate research programs, such as Genome Solver, GEP, and SEA_PHAGES, that seek to create networks of educators who are incorporating bioinformatics and genomics curricula into their classes.^{51,74,75,85} These programs stimulate students' interest in science, positively influence academic achievement, and enhance persistence (STEM) disciplines.

2.12 Summary

A review of the literature shows that multiple studies have demonstrated the value of incorporating UREs into the curriculum. Increased participation in authentic research can have far-reaching effects, including improvements in science self-efficacy, encouragement of STEM career persistence, enhanced student comprehension of science content, and increased graduation rates.^{26,134–136} Nevertheless, a significant portion of

these investigations have mainly concentrated on laboratory-based encounters, while others have just examined enhancements in student learning and, occasionally, confidence levels.

The primary objective of the CODE study was to investigate the impact of a discovery-based introductory-level bioinformatics project on pupils. The study sought to examine the shifts in student attitudes towards bioinformatics and their self-perceived levels of science self-efficacy and identity as a scientist, two constructs that can positively influence retention in STEM majors and careers. It has been demonstrated that bioinformatics research experiences positively affect the collaborations, research, publications, and career advancement of undergraduate students.^{82,137} Still, the question remains if computational biology research can also impact student interest, motivation, and retention. Can a bioinformatics-focused research project motivate students in the same ways seen with laboratory-based projects and CUREs?

Few recent papers directly address how participation in a bioinformatics-specific URE impacts students. Instead, the literature is more focused on providing examples of modules and lessons that can be implemented in classes,^{59,63–65,67,69} and sometimes assessing the students' knowledge gains or enjoyment of these implementations.^{61,62,68,70,72,73} Some large projects have been designed to provide training, curricula, and resources to support genomics and bioinformatics investigation at the undergraduate level.^{51,74,85} While these projects have not specifically assessed for constructs that drive persistence gains, they have been shown to generate student interest and knowledge gains. Determining effective methods for supporting student growth in

these constructs while supporting their training in bioinformatics skills will build a more robust workforce to support the STEM needs of the future.

Chapter 3. The Program Intervention

As described in Chapter 1, there is a need to increase student persistence in STEM majors and develop student bioinformatics skills. Studies show that undergraduate research can positively impact students' confidence in their ability to succeed in science (self-efficacy) and their scientific identity, leading to retention in science majors. The Characterizing Our DNA Exceptions (CODE) program was conceived to introduce the emerging field of bioinformatics to undergraduates while fostering their engagement through an authentic research project. Guided by the objectives of bolstering the bioinformatics workforce and offering accessible research opportunities, this ongoing initiative entails faculty training, resource provision, and showcasing student accomplishments through an annual conference. Bolstered by support from the NSF Improving Undergraduate STEM Education: Education and Human Resources grant (Award #2120918), the program has expanded its reach and faculty training and embarked on a research study to evaluate student engagement. CODE projects offer a potentially transformative option to provide an authentic bioinformatics research project to institutions lacking the funding for traditional wet lab research projects.

The CODE project aims to increase access to research-based learning experiences, particularly for students at small institutions with fewer undergraduate research opportunities. The program seeks to introduce a broad range of students to bioinformatics in a mentored project-based format and achieve the following:

- Increase awareness, interest in, and knowledge of bioinformatics.
- build confidence, self-efficacy, and scientific identity.
- encourage continued participation in science education and a STEM career path.

Few studies have evaluated how bioinformatics research can impact students' interest in the field and their intention to persist in science. The research assessment described in this work measured these variables in a wide range of students engaging in CODE projects at multiple colleges and universities in the Southeast. This chapter traces the program's origins, detailing its progression and presenting insights into facilitator training sessions, project implementation strategies, and the transformative nature of CODE Student Symposiums in cultivating participants' scientific identities.

3.1 History of CODE

What is now known as the CODE project originated in 2016 from a project initiated by Jeremy Prokop, Ph.D., with Athens State University and Sara Cline, Ph.D., to examine genomic variants using protein modeling and conservation analysis. Dr. Prokop brought the project to the Educational Outreach team at HudsonAlpha Institute for Biotechnology with a request to build out an initial data repository and expand the collaboration with more educational partners. This request grew into a more coordinated initiative that secured funding from the Alabama Power Foundation for two years. The pilot project kicked off in 2018 with five Alabama colleges to engage small groups of undergraduate students in authentic genomic research to sustain STEM interest and participation. Facilitators from each school were trained in the methods and tools needed to lead their students through a computational exploration of DNA variants of uncertain

significance (VUS) to determine their biological relevance and explore topics of their own interest.

The program expanded over the next two years with additional facilitator training workshops that introduced new faculty members to the research projects and brought new colleges and universities into the CODE community. In October 2021, the program was awarded funding from the National Science Foundation (NSF) Improving Undergraduate STEM Education (IUSE) initiative to support three years of continued development and expansion. Additional schools implemented CODE projects with their students as additional facilitators were trained at the semi-annual Facilitator Training Workshops. An annual Student Symposium began in 2019 to provide opportunities for students to share their findings and experiences. Resources and materials to support faculty and students were created and shared with the assistance of an advisory board of experienced CODE Facilitators. The assessment project began in the fall of 2021 with pre- and post-student surveys to measure the impacts of CODE participation. This study describes the measurement of two years of survey data, but the assessment project will continue and expand as the program grows.

3.2 The CODE Project Pathway

CODE projects lower the barriers to engaging in bioinformatics research by placing genomic analysis within the grasp of a broad and diverse audience. A typical project path mirrors the approach taken by clinical research analysts,¹³⁸ beginning with database research, followed by modeling and simulations, and culminating in data analysis as shown in Figure 3.1 CODE project pathway.

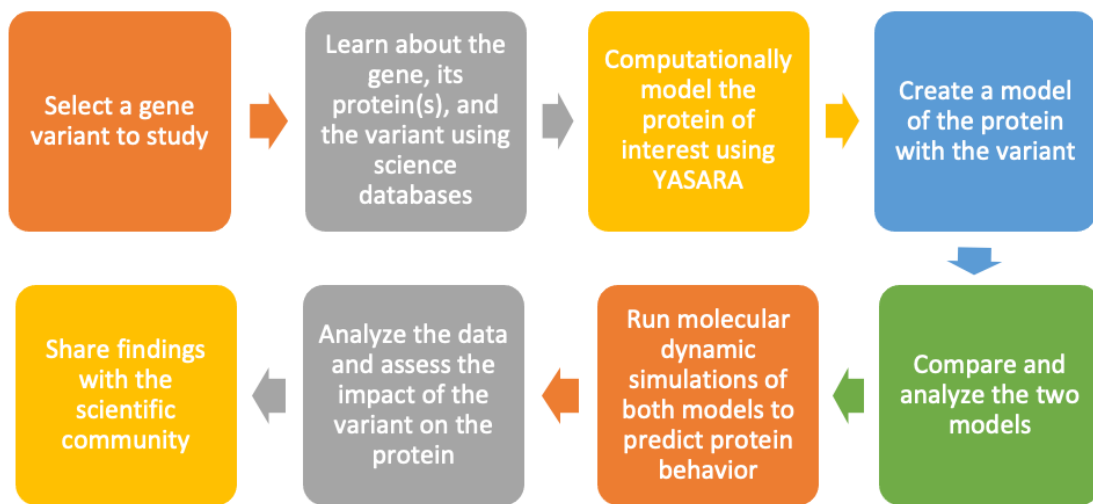


Figure 3.1 The CODE project pathway. Each CODE project is unique, but this flowchart outlines the most common path of a project.

Current sequencing technologies make it possible to obtain the entire genetic code of an individual or agriculturally important plant in a matter of days. The millions of genomic changes typically identified in these genomes must be examined and interpreted to detect the handful of variants with significant clinical or agricultural consequences. Often the process detects DNA variants that are poorly understood because they have not been studied previously. A genomic VUS must undergo extensive analysis and testing to determine if it has a functional role in a trait or disease. The VUS interpretation process is critically important yet can be painstakingly slow. Enormous amounts of genomic data

are being generated in clinical and research laboratories, meaning the number of variants awaiting deeper categorization continues to grow.

Students first identify a genetic variant of interest. They may begin with a particular disease or condition that interests them or a specific gene. The CODE project also provides students with a list of variants identified in HudsonAlpha Institute research projects. All personally identifying data are removed from variants found in human sequencing studies before being placed on the selection list. Participants can select a VUS from this curated list and contribute directly to the ongoing research studies at the Institute.

Students with more specific interests may identify and select a VUS from a publicly available genome database such as ClinVar.¹³⁹ Students use publicly available databases and relevant software to learn more about the gene with the variant and its functional product. To fully understand the biological functions of a protein, researchers need to know more than just the amino acid sequence. Students compare the amino acid sequence with similar segments from other organisms to study the evolutionary conservation of both DNA and protein. This information helps make predictions regarding the functional consequence of the DNA variant.

CODE project participants use software and protein-structure databases to visualize how a single-point mutation can change the molecule's three-dimensional conformation. These nucleotide-level variants can cause structure changes that lead to life-threatening diseases. Learning this information in a hands-on context, and mastering these computational skills, is far more powerful than a traditional lecture format or reading a textbook.¹¹⁷ Students employ molecular modeling to study how the VUS may

alter the protein. This type of molecular visualization provides essential support for formulating hypotheses related to molecular structure.¹⁴⁰ Current software tools such as YASARA, PyMOL, and UCSF Chimera allow students to build a 3-D model of a protein, insert a variant, and visualize the impacts of the variant change on protein folding and structure. Students then study the DNA variant in a computationally derived cellular environment, running molecular dynamics simulations (MDS) to predict how the variant-containing protein might behave inside a cell. This type of analysis can facilitate understanding the variant's effect and even provide insight into interventions to offset potentially damaging impacts.

Students document their work and findings about their genetic variant in a written report submitted to the CODE team. Reports of their results are shared with researchers to support ongoing genomic variant analysis. With student permission, research findings are also included on the CODE website to share with the scientific community. All students are encouraged to include their data files in a CODE repository library for other researchers to use. This data may consist of protein models, variant conservation analysis data, molecular dynamic simulation data, and their conclusions. A clearly defined policy of authorship for resulting scientific publications and intellectual property rights has been established, designed to recognize the significant role students and advisors play in contributing to the scientific body of knowledge.

CODE projects are unique in that students characterize the molecular structures of actual genomic variants identified by research studies and cataloged in publicly available scientific databases. Throughout this process, students are introduced to and gain experience with ten key bioinformatics concepts (Table 3.1), growing their research

skills. Students create hypotheses about how the variant may impact a protein's function, contributing to the general knowledge base around genomic variation.

Table 3.1 Bioinformatics techniques utilized by CODE projects. Ten key bioinformatics concepts are introduced to students during a CODE project.

Key concept	How students experience the concept
Database mining	gather information from databases such as UniProt, Ensembl, Protein Data Bank, ClinVar, Cosmic, NCBI, PubMed, OMIM, gnomAD, Varsome, ExPASy, Ensembl and dbSNP
Protein modeling	use homology modeling, <i>ab initio</i> , and threading techniques
3D protein structure and function analysis	examine variant impact on protein structures with powerful modeling tools
Multiple sequence alignments	identify similar regions in other species
Evolutionary conservation analysis	compare homologous sequences to determine functional and structural residues
Utilize open-source analysis tools	PyMol, Chimera, ESPript, BLAST, MEGA, ConSurf, STRING, and PolyPhen-2, and others
Protein molecular dynamics simulations (MDS)	simulate how the protein will move in a cellular environment; estimate the impact of a variant on the structure and behavior of the protein
High-resolution image and video creation of protein models	produce publication-quality materials of molecular models
Determine the quality of results	understand and utilize modeling results such as z-scores, TM-scores, and other statistics

Key concept	How students experience the concept
Data analysis and reporting	analyze, graph, and report findings from the molecular dynamics trajectory analysis, a dynamic cross-correlation matrix, and other results

Authentic learning techniques are the most effective way to learn and are highly motivating to students.¹⁴¹ CODE research projects encapsulate the ten design elements of an authentic learning experience:

1. grounded in real life and work
2. investigating a challenge that is ill-defined and requires students to identify the steps of the project
3. requiring a significant investment of time and intellectual resources
4. involves synthesizing data from a variety of resources
5. requiring students to learn and use skills that are expected of practicing scientists (e.g., technology, teamwork, problem-solving)
6. allowing the students to reflect and determine what the information they have acquired means
7. requiring students to apply knowledge from one or more disciplines or content areas
8. assessment of findings is conducted throughout the research, not just at the end
9. requiring students to present a finished product to justify their results with evidence from experiments or research
10. the study allows for diverse interpretations and competing solutions

Since students can be reluctant to venture into bioinformatics, the CODE project intentionally focuses on introductory-level research and developing students' skills. Although coding is a relevant and needed skill, the intimidation factor and learning curve may seem prohibitively high for some students, preventing them from starting down that career path. The CODE program emphasizes using open-source databases and tools rather than learning a coding language. Madlung argues that life scientists can use methods and programs developed by others—and apply them to novel research questions without having to be experts in the underlying disciplines in addition to biology and their research specialty.⁴⁹ CODE projects provide students with an entry point to the field, integrating user-friendly tools and graphical user interfaces that gradually build computational skills required for more advanced techniques, such as command-line interface.¹⁴² The CODE program works to overcome student reluctance and create a growing and dynamic population of students excited about incorporating bioinformatics into their education and career plans.

3.3 CODE Facilitator Training Workshop

Knowledge of bioinformatics techniques is becoming indispensable for biologists, yet incorporating these skills into the undergraduate curriculum is a slow process.^{34,49,50} Studies point to the fact that instructors are uncomfortable teaching bioinformatics concepts because they lack training, are unaware of available tools and resources, or are unsure how to use them, particularly at HBCUs.^{51,117,143} While faculty overwhelmingly agree that bioinformatics should be part of the curriculum, they need training, support, and robust protocols to overcome the difficulties of integrating these complicated concepts into their course load.^{34,50} Due to the rapid pace of genomic advances, many

instructors have not had the opportunity to be trained in computational techniques to manage, analyze, and learn from the massive amount of data available.

There is a need for bioinformatics activities that surmount the barriers faced by undergraduate faculty. Introductory-level programs, such as CODE, can prepare faculty to take on the challenge of introducing bioinformatics research experiences to their students. CODE is purposefully designed to be accessible to the introductory level bioinformatics student and reveal the world of bioinformatics and genomics research in an inquiry-based format that taps into student interests and allows them to shift their project in the directions that interest them most. The project sequence was deliberately designed to be user-friendly with enhanced algorithm development to minimize the number of demanding steps students must accomplish to reach results. Other “protein structure-to-function” courses have students navigate through a similar set of processes to understand the concepts,³¹ but that approach is particularly challenging for educators without a background in bioinformatics. As mentioned above, one of the primary factors limiting the inclusion of bioinformatics in the undergraduate curriculum is the reluctance and limited experience of the faculty.^{50,51,143} The CODE program allows faculty and students to quickly see progress in their goal to characterize genomic variants using a limited set of software tools to build high-quality protein models, visualize structure alignments, and run simulations.

Professional development for faculty is a crucial component of the CODE program. Faculty facilitators learn to use genetic databases during a two-day in-person workshop (or virtually if in-person training is not feasible). Although participants have expressed interest in teaching bioinformatics, no prior knowledge of bioinformatics is

required. The NSF IUSE grant funds registration, meals, and lodging. The goals of the Facilitator Training Workshops are to introduce professors from small colleges, regional universities, and historically black colleges and universities (HBCUs) to the methodology of genomic variant interpretation using computational biology tools and to equip participants with the skills, resources, and confidence to launch bioinformatics or computational biology research projects for their students.

Training sessions guide facilitators through the steps of a CODE project, providing them with the resources needed to computationally characterize a genomic VUS. The training workshop includes presentations on genomic variant classifications and various online databases to investigate genes, diseases and conditions, and their associated variants, homology modeling, databases and software tools, molecular dynamics simulations, variant impact prediction tools, genomic reference databases, data analysis methods, and implementation of student research projects.

Each CODE school is provided with a two-year group license to YASARA Structure, a modeling and analysis program.¹⁴⁴ Hands-on exercises familiarize facilitators with the YASARA Structure protein modeling software application. Participants learn how to locate and download a protein structure using UniProt and Protein Data Bank (PDB) and how to run a homology modeling experiment based on experimental data to create a model of a protein. Pre-computed data allow them to analyze the results of the modeling. Additional resources such as AlphaFold and ITASSER are introduced to provide predicted protein models if needed.

Each section begins with a short overview to provide the foundational biological concepts for the activity and an introduction to databases or software tools. Participants

work individually through guided hands-on analysis of provided protein sequences or models. Protocols are provided for each activity that can be used with any protein and variant(s), allowing the faculty to experience the same challenges student researchers face as they work with these evolving tools and databases. Each activity is provided to the facilitator and students in a manual containing background information, a detailed listing of analysis steps, and screenshots to ease comprehension. See Appendix A for an example training workshop schedule.

Support and training continue after the workshop through emails, virtual calls, website resources, and campus visits. CODE provides ongoing technical support from a dedicated technical advisor, available by email, phone, or video call, which can be particularly helpful for overcoming issues and roadblocks that might be unique to a school's project.

Facilitator training workshops are held twice a year. As of spring 2023, 70 facilitators have been trained at CODE workshops (66 from colleges, three from non-profit educational programs, and one high school teacher). This group was almost evenly split between males (44%) and females (56%). Approximately 40% of the group comprised faculty from underrepresented populations. Table 3.2 shows the demographic data for all of the facilitators that have been trained by the CODE program as of October 2023.

Table 3.2 Demographics of CODE facilitators. Seventy facilitators have been trained to date by the CODE program. The demographic data for the faculty are shown in this table.

All Trained CODE Facilitators (N = 70)	
Gender – N (%)	
Female	39 (56%)

All Trained CODE Facilitators (N = 70)	
Male	31 (44%)
Prefer not to answer	0
Race/ethnicity – N (%)	
Asian/SE Asian	9 (13%)
Black/African American	15 (21%)
White/Caucasian	42 (60%)
Two or more races	2 (3%)
Other	1 (1%)
Prefer not to answer	1 (1%)

These 70 facilitators represent 34 colleges and universities. The Carnegie classifications of the schools range from Associate’s Colleges to Doctoral Universities and Four-Year: Medical Schools & Centers and are depicted in Figure 3.2. The COVID pandemic impacted all aspects of the CODE program, especially training workshops. The two 2020 workshops and the spring 2021 session were canceled. Training resumed in December 2021. Two workshops were held in 2022, and a summer session in July 2023. Additional workshops are scheduled for December and July of each year going forward.

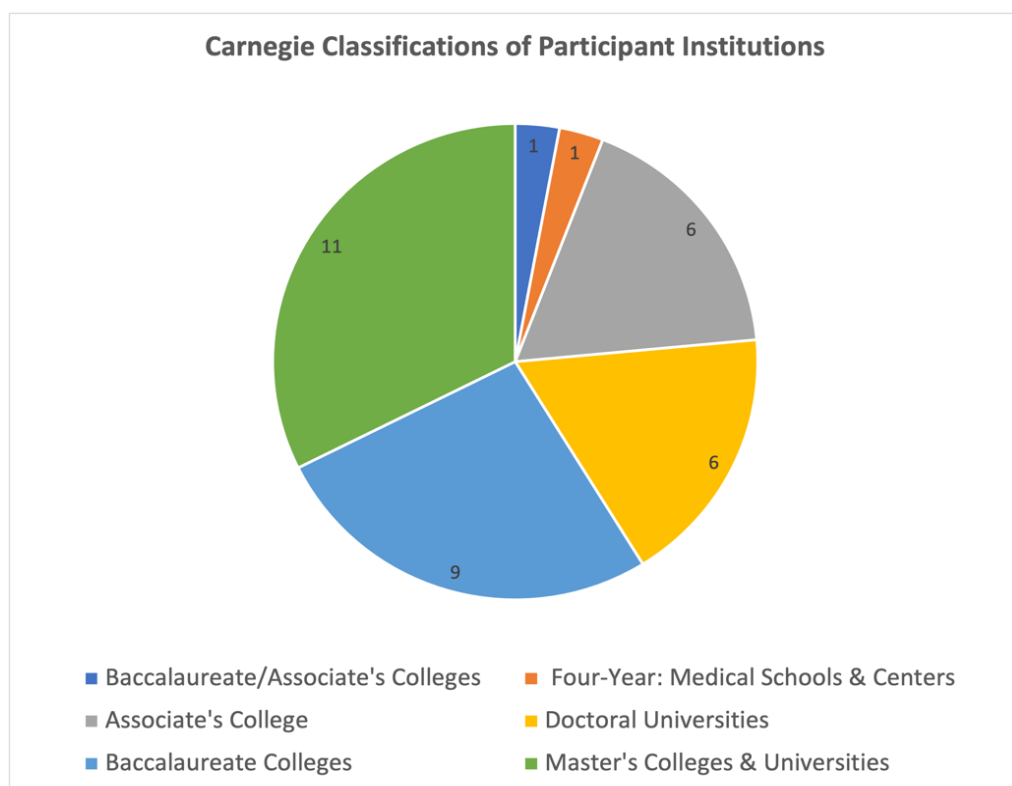


Figure 3.2. Carnegie classifications of participant institutions. Thirty-four institutions have participated in a CODE Facilitator Training Workshop. This figure depicts the Carnegie classification of each institution.

3.4 Implementation by Facilitators

A computer-based research experience has several advantages over the more traditional wet lab experiment. Facilitators only need to provide a computer and internet access for student projects. The YASARA Structure modeling software is provided by the CODE program, but a free version and other options are also available. These *in silico* projects have no major lab safety issues and lend themselves well to group work and collaboration. Data gathering and analysis from the open-source resources used by CODE is free, unlike the many associated costs of bench work, such as reagents, materials, and equipment.⁸⁵ Facilitators can structure CODE projects for a larger number

of students than can usually be effectively mentored in a research lab. All students can be taught a common set of techniques that they can then adapt to their unique research project.

Faculty mentors are encouraged to incorporate the elements of a CODE project in whatever format works best for their circumstances. This flexibility has resulted in short modules (2-4 weeks) embedded in biology, genetics, and other classes, dedicated courses based on the CODE pipeline, independent projects, and a weekly bioinformatics club meeting. Since the program's inception, over 500 students have been mentored in CODE projects at over thirty institutions. CODE projects and findings have led to peer-reviewed publications, including recent work exploring protein interactions between the SARS-CoV-2 virus and human extracellular protein receptors.^{145–148}

Facilitators are sometimes challenged in securing sufficient computer resources to run MDS for student projects. It is possible to run the MDS in YASARA Structure on a typical laptop computer, but the analysis can take up to two weeks and uses computer processing power that can potentially slow other applications. To address this issue, the CODE project partnered with HudsonAlpha's Information Technology department to provide cloud resources for running MDS. The team utilized Amazon Web Services (AWS) and deployed a private instance of YASARA Structure. AWS provides on-demand computing and storage, reducing the expense of computing resources. Students and faculty provided the CODE team with their protein structures, wild-type, and a structure containing the VUS. After the MDS was run in the cloud, the resulting data files were populated to a shared repository for analysis.

3.5 Annual CODE Student Symposium

Student and faculty participants are invited to attend an annual one-day symposium to showcase their work, interact with other participants, and learn from researchers in the field. All participating students, regardless of their level of progress, are encouraged to share their experiences and findings through a project poster, with selected students invited to give an oral presentation. A juried poster competition awards monetary prizes to top presentations. Additional sessions with scientific researchers and bioinformatics experts spotlight new findings and career paths. During breaks and meals, networking sessions provide further opportunities for students and faculty to learn what techniques, methods, challenges, and successes others have experienced in their research.

The first Student Symposium in 2019 welcomed 36 registered attendees from 10 colleges. Four students provided oral presentations, and 21 participants showcased their research with posters. Virtual symposiums were held in 2020 (17 student presenters and > 300 visits to the event) and 2021 (23 presenters and > 600 visits to the event) due to COVID-19 pandemic restrictions.

Forty-four attendees participated in the Symposium in April 2022 in the wake of lingering concerns around COVID-19 spread. Seven students gave oral slide presentations, and 26 shared their research in the poster session. In the spring of 2023, the Student Symposium expanded to a larger venue and audience. Approximately 100 students and faculty attended the event, along with ~50 public visitors to the student poster session. Seventy-two students presented their research from eight colleges and universities.

3.6 Quotes from Student Symposium Attendees

- *All of the speakers were interesting, and I appreciated the different perspectives and backgrounds that they provided. I also really enjoyed getting to see what all of the various student groups have been working on.*
- *I gained knowledge on different ways genetics is used to help people. I also gained a lot of self-confidence in my research by presenting my poster. It was really nice to get feedback from professionals about the work I did.*
- *The experience of presenting a poster is incredibly valuable and a skill that this symposium allowed us to develop as a group in a less competitive environment.*
- *I really enjoyed being able to present my poster to a knowledgeable audience. It helped me learn more about my project through exposure to other projects and professionals in that area.*
- *The diversity of thought processes, ethnicities, and educational backgrounds was enriching.*
- *I've gained a lot of insight into the world of different biological careers, and it's given me a deeper understanding than what I knew beforehand.*

- *I learned so much about careers in genomic-related fields. This was extremely useful information. I was also given the opportunity for valuable networking. It also allowed me to present to people who are in the field I want to enter.*
- *So well done. Thanks ... for all your hard work. The food and location were excellent. Judges were kind and challenging (always a hard balance). The whole experience was so supportive and friendly. My students had a blast and they're strutting now (-:*

3.7 Advisory and Curriculum Committee

A five-member Advisory and Curriculum Committee guides the CODE program. It works to develop a curriculum to enable new facilitators to implement CODE projects easily and quickly at their institutions. While faculty are encouraged to adapt the program to their needs, a vetted set of curriculum materials provides a solid foundation for new facilitators. The Committee members share examples of how they have implemented the project in their classes and collaborate to define a set of course background materials and protocols. Their experience and insight are invaluable in developing the most accessible lessons. The Committee's recommendations guide the creation of lessons with examples, images, and relevant links.

3.8 Summary

At the heart of every science education research assessment is the participant intervention measured by that assessment. The Characterizing Our DNA Exceptions (CODE) program aims to enhance student persistence in STEM majors and develop students' bioinformatics skills. There is a clear need for increased undergraduate student

engagement in authentic research experiences to boost confidence and scientific identity, ultimately leading to higher retention in STEM disciplines. The CODE program, supported by an NSF grant, is a model for introducing bioinformatics to undergraduates through mentored research projects. It emphasizes faculty training, resource provision, and student accomplishments showcased at an annual conference.

The CODE program has grown from a pilot project to a multi-institutional initiative. Students engage in authentic research by identifying genetic variants of uncertain significance, studying their biological relevance, and simulating their impact on proteins. This hands-on approach helps students develop key bioinformatics skills. An important element of the program is the faculty development and training, addressing the barriers instructors face in teaching bioinformatics. The CODE program offers faculty training workshops and ongoing support to ensure instructors are equipped to guide students effectively.

The annual CODE Student Symposium provides a platform for students and faculty to present their research, network with peers, and learn from experts in the field. The symposium encourages student engagement and showcases their findings. The entire CODE program is guided by the Advisory and Curriculum Committee, which plays a crucial role in developing curriculum materials and providing guidance for facilitators.

The CODE project intervention was assessed to measure its impacts on the participant researchers. Chapter 4 will comprehensively describe the methods used to complete the assessment. A thorough account of the surveys employed, and the analytical techniques applied to gauge the intervention's impact will be presented.

Chapter 4. Methodology

This study addresses four questions: 1) Does participating in a CODE research project increase bioinformatics awareness, interest, and knowledge? 2) Does participation increase students' self-efficacy? 3) Does participation increase students' scientific identity? 3) and 4) Does participation increase persistence in STEM? To answer these questions, the researcher quantified student-self-reported levels of these constructs from 14 colleges and universities. This strategy involved the collection and examination of both quantitative and qualitative data. As Creswell and Plano Clark point out, quantitative and qualitative methods can collaborate to provide a more comprehensive picture of the research issue.¹⁴⁹ This chapter includes a description of the participants, the assessment measures, and the data analysis plan relevant to the hypotheses being tested in this paper.

4.1 Quantitative Research - Design of Student Survey Instruments

To appropriately address the research questions, the author selected validated evaluation instruments from previous research studies to assess student research experience, science self-efficacy, scientific identity, bioinformatics interest, and persistence in STEM. To evaluate the first three of these constructs, a scale of psychosocial measures was drawn from Chemers *et al.*²⁴ with adaptations by Robnett *et al.*²⁵ whose study highlighted the importance of undergraduate research in enhancing science self-efficacy. These scales use a Likert rating system in a pre-/post-survey format to examine changes in students' self-reported ratings about research experience, science

self-efficacy, and identity as a scientist. Originally, the survey items for research experience, science self-efficacy, and identity as a scientist were implemented with a 5-point scale.²⁴ For each of these categories, the Robnett *et al.*²⁵ study employed psychometric analyses that indicated that the 5-point scale should be converted to a 4-point scale. These 4-point scales were used in this study. Four-point scales force respondents to choose a direction rather than providing a neutral response.¹⁵⁰

Statements to evaluate interest in the field were modified from the latest version of the Student Assessment of Learning Gains (SALG).¹⁵¹ Because the SALG is primarily a post-intervention instrument questions about the project's impact on attitudes were reframed for the pre-/post- format used in the CODE survey instrument. This section of the student survey was composed of three statements. Students rated how well the statement applied to them using a 5-point scale that ranged from Strongly Disagree to Strongly Agree.

To measure the participants' persistence in a STEM major and their commitment to a science career, statements from two validated instruments were combined to thoroughly cover both areas.^{108,128} Because CODE projects included students from interdisciplinary fields, the *intention to leave science* measures¹⁰⁸ were adapted to the *intention to leave STEM*, as exhibited in the study by Kuchynka *et al.*¹⁵² Participants rated six statements regarding their commitment to a science career.¹²⁸ These statements include items such as "I intend to work in a job related to science," and "I will work as hard as necessary to achieve a career in science." An additional two statements were specific to remaining in a STEM major, as most CODE participants are expected to be majoring in a STEM field. Four reverse-scored STEM persistence questions were

included and used as validation points for student responses but were not included in the analysis.

