DATA SCIENCE FOR MOLECULAR GENETICS AND COMMUNICATION IN THE NATURAL SCIENCES

A DISSERTATION SUBMITTED IN FULFILLMENT OF A REQUIREMENT FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

IN

MOLECULAR BIOSCIENCES AND BIOENGINEERING

UNIVERSITY OF HAWAI‘I AT MĀNOA

JULY, 2022

By

Bjarne R. Bartlett

Dissertation Committee:

Michael B. Kantar, Chairperson

Jon-Paul Bingham

Gernot Presting

Amy Hubbard

Monica Stitt-Bergh

ACKNOWLEDGEMENTS

My graduate education was supported by grants from the National Institutes of Health (NIH), National Institute of General Medical Sciences (NIGMS), IDeA Networks of Biomedical Research Excellence (INBRE, Award # P20GM103466) and the advanced computing resources from University of Hawai‘i Information Technology Services – Cyberinfrastructure – funded in part by the National Science Foundation (OAC, Award # NSCKLFSSABF2). I am grateful to the people who have played a role in my life and career up to this point. Both those who fostered rigorous, intellectual thought and those who embrace specious, unscientifc assumptions contributed to my growth as a scholar and reinforced my desire for knowledge. I thank my family and parents for teaching me the intrinsic value of schooling and an early introduction to computer skills, and I thank Dr. Chris Woo, for supporting me as my hānai uncle and dentist while I have been a resident of Hawaii.

I thank each of my mentors for their contribution to my pursuit for knowledge in bioinformatics, from cancer to agriculture, eventually leading me to pursue a graduate degree. Beginning with Drs. Luis Diaz, Bert Vogelstein, Kenneth Kinzler, Nicolas Papadopoulos, Shibin Zhao, Dung Le, Kevin “Wyatt” McMahon, Andrew Skora, Nicholas Roberts, and Chetan Bettegowda at Johns Hopkins, who modeled hard work and scholarship and encouraged me to pursue a research career. Dr. Youping Deng, who gave me my formal introduction to the field of bioinformatics. I have the utmost gratitude to my thesis advisor, Dr. Michael Kantar. His unbridled support and guidance provided me with the confidence and drive for excellence in all that I attempt. Other noteworthy contributors to my work include committee members Jon-Paul Bingham, Gernot Presting, Monica Stitt-Bergh, and Amy Hubbard. Each committee member supported me with expertise across the diverse areas of my research I am thankful for the Ideas Network for Biomedical Research Excelence (INBRE) and the Hawai‘i Data Science Institute (HI-DSI) for financial support and training during my graduate degree.

I am thankful to KTUH Honolulu and my mentors in public radio, Sarah Yap, Dale Machado, and Smee Wong. Many others not mentioned here have granted me the capacity to improve my personal and intellectual merits, of which I am infinitely thankful.

LIST OF FIGURES AND TABLES

GLOSSARY

- **Bibliometric:** The statistical analysis of analysis of academic journals, articles, and other publications.
- **Big Data:** Data that utilize existing analytical technologies but are applied faster and on a greater scale than before.
- **Biomarker:** A biological molecule that is used to signal information about an underlying biological process or disease.
- **Cross-disciplinary:** Representing more than 1 academic branch of knowledge.
- **Derived Principles:** Common piece of knowledge present across different disciplinary studies
- **Multidisciplinary:** Combining multiple academic branches of knowledge.
- **Student Learning Objectives:** Measurable, definable goals created for students by the instructor to be completed over a specific period.
- **Learning by Proxy:** Gathering information and skills by means of emulation as an alternative to structured written or verbal instruction.

PREFACE

"By 2025, it's estimated that 463 exabytes of data will be created each day globally – that's the equivalent of 212,765,957 DVDs per day!" -World Economic Forum

Data science refers to the study of increasingly large and complex datasets. Data that are too large for standard tools (e.g., Excel, Google Sheets) to analyze are often referred to as "big data." While big data exists across many areas and is thought to be the path to answering many questions, there is still no consensus on the fundamental principles and skills needed to interact with big data. Further, skills to study big data are not universally taught systematically at the college level–the resulting gap in skills leaves students unable to analyze the same big data that are touted as the way to answer complex questions.

This dissertation proposes a plan to close the big data knowledge gap by incorporating data science principles from diverse disciplines into a biology curriculum. Specifically, essential information was distilled from three independent study systems in cancer diagnostics, plant genomics, and academic publishing. Each study system contributed a different perspective on skills and knowledge from analyzing big data. From these systems, I identified three critical areas that are central to using big data effectively.

From these diverse perspectives, I developed a model to assist instructors in constructing curricula that will work in many different biological contexts. I piloted the use of these principles in a summer course. I found that by incorporating instruction developed across knowledge areas, meaningful data science instruction can occur in any curriculum at any student level.

CHAPTER 1: DATA SCIENCE AND COMMUNICATION IN THE NATURAL SCIENCES

1.1 The Role of Big Data

"Science" refers to knowledge gained through systematic study – "data science" is a specific area in which increasingly complex and heterogeneous data sets are studied. It differs from statistics because the data are the subject of study. Recent developments in data science allow practitioners to explore diverse subject areas (e.g., biology, science communication) with similar tools. As the amount of data available to investigate questions in the natural world has surged, studying the data itself has become a prerequisite to performing complex analysis. Many such datasets utilize existing analytical technologies but are applied faster and on a grander scale than before – these are often termed big data. Aspects of the philosophy and process of data science have been conserved across disciplines, with dozens of fields utilizing similar tools to interrogate big data (Miller, 2013). These observations align with gradual trends toward more interdisciplinary research across the sciences (Porter et al., 2009). The increasingly multidisciplinary nature of science has led to the diversification of research teams across fields. For example, including both sociologists and ecologists in developing ecosystem models creates better outcomes (Heemskerk et al., 2003). Merging expertise across disciplines allows for a broader, more fundamental range of questions to be answered.

In this dissertation, I apply tools for big data analysis to three diverse subject areas: science communication, plant genetics, and cancer biomarker development. Further, I distill common concepts from these research efforts into curriculum development for an undergraduate course on data science. While the subject areas' questions differ, understanding the tools that can be used to ask questions correctly and make inferences about data are central to modern science.

.

1.2 Curricula Design for Data Science

Benjamin Bloom and colleagues published one of the most famous approaches to understanding curriculum sixty-six years ago (Bloom, 1956). They created a framework for categorizing educational goals called *Taxonomy of Educational Objectives*. The framework, colloquially called Bloom's Taxonomy, has been applied by teachers for decades and consisted of six major categories: knowledge, comprehension, application, analysis, synthesis, and evaluation.

This original framework was revised in 2001 when it was published in "*A Taxonomy for Teaching, Learning, and Assessment*" which became known as Bloom's Revised Taxonomy (Anderson, et al., 2001). The revised model emphasizes verbs to reflect less static learning and more process-driven learning (Figure 1). The verbs in Bloom's Revised Taxonomy focus on cognitive processes underlying tasks required of students. Evolving models for learning reflect an increased understanding of educational psychology. These models are scaffolds to aid teachers in building practical lessons and learning environments. They ensure that classes contain meaningful assessment strategies. Such models can also assist in aligning instruction with the lesson objective.

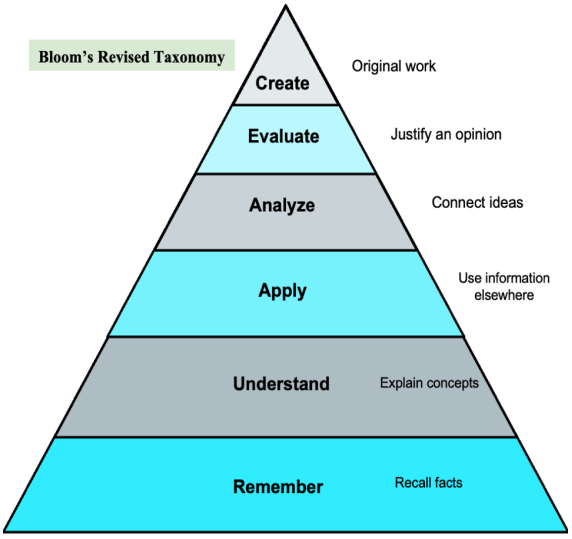


Figure 1: Bloom's Revised Taxonomy (Anderson, L.W., 2001)

While Bloom's taxonomy has been widely applied, it provides a one-dimensional way to explore a curriculum, that tends to be very hierarchical. This approach has some limitations. Another model that has been proposed provides a way to explore interactions: this model is called the model for significant learning developed by L.D. Fink and described in *Creating Significant Learning Experiences* (Figure 2; Fink, 2013).

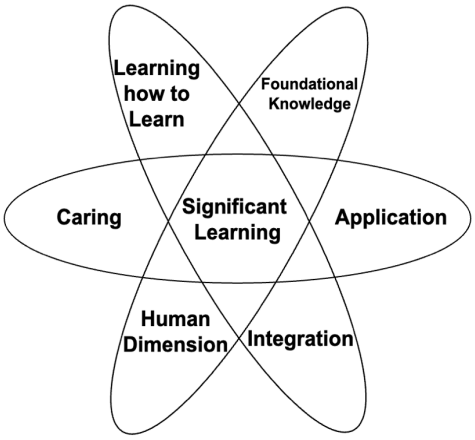


Figure 2. L. Dee Fink's Taxonomy for Significant Learning (Fink LD, 2013)

The Bloom and Fink models highlight an evolving understanding of how learning ties to underlying cognitive processes and provide the basis for how a broader context for learning

relates to learning data science. Fink's intersecting model for significant learning transforms the hierarchical, 1-dimensional Bloom model into a 2-dimensional design that emphasizes the integration of cognitive processes and content across different areas to promote significant learning – such a model hopefully encourages course designers to think about significant learning, not as a hierarchy, but as the simultaneous integration of cognitive processes across several areas.

To develop a holistic course based on educational models, it is essential to explore the empirical concepts in distinct disciplinary uses of data science methods. Understanding how to aggregate and analyze data are at the core of this thesis as they form the basis of all the questions being asked and how to gain insight from the data generated in all aspects of life. In the following sections of this chapter, the conceptual framework for each empirical study will be outlined, as will each chapter's connection to scientific and curriculum development goals.

1.3 The Role of Data Science in Cancer Research

Data science can be used to gain insight into cancer research, particularly in improving the diagnosis and treatment of the disease. The National Cancer Act of 1971 began the "war on cancer," a call to action and financial commitment by the United States to eradicate the disease. Figure 3 shows the death rates for top conditions in the United States, with the death rate for malignant neoplasms remaining remarkably consistent from the beginning of the war on cancer until 1998 (Figure 3A), despite the dramatic improvements in understanding the biology of human cancer. This begs the question: did the investment in science brought on by the war on cancer translate to patients? The big data accumulated from many years of cancer science, including genomics, addresses this question. As research moved into the 21st century, genetics

helped increase the rate of positive outcomes through new targeted therapies. Uncovering the genetic basis for cancer has brought our understanding of malignant neoplasms from a black box to a complex, modern version of the pathology behind the disease. Elucidating the genetics of human cancers has helped build a scaffold for targeted therapies to be developed (Vogelstein, 2013). One recent example was the FDA approval of the first tumor-agnostic cancer therapy based solely on a genomic biomarker (Le et al., 2015). Genomic information, supplied to scientists by the war on cancer, is beginning to result in progress for early diagnosis and targeted therapy of malignant tumors (Figure 3B). Genomic information is the type of big data that scientists can explore to help accelerate the understanding of cancer treatment.

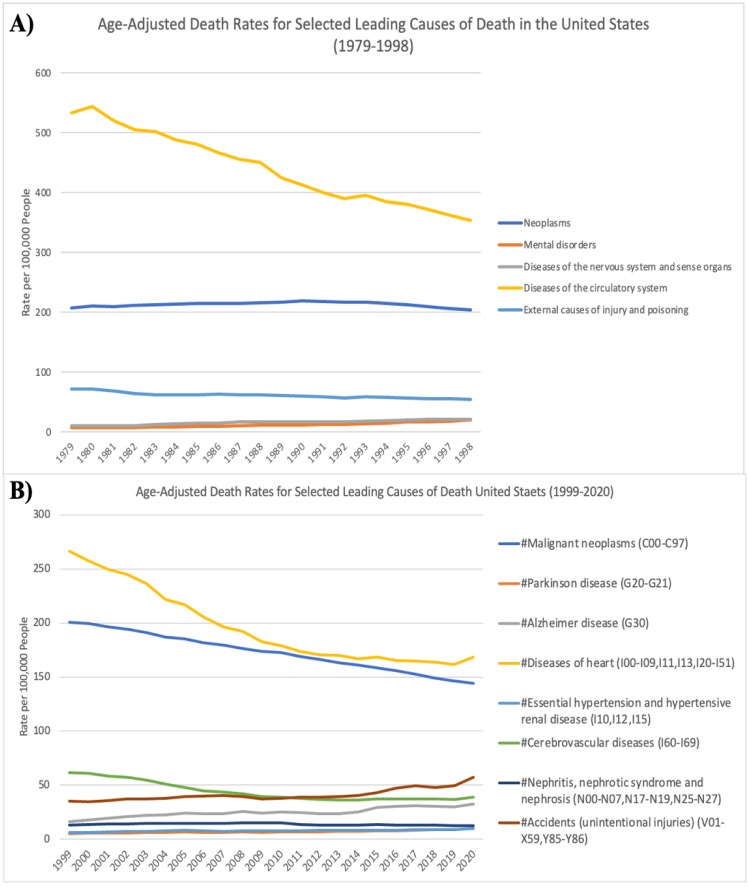


Figure 3: *Age-Adjusted Death Rates for Selected Leading Causes of Death in the United States During A: (1979-1998) B: (1999-2020). Causes of death are classified in accordance with the International Classification of Disease (ICD). Deaths for 1979-98 (A) are classified using the Ninth Revision (ICD-9). Deaths for 1999 and beyond (B) are classified using the Tenth Revision (ICD-10). Underlying cause-of-death is selected from the conditions entered by the physician on the cause of death section of the death certificate.*

SCIENTIFIC HYPOTHESIS

Companion therapeutics to current checkpoint inhibitors can be identified using the unique biology of MSI-H tumors.

CURRICULAR HYPOTHESIS

The unique structure of patient tumor genomic data (e.g., privacy issues, large numbers of individuals, small amount of sample) will provide critical insights into the ethical, technical, and reasoning aspects of data science that may be applied to an undergraduate informatics seminar.

1.4 Genomic Resources for **Macadamia tetraphylla** *and an examination of its historic use as a crop resource in Hawai‘i*

Most resources in science go to a small number of model systems (Farris, 2020). The development of big data has opened the door to exploring any species that a scientist is interested in. However, compared to other species, there is a significant disconnect between the ability to ask questions in a model system (e.g., drosophila, Arabidopsis, Human) and non-model systems. Many of the crop species in Hawai‘i are non-model species; this requires different approaches to understanding and using big data.

The impact of data science spans many disciplines—in addition to being on the front lines against cancer, it is also engaged in protecting our food supply. Macadamia nut is a high-value, nutrient-dense food crop (von Mueller, 1882; Lin et al., 2022). The most cultivated species is *Macadamia integrifolia.* In the future, macadamia nut farming is threatened by climate change (e.g., temperature and rising sea level -Arias, 2021) and increasing amounts of biotic stress (e.g., insect pests). Economic models estimate that, in the absence of effective

counteraction, climate change's overall costs and risks will equal a 5% decrease in the global gross domestic product (GDP) each year (Stern, 2006). The overall effect of climate change on agriculture is expected to be negative, despite potential gains in some crops in some regions of the world–posing a threat to global food security (Atkinson et al., 2008; Nelson et al., 2009).

Developing genomic resources for crop plants, such as Macadamia (e.g., genome -- Nock et al., 2020), offers the potential to improve production gains due to breeding (Rengel et al., 2015; Navarro and Rodrigues, 2016). Many factors threaten the Hawaiian macadamia industry, including pest and disease pressure; however, in many cases, there is no trait variation in *M. integrifolia*. Crop wild relatives (wild plants closely related to crop species) offer a significant reservoir of valuable traits (Castañeda-Álvarez et al., 2016; Vincent et al., 2019). The most crucial relative of *M. integrifolia* is *Macadamia tetraphylla* (Lin et al., 2022; Niu et al., 2022). Both *Macadamia* species are non-model systems, requiring new genomic and transcriptomic information to utilize them fully. To maximize the utility of this information, data science techniques are needed.

SCIENTIFIC HYPOTHESIS

Genomic resources for M. tetraphylla will allow for better characterization of introgression in Hawai‘i breeding material and remnant M. integrifolia orchard populations on O‘ahu.

CURRICULAR HYPOTHESIS

The structure of non-model species (e.g., crop wild relative) genomic data will provide critical insights into the ethical, technical, and reasoning aspects of data science that may be applied to an undergraduate informatics seminar.

1.5 Digital Technology Helps Remove Gender Bias in Academia

The same techniques used to evaluate the big data of genomes can be applied to other data types, such as metrics associated with academic publications. One goal of science is to identify general principles that result in long-lasting knowledge about the world. This knowledge

is intended to be objective, data-based, and helpful in stimulating new thought. Scientists achieve this goal through experimental methods, where researchers test specific, falsifiable hypotheses (Bratt et al., 2017) and conceptual, physical, mathematical, and computational models (Grimm et al., 2005). Some of this knowledge connects with the lay population; however, much of it does not, decreasing public trust in science (Pew Research Center, 2019). To understand the impact of science, much effort has been placed on measuring how the scientific community (e.g., h-index, i10) reacts to scientific work. Less effort has been put into how the public reacts to these scientific discoveries.

Big data are emerging in the areas of academic publishing and online attention. These datasets are increasing in size and scope, with more attention sources aggregated over time. Such data can democratize career evaluation, social mobility, and the amount of attention different types of material get (Raghavan et al., 2020; De Veirman et al., 2019; Enikolopov et al., 2018). There are direct ethical implications of working with data related to scientific publishing. A large body of recent literature has uncovered unconscious and conscious biases in metrics used to evaluate the proficiency of natural/social science and humanities scholars. For example, impact factors, h-indices, granting outcomes, and reference letters are repeatedly shown to present biases against women. The large amount of data collected by journals allow for testing if digital technologies amplify or mollify existing biases (Bakshy et al. 2015; Zou and Schiebinger 2018).

While journal-centric metrics are helpful, so are digital metrics that measure attention – the field of altmetrics (short for "alternative metrics") has emerged as a tool for quantifying digital attention (Erdt et al. 2016). The term "altmetric" refers to a variety of available metrics from digital media (e.g., blogs vs. Twitter vs. policy documents), with the Altmetric Attention Score (AAS) for individual journal articles being the most used aggregated metric (Bornmann et

al., 2018). It is commonly shown alongside citation scores and journal impact factors (Gumpenberger et al., 2016; Sopinka et al., 2020). The scale of the data (millions of data points across many years) provides opportunities to tease apart specific causes of increased attention and bias in who gets high scores.

SCIENTIFIC HYPOTHESIS

The gender biases that exist in traditional citation metrics will also exist in Altmetric Attention Scores.

CURRICULAR HYPOTHESIS

The structure of bibliometric data will provide critical insights into the ethical, technical, and reasoning aspects of data science that may be applied to an undergraduate informatics seminar.

1.6 The Cross-Disciplinary Value of Bioinformatics Curriculums in Higher Education

The data revolution in the life sciences has started shifting the discipline into a more quantitative era. Academia has responded by developing programs in quantitative methods for biology in both graduate and undergraduate programs (Attwood et al., 2019; Hack et al., 2005). Efforts to increase the number of informatics/data science courses in undergraduate curricula typically do not emphasize general principles of data science or develop skills required to perform analysis of multi-format data (McClatchy et al., 2020). Courses designed by ad-hoc practitioners without attention to a holistic curriculum run the risk of unbalanced treatment of subjects with the emphasis being placed on skills the instructor knows (DeMasi et al., 2020).

Successful curricula must articulate clearly defined general principles to give students the skills to choose the correct tools for analysis, assess the necessary computing resources, manage, and clean data, and apply ethically approved standards in data generation (Attwood et al., 2019). A pivotal element to reinforcing these general principles is to make learning active, where critical skills can be implemented along with theory (Shah et al., 2013; Freeman et al., 2014;

Neyhart JL., 2020). An active learning approach infuses theory into skill development through inquiry-based classroom experiences that can facilitate peer-to-peer learning (Smith et al., 2011). Understanding biological data has become a highly computational exercise (Ideker and Nussinov, 2017); key principles from different disciplinary contexts provide a crucial way to scaffold lessons to engage students. The curriculum analyzed in this chapter explores how a newly designed introductory bioinformatics seminar could leverage the combination of in-class instruction and independent research to build the practical research skill sets of undergraduate students beginning their research careers in the life sciences.

CURRICULAR HYPOTHESIS

Student Learning objectives based on multi-disciplinary synthesis of data science skills across diverse disciplines will result in positive student outcomes.

1.7 Summary

The empirical studies from diverse perspectives allowed me to interrogate the teaching of an introductory course. This enabled me to develop an instructional model for data science to assist instructors in designing a data science course for many diverse biological contexts. The model covered three key areas (Figure 4). The following chapters explore the specific scientific challenges that I undertook, the lessons I learned from teaching the course, and finally, the synthesis and presentation of the data science model that I found most parsimonious with the skills that students need to work with big data.

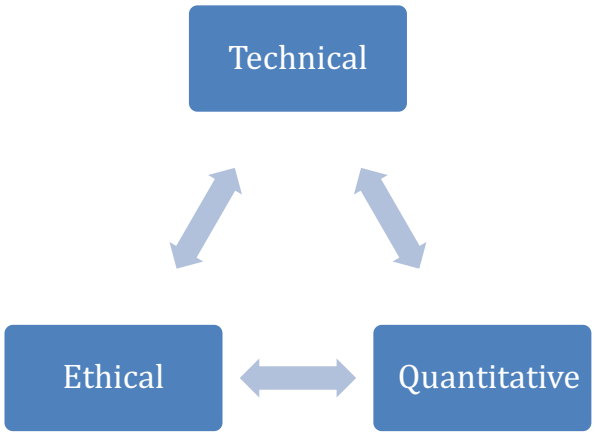


Figure 4: A Model for Data Science Instruction. The model spans 3 broad areas of knowledge: technical, ethical, and quantitative.

CHAPTER 2: THE MIRNA PROFILE OF INFLAMMATORY COLORECTAL TUMORS IDENTIFY TGF-B AS A COMPANION TARGET FOR CHECKPOINT BLOCKADE IMMUNOTHERAPY

"Now that we understand so much about the genetic basis of cancer, I'm optimistic we'll make progress in the years to come. But I also think that we need to readjust our efforts and spend more of our resources and intellectual energy on prevention and early detection."
~Bert Vogelstein, MD

Prologue

The "War on Cancer" saw little success in the first decades it was waged. In recent years, Bert Vogelstein's optimism in his statement above has been realized along with modest success in the "War on Cancer" waged by the United States of America. The genetic basis of many different cancers has been elucidated due to the vast amount of genomic sequence data that has been generated for patients at many different stages of the disease. This large-scale genomic data has helped cancer mortality to decline, and the first biomarkers based on tumor genetics have been approved by the FDA. Because many malignant tumors require multiple genetic changes and take about 20-30 years to develop – an enormous window of opportunity exists by which cancer can be detected based on genomic information.

In this study, an aggregated dataset consisting of genomic and clinical data from 549 colorectal cancer patients was interrogated. This effort sought to refine the current understanding of molecular biomarkers underpinning a subset of tumors with defective DNA mismatch—repair. Such tumors have been shown to respond to treatment by anti-PD-1 checkpoint blockade immunotherapy.

SCIENTIFIC HYPOTHESIS

Companion therapeutics to current checkpoint inhibitors can be identified using the unique biology of MSI-H tumors.

Analyzing diverse genomic data from these cancer datasets supplemented the overall curricular goals though: harmonizing data from multiple algorithms, addressing privacy concerns with human genomic data, and addressing technical limitations of different algorithms. Additionally, the scientific hypothesis was addressed through the development of new avenues for targeted therapy of cancers based on the unique biology of MSI-H tumors.

Published in Frontiers in Cell and Developmental Biology

Bartlett, B., Gao, Z., Schukking, M., Menor, M., Khadka, V. S., Fabbri, M., et al. (2021). The miRNA profile of inflammatory colorectal tumors identify TGF-β as a companion target for checkpoint blockade immunotherapy. *Frontiers in Cell and Developmental Biology*, 9.

2.1 Abstract

Extrinsic factors such as expression of PD-L1 (programmed death-ligand 1) in the tumor microenvironment (TME) have been shown to correlate with responses to checkpoint blockade therapy. More recently two intrinsic factors related to tumor genetics, microsatellite instability (MSI), and tumor mutation burden (TMB), have been linked to high response rates to checkpoint blockade drugs. These response rates led to the first tissue-agnostic approval of any cancer therapy by the FDA for the treatment of metastatic, MSI-H tumors with anti-PD-1 immunotherapy. But there are still very few studies focusing on the association of miRNAs with immune therapy through checkpoint inhibitors. Our team sought to explore the biology of such tumors further and suggest potential companion therapeutics to current checkpoint inhibitors. Analysis by Pearson Correlation revealed 41 total miRNAs correlated with mutation burden, 62 miRNAs correlated with MSI, and 17 miRNAs correlated with PD-L1 expression. Three miRNAs were correlated with all three of these tumor features as well as M1 macrophage polarization. No miRNAs in any group were associated with overall survival. TGF-β was predicted to be influenced by these three miRNAs ($p = 0.008$). Exploring miRNA targets as companions to treatment by immune checkpoint blockade revealed three potential miRNA targets predicted to impact TGF-β. M1 macrophage polarization state was also associated with tumors predicted to respond to therapy by immune checkpoint blockade.

2.2 Introduction

Despite therapeutic advances and declining mortality since 1990, an estimated 50,630 patients in the United States die annually from colorectal adenocarcinomas (Bray et al., 2018). New tools for precision medicine are necessary to build upon decades of progress in diagnosing and treating colon cancer. Immune Checkpoint inhibition (ICI) therapies, which block interactions between ligands and receptors, are one such innovation that has shown durable anti-tumor response. A combination of both intrinsic and extrinsic tumor features has been shown to correlate with response to checkpoint blockade therapy. Extrinsic factors, such as programmed cell-death ligand 1 (PD-L1) expression in the tumor microenvironment have been shown to correlate with responses to checkpoint blockade therapy (Topalian et al., 2012). More recently, two intrinsic factors related to tumor genetics, microsatellite instability (MSI), and tumor somatic mutation burden (TMB), have been linked to high ICI response rates (Snyder et al., 2014; Le et al., 2015). The high overall response rate (ORR) of solid tumors that are MSI-high (MSI-H) has led to the first tissue agnostic approval for a cancer therapy by the FDA in MSI-H metastatic tumors (Le et al., 2017; U.S. Food and Drug Administration, 2017). However, individual tumors continue to display a range of responses to checkpoint inhibition, highlighting the need for additional research to improve biomarkers and therapeutic approaches.

microRNAs (miRNA) are small, non-coding RNAs that usually function to regulate the expression of a particular gene by depleting the cellular protein contents. This is achieved post-transcriptionally through miRNA binding to a complementary part of the mRNA transcript for a specific protein. The binding of miRNA to mRNA primarily takes place in the 3′ untranslated region and results in either a particular mRNA not being translated or its degradation by the RNA interference effector complex (RISC) (Catalanotto et al., 2016). Because of their

importance in many cellular processes, the discovery of miRNAs has led to significant advances in the understanding and treatment of diseases including pharmacologic approaches. In the first pharmacologic use of miRNAs, Krutzfeldt et al. (2005) showed that a 23-nucleotide RNA molecule, complementary to the miR-122 target, could be delivered to liver tissue ablating endogenous miR-122.

Dysregulated miRNAs are a common feature of tumor cells that target oncogenes, tumor suppressor genes, and key immunologic pathways for tumorigenesis (Zhou et al., 2014; Chen et al., 2016; Fang et al., 2018; Vannini et al., 2018). miRNAs have been identified as important aspects of the molecular circuitry underlying cancer—miR-155, for example, has been found to be upregulated in many cancers. Van Roosbroeck et al. (2017) demonstrated that miR-155 directly targets *TP53*, thus functioning as an oncogene. Up till now, there have been several publications concerned with miRNA-based signatures in CRC screening programs. For example, miR-320d is found to be a promising non-invasive diagnostic biomarker that can significantly distinguish the metastatic from non-metastatic CRC patients (Tang et al., 2019). miR-378a-3p were identified as a potential circulating marker to differentiate the CRC patients from healthy subjects (Zanutto et al., 2020). Decreased exosomal miR-139-3p expression may take a role as a novel biomarker for early diagnosis monitoring in CRC patients (Liu et al., 2020). miRNAs have also been found to play an important role in regulating the immune environment. In addition to functioning as an oncogene, miR-155 was found by Lu et al. (2016) to promote M1 polarization along with miR-147-3p, and miR-9-5p. But there are still very few studies focusing on the association of miRNAs with immune therapy through checkpoint inhibitors in CRC.

In addition, the development of therapeutic targets that utilize RNA interference is an active area of pharmacologic research. Our team also sought to further explore the biology of

MSI-H tumors and suggest potential companion therapeutics to current checkpoint inhibitors. To do this, we initiated an *in silico* study to look at all three molecular phenotypes indicative of response to ICI therapeutics in the colon and rectal adenocarcinoma (CRC) cohorts from The Cancer Genome Atlas (TCGA) and further characterized changes in both the miRNA and transcriptomes.

2.3 Methods

Gathering Data

COAD data from The Cancer Genome Atlas was selected for analysis because many different types of analysis were available for the same patient cohort including somatic mutation burden, MSI status, mRNA analysis, and miRNA analysis. For our TCGA cohort, miRNA and mRNA expression data were procured from the Broad Firehose (Analysis-ready standardized TCGA data from Broad GDAC Firehose, 2016). Somatic mutation calls were obtained from the Genomic Data Commons for all CRC patients in TCGA (Grossman et al., 2016).

Obtaining Tumor Features

We chose tumor pathologies previously associated with response to checkpoint blockade immunotherapy for assessment in our CRC patient cohort from TCGA (Figure 5B). To compare miRNA expression between these tumor features, we also compared tumor phenotypes where one would expect a great deal of overlap, for example, MSI and TMB. MSI was assessed with the MicrOSAtellite Instability Classifier (MOSAIC) from Hause et al. (2016) to predict MSI status based on Whole Exome Sequencing (WES) data. The proportion of unstable microsatellite loci across the exome was correlated with the expression of miRNA. TMB was assessed using

Mutect2 and a 5% cutoff for allele frequency (Cibulskis et al., 2013). Expression of PD-L1 was assessed by quantifying gene expression—FPKM values from TCGA were used for this.

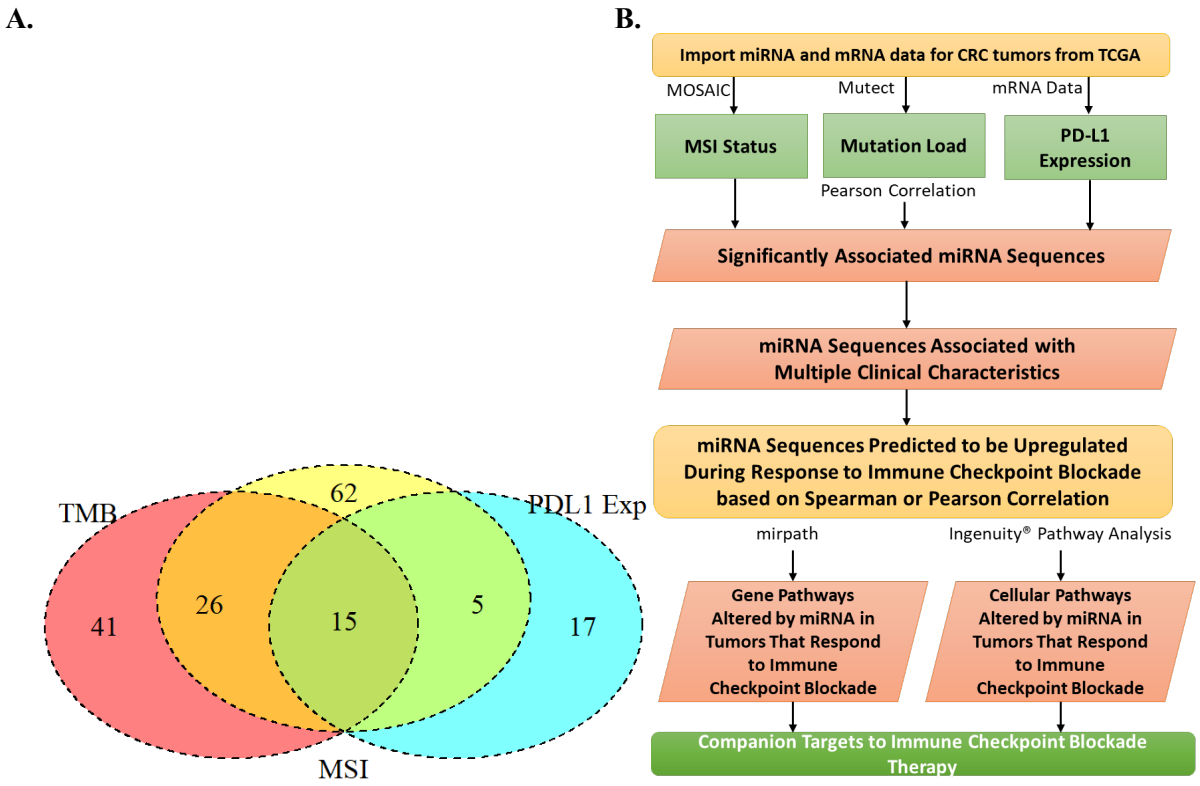


Figure 5. MiRNAs correlated with clinical features related to immunotherapy. (A) Tumor mutation burden, programmed death ligand 1 expression, CD8 fraction, and microsatellite instability were analyzed for a cohort of 549 colorectal cancer patients in The Cancer Genome Atlas. 15 miRNAs were identified that correlated with all 3 clinical features. (B) The whole analysis pipeline of the whole project.

Statistical Analysis

To assess whether each tumor feature was correlated with the presence of a particular miRNA, a Pearson correlation coefficient was used. miRNAs were individually assessed for correlation with each tumor feature. Once correlations were assessed for the different tumor features, miRNAs were pooled to look for miRNAs that were correlated with all 3 tumor features.

Immune Cell Deconvolution

In order to alleviate bias from any one algorithm, three separate tools were used to assess immune cell populations: xCell, TIMER, and CIBERSORT. CIBERSORT reports the fraction of 22 different immune cell lineages that are present in a particular RNA-Seq sample (Chen et al., 2018). xCell, similar to CIBESORT, is a gene signature-based method used to infer 64 immune and stromal cell types (Aran et al., 2017). The Tumor Immune Estimation Resource (TIMER) allows the calculation of six tumor-infiltrating immune subsets from gene expression data (Li et al., 2017). T-tests were used for each algorithm to determine whether the fraction of immune cells differed between phenotypic classifications of tumors. Once the group of miRNAs was determined to influence macrophage polarization in aggregate, each miRNA was individually assessed to determine whether it was correlated with macrophage polarization.

Pathway Analysis

mirPath (v3), a tool for predicting gene targets of miRNA sequences, was used to analyze which pathways the selected group of miRNA would preferentially (Vlachos et al., 2015). Once miRNA's were identified that correlated with macrophage polarization, these miRNAs were analyzed with mirPath to see which genes and pathways were targeted. TargetScan was queried using a conservation score of 0.1 to find genes and pathways intersected by miR-22, miR-155, or miR-146b (Karagkouni et al., 2018). Cancer-related genes and pathways were selected from those targeted by these miRNAs.

2.4 Results

Patient Cohort

The CRC patient cohort (n = 549) from TCGA was made up of 406 colon adenocarcinoma (COAD) patients and 143 rectal adenocarcinoma (READ) patients (Supplementary Table ST1). Typical immunotherapy recipients have late-stage cancers—we looked at stage in order to ensure a patient population representative of current immunotherapy recipients. We found that 14% of the total CRC patient cohort was advanced stage (IV). CRC patients were MSI-H at a rate of 18% in our CRC cohort, consistent with the literature. We chose three clinical features previously found to influence response to immunotherapy: MSI, tumor mutation burden (TMB), and PD-L1 expression. By aggregating these features, we aimed to predict an immunogenic subset of tumors from TCGA.

miRNAs Associated With Clinical Features Related to Immunotherapy

To characterize the relationship between miRNAs and the clinical features analyzed, a Pearson correlation was chosen. We measured linear correlations between each clinical variable and miRNA expression. Our Pearson correlation analysis resulted in 41 miRNAs significantly correlated with mutation burden, 62 miRNAs significantly correlated with MSI, and 17 miRNAs significantly correlated with PD-L1 expression. Of these three lists, 15 miRNA were overlapped and 12 of them were consistently positively correlated with the 3 tumor features and three of them were negatively correlated with the three tumor features (Figure 5A). 15 of these miRNAs were used for further analysis because they were correlated with all three tumor features (Table 1). To further characterize the 15 miRNA that were correlated with our clinical features, we

conducted pathway analysis revealing 2 immune-related pathways for further exploration: Colorectal cancer and TGF-β.

miRNA	**Association MMR- Tumor Features**	**Associated with Survival? (Z Score)**	**Associated with Macrophage Polarization**
let-7i	Up	No	No
mir-1266	Up	No	No
mir-132	Up	No	No
mir-146b	Up	No	Yes
mir-155	Up	No	Yes
mir-212	Up	No	No
mir-22	Up	No	Yes
mir-223	Up	No	No
mir-511(3p/5p)	Up	No	No
mir-625	Up	No	No
mir-629	Up	No	No
mir-335	Down	No	No
mir-552	Down	No	No
mir-92a-2	Down	No	No

Table 1: miRNAs found to be associated with 3 tumor phenotypes through the analysis described in Figure 5B. 15 miRNAs were found to have a common association with all 3 tumor phenotypes using Pearson's correlation. Whether the association was positive or negative was determined from the correlation coefficient, association with survival was determined using the R survival package, and association with macrophage polarization was determined using CIBERSORT and Pearson's correlation.

Association of Clinical Feature Related to Immunotherapy With Immune Cell Types

MSI, TMB, and PD-L1 expression were all separately assessed for Pearson correlations with the proportions of different immune cell types as reported by three separate immune cell deconvolution algorithms. Out of many cell types, only the proportions of plasma cells and M1

macrophages were significantly correlated with all three tumor features. The proportion of M1 macrophages was highly positively correlated with PD-L1 expression (Figure 6A, $p < 0.001$, Supplementary Figure SF2), MSI (Figure 6C, $p = 0.001$, Supplementary Figure SF1), and TMB (Figure 6E, $P < 0.001$, Supplementary Figure SF3). However, the proportion of plasma cells was negatively correlated with PD-L1 expression (Figure 6B, $p < 0.001$), MSI (Figure 6D, $p = 0.001$), and TMB (Figure 6F, $P < 0.001$). To further characterize the relationship between the proportion of M1 macrophages and the three tumors analyzed, we looked at correlations between M1 macrophage proportion and the expression of individual miRNAs. Among the 15 miRNAs, we found three miRNAs were significantly correlated with both the three clinical characteristics and M1 macrophage polarization: miR-22, miR-146b, and miR-155 (Figure 7). One miRNA, miR-220a, was excluded from further analysis because the correlation was based entirely on a single outlier.

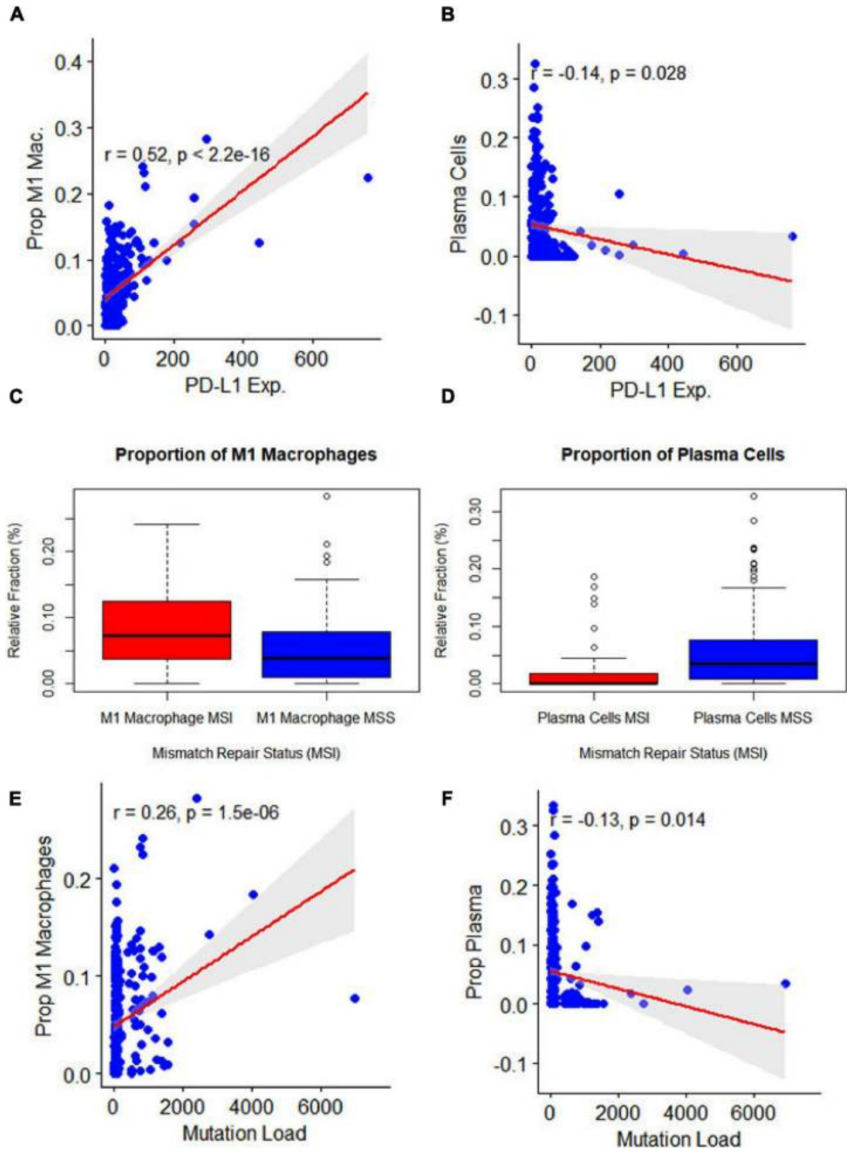


Figure 6: Association of clinical features related to immunotherapy with immune cell types **(A,B)**. Association of microsatellite instability status with: M1 macrophage polarization (p = 0.00) and plasma cells (p = 0.00). **(C,D)** Association of programmed death-ligand 1 expression with: M1 macrophage polarization (p = 0) and plasma cells (p = 0.03), y axis represented the immune cell deconvolution results as fraction relative to the immune-cell content: M1 macrophage and plasma cells. **(E,F)**: Association of mutation burden with: M1 macrophage polarization (p = 0.00) and plasma cells (p = 0.01).