Since no research-validated concept inventory existed to survey student bioinformatics awareness and learning as related to this project, a set of survey statements was drawn from a previous assessment of bioinformatics in life science curricula at the University of Wisconsin-La Crosse.⁶⁸ The researchers in this study implemented a program to integrate the study of bioinformatics into the undergraduate life science curricula at their university. The survey was conducted to assess student confidence in solving bioinformatics-related problems. The survey questions for the CODE project were drawn from this previous work because the questions in this instrument most closely evaluated the type of bioinformatics practices conducted during a CODE project. In particular, the questions in this paper included statements about protein domains, the three-dimensional structure of a protein, and genetic mutations responsible for disease. The author used a modified version of the questionnaire to assess the student's awareness and previous knowledge about bioinformatics techniques related to protein modeling to compare with their self-assessed knowledge after the research project had ended. Similarly, eight survey statements related to student comfort with and knowledge of computational modeling techniques were adapted from a study and assessment by Elmore *et al.*⁶⁹ that introduced college students to molecular dynamics and homology modeling.

4.2 Institutional Review Board (IRB) Approval and Ethical Compliance

As this project explored the impact of participating in an undergraduate research experience on student outcomes, human subject involvement was required. Study populations were college students, generally aged 19-23, and their faculty mentors from

the Southeastern region. In July 2021, an application for the use of human subjects in research was submitted for review to the Western Institutional Review Board (WIRB). Approval of the research protocol was granted on July 26, 2021. A modification to the protocol, including changes to the interview protocol, student consent, and communication to participants, was submitted for approval in May 2022 and granted on June 2, 2022. In November 2022, a modification to change the project PI from Neil Lamb, Ph.D., to Michele C. Morris, both of HudsonAlpha Institute, was requested. The revised protocol was granted on December 1, 2022. All research team members possessed CITI training. See Appendix B for IRB study approval documents.

An email sent to facilitators to share with their students also served as the information sheet for the study (see Appendix C for all IRB approved recruiting emails). The message informed the participants of the purpose of the study, described how the results would be used, clarified issues of confidentiality, informed them that participation was voluntary, and provided contact information for the study coordinators. The survey consent document was provided through an online link to the Qualtrics platform and outlined all of the same survey information in detail. Participants who consented to participate in the study were directed to the pre-survey. If consent was not provided, the survey was not opened. A similar online consent was implemented for the Facilitator surveys. (See Appendix D for Student Survey, Focus Interview, and Facilitator Survey Consent documents.)

4.3 Participants, Survey Administration, and Responses

Participants were recruited from the classes of trained CODE facilitators between August 2021 and May 2023. These facilitators taught at 17 colleges and universities in

multiple states. The institutions included community colleges, liberal arts schools, and universities. The surveys were administered online using Qualtrics Software, Version 2021. Facilitators were requested to share a provided online link and a QR code for the pre-survey with their students before beginning work on a CODE project. At the completion of their projects, facilitators were requested to provide a link and QR code to the student post-survey with their class. Students who completed the pre-survey were reminded by an email from the researcher to complete the post-survey approximately two weeks after their class or project work ended. Survey respondents provided their email in the pre-survey, allowing a post-survey request to be sent and facilitating the comparison between pre- and post-survey replies.

Data were collected from four schools in Fall 2021, 12 in 2022, and eight in 2023 (most schools participated during multiple semesters). Some CODE projects spanned multiple semesters, and post-survey submissions were provided in later semesters. Often the pre-survey was completed during class time and occasionally, time was provided during class to complete the post-survey. However, the post-survey response rate was $\sim 1/2$ of the pre-survey rate, most likely due to less emphasis on completing the post-survey by the instructor, disinterest, or survey fatigue by the student participants.¹⁵³

The study collected completed pre-surveys from 207 students at 17 institutions (2 community colleges, nine liberal arts institutions, and six universities). Of those institutions, eight were minority-serving. One hundred twelve unique individuals submitted completed student post-survey responses. Ninety-five post-surveys could be matched with pre-surveys and were used in subsequent analyses.

4.4 Data Preprocessing

The first step in the analysis process was data cleaning. The pre-and post-survey results were downloaded from the Qualtrics platform to a Microsoft Excel spreadsheet. A cutoff date for data collection of May 29, 2023, was selected due to the expectation that most students who were going to submit a post-survey had done so by that date. Each pre-survey response was given a three-digit ID code to anonymize the respondent. The author kept an ID code key for student names and schools in a password-protected file. Pre-survey responses that were incomplete were not included in further processing steps. If a response did not include identifying information that would allow their pre-survey to be matched to a post-survey, it was not included in subsequent analysis. If a student submitted more than one pre-survey, the original submission was retained, and the following submissions were deleted from further analysis. Of the 241 pre-survey responses, 207 met these requirements.

Post-survey data was downloaded to a Microsoft Excel spreadsheet and examined for completeness. Post-surveys that matched a pre-survey were given the matching ID code. If a post-survey was submitted that did not have a corresponding pre-survey, it was given an ID code and set aside for future consideration, if needed. Incomplete post-surveys were not used for analysis. Of the 141 submitted post-surveys, 95 were complete and could be matched with a pre-survey. One of these samples consistently skipped item responses, displayed conflicts with the reverse-coded items, and was deleted from the analysis, resulting in 94 matched responses. These 94 were used in the following analysis steps.

The Likert scale responses used in this survey were word-based (“strongly disagree” to “strongly agree” – 5-point scale; “not at all” to “a lot” – 4-point scale; “not at

all confident” to “a lot confident” – 4-point scale) and thus resulted in ordinal data. These data were converted to interval data using the mean value approach to facilitate data analysis. For example, “strongly disagree” was assigned a value of 1, “disagree” – 2; “neither agree or disagree” – 3; “agree” – 4; “strongly agree” – 5. Questions with 4-point Likert scale responses followed the same pattern.

4.5 Survey Constructs

The student survey consisted of 44 statements, divided into sections relating to bioinformatics awareness, interest, knowledge and comfort, research experience, science self-efficacy, science identity, and persistence in STEM. The author grouped the items into categories to facilitate analysis of each of the constructs of interest. The statistical analysis measure Cronbach’s alpha (α) was used to validate the reliability and internal consistency of each set of survey questions. Cronbach's alpha is a statistical measure that assesses the interrelatedness or correlation among items on a scale. It provides a value between 0 and 1, with higher values indicating greater internal consistency or reliability for the scale. This means that respondents who tended to select high scores for one item also chose high scores for the others. Similarly, respondents who selected low scores for one item tended to select low scores for the others. A Cronbach's alpha value of 0.7 or higher is generally acceptable for research purposes.¹⁵⁴ The pre-survey responses from the matched surveys were used to analyze each construct category using IBM SPSS Statistics (Version 29). All of the constructs scored ≥ 0.7 (Table 1).

The constructs measured the degree to which participation in a CODE project impacted shifts in bioinformatics awareness, interest, comfort, knowledge, research experience, self-efficacy, identity as a scientist, and persistence in STEM. The survey

statements are grouped into each construct, and the calculated Cronbach's alpha for each construct is shown in Table 4.1. The awareness, interest, comfort, and knowledge constructs examined four areas relating to bioinformatics. Students were asked to rank their agreement with their understanding of bioinformatics and how it is done in the bioinformatics awareness category of two questions ($\alpha = 0.847$). The interest construct included three statements that measured their agreement with enthusiasm about bioinformatics, interest in discussing bioinformatics with others, and interest in taking future classes in the subject ($\alpha = 0.841$). Responses in the bioinformatics comfort construct rated their comfort level in learning computational programs independently, working on a Unix/Linux-based computer, performing homology modeling, and performing molecular dynamics simulations ($\alpha = 0.700$). The bioinformatics knowledge category included nine items that measured if students felt they could find information about a gene, determine the domains of a protein sequence, translate DNA to an amino acid sequence, and if they thought they understood concepts such as computational modeling and homology modeling ($\alpha = 0.868$). All student survey items are listed in Table 4.1 and in Appendix E.

Although not a specific research question, the student surveys also included a category of items relating to research experience. Participants were asked to rate their participation in specific science-related activities using a 4-point scale (Not at all; A little; Quite a bit; A lot). As described above, the research experience items and scale were previously published and validated.^{24,25} These statements were included in the survey to evaluate the extent of prior research experience by CODE participants and if they felt that a CODE project provided an authentic research experience ($\alpha = 0.890$).

The items used to assess science self-efficacy and identity as a scientist were also drawn from the same studies by Chemers *et al.* and Robnett *et al.*^{24,25} Building on the work of Bandura⁹⁷ and Kardash,¹⁵⁵ the science self-efficacy scale assessed students' confidence in their abilities to function as a scientist. The six items relating to science self-efficacy included in this construct measured a student's confidence that they could complete the scientific tasks listed in the statements ($\alpha = 0.873$). These tasks included generating a research question, using scientific literature, explaining the results of a study, developing theories, and using scientific language and technical skills. Confidence levels were rated as *Not at all confident*, *A little confident*, *Somewhat confident*, and *A lot confident* (see Table 4.1 or Appendix E for a complete list of items).

The survey prompt¹³¹ for the identity as a scientist assessment section asked students to think about themselves and their personal identity. The prompt stated, "We want to understand how much you think being a scientist is part of who you are." Respondents rated how much they agreed with the five statements with the choices - *Not at all*, *A little*, *Somewhat*, and *A lot*. The statements in this measure included 'being a scientist is an important part of my self-image, being a scientist is an important reflection of who I am, I feel like I belong in the field of science, I have a strong sense of belonging to the community of scientists, and I am a scientist ($\alpha = 0.886$).

Lastly, students' retention intentions were measured in the persistence in STEM section of the survey. This construct contained 12 statements relating to the intention to major in a STEM field in college and pursue a STEM career ($\alpha = 0.914$). Students indicated their agreement with the statements using a 5-point Likert scale that ranged from *Strongly disagree* to *Strongly agree*, as was used in the studies by Syed *et al.*¹²⁸ and

Perez *et al.*¹⁰⁸ from which these items were drawn. The last four statements were reverse STEM persistence items intended to serve as a validation measure for the previous responses. They were not used in the analysis but can be found in Appendix E.

Table 4.1 Survey construct categories and internal reliability scores. The survey statements are grouped into each construct, and the calculated Cronbach's alpha for each construct is shown.

Construct	Item number	Survey Statement
Bioinformatics Awareness (Cronbach's alpha = 0.847)	14_1	I have a good understanding of what bioinformatics is.
	14_2	I have a good understanding of how bioinformatics is done.
Bioinformatics Interest (Cronbach's alpha = 0.841)	15_1	I am enthusiastic about bioinformatics.
	15_2	I am interested in discussing bioinformatics with my friends or family.
	15_3	I am interested in taking future classes relating to bioinformatics and computational biology.
Bioinformatics comfort (Cronbach's alpha = 0.700)	15_4	I am comfortable learning how to use computational programs on my own.
	15_7	I am comfortable working on a Unix/Linux-based computer.
	15_9	I feel comfortable doing homology modeling.
	15_11	I feel comfortable performing and analyzing molecular dynamics simulations.
Bioinformatics knowledge (Cronbach's alpha = 0.868)	14_3	I could use bioinformatics to find out information about a gene.
	14_4	I could use bioinformatics to find the domains in a protein sequence.
	14_5	I could use bioinformatics to determine the similarity between a group of genes or a group of proteins.
	14_6	

Construct	Item number	Survey Statement
	14_7	If given the nucleotide sequence of a gene, I could use bioinformatics to determine the amino acid sequence of the resulting protein.
	15_5	I could use bioinformatics to find a genetic mutation responsible for a disease or specific trait.
	15_6	I am familiar with the concepts of computational modeling methods.
	15_8	I am familiar with using computational modeling programs.
	15_10	I have a good conceptual understanding of homology modeling.
		I have a good conceptual understanding of molecular dynamics simulations.
Research Experience (Cronbach's alpha = 0.890)	17_1	I have worked on a research project in which I figured out what data to collect and how to collect it.
	17_2	I have reported my research results in an oral presentation or written report.
	17_3	I have learned scientific language and terminology.
	17_4	I have related my research results and explanations to the work of others.
	17_5	I have used scientific literature to guide a research project.
	17_6	I have had the opportunity to generate my own research question to answer.
	17_7	I have learned technical science skills.
	17_8	I have taken a leadership role in a scientific research team.
Science self-efficacy (Cronbach's alpha = 0.873)	19_1	Generate a research question to answer
	19_2	Use scientific literature to guide research
	19_3	Create explanations for the results of the study
	19_4	Develop theories (integrate results from multiple studies)
	19_5	

Construct	Item number	Survey Statement
	19_6	Use scientific language and terminology Use technical science skills
Identity as a scientist (Cronbach's alpha = 0.886)	21_1 21_2 21_3 21_4 21_5	In general, being a scientist is an important part of my self-image. Being a scientist is an important reflection of who I am. I feel like I belong in the field of science. I have a strong sense of belonging to the community of scientists. I am a scientist.
Persistence in STEM (Cronbach's alpha = 0.914)	23_1 23_2 23_3 23_4 23_5 23_6 23_7 23_8	I intend to work in a job related to science. I see the next steps in the field of science, and I intend to take them. I will work as hard as necessary to achieve a career in science. I expect a career in this field will be very satisfying. I feel that I am on a definite career path in science. I definitely want a career for myself in science. I am likely to remain in my STEM major through to graduation or completion of my program of study. I do not intend to leave my STEM major before I graduate or complete my program of study.

4.6 Analysis Methods

Each of the constructs showed excellent internal reliability for our sample ($\alpha \geq 0.70$). The author used each survey participant's average score for the items in the category (pre and post) to measure that construct. For example, the bioinformatics

interest grouping consisted of three items with an internal reliability score of 0.841. An average pre-survey response value was calculated for each student survey sample, and an average post-survey response value was calculated. The mean of the average pre-survey scores for each of the 94 samples was calculated and tested against the mean of the average post-survey scores for each respondent in the subsequent data analyses.

Descriptive and inferential analyses were executed on the student survey data for each construct. A histogram with a normal distribution curve and a boxplot graph was generated for the dataset of pre-survey means for each sample and repeated with the post-survey means dataset. These graphs were used to examine the data distribution, along with the Shapiro-Wilk test of normality and the interquartile range. A skewness test quantified the degree to which the data were tilted to one side of the distribution. A skewness value close to zero indicates that the data were approximately symmetric. Positive skewness suggests that the data were skewed to the right, while negative skewness suggests skewness to the left. The kurtosis test quantified the shape of the distribution, specifically focusing on the tails of the distribution relative to the rest of the data. A positive kurtosis value indicates heavier tails than a normal distribution, meaning the distribution has more extreme values and might peak more in the center. A negative kurtosis value indicates lighter tails than a normal distribution, suggesting fewer extreme values and potentially flatter or more spread-out tails. A higher kurtosis value signals outliers in the data.¹⁵⁶ These descriptive statistics were generated to understand the shape of the distribution, assess the departure from a normal distribution, and make informed decisions about data analysis approaches.

Samples identified as outliers by a boxplot and histogram in SPSS were carefully examined. The author considered the construct category and the possibility that the outlier might have been caused by a respondent selecting a response in haste to complete the survey or not providing thoughtful responses. For example, in the interest category, a sample was identified as an outlier with a mean pre-survey score of 1 and a post-survey score of 1. The author determined that it was entirely possible that a student had a very low interest in bioinformatics before their project and that those feelings did not change. Although several similar outliers were examined, all were accepted in the analyses and were classified as true outliers.

4.7 Related Samples Comparison Analysis

Paired samples t-tests were conducted for the pre-survey sample means and the post-survey sample means for each construct to test the null hypothesis that there was no difference between the means. The sample data did not meet the assumption of normality required for the t-test; however, the large sample size (94) allowed the usage of this statistical measure based on the central limit theorem.¹⁵⁷ The Wilcoxon Signed Rank test, a nonparametric test used when the data are not normally distributed, was also performed for each set of sample means to confirm the findings. The t-test is more powerful when assumptions are met, but the Wilcoxon Signed Rank test provides a non-parametric alternative robust to deviations from normality. These same measures were applied to the Facilitator Pre- and Post-Workshop surveys. All statistical tests were run using IBM SPSS Statistics (Version 29).

To reduce the risk of making a Type 1 error or rejecting the null hypothesis that is true, the significance level (p-value) was set at 0.05. At this level, there is a 95% chance

that the obtained results did not occur by chance, and the null hypothesis can be safely rejected. Statistical significance, defined by the p-value, is important in research and describes the likelihood of obtaining the results assuming the null hypothesis is true. In the case of this study, the null hypothesis holds that there is no difference between the pre-and post-survey means and no impact from participating in a CODE project. The one-sided significance result was used since the research questions ask if participation in the CODE project would increase student outcomes in several areas. This approach maximizes the test's statistical power by focusing the analysis on the specific direction of interest, which is an increase in changes in the examined measures in students.

A result may be statistically significant, but it is important to determine if it also has practical significance. A Cohen's d value was calculated for each paired-means analysis to measure effect size and understand the practical significance of the results. Cohen's d represents the standardized effect size and measures the size of the difference between the two groups (pre-survey responses and post-survey responses) relative to the variability in the data. The effect size is a measure of how meaningful the difference between groups is. A large effect size means the research finding has practical significance, while a small effect size indicates limited practical applications. Cohen's d value can also be used to compare studies to examine similar variables.¹⁵⁷

The 95% confidence interval (CI) for the pre-survey and post-survey mean in each category was calculated during these analyses. The confidence intervals are based on the sample data and are used to estimate the population parameter (the population mean). One can be reasonably confident the true population parameter lies within the CI value range 95% of the time.¹⁵⁷

4.8 Qualitative Assessment - Student Interviews

Qualitative focus interviews were used to gain a more in-depth understanding of the survey results. The focus interviews were designed to illustrate and confirm survey results and elaborate upon and clarify survey findings. All students who provided both pre- and post-surveys were sent an email requesting an interview. Eight survey respondents agreed to participate in an interview about their CODE project with the researcher. The discussions were held online through Zoom, and audio was recorded. A script of relevant open-ended questions derived directly from the survey measures was used for the interviews (Appendix E). The questions focused on the research themes of changes in awareness and interest in bioinformatics, self-efficacy, scientific identity, and intentions regarding persistence in a STEM field. Before the interviews, the participants' survey results were reviewed, and the follow-up interview instrument was edited to address any confusing or significant responses as needed.¹⁵⁸ All interview participants were asked the same questions, and the researcher requested follow-up comments for additional details as needed.

The participant interviews were semi-structured and ranged from 15 to 30 minutes. The semi-structured strategy helped delve deeply into a question to understand the responses thoroughly.¹⁵⁹ The audio recordings of each interview were transcribed by the *NVivo* Mac R1 (2020) transcription service. The researcher verified transcribed interviews by listening to the audiotape while reviewing the transcript word by word. During this process, all names or identifying information were replaced with descriptors and placed in parentheses. Video recordings and audio files were deleted. Transcribed interviews were coded with the students' ID codes to protect confidentiality and connect responses to the student's quantitative surveys.

A standard set of questions was posed to each interviewee, but the students also offered additional information and thoughts. Several questions asked in the interviews were similar to those in the survey to allow students to build on their previous responses. The focus interview discussions were designed primarily to enable students to provide free-form answers to the interviewer. Because the author led the interviews, there was a possibility of bias in the responses; therefore, the student responses were not evaluated systematically for this paper. In chapters 5 and 6, some intriguing and representative free-text answers are provided.

The comments were assessed qualitatively via inductive analysis using *NVivo* software to minimize rater bias. An initial set of codes was drawn from the prior surveys related to the research questions. Additional codes were added by the author each time a new discrete idea was encountered in the transcripts. The material from the transcripts was organized into 15 parent coding categories. Segments added to the code categories consisted of direct quotes from the transcript, allowing the students' voices to emerge and for direct interpretation of the data. After final coding, broad themes in the data were identified based on the frequency and patterns of code appearance. After identifying themes, representative quotes were chosen for many of the codes in this paper. They were lightly edited for confidentiality, clarity, and brevity before inclusion in the paper. After editing, the quotes were checked to ensure they represented their original meaning.

Additionally, *NVivo* was used for quantitative keyword analysis, explicitly looking for words and stems related to the research questions or appearing frequently. The author then evaluated these results to determine the context in which these words

appeared. Each student could make comments related to more than one code, so the number of comments exceeded the number of overall responses.

4.9 Facilitators Surveys for Formative Evaluation

College and university faculty were trained at several CODE facilitator training workshops. These one- or two-day sessions introduced CODE projects, the YASARA software program, and various other resources while teaching faculty to implement these projects at their institutions. A pre-and post-survey for the training workshop was developed based on the survey designed for The Genome Solver Project, which used a similar training format.⁵¹

Before each facilitator training workshop commenced, participants were asked to complete a preliminary six-item questionnaire addressing their grasp of bioinformatics, their level of interest, and their comfort with the subject. Following the workshop, a follow-up survey duplicated the initial six questions and introduced an additional 14 statements related to participants' perceptions of the workshop's benefits and their engagement in the CODE program. These statements were evaluated using a 5-point Likert scale (ranging from 1=Strongly disagree to 5=Strongly agree), and the survey concluded with four open-ended questions seeking their suggestions for enhancing the workshop. Demographic data were also collected in the pre-survey.

An online Facilitator Post-Implementation survey was designed using 14 questions from the Workshop survey modified to be answered retrospectively. Six new questions inquired about their experience implementing a CODE project and the impact on the students. All facilitator survey instruments are available in Appendix F. See Table 4.2 for details regarding the formative evaluation efforts.

Table 4.2 Facilitator formative evaluation methods. Formative evaluation methods were used to improve the facilitator training workshops. Details of the methods are shown in this table.

Formative research	Participant	Information Needed	Method
Assess the impacts of the Facilitator Training workshops on instructor knowledge and comfort level with instruction; feedback about quality of the training workshop	Facilitators / Mentors	Prior opinions of and experience with bioinformatics; shifts in comfort level teaching bioinformatics; Did the training meet your needs? What benefits did you derive from CODE training? Suggestions for changes	Pre- and post-training workshop survey; Likert scale and open-ended questions
Feedback about challenges	Facilitators / Mentors	Quality and usefulness of training and supporting resources	Surveys of trained facilitators >9 months post-training. Post-event Likert scale survey adapted from The Genome Solver Project
Feedback about the CODE Student Symposium	Student participants Mentors	Value of the Symposium to these audiences; suggestions for changes	Surveys immediately following the event; casual dialogues

Faculty facilitators completed the pre-and post-surveys during the facilitator training workshops to inform formative assessment of the training workshop. The surveys were provided online through the Qualtrics platform. The facilitator post-implementation

survey was shared with active facilitators to determine the long-term value of the training workshop. This follow-up survey was also administered online, approximately 6-9 months after a faculty member had attended a training workshop or when the researcher was made aware that the facilitator had completed an implementation of CODE with their students. An email was sent to the facilitators with the post-implementation link and QR code. Nine facilitators completed the post-implementation survey.

The pre-and post-surveys from the facilitator training workshop were examined for completeness and matched. Thirty-seven matching surveys were analyzed using paired samples t-tests in SPSS. The effect size measurement was also calculated. Due to a technical issue, the last question did not appear in the survey for some participants, and only 14 sample responses were available for that item.

Additional formative feedback was received from facilitators and student participants at the CODE Student Symposium. An online post-survey asked them to rank their impressions of the event speakers, the oral presentations, and the poster session. Open-ended items inquired as to what was the most valuable thing they gained from the symposium, how the symposium will affect their research, and also asked for additional comments. The responses (n=74) were all positive, with three main topics rising to the top: 1) Presentation experience (26 comments), 2) Networking opportunity (11 comments), and 3) Career options (26 comments).

4.10 Summary

The research questions for this study were considered through survey data collected from students who participated in a CODE project under the mentorship of a trained faculty facilitator at multiple colleges. Pre- and post-survey data collected from

CODE participants were matched for comparison analysis, resulting in 94 samples for quantitative analysis. Eight participants agreed to a subsequent post-interview with the researcher to provide qualitative responses to follow up on their survey results. The survey results were examined by descriptive and inferential statistical tests. The interview transcripts were coded for themes and organized into meaningful segments, including participant quotes. Surveys from faculty facilitators during training workshops and after project implementation provided valuable data for formative evaluation of the program. The results of the data analysis will be presented in Chapter 5.

Chapter 5. Results

This chapter presents a comprehensive analysis of the data collected during this research study, drawing from various sources, including pre- and post-surveys administered to CODE student participants, in-depth student interviews, and surveys conducted among faculty facilitators. The primary objective of this study was to critically assess the effectiveness of the CODE program as an intervention aimed at enhancing positive STEM-related outcomes. These outcomes encompass a range of key factors, such as bioinformatics awareness, interest, knowledge, science self-efficacy, developing an identity as a scientist, and the intent to persist in STEM fields. The assessment methods utilized in this research, as detailed in Chapter 4, have been employed to evaluate the impact of CODE research projects on students who have conducted research under the direction of trained faculty facilitators.

Surveys were administered to participating students before and after their engagement in CODE research projects. More than 200 students consented to participate in this study and submitted pre-survey responses. Among them, a subset of 95 students could be matched to post-survey answers, thus forming the basis for the analyses presented in this chapter. Eight students volunteered for in-depth focus interviews to augment the survey data, providing additional context and perspectives to complement the findings.

In addition to the student-centric data, this chapter delves into the faculty perspective. A group of faculty members engaged in CODE program facilitation

responded to pre- and post-training workshop surveys, offering valuable insights into the training administration. Subsequently, a post-implementation survey was administered to explore the faculty experience further. The outcomes of these surveys contribute an additional layer of understanding and complement the broader picture of the CODE program's impact on students and facilitators.

5.1 Student Participant Surveys - Demographic Data

The demographic questions were optional, but all of the students responded to those questions in the 95-matched surveys. Student participants were from 13 different undergraduate and one graduate school; 79% were upper-level students (juniors, seniors, and graduate), 28% identified as male, 71% as female, and 1% selected ‘Prefer not to answer’. Thirty-four percent self-identified as underrepresented minorities in science, 8% said they were living with a disability, and 25% were first-generation college students. Over 94% reported they were not Hispanic or Latino. The ethnicity breakdown was 60% White/Caucasian, 32% African American, 4% Asian or SE Asian, 1% American Indian, and 1% of two or more ethnicities. Table 5.1 contains the demographic data of the student participants).

Table 5.1 Demographic statistics of student respondents to the pre- and post-surveys.

CODE Student Survey Participants' Demographic Data	
Characteristic	Students N = 95
Gender	
Female	67 (71%)
Male	27 (28%)
Prefer not to answer	1 (1%)
Other	0
Ethnicity	

CODE Student Survey Participants' Demographic Data

Hispanic or Latino	5 (5%)
Not Hispanic or Latino	89 (94%)
Rather not say	1 (1%)

Race

American Indian or Alaskan Native	1 (1%)
Asian/SE Asian	4 (4%)
Black/African American	30 (32%)
Native Hawaiian or Pacific Islander	0
White/Caucasian	57 (60%)
Two or more races	1 (1%)
Prefer not to answer	2 (2%)
Unknown	0

Are you a person living with a disability?

Yes	8 (8%)
No	85 (87%)
Prefer not to answer	2 (2%)

Are you a first-generation college student?

Yes	24 (25%)
No	68 (72%)
Prefer not to say	3 (3%)
Not sure	0

Institution Type Represented by Respondents

Minority serving	7 (50%)
Non-minority serving	7 (50%)

Carnegie Classification of Respondents' Institutions

Baccalaureate Colleges	5 (36%)
Master's Colleges & Universities	4 (29%)
Associate's College	2 (14%)

CODE Student Survey Participants' Demographic Data	
Doctoral Universities	2(14%)
Medical Schools & Centers	1 (7%)
Student Level in School	
Sophomore	17 (18%)
Junior	20 (21%)
Senior	45 (47%)
Graduate student	12 (13%)

5.2 Student Participant Surveys – Descriptive Statistics

A set of descriptive statistics were calculated from the pre-and post-surveys to summarize and describe the main features of each dataset. To understand the central tendencies, variability, and distribution of the data from each construct, the following statistics were determined: mean, standard error of the mean, 95% confidence intervals for the mean, the standard deviation of the data set, the interquartile range, as well as skewness (asymmetry of the data distribution) and kurtosis ("tailedness" of the distribution) values. The mean values were calculated based on the numerical value assigned to each survey response. The bioinformatics awareness, interest, comfort, and knowledge sections were scored on a scale of 1 to 5, ranging from 1=strongly disagree and 5 = strongly agree. The persistence in STEM construct items were scored on this same scale. Research experience and identity as a scientist were measured on a 4-point scale (1 = Not at all; 4 = A lot), and science self-efficacy measured confidence on a scale of 'not at all confident' = 1 and 'a lot confident' = 4. The standard error of the mean (SEM) quantifies how far the calculated mean is likely to be from the true population mean. All of the SEM values are low (≤ 0.107), indicating that the mean is likely

representative of the population. Table 5.2 shows the descriptive statistics for the pre-and post-survey data for each construct. Based on the confidence intervals, the author is 95% confident that the survey means reported in Table 5.2 are between the shown confidence intervals.

To access the dispersion of the data for each survey the interquartile range (IQR), skewness, and kurtosis values were calculated. The IQR is the length of the scale where the middle 50% of the responses fall. Larger values indicate that the central portion of the data is spread out, while smaller values show that the middle values cluster more tightly.¹⁶⁰ Skewness indicates which direction the mean is pulled in a non-normal distribution of data. A normal distribution is symmetric and has a skewness value of 0. A positive value means the distribution is right-skewed. If the value is less than -1.0, the distribution is left-skewed, meaning the long tail would be toward the left, and the mean would be lower than the median. In general, a skewness value more than twice the value of its standard error indicates a non-normal distribution.¹⁵⁶ Of the survey datasets, only bioinformatics comfort (pre and post), bioinformatics knowledge (pre), research experience (pre and post), and science self-efficacy (pre) were determined to have a normal distribution by the Shapiro-Wilk test. See a complete list of values in Table 5.2.

Kurtosis is a measure of the "peakedness" or "flatness" of a distribution. A kurtosis value near zero indicates a data distribution close to normal. A negative value indicates a distribution that is more peaked than normal and has fewer extreme outliers, while a positive kurtosis indicates a shape flatter than normal and contains more extreme outliers than a normal distribution. The kurtosis values of this study's data sets are as

expected based on the distribution of the data sets as determined by other measures.¹⁵⁶

Refer to Table 5.2 for values.