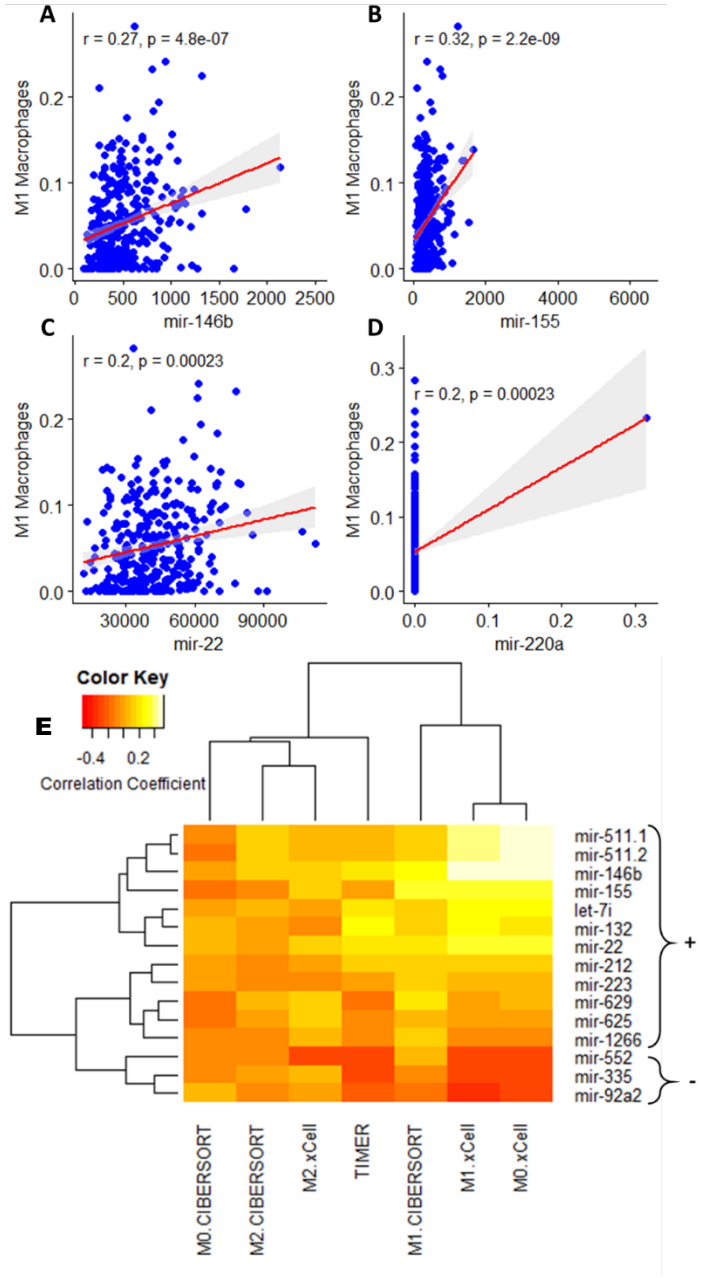


Figure 7: A-D. Association of M1 macrophage polarization with miRNAs. (A–D) Association of M1 macrophage polarization with: mir-146b (p = 0.00), mir-155 (p = 0.00), and mir-22 (p = 0.00). mir-220a was excluded as the correlation was the result of a single outlier. **E.** A heatmap of 15 miRNA sequences correlated with macrophage polarization.

Pathway Analysis

To characterize the crucial pathways for modulating the tumor immune environment, we predicted pathways that would be influenced by the three miRNAs related to both macrophage polarization and three tumor features. Unsurprisingly, these three miRNAs were predicted to influence the expression of genes in key immune and cancer-related pathways (Table 2). As a group, miR-155 and miR-22 were predicted to strongly influence pathways related to Colorectal Cancer ($p = 0.0001$) and TGF-β signaling ($p = 0.008$). Out of 21 genes predicted to be influenced by these miRNAs, three genes were shared between pathways related to COAD and TGF-β signaling: SMAD2, SMAD4, and TGFBR2.

Gene	Ensembl ID	TGF-β	CRC	hsa-miR-155-5p	hsa-miR-22-3p
SMAD2	ENSG00000175387	Yes	Yes	Yes	
ACVR1B	ENSG00000135503	Yes	No		Yes
SKP1	ENSG00000113558	Yes	No		Yes
ACVR2B	ENSG00000114739	Yes	No	Yes	Yes
SMAD4	ENSG00000141646	Yes	Yes		Yes
ZFYVE9	ENSG00000157077	Yes	No		Yes
ACVR2A	ENSG00000121989	Yes	No	Yes	
SP1	ENSG00000185591	Yes	No		Yes
EP300	ENSG00000100393	Yes	No		Yes
TGFBR2	ENSG00000163513	Yes	Yes	Yes	
FOS	ENSG00000170345	No	Yes	Yes	
GSK3B	ENSG00000082701	No	Yes	Yes	
PIK3CB	ENSG00000051382	No	Yes		Yes
KRAS	ENSG00000133703	No	Yes	Yes	
TP53	ENSG00000141510	No	Yes		Yes
PIK3CD	ENSG00000171608	No	Yes		Yes
CCND1	ENSG00000110092	No	Yes	Yes	
PIK3R1	ENSG00000145675	No	Yes	Yes	
AKT3	ENSG00000117020	No	Yes		Yes
PIK3CA	ENSG00000121879	No	Yes	Yes	
MAPK10	ENSG00000109339	No	Yes		Yes

Table 2: Genes targeted by miRNAs interact with in the TGF-β pathway and colorectal cancer pathway. MiRNA associated with microsatellite instability status, somatic tumor mutation burden, PD-L1 expression, M1 macrophage polarization that interact with the TGF-β signaling pathway ($p = 0.008$) and CRC pathways ($p = 0.0001$). Most of these genes interact with 2 miRNA sequences: hsa-miR-155-5p ($p = 0.004$) and hsa-miR-22-3p ($p = 0.038$). miRNA associations with genes were predicted by TargetScan (Conservation Score = 0.1). Results for TGF-β were merged by pathway union and results for CRC were merged by gene union.

2.5 Discussion

The aim of this study was to explore new targets for checkpoint blockade immunotherapy by exploring the unique biology of tumors known to respond to these drugs. Three features common to such tumors including high mutation burden, MSI, and PD-L1 expression were added into analysis. These features had 15 miRNAs in common, however, none of the 15 miRNAs predicted survival. M1 macrophage were found correlated with all three features through Pearson Correlation analysis. As a group, these 15 miRNAs predicted macrophage polarization. Individually assessing each of the 15 miRNAs for a correlation with macrophage polarization revealed three miRNAs that were strongly correlated with macrophage polarization: miRNA-146b, miRNA-155, and miRNA-22 (Figure 7:A-D). Subsequent pathway analysis revealed these three miRNAs as important components of the TGF-β and Colorectal Cancer pathways.

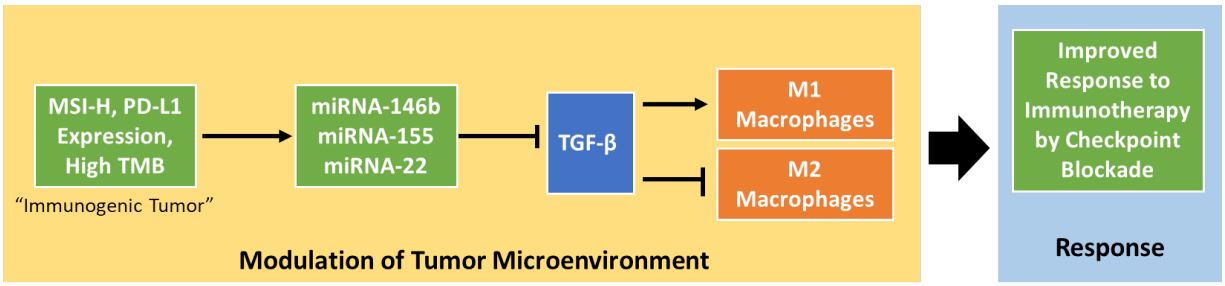


Figure 8: A potential mechanism by which MSI status, PD-L1 expression, and tumor mutation burden influence the tumor microenvironment. The proposed mechanism shows that these tumor phenotypes influence the tumor microenvironment in a TGF-β-dependent way to improve response to checkpoint blockade immunotherapy.

In this study, we searched all possible datasets from TCGA and GEO, TCGA is the only dataset that both contains miRNA and transcriptome, and our study analyzed TCGA data as controlling for bias using a randomly selected testing/training dataset. MicroRNA has been regarded as important promising molecular biomarkers in several tumor types (Zhang et al.,

2013; Chou et al., 2017). TGF-β has been identified as inhibiting the expansion and function of many components of the immune system (Batlle and Massague, 2019). A recent pair of papers has shown TGF-β to be an important modulator of the tumor microenvironment (Mariathasan et al., 2018; Tauriello et al., 2018). These experiments identify TGF-β signaling as an important aspect of response to PD-1-PD-L1 immunotherapy, connecting it to lower proportions of T cells in the tumor and poorer responses. This research supports the discovery of miRNAs targeting TGF-β in immunogenic tumors. TGF-β has also been shown to modulate the proportion of macrophages in the tumor microenvironment, promoting their polarization to an M2-like phenotype (Gong et al., 2012). Both ideas support a key role for suppressing TGF-β in immunogenic tumors that respond to checkpoint blockade therapy. Our research further characterizes this interaction by suggesting dysregulation of miRNA in immunogenic tumors as part of the biological system enabling responses to checkpoint blockade drugs (Figure 8).

Although, we didn't have validation of miRNA panel, TGF-β has been widely identified in many biological experiments which can be solid support evidence for the hypothesis we presented in this research. For example, miR-146b has been found to inhibit TGF-β by binding to the 3′ untranslated region (UTR) of SMAD4, an important member of the signaling pathway. Increased SMAD 4 levels and decreased cellular proliferation was observed by Geraldo et al. (2012) in human papillary carcinoma cells. Another study found the overexpression of SMAD4 in BCPAP cells, which is a validated target of miR-146b-5p and key protein in the TGF-β signaling pathway, significantly decreased migration and invasion to a degree very similar to that observed with the antagomir-146b-5p (Lima et al., 2016). miR-155 is one of the most extensively studied miRNAs and was the first miRNA shown to be oncogenic. An extensive body of research has established an important role for miR-155 throughout cellular process

related to human cancer (Costinean et al., 2006; Volinia et al., 2006). Geraldo et al. (2012) showed that miR-22 is significantly downregulated in TGF-β treated HT-29, a commonly used human colorectal cancer cell line (Cai et al., 2013).

In this study, we identified three miRNAs common in three immunotherapy-related clinical characteristics as well as M1 macrophage polarization, and function prediction of miRNAs showed SMAD2, SMAD4, and TGFBR2 were in common from COAD and TGF-β signaling pathways. miR-155 and miR-22 could influence pathways related to Colorectal Cancer and TGF-β signaling. Previous studies have already proved the regulation function of SMAD2, SMAD4, and TGFBR2 in cancers (Matsuzaki et al., 2009; Zhu et al., 2020), and these genes were also found related with miRNAs that strongly correlated with tumor features, indicating the potential function and clinical utility in immunotherapy.

2.6 Conclusion

Our comprehensive, integrated analysis of three miRNAs in colorectal cancer revealed a crucial component of TGF-β that modulate tumor immune environment and significantly correlated with macrophage polarization. The work highlights the important clinical implications of miRNAs functions in checkpoint blockade immunotherapy and helps develop potential therapeutical strategies for CRC patients.

CHAPTER 3: GENOMIC RESOURCES FOR *MACADAMIA TETRAPHYLLA* AND AN EXAMINATION OF ITS HISTORIC USE AS A CROP RESOURCE IN HAWAI ʻI

A man saw Nasrudin searching for something on the ground.
'What have you lost, Mulla?' He asked.
'My keys,' said the Mulla.
The man joined him asking : "Do you remember where you dropped them?"
Mulla answered: 'in my house.'
'Why are you looking here?', asked the man confused.
Mulla Nasrudin replied: 'Because there is more light here than in my house.'
~I. Shah, The Exploits of the Incomparable Mulla Nasrudin

Prologue

Model systems look for answers where there is light. When using conventional model species, biologists run the risk of constraining research in a way that fits that specific model, missing critical ecological information, unique to important, non-model species. While model species are often the most convenient system for research, this convenience does not necessarily make them the best choice for a particular experiment.

This chapter utilizes and combination of publicly available genomic resources for *Macadamia tetraphylla*, a crop-wild relative of *Macadamia integrifolia* as well as newly generated data on Hawaiian breeding line and treed found in remnant orchards. Through the development of a *de novo* transcriptome for the crop wild relative it was possible to explore historic hybridization in the newly collected material. Transcriptome construction relies on generating and manipulating large amounts of data and then simplifying this data into easy-to-understand pieces. New data can then be compared to these known parts to gain insight into the biology of the species.

SCIENTIFIC HYPOTHESIS

Genomic resources for M. tetraphylla will allow for better characterization of introgression in Hawai‘i breeding material and remnant M. integrifolia orchard populations on O‘ahu.

The analytical skills required for the *de novo* transcriptome and the hybridization analysis incorporated technical and quantitative interpretation that was unique to non-model genomic system that bring non-standard data to standardized tools. The scientific hypothesis was partially addressed through the development of new genomic resources, however travel limitations presented by the COVID-19 pandemic and sequencing-quality constraints limited the extent of these resources.

In preparation for BMC Research Notes

Bartlett, B., Cho, A., Presting, G., Laspisa, D., Gore, M.A., Kantar, M.B. 2022. Genomic Resources for Macadamia tetraphylla and an examination of its historic use as a crop resource in Hawai‘i. *BMC Res Notes* (In preparation)

3.1 Abstract

Macadamia tetraphylla is a wild relative of the economically valuable crop *Macadamia integrifolia.* Genomic knowledge of crop wild relatives is central to determining their possible role in breeding programs to mitigate biotic and abiotic stress in the future.

Objectives: The goal of this project was to develop a genomic resource for macadamia agriculture in Hawai'i through constructing a transcriptome of *M. tetraphylla* and testing for hybridity in University of Hawai'i breeding material.

Results: The advanced breeding lines (HI 862, HI 879) were confirmed to be hybrid with crop wild relatives using. The putative hybrid was shown to be hybrid; however, the percentage ancestry indicated the possibility of being an F_1 with some segregation distortion or a late generation hybrid. Additionally, the hybridity of putative pure *M. tetraphylla* from a remnant orchard planting on O'ahu did not match anecdotal accounts of the orchard. Further, the transcriptome assembly of *M. tetraphylla* showed there to be large differences in expression based on tissue type.

3.2 Introduction

Macadamia nut (*Macadamia integrifolia*; 2n = 2x =28) is an important global crop originating in Australia (von Mueller, 18; Lin et al., 2022). Recently, there has been increased interest in developing genomic resources for this crop (e.g., genome -- Nock et al., 2020) due to its importance as a high-value, nutrient-dense food (Rengel et al., 2015; Navarro and Rodrigues, 2016) that has undergone relatively little commercial breeding. For example, in Hawai‘i, there has been an increase in resources dedicated to the crop due to its large land uses (17,100 acres in 2019) and high value ($42 million USD) (USDA-NASS, 2019). These efforts have identified many agronomic problems in the current orchards that can be attributed to the need for genetic improvement (Gutierrez-Coarite et al., 2021). There is, however, limited genetic diversity within Hawaiian *M. integrifolia* that can be used for breeding purposes (Hardner, 2016). Many factors threaten the livelihoods of the macadamia industry in Hawai‘i including pest and disease pressure, however, there is no adapted *M. integrifolia* tolerance for the most important pressures, requiring the use of a different donor species.

In species with limited genetic diversity, the field of plant breeding commonly uses related species as a donor to improve specific traits (Castañeda-Álvarez et al., 2016). In fact, crop wild relatives and the genetic material that comes from them have been valued at over $100 billion U.S. dollars annually (PwC, 2013). The most important relative of *M. integrifolia* is *Macadamia tetraphylla* (2n = 2x = 28), a species that is occasionally cultivated but more frequently used as a donor for useful biotic stress tolerance traits (Lin et al., 2022; Niu et al., 2022). To best operationalize the use of unadapted plant material for specific localities, genomic information is required (Wambugu and Henry, 2022). While there is a history of breeding in Hawai‘i, gaps in staffing have caused records to be lost, therefore there is also a need to

corroborate ancestry in current promising breeding lines. Recently, a genome of *M. tetraphylla* was released (Niu et al., 2022), providing a useful resource to characterize Hawaiian *M. tetraphylla.* Therefore, the goals of this study were to i) characterize the *M. tetraphylla* transcriptome and ii) explore the ancestry of promising University of Hawai‘i breeding lines.

3.3 Methods

A transcriptome assembly of *M. tetraphylla* was constructed from existing public data along with new RNA-sequencing data generated from wild-collected, feral *M. tetraphylla* and putative hybrid breeding lines in Hawai‘i with RNA from leaf tissue was extracted using RNAeasy Plant Mini Kit (Qiagen, Valencia, California). RNA was then sent to Novogene Bioinformatics Technology Co., Ltd., for sequencing, which was conducted on an Illumina NovaSeq with 150 bp paired-end reads. Next-generation sequencing data (PRJNA587821) was obtained for *M. tetraphylla* for 5 different tissue types and assembled using Trinity (Haas et al., 2013). Coding regions within transcripts were identified with TransDecoder V.5.5.0 (Haas & Papanicolaou, 2012). Annotation of transcripts was done with BLASTX (Camacho et al. 2009) using the UniProt database to identify homologous sequences (UniProt, 2021). HMMER (Eddy, 2011) and the PFAM (Mistry et al. 2021) database were used to identify protein domains. Data for all analyses can be seen in Table 3. A read set was generated using combined RNAseq data from 5 tissues (bark, proteoid root, flowering inflorescence, young inflorescence and leaves, SRR10424518-SRR10424522) that were trimmed, and quality checked using the same methods as for the Trinity assembly. The combined read set was mapped to the *M. tetraphilla* physical map (ASM2298504v1) using TopHat2 (default parameters). The resulting alignment file was filtered for unmapped reads (samtools view -F 0x04), sorted and indexed using samtools.

Annotations were generated from the sorted alignment file using Cufflinks 2.2.1 (default parameters) which generated the GTF file used in the analysis. The alignment file and annotations were manually inspected in IGV for validation. The tool sppIDer (Langdon et al., 2018) was used to assess hybridity in four putative *M. integrifolia* × *M. tetraphylla* hybrids from the University of Hawai'i accessions sequenced here.

Table	Name of data file	File type
Transcriptome Leaf of *M. tetraphylla*	SRR10424522_leaves.fasta	.fasta
Transcriptome young inflorescence of *M. tetraphylla*	SRR10424521_younginflorescence.fasta	.fasta
Transcriptome flowering inflorescence of *M. tetraphylla*	SRR10424520_floweringinflorescence.fasta	.fasta
Transcriptome proteoid root of *M. tetraphylla*	SRR10424519_proteoidroot.fasta	.fasta
Transcriptome bark of *M. tetraphylla*	SRR10424518_bark.fasta	.fasta
VCF file of *M. integrifolia* × *M. tetraphylla* hybrid Hawai'i 879	MT1_879.vcf	.vcf
VCF file of O'ahu Tetraphylla	MT5_OT.vcf	.vcf
VCF file of O'ahu putative M. integrifolia × M. tetraphylla hybrid	MT9_OTPH.vcf	.vcf
VCF file of M. integrifolia × M. tetraphylla hybrid Hawai'i 862	MT10_862.vcf	.vcf

Table 3: Overview of data files. Repository at github.com/bjarnebartlett/Interspecific-Breeding-History-of-Hawaiian-Macadamia

3.4 Results

Differential gene Expression across tissue type

The de novo *M. tetraphylla* transcriptomes showed a high level of completeness (Figure 9A) using Benchmarking Universal Single-Copy Orthologous (BUSCO) for bark (97.65% complete), proteoid root (98.43% complete), flowering inflorescence (98.43% complete), young inflorescence (99.61% complete), and leaves (98.82% complete). For validation and annotation, the assembled sequence was compared to the KEGG and PFAM databases (Ogata et al., 1998; Mistry et al., 2021). Combining all of the gene clusters across tissues a total of 46,967 transcripts were called representing 34,444 unigenes and 12,523 isoforms, this is similar to previous reports (Niu et al. 2022). The additional genes called here may be the result of TE/repeat derived transcripts called because of the multi-mapping approach used or the use of default parameters that may not be optimal. There was clear differential expression between tissue types with a

transcript being considered expressed if FPKM greater than 10 (Figure 9B). There was also clear enrichment for different KEGG pathways in each tissue type; pathways were considered expressed if each member in the pathway had an FPKM greater than 10 (Figure 9C).

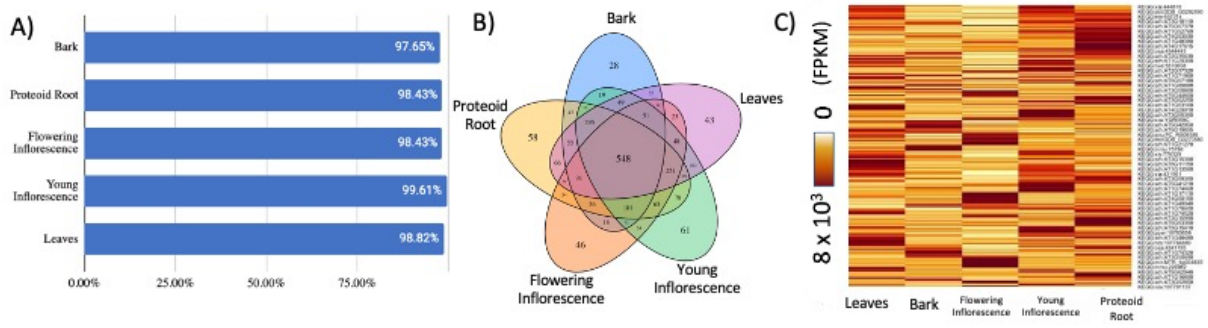


Figure 9. A) The de novo M. tetraphylla transcriptomes showed high completeness based on Benchmarking Universal Single-Copy Orthologous (BUSCO) quality assessments. B) Number of genes expressed in different tissues by KEGG Pathway in M. tetraphylla. Total number of genes with FPKM > 10 across proteoid root (PR), flowering inflorescence (FI), young inflorescence (YI), and leaves. C) Gene expression of KEGG pathways represented as FPKM (fragments per kilobase million) in M. tetraphylla for leaves, bark, flowering inflorescence, young inflorescence, and proteoid root, clear enrichment for different functional pathways is seen across different tissue types.

Hybridity of UH Mānoa Breeding Lines

Combining the information from the de novo assembly and the previous sequence showed that the new assembly has potential to inform *M. integrifolia* breeding. The advanced breeding lines (HI 862, HI 879) were confirmed to be hybrid with crop wild relatives using the tool sppIDer (Figure 10; Supplemental Table ST2). The breeding materials were shown to be BC_1 lines, despite their pedigrees indicating that they should be later generation hybrids. The putative hybrid was shown to be hybrid, however, the percentage ancestry (~35% *M. integrifolia*) may mean it is an F_1 with some segregation distortion or a late generation hybrid. Surprisingly, the putative pure *M. tetraphylla* from a remnant orchard planting on O‘ahu also was found to be a hybrid, again likely an F_1 or a later generation hybrid. Using the HyDe tool

overall hybridity was confirmed for HI 879 (HyDe Z-score = 48.53, $p < 0.01$), but not for HI 862.

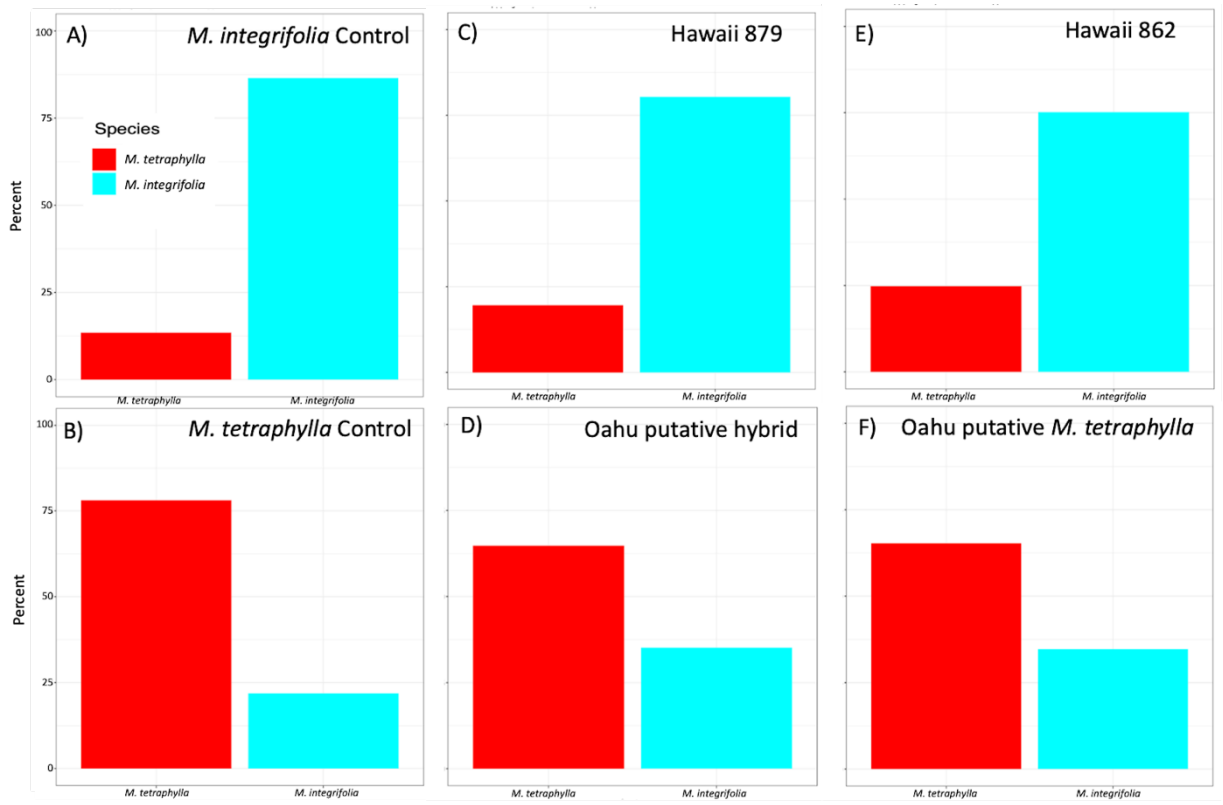


Figure 10. Hybridity of breeding lines, wild collected putative hybrid, and a wild collected putative pure M. tetraphylla. A) represents a M. integrifolia control , B) represents a M. tetraphylla control, C) represents a putative hybrid breeding line from UH Manoa, D) represents a putative hybrid collected from a remnant orchard, E) represents a putative hybrid breeding line from UH Manoa, and F) represents a putative pure M. tetraphylla collected from O'ahu. All tested samples (C,E, F) were identified as putative hybrids..

3.5 Discussion

It is now a routine procedure to construct genomic resources in non-model plants (Wambugu and Henry, 2022). This is an important step in increasing the efficiency of interspecific breeding, especially in species that have long generation time (Valle-Echevarria et al., 2021). Transcriptomes, which provide not only a rich source of expression dynamics, but

also insight into gene function is a rich source of information on crop relatives. Here we have found there is considerable tissue-specific expression across functional pathways.

Further, while the historic record suggested that there was both the growth of pure wild relatives and widespread hybridization between the important macadamia species (*M. integrifolia* and *M. tetraphylla*), we only found evidence of hybridization not only in the breeding lines but also in remnant orchard material on O‘ahu. Introgressions appeared to be consistent between individuals (HI 862 and HI 879), appearing to occur from chromosome 2, 5, and 9 of *M. tetraphylla* with additional introgression on chromosomes 7 and 8 in HI 862 (Supplemental Figure SF4.) There was more than expected (based on breeding records) introgressions within remnant individuals indicating admixture in the past. The breeding lines had higher levels of *M. tetraphylla* than expected when compared to pedigree histories in breeder records. While the two tools did not agree, there appears to be a meaningful amount of introgression in all lines tested on O‘ahu. This suggests that for new breeding material to meet the requirements of the industry in terms of yield and quality, new backcrosses to *M. integrifolia* will need to be initiated to fully gain the benefits of the biotic stress tolerance of *M. tetraphylla* while not having linkage drag associated with the wilder characteristics of *M. tetraphylla* that make cultivation difficult.

CHAPTER 4: DIGITAL TECHNOLOGY HELPS REMOVE GENDER BIAS IN ACADEMIA

"Women's rights are an essential part of the overall human rights agenda, trained on the equal dignity and ability to live in freedom all people should enjoy." ~ Ruth Bader Ginsburg

Prologue

Science attempts to be a meritocracy, to achieve this there have always been measurements of the visibility and quality of scientific research. Gender disparities in higher education are well known. Many of the traditional measurements have been shown time and again to favor men over women, exacerbating known biases within academic institutions. There is hope that new measurements will ensure the equity that Justice Ginsburg describes above. The advent of personal brands, social media, and the internet has led to a new way of exploring research impact. There is hope that newer measurements such as the altmetric attention score, which explores popular attention that scientific articles generate, will reduce bias.

To create the dataset for this analysis, 12 million articles were considered. Data were filtered to articles from seven journals over a ~10-year period resulting in 200,000 data points. This large dataset showed that contrary to traditional metrics, there appeared to be little gender bias in this scholarly metric.

SCIENTIFIC HYPOTHESIS

The gender biases that exist in traditional citation metrics will also exist in Altmetric Attention Scores.

Interrogating the altimetric attention score for bias informed the instructional model for data science in the interpretation of the score of individual articles and the associated prediction of the gender of the authors based their first name. This required sub-setting, cleaning, and generating of new data with the appending of the relevant new variable for analysis. This dataset presented unique challenges around ethics and analysis for unconscious bias, as it was a proprietary algorithmic product from a private company. The scientific hypothesis is not supported by the data, except in the journal *Science* during 2017-2018.

Published in Scientometrics

Fortin, J*., **Bartlett, B.***, Kantar, M., Tseng, M., & Mehrabi, Z. (2021). Digital technology helps remove gender bias in academia. *Scientometrics*, 126(5), 4073-4081.

*co-first author

4.1 Abstract

Science attempts to be a meritocracy; however, in recent years, there has been increasing evidence of systematic gender bias against women. This bias is present in many metrics commonly used to evaluate scientific productivity, influencing hiring and career success. Here we explore a new metric, the Altmetric Attention Score, and find no evidence of bias across many major journals (Nature, PNAS, PLOS One, New England Journal of Medicine, Cell, and BioRxiv), with equal attention afforded to articles authored by men and women alike. The exception to this rule is the journal Science, which has marked gender bias against women in 2018, equivalent to a mean of 88 more tweets or 11 more news articles and a median of 20 more tweets or 3 more news articles for male than female first authors. Our findings qualify Altmetric, for many types and disciplines of journals, as a potentially unbiased measure of science communication in academia and suggest that new technologies, such as those on which Altmetric is based, might help to democratize academic evaluation.

4.2 Introduction

A large body of recent literature has uncovered unconscious or conscious biases in several metrics used to evaluate the proficiency of natural/social science and humanities scholars. For example, impact factors, h-indices, granting outcomes, and reference letters are repeatedly shown to present biases against women (see our review in Table 4). Because these metrics are all used in decision-making within academic institutions, these biases present a severe impediment for equal opportunity among genders throughout their scientific careers.

The emergence of big data and new technologies has the potential to democratize career evaluation and social mobility. For example, machine learning algorithms have been used to reduce biased decisions in hiring, evaluation, and promotion (Raghavan et al. 2020). Additionally, social media has reduced employment barriers, as witnessed by the rise of professional influencers in advertising and has also increased the accountability of corporations in countries with otherwise highly censored traditional media (De Veirman et al., 2019; Enikolopov et al. 2018). While the benefits of these technological advances are significant, there is also concern that new digital technologies may also amplify existing biases, as seen in selective content exposure in social media platforms like Facebook (Bakshy et al., 2015) and in the way algorithms trained on biased data can default to male-gendered pronouns (Zou and Schiebinger, 2018).

Scientists now use digital media as a critical platform for disseminating scientific findings, and the field of altmetrics (short for "alternative metrics") has emerged as a tool for quantifying the digital attention received by scientific papers (Erdt et al. 2016). The term "altmetric" refers to a variety of available metrics that differ according to how they aggregate 'mentions' of scientific output on various digital media (e.g., blogs vs. Twitter vs. policy

documents). Altmetric.com is one of the largest aggregators of altmetric scores, and hundreds of journals now publish the Altmetric Attention Score (AAS) for individual journal articles (Bornmann et al. 2018). Altmetric scores have been shown to be positively correlated with traditional measures of impact, such as citations, h-indices, or impact factors in some fields (Kunze et al. 2020; Nocera et al. 2019; Thelwall and Nevill 2018). Previous studies have suggested that altmetrics such as the AAS are difficult to interpret due to lack of normalization and standardization, as well as the proprietary nature of the algorithms used for web scraping. Nevertheless, AAS in particular has become an important measure of how articles are perceived, and it is commonly shown alongside citation scores and journal impact factor (Gumpenberger et al. 2016).

Given the widespread use of altmetrics, scientists and policymakers have advocated for the incorporation of such scores when evaluating the overall impact of scientific papers (Sopinka et al. 2020). However, whether altmetric scores carry the same gender biases as traditional metrics used in scientific evaluation is unclear. To address this knowledge gap, here we investigated gender bias in AAS in seven major scientific publications for the years 2011–2018. A total of 208,804 journal articles were analyzed, representing ~ 1.6% of the 12 million research works covered by altmetric.com.

Type of evidence	Bias in favor of males	Bias in favor of females	Unbiased	Citation
h-index	x			García-Pérez et al. (2009)
Cite score	x			Dion et al. (2018)
Reference letters	x			Madera (2019)
Interviews	x			Quadlin (2018)
Grants	x			Morgan et al., (2018)
Invited papers	x			Holman et al. (2018)
Invited speakers	x			Nittrouer et al. (2018)
Altmetrics			x(for all journals except science)	(Present study)
Quotes in media	x			Morris (2016)
% first authorships	x			Filardo et al, (2015)
% last authorships	x			West et al. (2013)
Number of publications	x (given existing distribution of resources)		x (if controlling for position & funding)	Holliday, et al. (2014)

Table 4. Identification of performance areas where bias exists from published examples in the literature. Most of the common performance metrics within the Academy show bias in favor of men.

4.3 Materials and Methods

We targeted articles from: Science, Proceedings of the National Academy of Sciences (PNAS), PLOS One, New England Journal of Medicine (NEJM), Nature, Cell, and BioRxiv (Figure SF5). These journals represent idiotypic journal types, specifically general interest (Nature, Science, PNAS), open access (PLOS One), disciplinary (NEJM, Cell), and a preprint server (BioRxiv). This sample enabled us to explore bias across a range of journals in which scholars publish their work.

We extracted the following variables from the Altmetric data base: author names, journal, publication date (which we used to create independent variables in our models, see below) and Altmetric score 1 year after publication (which we used to create the dependent variables in our models). While AAS are available at different time intervals after article publication (e.g. 5 days,

1 month, 1 year, all-time), we used 1-year scores in our analysis, to both maintain a controlled exposure time, and to enable reasonable length of time for scores to amount.

We genderized author names using the genderizeR package in R (Wais, 2016). We extracted all author names and inputted first names into the genderize.io API, which returns a suggested gender and probability, based on the proportion of references in the genderize.io database. Names that can be considered unisex are assigned a gender based on the gender with the highest probability score, with a threshold of > 0.5. Names that do not appear in the genderize.io database are listed as "unknown". The genderize.io database includes names from over 79 countries and 89 languages. The error rate of genderize.io predictions has previously been estimated at 5.02% on a test dataset including names of European, African and Asian origin (Santamaría and Mihaljevic 2018). We created three variables using this method: first author gender, last author gender and proportion of female authors for each article. Our genderized dataset consists of 31% female-first-authored papers (61% male-first-authored, 7% unknown) and 21% female-last-authored papers (69% male-last-authored, 9% unknown).

As the majority (139,035 out of 208,804 = 66.6%) of articles in our dataset, with1-year exposure times, had an AAS of 0 (introducing error that cannot be normalized through a ln(x + 1) transform for Altmetric data as has been suggested elsewhere, see Thelwall 2020, we split our analysis into two simple questions: (1) does gender explain whether an article receives an AAS greater than 0; (2) does gender explain the magnitude of the score. We evaluated these two hypotheses using logistic and linear regression models, respectively.

The first model (model 1) had a binary dependent variable indicating whether an article received a score greater than 0, and a series of independent variables:

$$y_i = Bin(1, \boldsymbol{p_i})$$

$$p_i = logit^{-1}(\beta_0 + \beta_{1..m-1} Gender \cdot Journal \cdot Year_i + \gamma_1 Month_i + \gamma_2 No.Auth_i$$

where pi is the fraction of articles that have a score, β0 is an intercept, Gender × Journal × Year are dummy variables representing m 3rd order interactions between gender of the first or last author, journal and year, Month is the publication month, No.Auth is the total number of authors of the article and Prop.Fem is the proportion of authors in the author list that are female.

The second model (model 2) has the magnitude of an article's 1-year AAS (log-transformed) as a continuous response variable and included the same independent variables as the binomial model above:

$$log(y_i) = N(\mu_i, \sigma_i^2)$$

$$\mu_i = \beta_0 + \beta_{1..m-1} Gender \cdot Journal \cdot Year_i + \gamma_1 Month_i + \gamma_2 No.Auth_i$$

where $\mu_i$is E(log(y_i)).

We used these models for inference on the a priori null hypothesis of no difference between male and female first (or last) authors in the probability of obtaining an AAS(model 1), and in the magnitude of the score received (model 2). We computed simultaneous confidence intervals on the differences to account for multiple testing across journals, years, and for first and last authors.

BioRxiv was modeled separately because its available time series was shorter than the other journals in our sample (the repository was launched in late 2013). For all journals, we conducted standard model checking through visual diagnostics. The overall goodness of fit for the binomial model for journals excluding BioRxiv was a pseudo-R 2 of 0.29, and for the linear model the adjusted R 2 was 0.26; while BioRxiv had a pseudo-R 2 of 0.64 and an adjusted R 2 of 0.06.

We conducted a sensitivity analysis to check how the accuracy of the gender assignment impacted the model results. We found that the effect of misclassification error was likely

negligible on our results: assuming a genderizing error rate (proportion of males misclassified as females and females misclassified as males) of 5% and a gender bias error rate (difference between misclassification of males and females) of 2% (Santamaría and Mihaljevic 2018), our observed difference in scores would be 10% larger than the true difference in score, indicating our results are not false negative due to inaccuracy in the genderize algorithm. All scripts are available as supplemental files and on Github (https://github.com/bjarnebartlett/AltmetricAnalysis)

4.4 Results

The objective of this study was to examine if author gender explains (a) whether a study receives an AAS greater than zero, and (b) the magnitude of the score. No substantial trends in gender bias were found in any journal regarding whether an article received an AAS, the only exception being Science in 2012, which was associated with women first authors obtaining a score (log odds ratio = 0.63, 95% confidence interval [0.09,1.17]) (Figure SF6). We found these model results to be robust to the presence of outliers. We further found that there was no clear evidence of gender bias against women in the magnitude of AAS across six of the seven journals for 2011–2018 (Figure 11). While we identified potential emerging effects in favor of women appearing in 2018 for the first author in Nature (difference in log(AAS) = 0.28, 95% CI [− 0.1, 0.65]) and PNAS (difference in log(AAS) = 0.34, 95% CI [0.01, 0.68]), these effects did not carry high confidence. We conclude, based on our analysis, that in general, there is little current evidence for gender bias in AAS across major academic journals.

The one exception to these findings was for Science magazine—which showed biases against women in 2017 and 2018. Specifically, women scored lower than men for first authors, with a difference in log(AAS) of − 0.71 (95% CI [− 0.97, − 0.45]) and − 1.88 (95% CI [− 2.19, − 1.56]), respectively. While we did not undertake a separate analysis of the individual contributing components of the AAS, these mean differences for Science equate to roughly 22 more tweets or 3 more news mentions per article for men than women in 2017, and 88 more tweets or 11 more news mentions per article in 2018. We re-ran this analysis with our inference based on quantile estimates and found very similar results with respect to the mean differences, albeit with the median bias for first authors in Science in 2018 less prominent than the mean bias (equivalent to 5 more tweets for men than women in 2017, and to 20 more tweets or 2.5 more news articles in 2018) (Figure SF7).

The results we found for last authors were generally consistent with those for first authors (Figure S5B and S5C). Beyond these effects of gender of the first and last author, we found no impact of publication month or proportion of female authors on AAS (Figure SF9). As would be expected, the year of publication carried an important effect, with recent publications receiving higher scores (the mean 1-year score across all journals in 2011 was 10.5; in 2018, it was 42.1), as did having many authors, with scores increasing as the number of authors increased (each additional author leads to a 0.018 increase in log score for BioRxiv and a 0.005 increase in log score for all other journals) (Figure SF9).

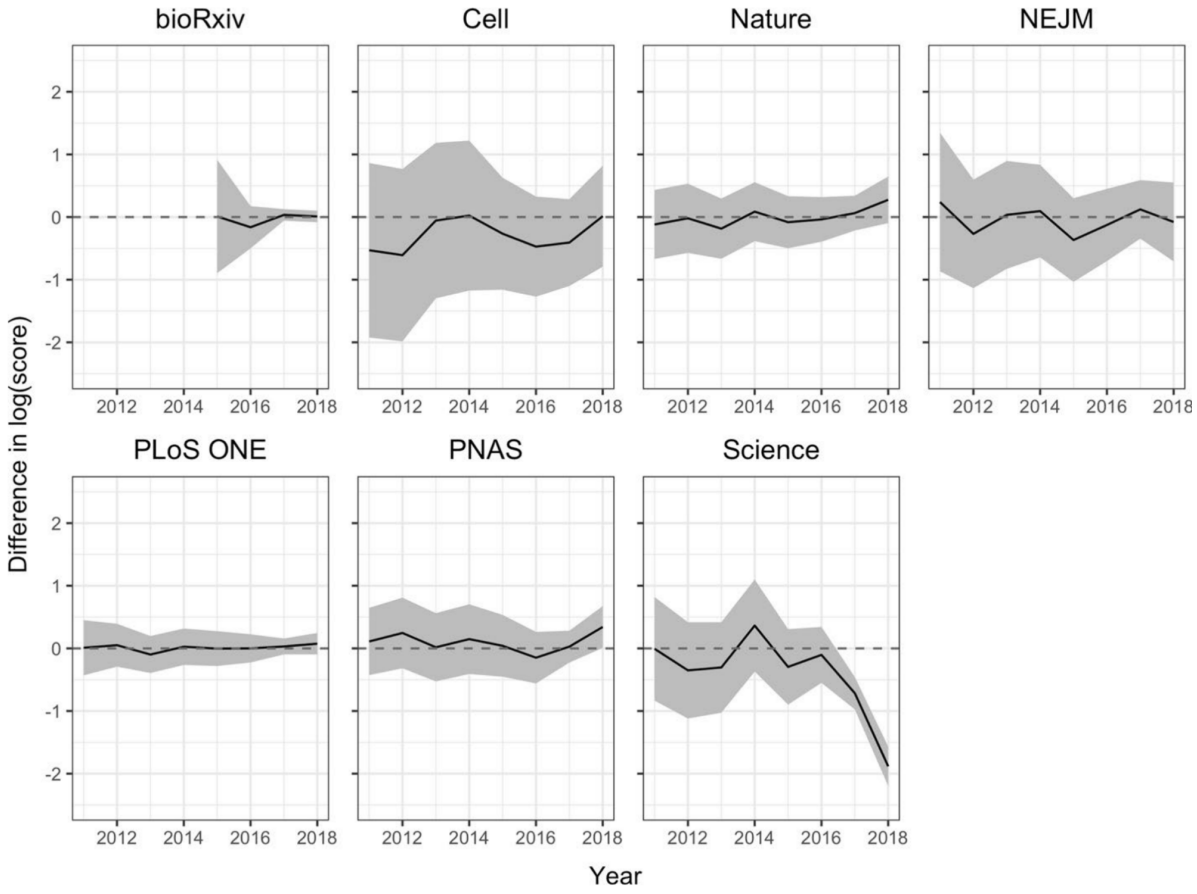


Figure 11: Gender bias in seven idiotypic journals for Altmetric Attention Scores (AAS) for first authors. The black line represents the mean difference in log(AAS) score between female and male first authors; where positive numbers represent a higher score for females and vice versa for males. The gray shading is the 95% confidence interval. With the exception of Science magazine which shows bias in favor of male authors in 2017 and 2018, all other journals show no clear evidence of bias.