Table 5.2 Descriptive statistics for the pre-and post-survey data for each construct.

Descriptive Statistics	Mean	Std. Error of Mean	95% Confidence interval for Mean - Lower Upper	Standard deviation	Inter-quartile Range	Skewness (Std. Error)	Kurtosis (Std. Error)
Bioinformatics Awareness Pre-survey (n=94)	2.664	0.107	2.452 2.878	1.041	1.50	0.045 (0.249)	-1.145 (0.493)
Bioinformatics Awareness Post-survey (n=94)	4.170	0.068	4.035 4.305	0.658	0.50	-1.437 (0.249)	5.603 (0.493)
Bioinformatics Interest Pre-survey (n=94)	3.635	0.077	3.483 3.788	0.745	0.67	-.653 (0.249)	1.303 (0.493)
Bioinformatics Interest Post-survey (n=94)	3.777	0.094	3.591 3.963	0.909	1.34	-0.730 (0.249)	0.521 (0.493)
Bioinformatics Comfort Pre-survey (n=94)	2.713	0.080	2.553 2.872	0.779	0.88	-0.007 (0.249)	0.045 (0.493)
Bioinformatics Comfort Post-survey (n=94)	3.497	0.080	3.340 3.656	0.772	1.00	-0.274 (0.249)	0.197 (0.493)

Descriptive Statistics	Mean	Std. Error of Mean	95% Confidence interval for Mean - Lower Upper	Standard deviation	Inter-quartile Range	Skewness (Std. Error)	Kurtosis (Std. Error)
Bioinformatics Knowledge Pre-survey (n=94)	2.915	0.075	2.764 3.065	0.735	1.00	-0.172 (0.249)	-0.036 (0.493)
Bioinformatics Knowledge Post-survey (n=94)	4.028	0.064	3.900 4.154	0.619	0.78	-0.848 (0.249)	1.743 (0.493)
Research Experience Pre-survey (n=93)	2.415	0.073	2.270 2.560	0.705	1.00	0.195 (0.250)	-0.707 (0.495)
Research Experience Post-survey (n=93)	3.020	0.065	2.891 3.148	0.623	1.00	-0.271 (0.250)	-0.535 (0.495)
Science self-efficacy Pre-survey (n=93)	2.824	0.069	2.688 2.961	0.662	1.00	-0.173 (0.250)	-0.517 (0.495)
Science self-efficacy Post-survey (n=93)	3.337	0.057	3.223 3.451	0.553	0.83	-0.478 (0.250)	-0.557 (0.495)
Identity as a Scientist Pre-survey (n=93)	2.940	0.0746	2.791 3.088	0.719	1.20	-0.235 (0.250)	-0.920 (0.495)

Descriptive Statistics	Mean	Std. Error of Mean	95% Confidence interval for Mean - Lower Upper	Standard deviation	Inter-quartile Range	Skewness (Std. Error)	Kurtosis (Std. Error)
Identity as a Scientist Post-survey (n=93)	3.256	0.076	3.105 3.406	0.731	1.20	-0.785 (0.250)	-0.122 (0.495)
Persistence in STEM Pre-survey (n=92)	4.375	0.063	4.251 4.499	0.601	0.88	-1.129 (0.251)	1.244 (0.498)
Persistence in STEM Post-survey (n=92)	4.498	0.065	4.369 4.626	0.620	1.00	-1.013 (0.251)	0.005 (0.498)

5.3 Student Participant Surveys – Quantitative Findings

Quantitative data were collected from pre- and post-survey of students who participated in a CODE project under the mentorship of a trained faculty facilitator. Eight constructs were developed from the survey questions. The mean of each student's responses to a set of construct items was calculated. Figure 5.1 depicts the differences between the pre- and post-survey means for each construct that was ranked on the 5-point Likert scale (1=strongly disagree, 3=neither agree or disagree, 5=strongly agree). Figure 5.2 depicts the differences between the pre- and post-survey means for each construct that was ranked on the 4-point Likert scale (1=Not at all, 4=A lot; or 1=Not at all confident, 4=A lot confident.)

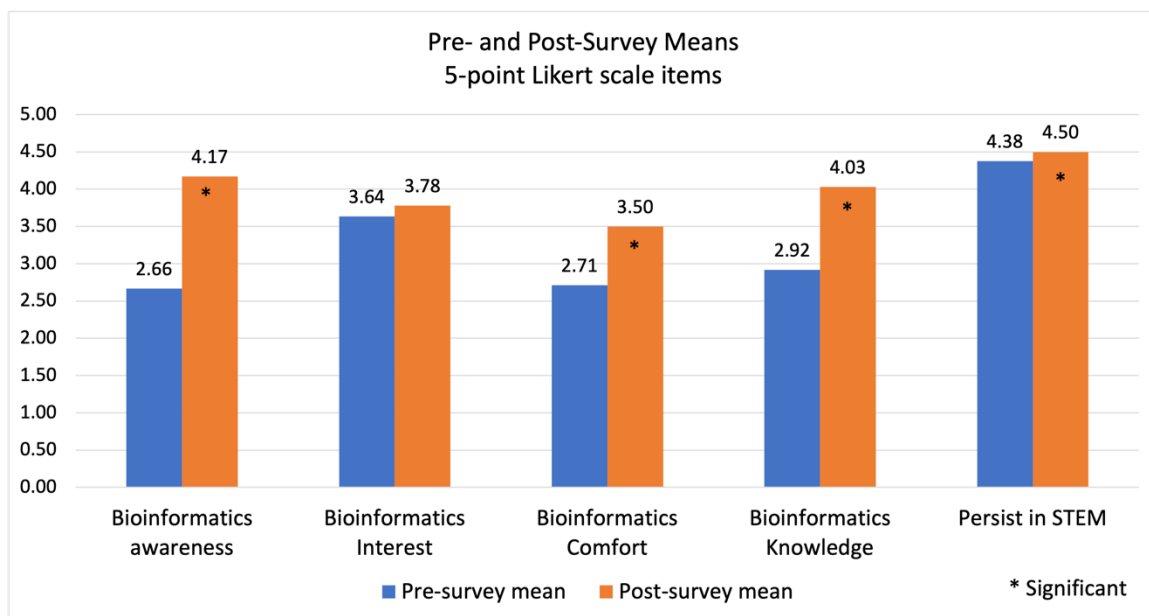


Figure 5.1 Pre-and post-survey means: 5-point scale items. The mean values for each construct are shown. Significant increases in the post-survey means are indicated with an asterisk (*).

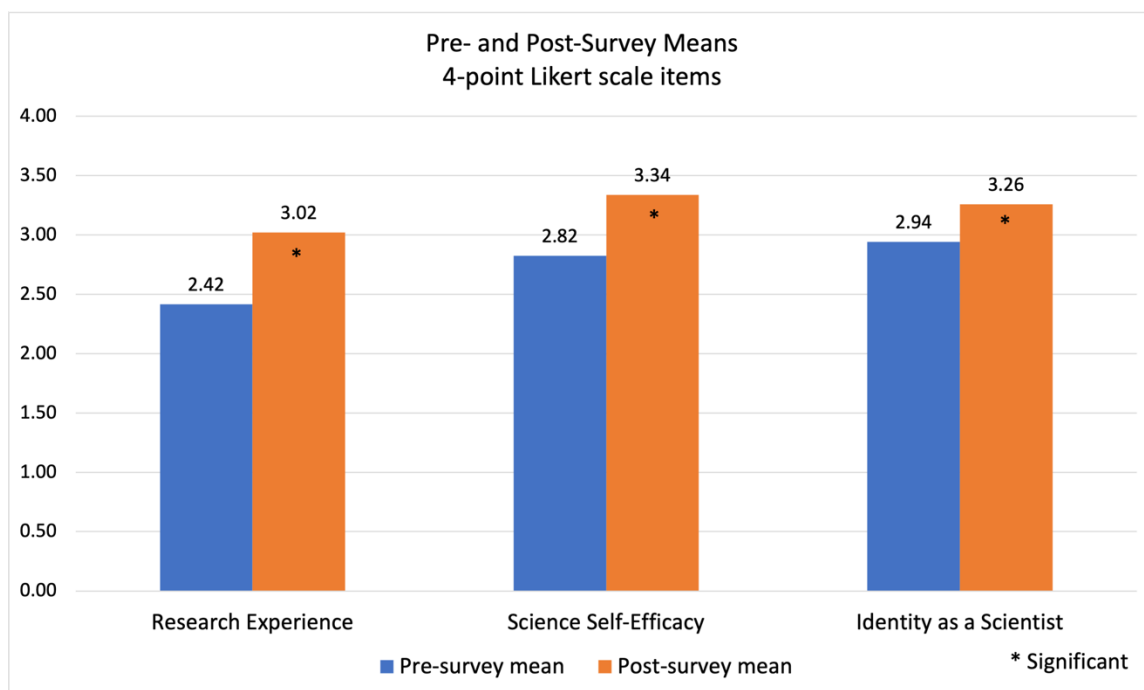


Figure 5.2 Pre-and post-survey means: 4-point scale items. The mean values for each construct are shown. Significant increases in the post-survey means are indicated with an asterisk (*).

The values for the means, along with the standard deviation and one-tailed significance values, are shown in Table 5.3. The means of all samples for pre- and post-survey answers were calculated and used in paired samples t-tests. The data for several constructs were not normally distributed, indicating that a non-parametric analysis such as the Wilcoxon Signed Rank Test would be appropriate to assess the difference between the means. However, due to the large sample size for each construct ($N \geq 92$), the central limit theorem holds that the distribution of the sample distribution will be normal, allowing for the more robust paired-samples t-test, a parametric test, to be performed. Both the Wilcoxon Signed Rank and paired-samples t-test were calculated for each construct and are shown in Table 5.3. An alpha level of .05 was used for all statistical tests, and a one-tailed t-test was employed since this study's hypotheses presume an increase in the surveyed measures after participation in a CODE research project. The effect size was determined using Cohen's d value in SPSS27 to assess the practical significance of the student pre-and post-survey means. See Table 5.3 for all paired sample comparison data.

Table 5.3 Student pre- and post-survey means and differences.

	Pre-survey mean	Post-survey mean	Difference	% change	Standard Deviation	Significance (one-tailed p-value)
Bioinformatics awareness (N=95; scale = 1-5)	2.664	4.170	1.506	57%	1.163	<.001
Bioinformatics Interest (N=94; scale = 1-5)	3.635	3.777	.142	4%	.914	.068

Bioinformatics Comfort (N=94; scale = 1-5)	2.713	3.497	.784	29%	.921	<.001
Bioinformatics Knowledge (N=94; scale = 1-5)	2.915	4.028	1.113	38%	.867	<.001
Research Experience (N=93; scale = 1-4)	2.415	3.020	.605	25%	.693	<.001
Science Self- Efficacy (N=93; scale = 1-4)	2.824	3.337	.513	18%	.701	<.001
Identity as a Scientist (N=93; scale = 1-4)	2.940	3.256	.316	11%	.716	<.001
Persist in STEM (N =92; scale = 1-5)	4.375	4.498	.123	3%	.570	.021

5.4 Bioinformatics Awareness and Interest

The first research question was addressed by the student survey items that focused on bioinformatics awareness, interest, comfort, and knowledge. As noted in Chapter 1, bioinformatics skills have become increasingly essential for scientific research and need to be integrated into post-secondary education.^{11,31–33} Four constructs were included in the bioinformatics section – awareness of what bioinformatics is and how it is done; interest in participating in bioinformatics-related activities; comfort level using basic bioinformatics tools and techniques; the acquisition of skills and knowledge related to bioinformatics research. The items for this construct were ranked on a 5-point Likert

scale (1=strongly disagree, 2=disagree, 3=neither agree nor disagree, 4=agree, 5 = strongly agree).

The measure for bioinformatics awareness included two survey statements, "I have a good understanding of what bioinformatics is" and "I have a good understanding of how bioinformatics is done." A paired samples t-test was conducted to determine the effect of participation in a CODE project on the awareness construct. The results indicated a significant improvement in students' self-reported awareness of bioinformatics pre-CODE ($M=2.66$; $SD=1.04$) and post-CODE ($M=4.17$, $SD=0.68$, $t(94) = 12.55$, $p < .001$) with a 1.505 point mean difference between the pre-and post-survey responses. A Wilcoxon-Signed Rank test confirmed these findings ($Z = 7.63$, $p = < .001$), negating the null hypothesis that the differences between the pre-and post-survey mean equals 0. The effect size of this result, measured by Cohen's d , was very large at 1.3.

The construct for bioinformatics interest presented three statements: *I am enthusiastic about bioinformatics*, *I am interested in discussing bioinformatics with my friends or family*, and *I am interested in taking future bioinformatics and computational biology classes*. The Wilcoxon Signed Rank non-parametric test ($Z = 1.86$, $p = .063$) for the means of the Interest items in the survey was not below the assigned significance level of 0.05, indicating that the students' pre-survey mean ($M = 3.64$, $SD = .75$) and post-survey mean ($M = 3.78$, $SD = .91$) did not have a significant difference. The paired samples t-test produced a similar result ($t(93) = 1.50$, $p = .068$), indicating a non-significant trend in the predicted direction. The effect size of 0.155 for this construct signifies a small effect of participation on an increase in interest.

5.5 Comfort Level with Bioinformatics

The survey questions also gauged the students' comfort levels when working with bioinformatics. Although their comfort with bioinformatics techniques and tools was not explicitly included in the research questions, it is essential to recognize that this aspect significantly influences a student's confidence and ability to thrive in the field of bioinformatics. Survey items related to how comfortable they felt using computational programs independently, working on a Unix/Linux-based computer, doing homology modeling, and performing molecular dynamics simulations – all concepts related to CODE projects

The pre-survey mean for comfort level was 2.71 ($SD = .78$), while the post-survey mean was 3.50 ($SD = .77$). These data were subjected to the t-test for paired samples, with the results showing a statistically significant gain ($t(93) = 8.26$; $p = <.001$). The Wilcoxon Signed Rank test corroborated this finding. Cohen's d measurement of effect size was 0.85, considered a high effect size.

5.6 Bioinformatics Knowledge

Similar to the comfort level survey statements, the knowledge construct included nine items related to computational tasks related to bioinformatics and genetic characterization. There was a statistically significant improvement in bioinformatics knowledge following participation in a CODE project from $2.92 \pm .74$ (pre-survey) to $4.03 \pm .62$ (post-survey) with a t -statistic of 12.45 ($df = 93$) ($p < .001$). This change was an improvement of $1.11 \pm .87$ Likert scale points. The Cohen's d effect size measure of 1.28 is very large for this construct. The Wilcoxon-Signed Rank test confirmed the significance result.

5.7 Research Experience

The CODE program aimed to provide students with the opportunity to conduct authentic research, particularly at smaller schools with fewer research options. To gauge if the availability of the CODE projects was increasing the students' sense of garnering research experience, eight items on the survey addressed this topic and were scored on a 4-point Likert scale (1 = Not at all; 4 = A lot). The dependent t-test results for research experience revealed significant differences between the pre-survey mean ($M = 2.42$, $SD = .71$) and the post-survey mean ($M = 3.02$, $SD = .62$), $t(92) p = <.001$) with a .605 Likert scale point difference between the means of the two surveys. The non-parametric test also determined the difference in the means was significant. The effect size value for research experience is 0.873, generally interpreted as a large effect of the intervention.

5.8 Science Self-efficacy

The second research question asked if students would experience increased science self-efficacy after completing a genetic VUS characterization as per the guidelines outlined in the CODE curriculum. The six statements in this construct were drawn from well-established, validated studies,^{24,25} enabling students to assess and rank their confidence levels when undertaking scientific research objectives. Students rated their confidence levels for these items on a 4-point Likert scale (1 = Not at all confident; 4 = A lot confident). See Appendix E for a complete list of the student survey items.

This study's results indicate a statistically significant difference between the mean pre- and post-survey scores for the science self-efficacy construct. Specifically, the post-survey mean was higher than the pre-survey mean. The mean pre-survey score was 2.82 ($SD = .66$). The mean post-survey score for self-efficacy was 3.34 ($SD = .55$). The

paired-sample t-test was conducted to compare the two groups. The t -statistic was 7.05 ($df = 92$) and a significance level of $p < .001$. The effect size for the difference between the pre-and post-responses was calculated using Cohen's d , resulting in a value of 0.73, considered a medium to large effect on science self-efficacy improvement.

5.9 Identity as a Scientist

"Identity as a scientist" refers to an individual's sense of self and belonging within the scientific community, encompassing their values, beliefs, and roles as a practitioner of science. Exploring improvements in this concept was the focus of the third research question. The students ranked the items on a Likert scale from 1 = Not at all to 4 = A lot. The five survey statements included how much a respondent feels they belong in science and how important being a scientist is to their self-image.

The mean pre-survey score for science identity was 2.94 ($SD = .72$), while the mean post-survey score was 3.26 ($SD = .73$). The non-parametric Wilcoxon-Signed Rank test revealed that the difference in the means was significant ($< .001$). These data were subjected to the t-test for paired samples, with the results showing a statistically significant gain ($t(92) = 4.27$; $p = < .001$). The effect size was .44, indicating a medium effect size for this intervention.

5.10 Intention to Persist in STEM

Persistence in STEM majors is crucial because it fosters a pipeline of skilled professionals who can drive innovation, solve complex problems, and advance society in areas critical to our future. Providing students with the interest and confidence to remain in a STEM field also promotes diversity and inclusion in these fields, ultimately ensuring a broader range of perspectives and ideas in these career fields. The last research question

asked if participation in a CODE research project would increase the students' intention to remain in science majors and career tracks. Students rated their agreement with eight items on a 5-point Likert scale (1=strongly disagree; 5 = strongly agree). These statements addressed their intention to work in science in the future and remain in a STEM major. See Appendix E for a complete listing of the student survey statements.

The pre-survey mean for this construct was high, at 4.36 ($SD = .60$). This pre-survey mean was the largest of all the constructs measured in this survey. There was a slight increase in the post-survey mean of .122 Likert scale points ($M = 4.50$, $SD = .62$) for a difference of 3%. See Table 5.3.

The Wilcoxon Signed Rank non-parametric test ($Z = 2.45$, $p = .015$) for the means of the Persistence items indicated that the change in the means was significant. The paired samples t-test produced a similar result ($t(91) = 2.05$, $p = .021$), indicating a significant increase in the predicted direction. The effect size of 0.214 for this construct signifies a small effect. All values for the paired samples comparison analysis of the student survey data are shown in Table 5.4.

Table 5.4 Student pre- and post-surveys - Paired samples comparison analysis data.

	Bioinformatics Awareness N = 94	Bioinformatics Interest N = 94	Bioinformatics Comfort N = 94	Bioinformatics Knowledge N = 94	Research Experience N = 93	Science Self-Efficacy N = 93	Identity as a Scientist N = 93	Persist in STEM N = 92
Mean difference (post - pre)	1.505	.142	.785	1.113	.605	.513	.316	.122
Std. Dev.	1.163	.914	.921	.867	.693	.701	.716	.570
Std. Error Mean	.120	.094	.095	.089	.072	.072	.074	.059

	Bioinformatics Awareness N = 94	Bioinformatics Interest N = 94	Bioinformatics Comfort N = 94	Bioinformatics Knowledge N = 94	Research Experience N = 93	Science Self-Efficacy N = 93	Identity as a Scientist N = 93	Persist in STEM N = 92
95% Confidence Interval of Difference, -								
- Lower	1.267	.046	.596	.935	.462	.368	.169	.004
- Upper	1.743	.329	.973	1.290	.747	.657	.463	.240
Wilcoxon Signed Rank Test (Z value, <i>p</i> value ^a)	7.63 <.001	1.86 .063	6.37 <.001	7.81 <.001	6.76 <.001	6.36 <.001	4.21 <.001	2.45 .015
<i>t</i> statistic	12.551	1.502	8.262	12.446	8.418	7.052	4.266	2.054
<i>df</i>	93	93	93	93	92	92	92	91
Significance One-sided <i>p</i>	<.001	.068	<.001	<.001	<.001	<.001	<.001	.021
Significance Two-sided <i>p</i>	<.001	.137	<.001	<.001	<.001	<.001	<.001	.043
Cohen's <i>d</i>	1.295	.155	.852	1.284	.873	.731	.442	.214

^aAsymptotic Sig. (2-sided test)

Of the nine participants who responded with either a strongly disagree, disagree, or neither agree or disagree to “*I am likely to remain in my STEM major through to graduation or completion of my program of study*” 100% of them responded with either agree or strongly agree on the post-survey. Conversely, five students who responded

either agree or strongly agree to that question responded with either a strongly disagree, disagree, or neither agree or disagree in the post-survey.

Of the eight participants who responded with either a “strongly disagree”, “disagree”, or “neither agree or disagree” to “*I intend to work in a job related to science,*” three of them responded with “agree” on the post-survey, and the other five did not change their responses. Two students responded “agree” to that question and changed their answer to “neither agree or disagree” in the post-survey. The remaining respondents responded either ‘agree’ or “strongly agree”) on both surveys

5.11 Qualitative Data – Student Interviews

All students who completed both the pre- and post-surveys were invited to participate in a short interview with the researcher to elaborate on their experience with the CODE project. Eight students agreed to be interviewed, three males and five females. Three of the students were Black/African American, and five were White/Caucasian. The schools represented included two liberal arts state universities and one HBCU. A standard set of questions, based on the research questions, were asked of each student volunteer. There were eight focus interview participants, and they all expressed positive impressions of the program. The audio files were transcribed, coded, and analyzed as described in Chapter 4.

5.12 Summary of the Interview Findings

- Career in science

All of the eight students were planning on careers in the science field. Two of them said the CODE project helped them to expand their plans to a potential research career.

- Collaborated with classmates

Four students discussed working collaboratively with their classmates and how that was helpful and rewarding.

"We were able to collaborate and share our findings together. So, it was great."

- Presentation skills

Three interviewees spoke about how their communication and presentation skills had improved through the CODE program. They shared their work in class, in school research seminars, and at the CODE Student Symposium and stressed how much confidence they had gained from these experiences.

- Identity as a scientist and confidence

The students usually discussed these two constructs together and agreed that they had gained confidence and a stronger sense of their scientific identity.

- *"Yes, I definitely think of myself as a scientist more than I did before this project."*
- *"I think it gave me some more confidence in areas that I didn't have confidence in, like communication for sure, and then being able to complete a project all the way through from where you have a goal and being able to finish and complete that goal, being able to present your completed findings at a scientific research fair. It definitely improved my confidence there."*
- *"It just made me a lot more confident, really kind of made research, more of an idea for something to do."*
- *"Before this course, I remember I checked "No" to "I was a scientist" because, you know, there are people who have all of these papers and they do research a*

lot. And I would say because I did do research and I did present it and I feel like it means something, that I am a scientist."

- Enjoyment

The coding category that collected the most comments was "Enjoyed project or had fun." Every interviewed participant shared at least one comment related to their enjoyment of the project, the fun they had using the YASARA Structure modeling software, working with diseases that were personal to them, learning genetics, or helping their classmates.

- Influenced by peers

Two interviewed students mentioned that they sought out the CODE class after seeing their peers present their projects in a club meeting or poster presentation.

- Interest in bioinformatics

All of the interviewed students said they would be interested in continuing to learn more about bioinformatics content in the future. Some intended to seek out courses, while others were amenable to activities that might be included in their required courses or would seek additional learning if they had time.

- Positive learning experience

Various comments about their learning experience surfaced during the interviews, although no questions were explicitly asked about learning. Seven of the students shared thoughts about how the program's hands-on structure had helped them learn and made learning enjoyable. Several mentioned that the challenging nature of the projects was a factor in increasing their learning and enjoyment.

- Research Experience

All of the participants stated in their interviews that they felt they had gained research experience and were more confident about conducting research in the future.

- *"I would be okay with conducting a research project. I would say that I'm not 100% confident, but I'm not like, I can't do it. So, like a happy medium."*
- *"I would say it definitely had an effect on me being able to perform my own research in whole and being able to participate in other people's research as well. It definitely improves my communication skills."*

5.13 Facilitator Surveys for Formative Assessment

Prior to the beginning of a facilitator training workshop, all faculty participants were surveyed in a six-item questionnaire about their bioinformatics understanding, interest, and comfort level. A post-workshop survey repeated the same six items and included 14 additional statements related to the perceived benefits of the workshop and participation in the CODE program. These statements were ranked on a 5-point Likert scale (1=Strongly disagree, 2=Disagree, 3=Neither agree nor disagree, 4=Agree, 5=Strongly agree). Finally, four free-response questions solicited their opinions about improving the workshop. The statements from the surveys can be found in Appendix F.

The training workshop pre-survey was completed by 39 facilitators. These pre-surveys could be matched with 37 of the 39 post-surveys submitted. A pre-survey mean for the responses for each question was calculated and repeated with the responses for each post-survey question. These data were analyzed with a paired samples t-test. Only two questions resulted in significant differences in the means (I am comfortable finding biological information in publicly available databases - $t(37) = 4.10, p = <.001$; I am comfortable using computer programs to visualize 3D images of molecules - $t(37) = 4.62$,

$p = <.001$). Each of these questions showed medium effect sizes. Table 5.5 reports the data from the analysis of the facilitator pre- and post-training workshop surveys.

Table 5.5 Facilitator's Training Workshop survey analysis. Pre- and post-survey paired differences comparison data.

	I have a good understanding of what bioinformatics is (N=37)	I am interested in mentoring bioinformatics research projects for my students (N = 37)	I am comfortable finding biological information in publicly available databases (N = 37)	I am interested in analyzing biological information (N = 37)	I am comfortable using computer programs to visualize 3D images of molecules (N = 37)	I am comfortable mentoring bioinformatics research projects for undergraduates (N = 14)
Pre-survey mean	3.892	4.405	3.919	4.514	3.216	3.929
Post-survey mean	4.270	4.649	4.459	4.730	4.135	4.189
Mean difference (post - pre)	0.378	0.243	0.541	0.216	0.919	0.261
Std. Dev.	.953	.863	.803	.821	1.211	1.204
95% Confidence Interval of Difference, - Lower - Upper	.061 .696	.044 .531	.273 .808	.058 .490	.515 1.323	.410 .981
<i>t</i> statistic	2.415	1.715	4.097	1.602	4.617	.888
<i>df</i>	36	36	36	36	36	13
Significance one-sided <i>p</i>	.010	.048	<.001	.059	<.001	.195
Significance two-sided <i>p</i>	.021	.095	<.001	.118	<.001	.391
Cohen's <i>d</i>	.397	.282	.673	.263	.759	.237

5.14 Facilitator Post-Implementation Survey Results

The post-implementation survey was completed by nine facilitators who had mentored their students with CODE projects. The six questions from the workshop surveys related to their knowledge of and comfort with bioinformatics were repeated in this survey. Of the seven survey respondents that could be matched with workshop surveys, there was very little change in their ranking of these six statements. Most responded with either 'agree' or 'strongly agree' again.

Sixteen additional questions in the post-implement survey related to the facilitators' impressions of the CODE program and its value to their teaching and students. The results were primarily either 'agree' or 'strongly agree'. Three items whose means were closer to the 'neither agree or disagree' category were related to the CODE website usage, interest from other colleagues at their schools, and gaining new colleagues in the genomics arena through CODE. See the complete list of Facilitator Post-Implementation Survey items and their means in Table 5.6.

Table 5.6 Facilitator Post-Implementation Survey items and means (n =97).

Scale: 1 = strongly disagree, 3 = neither agree or disagree, 5 = strongly agree	Mean (n = 9)
I have a good understanding of what bioinformatics is.	4.56
I am interested in mentoring bioinformatics research projects for my students.	4.67
I am comfortable finding biological information in publicly available databases.	4.22
I am interested in analyzing biological information.	4.67
I am comfortable using computer programs to visualize 3D images of molecules.	4.22

Scale: 1 = strongly disagree, 3 = neither agree or disagree, 5 = strongly agree	Mean (n = 9)
I am comfortable mentoring bioinformatics research projects for undergraduates.	4.44
I learned new bioinformatics applications or techniques through the CODE Facilitator Training workshop.	4.78
The workshop increased my understanding of the concepts and skills relating to protein modeling and variant research.	4.78
I am comfortable using the bioinformatics approaches included in the CODE program to teach my course(s).	4.44
The CODE workshop has been important in helping me plan my teaching of bioinformatics.	4.56
Because of the CODE program, I implemented new content or teaching approaches in my course(s) sooner than I otherwise would have.	4.44
Other faculty in my department have expressed interest in applying this approach to their area.	3.89
My participation in the CODE program has increased opportunities for undergraduate research projects in my department.	4.89
I believe that my participation in CODE has positively affected my reputation in my department.	4.11
I feel that my students have a better understanding of bioinformatics after participating in a CODE project.	4.63
My students are better prepared to use bioinformatics approaches than before the initiation of the CODE program.	4.63
The skills and activities covered in the CODE workshop have been beneficial in developing research projects with my students.	4.50
CODE provides support for flexible implementation and curriculum development.	4.50
I have used the website HudsonAlphaCODE.org in my teaching and research endeavors.	4.13

Scale: 1 = strongly disagree, 3 = neither agree or disagree, 5 = strongly agree	Mean (n = 9)
I have used the connections I made during the workshop in my teaching and research endeavors.	3.88
The CODE workshop provided a good network of technical expertise and support.	4.25
Through the CODE program, I have gained colleagues in genomics.	3.50

Some facilitators included comments in the optional free-response section of the post-implementation survey.

- I participated in the workshop twice: in 2018, and in 2021. My undergrad student who is now at NIH participated instead of me once in 2019. It was very very helpful. I am teaching my students that bioinformatics is an exploding field with basically no textbooks available. CODE program helped me a lot to shape up my course. I have noticed that the program and workshop are mostly focused on the bioinformatics of proteins. My research project is computed genomics of small RNA molecules. Still, the concepts I learned during the workshops are very helpful and applicable to the relatively under-developed field of the bioinformatics of small RNAs molecules. My grad student is defending her MSc thesis tomorrow (May 4, 2023) and will proceed to work at the NIH ... in computed genomics of piwiRNAs. Thank you a lot!*
- The CODE project allows me to implement undergraduate research projects that are suitable for students that are both early and advanced in their research experience. The CODE project is a perfect complement to the "wet lab" research*

but has the added benefit of a more predictable experimental success compared to the "wet lab" experiments.

- *My students really loved CODE.*

5.15 Summary

This chapter examined the data collected in this research study, which encompassed pre- and post-surveys administered to CODE student participants, student interviews, and surveys conducted among faculty facilitators. The central aim of this study was to critically evaluate the effectiveness of the CODE program as an intervention aimed at enhancing various positive STEM-related outcomes, including bioinformatics awareness, interest, and knowledge, as well as science self-efficacy, the development of a scientist's identity, and the intent to persist in STEM fields. Employing the assessment methods outlined in Chapter 4, this research has sought to assess the impact of CODE research projects conducted under the guidance of trained faculty facilitators on student participants.

The analysis of student participant data revealed significant improvements in multiple key areas. Bioinformatics awareness showed a highly significant increase after they participated in CODE projects. Similarly, students' comfort levels with bioinformatics tools and techniques significantly improved statistically. Furthermore, the analysis revealed a significant enhancement in bioinformatics knowledge among students. These findings are supported by large effect sizes, emphasizing the practical significance of these improvements.

Regarding science self-efficacy, students' confidence in their scientific abilities experienced a significant increase following their engagement in CODE projects. Identity as a scientist also exhibited growth, with students reporting a statistically significant improvement in their sense of belonging and importance within the scientific community. Lastly, the analysis addressed the intention to persist in STEM fields, revealing a small but significant increase in students' intent to remain in science majors and careers. While the change was relatively small, it's essential to note that students' pre-survey scores were high.

Additionally, the facilitator surveys provided valuable insights into faculty perspectives on CODE program participation. Facilitators reported increased comfort levels with various bioinformatics concepts after attending training workshops. The following chapter will discuss these results, along with the author's conclusions and plans for future work.

Chapter 6. Discussion and Conclusions

Over the past 25 years, numerous studies have proclaimed the need to integrate bioinformatics education into the college curriculum to support this growing scientific field. Yet, integration has been slow due to barriers such as untrained faculty and student hesitancy. Understanding the need for graduates with bioinformatics skills is clear, but elucidating the psychological constructs that support students along the path to graduation and a career related to computational biology is less clear. The Characterizing Our DNA Exceptions (CODE) program has been introduced in this study, and a series of psychological variables significant to STEM persistence have been examined.

The CODE project sought to examine these variables through participation in active learning and real-world research experiences in an interdisciplinary mix of courses at seventeen community colleges, liberal arts colleges, and universities. The study examined the impact an introductory bioinformatics project characterizing genomic variants of uncertain significance could have on college students' awareness, interest, and knowledge of the field, as well as their perceptions of science self-efficacy, identity as a scientist, and intention to remain in the science arena. The results indicate that the CODE program is a promising initiative for increasing measures of STEM retention at the undergraduate level. Additionally, the study has shown that the program can be successfully scaled for implementation at different universities and serves as a valuable initiative to introduce bioinformatics concepts and inquiry-driven research experiences to students.

The following chapter will discuss the results from a student survey before and after participating in a CODE project and student interviews as they relate to the study's research questions. Additionally, findings from faculty facilitator surveys related to a training workshop and their implementation of a CODE project will be reviewed. The chapter will discuss the study's limitations and plans for future work, followed by concluding remarks.

6.1 Restatement of the Problem

In recent years, significant efforts and resources have been dedicated to reshaping biology education to meet the evolving demands of the global workforce. Initiatives like the AAAS Vision and Change reports have catalyzed progress in science education.^{1,4} However, despite these strides, the United States still faces challenges in STEM proficiency compared to other nations.⁵ The urgency of addressing this issue stems from its implications for the national economy, as scientific innovation is intertwined with global competitiveness. STEM careers, known for their high wages and public esteem, contribute to economic growth.⁷⁻⁹ Addressing the STEM workforce gap becomes increasingly vital as STEM occupations continue to grow faster than the overall workforce.^{8,11} Although there has been a gradual increase in STEM graduates, specific gaps persist, particularly in fields like bioinformatics and computer science.¹³ These challenges emphasize the need to revitalize the STEM education system, reduce barriers to access, and promote diversity, all factors essential for innovation and competitiveness.

One critical strategy for mitigating these challenges involves undergraduate populations. Undergraduate research experiences have proven highly effective in engaging students, enhancing their confidence and scientific identity, and ultimately

increasing their persistence in STEM majors.^{19–21} As the demand for STEM professionals rises, boosting the number of STEM college graduates becomes imperative. Research indicates that students' interest in STEM fields and their ability to envision themselves in STEM careers significantly influence their retention in these disciplines.^{27–29} As the U.S. strives for scientific competitiveness on the global stage, promoting student persistence in STEM is pivotal. Additionally, advances in molecular biology and bioinformatics tools have made authentic research experiences accessible to students across diverse educational institutions, democratizing participation in bioinformatics research.

Inquiry-based learning experiences, such as a CODE project, enable undergraduates to engage with the culture of scientific research by participating in the same work as that done by scientists to produce knowledge.^{45,84,86} Participation in authentic learning experiences has been shown in numerous studies to increase students' interest and retention in science, as well as their confidence and motivation.^{83,85,88} The research questions posed by this study sought to understand if CODE project participation would shift these student self-impressions.

Does participating in a CODE project increase:

1. awareness, interest in, and knowledge of bioinformatics?
2. science self-efficacy?
3. scientific identity?
4. intention to persist in STEM?

The author hypothesized that positive shifts in these areas would be seen in the pre- and post-survey results. The following section will address the study findings and the author's thoughts related to each of these research questions.

6.2 Research Question 1: Does Participating in a CODE Project Increase Students' Awareness, Interest in, and Knowledge of Bioinformatics?

The first research question focuses on changes in a student's bioinformatics awareness, interest, comfort, and knowledge. The findings in this study show that participating in a CODE project under the guidance of a trained facilitator can create significant changes in most of these areas for student participants. Bioinformatics awareness and knowledge constructs showed the largest significant shifts from pre- to post-survey means, 57% and 38%, respectively. Becoming more comfortable with bioinformatics concepts and tools also saw a significant increase in the post-survey among CODE participants. The results for interest in bioinformatics increased slightly, but not significantly, after project participation.

The development of self-efficacy relies heavily on awareness and interest. As students are introduced to STEM professions, such as bioinformatics, they engage in activities that enhance their sense of competence and self-efficacy in the field. This form of participation causes them to anticipate positive outcomes from continuing the activity, thereby increasing their likelihood of pursuing a STEM career.

The fact that the students in this study significantly increased their knowledge of bioinformatics was expected, given past research on similar bioinformatics educational programs. For example, several studies have described bioinformatics education interventions that assessed the participants for knowledge gains.^{69,73,82} In all of these interventions, student knowledge increased. It is understandable that students new to a topic will learn new concepts during an activity. However, the implementation of the activity and their success with the challenges placed before them can substantially impact

their learning. Feedback from the facilitators and student comments indicate that a valuable feature of a CODE project is the student's ability to integrate known data and new findings in an impactful way as they characterize genomic variants.

In terms of effect size, increased bioinformatics awareness was the most robust student outcome ($\alpha = 1.295$), closely followed by bioinformatics knowledge ($\alpha = 1.284$) and bioinformatics comfort ($\alpha = .852$). These high effect sizes clearly indicate that bioinformatics research is approachable and achievable by undergraduate students.

The interest construct did not show a significant change but was trending toward increased interest. It was unexpected that student interest in bioinformatics did not increase significantly after participation in a CODE research project. In most cases, students self-selected for classes that included CODE projects, possibly indicating that they had an initial level of interest in bioinformatics that was not impacted by the project. The pre-survey mean for bioinformatics interest (3.635 on a 5-point scale) was the second highest pre-survey mean, just below the persistence in STEM pre-survey mean of 4.375 (5-point scale). It is encouraging to see the high levels of interest in bioinformatics, even in students who haven't participated in a computational biology activity in the past. Based on the students' responses in the interviews, there was a high level of enjoyment in the CODE project elements, in part because the instructions were "straightforward" and the modules were not overwhelming. As described by Bandura in his social cognitive theory (2001), finding success at each step as they move through the VUS analysis process contributes to a sense of achievement, supporting growth in science self-efficacy.

The focus interviews addressed these topics with eight participants to provide commentary to expand the survey findings. Seven of the students reported having no

bioinformatics knowledge or experience prior to participating in a CODE project, and all of them said they enjoyed the bioinformatics portions of the program, especially using the YASARA Structure modeling program. It has become clear that these types of hands-on learning activities are impactful and also enjoyable for the students.

- *"When the professor was like "a CODE project," I'm thinking I actually had to be at a computer and do a coding project, but this was much, much easier and smoother than a regular coding project that I had usually been used to. So it was fun to me. I mean, it crushed, so I think that's why I enjoyed it."*
- *"...it was really cool to kind of decipher and see what I could get out of it and then could pull the points that I needed. It was really fun. Especially playing around in YASARA and seeing how the protein folded, how it was meant to be, and going in and changing the side chains and everything was really fun."*
- *"Yes, I definitely did (enjoy the project), especially using the modeling software and doing things like the molecular dynamics simulations. Being able to do something where we could actually get usable and impactful data right in class was really, really interesting."*

Although the survey questions were designed to provide data allowing the researcher to draw conclusions about a student's experience during a bioinformatics-focused research project, another objective of the CODE project was to present the science of bioinformatics to students with no (or minimal) previous contact with the field.

Discussions with students have indicated that bioinformatics can seem like an intimidating and unappealing field. The CODE projects provided a shift in perspective for many students, leaving them with a positive impression of bioinformatics and interest in continued study of these processes and tools.

It is not surprising that this study showed increases in participants' impressions of bioinformatics because many schools, especially smaller colleges and universities, do not offer bioinformatics majors or classes, particularly at the undergraduate level. The field of bioinformatics has become more prominent since the completion of the human genome project. Educational opportunities are still making their way into the post-secondary curriculum and are rarely seen at the high school level. Once introduced to the tools and data analysis techniques involved with genomic variant analysis under the guidance of a trained facilitator, student engagement and excitement grow as they expand their projects into directions that interest them most. Increases in knowledge and comfort with bioinformatics are evident in students who present their work as a poster at the CODE Student Symposium. The conference is a forum for their work as well as an incubator for their confidence.

6.3 Research Question 2: Does Participating in a CODE Project Increase Students' Science Self-Efficacy?

The model of Social Cognitive Career Theory (SCCT)¹⁶¹ sees self-efficacy as a required precursor to developing an interest in a specific career. Self-efficacy can be conceptualized as an individual's internal conviction in their ability to successfully execute a certain activity or achieve a specific goal. This study was based on the theoretical construct of SCCT, which provided a path for the design of the research study and grounded the work in the principles of the psychological theory. The development of

self-efficacy in the realm of science has been shown to be critical to increased interest and commitment to a career in STEM, particularly after participation in undergraduate research experiences.^{128,162} The pre-post design of this study allowed the researcher to directly test whether shifts in science self-efficacy can be attributed to participation in a CODE project.

As described in Chapter 5, this study's results showed a statistically significant increase in students' sense of science self-efficacy, with a medium to large practical effect size (Cohen's $d = 0.73$), indicating that the effect is substantial and likely to have real-world significance. This gratifying finding is not unexpected due to the many elements of the CODE program that support student growth in this psychological mediator.

A key component to building self-efficacy is mastery – the practice of completing tasks and challenges successfully.^{95–97} Mastery is an influential source of efficacy, allowing students to gain confidence and belief in positive outcomes. The six items on the student surveys that related to self-efficacy all addressed mastery concepts in science and are shown below.

Survey prompt:

Indicate the extent to which you are confident that you can complete the following tasks. (*Scale: 1 = Not at all confident, 2 = A little confident, 3 = Somewhat confident, 4 = A lot confident.*)

Survey items:

- Generate a research question to answer
- Use scientific literature to guide research
- Create explanations for the results of the study

- Develop theories (integrate results from multiple studies)
- Use scientific language and terminology
- Use technical science skills

Throughout the progress of a CODE project, like many undergraduate research experiences, students engage with each of these mastery skills and many others. Unlike lab experiences with pre-determined results, so-called "cookbook" labs, a CODE project leads the student down the path of novel discovery. Beginning with a question they create by selecting a disease or condition and an associated VUS, they master multiple tools and databases to collect information to support their hypothesis. Each step provides them with an opportunity to master new scientific terminology, online resources, the process of experimental design, analytical thinking, and problem-solving.

Several students referred to their feelings of self-efficacy and mastery during the interview sessions.

- *"I think it gave me some more confidence in areas that I didn't have confidence in, like communication for sure, and then being able to complete a project all the way through from where you have a goal and being able to finish and complete that goal, being able to present your completed findings at a scientific research fair. It definitely improved my confidence there."*
- *"Letting us be hands-on and telling us that these are the platforms that are available and then here's how to get to it. And then letting us be free with it. And so we were shown this is how you do it, and then we were able to go and*

troubleshoot and do it on our own and kind of work through our problems individually."

Another key opportunity for skills development during most CODE programs arose when students shared their research findings with others. Presentations are the source of a considerable amount of anxiety for many people. Most participants presented their work in class or at a school-sponsored event. Many attended and presented at a public conference setting, such as the CODE Student Symposium, affording the students the opportunity to manage their anxiety and have a positive outcome by sharing their research experiences. Multiple attendees at the 2023 CODE Student Symposium responded on a post-event survey that the most valuable thing they gained was the experience of presenting their work. Self-efficacy is increased when success can be attributed to internal or controllable causes, such as ability or effort, rather than to external factors, such as luck or the intervention of others.^{100,101} One student shared their thoughts about building their presentation skills in their interview:

"I think the biggest thing for me, actually, I learned a lot about biology, but I think the biggest change for me was in presenting skills and being able to communicate with my peers and the general public about my research. For our class specifically, we focused a lot on being able to present your ideas each week. Getting used to presenting and talking with other people is something that was really unfamiliar to me, but I think it was definitely the most important skill I've gotten from this program."

A great deal of credit is due to the committed faculty facilitators who acted as mentors for students during their CODE projects, which resulted in observable improvements in science self-efficacy. In this regard, social persuasion becomes a critical component in supporting self-efficacy levels since it entails the influential function of teachers and mentors who praise people's talents and achievements, encouraging them to continue their pursuits. The author has confidence in the guidance and assistance these facilitators have constantly provided, having worked closely with and trained these facilitators while also keeping regular contact throughout the projects.

6.4 Research Question 3: Does Participating in a CODE Project Increase Students' Scientific Identity?

A science identity refers to an individual's perception of themselves as a scientist or someone who belongs to the scientific community. It encompasses their beliefs, values, and roles related to practicing science. Forming a sense of self as a scientist is a crucial factor for persistence in a STEM major,^{162,163} and the lack of a science identity has been identified as a factor contributing to an exit from STEM majors.^{110,164} Carlone and Johnson suggested that developing an identity as a scientist consists of three overlapping aspects: proficiency in important scientific techniques, putting these skills into practice in a way that is apparent to others, and receiving acknowledgment for these efforts from influential individuals.²⁷ The sense of being a part of the scientific community, or "fitting in" with other academics and the academic community at large, is linked to a student's identity, according to studies on student persistence and achievement in educational settings.^{24,25,165}

Many elements can contribute to shaping an individual's scientific identity. Hands-on learning opportunities, and especially the process of conducting inquiry-driven

research, can contribute to feeling like a true scientist. Huffmyer *et al.*¹¹¹ contended that authentic research experiences can foster a scientific identity. These experiences affect a student's sense of belonging, recognition, interest, and performance in science, as well as their scientific competence.

The results of the pre- and post-surveys revealed a significant increase in science identity among students who participated in a CODE initiative. This result was not unanticipated because CODE initiatives provide the opportunity to develop the skills required to promote a strong scientific identity. A supportive and inclusive environment that offers opportunities for exploration, learning, and mentorship is essential for developing a positive and robust science identity. These elements are crucial to the success of CODE initiatives and have been incorporated into the training for facilitators, the curriculum materials, and the Student Symposium.

Intrinsic motivation and a genuine interest in science, as well as positive experiences in science classes, research projects, and STEM-related activities, can contribute to the development of a strong science identity. Excerpts from the student interviews highlighted their positive experiences.

- *"I really enjoyed CODE. A lot of people can get a lot out of it. Even if they find out research isn't for them. I still think it's good to try it out . . ."*
- *"I think specifically (I enjoyed the project) because it allowed us to focus on things that were relevant to us and a little personal to us. And so being able to look at those diseases in that way and then see the more science and genetics part, that's what I enjoyed."*

- *"It was it was a great experience for me personally. I was able to, like I said, I was able to help my other classmates. We were able to collaborate and share our findings together. So, it was it was great."*

When asked about any changes to their identity as a scientist in the interviews, the students had several enlightening comments.

- *"Before this course, I remember I checked "No" to "I am a scientist" because, you know, there are people who have all of these papers, and they do research a lot. And I would say because I did do research and I did present it, and I feel like it means something that I am a scientist."*
- *"This is my last semester. And over the four years of being here, it's kind of been a little rough around the edges. This CODE project just made me think, okay, if I could do this and start from scratch, then maybe I could go forward and start somewhere else as a scientist because I pretty much did it on my own, and not only on my own, I helped my classmates. I feel like it did improve me as a scientist because before, I didn't talk to anybody, I was very shy, but now, it helped me with my characteristic as far as being able to be more open and being able to more talk about the fundamentals of biology and science, like it just helped me all around, more than I thought it was going to help me."*

- *"I definitely think of myself as a scientist more than I did before this project because when I was introduced to it, I was like, you know, "Oh gosh, what is this going to entail?" But after completing this project, it gave me a lot more confidence. And in my class, we actually conducted two assignments with the CODE project. So that gave us more practice with it. I definitely feel like, if I was to do a research internship or something of that nature, I would be more confident in completing the job."*

6.5 Research Question 4: Does Participating in a CODE Project Increase Students' Intention to Persist in STEM?

A myriad of factors has been shown to impact student persistence in STEM. Students are more likely to leave STEM when they have limited access to role models, poor peer support,²⁵ a challenging institutional environment, or face discrimination.^{166–168} Individual factors such as low confidence in their abilities^{169,170} or a lack of knowledge of potential STEM careers contribute to declining retention in STEM majors.¹⁷¹

CODE projects integrate learning with authentic research and stimulate students' confidence and interest in seeking additional science-related education and careers. In each instance of the project groups that participated in this study, the students were either science majors or self-selected to register for an interdisciplinary course based on CODE. Considering the high pre-survey mean for STEM persistence (4.375 on a 5-point scale) and taking the student interview responses as an example, the majority of these participants had already decided to pursue careers in science prior to CODE participation. Interestingly, one student said in their interview that participation in CODE had shifted their interest from the medical field to research.

"I came to college thinking I was going to get into medical laboratory science, also, I considered going to medical school for pathology. But now I am definitely considering more going to graduate school and making research my career because of this class."

According to SCCT, an increase in persistence in the science career domain would be expected, given the general gains in scientific research self-efficacy and science identity seen in this participant cohort.⁸⁹ This study quantified a small but significant change in students' future academic and career plans following CODE research experiences. Similar persistence results have been seen in other studies.^{66,78}

It was anticipated that CODE participants would enter the program with a high degree of interest in and intention to pursue a career in science. Indeed, 91.5% (n=94) responded with either "agree" or "strongly agree" to the pre-survey item "I intend to work in a job related to science," leaving little space for the slight post-survey increase (92.5%) on this measure. These results are highly encouraging and support the importance of providing students, even science majors, with increased research opportunities to build their conviction about pursuing a science career.

Students leave STEM majors for many reasons, some unrelated to their interest or desire to have a career in STEM. Other factors that reduce retention are more closely related to their pedagogical experience: the perceived difficulty of STEM classes, poor learning environment with a lack of a community of learners, low science self-efficacy, lack of science identity, and poor instruction. The cultivation of STEM self-efficacy and a scientific identity are essential factors in the continuation of a STEM major and the

progression to a STEM career.^{162,164} These issues can be addressed with interventions verified by research to positively impact students.

Inquiry-based learning, particularly in the field of bioinformatics, has been suggested as an effective method for increasing interest and persistence in science among undergraduates. Incorporating bioinformatics into the undergraduate curriculum^{61,172} as well as interdisciplinary research projects can increase students' engagement, skills, and process-specific knowledge, all factors that lead to overall increases in measures that relate to STEM persistence.

Participation in an authentic research experience can have a significant, positive effect on persistence, considering the multiple factors that influence STEM student retention. The aspects of CODE that make it so appealing and approachable for faculty - bioinformatics content, low-cost to implement, faculty training, curricula materials, open-source tools and software, a network of support, inquiry-driven research, and a student conference - provide students with a solid foundation for developing perseverance.

The persistence framework developed by Graham *et al.*¹⁹ lists eight educational interventions that promote persistence (early research experiences, active learning, content relevant to experiences of diverse students, constructive and encouraging feedback, mentors, role models, study groups, and extracurricular activities in STEM.) While no single intervention is likely to incorporate all of these persistence factors, CODE projects provide a solid foundation to encourage retention in STEM. The opportunity for students to engage in an early research experience provides context for student knowledge gains but also project ownership and the thrill of discovery, boosting their identity as scientists because they are actually doing science. Additionally, the

facilitator training helps faculty become strong mentors that can enhance the student's research skills and confidence. Lastly, student participation in the CODE Student Symposium is a valuable extracurricular STEM activity that stimulates confidence and allows students to identify with established scientists.

6.6 Student Interviews

The eight student interviews were enlightening and informative. The student comments were overwhelmingly positive, even when asked to give feedback to improve the program that might be negative. As seen in the summary data in Chapter 5 and the quotes sprinkled throughout this chapter, the students discussed their enjoyment with the challenging nature of the research and the chance to learn new bioinformatics tools. Every student expressed their interest in participating in future bioinformatics activities, given the opportunity. This finding is supportive of the importance of introducing students to bioinformatics in an engaging, non-programming-based format to allow them to form positive associations with the field and be open to further exploration.

All students who completed both a pre- and post-survey were invited to participate in an interview, however, only eight participants agreed to be interviewed, possibly because they are highly engaged students who had a positive experience with the CODE project. Finding students to agree to be interviewed was challenging, and some type of incentive may be beneficial in the future. The small sample set of interviews needs to be expanded to provide additional insights into the students' impressions of their experiences. Engaging the facilitators to select a more random array of interview participants would be informative.

6.7 Research Experience

Changes in the participants' self-perceived valuations of research were also assessed in the surveys. Of particular interest was the increase in their confidence in their ability to conduct research investigations after completing their CODE projects. Overall, there was a significant increase in their post-survey responses to items relating to research experiences, with a high practical effect. For example, on one survey item, students were asked to rank their agreement with the research-focused statement "I have worked on a research project in which I figured out what data to collect and how to collect it" on a 4-point scale (1 = not at all, 2 = a little, 3 = quite a bit, 4 = a lot). Fifty-three percent (n = 94) reported 'not at all' or 'a little' in their pre-survey responses. Of this group, 84% reported an increase in these research techniques in their post-survey. One student quoted in their interview, *"It just made me a lot more confident, really kind of made research, more of an idea for something to do."*

As seen in the discussions of science self-efficacy and scientific identity, research experience can contribute to building confidence in one's skills and anticipation of positive future research outcomes, as well as firmly establishing a student in the scientific community through their contributions of knowledge and sharing of their findings. Students enrolled in small liberal arts colleges or institutions with a limited research budget often do not have access to undergraduate research. The computer-based bioinformatics of a CODE project provides faculty with a welcome opportunity to incorporate authentic research experiences for their students. Faculty members supplied these comments on their applications to the CODE Facilitator Training Workshop when asked why they were interested in attending.

- *I want to offer students an opportunity to perform research.*

- *I want to give students experience in research, but I do not have space or equipment to do this. I have been actively looking for ways to increase my students' exposure to the possibility of research as a career, and I believe that this program will greatly help me and my students to attain this goal.*

- *The advent of modern molecular techniques has changed many aspects of biological sciences. However, the application of these molecular techniques and subsequent analyses using bioinformatics tools have not been reflected in the curricula as well as undergraduate research at (name of college). Thus, I'd like to participate in this bioinformatics training and gain related knowledge and resources so I can provide a better research opportunity for the undergraduate students in genomics research at (name of college).*

- *I would like to gain bioinformatics experience because I am currently limited in the approaches I can use in the research lab and classroom settings, and proficiency in the use of bioinformatics approaches is a key skill set for emerging scientists.*

6.8 Implementation of CODE Projects by Facilitators

A bioinformatics-based project can be an attractive alternative to the traditional wet lab-based study. Bioinformatics research can be implemented inexpensively due to freely available data and software, making it an appealing way for students to engage in authentic research and inquiry-based learning.^{21,97} There is a wealth of data to be

interpreted, resulting from dramatic increases in genomic sequencing. Expanding the populations of students who have the opportunity to explore genomic research is an effective way to increase bioinformatics education and work towards meeting the need for graduates with knowledge of molecular structure and bioinformatics.⁴⁹ Additionally, embedding undergraduate research projects within a course, as CODE projects are commonly implemented, is especially helpful for low-income, first-generation, and minority students who are often unable to dedicate time outside of class to research due to personal and financial constraints.¹¹⁷

Individual faculty facilitators are free to implement a CODE initiative in their institutions as they see fit. Facilitators have used diverse approaches in designing methods to implement the CODE curriculum. Depending on the instructional structure of the institution and the availability of computers, the following implementation strategies have been utilized:

- developing a new semester-long interdisciplinary course based on CODE projects.
- integrating a CODE project into a broader genetics or molecular biology course as a short-term project.
- incorporating a CODE project into a laboratory section of a science course, sometimes presented as an online implementation.
- developing independent projects for summer research or capstone initiatives.

6.9 Faculty Facilitator Outcomes

To increase awareness and interest in bioinformatics in students, faculty members must first be trained to confidently take the curriculum into their classrooms. The CODE facilitator training workshops introduced the tools and resources for characterizing

genomic VUSs. For many facilitators, the CODE program was their first experience with bioinformatics and genomic variants. Most facilitators teach at small institutions with few resources for experimental research. Participating in the training workshops and becoming a part of the CODE community opened the door for professional networking and collaborations. Several of the questions on the facilitator pre- and post-workshop surveys, adapted from the Genomics Education Partnership program, provided insight into outcomes related to these themes.⁸⁵

Pre- and post-surveys were administered at the facilitator training workshops. Faculty reported a significant increase in their understanding of bioinformatics, interest in mentoring research projects, and comfort in finding information on public databases and using computer modeling programs. Two items that increased from pre- to post-survey, but not significantly, were interest in analyzing bioinformatic information and comfort in mentoring bioinformatics projects. A large amount of information is presented in the two-day workshop, potentially leading to hesitancy in implementing the program without some planning and thought. This hesitancy may have influenced the responses to these questions. Facilitators sometimes say the training sessions are intense, and they need time to process all the new information they have learned. Providing some basic 'homework' could reduce the amount of information that needs to be covered in the training. Also, extending the length of the training could be impactful.

Four open-ended questions gathered faculty participants' opinions on what they would like to see added, what was beneficial, what could be improved, and any other thoughts. Individual comments and suggestions gathered through post-event surveys are essential for the iterative improvement of educational programs. Based on the open-

response comments, the workshops were effective and engaging, but there is room for improvement. Some participants requested additional hands-on training activities and examples, as well as extending the workshop to three days and adding an advanced training. Based on these comments, plans for adding additional hands-on activities to the workshops are being developed, along with the possibility of expanding the training to three days or providing advance work for the participants. An advanced workshop is also in the works, in either a one-day format or a synchronous virtual arrangement where faculty can choose to attend the sessions featuring topics that interest them.

The facilitator post-implementation survey sought to obtain feedback about the quality and usefulness of the training, supporting resources, and the success of their CODE research projects. While there were only nine respondents, the findings showed that they are implementing CODE projects with success. Non-active facilitators likely opted not to return this survey, possibly leading to a biased view of the facilitators' actual implementation results and opinions. A revised survey to gather input from less-active facilitators would be beneficial for structuring a more supportive program in the future.

6.10 Benefits of a Computational Research Program

A computer-based research experience offers distinct advantages over traditional wet lab experiments. Facilitators only require a computer with internet access for student projects, with the provided YASARA Structure modeling software from the CODE program or free alternatives. These *in silico* projects pose minimal lab safety concerns and promote collaborative group work. Data gathering and analysis from open-source resources entail no expenses, in contrast to the significant costs associated with bench work, including reagents and equipment. This approach allows facilitators to structure

projects for larger student cohorts, teaching a common set of techniques adaptable to individual research projects.

When students have a personal connection to their research projects, they become more engaged in and enthusiastic about learning. This engagement is evident in comments from CODE students. Bioinformatics can be an intimidating subject for many students, so it is vital to encourage them to interact with databases and tools through an engaging project. Students frequently select CODE project variants that pertain to a disease or condition that affects their family or themselves. This connection motivates them to dig a little deeper and learn the techniques that will enable them to answer their queries about the phenotype associated with the gene corresponding to their study variant.

6.11 Summation of Significant Findings

The study's significant findings possess the potential for far-reaching impacts that extend beyond the boundaries of the CODE project. The students' advancements in the examined psychological variables will likely yield long-term consequences. Table 6.1 summarizes the percentage changes corresponding to each of these findings and their potential implications.

Table 6.1 Significant findings of the CODE project study. The following table provides an executive snapshot of the study findings and impacts from the quantitative assessment.

Significant Finding	% change	Impact
Increased awareness of bioinformatics	57%	Contributes to interest in bioinformatics and knowledge of career opportunities

Significant Finding	% change	Impact
Increased comfort with bioinformatics activities	29%	Builds confidence and expands the depth of techniques one can utilize for research projects, as well as provides an opportunity to teach others
Increased bioinformatics knowledge	38%	Supports the skills needed to analyze and interpret complex biological data and provides access to a broader range of job opportunities
Increased research experience	25%	Enhanced critical thinking, improved problem-solving skills and self-confidence, contribution to knowledge, and much more
Increased science self-efficacy	18%	Empowers individuals to pursue scientific interests, overcome challenges, and make meaningful contributions to the field of science
Increased identity as a scientist	11%	Enhances motivation, confidence, and academic performance while also promoting a sense of belonging, innovation, and career success in the scientific community
Intention to persist in STEM	3%	Supports completion of STEM degree, workforce readiness, and economic growth and global competitiveness

6.12 Limitations

Despite this study's strengths and promising findings, there were some limitations. First, a randomized controlled trial was not feasible in the present context; therefore, this assessment could not establish the CODE program's true causal effects. The data in this study were collected via survey methods, which can introduce inaccurate and incomplete responses.^{173,174} To ameliorate any potential issues, incomplete survey responses were

eliminated from the analysis, and a sample displaying conflicts with the reverse-coded items was deleted from the analysis.