4.5 Discussion

Our results indicate a promising shift in new digital metrics relative to traditional metrics. Alternative metrics, like those consolidated by altmetric.com, leverage digital technology to attempt to quantify scientific reach, and aggregating mentions from vast user bases may serve to democratize the evaluation of scholarly outputs. However, there are arguments on either side that this digitization of evaluation could favor either gender. For instance, men are quoted more

frequently in the news (Morris, 2016) and are more likely to self-promote on platforms like Twitter (Duggan et al., 2015; Mancuso et al., 2017). Yet some communication patterns favor women: working on questions of interest to the public (Milkman and Berger, 2014), attracting more student readers (Thelwall, 2018), having higher crowdfunding success rates (Terry et al., 2017). Our findings corroborate that the balance does not appear to be tipped in either direction; men and women authors of peer-reviewed articles get similar reach on digital platforms.

Gender plays a contributory role in professional advancement; moving toward less biased metrics is essential to create equitable workplaces. Traditional citation metrics, such as the h-index and i10-index, favor the male academic community (King et al. 2017). Other factors within the academy also favor the male population: differences in salary, space, awards, and resources have resulted in the marginalization of women faculty with women receiving less despite professional accomplishments equal to those of their male colleagues (MIT Committee on Women Faculty in the School of Science, 1999). Some communication patterns favor women, as discussed above, but many of these factors are most present inf digital spaces outside the academy and have little influence on tenure and promotion.

Citation metrics are particularly relevant to tenure and promotion in fields within science, technology, engineering, and math (STEM). Obstacles within STEM fields synergize to form an achievement disparity between men and women working in STEM, where less-qualified men are retained over more qualified women (Cimpian et al., 2020). Our analysis highlights that, while gender bias is rampant in academia, new digital metrics, carrying the balancing effect of a large, diverse group, have the potential to be democratizing and might provide less biased ways of assessing the impact of an academic article (Harambam et al., 2018). However, bias is not totally absent, as the worrying trend in Science shows. It is unclear where this bias arises from for

Science, but one possibility is that the media team at Science themselves are biased in their promotion of men over women authors. This and other possible causes should be investigated further.

There are some potential limitations to our study. First, our methods rely on being able to ascribe binary gender based on names, which is limited according to references in the genderize.io database. Second there may be additional heterogeneity that we did not capture. For instance, field, article topic or open access status may have an influence on AAS in a way that was not adequately encapsulated in our selection of idiotypic journals. Third, we recognize the importance of recognizing non-binary genders in STEM and that current methods for genderizing names do not reflect this. The third issue in particular warrants further attention and study as it implies that our approach does not answer the question of whether authors' gender identity influences AAS, but rather whether the likely binary perception of authors' gender by others influences AAS.

4.6 Conclusion

Our results suggest that while traditional metrics used by academic institutions exacerbate bias, new metrics, such as the AAS, may present a leveling of the playing field. Unlike the findings for most indices used to measure competency in academia, we found no clear gender bias in AAS in six of the seven journals. While the AAS is not a definitive measure of the quality of the research or the researcher, our results present the first quantitative assessment of whether bias exists in this new metric, highlight its value as a complementary index, and suggest that new technologies, such as those on which altmetrics is based, may indeed help to reduce the institutional biases reflected in traditional metrics used in academic evaluation.

CHAPTER 5: A DATA SCIENCE PRIMER TO ENGAGE UNDERGRADUATE STUDENTS IN RESEARCH

"Education is a social process; education is growth; education is not preparation for life but is life itself." ~John Dewey

Prologue

Curricula in higher education strive to impart skills to improve both the intellectual capacity and professional abilities of students. In the life sciences there are subdisciplines that draw from the same set of skills. However, frequently these subdisciplines use different names for the same skill or do not recognize the similarity in techniques. This can be due to narrow research silos or lack of rigorous examination of student learning objectives. To live up to the ideals of creating lifelong learning, as exemplified by the philosophy of John Dewey, it is essential that course development is ongoing, and that student learning is measured in meaningful ways for all the skills that instructors wish to impart.

There are many technical tools that are needed to interrogate big data, the goal of this introductory course was to create easy to understand modules that would provide students the skills necessary to work with these big data. To assess how well the original course student learning objectives functioned a comprehensive survey was conducted to rigorously identify success and failure of specific lessons.

Understanding what parts of the curriculum were successful coupled with the empirical data analysis of the previous chapters allowed for a rethinking of the most appropriate model for teaching introductory course on big data.

*In review Biochemistry and Molecular Biology Education
Bartlett, B, Kantar, MB, Stitt-Bergh, M, Bingham, JP. 2022. A Data Science Primer to Engage Undergraduate Students in Research. (In Review Biochemistry and Molecular Biology Education)

5.1 Abstract

An explosion of data available in the life sciences has shifted the discipline towards genomics and quantitative data science research. Institutions of higher learning have been addressing this shift by modifying undergraduate curriculums resulting in an increasing number of bioinformatics courses and research opportunities for undergraduates. The goal of this study was to explore how a newly designed introductory bioinformatics seminar could leverage the combination of in-class instruction and independent research to build the practical skill sets of undergraduate students beginning their careers in the life sciences. Participants were surveyed to assess learning perceptions towards the dual curriculum. Most students had a neutral or positive interest in these topics before the seminar and reported increased interest after the seminar. Students had increases in confidence level in their bioinformatic proficiency and understanding of ethical principles for data/genomic science. By combining undergraduate research with directed bioinformatics skills, classroom seminars facilitated a connection between student's life sciences knowledge and emerging research tools in computational biology.

5.2 Introduction

The data revolution in the life sciences has initiated a gradual paradigm shift from a descriptive science to a primarily quantitative discipline. To address this, academia has gradually developed programs in bioinformatics and quantitative methods for biology (Attwood et al., 2019). These programs began at the graduate level and have percolated into undergraduate curriculums (Hack et al., 2005). Efforts to increase the number of informatics/data science courses in undergraduate curricula often fail to emphasize general principles of data science or consider skills required to perform analysis of multi-format data (McClatchy et al., 2020). To be sure, articulating clearly defined general principles is imperative for students to choose the correct tools for analysis, assess the necessary computing resources, manage and clean data, and apply ethically approved standards in data collection, analysis, and storage (Attwood et al., 2019). Imparting fundamental disciplinary skills and integrating a solid theoretical background with more specialized knowledge creates a solid foundation for students in the life sciences.

Learning through traditional introductory molecular biology courses has historically been passive (Shah et al., 2013); however, it has been shown that activity-based learning increases student interest in the natural sciences (Freeman et al., 2014; Neyhart JL., 2020). An active learning approach can infuse a theoretical background of the significant concepts in molecular biology while providing skills that mirror knowledge creation within a particular domain (Table 5). Moreover, active learning of foundational biological concepts through inquiry-based labs with data-oriented exercises can facilitate peer-to-peer learning. Students learn from talking to their peers during group discussions of topical issues. Such conversations can increase student interest while achieving a broader understanding by amalgamating an instructor's explanation (Smith et al., 2011).

Table 5: *Selected Modes of Instruction vs. Active Modes of Instruction.* Examples from the seminar comparing classic (or traditional) modes of instruction with active modes of instruction with greater potential for student learning. Implementation of active learning techniques is also described.

Seminar Topic	Classic Mode of Instruction	Active Mode of Instruction	Seminar Implementation
Use R to enter and edit expressions and scripts.	*Students listen to a lecture on R expressions to manipulate data*	*Students learn by doing, crafting R expressions to manipulate different types of data*	*For their assignments, students started with code in pre-made R markdown documents*
Make figures and tables from data	*Students read an article and study the theory describing how to use a particular figure*	*Students compare software options to produce figures and practice using the software*	*Students worked in groups to design and make figures for their projects*
Gain a basic understanding of bioinformatic data	*Students read a textbook describing types of bioinformatic data and databases*	*Students compare types of bioinformatic data and databases and analyze the differences*	*Students engaged in online discussions of the ethical implications for storing biological data*
Know the general principles of designing a bioinformatic study	*Students listen to a lecture describing methods of bioinformatic study design*	*Students design a unique bioinformatic study*	*Students worked in groups to design studies around bioinformatic datasets*

One active method is through dynamic programming exercises for students, where students can combine molecular biology concepts with basic computational skills (Buitrago Flórez et al., 2017). As the size of data increases, it is critical that students have functional skills related to technology and computing to interact with and explore these data (Patacsil and Tablatin, 2017). Understanding next-generation biological tools and the full extent of biological data has become a highly computational exercise (Ideker and Nussinov, 2017). Several different computing languages are common in biology; these include R (R Core Team, 2021), Python (Van Rossum and Drake, 2009), and MatLab (MATLAB, 2020). In bioinformatics, the currently ascendant language is R (Fourment and Gillings, 2008). Many students do not have computing

expertise; thus, programming concepts must be integrated into the active learning curriculum simultaneously to introduce biological concepts (National Research Council, 2010).

Our study aimed to explore how a newly designed introductory bioinformatics seminar could leverage the combination of in-class instruction and independent research to build the practical research skill sets of undergraduate students beginning their research careers in the life sciences. The seminar design included inquiry-based labs for teaching data science together with foundational concepts in molecular biology. This new curriculum was introduced within the INBRE IV (Institutional Biomedical Research Excellence - NIH project 5P20GM103466-18) student undergraduate research program to understand core concepts while teaching independent research practices through an independent research project. The 10-week undergraduate seminar's core concepts were gleaned from data science projects that traversed data with varying characteristics (Fortin J., 2021; Bartlett B., 2021). We assessed the impact of the active learning activities (inquiry-based labs, data-oriented exercises, peer discussion, group projects) format that integrated data science and R markdown with molecular biology content. We hypothesized that this method of instruction would improve student perspectives of core genomic concepts. Student perceptions about the domain knowledge they are learning impact their approach to the material and how they learn it (Hammer, D., 1994). We tested this by surveying the student participants at the end of the seminar to understand their knowledge, perceptions of their knowledge, and the relevance of this knowledge to their research. The goal is to enhance student research outcomes by offering multidisciplinary training in genomics to complement student independent research projects.

5.3 Materials and Methods

Student Background

Students in the program were part of the INBRE program 'housed' at the University of Hawai'i. The student cohort spanned two University of Hawai'i campuses and 2-year (UH Community Colleges) and 4-year institutions (Chaminade and Hawai'i Pacific University), characterized by a collection of students possessing disparate biomedical research backgrounds (Table 6). 26 Students participated in the seminar, 12 responded to the anonymous survey. Students represented 12 different majors, reflecting the diverse academic interests of students enrolled in a biomedical research training program such as INBRE. Thus, the large cross-section of majors requires an all-encompassing general curriculum to globally address the skills needed for students in this unique INBRE research training and education program.

Table 6: *Student Demographics.* Overview of the demographics of 26 total student seminar participants.

Type of Institution	
2-Year	*3*
4-Year	*22*
Unknown	*1*
Major	
Applied Science - Health Professions	*4*
Biochemistry & Molecular Biology	*1*
Bioengineering	*1*
Biology	*9*
Chemistry	*2*
Health/Exercise Science & Lifestyle Management	*2*
Marine Biology & Biomedical Engineering	*1*
Microbiology	*1*
Molecular Biosciences & Biotechnology	*1*
Molecular Cell Biology	*1*
Natural Science	*1*
Non-degree post-baccalaureate	*1*
Unknown	*1*
Class Standing	
Sophomore	*1*
Junior	*4*
Senior	*14*
Post-baccalaureate	*2*
Unknown	*5*

Seminar Description

Bioinformatics seminars at The University of Hawai‘i introduce core concepts early in the college curriculum, i.e., in the first and second years of college. Our INBRE (IDeA Networks of Biomedical Research Excellence) program consisted of cross-disciplinary seminars designed to give undergraduate students across the University of Hawai‘i system and INBRE partner institutes (Chaminade and Hawai‘i Pacific University) an introduction to bioinformatics concepts and resources during their college education. Students derive skills in data science from empirical research. Activities include introducing students to biological databases, R, and phylogenetic analysis. The mode of seminar instruction is entirely online. Students' comprehension and mastery of bioinformatics are assessed through online lessons, labs, quizzes, and inquiry-based group activities. The seminar's active learning components are designed around embedded interactions with bioinformatics databases during lecture time culminating in a final group project. Group size ranged from 1 - 13 students. Student Learning Objectives (SLOs) included the following: utilization of R to enter and edit expressions and scripts; read, subset, and reshape tabular data; find and install external R packages; make figures and tables from data; gain a fundamental understanding of bioinformatic data; and learn the general principles of designing a bioinformatic study (Syllabus, Supplementary Document SD1).

The following criteria were considered in seminar design: operating system, programming languages, technology requirements, and student demographics. Seminar activities incorporated data retrieval, data cleaning, and data processing. Online labs were developed to provide practical experience in the data science process through bioinformatics. Students identified relevant topics and designed scripts to execute the appropriate strategies, including an inquiry-based final project. R may impose a learning curve that is too advanced for

demonstrations in an introductory bioinformatics seminar. To overcome this, our team developed R modules that were made available using R markdown so that students could complete them with minimal knowledge of R programming. Seminar labs focused on several core concepts: working with data, making inferences, and prediction modeling. A final project spanned the entire seminar. The project began with a biological rationale and subsequently utilized MEGA (Kumar et al., 2018), a multifaceted bioinformatics software for phylogenetics, to generate data about the chosen hypothesis.

Survey Design

At the end of the seminar, students completed a survey that included a set of questions assessing their interest level and knowledge before and post-seminar, and questions on knowledge gained, level of confidence in subject ability, the relevance of the seminar's topics/techniques to their (planned) research, and seminar elements (e.g., group project). The survey was covered under the University of Hawai'i IRB Protocol #2020-00940. A retrospective pre/post survey was selected because students entered the seminar with little knowledge about the topics and thus could not accurately answer questions on a pre-seminar study—the questions aimed to understand the amassing and mastery of bioinformatics skills in undergraduate students during the seminar. Specifically, the survey addressed bioinformatic tools and ethical concepts in data science, including phylogenetics, NCBI databases, and genomic privacy. Each closed-ended question utilized a 5-point Likert scale.

Survey Analysis

Survey results were analyzed using a scoring system as follows: very disinterested/not at all useful/not confident/strongly disagree, -2; disinterested/somewhat useful/somewhat

confident/disagree, -1; do not know/no opinion/neutral, 0; interested/moderately useful/confident/agree, +1; very interested/very useful/extremely confident/strongly agree, +2. A Standard error was calculated for the percent of students responding with a particular answer (Tables 7-12). A Pearson correlation was used to examine the relationships between self-reported student confidence, interest, and utility. Effect size was estimated using a one-sample Wilcoxin signed-rank test. The percent change was used to compare student responses for before the seminar to after the seminar (Figure 13).

5.4 Results

Student Interest and Confidence

The seminar covered a variety of topics spanning genetic data analysis, programming, and data science. Aggregate student interest increased across all topics covered during instruction, except for NCBI tools (Table 7). The most significant gain in student interest was in sequence alignment (Wilcoxon effect size estimate, 0.890.01), with 4 students self-reporting increased interest in the topic during the seminar (Table 7). Learning to use NCBI tools also showed an overall increase, this topic also had the highest proportion of interested students before the seminar (Wilcoxon effect size estimate, 0.90.03). The pattern regarding student confidence was less clear; more common terms and concepts to biology such as sequence alignment and NCBI tools seemed to garner higher confidence among students as compared to more specialized downstream analytical processes like multiple sequence alignment and phylogenetic analysis, however there was no statistical difference between these 2 groups (Tables 8).

Table 7, Question 1: Please rate your level of interest in the following subjects in biological sequence data before you took this seminar and now, after the seminar.

	Interest Rating Before/After Seminar	**Very Disinterested**	**Disinterested**	**No Opinion/Neutral**	**Interested**	**Very Interested**
Types of Biological Sequence Data	Before	8 ± 8%	17 ± 8%	42 ± 8%	33 ± 8%	0 ± 8%
	After	0 ± 9%	25 ± 9%	33 ± 9%	42 ± 9%	0 ± 9%
Database Structures	Before	8 ± 7%	25 ± 7%	42 ± 7%	25 ± 7%	0 ± 7%
	After	0 ± 9%	33 ± 9%	25 ± 9%	42 ± 9%	0 ± 9%
Sequence Alignment	Before	8 ± 9%	8 ± 9%	50 ± 9%	33 ± 9%	0 ± 9%
	After	0 ± 12%	17 ± 12%	17 ± 12%	67 ± 12%	0 ± 12%
Phylogenetics of Organisms	Before	8 ± 9%	17 ± 9%	25 ± 9%	50 ± 9%	0 ± 9%
	After	0 ± 11%	17 ± 11%	25 ± 11%	58 ± 11%	0 ± 11%
Sequence Homology	Before	8 ± 8%	17 ± 8%	33 ± 8%	42 ± 8%	0 ± 8%
	After	0 ± 10%	17 ± 10%	33 ± 10%	50 ± 10%	0 ± 10%
NCBI Tools	Before	8 ± 10%	8 ± 10%	17 ± 10%	58 ± 10%	8 ± 10%
	After	0 ± 9%	8 ± 9%	25 ± 9%	50 ± 9%	17 ± 9%

Table 8, Question 2: As a result of this seminar, what is your confidence level in your knowledge of the following subjects in bioinformatics.

	Not Confident	**Somewhat Confident**	**Confident**	**Extremely Confident**	**Don't Know**
NCBI Databases (ex. Rentrez R package)	17 ± 8%	42 ± 8%	33 ± 8%	0 ± 8%	8 ± 8%
Biological Sequence Alignment (ex. BLAST, Biostrings)	17 ± 8%	33 ± 8%	42 ± 8%	8 ± 8%	0 ± 8%
Multiple Sequence Alignment (ex. MEGA, MUSCLE, CLUSTAL)	25 ± 9%	50 ± 9%	25 ± 9%	0 ± 9%	0 ± 9%
Phylogenetic Analysis (ex. MEGA, CLUSTAL, MAFFT)	33 ± 7%	33 ± 7%	25 ± 7%	8 ± 7%	0 ± 7%

Perceived Utility by Students for Research

The seminar took place during the backdrop of the SARS-CoV-2 pandemic, with restrictions on in-person learning, coupled with an online format that was new to the INBRE program. There was a statistically significant correlation between students' interest and perceived usefulness (R2=0.57, P=0.05), suggesting the importance of practical applications for lab exercises. This relationship was not significant for student confidence (R2=0.80, P=0.16) (Table 9, Figure 12).

Table 9, Question 3: How useful/not useful are the following regarding the INBRE research you are conducting/planning to conduct.

	Not Useful	**Somewhat Useful**	**Useful**	**Extremely Useful**	**Don't Know**
[NCBI Databases (ex. Rentrez R package)]	33 ± 4%	17 ± 4%	17 ± 4%	25 ± 4%	8 ± 4%
[Biological Sequence Alignment (ex. BLAST, Biostrings)]	25 ± 7%	8 ± 7%	33 ± 7%	33 ± 7%	0 ± 7%
[Multiple Sequence Alignment (ex. MEGA, MUSCLE, CLUSTAL)]	33 ± 6%	17 ± 6%	33 ± 6%	17 ± 6%	0 ± 6%
[Phylogenetic Analysis (ex. MEGA, CLUSTAL, MAFFT)]	33 ± 6%	17 ± 6%	33 ± 6%	17 ± 6%	0 ± 6%

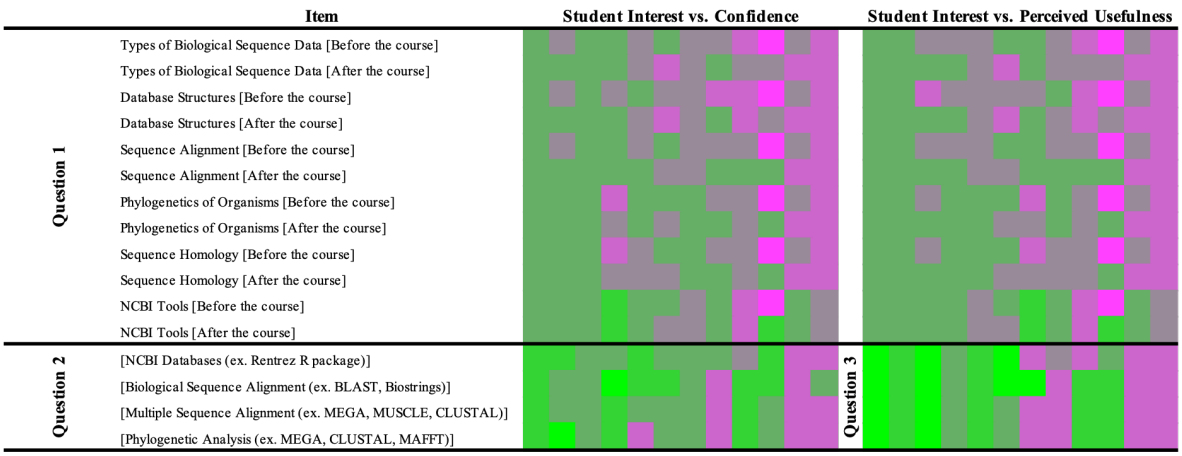


Figure 12: *Heatmaps Comparing Student Interest vs. Confidence vs. Perceived Usefulness.* Self-reported usefulness and confidence of different seminar topics for 12 student survey respondents compared to self-reported interest. Survey responses were scored as follows: very disinterested/not at all useful/not at all confident, -2, light orange; disinterested/somewhat useful/somewhat confident, -1, orange; do not know/no opinion, 0, brown; interested/moderately useful/confident, +1, medium green; very interested/very useful/very confident, +2, dark green. A statistically significant correlation was observed (R = 0.57, p=0.05).