In most instances, the individuals who took part in the study voluntarily chose to enroll in the courses that incorporated CODE projects, potentially leading to the introduction of self-selection bias. There is a potential for variation in the interests, goals, and attributes of students who opted to enroll in classes incorporating CODE projects compared to those who did not select such courses. Consequently, this difference may provide dissimilar outcomes when compared to other student cohorts. Future studies would benefit from examining the effects of participation in a CODE project on students enrolled in required general science courses necessary for graduation as compared to those in elective courses featuring a CODE project.

Similar to most research studies that rely on surveys, this study incorporates a sample of the population rather than including the complete population. The objective is to extend the findings obtained from students introduced to bioinformatics and scientific research via a CODE project to a broader population of college students. It's important to note that the conclusions derived from this study apply primarily to individuals sharing similar characteristics with the participants. However, as the program grows and assesses a larger number of students, the sample group's size and diversity will expand, enabling more robust predictions regarding the impacts on the population at large.

Likert scales, commonly employed in surveys, are susceptible to several response biases. Respondents may exhibit social desirability bias by avoiding extreme responses to present themselves more favorably, leading to response bias. Additionally, in lengthy surveys, fatigue or inattention can cause individuals to provide careless or inattentive

responses, undermining the validity of the data. Subjective interpretation is another challenge, as vague terms like "somewhat" or "fair" may yield varying interpretations among respondents. Lastly, the closed-ended nature of Likert-type questions may force respondents to select the most relevant answer, even if it doesn't accurately represent their true perspectives, thus limiting response accuracy.^{175,176} It is possible that these biases impacted the responses to this project's surveys.

The potential for non-responders to have different outcomes after a CODE research project than those students who did respond is another form of bias that must be considered when evaluating the results. Efforts to mitigate nonresponse bias included offering the survey through an online computer browser or on a mobile device. Participants were assured confidentiality in the consent forms and interviews. Multiple email reminders were sent to students in an effort to solicit responses to the post-survey.

Concerning the focus interviews, the interviewer was known to the students as the director of the CODE program, which could have influenced their responses. For example, the student may have provided more positive answers in an effort to impress the interviewer. In the future, soliciting an unrelated individual to conduct the interview could provide more reliable data.

Finally, one must consider the wide variation in the experiences the CODE students encountered due to each group being facilitated by a different faculty member at a different school. These projects were implemented in multiple formats (in-person semester-long class, online class, short project as part of a genetics or biology class, summer project (individual), small group project) and included different curricula at the professor's discretion. It is inevitable that the learning experiences will vary. While there

is no way to control for these variables, in the future it may be possible to compare the groups to one another with a larger sample size of students. However, since the goal of CODE is to implement a program to train facilitators and introduce bioinformatics research to a wide range of schools and students, perhaps the emphasis of the survey analysis is appropriately placed on examining all the student participants as a population rather than individual school groups.

6.13 Future Work

The research assessment surveys for both students and facilitators will be ongoing over the next year, allowing for the accumulation of more extensive and nuanced data. Expanding the variables under investigation would offer valuable insights for the CODE program and its participants. Conducting a gender-based data analysis to identify potential disparities between male and female participants may prove enlightening. Additionally, categorizing participant responses by their respective schools and mentoring facilitators and subsequently comparing these groups could reveal outcome variations based on the program's specific teaching and learning experiences.

On a separate note, prior research has indicated that students often forge strong connections with their peers during group research projects or when collaborating on similar research endeavors. This phenomenon may contribute to the observed link between research experience and the development of a student's identity as a scientist.¹⁷⁷ Forming close social bonds with fellow motivated students can intensify one's identification with the field of science. Previous studies have underscored the role of peer connectedness in students' pursuit of careers in science.^{178,179} Investigating these connections between peer connectedness and CODE research projects holds the potential

to provide valuable insights for facilitating the program in the future, ultimately benefiting both facilitators and students alike.

CODE joins other programs, such as GCAT-SEEK,^{180,181} Genomics Education Partnership,^{75,85} SEA_PHAGES,⁷⁴ and Genome Solver⁵¹ that seek to create networks of educators who are incorporating bioinformatics and genomics curricula into their classes. Each of these programs is built on different research foci and methods. Still, all share a similar philosophy of increasing student engagement with genomics and bioinformatics and supporting faculty in implementing authentic research experiences. CODE will continue to expand its facilitator networks through collaborations with these programs, some of which are already in place, and learning from their successful examples. Resources such as the BioQUEST Curriculum Consortium provide platforms to support STEM education programs by facilitating team projects and making information and resources easily accessible. This partnership will provide access to a broader audience and grow the CODE community.

The future is bright for the CODE program, with strong interest among faculty and positive impacts on students. The current funding for CODE will continue for another year, and a proposal submission for a next-level grant is planned. The next two Facilitator Training Workshops have reached maximum registration (17 participants) and have waiting lists. The annual Student Symposium will be held in April 2024, and record attendance is expected. The CODE Advisory and Curriculum Committee is continuing to develop support materials for faculty to ease implementation in their classrooms. The findings from this research study are evidence that students who participate in CODE

projects reap positive impacts that solidify their science self-efficacy and identity, potentially leading to their continued path to a STEM career.

6.14 Conclusions

This study aimed to understand how participation in a CODE project impacts a student's awareness, interest, knowledge, science self-efficacy, and scientific identity, with the understanding that these factors are crucial in determining a student's willingness to remain in a STEM major and seek a career in a STEM field.

The findings of this study underscore the CODE program's substantial positive impacts on students' bioinformatics awareness, knowledge, self-efficacy, identity as scientists, and intention to persist in STEM. Furthermore, the faculty facilitator surveys confirm the program's success in enhancing their comfort and competence in bioinformatics, emphasizing the multifaceted benefits of the CODE program across both student and faculty dimensions.

The current study contributes to the literature by quantifying the student gains associated with participation in an introductory bioinformatics research project characterizing genomic variants. Multiple examples of bioinformatics education interventions employing generalized assessment strategies have been exemplified in the literature^{70,72,76}; however, very few studies have included measures to assess the crucial constructs of science self-efficacy, identity as a scientist, and intention to remain in a STEM field.^{74,102,127,132} This study contributes to the literature by providing evidence that the CODE bioinformatics intervention can significantly increase these predictors of persistence in science.

This low-cost and easy-to-implement project can be widely distributed and is especially valuable to faculty who are interested in adding bioinformatics to their curriculum but have no formal training and to institutions without ample research opportunities for undergraduates. The knowledge generated from this evaluation will serve as a case study on the effectiveness of increasing science self-efficacy, scientific identity, and authentic research opportunities in computational biology using a collaborative development, training, and implementation approach.

References

1. Vision and Change in Undergraduate Biology Education: A View for the 21st Century. In: *American Association for the Advancement of Science (AAAS) Conference Homepage*. ; 2009. Accessed July 17, 2023. https://www.aaas.org/sites/default/files/content_files/VC_report.pdf
2. Woodin T, Carter VC, Fletcher L. Vision and change in biology undergraduate education, a call for action--initial responses. *CBE-Life Sciences Education*. 2010;9(2):71-73. doi:10.1187/cbe.10-03-0044
3. *Vision and Change: A Call to Action, Final Report*.; 2011.
4. AAAS. *Vision and Change in Undergraduate Biology Education: Chronicling Change, Inspiring the Future*.; 2015.
5. Athanasia G, Cota J. The U.S. Should Strengthen STEM Education to Remain Globally Competitive. Center for Strategic & International Studies; Perspectives on Innovation Blog. Published 2022. Accessed August 6, 2023. <https://www.csis.org/blogs/perspectives-innovation/us-should-strengthen-stem-education-remain-globally-competitive>
6. *Science and Engineering Indicators: Science and Engineering Labor Force*.; 2022. <https://nces.nsf.gov/pubs/nsb20198/>
7. Zakaria F. *The Post-American World*. 1st ed. W. W. Norton & Company; 2008.
8. *Employment in STEM Occupations*.; 2023. Accessed September 25, 2023. <https://www.bls.gov/emp/tables/stem-employment.htm>
9. Science and Technology: Public Perceptions, Awareness, and Information Sources. Science and Engineering Indicators 2022. Published 2022. Accessed October 5, 2023. <https://nces.nsf.gov/pubs/nsb20227/public-perceptions-of-science-and-technology>
10. Handelsman J, Elgin S, Estrada M, *et al*. Achieving STEM diversity: Fix the classrooms. *Science*. 2022;376(6597):1057-1059. doi:10.1126/SCIENCE.ABN9515
11. Beardsley TM, Gropp RE, Verdier JM. Addressing Biological Informatics Workforce Needs: A Report from the AIBS Council. *Bioscience*. 2018;68(11):847-853. doi:10.1093/biosci/biy116
12. National Center for Education Statistics. Undergraduate Degree Fields. Condition of Education. U.S. Department of Education, Institute of Education Sciences. Published 2023. Accessed August 5, 2023. <https://nces.ed.gov/programs/coe/indicator/cta>
13. National Center for Education Statistics. Undergraduate Retention and Graduation Rates. Condition of Education. U.S. Department of Education, Institute of Education Sciences. Published 2022. Accessed September 25, 2023. <https://nces.ed.gov/programs/coe/indicator/ctr>

14. Zilberman A, Ice L. Why computer occupations are behind strong STEM employment growth in the 2019–29 decade: Beyond the Numbers: U.S. Bureau of Labor Statistics. Beyond the Numbers: Bureau of Labor Statistics. Published 2021. Accessed August 27, 2023. <https://www.bls.gov/opub/btn/volume-10/why-computer-occupations-are-behind-strong-stem-employment-growth.htm>
15. Riegle-Crumb C, King B, Irizarry Y. Does STEM Stand Out? Examining Racial/Ethnic Gaps in Persistence Across Postsecondary Fields. *Educational Researcher*. 2019;48(3):133-144. doi:10.3102/0013189X19831006
16. Women, Minorities, and Persons with Disabilities in Science and Engineering. National Center for Science and Engineering Statistics. Published 2019. <https://nces.nsf.gov/pubs/nsf19304/digest>
17. Hewlett SA, Marshall M, Sherbin L. How Diversity Can Drive Innovation. *Harv Bus Rev*. Published online December 2013.
18. Ellison SF, Mullin WP. Diversity, Social Goods Provision, and Performance in the Firm. *J Econ Manag Strategy*. 2014;23(2):465-481. doi:10.1111/jems.12051
19. Graham MJ, Frederick J, Byars-Winston A, Hunter AB, Handelsman J. Increasing Persistence of College Students in STEM. *Science (1979)*. 2013;341(6153):1455-1456. doi:10.1126/science.1240487
20. Estrada M, Hernandez PR, Schultz PW. A longitudinal study of how quality mentorship and research experience integrate underrepresented minorities into STEM careers. *CBE Life Sci Educ*. 2018;17(1). doi:<https://doi.org/10.1187/cbe.17-04-0066>
21. Peters AW, Tisdale VA, Swinton DJ. High-impact Educational Practices that Promote Student Achievement in STEM. *Broadening Participation in STEM*. 2019;22:183-196. doi:10.1108/s1479-364420190000022008
22. Mervis J. Genuine research keeps students in science. *Science (1979)*. 2016;352(6291):1266.
23. Chen X. STEM Attrition: College Students' Paths Into and Out of STEM Fields. *National Center for Education Statistics*. 2013;(Statistical Analysis Report. NCES 2014-001):102. <https://nces.ed.gov/pubs2014/2014001rev.pdf>
24. Chemers MM, Zurbriggen EL, Syed M, Goza BK, Bearman S. The Role of Efficacy and Identity in Science Career Commitment Among Underrepresented Minority Students. *Journal of Social Issues*. 2011;67(3):469-491. doi:10.1111/j.1540-4560.2011.01710.x
25. Robnett RD, Chemers MM, Zurbriggen EL. Longitudinal Associations Among Undergraduates' Research Experience, Self-Efficacy, and Identity. *J Res Sci Teach*. 2015;52(6):847-867. doi:10.1002/tea.21221
26. Rodenbusch SE, Hernandez PR, Simmons SL, Dolan EL. Early engagement in course-based research increases graduation rates and completion of science, engineering, and mathematics degrees. *CBE Life Sci Educ*. 2016;15(2). doi:10.1187/cbe.16-03-0117

27. Carlone HB, Johnson A. Understanding the science experiences of successful women of color: Science identity as an analytic lens. *J Res Sci Teach.* 2007;44(8):1187-1218. doi:10.1002/TEA.20237
28. Herrera FA, Hurtado S, Garcia GA, Gasiewski J. A Model for Redefining STEM Identity For Talented STEM Graduate Students. Published online 2005. Accessed October 5, 2023. <https://www.researchgate.net/publication/265350950>
29. Singer A, Montgomery G, Schmoll S. How to foster the formation of STEM identity: studying diversity in an authentic learning environment. *Int J STEM Educ.* 2020;7(1):1-12. doi:10.1186/S40594-020-00254-Z/TABLES/5
30. Altman RB. A curriculum for bioinformatics: the time is ripe. *Bioinformatics.* 1998;14(7):549-550. doi:10.1093/BIOINFORMATICS/14.7.549
31. Centeno NB, Villà-Freixa J, Oliva B. Teaching Structural Bioinformatics at the Undergraduate Level. *Biochemistry and Molecular Biology Education.* 2003;31(6):386-391. doi:10.1002/bmb.2003.494031060287
32. Gatherer D. Reflections on integrating bioinformatics into the undergraduate curriculum: The Lancaster experience. *Biochemistry and Molecular Biology Education.* 2020;48(2):118-127. doi:10.1002/bmb.21320
33. Bain SA, Plaisier H, Anderson F, *et al.* Bringing bioinformatics to schools with the 4273pi project. *PLoS Comput Biol.* 2022;18(1). doi:10.1371/JOURNAL.PCBI.1009705
34. Wilson Sayres MA, Hauser C, Sierk M, *et al.* Bioinformatics core competencies for undergraduate life sciences education. Dalby AR, ed. *PLoS One.* 2018;13(6):e0196878. doi:10.1371/journal.pone.0196878
35. Lopatto D, Alvarez C, Barnard D, *et al.* Undergraduate research. Genomics Education Partnership. *Science.* 2008;322(5902):684-685. doi:10.1126/SCIENCE.1165351
36. Işık EB, Brazas MD, Schwartz R, *et al.* Grand challenges in bioinformatics education and training. *Nature Biotechnology* 2023 41:8. 2023;41(8):1171-1174. doi:10.1038/s41587-023-01891-9
37. Zhan YA, Wray CG, Namburi S, Glantz ST, Laubenbacher R, Chuang JH. Fostering bioinformatics education through skill development of professors: Big Genomic Data Skills Training for Professors. Ouellette F, ed. *PLoS Comput Biol.* 2019;15(6):e1007026. doi:10.1371/journal.pcbi.1007026
38. Badotti F, Barbosa AS, Reis ALM, do Valle ÍF, Ambrósio L, Bitar M. Comparative modeling of proteins: A method for engaging students' interest in bioinformatics tools. *Biochemistry and Molecular Biology Education.* 2014;42(1):68-78. doi:10.1002/bmb.20721
39. Magana AJ, Taleyarkhan M, Alvarado DR, Kane M, Springer J, Clase K. A Survey of Scholarly Literature Describing the Field of Bioinformatics Education and Bioinformatics Educational Research. *CBE Life Sci Educ.* 2014;13(4):607. doi:10.1187/CBE.13-10-0193

40. Dinsdale E, Elgin SCR, Grandgenett N, *et al.* NIBLSE: A Network for Integrating Bioinformatics into Life Sciences Education. *CBE—Life Sciences Education*. 2015;14(4):14:le3, 1-4. doi:10.1187/cbe.15-06-0123
41. Hernandez PR, Woodcock A, Estrada M, Schultz PW. Undergraduate Research Experiences Broaden Diversity in the Scientific Workforce. *Bioscience*. 2018;68(3):204-211. doi:10.1093/biosci/bix163
42. Rehm HL, Berg JS, Plon SE. ClinGen and ClinVar – Enabling Genomics in Precision Medicine. *Hum Mutat*. 2018;39(11):1473-1475. doi:10.1002/HUMU.23654
43. Oliver GR, Zimmermann MT, Klee EW, Urrutia RA. “The molecule’s the thing”: The promise of molecular modeling and dynamic simulations in aiding the prioritization and interpretation of genomic testing results. *F1000Res*. 2016;5. doi:10.12688/F1000RESEARCH.8600.3
44. Richards S, Aziz N, Bale S, *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-423. <http://dx.doi.org/10.1038/gim.2015.30>
45. Auchincloss LC, Laursen SL, Branchaw JL, *et al.* Assessment of Course-Based Undergraduate Research Experiences: A Meeting Report. *CBE—Life Sciences Education*. 2014;13(1):29-40. doi:10.1187/cbe.14-01-0004
46. Corwin LA, Graham MJ, Dolan EL. Modeling Course-Based Undergraduate Research Experiences: An Agenda for Future Research and Evaluation. *CBE Life Sci Educ*. 2015;14(1):1-13. doi:10.1187/cbe.14-10-0167
47. *Vision and Change - A Call to Action.*; 2009. https://www.aaas.org/sites/default/files/content_files/VC_report.pdf
48. Occupational Outlook Handbook, Computer and Information Research Scientists. Bureau of Labor Statistics, U.S. Department of Labor. Published 2023. Accessed September 5, 2023. <https://www.bls.gov/ooh/computer-and-information-technology/computer-and-information-research-scientists.htm>
49. Madlung A. Assessing an effective undergraduate module teaching applied bioinformatics to biology students. *PLoS Comput Biol*. 2018;14(1). doi:10.1371/journal.pcbi.1005872
50. Williams JJ, Drew JC, Galindo-Gonzalez S, *et al.* Barriers to integration of bioinformatics into undergraduate life sciences education: A national study of US life sciences faculty uncover significant barriers to integrating bioinformatics into undergraduate instruction. *PLoS One*. 2019;14(11):e0224288. doi:10.1371/journal.pone.0224288
51. Mathur V, Arora GS, McWilliams M, Russell J, Rosenwald AG. The Genome Solver Project: Faculty Training and Student Performance Gains in Bioinformatics. *J Microbiol Biol Educ*. 2019;20(1). doi:10.1128/jmbe.v20i1.1607

52. History of EMBL | EMBL.org. Accessed September 13, 2023. <https://www.embl.org/about/history/>
53. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2010;38(Database issue):D46. doi:10.1093/NAR/GKP1024
54. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol.* 1981;147(1):195-197. doi:10.1016/0022-2836(81)90087-5
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol.* 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2
56. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. *Brief Bioinform.* 2019;20(6):1981-1996. doi:10.1093/BIB/BBY063
57. Salter H. Teaching bioinformatics. *Biochem Educ.* 1998;26(1):3-10. doi:10.1016/S0307-4412(97)00163-5
58. Stephan PE, Black G, Agresti J, *et al.* Bioinformatics: Does the US system lead to missed opportunities in emerging fields? A case study. *Sci Public Policy.* 1999;26(6):382-392. doi:10.3152/147154399781782211
59. Smith TM, Emmeluth DS. Introducing Bioinformatics into the Biology Curriculum: Exploring the National Center for Biotechnology Information. *American Biology Teacher.* 2001;64(2):93-99. Accessed September 1, 2015. <http://eric.ed.gov/?id=EJ641922>
60. Feig AL, Jabri E. Incorporation of Bioinformatics Exercises into the Undergraduate Biochemistry Curriculum. *Biochemistry and Molecular Biology Education.* 2002;30.
61. Honts JE. Evolving Strategies for the Incorporation of Bioinformatics Within the Undergraduate Cell Biology Curriculum. *Cell Biol Educ.* 2003;2(4):233-247. doi:10.1187/cbe.03-06-0026
62. Bednarski AE, Elgin SCR, Pakrasi HB. An inquiry into protein structure and genetic disease: Introducing undergraduates to bioinformatics in a large introductory course. *Cell Biol Educ.* 2005;4(3):207-220. doi:10.1187/cbe.04-07-0044
63. Boyle JA. Bioinformatics in undergraduate education: Practical examples. *Biochemistry and Molecular Biology Education.* 2004;32(4):236-238. doi:10.1002/bmb.2004.494032040376
64. Cattley S, Arthur JW. BioManager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training. *Brief Bioinform.* 2007;8(6):457-465. doi:10.1093/bib/bbm039
65. Sutcliffe IC, Cummings SP. Making bioinformatics projects a meaningful experience in an undergraduate biotechnology or biomedical science programme. *Bioscience Education.* 2007;10(1):1-2. doi:10.3108/BEEJ.10.2

66. Drew JC, Triplett EW. Whole Genome Sequencing in the Undergraduate Classroom: Outcomes and Lessons from a Pilot Course. *J Microbiol Biol Educ.* 2008;9(1):3. doi:10.1128/jmbe.v9.89
67. Robertson AL, Phillips AR. Integrating PCR theory and bioinformatics into a research-oriented primer design exercise. *CBE Life Sci Educ.* 2008;7(1):89-95. doi:10.1187/cbe.07-07-0051
68. Howard DR, Miskowski JA, Grunwald SK, Abler ML. Assessment of a Bioinformatics Across Life Science Curricula Initiative. *Biochemistry and Molecular Biology Education.* 2007;35(1):16-23. doi:10.1002/bmb.13
69. Elmore DE, Guayasamin RC, Kieffer ME. A series of molecular dynamics and homology modeling computer labs for an undergraduate molecular modeling course. *Biochemistry and Molecular Biology Education.* 2010;38(4):216-223. doi:10.1002/BMB.20396
70. Gelbart H, Yarden A. Learning genetics through an authentic research simulation in bioinformatics. *J Biol Educ.* 2006;40(3):107-112. doi:10.1080/00219266.2006.9656026
71. Shortlidge EE, Brownell SE. How to Assess Your CURE: A Practical Guide for Instructors of Course-Based Undergraduate Research Experiences. *J Microbiol Biol Educ.* 2016;17(3):399-408. doi:10.1128/jmbe.v17i3.1103
72. Wightman B, Hark AT. Integration of bioinformatics into an undergraduate biology curriculum and the impact on development of mathematical skills. *Biochemistry and Molecular Biology Education.* 2012;40(5):310-319. doi:10.1002/BMB.20637
73. Brame CJ, Pruitt WM, Robinson LC. A Molecular Genetics Laboratory Course Applying Bioinformatics and Cell Biology in the Context of Original Research. *CBE Life Sci Educ.* 2008;7(4):410-421. doi:10.1187/CBE.08-07-0036
74. Jordan TC, Burnett SH, Carson S, *et al.* A Broadly Implementable Research Course in Phage Discovery and Genomics for First-Year Undergraduate Students. *mBio.* 2014;5(1):e01051-13. doi:10.1128/mBio.01051-13
75. Shaffer CD, Alvarez CJ, Bednarski AE, *et al.* A course-based research experience: How benefits change with increased investment in instructional time. *CBE-Life Sciences Education.* 2014;13(1):111-130. doi:10.1187/cbe-13-08-0152
76. Smith JT, Harris JC, Lopez OJ, Valverde L, Borchert GM. “On the Job” Learning: A Bioinformatics Course Incorporating Undergraduates in Actual Research Projects and Manuscript Submissions. *Biochemistry and Molecular Biology Education.* 2015;43(3):154-161. doi:10.1002/bmb.20848
77. Brownell SE, Kloser MJ. Toward a conceptual framework for measuring the effectiveness of course-based undergraduate research experiences in undergraduate biology. *Studies in Higher Education.* 2015;40(3):525-544. doi:10.1080/03075079.2015.1004234

78. Frantz KJ, Demetrikopoulos MK, Britner SL, *et al.* A comparison of internal dispositions and career trajectories after collaborative versus apprenticed research experiences for undergraduates. *CBE Life Sci Educ.* 2017;16(1). doi:<https://doi.org/10.1187/cbe.16-06-0206>
79. Indorf JL, Weremijewicz J, Janos DP, Gaines MS. Adding Authenticity to Inquiry in a First-Year, Research-Based, Biology Laboratory Course. *CBE Life Sci Educ.* 2019;18(3). doi:<https://doi.org/10.1187/cbe.18-07-0126>
80. Shuster MI, Curtiss J, Wright TF, Champion C, Sharifi M, Bosland J. Implementing and Evaluating a Course-Based Undergraduate Research Experience (CURE) at a Hispanic-Serving Institution. *Interdisciplinary Journal of Problem-Based Learning.* 2019;13(2):1. doi:10.7771/1541-5015.1806
81. Dewey J, Evers A, Schuchardt A. Students' Experiences and Perceptions of the Scientific Research Culture after Participating in Different Course-Based Undergraduate Research Experience Models. *CBE Life Sci Educ.* 2022;21(2). doi:10.1187/CBE.21-10-0304
82. Gao L, Guo M. A course-based undergraduate research experience for bioinformatics education in undergraduate students. *Biochemistry and Molecular Biology Education.* 2023;51(2):189-199. doi:10.1002/BMB.21710
83. Dolan EL. *Course-Based Undergraduate Research Experiences: Current Knowledge and Future Directions.*; 2016. https://sites.nationalacademies.org/cs/groups/dbassesite/documents/webpage/dbasse_177288.pdf
84. Lave J, Wenger E. *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press; 1991. doi:<https://doi.org/10.1017/CBO9780511815355>
85. Shaffer CD, Alvarez C, Bailey C, *et al.* The Genomics Education Partnership: Successful Integration of Research into Laboratory Classes at a Diverse Group of Undergraduate Institutions. *CBE Life Sci Educ.* 2010;9(1):55-69. doi:10.1187/09-11-0087
86. Lunetta VN, Hofstein A, Clough MP. Learning and Teaching in the School Science Laboratory: An Analysis of Research, Theory, and Practice. In: Lederman N, Abel S, eds. *Handbook of Research on Science Education.* 1st ed. Lawrence Erlbaum; 2007:393-441. Accessed September 13, 2023. <https://www.taylorfrancis.com/chapters/edit/10.4324/9780203824696-18/learning-teaching-school-science-laboratory-analysis-research-theory-practice-vincent-lunetta-avi-hofstein-michael-clough>
87. Olimpo JT, Fisher GR, Dechenne-Peters SE. Development and evaluation of the tigriopus course-based undergraduate research experience: Impacts on students' content knowledge, attitudes, and motivation in a majors introductory biology course. *CBE Life Sci Educ.* 2016;15(4). doi:10.1187/cbe.15-11-0228
88. Shanle EK, Tsun IK, Strahl BD. A course-based undergraduate research experience investigating p300 bromodomain mutations. *Biochemistry and Molecular Biology Education.* 2016;44(1):68-74. doi:10.1002/BMB.20927

89. Lent RW, Brown SD, Hackett G. Toward a Unifying Social Cognitive Theory of Career and Academic Interest, Choice, and Performance. *J Vocat Behav.* 1994;45(1):79-122. doi:10.1006/JVBE.1994.1027
90. Bandura A. *Social Foundations of Thought and Action: A Social Cognitive Theory.* Prentice-Hall, Inc.; 1986.
91. McDonough E, Sawyer KS, Wilks JC, Jacque B. Students Attitudes Surrounding STEM: A Social Cognitive Career Theory Instrument for High School. *bioRxiv.* Published online 2021. doi:10.1101/2021.11.29.470294
92. Lent RW, Brown SD. On Conceptualizing and Assessing Social Cognitive Constructs in Career Research: A Measurement Guide. *J Career Assess.* 2006;14(1):12-35. doi:10.1177/1069072705281364
93. Betz NE, Hackett G. The relationship of mathematics self-efficacy expectations to the selection of science-based college majors. *J Vocat Behav.* 1983;23(3):329-345. doi:10.1016/0001-8791(83)90046-5
94. Gainor KA, Lent RW. Social Cognitive Expectations and Racial Identity Attitudes in Predicting the Math Choice Intentions of Black College Students. *J Couns Psychol.* 1998;45(4):403-413. doi:10.1037/0022-0167.45.4.403
95. Bandura A. Self-efficacy: Toward a unifying theory of behavioral change. *Psychol Rev.* 1977;84(2):191-215. doi:10.1037/0033-295X.84.2.191
96. Bandura A. Self-efficacy mechanism in human agency. *American Psychologist.* 1982;37(2):122-147. doi:10.1037/0003-066X.37.2.122
97. Bandura A. *Self-Efficacy: The Exercise of Control.* W.H. Freeman; 1997. doi:10.4278/0890-1171-12.1.8
98. Pajares F. Self-Efficacy Beliefs in Academic Settings. *Rev Educ Res.* 1996;66(4):543. doi:10.2307/1170653
99. Lopez-Garrido G. Bandura's Self-Efficacy Theory Of Motivation In Psychology. SimplyPsychology. Published 2023. Accessed October 6, 2023. <https://www.simplypsychology.org/self-efficacy.html>
100. Bandura A. Perceived Self-Efficacy in Cognitive Development and Functioning. *Educ Psychol.* 1993;28(2):117-148. doi:10.1207/S15326985EP2802_3
101. Schunk DH, DiBenedetto MK. Motivation and social cognitive theory. *Contemp Educ Psychol.* 2020;60:101832. doi:10.1016/J.CEDPSYCH.2019.101832
102. Kovarik DN, Patterson DG, Cohen C, *et al.* Bioinformatics education in high school: Implications for promoting Science, Technology, Engineering, and Mathematics Careers. *CBE Life Sci Educ.* 2013;12(3):441-459. doi:10.1187/cbe.12-11-0193
103. Erikson EH. *Identity, Youth, and Crisis.* 1st ed. W. W. Norton; 1968.