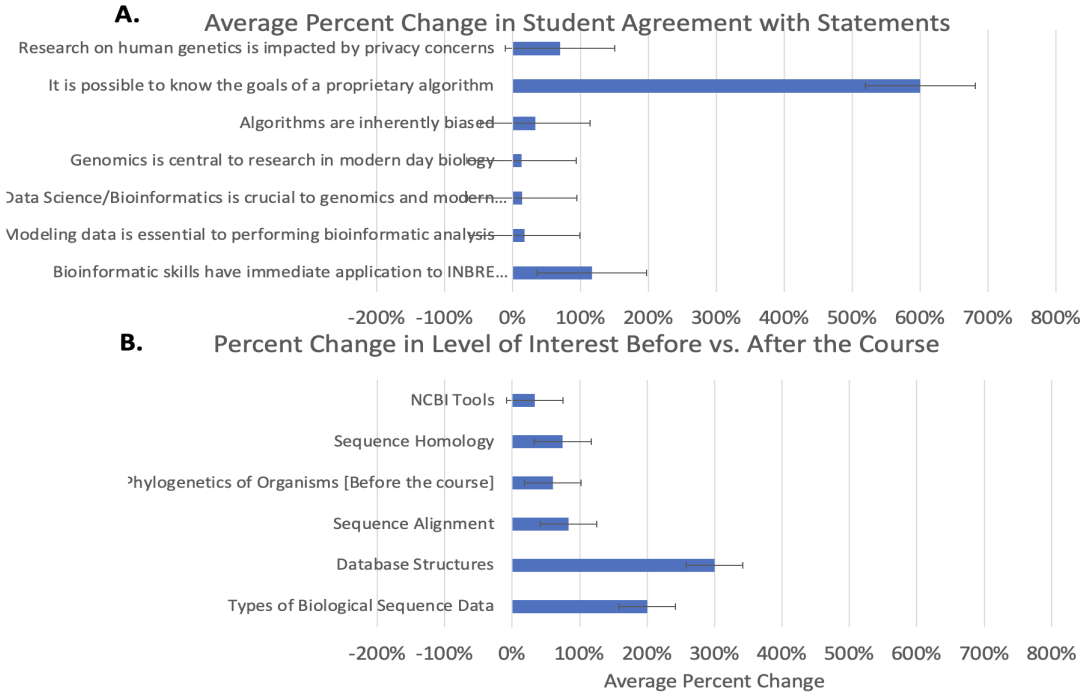


Figure 13: Average Percent Change in (A) Self-Reported Agreement by Students to Statements about Seminar Material and (B) Self-Reported Level of Interest in Computational Biology Topics. A. Self-reported agreement with statements related to seminar material. Some statements were explicitly addressed in the seminar, while some statements were topically related but not addressed. B. Self-reported interest in computational biology topics addressed in the seminar.

Self-reported Understanding of Core Concepts

Students self-reported a greater understanding of bioinformatics, and the majority indicated they would like to learn more about the discipline after taking the seminar (Table 10). Students also indicated that the frequent practice of exacting computational techniques was helpful to them in improving their understanding. Finally, most students said that both exercises using R markdown documents and group projects were helpful in their learning of bioinformatics (Table 11).

Table 10, Question 4: Please indicate your level of agreement with the following statements before you took this seminar and now, after the seminar.

	Before/After the Seminar	Strongly Disagree	Disagree	Agree	Strongly Agree	Don't Know
I have a good understanding of bioinformatics and the associated R tools.	Before	50 ± 11%	42 ± 11%	8 ± 11%	0 ± 11%	0 ± 11%
	After	8 ± 8%	17 ± 8%	50 ± 8%	8 ± 8%	17 ± 8%
I would like to learn more about biological databases, R, and other statistical software.	Before	0 ± 13%	67 ± 13%	33 ± 13%	0 ± 13%	0 ± 13%
	After	0 ± 12%	17 ± 12%	67 ± 12%	8 ± 12%	8 ± 12%

Table 11, Question 5: In thinking about this seminar, indicate your level of agreement with the following statements.

	Strongly Disagree	**Disagree**	**Agree**	**Strongly Agree**
The R tools provided in the homework have enhanced my learning as an INBRE student.	8 ± 15%	25 ± 15%	67 ± 15%	0 ± 15%
The R homework assignments were easy to use.	0 ± 15%	42 ± 15%	58 ± 15%	0 ± 15%
The group projects were a useful addition to the course content	17 ± 14%	17 ± 14%	67 ± 14%	0 ± 14%

Student Understanding of General Data Science Concepts

To measure the impact of student discussions, students were asked to rate their level of agreement with statements related the discussion topics – as a comparison, students were also asked to rate their level of agreement with statements not discussed, but in a similar discipline. After completing the seminar, most of the students agreed with factual statements about the science concepts explicitly addressed in the curriculum (marked with an asterisk in Table 12). For example, students reported a better understanding of privacy concerns in genomics after completing group discussions on this topic. Algorithmic bias, a subject not explicitly discussed, was not understood better (Table 12).

Table 12, Question 6: Please indicate your level of agreement with the following statements before you took this seminar and now, after the seminar.

	Before/After the Seminar	Strongly Disagree	Disagree	Agree	Strongly Agree	Don't Know
Bioinformatic skills have immediate application to INBRE research projects.*	Before	0 ± 14%	17 ± 14%	67 ± 14%	0 ± 14%	17 ± 14%
	After	0 ± 18%	0 ± 18%	92 ± 18%	8 ± 18%	0 ± 18%
Modeling data is essential to performing bioinformatic analysis.	Before	0 ± 18%	0 ± 18%	92 ± 18%	0 ± 18%	8 ± 18%
	After	0 ± 18%	0 ± 18%	92 ± 18%	8 ± 18%	0 ± 18%
Data Science/Bioinformatics is crucial to genomics and modern-day biology.*	Before	0 ± 13%	0 ± 13%	67 ± 13%	25 ± 13%	8 ± 13%
	After	0 ± 13%	0 ± 13%	67 ± 13%	33 ± 13%	0 ± 13%
Genomics is central to research in modern day biology.*	Before	0 ± 11%	0 ± 11%	58 ± 11%	33 ± 11%	8 ± 11%
	After	0 ± 13%	0 ± 13%	58 ± 13%	42 ± 13%	0 ± 13%
Algorithms are inherently biased.	Before	0 ± 12%	8 ± 12%	17 ± 12%	8 ± 12%	67 ± 12%
	After	0 ± 8%	17 ± 8%	33 ± 8%	8 ± 8%	42 ± 8%
It is possible to know the goals of a proprietary algorithm.	Before	0 ± 11%	17 ± 11%	25 ± 11%	0 ± 11%	58 ± 11%
	After	0 ± 13%	8 ± 13%	67 ± 13%	0 ± 13%	25 ± 13%
Research on human genetics is impacted by privacy concerns.*	Before	0 ± 10%	0 ± 10%	50 ± 10%	17 ± 10%	33 ± 10%
	After	0 ± 13%	0 ± 13%	58 ± 13%	42 ± 13%	0 ± 13%

*These topics were explicitly addressed in the seminar.

5.5 Discussion

The results from this small study are encouraging they suggest that in-class instruction, active learning, and connections to students (planned) research did build the intended skills and interest. While not every aspect of every lesson contained active learning, many SLOs were addressed with active learning components – an instructional style shown to increase student learning (Brown et al. 2014). In addition, the seminar structure allowed students to explore varied disciplinary techniques and identify where they thought they had understanding, deficiencies, or mastery. The seminar's inclusion of critical cross-disciplinary skills such as computer science and data ethics helps students gain a more accurate perception of life sciences and their training as scientists.

Well-designed courses incorporate a flexible, goal-oriented approach with a first step in course design focused on identifying desired results (Wiggins and McTighe, 2005). Once the end goals of teaching are clear, the designer can determine acceptable evidence and plan learning experiences and instruction. In practice, we measured our end goal of enhancing student research outcomes through multidisciplinary training in genomics by means of group projects and a final pre/post survey. When running a decentralized online seminar that is cross-disciplinary, the limitations concerning the completion of student activities can provide an opportunity to assess student performance through authentic activities such as these students' group projects (Sambell et al. 2012). In a research-focused program, such as INBRE, it is possible that students became more interested in topics as they were perceived to be relevant to their laboratory work. This evolution in interest can be attributed to understanding new methods of how to ask/answer biological questions.

Most students self-reported an increased understanding across all topics in genomics covered in the post-course survey (Table 4). Students reported the greatest increase in agreement with statements covered in the seminar including the privacy concerns impacting human genetics, and whether bioinformatic skills could be applied to their INBRE research projects. Students were most engaged when the seminar combined theory and practice in biology, tying their research to the lessons in bioinformatics and the ethical implications of those technologies—underscoring a need to design computational biology courses around current biological concepts and the ethical implications of these concepts (National Academies of Sciences, 2020). The essential combination of theory and practice could explain why student interest was correlated with perceived usefulness, while not associated with increased student confidence (Figure 1). Engaging in peer discussions was a central component in improving student understanding of ethical concepts—students reported a better understanding of privacy concerns in genomics after completing group discussions on this topic. Incorporating ethics into college curricula has been demonstrated to be essential to student engagement and an ability to connect learning in the classroom to a broader context (Whitley et al. 2020).

While this seminar's format was largely successful, we did receive student feedback that could help future cohorts. First, because coding was new to many students, students noted that introducing R sacrificed valuable class time that could be used to explore a single tool, in-depth without requiring R skills. In bioinformatic analysis, selecting the correct tool is critical (Welch, L. 2014). In future iterations of the seminar, we hope to place additional emphasis on selecting between tools and programs that perform similar tasks. Previous studies have found self-taught R skills to be inadequate, therefore we plan to continue emphasizing introductory R skills (Eglen, 2009). Second, students asked for group size to be limited for the final project.

Crossdisciplinarity group work contributes to success in computational biology (Aikens & Dolan, 2014). We hope to incorporate this feedback and continue group projects but experiment with smaller groups. Third, students came into the program with different expectations for conducting in silico research projects through INBRE. Some expected to augment their laboratory experience with such approaches, and others expected to work solely in a wet lab (Supplementary Document SD2).

In addition to student feedback, we observed that additional effort is needed to reach the target student demographic. Though we sought to reach freshman and sophomore students, 20 of 26 students were upperclassmen (junior and above). The seminar was advertised equally to all levels of students at participating institutions. It is possible that freshman/sophomore students are still exploring their career aspirations and less interested in skills-based preparation for biomedical research. These challenges can be managed by identifying the interests of first and second-year college students and through diligent communication about the expectations of the INBRE program.

5.6 Conclusion

Utilizing modern tools, data science requires that theory be interwoven tightly with concrete skill development. In this study, students improved understanding of biology and genomics through practical exercises and discussion questions. Students were instructed on using multiple tools during lectures for each biological concept to ascertain when a specific technical solution is needed. Future iterations of our seminar series will see some changes—firstly, the groups for our seminar were too large and varied in size (from 1 to 13 students). Groups were not assigned. Students with a similar interest formed a group, regardless of how many students were

in each group. This system did not give each student adequate exposure to computer skills and software, which groups of 3-4 would have done. Though they were engaged in the discussion, students expressed little disagreement. Solutions for future iterations of the seminar include having the student groups formed based on which discussion questions they choose and providing instruction on having productive debates. A secondary benefit would be grouping students with differing levels of computer skills. Given the size of the seminar and the number of survey respondents (n=12), additional iterations are needed to understand the complex interplay between the seminar's concepts (e.g., the ethics of technology by genetics interaction). Further iterations are essential to understanding any links between demographic factors and levels of interest, the correlation between different sub-subjects, and the relationship between areas of interest with computer skills. Having a better understanding of these complex interactions will allow for the continued improvement of future seminars and tailored attention to students' needs, resulting in better student retention and job preparedness.

CHAPTER 6: CONCLUSIONS AND FUTURE DIRECTION

6.1 The Role of Data in Higher Education

Including critical, cross-disciplinary skills (e.g., cancer biology, plant genomics, bibliometrics) and data ethics help students build a more accurate perception of life sciences and improve scientists' training. The first step in the course design is identifying desired results (Wiggins and McTighe, 2005); in my dissertation, I identified critical components for biological data science by conducting three empirical big data studies. Teaching a course based on these components proved to be an excellent opportunity to apply a broad skill base in data science derived from the three empirical studies. While I was pleased with the outcome and responses from student-participants in the seminar course, I saw a significant need to revise the SLOs for future seminar iterations based on course assessments. This led to the development of a new data science model and new revised SLOs. When taken in a broader context, analysis of student feedback and course participation led me to identify broad data science skills that could be taught within customary course frameworks to improve student engagement and learning outcomes (Figure 14). These skills broadly can be classified as 'Technical Knowledge,' 'Quantitative Interpretation,' and 'Ethical Knowledge.' Good SLOs should have specific, measurable outcomes; however, the original iteration of the course used SLOs that were ambiguous and did not provide enough measures or benchmarks for a sufficient measure of learning. The newly developed SLOs could allow any future instructor to measure how successfully students gain competence in the distilled data science principles. In the following section, I define the core concepts and the specific SLOs derived for each.

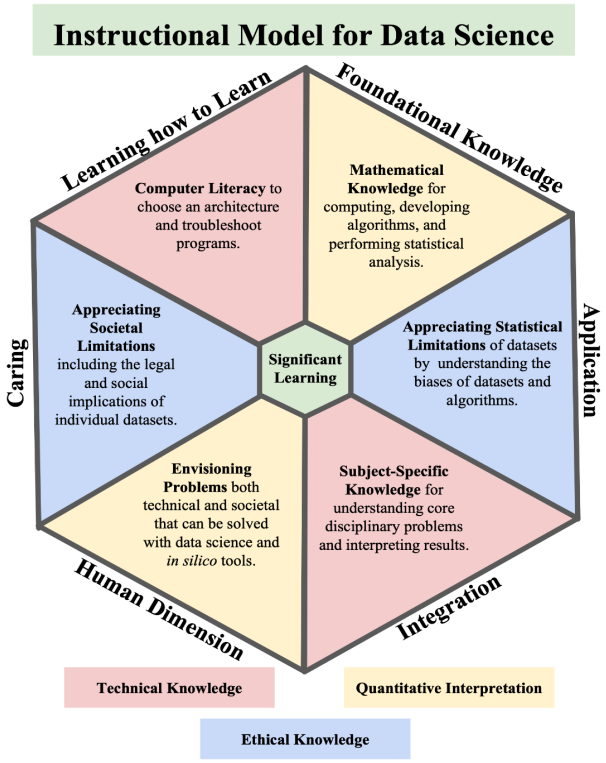


Figure 14: A Proposed Instructional Model for Data Science. This model is adapted from the model for significant learning (Fink, 2013).

Technical Knowledge

Technical knowledge refers to the computer literacy and subject-specific knowledge necessary to derive robust answers to *in silico* research questions through experimentation and analysis. I used three empirical study systems to derive specific learning outcomes on technical skills. Research has found that relying on self-taught R skills when introducing students to computational biology is inadequate (Eglen, 2009). I sought to incorporate specific skills into the curriculum, leading to revised SLOs based on both computer literacy and subject-specific knowledge:

Computer Literacy

SLO: Recognize computing principles of R to apply the appropriate methods for editing expressions and scripts; read, subset, and reshape tabular data.

SLO: Demonstrate R knowledge through installing and using external R packages.

Subject-Specific Knowledge

SLO: Compare and contrast similar R packages to examine which packages apply to a particular dataset.

SLO: Appraise a particular dataset's quality and collection methods and defend its inclusion in the study of a broader research question.

Quantitative Interpretation Knowledge

Quantitative interpretation refers to the discipline/problem-specific knowledge and mathematical skills required for developing algorithms, performing statistical analysis, and envisioning the vital link between a particular dataset and the problems to which that data can be applied. Quantitative interpretation skills are necessary to troubleshoot software across many research areas and ultimately choose appropriate software options for the dataset.

Mathematical Knowledge

SLO: State foundational knowledge of mathematics and statistics and recognize strategies to reshape and subset tabular data.

SLO: Apply foundational mathematical knowledge to check the quality and integrity of a dataset and develop basic figures and tables to do this.

Envisioning Problems

SLO: Formulate a research question that addresses a particular disciplinary problem and develop an experiment to answer the question.

SLO: Recognize problems in a particular area and select specific quantitative questions that can be investigated through research to address these problems.

Ethical Knowledge

Appreciating the societal limitations of data science, including the legal, social, and statistical implications of individual datasets, helps students combine theory and practice in the classroom by engaging in reflection and discussions on ethical concepts around research.

Appreciating Statistical Limitations

SLO: Use quantitative methods to examine a particular dataset for biases to defend its use.

SLO: Differentiate between datasets that inform with new knowledge and those that exacerbate structural biases.

Appreciating Societal Limitations

SLO: State the legal implications of a particular dataset and identify how these implications extend to society.

SLO: Recognize any social impacts of a dataset and formulate a plan to address these impacts in an equitable way.

6.2 Cancer Genomics

In chapter 2 the process of cancer biomarker discovery required choosing appropriate packages for analyzing gene expression data and developing custom code for visualizing the gene expression data in the R environment. This analysis and accompanying visualization allowed for identifying promising biomarkers. Chapter 2 necessitated technical knowledge to address the poor performance of relatively new algorithms for immune cell deconvolution. While not a perfect solution, I attempted to address these algorithms' deficiencies by combining three different algorithms' predictions. Chapter 2 utilized human genomic data; this engenders unique privacy concerns. This data used genomic information from a publicly available dataset (The Cancer Genome Atlas); despite being public, a request and approval are required to access the data. These privacy concerns were directly used in the course, where students were asked about

privacy concerns arising from public genomic databases–and specifically asked to discuss the impact of short tandem repeats (STRs) on the Y chromosome and the availability of this information in public genealogy databases as these data were previously used to derive surnames from deidentified public genomes and correctly identify the individual when combined with age and state of residency (Gymrek, 2013).

CURRICULAR HYPOTHESIS

The unique properties of patient tumor genomic data (e.g., privacy, large numbers of individuals, small amount of sample) will provide critical insights into the ethical, technical, and reasoning aspects of data science.

IMPLICATIONS FOR DATA SCIENCE INSTRUCTION

The key technical take-homes from this experiment were that RNA sequencing and immunology data could be synthesized and taken together. When multiple algorithms are available, the limitations of any single algorithm can be reduced by considering the results together. The critical quantitative reasoning take-homes were the limitations of the individual algorithms used for immune cell deconvolution and the overall potential of early detection in cancer as a viable way to manage the disease. This addressed both the technical and ethical curricular hypotheses including ethical considerations that a human-subjects dataset could have known privacy concerns depending on the source.

6.3 Plant Genomics

Chapter 3 was conducted in a non-model system. The standard tools that are used to quantitatively look at hybridization required modifications. For example, the tool SppIDer, a software package developed to detect hybrids was used, however, this tool was validated using species with small genomes (e.g., *Saccharomyces*), to have it function on large genomes the software had to be modified in a way that required both subject-specific knowledge of the macadamia genome and technical understanding of how SppIDer was processing the genomic information and how this was impacted by genome size. In chapter 3, large amounts of data had to be integrated between different libraries that combined code in two different languages, R and

Python. From a quantitative interpretation standpoint, selecting the appropriate tools was a central task in assessing the hybridization, however, even after selecting appropriate tools, conflicting results were in one case observed. This conflict was resolved by exploring the tool output and historical data on the samples collected. Further, the often-controversial intellectual property status of plants was considered due to the high value of macadamia, specifically the delicate balance between the rights of the community and the rights of the inventor.

CURRICULAR HYPOTHESIS

The properties of non-model species (e.g., crop wild relative) genomic data will provide critical insights into the ethical, technical, and reasoning aspects of data science.

IMPLICATIONS FOR DATA SCIENCE INSTRUCTION

The critical technical curricular considerations from this experiment were the importance of the quality of genomic data to downstream analysis, the challenges of de novo *transcriptome assembly, and the different architecture of tools used for smaller genomes compared to larger genomes. The critical quantitative interpretation take home was that two tools for hybridity can give different signals, addressing the curricular hypothesis with respect to the unique challenges of developing genomic resources in non-model species. There were no critical insights regarding ethical aspects of data science.*

6.4 Bibliometrics

The implications of the data used in Chapter 4 applied most to ethics, as it used data science to directly assess diversity, equity, and inclusion. Chapter 4 used R to analyze altmetric data sourced from the company with the same name. The major technical issue associated with this study was partitioning the data to a useable size. Tools, such as GenderizeR, help filter information to a smaller dataset–that can be used to ask specific questions, which can be answered using complex statistical analysis. While the GenderizeR tool was useful, it required validation, which was done with spot-checking the results to independently validate the software's database. The unique problem of searching names in an author list was compounded

by the proprietary algorithm that generated the response variable (AAS), this brought up more ethical issues that required careful interpretation.

Curricular Hypothesis

The structure of bibliometric data will provide critical insights into the ethical, technical, and reasoning aspects of data science.

Implications for Data Science Instruction

The critical quantitative interpretation take home from this study system was using statistics to analyze a dataset for bias. The critical ethical curricular considerations from this study were that sources of bias are often unknown and unconscious, including that the goals of proprietary algorithms are unknown to the user—the presence of bias necessitates the need to both collect data and check results for both bias and errors. There were no critical insights into the technical aspects of data science.

Implications for generating attention around scientific discoveries

The results from this study underscore the need to look closely at the metrics that drive how science is conducted and how the public understands discoveries. This matters to scientists, funders, and those evaluating the products of science. There may not be a clear right or wrong answer, as different institutes value different outcomes from their researchers. Still, if these results hold over time, they do suggest that a broader, more holistic view of research and its impacts is worthwhile. The current citation metrics have been proven to have significant, measurable bias in multiple ways – new metrics relying on public attention might prove more complex to manipulate because they aggregate information from a larger number of sources and demographics. This being said, the new metrics can create artificial incentives to work on large consortium-based projects or on controversial scientific topics that will garner public attention. As with any algorithmic approach that proposes objectivity, there is the potential for both unconscious bias and deliberate manipulation of results by gaming the algorithm.

6.5 Future Directions

Curricular Hypothesis

Student Learning objectives based on multidisciplinary synthesis will result in positive student outcomes.