104. Sica LS, Sestito LA. Personal Skills for Optimal Identity Development: A Person-Centered Approach in Italian Late-Adolescents. *J Pers Oriented Res*. 2021;7(1):36. doi:10.17505/JPOR.2021.23450
105. Brickhouse N, Lowery PA, Schultz K. What Kind of a Girl Does Science? The Construction of School Science Identities. *J Res Sci Teach*. Published online 2000. doi:10.1002/(SICI)1098-2736(200005)37:5
106. Aschbacher PR, Li E, Roth EJ. Is science me? High school students' identities, participation and aspirations in science, engineering, and medicine. *J Res Sci Teach*. 2010;47(5):564-582. doi:10.1002/tea.20353
107. Hazari Z, Sonnert G, Sadler PM, Shanahan MC. Connecting high school physics experiences, outcome expectations, physics identity, and physics career choice: A gender study. *J Res Sci Teach*. 2010;47(8):978-1003. doi:10.1002/tea.20363
108. Perez T, Cromley JG, Kaplan A. The Role of Identity Development, Values, and Costs in College STEM Retention. *J Educ Psychol*. 2014;106(1):315-329. doi:10.1037/a0034027
109. Tobias S. They're not dumb. They're different.: A new "tier of talent" for science. *Change: The Magazine of Higher Learning*. 1990;22(4):11-30.
110. Seymour E, Hewitt NM. *Talking about Leaving*. Westview Press; 1997. <http://www-personal.umich.edu/~glarose/dept/teaching/resources/talkingAboutLeavingSeymourHewittOverview.pdf>
111. Huffmyer AS, O'Neill T, Lemus JD. Evidence for Professional Conceptualization in Science as an Important Component of Science Identity. *CBE Life Sci Educ*. 2022;21(4). doi:https://doi.org/10.1187/cbe.20-12-0280
112. Hunter AB, Laursen SL, Seymour E. Becoming a scientist: The role of undergraduate research in students' cognitive, personal, and professional development. *Sci Educ*. 2006;91(1):36-74. doi:10.1002/sce.20173
113. Nagda BA, Gregerman SR, Jonides J, Hippel W von, Lerner JS. Undergraduate Student-Faculty Research Partnerships Affect Student Retention. *Rev High Ed*. 1998;22(1):55-72. Accessed November 5, 2020. https://conferences.wcfia.harvard.edu/sites/projects.iq.harvard.edu/files/lernerlab/files/nagda_gregerman_jonides_vonhippel_lerner_1998.pdf
114. Wang D, Liu X, Deng H. The perspectives of social cognitive career theory approach in current times. *Front Psychol*. 2022;13. doi:10.3389/FPSYG.2022.1023994/PDF
115. Lent RW, Miller MJ, Smith PE, Watford BA, Lim RH, Hui K. Social cognitive predictors of academic persistence and performance in engineering: Applicability across gender and race/ethnicity. *J Vocat Behav*. 2016;94:79-88. doi:10.1016/j.jvb.2016.02.012
116. Campbell AM. Meeting report: genomics in the undergraduate curriculum--rocket science or basic science? *Cell Biol Educ*. 2002;1(3):70-72. doi:10.1187/CBE.02-06-0014

117. Holtzclaw JD, Eisen A, Whitney EM, Penumetcha M, Hoey JJ, Kimbro KS. Incorporating a New Bioinformatics Component into Genetics at a Historically Black College: Outcomes and Lessons. *CBE—Life Sciences Education*. 2006;5(1):52-64. doi:10.1187/cbe.05-04-0071
118. Campbell AM. Public access for teaching genomics, proteomics, and bioinformatics. *Cell Biol Educ*. 2003;2(2):98-111. doi:https://doi.org/10.1187/cbe.03-02-0007
119. Takayama K. Three-Dimensional Visualizations in Teaching Genomics and Bioinformatics: Mutations in HIV Envelope Proteins and Their Consequences for Vaccine Design. *Microbiol Educ*. 2004;5(1):3-12. doi:10.1128/jmbe.v5.72
120. Boomer SM, Lodge DP, Dutton BE. Bacterial Diversity Studies Using the 16S rRNA Gene Provide a Powerful Research-Based Curriculum for Molecular Biology Laboratory. *Microbiol Educ*. 2002;3(1):18-25. doi:10.1128/ME.3.1.18-25.2002
121. Lopatto D, Alvarez C, Barnard D, *et al*. Genomics Education Partnership. *Science* (1979). 2008;322(5902):684-685. doi:10.1126/science.1165351
122. Hauser C, Bailey C, Elgin SC, Lopatto D. The Genomics Education Partnership. *The FASEB Journal*. 2009;23(S1):84.5-84.5. doi:10.1096/FASEBJ.23.1_SUPPLEMENT.84.5
123. Reeves NL, Barral AM, Sharif KA, *et al*. The Genomics Education Partnership: Assessing and Improving a Course-based Undergraduate Research Experience (CURE). *The FASEB Journal*. 2016;30(S1):lb197-lb197. doi:10.1096/FASEBJ.30.1_SUPPLEMENT.LB197
124. Goodman AL, Stry M Van, Lopatto D, Leung W, Reed LK, Community GF. Growing and Sustaining a Nationwide CURE: Genomics Education Partnership Enhances Research Opportunities for Students and Faculty at Diverse Institutions. *The FASEB Journal*. 2020;34(S1):1-1. doi:10.1096/FASEBJ.2020.34.S1.00672
125. Lopatto D, Rosenwald AG, Burgess RC, *et al*. Student Attitudes Contribute to the Effectiveness of a Genomics CURE. *J Microbiol Biol Educ*. 2022;23(2). doi:10.1128/JMBE.00208-21/SUPPL_FILE/JMBE.00208-21-S001.PDF
126. Lopatto D, Rosenwald AG, DiAngelo JR, *et al*. Facilitating Growth through Frustration: Using Genomics Research in a Course-Based Undergraduate Research Experience. *J Microbiol Biol Educ*. 2020;21(1). doi:10.1128/JMBE.V21I1.2005
127. Findlely-Van Nostrand D, Pollenz RS. Evaluating psychosocial mechanisms underlying STEM persistence in undergraduates: Evidence of impact from a six-day pre-college engagement STEM academy program. *CBE Life Sci Educ*. 2017;16(2). doi:10.1187/cbe.16-10-0294
128. Syed M, Zurbriggen EL, Chemers MM, *et al*. The Role of Self-Efficacy and Identity in Mediating the Effects of STEM Support Experiences. *Analyses of Social Issues and Public Policy*. 2019;19(1):7-49. doi:10.1111/asap.12170
129. SEA-PHAGES | Home. Accessed September 15, 2023. <https://seaphages.org/>

130. Lopatto D. Survey of Undergraduate Research Experiences (SURE): First findings. *Cell Biol Educ.* 2004;3(4):270-277. doi:10.1187/cbe.04-07-0045
131. Lopatto D. CURE – Undergraduate Research Experience Surveys. Grinnell College. Accessed October 6, 2023. <https://sure.sites.grinnell.edu/cure-survey/>
132. Newell MJ, Ulrich PN. Gains in Scientific Identity, Scientific Self-Efficacy, and Career Intent Distinguish Upper-Level CUREs from Traditional Experiences in the Classroom. *J Microbiol Biol Educ.* 2022;23(3). doi:10.1128/JMBE.00051-22
133. Estrada M, Woodcock A, Hernandez PR, Schultz PW. Toward a Model of Social Influence that Explains Minority Student Integration into the Scientific Community. *J Educ Psychol.* 2011;103(1):206-222. doi:10.1037/A0020743
134. Harrison M, Dunbar D, Ratmansky L, Boyd K, Lopatto D. Classroom-based science research at the introductory level: Changes in career choices and attitude. *CBE Life Sci Educ.* 2011;10(3):279-286. doi:10.1187/cbe.10-12-0151
135. Thiry H, Weston TJ, Laursen SL, Hunter AB. The Benefits of Multi-Year Research Experiences: Differences in Novice and Experienced Students' Reported Gains from Undergraduate Research. Ledbetter MLS, ed. *CBE—Life Sciences Education.* 2012;11(3):260-272. doi:10.1187/cbe.11-11-0098
136. Beck C, Butler A, da Silva KB. Promoting inquiry-based teaching in laboratory courses: Are we meeting the grade? *CBE Life Sci Educ.* 2014;13(3). doi:10.1187/cbe.13-12-0245
137. Brazas MD, Ouellette BFF. Continuing Education Workshops in Bioinformatics Positively Impact Research and Careers. *PLoS Comput Biol.* 2016;12(6). doi:10.1371/JOURNAL.PCBI.1004916
138. Prokop JW, Lazar J, Crapitto G, Smith DC, Worthey EA, Jacob HJ. Molecular Modeling in the Age of Clinical Genomics, The Enterprise of the Next Generation. *J Mol Model.* 2017;23(3). doi:10.1007/s00894-017-3258-3
139. Landrum MJ, Lee JM, Riley GR, *et al.* ClinVar: Public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42(D1):D980-D985. doi:10.1093/nar/gkt1113
140. Martinez X, Chavent M, Baaden M. Visualizing protein structures — tools and trends. *Biochem Soc Trans.* 2020;48(2):499-506. doi:10.1042/BST20190621
141. Lombardi MM. Authentic Learning for the 21st Century: An Overview Authentic Learning for the 21st Century. Educause Learning Initiative. Published 2007. Accessed August 18, 2023. <https://library.educause.edu/resources/2007/1/authentic-learning-for-the-21st-century-an-overview>
142. Welch L, Lewitter F, Schwartz R, *et al.* Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies. *PLoS Comput Biol.* 2014;10(3). doi:10.1371/journal.pcbi.1003496

143. Cummings MP, Temple GG. Broader incorporation of bioinformatics in education: Opportunities and challenges. *Brief Bioinform.* 2010;11(6):537-543. doi:10.1093/bib/bbq058
144. Krieger E. YASARA - Yet Another Scientific Artificial Reality Application. Published online 2018. Accessed July 7, 2020. <http://yasara.org/>
145. Joy Shepard C, Cline SG, Hinds D, Jahanbakhsh S, Prokop JW. Breakdown of multiple sclerosis genetics to identify an integrated disease network and potential variant mechanisms. *Physiol Genomics.* 2019;51(11):562-577. doi:10.1152/physiolgenomics.00120.2018
146. Sanders M, Lawlor MJ, Li X, *et al.* Genomic, transcriptomic, and protein landscape profile of CFTR and cystic fibrosis. *Hum Genet.* 2020;140:423-439 (2021). doi:10.1007/s00439-020-02211-w
147. Gupta R, Charron J, Stenger CL, *et al.* SARS-CoV-2 (COVID-19) structural and evolutionary dynamicome: Insights into functional evolution and human genomics. *Journal of Biological Chemistry.* 2020;295(33):11742-11753. doi:10.1074/jbc.ra120.014873
148. Bauss J, Morris M, Shankar R, *et al.* CCR5 and Biological Complexity: The Need for Data Integration and Educational Materials to Address Genetic/Biological Reductionism at the Interface of Ethical, Legal, and Social Implications. *Front Immunol.* 2021;12. doi:10.3389/fimmu.2021.790041
149. Creswell JW, Plano Clark VL. *Designing and Conducting Mixed Methods Research.* Sage Publications; 2007. Accessed August 1, 2023. <https://onlinelibrary.wiley.com/doi/10.1111/j.1753-6405.2007.00096.x>
150. Singleton RA, Straits BC. *Approaches to Social Research.* 4th ed. Oxford University Press; 2004.
151. Seymour E, Wiese D, Hunter AB, Daffinrud S. *Student Assessment of Learning Gains (SALG) CAT.*; 1999. Accessed November 12, 2020. <http://archive.wceruw.org/cl1/flag/extra/download/cat/salg/salg.pdf>
152. Kuchynka S, Nostrand DF Van, Pollenz RS. Evaluating Psychosocial Mechanisms Underlying STEM Persistence in Undergraduates: Scalability and Longitudinal Analysis of Three Cohorts from a Six-Day Pre-College Engagement STEM Academy Program. *CBE Life Sci Educ.* 2019;18(3):18:ar41, 1-13. doi:10.1187/cbe.19-01-0028
153. Wu MJ, Zhao K, Fils-Aime F. Response rates of online surveys in published research: A meta-analysis. *Computers in Human Behavior Reports.* 2022;7:100206. doi:10.1016/J.CHBR.2022.100206
154. Tavakol M, Dennick R. Making sense of Cronbach's alpha. *Int J Med Educ.* 2011;2:53. doi:10.5116/IJME.4DFB.8DFD
155. Kardash CAM. Evaluation of an undergraduate research experience: Perceptions of undergraduate interns and their faculty mentors. *J Educ Psychol.* 2000;92(1):191-201. doi:10.1037/0022-0663.92.1.191

156. e-Handbook of Statistical Methods. NIST/SEMATECH. doi:10.18434/M32189
157. Bremer M, Doerge RW. *Statistics at the Bench: A Step-by-Step Handbook for Biologists*. Cold Spring Harbor Laboratory Press; 2010.
<https://books.google.com/books?id=SCIFAQAIAAJ>
158. Driscoll D, Appiah-Yeboah A, Salib P, Rupert DJ. Merging Qualitative and Quantitative Data in Mixed Methods Research: How To and Why Not. *Ecological and Environmental Anthropology*. 2007;3(1):19-28.
<https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1012&context=icwdmeea>
159. Harrell MC, Bradley MA. *Data Collection Methods: Semi-Structured Interviews and Focus Groups*. RAND Corporation; 2009. Accessed October 6, 2023.
https://www.rand.org/pubs/technical_reports/TR718.html
160. David M. *Fundamentals of Analysis*.; 2019. Accessed September 26, 2023.
<https://dataschool.com/fundamentals-of-analysis/what-is-the-interquartile-range/>
161. Brown SD, Lent RW. A Social Cognitive Framework for Career Choice Counseling. *Career Development Quarterly*. 1996;44(4):354-366. doi:10.1002/J.2161-0045.1996.TB00451.X
162. Goode CT, Britner SL, Gagné P, *et al*. Scientific research self-efficacy among undergraduates: Current contexts and approaches for measurement. In: *Self-Efficacy in School and Community Settings*. Nova Science Publishers; 2012:21-52.
163. Mickelson RA, Mikkelsen I, Dorodchi M, Cukic B, Horn T. Fostering Greater Persistence Among Underserved Computer Science Undergraduates: A Descriptive Study of the I-PASS Project. *J Coll Stud Ret*. Published online March 31, 2022.
doi:<https://doi.org/10.1177/15210251221086928>
164. Chambers B, Salter A, Muldrow L. Getting Past the Gateway: An Exploratory Case on Using Utilitarian Scientific Literacy to Support First-Year Students At-Risk of Leaving STEM. *Education Sciences 2019, Vol 9, Page 265*. 2019;9(4):265. doi:10.3390/EDUCSCI9040265
165. Syed M, Goza BK, Chemers MM, Zurbriggen EL. Individual Differences in Preferences for Matched-Ethnic Mentors Among High-Achieving Ethnically Diverse Adolescents in STEM. *Child Dev*. 2012;83(3):896-910. doi:10.1111/J.1467-8624.2012.01744.X
166. Chang MJ, Kevin Eagan M, Lin MH, Hurtado S. Considering the Impact of Racial Stigmas and Science Identity: Persistence Among Biomedical and Behavioral Science Aspirants. *J Higher Educ*. 2011;82(5):564. doi:10.1353/JHE.2011.0030
167. Fouad NA, Santana MC. SCCT and Underrepresented Populations in STEM Fields. *J Career Assess*. 2016;25(1):24-39. doi:<https://doi.org/10.1177/1069072716658324>
168. Gayles JG, Ampaw F. The Impact of College Experiences on Degree Completion in STEM Fields at Four-Year Institutions: Does Gender Matter? *J Higher Educ*. 2014;85(4):439-468. doi:10.1080/00221546.2014.11777336

169. Shaw EJ, Barbuti SM. Patterns of Persistence in Intended College Major with a Focus on STEM Majors. *NACADA Journal*. 2010;30(2):19-34. doi:10.12930/0271-9517-30.2.19
170. Wang X. Why Students Choose STEM Majors: Motivation, High School Learning, and Postsecondary Context of Support. *Am Educ Res J*. 2013;50(5). doi:10.3102/0002831213488622
171. Cohen C, Patterson D, Kovarik D, Kappan JCWS, 2013 U. Fostering STEM career awareness: emerging opportunities for teachers. *WA State Kappan*. 2013;6:1-17. Accessed October 6, 2023. [https://www.nwabr.org/sites/default/files/pagefiles/PDK STEM Career Awareness.pdf](https://www.nwabr.org/sites/default/files/pagefiles/PDK%20STEM%20Career%20Awareness.pdf)
172. Maloney M, Parker J, Leblanc M, Woodard CT, Glackin M, Hanrahan M. Bioinformatics and the undergraduate curriculum. *CBE-Life Sciences Education*. 2010;9(3):172-174. doi:10.1187/cbe.10-03-0038
173. Lefever S, Dal M, Matthíasdóttir Á. Online data collection in academic research: advantages and limitations. *British Journal of Educational Technology*. 2007;38(4):574-582. doi:10.1111/J.1467-8535.2006.00638.X
174. Wright KB. Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*. 2005;10(3). doi:10.1111/J.1083-6101.2005.TB00259.X/4614509
175. Subedi B. Using Likert Type Data in Social Science Research: Confusion, Issues and Challenges. *International Journal of Contemporary Applied Sciences*. 2016;3(2):36-49. Accessed September 29, 2023. <https://ijcar.net/assets/pdf/Vol3-No2-February2016/02.pdf>
176. Theofanidis D, Fountouki A. Limitations and Delimitations in the Research Process. *Perioperative Nursing - Quarterly scientific, online official journal of GORNA*. 2018;7(3):155-163. doi:10.5281/ZENODO.2552022
177. Stake JE, Nickens SD. Adolescent girls' and boys' science peer relationships and perceptions of the possible self as scientist. *Sex Roles*. 2005;52(1-2):1-11. doi:10.1007/S11199-005-1189-4/METRICS
178. Cohen GL, Garcia J. Identity, belonging, and achievement: A model, interventions, implications. *Curr Dir Psychol Sci*. 2008;17(6):365-369. doi:10.1111/J.1467-8721.2008.00607.X
179. Robnett RD, Leaper C. Friendship Groups, Personal Motivation, and Gender in Relation to High School Students' STEM Career Interest. *Journal of Research on Adolescence*. 2013;23(4):652-664. doi:10.1111/JORA.12013
180. Buonaccorsi VP, Boyle MD, Grove D, et al. GCAT-SEEKquence: Genome Consortium for Active Teaching of Undergraduates through Increased Faculty Access to Next-Generation Sequencing Data. *CBE Life Sci Educ*. 2011;10(4):342-345. doi:10.1187/cbe.11-08-0065

181. Buonaccorsi V, Peterson M, Lamendella G, *et al.* Vision and Change through the Genome Consortium for Active Teaching Using Next-Generation Sequencing (GCAT-SEEK): Letter to the Editor. *CBE—Life Sciences Education*. 2014;13:1-2. doi:10.1187/cbe.13-10-01

Appendix A. CODE Facilitator Training Workshop Schedule Example



Facilitator Training Workshop
HudsonAlpha Institute for Biotechnology
Paul Propst Center
800 Hudson Way NW, Huntsville, AL 35806

Day 1

Time	Event
8:30 am	Breakfast and check-in
9:00 am	Welcome and introductions
	Intro to HudsonAlpha and CODE
	Facilitator Training Workshop Pre-Survey - online link
	Technology check – Loading YASARA, project files on the flash drive, Wi-Fi
	Important forms – in the provided folder
	Amino Acid Tubers Activity
	Student Project Workflow • High level overview • More details on Friday!
	Break
	Variants and Their Characterization
	Group Photo - Lobby Staircase
12:00	Lunch
12:45	Walk-through of a CODE Mini-project • We would like for 8 participants to volunteer to share their mini-project with the group tomorrow. Willing to show us your skills? <i>Please sign-up to present.</i>
	Time to build your Simple ClinVar slide presentation
	Break

Time	Event
	Finding a VUS of Interest – pg. xx in Handbook ● NCBI Gene & ClinVar ● HudsonAlpha VUS List – Link in VUS List folder on flash drive
	UniProt Tutorial Finding TP53 structures online – pg. xx in Handbook
	Intro to YASARA Structure software CFTR Tutorial and exercise
	Exercise: Working with a PDB file
	Homology Modeling - pg. xx in Handbook
	Exercise: Homology Modeling with BRCA1
	Exercise: I-TASSER with NOD2 - pg. xx in Handbook
	Exercise: Making a Mutation (swapping residues) – pg. xx in Handbook
	Exercise: Energy Minimization with net1 – pg. xx in Handbook
	Pathogenicity Prediction Software – pg. xx in Handbook SIFT, Polyphen2, REVEL, Mutation Assessor, CADD, GERP, RVIS
5:00	Leave for hotel
6:00	Dinner

Day 2

Time	Event
8:00 am	Breakfast and check-in
	Mini-Project Presentations
	Do You Remember How to . . . in YASARA?
	Curriculum Modules for Student Projects
	Break
	Molecular Dynamic Simulations – pg. xx in Handbook
	Exercise: Molecular Dynamic Simulation (MDS) with TP53 – pg. xx in Handbook
	How to Request an MDS run from HudsonAlpha – pg. xx in Handbook

Time	Event
12:00	Lunch
12:45 pm	Analysis of MDS results of TP53 ● Follow Handbook protocol pg. xx
	Discussion, Questions, and Brainstorming
	Facilitator Training Workshop Post-Survey – online link
3:00	Training ends

Appendix B. IRB Protocol and Certificate of Action Approvals for Study

HudsonAlpha Institute for Biotechnology

IRB Protocol # 0106

PROTOCOL: Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format

DATE: November 15, 2022

SPONSOR: HudsonAlpha Institute for Biotechnology

SPONSOR PROTOCOL NO.: 0106

**PRINCIPAL INVESTIGATOR
AND PRIMARY SITE:**

Michele C. Morris
Director for Workforce Development
(256) 327-0576, mmorris@hudsonalpha.org
HudsonAlpha Institute for Biotechnology
The Paul Propst Center
800 Hudson Way, Huntsville, AL 35806

I. Purpose

This study aims to determine the effectiveness of an undergraduate research program at introducing bioinformatics and increasing awareness, interest, and persistence in STEM fields.

II. Background

The Characterizing Our DNA Exceptions (CODE) program introduces authentic bioinformatics research to undergraduate students from schools historically underrepresented in this type of experience. Students participating in CODE use computational modeling and database analyses to discover the impact of DNA differences (“variants”) found in clinical and agricultural studies. CODE faculty mentors (“facilitators”) receive training to strengthen their ability and confidence to present and teach the bioinformatics content.

This project will build a network of faculty at small colleges and universities across the Southeastern US and provide the educational resources to implement entry-level bioinformatics and protein modeling experiences to their students. The faculty facilitators will implement CODE programs to engage students in bioinformatics research and increase the number and diversity of those persisting in STEM, particularly in bioinformatics and related fields.

Program evaluation will measure shifts in student awareness and interest in bioinformatics, self-efficacy, and scientific identity - all factors that positively impact

STEM persistence. This project's success will provide a sustainable model for low-cost, authentic undergraduate bioinformatics research projects for institutions of all sizes.

Study design

Undergraduate students will participate in a research project based on the CODE parameters under the guidance of a CODE-trained faculty member, referred to as a Facilitator. The students will complete online assessment surveys before and after the project. We will invite a random subset of 3-6 students from each school to video conference interviews. These surveys and interviews will assess changes in student knowledge of bioinformatics concepts, interest in science as a career, and their intention to and persistence towards remaining in a science field.

We will ask the faculty members attending CODE Facilitator Training Workshops to complete an online assessment survey before the training begins, immediately after the workshop, and 6-9 months later. These surveys assess changes in bioinformatics knowledge, comfort with mentoring bioinformatics research projects, and the quality and value of the CODE Training Workshop.

Intent

The CODE assessment surveys intend to improve the program and generate knowledge about its effectiveness at introducing bioinformatics as an undergraduate research project and increasing awareness, interest, and persistence in STEM fields. This study expands the knowledge base of a scientific discipline and produces results that apply to a larger population beyond the data collection sites and the specific subjects studied. It can be used to develop, test, or support theories, principles, and statements of relationships or inform policy beyond the study. Additionally, the project is a preliminary, exploratory, or research development activity.

III. Criteria for Subject Selection

Number of subjects

- Faculty Facilitator participants: Up to 75 Facilitators, including faculty, postdoctoral researchers, and graduate students.
- Student participants: Up to 800 undergraduates participating in CODE projects.

Gender of subjects

Facilitators and students who meet the inclusion criteria will not be excluded based on gender.

Age of subjects

All subjects will be adults 19 years of age or older.

Racial and ethnic origin

Facilitators and students who meet the inclusion criteria will not be excluded based on racial or ethnic origin.

Inclusion criteria

- Undergraduate students participating in a class or independent project mentored by a CODE facilitator.
- Faculty members participating in a CODE Facilitator Training Workshop.

- Access to the internet, email, and a computer that supports video conferencing (for those who wish to participate in interviews).

Exclusion criteria

- Anyone who is less than 19 years of age.

Vulnerable subjects

- Pregnant women: Pregnant women may participate as a mentor or student. We will not exclude pregnant women from this study, and we will not ascertain the pregnancy status of the subjects. Pregnancy status should not impact study results. Participating in the study does not impose a greater risk to pregnant women than other participants, such as women of childbearing age.
- Children: We will exclude children from this study.
- Prisoners: We will exclude prisoners from this study.
- Employees: We will exclude HudsonAlpha employees from the study.
- Students: We will recruit undergraduate students participating in CODE programs at various colleges and universities. Study participation is not required as a condition of taking a class or participating in a CODE program. Trained CODE facilitators, who may be the students' teachers or mentors, may help recruit students.

IV. Methods and Procedures

Study procedure

The evaluation plan for the CODE project primarily addresses changes in student awareness and interest in bioinformatics, science self-efficacy, scientific identity, and the intention to enter or remain in a STEM major. Additionally, the study evaluates the impacts of the Facilitator Training Workshops on instructor knowledge and comfort level with instruction and feedback about the quality of the workshop.

The summative evaluation assesses whether the program positively shapes awareness and interest in informatics and positively impacts student-reported measures of self-efficacy and resilience in the STEM career pipeline. We will use a sequential mixed methods design of quantitative online surveys followed by qualitative interviews. The interviews will help further explain student survey responses and add a deeper understanding of the research experience's impacts.

This project will gather information through online surveys and teleconference interviews. The pre-CODE student survey collects general demographic information and data on their interest in and comfort with bioinformatics, their confidence level with science research, their sense of being part of the scientific community, their interest in STEM fields as careers, and their intention to persist in STEM. The post-CODE survey consists of the same questions, minus the demographic information. The study coordinator will de-identify and code the survey data to allow pairing of pre- and post-CODE surveys.

We will invite a random subset of 3-6 survey respondents from each school to participate in teleconference interviews. The discussions will be held online through Zoom, audio-

recorded, and transcribed. The interviewers will use a script of relevant open-ended questions derived directly from the survey measures. All interview participants will be asked the same questions, and interviewers will follow up on participant comments for additional details as needed.

Participant identities will be coded, and the interview results will be non-identifiable to all but the study coordinator. The coordinator will store the key linking names to codes in a password-protected digital file or a locked cabinet accessible only by the coordinator. The coordinator will store coded recordings and transcripts in a password-protected digital file accessible only by the research study personnel who need access to these files.

Faculty members attending a CODE Facilitator Training Workshop will be invited to complete online surveys before the training begins, immediately following the workshop, and 6-9 months later. These surveys assess changes in knowledge of bioinformatics, comfort mentoring bioinformatics research projects, and the quality and value of the CODE Training Workshop.

The pre-workshop survey collects demographic information and data on the Facilitators' comfort with bioinformatics. The post-workshop survey is expanded to assess their estimation of the value of the workshop and includes open-ended questions regarding the workshop content. The post-workshop surveys do not collect demographic information. The extended follow-up survey administered 6-9 months after the workshop consists of the same Likert-scale questions as the post-workshop survey evaluating the value of the CODE training.

As with the students, the study coordinator will remove personal identifiers and assign a subject code to protect the respondent's confidentiality and allow for pairing pre-/post- and extended follow-up facilitator surveys. Results will be non-identifiable to all but the study coordinator and will be stored in the same manner as the student data.

Data Variables

The evaluation model incorporates outcomes for student participants and faculty facilitators who will mentor student research projects. Repeated assessments (pre-project, post-project, and interviews) will provide data on student shifts in opinions and confidence levels. Student survey questions use Likert scales and include ranges of agreement, activity levels in research actions, and confidence levels. The Facilitator surveys consist of Likert-scale questions to indicate their agreement with statements concerning their comfort with bioinformatics and opinions about the value of the workshop. The post-workshop survey includes open-ended questions regarding the workshop content.

The student survey analysis results will be integrated with results from the interviews, creating a subset of cases with both survey data and a transcript. All the data collected on a single case will be studied together and examined in detail - for example, comparing a student's survey responses to their interview transcript. Findings will

include statistical information from the quantitative surveys and quotes from students about how they perceive the impact of participating in CODE. Facilitator responses will be compared across their surveys to measure shifts in their confidence levels and to assess the value of the CODE training workshops.

Data sources

Data will be obtained from participant survey responses (undergraduate students and faculty mentors) and audio-recorded video-conference interviews.

Identifiability of data and specimens

Only the study coordinator will have access to the names and emails of the students and facilitators. The coordinator will only use names and emails to match surveys answered by the same participants and interview responses. The study coordinator will have access to survey responses and corresponding subject names. She will anonymize survey responses by assigning each a subject number. She will maintain the key to the code and all identifiers separately from the survey responses, interview transcript, and any other study data.

No biological specimens will be collected.

Data storage and confidentiality

Students and facilitators can only complete the online surveys if they receive a URL link from the study coordinator. All personal identifying information will be kept strictly confidential by the coordinator, and they will code all research data to exclude names, emails, and other identifying information.

The electronic database will be available only to study researchers. The online tools (Qualtrics, NVIVO, Google Forms) are accessible by the study coordinator's private, password-protected accounts. We will not identify study participants by name or use information from which their identities could be inferred in any research or evaluation report, publication, presentation, or article based on information gathered during the project. We will retain all data for at least three years in compliance with federal regulations.

V. Risk/Benefit Assessment

Risk category

The research presents minimal risk to research participants.

Potential risk

We will ask study participants to share their experience participating as students in an undergraduate research project or as facilitators for these research projects. Depending on their experience and how much they wish to share, some participants may experience anxiety or distress. Any discomfort experienced is not expected to exceed that encountered in everyday life. We anticipate no physical, psychological, or sociological risks.

Protection against risks

Participation is voluntary. Any subject who perceives a more than minimal risk to participating in the surveys can simply choose not to participate. For those who do

participate, all survey questions are optional. They may decide not to answer questions or end their participation at any moment during the surveys. There will be no negative consequence to ending participation.

Names and contact information will be stored solely for the study coordinator to contact the participants to ask or remind them to complete a survey. The study coordinator will code survey responses and maintain a key to the code in a locked cabinet or a password-protected file. Consent forms will be collected online and will be stored in password-protected electronic folders. Only the study coordinator will have access to names and emails. All survey responses will be de-identified, and any information that could identify the participant will be removed.

Potential benefits to the participants

Participants receive no direct benefits, and increased general knowledge is the most likely project outcome. Responses to the surveys may help future students who participate in similar undergraduate research projects. Some people find it beneficial to be able to share and evaluate their research experience.

Alternatives to participation

There are no alternative courses of action available to subjects who elect not to participate in the study. The alternative to participating is not to participate.

VI. Study Population and Recruitment

Method of subject identification and recruitment

We will invite CODE facilitators and students to participate in the study. Facilitators will be invited by email to complete the online workshop surveys when they sign up to attend a CODE Facilitator Training Workshop. They will receive an email 6-9 months later after the workshop asking them to complete the Extended Follow-up Survey. Facilitators trained before the start of this study may be asked by email to participate in the Extended Follow-up Survey.

Depending on the number of survey responses received, we may send up to three weekly follow-up reminder emails to all current or prospective participants who have not completed the surveys.

The Facilitator or study coordinator will communicate to the CODE students that they have the opportunity to participate in a research study that uses survey methods to evaluate their experience with CODE. We will explain that the study is optional and does not impact their participation in CODE.

The study coordinator will provide a URL to a Google form or similar secure platform for the students to volunteer their name and email to receive the email invitation to the study. The invitation email will include a unique link to the survey consent. Students who provide electronic consent will then be directed to complete the survey.

No preliminary data or biological specimens will be collected for recruiting and screening participants.

Process of consent

Consent will be obtained from all survey participants using an electronic Consent Form. Facilitators and students will receive an email with a unique link to a consent form. If they choose not to consent to participate, they can select that option, and they will not access the survey. If they consent to participate, they will proceed to the survey. The completed consent form data will be stored electronically in a password-protected file only accessible by the study coordinator.

Prospective interview subjects will receive an email invitation explaining the format and purpose of the interview. The email will have an attached document with the consent language that will be used to obtain verbal consent from participants prior to beginning the interview. Those who are interested in participating will follow a URL link to provide their name, email address, and phone number so that the study coordinator can schedule their interview. The study coordinator will use a script of the consent language provided in the recruitment email to obtain verbal consent for the interview participants.

We will not deliberately withhold information or provide false information to the subjects during the consent process.

Subject capacity

We expect that all subjects will have the capacity to give informed consent. Subjects that require the consent of a legally authorized representative (LAR) will be excluded from the study.

Subject comprehension

To participate in the study, potential participants will have to read and sign a consent form or provide verbal consent after having a consent form read to them (for teleconference interviews). Contact information for the principal investigator and study coordinator will be provided on the consent form so that potential participants may reach out with questions before giving their consent to participate. Participants providing verbal consent will have ample time to ask questions.

Costs to the subject

The subjects should not incur any cost as a result of participating in the study. The sponsor will not be responsible for any costs incurred by the participants.

Payment for participation

The subjects will not receive any payment or compensation for participating in the study.

Date: _____

Signature: _____
Michele Morris, Principal Investigator

Lamb, Neil - Certificate of Action for Study#_ 1313664, Panel Meeting#_ 313334,
WO#_ 1-1453737-1



Certificate of Action

Investigator Name: Neil E. Lamb, PhD	Board Action Date: 07/26/2021
Investigator Address: The Paul Propst Center, 800 Hudson Way Huntsville, AL 35806, United States	Approval Expires: 07/26/2022 Continuing Review Frequency: No CR Required
Sponsor: HudsonAlpha Institute for Biotechnology Institution Tracking Number:	Sponsor Protocol Number: 0106 Amended Sponsor Protocol Number:
Study Number: 1313664	IRB Tracking Number: 20213554
Work Order Number: 1-1453737-1	
Protocol Title: Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format	

THE FOLLOWING ITEMS ARE APPROVED:

Investigator
Protocol (07-09-2021)
Consent Form - Facilitator [S0]
Consent Form - Student Focus Group [S0]
Consent Form - Students [S0]
Advertisement - Recruitment Email - The creators of CODE at #31444293.0 - As Submitted
Evaluation Instruments - Facilitator #31465093.0 - As Submitted
Evaluation Instruments - Student #31444296.0 - As Submitted

Please note the following information:

The Board directed that no minors be enrolled in this study.

Under the revised common rule (effective 1-21-2019), continuing review by the Board of the above referenced research is not required; however, the IRB will maintain our records and continue responsibility for exercising administrative and regulatory oversight of this research. The IRB will automatically charge an Ongoing Oversight fee for this administrative effort unless we are notified the research is closing. To avoid unnecessary fees due to closure, a closure form must be submitted for each site 30 days prior to expiration.

The Board found that this research meets the requirements for a waiver of documentation of consent under 21 CFR 56.109(c)

The Board requires that all subjects must be able to consent for themselves to be enrolled in this study. This means that you cannot enroll incapable subjects who require enrollment by consent of a legally authorized representative.

Request for Alternative Consent Process

THE IRB HAS APPROVED THE FOLLOWING LOCATIONS TO BE USED IN THE RESEARCH:

HudsonAlpha Institute for Biotechnology, The Paul Propst Center, 800 Hudson Way, Huntsville, Alabama 35806

ALL IRB APPROVED INVESTIGATORS MUST COMPLY WITH THE FOLLOWING:

As a requirement of IRB approval, the investigators conducting this research will:

- Comply with all requirements and determinations of the IRB.

This is to certify that the information contained herein is true and correct as reflected in the records of WCG IRB. WE CERTIFY THAT WCG IRB IS IN FULL COMPLIANCE WITH GOOD CLINICAL PRACTICES AS DEFINED UNDER THE U.S. FOOD AND DRUG ADMINISTRATION (FDA) REGULATIONS, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (HHS) REGULATIONS, AND THE INTERNATIONAL CONFERENCE ON HARMONISATION (ICH) GUIDELINES.



Board Action: 07/26/2021

- Protect the rights, safety, and welfare of subjects involved in the research.
- Personally conduct or supervise the research.
- Conduct the research in accordance with the relevant current protocol approved by the IRB.
- Ensure that there are adequate resources to carry out the research safely.
- Ensure that research staff are qualified to perform procedures and duties assigned to them during the research.
- Submit proposed modifications to the IRB prior to their implementation.
 - Not make modifications to the research without prior IRB review and approval unless necessary to eliminate apparent immediate hazards to subjects.
- For research subject to continuing review, submit continuing review reports when requested by the IRB.
- Submit a closure form to close research (end the IRB's oversight) when:
 - The protocol is permanently closed to enrollment
 - All subjects have completed all protocol related interventions and interactions
 - For research subject to federal oversight other than FDA:
 - No additional identifiable private information about the subjects is being obtained
 - Analysis of private identifiable information is completed
- For research subject to continuing review, if research approval expires, stop all research activities and immediately contact the IRB.
- Promptly (within 5 days) report to the IRB the information items listed in the IRB's "Prompt Reporting Requirements" available on the IRB's Web site.
- Not accept or provide payments to professionals in exchange for referrals of potential subjects ("finder's fees.")
- Not accept payments designed to accelerate recruitment that are tied to the rate or timing of enrollment ("bonus payments") without prior IRB approval.
- When required by the IRB ensure that consent, permission, and assent are obtained and documented in accordance with the relevant current protocol as approved by the IRB.
- Promptly notify the IRB of any change to information provided on your initial submission form.

Consistent with AAHRPP's requirements in connection with its accreditation of IRBs, the individual and/or organization shall promptly communicate or provide, the following information relevant to the protection of human subjects to the IRB in a timely manner:

- Upon request of the IRB, a copy of the written plan between sponsor or CRO and site that addresses whether expenses for medical care incurred by human subject research subjects who experience research related injury will be reimbursed, and if so, who is responsible in order to determine consistency with the language in the consent document.
- Any site monitoring report that directly and materially affects subject safety or their willingness to continue participation. Such reports will be provided to the IRB within 5 days.
- Reports from any data monitoring committee, data and safety monitoring board, or data and safety monitoring committee in accordance with the time frame specified in the research protocol.
- Any findings from a closed research when those findings materially affect the safety and medical care of past subjects. Findings will be reported for 2 years after the closure of the research.

For Investigator's Brochures, an approval action indicates that the IRB has the document on file for the research.

If the IRB approved an e-consent process that involves uploading the approved consent form to an e-consent platform, please ensure that the consent form(s) approved for your site is the version of the consent form that gets uploaded to the platform. If the board approves a change of Principal Investigator - Once approved, the new Principal Investigator is authorized by WCG IRB to carry out the study as previously approved for the prior Principal Investigator (unless the Board provides alternate instructions to the new Principal Investigator). This includes continued use of the previously approved study materials. The IRB considers the approval of the new PI a continuation of the original approval, so the identifying information about the study remains the same.

If your research site is a HIPAA covered entity, the HIPAA Privacy Rule requires you to obtain written authorization from each research subject for any use or disclosure of protected health information for research. If your IRB-approved consent form does not include such HIPAA authorization language, the HIPAA Privacy Rule requires you to have each research subject sign a separate authorization agreement. "

For research subject to continuing review, you will receive Continuing Review Report forms from WCG IRB when the expiration date is approaching.

Thank you for using this WCG IRB to provide oversight for your research project.

DISTRIBUTION OF COPIES:

Contact, Company

Neil E. Lamb, PhD, HudsonAlpha Institute for Biotechnology
Michael Bryan, JD, HudsonAlpha Institute for Biotechnology

COA Template 12-2020

Florencia P. Behn, PhD, HudsonAlpha Institute for Biotechnology
Michele Morris, HudsonAlpha Institute for Biotechnology

Interview Protocol revised_June 2022_Lamb, Neil - Certificate of Action for Study#_1313664, Panel Meeting#_339388, WO#_1-1552820-1



Certificate of Action

Investigator Name: Neil E. Lamb, PhD	Board Action Date: 06/02/2022
Investigator Address: The Paul Propst Center, 800 Hudson Way Huntsville, AL 35806, United States	Approval Expires: 07/26/2022 Continuing Review Frequency: No CR Required
Sponsor: HudsonAlpha Institute for Biotechnology Institution Tracking Number:	Sponsor Protocol Number: 0106 Amended Sponsor Protocol Number:
Study Number: 1313664	IRB Tracking Number: 20213554
Work Order Number: 1-1552820-1	
Protocol Title: Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format	

THE FOLLOWING ITEMS ARE APPROVED:

Revised Protocol (05-27-2022)
Consent Form - Student Focus Group [S1]
Consent Form - Students [S1]
Advertisement - Recruitment Email - The creators of CODE at #31444293.1 - As Submitted
Evaluation Instruments - Student #31444296.1 - As Submitted

Please note the following information:

Please have all future subjects sign the Consent Form(s) specified in this approval.

THE IRB HAS APPROVED THE FOLLOWING LOCATIONS TO BE USED IN THE RESEARCH:

HudsonAlpha Institute for Biotechnology, The Paul Propst Center, 800 Hudson Way, Huntsville, Alabama 35806

ALL IRB APPROVED INVESTIGATORS MUST COMPLY WITH THE FOLLOWING:

As a requirement of IRB approval, the investigators conducting this research will:

- Comply with all requirements and determinations of the IRB.
- Protect the rights, safety, and welfare of subjects involved in the research.
- Personally conduct or supervise the research.
- Conduct the research in accordance with the relevant current protocol approved by the IRB.
- Ensure that there are adequate resources to carry out the research safely.
- Ensure that research staff are qualified to perform procedures and duties assigned to them during the research.
- Submit proposed modifications to the IRB prior to their implementation.
 - Not make modifications to the research without prior IRB review and approval unless necessary to eliminate apparent immediate hazards to subjects.
- For research subject to continuing review, submit continuing review reports when requested by the IRB.
- Submit a closure form to close research (end the IRB's oversight) when:
 - The protocol is permanently closed to enrollment
 - All subjects have completed all protocol related interventions and interactions
 - For research subject to federal oversight other than FDA:
 - No additional identifiable private information about the subjects is being obtained
 - Analysis of private identifiable information is completed
- For research subject to continuing review, if research approval expires, stop all research activities and immediately contact the IRB.

This is to certify that the information contained herein is true and correct as reflected in the records of WCG IRB. WE CERTIFY THAT WCG IRB IS IN FULL COMPLIANCE WITH GOOD CLINICAL PRACTICES AS DEFINED UNDER THE U.S. FOOD AND DRUG ADMINISTRATION (FDA) REGULATIONS, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (HHS) REGULATIONS, AND THE INTERNATIONAL CONFERENCE ON HARMONISATION (ICH) GUIDELINES.



- Promptly (within 5 days) report to the IRB the information items listed in the IRB's "Prompt Reporting Requirements" available on the IRB's Web site.
- Not accept or provide payments to professionals in exchange for referrals of potential subjects ("finder's fees.")
- Not accept payments designed to accelerate recruitment that are tied to the rate or timing of enrollment ("bonus payments") without prior IRB approval.
- When required by the IRB ensure that consent, permission, and assent are obtained and documented in accordance with the relevant current protocol as approved by the IRB.
- Promptly notify the IRB of any change to information provided on your initial submission form.

Consistent with AAHRPP's requirements in connection with its accreditation of IRBs, the individual and/or organization shall promptly communicate or provide, the following information relevant to the protection of human subjects to the IRB in a timely manner:

- Upon request of the IRB, a copy of the written plan between sponsor or CRO and site that addresses whether expenses for medical care incurred by human subject research subjects who experience research related injury will be reimbursed, and if so, who is responsible in order to determine consistency with the language in the consent document.
- Any site monitoring report that directly and materially affects subject safety or their willingness to continue participation. Such reports will be provided to the IRB within 5 days.
- Any findings from a closed research when those findings materially affect the safety and medical care of past subjects. Findings will be reported for 2 years after the closure of the research.

For Investigator's Brochures, an approval action indicates that the IRB has the document on file for the research.

If the IRB approved an e-consent process that involves uploading the approved consent form to an e-consent platform, please ensure that the consent form(s) approved for your site is the version of the consent form that gets uploaded to the platform.

If the board approves a change of Principal Investigator - Once approved, the new Principal Investigator is authorized by WCG IRB to carry out the study as previously approved for the prior Principal Investigator (unless the Board provides alternate instructions to the new Principal Investigator). This includes continued use of the previously approved study materials. The IRB considers the approval of the new PI a continuation of the original approval, so the identifying information about the study remains the same.

If your research site is a HIPAA covered entity, the HIPAA Privacy Rule requires you to obtain written authorization from each research subject for any use or disclosure of protected health information for research. If your IRB-approved consent form does not include such HIPAA authorization language, the HIPAA Privacy Rule requires you to have each research subject sign a separate authorization agreement.

If this study includes data monitoring committee/data safety monitoring board, please note that the reports of all meetings of this committee should be submitted to the IRB even if the outcome of the meeting results in no changes to the study.

For research subject to continuing review, you will receive Continuing Review Report forms from WCG IRB when the expiration date is approaching.

Thank you for using this WCG IRB to provide oversight for your research project.

DISTRIBUTION OF COPIES:

Contact, Company

Neil E. Lamb, PhD, HudsonAlpha Institute for Biotechnology
 Michael Bryan, JD, HudsonAlpha Institute for Biotechnology
 Florencia P. Behn, PhD, HudsonAlpha Institute for Biotechnology
 Michele Morris, HudsonAlpha Institute for Biotechnology

Morris, Michele - Certificate of Action for Study#_ 1313664, Panel Meeting#_ 351879, WO#_ 1-1605323-1



Certificate of Action

Investigator Name: Michele Morris, BA, BS, AS	Board Action Date: 12/01/2022
Investigator Address: 800 Hudson Way, The Paul Propst Center Huntsville, AL 35806, United States	Approval Expires: 07/26/2023 Continuing Review Frequency: No CR Required
Sponsor: HudsonAlpha Institute for Biotechnology Institution Tracking Number:	Sponsor Protocol Number: 0106 Amended Sponsor Protocol Number:
Study Number: 1313664	IRB Tracking Number: 20213554
Work Order Number: 1-1605323-1	
Protocol Title: Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format	

THE FOLLOWING ITEMS ARE APPROVED:

Investigator
Consent Form - Facilitator [S0-1]
Consent Form - Student Focus Group [S1-1]
Consent Form - Students [S1-1]
Advertisement - Recruitment Email - The creators of CODE at #31444293.2 - As Modified

Please note the following information:

Under the revised common rule (effective 1-21-2019), continuing review by the Board of the above referenced research is not required; however, the IRB will maintain our records and continue responsibility for exercising administrative and regulatory oversight of this research. The IRB will automatically charge an Ongoing Oversight fee for this administrative effort unless we are notified the research is closing. To avoid unnecessary fees due to closure, a closure form must be submitted for each site 30 days prior to expiration.

Please have all future subjects sign the Consent Form(s) specified in this approval.

THE IRB HAS APPROVED THE FOLLOWING LOCATIONS TO BE USED IN THE RESEARCH:

HudsonAlpha Institute for Biotechnology, 800 Hudson Way, The Paul Propst Center, Huntsville, Alabama 35806

ALL IRB APPROVED INVESTIGATORS MUST COMPLY WITH THE FOLLOWING:

As a requirement of IRB approval, the investigators conducting this research will:

- Comply with all requirements and determinations of the IRB.
- Protect the rights, safety, and welfare of subjects involved in the research.
- Personally conduct or supervise the research.
- Conduct the research in accordance with the relevant current protocol approved by the IRB.
- Ensure that there are adequate resources to carry out the research safely.
- Ensure that research staff are qualified to perform procedures and duties assigned to them during the research.
- Submit proposed modifications to the IRB prior to their implementation.
 - Not make modifications to the research without prior IRB review and approval unless necessary to eliminate apparent immediate hazards to subjects.
- For research subject to continuing review, submit continuing review reports when requested by the IRB.
- Submit a closure form to close research (end the IRB's oversight) when:
 - The protocol is permanently closed to enrollment
 - All subjects have completed all protocol related interventions and interactions

This is to certify that the information contained herein is true and correct as reflected in the records of WCG IRB. WE CERTIFY THAT WCG IRB IS IN FULL COMPLIANCE WITH GOOD CLINICAL PRACTICES AS DEFINED UNDER THE U.S. FOOD AND DRUG ADMINISTRATION (FDA) REGULATIONS, U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES (HHS) REGULATIONS, AND THE INTERNATIONAL CONFERENCE ON HARMONISATION (ICH) GUIDELINES.



- For research subject to federal oversight other than FDA:
 - No additional identifiable private information about the subjects is being obtained
 - Analysis of private identifiable information is completed
- For research subject to continuing review, if research approval expires, stop all research activities and immediately contact the IRB.
- Promptly (within 5 days) report to the IRB the information items listed in the IRB's "Prompt Reporting Requirements" available on the IRB's Web site.
- Not accept or provide payments to professionals in exchange for referrals of potential subjects ("finder's fees.")
- Not accept payments designed to accelerate recruitment that are tied to the rate or timing of enrollment ("bonus payments") without prior IRB approval.
- When required by the IRB ensure that consent, permission, and assent are obtained and documented in accordance with the relevant current protocol as approved by the IRB.
- Promptly notify the IRB of any change to information provided on your initial submission form.

Consistent with AAHRPP's requirements in connection with its accreditation of IRBs, the individual and/or organization shall promptly communicate or provide, the following information relevant to the protection of human subjects to the IRB in a timely manner:

- Upon request of the IRB, a copy of the written plan between sponsor or CRO and site that addresses whether expenses for medical care incurred by human subject research subjects who experience research related injury will be reimbursed, and if so, who is responsible in order to determine consistency with the language in the consent document.
- Any site monitoring report that directly and materially affects subject safety or their willingness to continue participation. Such reports will be provided to the IRB within 5 days.
- Any findings from a closed research when those findings materially affect the safety and medical care of past subjects. Findings will be reported for 2 years after the closure of the research.

For Investigator's Brochures, an approval action indicates that the IRB has the document on file for the research.

If the IRB approved an e-consent process that involves uploading the approved consent form to an e-consent platform, please ensure that the consent form(s) approved for your site is the version of the consent form that gets uploaded to the platform.

If the board approves a change of Principal Investigator - Once approved, the new Principal Investigator is authorized by WCG IRB to carry out the study as previously approved for the prior Principal Investigator (unless the Board provides alternate instructions to the new Principal Investigator). This includes continued use of the previously approved study materials. The IRB considers the approval of the new PI a continuation of the original approval, so the identifying information about the study remains the same.

If your research site is a HIPAA covered entity, the HIPAA Privacy Rule requires you to obtain written authorization from each research subject for any use or disclosure of protected health information for research. If your IRB-approved consent form does not include such HIPAA authorization language, the HIPAA Privacy Rule requires you to have each research subject sign a separate authorization agreement.

If this study includes data monitoring committee/data safety monitoring board, please note that the reports of all meetings of this committee should be submitted to the IRB even if the outcome of the meeting results in no changes to the study.

For research subject to continuing review, you will receive Continuing Review Report forms from WCG IRB when the expiration date is approaching.

Thank you for using this WCG IRB to provide oversight for your research project.

DISTRIBUTION OF COPIES:

Contact, Company

Michael Bryan, JD, HudsonAlpha Institute for Biotechnology
 Florencia P. Behn, PhD, HudsonAlpha Institute for Biotechnology
 Michele Morris, BA, BS, AS, HudsonAlpha Institute for Biotechnology

Appendix C. Study Recruitment Emails

#31444293.2

IRB Approved As
Modified at the
Study Level
Dec 01, 2022

[SUBJECT RECRUITMENT EMAIL FOR STUDENTS AND FACILITATORS]

Hello!

The creators of CODE at the HudsonAlpha Institute for Biotechnology invite you to participate in a study *exploring your experiences with the Characterizing Our DNA Exceptions (CODE) program*.

You received this email because you participated in CODE facilitator training or as a CODE student and gave us permission to contact you for future research studies.

Your participation in the survey will help us evaluate CODE activities and improve the program for future facilitators and students.

The possible risks or discomforts of participating in the study are minimal. You may feel uncomfortable answering the questions, but not any more so than in everyday life experiences.

Participation in this study is voluntary. You can choose to answer all, some, or none of the questions. If you participate, you will allow researchers at HudsonAlpha to use your de-identified survey responses for research.

Please complete the study consent form and survey at [\[insert link\]](#)

Thank you in advance for your time and participation; we greatly appreciate it. If you have questions or concerns, please do not hesitate to contact me.

Sincerely,

Michele Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
256.327.0576
mmorris@hudsonalpha.org

[POST-CODE SURVEY EMAIL FOR STUDENTS - to be sent immediately after the end of the students' CODE projects]

Hello!

The creators of CODE at the HudsonAlpha Institute for Biotechnology invite you to continue your participation in a study *exploring your experiences with the Characterizing Our DNA Exceptions (CODE) program*.

You received this email because you were a CODE program student and agreed to participate in a research study called "Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format."

As part of this study, you provided consent and completed a survey before the start of your CODE experience.

Please review the study consent form and complete the post-CODE survey at [\[insert link\]](#). This is the last survey that we will ask you to complete as part of this study.

Thank you in advance for your time and participation; we greatly appreciate it. If you have questions or concerns, please do not hesitate to contact us.

Thank you in advance for your time and participation; we greatly appreciate it. If you have questions or concerns, please do not hesitate to contact me.

Sincerely,

Michele Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
256.327.0576
mmorris@hudsonalpha.org

[INTERVIEW INVITATION EMAIL FOR STUDENTS]

Hello!

The creators of CODE at the HudsonAlpha Institute for Biotechnology invite you to participate in a study *exploring your experiences with the Characterizing Our DNA Exceptions (CODE) program*.

You received this email because you took part in a research study where you answered questions about participating in a CODE program and gave us permission to contact you to participate in an interview.

We invite you to participate in a 20-30 minute video conference interview about your experience with CODE. There is no known risk to participating and it is completely voluntary.

If you would like to participate, please enter your contact information in this form [\[insert link\]](#) so we may contact you and schedule the interview.

Thank you in advance for your time and participation; we greatly appreciate it. If you have questions or concerns, please do not hesitate to contact me.

Sincerely,

Michele Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
256.327.0576
mmorris@hudsonalpha.org

**[POST-WORKSHOP SURVEY EMAIL FOR FACILITATORS - to be emailed
immediately after the CODE Facilitator Training Workshop]**

Hello!

The creators of CODE at the HudsonAlpha Institute for Biotechnology invite you to continue your participation in a study *exploring your experiences with the Characterizing Our DNA Exceptions (CODE) program*.

You received this email because you received CODE facilitator training and agreed to participate in a research study called “Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format.”

As part of this study, you provided consent and completed a survey before the start of your code facilitator training.

Please review the study consent form and complete the post-CODE training workshop survey at [\[insert link\]](#)

Thank you in advance for your time and participation; we greatly appreciate it. If you have questions or concerns, please do not hesitate to contact us.

Sincerely,

Michele Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
256.327.0576
mmorris@hudsonalpha.org

**[EXTENDED POST-WORKSHOP SURVEY EMAIL FOR FACILITATORS - to be
emailed 6-9 months after the CODE Facilitator Training Workshop]**

Hello!

The creators of CODE at the HudsonAlpha Institute for Biotechnology invite you to continue your participation in a study *exploring your experiences with the Characterizing Our DNA Exceptions (CODE) program*.

You received this email because you received CODE facilitator training and agreed to participate in a research study called “Characterizing Our DNA Exceptions (CODE): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format.”

As part of this study, you provided consent and completed a survey before the start of your code facilitator training and another immediately after the workshop.

Please review the study consent form and complete the extended post-CODE training workshop survey at [\[insert link\]](#). This is the final survey that we will ask you to complete for this study.

Thank you in advance for your time and participation; we greatly appreciate it. If you have questions or concerns, please do not hesitate to contact us.

Sincerely,

Michele Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
256.327.0576
mmorris@hudsonalpha.org

Appendix D. Student Survey Consent

IRB APPROVED
AS MODIFIED
Dec 01, 2022

RESEARCH SUBJECT INFORMATION AND CONSENT FORM

TITLE: Characterizing Our DNA Exceptions (**CODE**): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format

PROTOCOL NO.: 0106
WCG IRB Protocol #20213554

SPONSOR: HudsonAlpha Institute for Biotechnology

PRINCIPAL INVESTIGATOR: Michele C. Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
The Paul Propst Center
800 Hudson Way
Huntsville, Alabama 35806
mmorris@hudsonalpha.org

STUDY-RELATED PHONE NUMBER(S): Michele C. Morris
(256) 327-0576

We are asking you to be in a research study. The purpose of this consent form is to give you the information you need to decide whether to participate in the study. Please read the form carefully.

What should I know about this research?

- Someone will explain this research to you.
- Taking part in this research is voluntary. Whether you take part is up to you.
- If you don't take part, it won't be held against you. You can also take part now, and stop participating at any time. No matter what your decision is, there will be no penalty or loss of benefits to which you are otherwise entitled.
- If you don't understand, ask questions.
- Ask all the questions you want before you decide.

You may ask questions about the purpose of the research, the tasks of a participant, the possible risks and benefits, your rights as a volunteer, and anything else about the study

or this form that is not clear. When we have answered all your questions, you can decide whether you want to participate. You may save a digital copy of this form for your records.

You must be at least 19 years old to participate in this study. Do not continue if you are younger than 19 years of age.

PURPOSE OF THE STUDY

CODE aims to increase access to research-based learning experiences and introduce a broad range of students to bioinformatics in a mentored, project-based format that captures interest, builds confidence, and encourages continued participation in science education and a STEM career path.

This research study will evaluate whether participating in a CODE project increases awareness and interest in bioinformatics, science self-efficacy, scientific identity, and intention to enter or remain in a STEM major - all factors that positively impact STEM persistence. The data we collect will help improve CODE and make the model available to more colleges and universities.

PARTICIPATION

Participation in this study is voluntary. You may refuse to take part in the study or exit the survey at any time without penalty. You can choose to answer all, some, or none of the questions for any reason. You can withdraw your consent by emailing the Study Coordinator named above.

We will ask you to complete a survey ***before beginning*** and ***after completing*** your CODE project. The survey has 7 demographic questions (age, gender, etc.) and 52 rating scale questions about bioinformatics awareness and interest, research experience, science self-efficacy, scientific identity, and interest and persistence in STEM.

We may also ask you to participate in a 20-30 minute video conference interview to understand your survey responses better. You will complete a separate consent process if you decide to participate in an interview.

BENEFITS

You may not benefit directly from taking part in this study. However, the study may help the education community better understand the efficiency of programs designed to engage students in STEM careers. We will also use the information from this study to improve the CODE project and provide a sustainable model for low-cost, authentic undergraduate bioinformatics research projects for institutions of all sizes.

PAYMENT FOR PARTICIPATION

Participation in this research study is voluntary, and you will not receive payment or compensation for your participation. There are no costs to you for your participation.

STUDY DURATION

Your participation in this study will be over when you have completed all the surveys described in this consent form.

RISKS AND DISCOMFORTS

There are no known risks to participating in this research study. It may be inconvenient, or you may feel uncomfortable when answering the survey questions. This discomfort should not be greater than that experienced in everyday life.

There is a small risk for a breach of confidentiality of your personal information in the study. To minimize this risk, files will only be available to study personnel and will be password-secured. Whenever possible, we will remove identifiers from study-related information.

What other choices do I have besides taking part in this research?

This research is not a treatment study. Your alternative is to not participate in this research.

CONFIDENTIALITY

Information obtained about you for this study will be kept confidential to the extent allowed by law. Research information that identifies you may be shared with WGC IRB and others responsible for ensuring compliance with laws and regulations related to research. We may publish our findings for scientific purposes or reports to the funding agency; however, we will not share your name, email, or other information that could easily identify you.