This dissertation explored a multidisciplinary approach to understanding data and how data science methods can be used to understand broad questions and help students gain a dynamic understanding of biology. From each of the empirical studies, there were clear

implications for how to create significant learning for students – the relationship between multidisciplinary synthesis and student outcomes lent support to the curricular hypothesis. This led to the development of cross-disciplinary principles that could be implemented in any data science curriculum that looked at the intersection of the broad areas of 'Technical Knowledge', 'Quantitative Interpretation', and 'Ethical Knowledge.' These areas were further developed and expanded to encompass specific learning goals that became part of a distilled model that expanded on the six categories from the original Fink model, providing clear approaches and goals for an instructor creating an introductory course for biology students (Figure 14).

College curriculums outside of education departments often do not include formal training in teaching. University instructors, who received their training in these programs, will gravitate toward the models they were taught under by emulating their own teachers: a so-called "learning by proxy" model. While a valuable tool, "learning by proxy" is not rooted in the latest ideas or literature. For advances in pedagogy to be translated to the classroom, it will be imperative that teachers be encouraged and helped to use newer models as curricula are developed and evolved. Achieving the learning outcomes that will help students intellectually and technically will require both continued innovation and the implementation of these new ideas by teachers.

REFERENCES

Aikens, Melissa L., and Erin L. Dolan. 2014. "Teaching Quantitative Biology: Goals, Assessments, and Resources." *Molecular Biology of the Cell* 25 (22): 3478–81.

Analysis-Ready Standardized TCGA Data from Broad GDAC Firehose. 2016. Broad Institute TCGA Genome Data Analysis Center. Broad Institute of MIT and Harvard. https://doi.org/10.7908/C11G0KM9.

Anderson, Lorin W., and David R. Krathwohl. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman,.

Aran, Dvir, Zicheng Hu, and Atul J. Butte. 2017. "XCell: Digitally Portraying the Tissue Cellular Heterogeneity Landscape." *Genome Biology* 18 (1): 1–14.

Arias, Paola, Nicolas Bellouin, Erika Coppola, Richard Jones, Gerhard Krinner, Jochem Marotzke, Vaishali Naik, Matthew Palmer, G.-K. Plattner, and Joeri Rogelj. 2021. "Climate Change 2021: The Physical Science Basis. Contribution of Working Group14 I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change; Technical Summary."

Atkinson, M. D., P. S. Kettlewell, P. R. Poulton, and P. D. Hollins. 2008. "Grain Quality in the Broadbalk Wheat Experiment and the Winter North Atlantic Oscillation." *The Journal of Agricultural Science* 146 (5): 541–49.

Attwood, Teresa K., Sarah Blackford, Michelle D. Brazas, Angela Davies, and Maria Victoria Schneider. 2019. "A Global Perspective on Evolving Bioinformatics and Data Science Training Needs." *Briefings in Bioinformatics* 20 (2): 398–404.

Bakshy, Eytan, Solomon Messing, and Lada A. Adamic. 2015. "Exposure to Ideologically Diverse News and Opinion on Facebook." *Science* 348 (6239): 1130–32.

Bartlett, Bjarne, Zitong Gao, Monique Schukking, Mark Menor, Vedbar S. Khadka, Muller Fabbri, Peiwen Fei, and Youping Deng. 2021. "The MiRNA Profile of Inflammatory Colorectal Tumors Identify TGF-β as a Companion Target for Checkpoint Blockade Immunotherapy." *Frontiers in Cell and Developmental Biology* 9.

Batlle, Eduard, and Joan Massagué. 2019. "Transforming Growth Factor-β Signaling in Immunity and Cancer." *Immunity* 50 (4): 924–40.

Bloom, Benjamin Samuel. 1956. "Taxonomy of Educational Objectives: The Classification of Educational Goals." *Cognitive Domain*.

Bornmann, Lutz, and Robin Haunschild. 2018. "Do Altmetrics Correlate with the Quality of Papers? A Large-Scale Empirical Study Based on F1000Prime Data." *PloS One* 13 (5): e0197133.

Bratt, Sarah, Jeff Hemsley, Jian Qin, and Mark Costa. 2017. "Big Data, Big Metadata and Quantitative Study of Science: A Workflow Model for Big Scientometrics." *Proceedings of the Association for Information Science and Technology* 54 (1): 36–45.

Bray, Freddie, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L. Siegel, Lindsey A. Torre, and Ahmedin Jemal. 2018. "Global Cancer Statistics 2018: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries." *CA: A Cancer Journal for Clinicians* 68 (6): 394–424.

Brown, Peter C., Henry L. Roediger III, and Mark A. McDaniel. 2014. *Make It Stick: The Science of Successful Learning*. Harvard University Press.

Buitrago Flórez, Francisco, Rubby Casallas, Marcela Hernández, Alejandro Reyes, Silvia Restrepo, and Giovanna Danies. 2017. "Changing a Generation's Way of Thinking:

Teaching Computational Thinking through Programming." *Review of Educational Research* 87 (4): 834–60.

Cai, Zhi-Gang, Shao-Ming Zhang, Hang Zhang, Yi-Yong Zhou, Hai-Bo Wu, and Xiao-Ping Xu. 2013. "Aberrant Expression of Micro RNA s Involved in Epithelial–Mesenchymal Transition of HT-29 Cell Line." *Cell Biology International* 37 (7): 669–74.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. "BLAST+: Architecture and Applications." *BMC Bioinformatics* 10 (1): 1–9.

Castañeda-Álvarez, Nora P., Colin K. Khoury, Harold A. Achicanoy, Vivian Bernau, Hannes Dempewolf, Ruth J. Eastwood, Luigi Guarino, Ruth H. Harker, Andy Jarvis, and Nigel Maxted. 2016. "Global Conservation Priorities for Crop Wild Relatives." *Nature Plants* 2 (4): 1–6.

Catalanotto, Caterina, Carlo Cogoni, and Giuseppe Zardo. 2016. "MicroRNA in Control of Gene Expression: An Overview of Nuclear Functions." *International Journal of Molecular Sciences* 17 (10): 1712.

Chen, Binbin, Michael S. Khodadoust, Chih Long Liu, Aaron M. Newman, and Ash A. Alizadeh. 2018. "Profiling Tumor Infiltrating Immune Cells with CIBERSORT." In *Cancer Systems Biology*, 243–59. Springer.

Chou, Chen-Kai, Rue-Tsuan Liu, and Hong-Yo Kang. 2017. "MicroRNA-146b: A Novel Biomarker and Therapeutic Target for Human Papillary Thyroid Cancer." *International Journal of Molecular Sciences* 18 (3): 636.

Cibulskis, Kristian, Michael S. Lawrence, Scott L. Carter, Andrey Sivachenko, David Jaffe, Carrie Sougnez, Stacey Gabriel, Matthew Meyerson, Eric S. Lander, and Gad Getz. 2013. "Sensitive Detection of Somatic Point Mutations in Impure and Heterogeneous Cancer Samples." *Nature Biotechnology* 31 (3): 213–19.

Cimpian, Joseph R., Taek H. Kim, and Zachary T. McDermott. 2020. "Understanding Persistent Gender Gaps in STEM." *Science* 368 (6497): 1317–19.

Committee on Women Faculty in the School of Science. 1999. "A Study on the Status of Women Faculty in Science at MIT." *MIT Faculty Newsletter* 11 (4): 4–15.

Costinean, Stefan, Nicola Zanesi, Yuri Pekarsky, Esmerina Tili, Stefano Volinia, Nyla Heerema, and Carlo M. Croce. 2006. "Pre-B Cell Proliferation and Lymphoblastic Leukemia/High-Grade Lymphoma in Eμ-MiR155 Transgenic Mice." *Proceedings of the National Academy of Sciences* 103 (18): 7024–29.

Council, National Research. 2010. "Committee on Undergraduate Biology Education to Prepare Research Scientists for the 21st Century." *Bio2010: Transforming Undergraduate Education for Future Research Biologists*.

De Veirman, Marijke, Liselot Hudders, and Michelle R. Nelson. 2019. "What Is Influencer Marketing and How Does It Target Children? A Review and Direction for Future Research." *Frontiers in Psychology* 10: 2685.

DeMasi, Orianna, Alexandra Paxton, and Kevin Koy. 2020. "Ad Hoc Efforts for Advancing Data Science Education." *PLoS Computational Biology* 16 (5): e1007695.

Duggan, Maeve, Nicole B. Ellison, Cliff Lampe, Amanda Lenhart, and Mary Madden. 2015. "Demographics of Key Social Networking Platforms." *Pew Research Center* 9.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Computational Biology* 7 (10): e1002195.

Eglen, Stephen J. 2009. "A Quick Guide to Teaching R Programming to Computational Biology Students." *PLoS Computational Biology* 5 (8): e1000482.

Enikolopov, Ruben, Maria Petrova, and Konstantin Sonin. 2018. "Social Media and Corruption." *American Economic Journal: Applied Economics* 10 (1): 150–74.

Erdt, Mojisola, Aarthy Nagarajan, Sei-Ching Joanna Sin, and Yin-Leng Theng. 2016. "Altmetrics: An Analysis of the State-of-the-Art in Measuring Research Impact on Social Media." *Scientometrics* 109 (2): 1117–66.

Fang, Rui, Yong Zhu, Ling Hu, Vedbar Singh Khadka, Junmei Ai, Hanqing Zou, Dianwen Ju, Bin Jiang, Youping Deng, and Xiamin Hu. 2019. "Plasma MicroRNA Pair Panels as Novel Biomarkers for Detection of Early Stage Breast Cancer." *Frontiers in Physiology* 9: 1879.

Farris, Sarah M. 2020. "The Rise to Dominance of Genetic Model Organisms and the Decline of Curiosity-Driven Organismal Research." *Plos One* 15 (12): e0243088.

FDA News Release. 2017. "FDA Approves First Cancer Treatment for Any Solid Tumor with a Specific Genetic Feature," May 23, 2017. https://www.fda.gov/news-events/press-announcements/fda-approves-first-cancer-treatment-any-solid-tumor-specific-genetic-feature.

Fink, L. Dee. 2013. *Creating Significant Learning Experiences: An Integrated Approach to Designing College Courses*. John Wiley & Sons.

Fortin, Julie, Bjarne Bartlett, Michael Kantar, Michelle Tseng, and Zia Mehrabi. 2021. "Digital Technology Helps Remove Gender Bias in Academia." *Scientometrics* 126 (5): 4073–81.

Fourment, Mathieu, and Michael R. Gillings. 2008. "A Comparison of Common Programming Languages Used in Bioinformatics." *BMC Bioinformatics* 9 (1): 1–9.

Freeman, Scott, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. "Active Learning Increases Student Performance in Science, Engineering, and Mathematics." *Proceedings of the National Academy of Sciences* 111 (23): 8410–15.

Funk, Cary. 2020. "Key Findings about Americans' Confidence in Science and Their Views on Scientists' Role in Society."

Geraldo, Murilo Vieira, Alex Shimura Yamashita, and Edna Teruko Kimura. 2012. "MicroRNA MiR-146b-5p Regulates Signal Transduction of TGF-β by Repressing SMAD4 in Thyroid Cancer." *Oncogene* 31 (15): 1910–22.

Gong, Dapeng, Wei Shi, Sun-ju Yi, Hui Chen, John Groffen, and Nora Heisterkamp. 2012. "TGFβ Signaling Plays a Critical Role in Promoting Alternative Macrophage Activation." *BMC Immunology* 13 (1): 1–10.

Grimm, Volker, Eloy Revilla, Uta Berger, Florian Jeltsch, Wolf M. Mooij, Steven F. Railsback, Hans-Hermann Thulke, Jacob Weiner, Thorsten Wiegand, and Donald L. DeAngelis. 2005. "Pattern-Oriented Modeling of Agent-Based Complex Systems: Lessons from Ecology." *Science* 310 (5750): 987–91.

Grossman, Robert L., Allison P. Heath, Vincent Ferretti, Harold E. Varmus, Douglas R. Lowy, Warren A. Kibbe, and Louis M. Staudt. 2016. "Toward a Shared Vision for Cancer Genomic Data." *New England Journal of Medicine* 375 (12): 1109–12.

Gumpenberger, Christian, Wolfgang Glänzel, and Juan Gorraiz. 2016. "The Ecstasy and the Agony of the Altmetric Score." *Scientometrics* 108 (2): 977–82.

Gutierrez-Coarite, Rosemary, Alyssa H. Cho, Javier Mollinedo, Ishakh Pulakkatu-Thodi, and Mark G. Wright. 2021. "Macadamia Felted Coccid Impact on Macadamia Nut Yield in the

Absence of a Specialized Natural Enemy, and Economic Injury Levels." *Crop Protection* 139: 105378.

Gymrek, Melissa, Amy L. McGuire, David Golan, Eran Halperin, and Yaniv Erlich. 2013. "Identifying Personal Genomes by Surname Inference." *Science* 339 (6117): 321–24.

Haas, B., and A. Papanicolaou. 2012. "TransDecoder (Find Coding Regions within Transcripts)[WWW Document]." *URL Https://Transdecoder. Github. Io*.

Haas, Brian J., Alexie Papanicolaou, Moran Yassour, Manfred Grabherr, Philip D. Blood, Joshua Bowden, Matthew Brian Couger, David Eccles, Bo Li, and Matthias Lieber. 2013. "De Novo Transcript Sequence Reconstruction from RNA-Seq Using the Trinity Platform for Reference Generation and Analysis." *Nature Protocols* 8 (8): 1494–1512.

Hack, Catherine, and Gary Kendall. 2005. "Bioinformatics: Current Practice and Future Challenges for Life Science Education." *Biochemistry and Molecular Biology Education* 33 (2): 82–85.

Hammer, David. 1994. "Epistemological Beliefs in Introductory Physics." *Cognition and Instruction* 12 (2): 151–83.

Harambam, Jaron, Natali Helberger, and Joris van Hoboken. 2018. "Democratizing Algorithmic News Recommenders: How to Materialize Voice in a Technologically Saturated Media Ecosystem." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376 (2133): 20180088.

Hardner, Craig. 2016. "Macadamia Domestication in Hawai 'i." *Genetic Resources and Crop Evolution* 63 (8): 1411–30.

Hause, Ronald J., Colin C. Pritchard, Jay Shendure, and Stephen J. Salipante. 2016. "Classification and Characterization of Microsatellite Instability across 18 Cancer Types." *Nature Medicine* 22 (11): 1342–50.

Heemskerk, Marieke, Karen Wilson, and Mitchell Pavao-Zuckerman. 2003. "Conceptual Models as Tools for Communication across Disciplines." *Conservation Ecology* 7 (3).

Ideker, Trey, and Ruth Nussinov. 2017. *Network Approaches and Applications in Biology*. *PLoS Computational Biology*. Vol. 13. Public Library of Science San Francisco, CA USA.

Karagkouni, Dimitra, Maria D. Paraskevopoulou, Serafeim Chatzopoulos, Ioannis S. Vlachos, Spyros Tastsoglou, Ilias Kanellos, Dimitris Papadimitriou, Ioannis Kavakiotis, Sofia Maniou, and Giorgos Skoufos. 2018. "DIANA-TarBase v8: A Decade-Long Collection of Experimentally Supported MiRNA–Gene Interactions." *Nucleic Acids Research* 46 (D1): D239–45.

King, Molly M., Carl T. Bergstrom, Shelley J. Correll, Jennifer Jacquet, and Jevin D. West. 2017. "Men Set Their Own Cites High: Gender and Self-Citation across Fields and over Time." *Socius* 3: 2378023117738903.

Kodama, Yuichi, Jun Mashima, Takehide Kosuge, Eli Kaminuma, Osamu Ogasawara, Kousaku Okubo, Yasukazu Nakamura, and Toshihisa Takagi. 2018. "DNA Data Bank of Japan: 30th Anniversary." *Nucleic Acids Research* 46 (D1): D30–35.

Krützfeldt, Jan, Nikolaus Rajewsky, Ravi Braich, Kallanthottathil G. Rajeev, Thomas Tuschl, Muthiah Manoharan, and Markus Stoffel. 2005. "Silencing of MicroRNAs in Vivo with 'Antagomirs.'" *Nature* 438 (7068): 685–89.

Kumar, Sudhir, Glen Stecher, Michael Li, Christina Knyaz, and Koichiro Tamura. 2018. "MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms." *Molecular Biology and Evolution* 35 (6): 1547.

Kunze, Kyle N., Evan M. Polce, Amar Vadhera, Brady T. Williams, Benedict U. Nwachukwu, Shane J. Nho, and Jorge Chahla. 2020. "What Is the Predictive Ability and Academic Impact of the Altmetrics Score and Social Media Attention?" *The American Journal of Sports Medicine* 48 (5): 1056–62.

Langdon, Quinn K., David Peris, Brian Kyle, and Chris Todd Hittinger. 2018. "SppIDer: A Species Identification Tool to Investigate Hybrid Genomes with High-Throughput Sequencing." *Molecular Biology and Evolution* 35 (11): 2835–49.

Le, Dung T., Jennifer N. Durham, Kellie N. Smith, Hao Wang, Bjarne R. Bartlett, Laveet K. Aulakh, Steve Lu, Holly Kemberling, Cara Wilt, and Brandon S. Luber. 2017. "Mismatch Repair Deficiency Predicts Response of Solid Tumors to PD-1 Blockade." *Science* 357 (6349): 409–13.

Le, Dung T., Jennifer N. Uram, Hao Wang, Bjarne R. Bartlett, Holly Kemberling, Aleksandra D. Eyring, Andrew D. Skora, Brandon S. Luber, Nilofer S. Azad, and Dan Laheru. 2015. "PD-1 Blockade in Tumors with Mismatch-Repair Deficiency." *New England Journal of Medicine* 372 (26): 2509–20.

Li, Taiwen, Jingyu Fan, Binbin Wang, Nicole Traugh, Qianming Chen, Jun S. Liu, Bo Li, and X. Shirley Liu. 2017. "TIMER: A Web Server for Comprehensive Analysis of Tumor-Infiltrating Immune Cells." *Cancer Research* 77 (21): e108–10.

Lima, Cilene Rebouças, Murilo Vieira Geraldo, Cesar Seigi Fuziwara, Edna Teruko Kimura, and Marinilce Fagundes Santos. 2016. "MiRNA-146b-5p Upregulates Migration and Invasion of Different Papillary Thyroid Carcinoma Cells." *BMC Cancer* 16 (1): 1–13.

Lin, Jishan, Wenping Zhang, Xingtan Zhang, Xiaokai Ma, Shengcheng Zhang, Shuai Chen, Yibin Wang, Haifeng Jia, Zhenyang Liao, and Jing Lin. 2022. "Signatures of Selection in Recently Domesticated Macadamia." *Nature Communications* 13 (1): 1–12.

Liu, Wanchao, Dianyu Yang, Longmei Chen, Qingqing Liu, Wenhui Wang, Zhenghua Yang, Anquan Shang, Wenqiang Quan, and Dong Li. 2020. "Plasma Exosomal MiRNA-139-3p Is a Novel Biomarker of Colorectal Cancer." *Journal of Cancer* 11 (16): 4899.

Lu, Liangqun, Sara McCurdy, Sijia Huang, Xun Zhu, Karolina Peplowska, Maarit Tiirikainen, William A. Boisvert, and Lana X. Garmire. 2016. "Time Series MiRNA-MRNA Integrated Analysis Reveals Critical MiRNAs and Targets in Macrophage Polarization." *Scientific Reports* 6 (1): 1–14.

Mancuso, Julio, Ananta Neelim, and Joseph Vecci. 2017. "Gender Differences in Self-Promotion: Understanding the Female Modesty Constraint." *Available at SSRN 3039233*.

Mariathasan, Sanjeev, Shannon J. Turley, Dorothee Nickles, Alessandra Castiglioni, Kobe Yuen, Yulei Wang, Edward E. Kadel III, Hartmut Koeppen, Jillian L. Astarita, and Rafael Cubas. 2018. "TGFβ Attenuates Tumour Response to PD-L1 Blockade by Contributing to Exclusion of T Cells." *Nature* 554 (7693): 544–48.

MATLAB (version 2020a). 2020. The Math Works, Inc. https://www.mathworks.com/.

Matsuzaki, Koichi, Chiaki Kitano, Miki Murata, Go Sekimoto, Katsunori Yoshida, Yoshiko Uemura, Toshihito Seki, Shigeru Taketani, Jun-ichi Fujisawa, and Kazuichi Okazaki. 2009. "Smad2 and Smad3 Phosphorylated at Both Linker and COOH-Terminal Regions Transmit Malignant TGF-β Signal in Later Stages of Human Colorectal Cancer." *Cancer Research* 69 (13): 5321–30.

McClatchy, Susan, Kristin M. Bass, Daniel M. Gatti, Adam Moylan, and Gary Churchill. 2020. "Nine Quick Tips for Efficient Bioinformatics Curriculum Development and Training." *PLoS Computational Biology* 16 (7): e1008007.

MEGA, X. 2018. "Molecular Evolutionary Genetics Analysis across Computing Platforms; S Kumar, G Stecher, M Li, C Knyaz, K Tamura." *Molecular Biology and Evolution* 35 (1): 1547–49.

Milkman, Katherine L., and Jonah Berger. 2014. "The Science of Sharing and the Sharing of Science." *Proceedings of the National Academy of Sciences* 111 (supplement_4): 13642–49.

Miller H.E. n.d. *Big-Data in Cloud Computing: A Taxonomy of Risks*.

Mistry, Jaina, Sara Chuguransky, Lowri Williams, Matloob Qureshi, Gustavo A. Salazar, Erik LL Sonnhammer, Silvio CE Tosatto, Lisanna Paladin, Shriya Raj, and Lorna J. Richardson. 2021. "Pfam: The Protein Families Database in 2021." *Nucleic Acids Research* 49 (D1): D412–19.

Morris, Marika. 2016. "Gender of Sources Used in Major Canadian Media." *Ottawa, ON: Informed Opinions. URL: Http://Www. Informedopinions. Org/Wp-Content/Uploads/2016/02/Genderof-Sources-in-Canadian-Media-. Pdf [June 3, 2017]*.

Mortality Statistics Branch. n.d. *The Underlying Cause of Death Data from CDC WONDER*. Division of Vital Statistics, National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), United States Department of Health and Human Services (US DHHS). https://wonder.cdc.gov/.

Mueller, Ferdinand Baron von. 1882. *Census of the Genera of Plants Hitherto Known as Indigenous to Australia*. Government Printer, South Africa.