Your survey answers will be linked to your email address and sent to the survey website, where we will store your data in a password-protected electronic format. **Please do not share your survey link with anyone.**

Only the study coordinator will have access to your name and email. She will remove these and any other identifiers and share your survey responses with researchers at HudsonAlpha who are conducting this study. We will retain study data for at least three years at HudsonAlpha Institute.

What happens to the information collected for this research?

Your private information and your research record will be shared with individuals and organizations that conduct or watch over this research, including:

- The research sponsor
- People who work with the research sponsor
- Government agencies, such as the Food and Drug Administration
- WCG IRB, the Institutional Review Board (IRB) that reviewed this research

SOURCE OF FUNDING FOR THE STUDY

This study is paid for with a grant from the National Science Foundation (NSF) to the HudsonAlpha Institute for Biotechnology.

QUESTIONS

Contact Ms. Michele Morris, Principal Investigator (256-327-0576, mmorris@hudsonalpha.org) for any of the following reasons:

- If you have any questions about your participation in this study
- If you feel you have had a research-related injury
- If you have questions, concerns, or complaints about the research

If you have questions about your rights as a research subject or if you have questions, concerns, or complaints about the research, you may contact:

WCG IRB
 1019 39th Avenue SE Suite 120
 Puyallup, Washington 98374-2115
 Telephone: 855-818-2289
 E-mail: researchquestions@wcgirb.com

WCG IRB is a group of people who independently review research.

WCG IRB will not be able to answer some study-specific questions, such as questions about appointment times. However, you may contact WCG IRB if you cannot reach the research staff or if you wish to talk to someone other than the research staff.

ELECTRONIC CONSENT

Please select your choice below and print a copy of this consent form for your records.

Clicking on the “Agree” button indicates that:

- You have read and understood the above information
- You voluntarily agree to participate
- You are 19 years of age or older

☐ Agree

☐ Disagree

RESEARCH SUBJECT INFORMATION AND CONSENT SCRIPT

TITLE: Characterizing Our DNA Exceptions (**CODE**): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format

PROTOCOL NO.: 0106
WCG IRB Protocol #20213554

SPONSOR: HudsonAlpha Institute for Biotechnology

PRINCIPAL INVESTIGATOR: Michele C. Morris
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
The Paul Propst Center
800 Hudson Way
Huntsville, Alabama 35806
mmorris@hudsonalpha.org

STUDY-RELATED PHONE NUMBER(S): Michele C. Morris
(256) 327-0576

The study coordinator will use this script to obtain consent from CODE students (at least 19 years old) who wish to participate in an interview.

What should I know about this research?

- Someone will explain this research to you.
- Taking part in this research is voluntary. Whether you take part is up to you.
- If you don't take part, it won't be held against you. You can also take part now, and stop participating at any time. No matter what your decision is, there will be no penalty or loss of benefits to which you are otherwise entitled.
- If you don't understand, ask questions.
- Ask all the questions you want before you decide.

Purpose of the Study

We are asking you to participate in a virtual interview to discuss your CODE project experience. The online discussion will last 20-30 minutes and will include questions about your interest in bioinformatics, your CODE project experience, and your thoughts about a career in science. If you decide to take part in this research study, the general procedures include participation in this interview. During the interview, we will ask you several open-ended questions derived directly from the surveys you already completed. Discussion from the interview will be recorded for transcription and analysis.

Study Risk

Participating in this study has minimal risk. The harm or discomfort you may experience is not greater than what you would encounter daily. You might feel embarrassed to discuss issues around identity and science learning. However, you do not have to answer any questions or share anything you do not want to share. You can stop participating in the study at any time without penalty.

The study investigators take precautions to keep your information confidential and prevent anyone from discovering or guessing your identity. For example, we use an identification code instead of your name and keep all data on a password-protected computer or electronic files. There is still a potential risk of loss of confidentiality.

Study Benefits

There is no direct benefit to you for participating in this study. Participation may benefit undergraduate education and bioinformatics research to learn about the best ways to implement undergraduate research projects in the future.

Study Compensation

We will not pay you to participate in the study. Other than your time, there are no costs to you for taking part in this study.

Study Duration

The study is over when you have completed the interview. Your participation is voluntary, and you can end the interview at any time, even if we are not finished.

What other choices do I have besides taking part in this research?

This research is not a treatment study. Your alternative is to not participate in this research.

Protection of your Confidentiality

All interviews will be held online. We will audio record the session to help us remember exactly what you say. The audio recording will be transcribed (written down), and the audio recording will then be destroyed.

We will code your identity, and the interview results will be non-identifiable to all but the study coordinator. The study coordinator will store the coded transcripts in a password-protected digital file accessible only to the researchers in this study who need access to the files.

What happens to the information collected for this research?

Your private information and your research record will be shared with individuals and organizations that conduct or watch over this research, including:

- The research sponsor
- People who work with the research sponsor
- Government agencies, such as the Food and Drug Administration
- WCG IRB, the Institutional Review Board (IRB) that reviewed this research

Use of Research Results

Evaluation data will measure shifts in student awareness and interest in bioinformatics, self-efficacy, and scientific identity - all factors that positively impact STEM persistence. This project's success will provide a sustainable model for low-cost, authentic undergraduate bioinformatics research projects for institutions of all sizes.

We may publish the results of this study in journals and present our findings at academic conferences. Your identity will be removed from all data before publication or use for educational purposes.

Consent for Audio Recording

Audio recording is part of this research study. The audio recording will be deleted after it is transcribed. We will not share our notes or the audio recordings outside of the CODE project team. If you do not wish to be audio-recorded, you cannot participate in this study.

Contact Information: If you have any questions, concerns, or complaints about the research, or you think you have been injured as a result of your participation in this research study, you can contact me at mmorris@hudsonalpha.org or by phone at 256-327-0576.

This research is being overseen by WCG IRB. An IRB is a group of people who perform independent review of research studies. You may talk to them at 855-818-2289 or researchquestions@wcgirb.com if:

- You have questions, concerns, or complaints that are not being answered by the research team.
- You are not getting answers from the research team.
- You cannot reach the research team.
- You want to talk to someone else about the research.
- You have questions about your rights as a research subject.

Participant's Rights

1. Someone has discussed the study and read the informed consent to me. I have had ample opportunity to ask questions about the purposes, procedures, risks, and benefits of this research study.
2. I understand that my participation is voluntary. The investigator may withdraw me from the research if they feel that I cannot complete all parts of the study or become severely distressed from the study.
3. If significant new information becomes available during the study, which may relate to my willingness to continue, the investigator will provide this information to me.
4. Any information derived from the research study that identifies me will not be voluntarily released or disclosed without my separate consent, except as expressly required by law.
5. I should receive a copy of the Informed Consent document.

Do you agree to participate in this study, are at least 19 years old, and understand that the session will be audio recorded?

RESEARCH SUBJECT INFORMATION AND CONSENT FORM

TITLE: Characterizing Our DNA Exceptions (**CODE**): Increasing Undergraduate Engagement in Bioinformatics through a Mentored, Project-Based Format

PROTOCOL NO.: 0106
WCG IRB Protocol #20213554

SPONSOR: HudsonAlpha Institute for Biotechnology

PRINCIPAL INVESTIGATOR: Michele C. Morris,
Director for Workforce Development
HudsonAlpha Institute for Biotechnology
The Paul Propst Center
800 Hudson Way
Huntsville, Alabama 35806
mmorris@hudsonalpha.org

STUDY-RELATED PHONE NUMBER(S): Michele C. Morris
(256) 327-0576

We are asking you to be in a research study. The purpose of this consent form is to give you the information you need to decide whether to participate in the study. Please read the form carefully.

What should I know about this research?

- Someone will explain this research to you.
- Taking part in this research is voluntary. Whether you take part is up to you.
- If you don't take part, it won't be held against you. You can also take part now, and stop participating at any time. No matter what your decision is, there will be no penalty or loss of benefits to which you are otherwise entitled.
- If you don't understand, ask questions.
- Ask all the questions you want before you decide.

You may ask questions about the purpose of the research, the tasks of a participant, the possible risks and benefits, your rights as a volunteer, and anything else about the study or this form that is not clear. When we have answered all your questions, you can decide whether you want to participate. You may save a digital copy of this form for your records.

You must be at least 19 years old to participate in this study. Do not continue if you are younger than 19 years of age.

PURPOSE OF THE STUDY

CODE aims to increase access to research-based learning experiences and introduce a broad range of students to bioinformatics in a mentored, project-based format that captures interest, builds confidence, and encourages continued participation in science education and a STEM career path. This research study will evaluate whether participating in a CODE project increases awareness and interest in bioinformatics, science self-efficacy, scientific identity, and intention to enter or remain in a STEM major - all factors that positively impact STEM persistence. The data we collect will help improve CODE and make the model available to more colleges and universities.

The facilitator surveys we are asking you to participate in will measure shifts in comfort with bioinformatics content and the value of the CODE program in providing an undergraduate research experience for students.

PARTICIPATION

Participation in this study is voluntary. You may refuse to take part in the research or exit the survey at any time without penalty. You can choose to answer all, some, or none of the questions for any reason. You can withdraw your consent by emailing the Study Coordinator named above.

We will ask you to complete a survey ***before beginning, after completing, and 6-9 months after*** your CODE Facilitator Training Workshop. The first survey has 10 demographics questions (age, gender, etc.), 3 multiple-choice questions, and 6 rating scale questions about your comfort level with bioinformatics and the training workshop. The post-workshop survey has 20 rating scale questions and four open response questions on workshop content and improvement. The extended follow-up survey 6-9 months after training has 21 rating scale questions.

BENEFITS

You may not benefit directly from taking part in this study. However, this study may help the education community better understand the efficiency of programs designed to engage students in STEM careers. We will also use the information from this study to improve the CODE project and provide a sustainable model for low-cost, authentic undergraduate bioinformatics research projects for institutions of all sizes.

PAYMENT FOR PARTICIPATION

Participation in this research study is voluntary, and you will not receive payment or compensation for your participation. There are no costs to you for your participation.

STUDY DURATION

Your participation in this study will be over when you have completed all the surveys described in this consent form.

RISKS AND DISCOMFORTS

There are no known risks to participating in this research study. It may be inconvenient, or you may feel uncomfortable when answering the survey questions. This discomfort should not be greater than that experienced in everyday life.

There is a small risk for a breach of confidentiality of your personal information in the study. To minimize this risk, files will only be available to study personnel and will be password-secured. Whenever possible, we will remove identifiers from study-related information.

What other choices do I have besides taking part in this research?

This research is not a treatment study. Your alternative is to not participate in this research.

CONFIDENTIALITY

Information obtained about you for this study will be kept confidential to the extent allowed by law. Research information that identifies you may be shared with WGC IRB and others responsible for ensuring compliance with laws and regulations related to research. We may publish our findings for scientific purposes or reports to the funding agency; however, we will not share your name, email, or other information that could easily identify you.

Your survey answers will be linked to your email address and sent to the survey website, where we will store your data in a password-protected electronic format. **Please do not share your survey link with anyone.**

Only the study coordinator will have access to your name and email. She will remove these and any other identifiers and share your survey responses with researchers at HudsonAlpha who are conducting this study. We will retain study data for at least three years at HudsonAlpha Institute.

What happens to the information collected for this research?

Your private information and your research record will be shared with individuals and organizations that conduct or watch over this research, including:

- The research sponsor
- People who work with the research sponsor
- Government agencies, such as the Food and Drug Administration
- WCG IRB, the Institutional Review Board (IRB) that reviewed this research

SOURCE OF FUNDING FOR THE STUDY

This study is paid for with a grant from the National Science Foundation (NSF) to the HudsonAlpha Institute for Biotechnology.

QUESTIONS

Contact Michele Morris, Principal Investigator (256-327-0576), for any of the following reasons:

- If you have any questions about your participation in this study
- If you feel you have had a research-related injury
- If you have questions, concerns, or complaints about the research

If you have questions about your rights as a research subject or if you have questions, concerns, or complaints about the research, you may contact:

WCG IRB
1019 39th Avenue SE Suite 120

Puyallup, Washington 98374-2115
Telephone: 855-818-2289
E-mail: researchquestions@wcgirb.com

WCG IRB is a group of people who independently review research.

WCG IRB will not be able to answer some study-specific questions, such as questions about appointment times. However, you may contact WCG IRB if you cannot reach the research staff or if you wish to talk to someone other than the research staff.

ELECTRONIC CONSENT

Please select your choice below and print a copy of this consent form for your records.

Clicking on the “Agree” button indicates that:

- You have read and understood the above information
- You voluntarily agree to participate
- You are 19 years of age or older

☐ Agree

☐ Disagree

Appendix E. Summative Instruments for Student Participants

#31444296.0

IRB Approved at the Study Level

CODE Project Evaluation Summative Instruments for Student Participants

Pre-CODE survey for student participants in CODE-based research projects

Thank you for participating in our research study. You may answer any or all of the questions below.

The demographic questions are for study purposes only. No individual information will be shared with or reported to anyone other than CODE research study staff.

1. Gender
 1. Male
 2. Female
 3. Prefer not to answer
 4. Other
2. Current level in school
 1. Freshman
 2. Junior
 3. Sophomore
 4. Senior
 5. Graduate student
3. Ethnicity
 1. Hispanic or Latino
 2. Not Hispanic or Latino
 3. Rather not say
4. Which race best describes you? (Please choose only one.)
 1. American Indian or Alaskan Native
 2. Asian/SE Asian
 3. Black or African American
 4. Native Hawaiian or Pacific Islander
 5. White/Caucasian
 6. Two or more races
 7. Prefer not to say
 8. Unknown
5. Are you a person living with a disability?
 1. Yes
 2. No

3. Prefer not to answer
6. What is the highest level of education you have completed?
 1. Graduated from high school
 2. 1 year of college
 3. 2 years of college
 4. 3 years of college
 5. Graduated from college
 6. Some graduate school
 7. Completed graduate school
7. Are you a first-generation college student?
 1. Yes
 2. No
 3. Not sure
 4. Prefer not to answer

Bioinformatics Awareness, Interest, and Learning

For each statement, indicate how well the statement applies to you using the scale below:

- 1 – strongly disagree
- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

1. I have a good understanding of what bioinformatics is.
2. I have a good understanding of how bioinformatics is done.
3. I could use bioinformatics to find out information about a gene.
4. I could use bioinformatics to find the domains in a protein sequence.
5. I could use bioinformatics to determine the similarity between a group of genes or a group of proteins.
6. I could use bioinformatics to view the three-dimensional structure of a protein.
7. If given the nucleotide sequence of a gene, I could use bioinformatics to determine the amino acid sequence of the resulting protein.
8. I could use bioinformatics to find a genetic mutation responsible for a disease or specific trait.
9. I am enthusiastic about bioinformatics.
10. I am interested in discussing bioinformatics with my friends or family.
11. I am interested in taking future classes relating to bioinformatics and computational biology.
12. I am comfortable learning how to use computational programs on my own.
13. I am familiar with the concepts of computational modeling methods.
14. I am familiar with using computational modeling programs.
15. I am comfortable working on a Unix/Linux based computer.
16. I have a good conceptual understanding of homology modeling.
17. I feel comfortable doing homology modeling.
18. I have a good conceptual understanding of molecular dynamics simulations.

19. I feel comfortable performing and analyzing molecular dynamics simulations.

Research Experience

Please describe how active you have been in the following science related activities in the past using the scale below.

- 1 – not at all
- 2 – a little
- 3 – quite a bit
- 4 – a lot

1. I have worked on a research project in which I figured out what data to collect and how to collect it.
2. I have reported my research results in an oral presentation or written report.
3. I have learned scientific language and terminology.
4. I have related my research results and explanations to the work of others.
5. I have used scientific literature to guide a research project.
6. I have had the opportunity to generate my own research question to answer.
7. I have learned technical science skills.
8. I have taken a leadership role in a scientific research team.

Science Self-Efficacy

Indicate the extent to which you are confident that you can complete the following tasks using the scale below.

- 1 – not at all confident
- 2 – a little confident
- 3 – somewhat confident
- 4 – a lot confident

1. Generate a research question to answer
2. Use scientific literature to guide research
3. Create explanations for the results of the study
4. Develop theories (integrate results from multiple studies)
5. Use scientific language and terminology
6. Use technical science skills

Identity as a Scientist

The following questions ask how you think about yourself and your personal identity. We want to understand how much you think that being a scientist is part of who you are. Rate much you agree with the following statements using the scale below.

- 1 – not at all
- 2 – a little
- 3 – somewhat

4 – a lot

1. In general, being a scientist is an important part of my self-image.
2. Being a scientist is an important reflection of who I am.
3. I feel like I belong in the field of science.
4. I have a strong sense of belonging to the community of scientists.
5. I am a scientist.

Interest and Persistence in STEM

Rate your level of agreement with each statement using the scale below.

- 1 – strongly disagree
- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

1. I intend to work in a job related to science.
2. I see the next steps in the field of science, and I intend to take them.
3. I will work as hard as necessary to achieve a career in science.
4. I expect a career in this field will be very satisfying.
5. I feel that I am on a definite career path in science.
6. I definitely want a career for myself in science.
7. Science is the ideal field of study for my life.
8. I am likely to switch to a major that is NOT in a STEM field.
9. I am likely to remain in my STEM major through to graduation or completion of my program of study.
10. I intend to leave my science major or science related track.
11. I do not intend to leave my STEM major before I graduate or complete my program of study.
12. I am not going to remain in a major or track in the STEM fields.
13. I am going to remain in a major or track in the STEM fields,
14. I have considered switching majors to one that is NOT in a STEM field.

Thank you for completing this survey!

If you have any questions about this study or further comments, you may contact Michele Morris at 256.327.0576 or mmorris@hudsonalpha.org

Post-CODE survey for student participants in CODE-based research projects

Thank you for participating in our research study. You may review the consent document that you agreed to for this study at this link [insert link to PDF of student consent form].

The questions below are for research purposes only. No individual information will be shared or reported with anyone other than CODE project staff.

Bioinformatics Awareness, Interest, and Learning

For each statement, indicate how well the statement applies to you using the scale below:

- 1 – strongly disagree
- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

1. I have a good understanding of what bioinformatics is.
2. I have a good understanding of how bioinformatics is done.
3. I could use bioinformatics to find out information about a gene.
4. I could use bioinformatics to find the domains in a protein sequence.
5. I could use bioinformatics to determine the similarity between a group of genes or a group of proteins.
6. I could use bioinformatics to view the three-dimensional structure of a protein.
7. If given the nucleotide sequence of a gene, I could use bioinformatics to determine the amino acid sequence of the resulting protein.
8. I could use bioinformatics to find a genetic mutation responsible for a disease or specific trait.
9. I am enthusiastic about bioinformatics.
10. I am interested in discussing bioinformatics with my friends or family.
11. I am interested in taking future classes relating to bioinformatics and computational biology.
12. I am comfortable learning how to use computational programs on my own.
13. I am familiar with the concepts of computational modeling methods.
14. I am familiar with using computational modeling programs.
15. I am comfortable working on a Unix/Linux based computer.
16. I have a good conceptual understanding of homology modeling.
17. I feel comfortable doing homology modeling.
18. I have a good conceptual understanding of molecular dynamics simulations.
19. I feel comfortable performing and analyzing molecular dynamics simulations.

Research Experience

Please describe how active you have been in the following science related activities **while working on a CODE project** using the scale below.

- 1 – not at all

- 2 – a little
- 3 – quite a bit
- 4 – a lot

1. I worked on a research project in which I figured out what data to collect and how to collect it.
2. I reported my research results in an oral presentation or written report.
3. I learned scientific language and terminology.
4. I related my research results and explanations to the work of others.
5. I used scientific literature to guide a research project.
6. I had the opportunity to generate my own research question to answer.
7. I learned technical science skills.
8. I took a leadership role in a scientific research team.

Science Self-Efficacy

Indicate the extent to which you are confident that you can complete the following tasks using the scale below.

- 1 – not at all confident
- 2 – a little confident
- 3 – somewhat confident
- 4 – a lot confident

7. Generate a research question to answer
8. Use scientific literature to guide research
9. Create explanations for the results of the study
10. Develop theories (integrate results from multiple studies)
11. Use scientific language and terminology
12. Use technical science skills

Identity as a Scientist

The following questions ask how you think about yourself and your personal identity. We want to understand how much you think that being a scientist is part of who you are. Rate much you agree with the following statements using the scale below.

- 1 – not at all
- 2 – a little
- 3 – somewhat
- 4 – a lot

6. In general, being a scientist is an important part of my self-image.
7. Being a scientist is an important reflection of who I am.
8. I feel like I belong in the field of science.
9. I have a strong sense of belonging to the community of scientists.
10. I am a scientist.

Persistence and Interest in STEM

Rate your level of agreement with each statement using the scale below.

- 1 – strongly disagree
- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

- 1. I intend to work in a job related to science.
- 2. I see the next steps in the field of science, and I intend to take them.
- 3. I will work as hard as necessary to achieve a career in science.
- 4. I expect a career in this field will be very satisfying.
- 5. I feel that I am on a definite career path in science.
- 6. I definitely want a career for myself in science.
- 7. Science is the ideal field of study for my life.
- 8. I am likely to switch to a major that is NOT in a STEM field.
- 9. I am likely to remain in my STEM major through to graduation or completion of my program of study.
- 10. I intend to leave my science major or science related track.
- 11. I do not intend to leave my STEM major before I graduate or complete my program of study.
- 12. I am not going to remain in a major or track in the STEM fields.
- 13. I am going to remain in a major or track in the STEM fields,
- 14. I have considered switching majors to one that is NOT in a STEM field.

Thank you for completing this survey!

If you have any questions about this study or further comments, you may contact Michele Morris at 256.327.0576 or mmorris@hudsonalpha.org

Student Focus Groups Interview Script & Questions (to be used after each subject provides verbal consent)

Thanks for taking the time to talk with me today. I'm interested to hear from all of you, so please don't be shy. I'd like to ask you a few questions to elaborate on the CODE survey you took recently and help me understand your responses better.

1. What did you know about bioinformatics before the CODE project?
2. Did you enjoy the bioinformatics aspects of this project? If not, why not? If so, why?
3. Do you think you will continue to learn more computational biology techniques? If not, why not? If so, why?
4. Tell me how you think you have changed, if at all, since participating in this research project.
5. How qualified do you feel about conducting a science research project now?
 - a. Why do you feel this way?
 - b. How do you think the CODE project contributed to how qualified you feel about conducting science research?
6. How did participating in this research project change your image of yourself as a scientist?
7. Did this experience change your preferences for a future career?
8. What are your plans for your future concerning your major and career field?
9. Would you like to participate in other bioinformatics activities?
10. Would you like to add anything?

Appendix F. Survey Instruments for Facilitator Participants

#31465093.0

IRB Approved at the

Study Level

CODE Project Evaluation Summative Instruments for Facilitator Participants

CODE Facilitator Training Pre-Workshop Survey

Thank you for helping us better understand the effects of participating in the CODE Facilitator Training Workshops. We value your responses. These surveys are confidential.

1. Gender
 - a. Female
 - b. Male
 - c. Prefer to not answer
 - d. Other
2. Ethnicity
 - a. Hispanic or Latino
 - b. Not Hispanic or Latino
 - c. Rather not say
3. Which race best describes you? (Please choose only one.)
 - a. American Indian or Alaskan Native
 - b. Asian/SE Asian
 - c. Black or African American
 - d. Native Hawaiian or Pacific Islander
 - e. White/Caucasian
 - f. Two or more races
 - g. Prefer not to say
 - h. Unknown
4. Are you a person living with a disability?
 - a. Yes
 - b. No
 - c. Prefer not to answer
5. Highest Degree Earned
 - a. B.S. (or equivalent)
 - b. M.S. (or equivalent)
 - c. Professional degree (e.g., M.D.)
 - d. Ph.D. (or equivalent)
 - e. Other, please explain: _____

6. Please select the statement below that best describe you
 - a. I teach at a 4-year college or university
 - b. I teach at a 2-year college or technical school
 - c. I am a graduate student
 - d. Other: _____

7. Is your institution classified as minority-serving?
 - a. Yes
 - b. No
 - c. Don't know

8. What is the total number of undergraduate students at your institution?
 - a. < 5,000 students
 - b. 5,000 – 15,000 students
 - c. >15,000 students
 - d. Don't know

9. What is the name of your department/unit (e.g., Department of Biology, Department of Biochemistry, School of Interdisciplinary Informatics)?

10. Which of the following best describes your level of bioinformatics training?
 - a. No training/experience
 - b. No formal training (self-taught)
 - c. Short workshop/bootcamp
 - d. Some undergraduate courses
 - e. Undergraduate certificate
 - f. Undergraduate degree
 - g. Post-graduate certificate
 - h. Graduate courses
 - i. Graduate degree

11. I currently or have in the past included bioinformatics training or activities in my classes.
 - a. Yes
 - b. No
 - c. Not applicable

12. What is the level of the courses with bioinformatics content that you teach?
 (Check all that apply)
 - a. Freshman
 - b. Sophomore
 - c. Junior
 - d. Senior
 - e. Graduate level
 - f. I do not teach courses with bioinformatics content

For each statement, indicate how well the statement applies to you using the scale below:

1 – strongly disagree

- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

- 1. I have a good understanding of what bioinformatics is.
- 2. I am interested in mentoring bioinformatics research projects for my students.
- 3. I am comfortable finding biological information in publicly available databases.
- 4. I am interested in analyzing biological information.
- 5. I am comfortable using computer programs to visualize 3D images of molecules.
- 6. I am comfortable mentoring bioinformatics research projects for undergraduates.

CODE Facilitator Post-Workshop Survey (immediately following workshop)

Thank you for helping us better understand the effects of participating in the CODE Facilitator Training Workshops. We value your responses. These surveys are confidential.

You may review the consent document that you agreed to for this study at this link [insert link to PDF of student consent form].

For each statement, indicate how well the statement applies to you using the scale below:

- 1 – strongly disagree
- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

1. I have a good understanding of what bioinformatics is.
2. I am interested in mentoring bioinformatics research projects for my students.
3. I am comfortable finding biological information in publicly available databases.
4. I am interested in analyzing biological information.
5. I am comfortable using computer programs to visualize 3D images of molecules.
6. I am comfortable mentoring bioinformatics research projects for undergraduates.
7. This workshop has increased my understanding of the concepts and skills relating to protein modeling and variant research.
8. The workshop content and activities were stimulating and challenging.
9. The pace of the workshop was appropriate.
10. The amount of material included in the workshop was appropriate.
11. Overall, I gained a good deal of knowledge and new skills from this workshop.
12. The skills and activities covered in this workshop will be beneficial in developing research projects with my students
13. I believe that my participation in the CODE program will help increase opportunities for undergraduate research projects in my department.
14. I anticipate that my participation in CODE will positively affect my reputation in my department.
15. CODE provides support for flexible implementation and curriculum development.
16. I plan to use the website HudsonAlphaCODE.org and the connections I made here in my teaching and research endeavors.
17. The CODE workshop provided a good network of technical expertise and support.
18. Through the CODE workshop I have gained colleagues in genomics.
19. The CODE workshop will be important in helping me plan my teaching of bioinformatics.
20. The CODE workshop contributed to my confidence in applying these techniques in my individual research.

Open response questions:

1. What additional information would you like to see included in this workshop?
2. What has been the most beneficial aspect of this workshop?
3. If you were the organizer, is there anything you would have done to improve this training session?
4. Do you have any additional comments? If so, please provide them here.

CODE Facilitator Extended Post-Workshop Survey (6-9 months following Workshop)

Thank you for participating in a final survey about the CODE project resources and how you may have applied the training you received in the Facilitator Training Workshop. We would appreciate your responses to the following questions that will be used to improve and evaluate the program. These surveys are confidential.

You may review the consent document that you agreed to for this study at this link [insert link to PDF of student consent form].

For each statement, indicate how well the statement applies to you using the scale below:

- 1 – strongly disagree
- 2 – disagree
- 3 – neither agree nor disagree
- 4 – agree
- 5 – strongly agree

1. I have a good understanding of what bioinformatics is.
2. I am interested in mentoring bioinformatics research projects for my students.
3. I am comfortable finding biological information in publicly available databases.
4. I am interested in analyzing biological information.
5. I am comfortable in using computer programs to visualize 3D images of molecules.
6. I learned new bioinformatics applications or techniques through the CODE Facilitator Training workshop.
7. The workshop increased my understanding of the concepts and skills relating to protein modeling and variant research.
8. I am comfortable using the bioinformatics approaches included in the CODE program to teach my course(s).
9. Because of the CODE program, I implemented new content or teaching approaches in my course(s) sooner than I otherwise would have.
10. I feel that my students have a better understanding of bioinformatics after participating in a CODE project.
11. My students are better prepared to use bioinformatics approaches than before the initiation of the CODE program.
12. Other faculty in my department have expressed interest in applying this approach to their area.
13. The skills and activities covered in this workshop have been beneficial in developing research projects with my students.
14. My participation in the CODE program has increased opportunities for undergraduate research projects in my department.

15. I believe that my participation in CODE has positively affected my reputation in my department.
16. CODE provides support for flexible implementation and curriculum development.
17. I have used the website HudsonAlphaCODE.org and the connections I made during the workshop in my teaching and research endeavors.
18. The CODE workshop provided a good network of technical expertise and support.
19. Through the CODE workshop I have gained colleagues in genomics.
20. The CODE workshop has been important in helping me plan my teaching of bioinformatics.
21. The CODE workshop contributed to my confidence in applying these techniques in my individual research.

ProQuest Number: 30693533

INFORMATION TO ALL USERS

The quality and completeness of this reproduction is dependent on the quality and completeness of the copy made available to ProQuest.



Distributed by ProQuest LLC (2023).

Copyright of the Dissertation is held by the Author unless otherwise noted.

This work may be used in accordance with the terms of the Creative Commons license or other rights statement, as indicated in the copyright statement or in the metadata associated with this work. Unless otherwise specified in the copyright statement or the metadata, all rights are reserved by the copyright holder.

This work is protected against unauthorized copying under Title 17,
United States Code and other applicable copyright laws.

Microform Edition where available © ProQuest LLC. No reproduction or digitization of the Microform Edition is authorized without permission of ProQuest LLC.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346 USA