Murphy, Sherry L., Jiaquan Xu, Kenneth D. Kochanek, S. C. Curtin, and E. Arias. 2013. "National Vital Statistics Reports." *National Vital Statistics Reports* 61 (4).

National Academies of Sciences, Engineering. 2020. "Roundtable on Data Science Postsecondary Education: A Compilation of Meeting Highlights."

Navarro, Sandra LB, and Christianne EC Rodrigues. 2016. "Macadamia Oil Extraction Methods and Uses for the Defatted Meal Byproduct." *Trends in Food Science & Technology* 54: 148–54.

"NCBI Sequence Read Archive." 2017. Https://Identifiers.Org/Ncbi/Insdc.Sra:SRP121625. 2017.

Nelson, Gerald C., Mark W. Rosegrant, Jawoo Koo, Richard Robertson, Timothy Sulser, Tingju Zhu, Claudia Ringler, Siwa Msangi, Amanda Palazzo, and Miroslav Batka. 2009. *Climate Change: Impact on Agriculture and Costs of Adaptation*. Vol. 21. Intl Food Policy Res Inst.

Neyhart, Jeffrey L., and Eric Watkins. 2020. "An Active Learning Tool for Quantitative Genetics Instruction Using R and Shiny." *Natural Sciences Education* 49 (1): e20026.

Niu, Yingfeng, Guohua Li, Shubang Ni, Xiyong He, Cheng Zheng, Ziyan Liu, Lidan Gong, Guanghong Kong, Wei Li, and Jin Liu. 2022. "The Chromosome-Scale Reference Genome of Macadamia Tetraphylla Provides Insights Into Fatty Acid Biosynthesis." *Frontiers in Genetics* 13.

Nocera, Alexander P., Carter J. Boyd, Hunter Boudreau, Ornin Hakim, and Soroush Rais-Bahrami. 2019. "Examining the Correlation between Altmetric Score and Citations in the Urology Literature." *Urology* 134: 45–50.

Nock, Catherine J., Abdul Baten, Ramil Mauleon, Kirsty S. Langdon, Bruce Topp, Craig Hardner, Agnelo Furtado, Robert J. Henry, and Graham J. King. 2020. "Chromosome-Scale Assembly and Annotation of the Macadamia Genome (Macadamia Integrifolia HAES 741)." *G3: Genes, Genomes, Genetics* 10 (10): 3497–3504.

Ogata, Hiroyuki, Susumu Goto, Wataru Fujibuchi, and Minoru Kanehisa. 1998. "Computation with the KEGG Pathway Database." *Biosystems* 47 (1–2): 119–28.

Ov12053, Octopus vulgaris isolate. 2018. *Whole Genome Shotgun Sequencing Project. GenBank*.

Patacsil, Frederick F., and Christine Lourrine S. Tablatin. 2017. "Exploring the Importance of Soft and Hard Skills as Perceived by IT Internship Students and Industry: A Gap Analysis." *Journal of Technology and Science Education* 7 (3): 347–68.

Pew Research Center. 2019. "Trust and Mistrust in Americans' Views of Scientific Experts." PewResearch.Org. August 2, 2019. https://www.pewresearch.org/science/wp-content/uploads/sites/16/2019/08/PS_08.02.19_trust.in_.scientists_FULLREPORT_8.5.19.pdf.

Porter, Alan, and Ismael Rafols. 2009. "Is Science Becoming More Interdisciplinary? Measuring and Mapping Six Research Fields over Time." *Scientometrics* 81 (3): 719–45.

PwC, P. 2013. "Crop Wild Relatives: A Valuable Resource for Crop Development." 2013. https://pwc.blogs.com/files/pwc-seed-bank-analysis-for-msb-0713.pdf.

Raghavan, Manish, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. "Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices." In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–81.

Rengel, Alejandra, Elevina Pérez, Georges Piombo, Julien Ricci, Adrien Servent, María Soledad Tapia, Olivier Gibert, and Didier Montet. 2015. "Lipid Profile and Antioxidant Activity of Macadamia Nuts (Macadamia Integrifolia) Cultivated in Venezuela."

Sambell, Kay, Liz McDowell, and Catherine Montgomery. 2012. *Assessment for Learning in Higher Education*. Routledge.

Santamaría, Lucía, and Helena Mihaljević. 2018. "Comparison and Benchmark of Name-to-Gender Inference Services." *PeerJ Computer Science* 4: e156.

Shah, Samit, Arthur G. Cox, and Martin M. Zdanowicz. 2013. "Student Perceptions of the Use of Pre-Recorded Lecture Modules and Class Exercises in a Molecular Biology Course." *Currents in Pharmacy Teaching and Learning* 5 (6): 651–58.

Smith, Michelle K., and William B. Wood. 2016. "Teaching Genetics: Past, Present, and Future." *Genetics* 204 (1): 5–10.

Smith, Michelle K., William B. Wood, Ken Krauter, and Jennifer K. Knight. 2011. "Combining Peer Discussion with Instructor Explanation Increases Student Learning from In-Class Concept Questions." *CBE—Life Sciences Education* 10 (1): 55–63.

Snyder, Alexandra, Vladimir Makarov, Taha Merghoub, Jianda Yuan, Jesse M. Zaretsky, Alexis Desrichard, Logan A. Walsh, Michael A. Postow, Phillip Wong, and Teresa S. Ho. 2014. "Genetic Basis for Clinical Response to CTLA-4 Blockade in Melanoma." *New England Journal of Medicine* 371 (23): 2189–99.

Sopinka, Natalie M., Laura E. Coristine, Maria C. DeRosa, Chelsea M. Rochman, Brian L. Owens, and Steven J. Cooke. 2020. "Envisioning the Scientific Paper of the Future." *Facets* 5 (1): 1–16.

Stern, Nicholas. 2006. "Stern Review: The Economics of Climate Change."

"SuccessFactors: Human Capital Management." n.d. HXM: The Evolution of HCM Software. Accessed June 30, 2022. https://www.sap.com/products/human-resources-hcm.html.

Tang, Youyong, Yajing Zhao, Xingguo Song, Xianrang Song, Limin Niu, and Li Xie. 2019. "Tumor-derived Exosomal MiRNA-320d as a Biomarker for Metastatic Colorectal Cancer." *Journal of Clinical Laboratory Analysis* 33 (9): e23004.

Tauriello, Daniele VF, Sergio Palomo-Ponce, Diana Stork, Antonio Berenguer-Llergo, Jordi Badia-Ramentol, Mar Iglesias, Marta Sevillano, Sales Ibiza, Adrià Cañellas, and Xavier Hernando-Momblona. 2018. "TGFβ Drives Immune Evasion in Genetically Reconstituted Colon Cancer Metastasis." *Nature* 554 (7693): 538–43.

Team, R. Core. 2013. "R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria." *Http://Www. R-Project. Org/*.

Terry, J., M. Kashyap, S. Davies, A. Flood, S. Karve, and B. Sethi. 2017. "Women Unbound: Unleashing Female Entrepreneurial Potential." In . Pwc.

Thelwall, Mike. 2018. "Does Female-Authored Research Have More Educational Impact than Male-Authored Research?"

———. 2021. "Measuring Societal Impacts of Research with Altmetrics? Common Problems and Mistakes." *Journal of Economic Surveys* 35 (5): 1302–14.

Thelwall, Mike, and Tamara Nevill. 2018. "Could Scientists Use Altmetric. Com Scores to Predict Longer Term Citation Counts?" *Journal of Informetrics* 12 (1): 237–48.

Topalian, Suzanne L., F. Stephen Hodi, Julie R. Brahmer, Scott N. Gettinger, David C. Smith, David F. McDermott, John D. Powderly, Richard D. Carvajal, Jeffrey A. Sosman, and Michael B. Atkins. 2012. "Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer." *New England Journal of Medicine* 366 (26): 2443–54.

"UniProt: The Universal Protein Knowledgebase in 2021." 2021. *Nucleic Acids Research* 49 (D1): D480–89.

United States Department of Agriculture- National Agricultural Statistics Service. 2019. "Pacific Region- Hawaii Macadamia Nuts Final Season Estimates." 2019. https://www.nass.usda.gov/Statistics_by_State/Hawaii/Publications/Fruits_and_Nuts/072019MacNutFinal.pdf.

Valle-Echevarria, Angel Del, Nathan Fumia, Michael A. Gore, and Michael Kantar. 2021. "Accelerating Crop Domestication in the Era of Gene Editing." *Plant Breeding Reviews* 45: 185–211.

Van, R. Guido, and F. Drake. 2009. "Python 3 Reference Manual." *Scotts Valley, CA: CreateSpace* 10: 1593511.

Van Roosbroeck, Katrien, Francesca Fanini, Tetsuro Setoyama, Cristina Ivan, Cristian Rodriguez-Aguayo, Enrique Fuentes-Mattei, Lianchun Xiao, Ivan Vannini, Roxana S. Redis, and Lucilla D'Abundo. 2017. "Combining Anti-MiR-155 with Chemotherapy for the Treatment of Lung Cancers." *Clinical Cancer Research* 23 (11): 2891–2904.

Vannini, Ivan, Francesca Fanini, and Muller Fabbri. 2018. "Emerging Roles of MicroRNAs in Cancer." *Current Opinion in Genetics & Development* 48: 128–33.

Victoria, Philosophical Institute of. 1858. *Transactions of the Philosophical Institute of Victoria from January to December...* Vol. 2. The Institute.

Vincent, Holly, Ahmed Amri, Nora P. Castañeda-Álvarez, Hannes Dempewolf, Ehsan Dulloo, Luigi Guarino, David Hole, Chikelu Mba, Alvaro Toledo, and Nigel Maxted. 2019. "Modeling of Crop Wild Relative Species Identifies Areas Globally for in Situ Conservation." *Communications Biology* 2 (1): 1–8.

Vlachos, Ioannis S., Konstantinos Zagganas, Maria D. Paraskevopoulou, Georgios Georgakilas, Dimitra Karagkouni, Thanasis Vergoulis, Theodore Dalamagas, and Artemis G. Hatzigeorgiou. 2015. "DIANA-MiRPath v3. 0: Deciphering MicroRNA Function with Experimental Support." *Nucleic Acids Research* 43 (W1): W460–66.

Vogelstein, Bert, Nickolas Papadopoulos, Victor E. Velculescu, Shibin Zhou, Luis A. Diaz Jr, and Kenneth W. Kinzler. 2013. "Cancer Genome Landscapes." *Science* 339 (6127): 1546–58.

Volinia, Stefano, George A. Calin, Chang-Gong Liu, Stefan Ambs, Amelia Cimmino, Fabio Petrocca, Rosa Visone, Marilena Iorio, Claudia Roldo, and Manuela Ferracin. 2006. "A MicroRNA Expression Signature of Human Solid Tumors Defines Cancer Gene Targets." *Proceedings of the National Academy of Sciences* 103 (7): 2257–61.

Wais, Kamil. 2016. "Gender Prediction Methods Based on First Names with GenderizeR." *R J.* 8 (1): 17.

Wambugu, Peterson W., and Robert Henry. 2022. "Supporting in Situ Conservation of the Genetic Diversity of Crop Wild Relatives Using Genomic Technologies." *Molecular Ecology* 31 (8): 2207–22.

Welch, Lonnie, Fran Lewitter, Russell Schwartz, Cath Brooksbank, Predrag Radivojac, Bruno Gaeta, and Maria Victoria Schneider. 2014. "Bioinformatics Curriculum Guidelines: Toward a Definition of Core Competencies." *PLOS Computational Biology* 10 (3): e1003496.

Whitley, Kiara V., Josie A. Tueller, and K. Scott Weber. 2020. "Genomics Education in the Era of Personal Genomics: Academic, Professional, and Public Considerations." *International Journal of Molecular Sciences* 21 (3): 768.

Wiggins, Grant, Grant P. Wiggins, and Jay McTighe. 2005. *Understanding by Design*. Ascd.

Zanutto, Susanna, Chiara Maura Ciniselli, Antonino Belfiore, Mara Lecchi, Enzo Masci, Gabriele Delconte, Massimo Primignani, Giulia Tosetti, Marco Dal Fante, and Linda Fazzini. 2020. "Plasma MiRNA-based Signatures in CRC Screening Programs." *International Journal of Cancer* 146 (4): 1164–73.

Zhang, Xiaoping, Maoquan Li, Keqiang Zuo, Dan Li, Meng Ye, Lanbao Ding, Haidong Cai, Da Fu, Youben Fan, and Zhongwei Lv. 2013. "Upregulated MiR-155 in Papillary Thyroid Carcinoma Promotes Tumor Growth by Targeting APC and Activating Wnt/β-Catenin Signaling." *The Journal of Clinical Endocrinology & Metabolism* 98 (8): E1305–13.

Zhou, Hong, Jayson X. Chen, Chung S. Yang, Mary Qu Yang, Youping Deng, and Hong Wang. 2014. "Gene Regulation Mediated by MicroRNAs in Response to Green Tea Polyphenol EGCG in Mouse Lung Cancer." *BMC Genomics* 15 (11): 1–10.

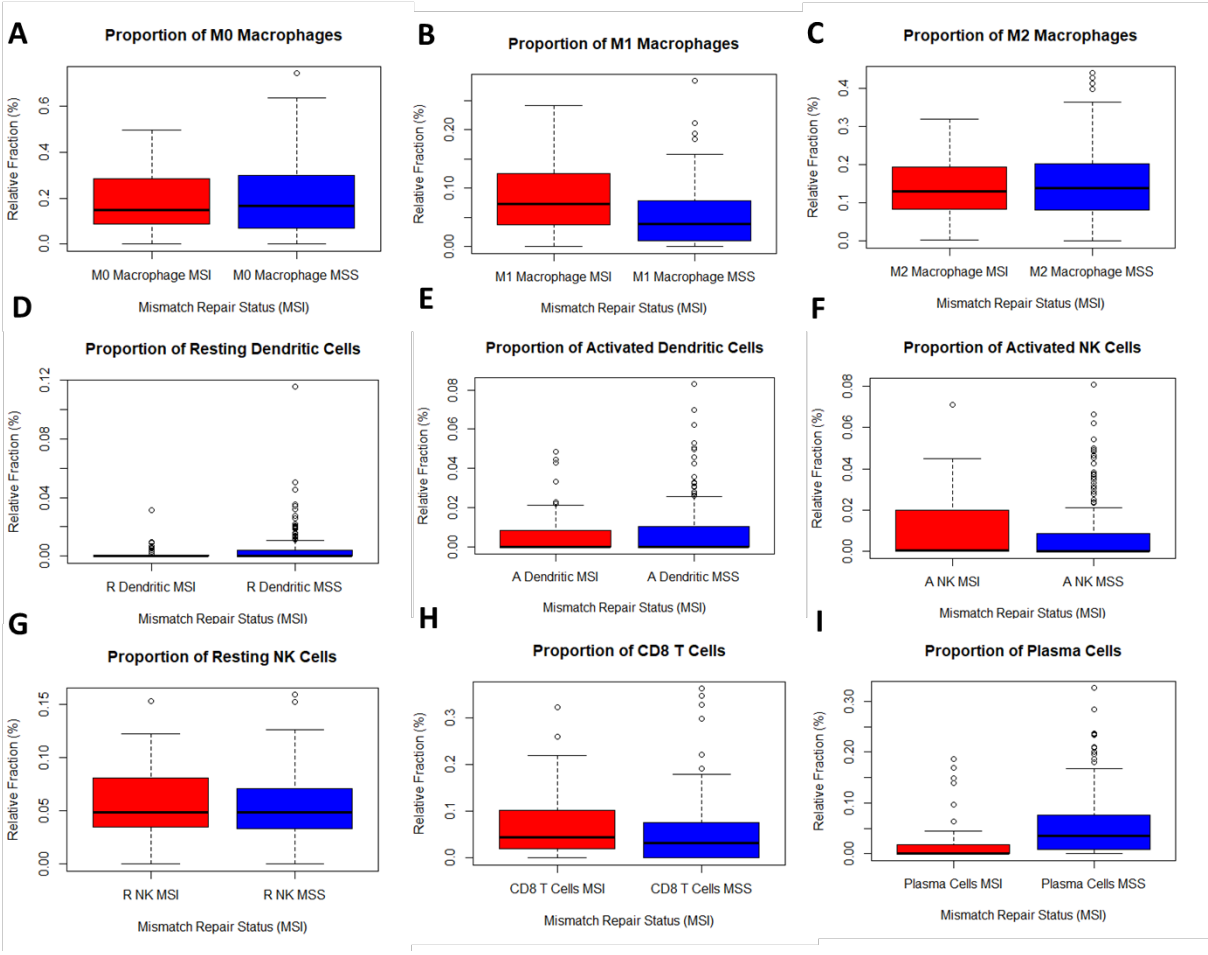
Zhu, Xiaolin, Tingting Zhang, Ye Zhang, Hao Chen, Jianbo Shen, Xinxin Jin, Jinhuan Wei, Erhao Zhang, Mingbing Xiao, and Yihui Fan. 2020. "A Super-Enhancer Controls TGF-β Signaling in Pancreatic Cancer through Downregulation of TGFBR2." *Cellular Signalling* 66: 109470.

Zou, James, and Londa Schiebinger. 2018. *AI Can Be Sexist and Racist—It's Time to Make It Fair*. Nature Publishing Group.

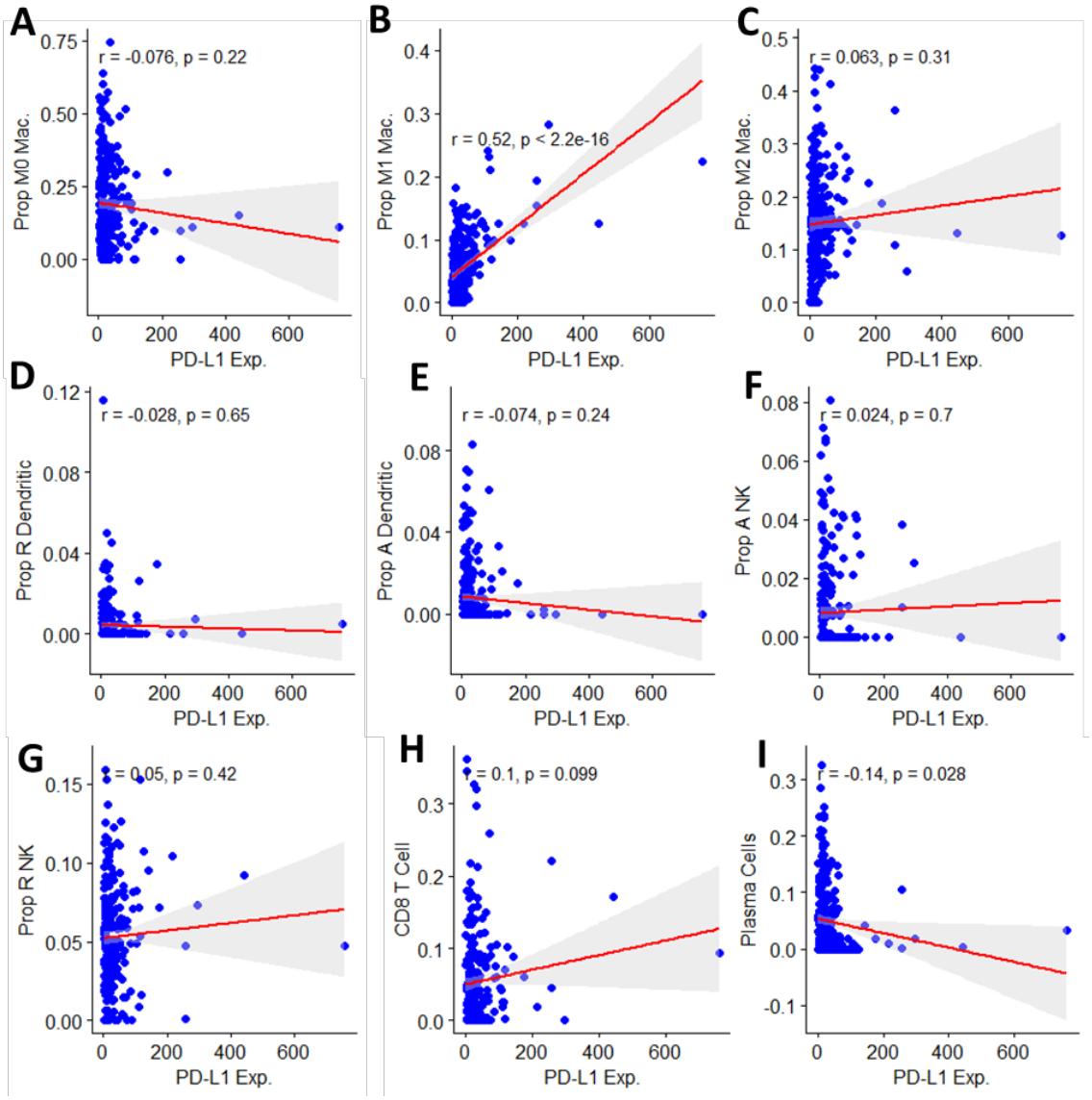
APPENDIX A: SUPPLEMENTARY INFORMATION CHAPTER 2

Feature		COAD	READ	COAD % Total	READ % Total
Stage	i	67	28	16.50%	19.58%
	ii	158	42	38.92%	29.37%
	iii	113	43	27.83%	30.07%
	iv	58	21	14.29%	14.69%
	No Data	10	9	2.46%	6.29%
MSI Status	MSS/MSI-L	332	140	81.77%	97.90%
	MSI-H	74	2	18.23%	1.40%
	No Data	0	1	0.00%	0.70%
Tissue Location	Ascending Colon	77	0	18.97%	0.00%
	Cecum	95	0	23.40%	0.00%
	Descending Colon	17	0	4.19%	0.00%
	Hepatic Flexure	27	0	6.65%	0.00%
	Rectosigmoid Junction	1	40	0.25%	27.97%
	Sigmoid Colon	134	3	33.00%	2.10%
	Splenic Flexure	6	0	1.48%	0.00%
	Transverse Colon	33	0	8.13%	0.00%
	Recum	0	97	0.00%	67.83%
	No Data	16	3	3.94%	2.10%

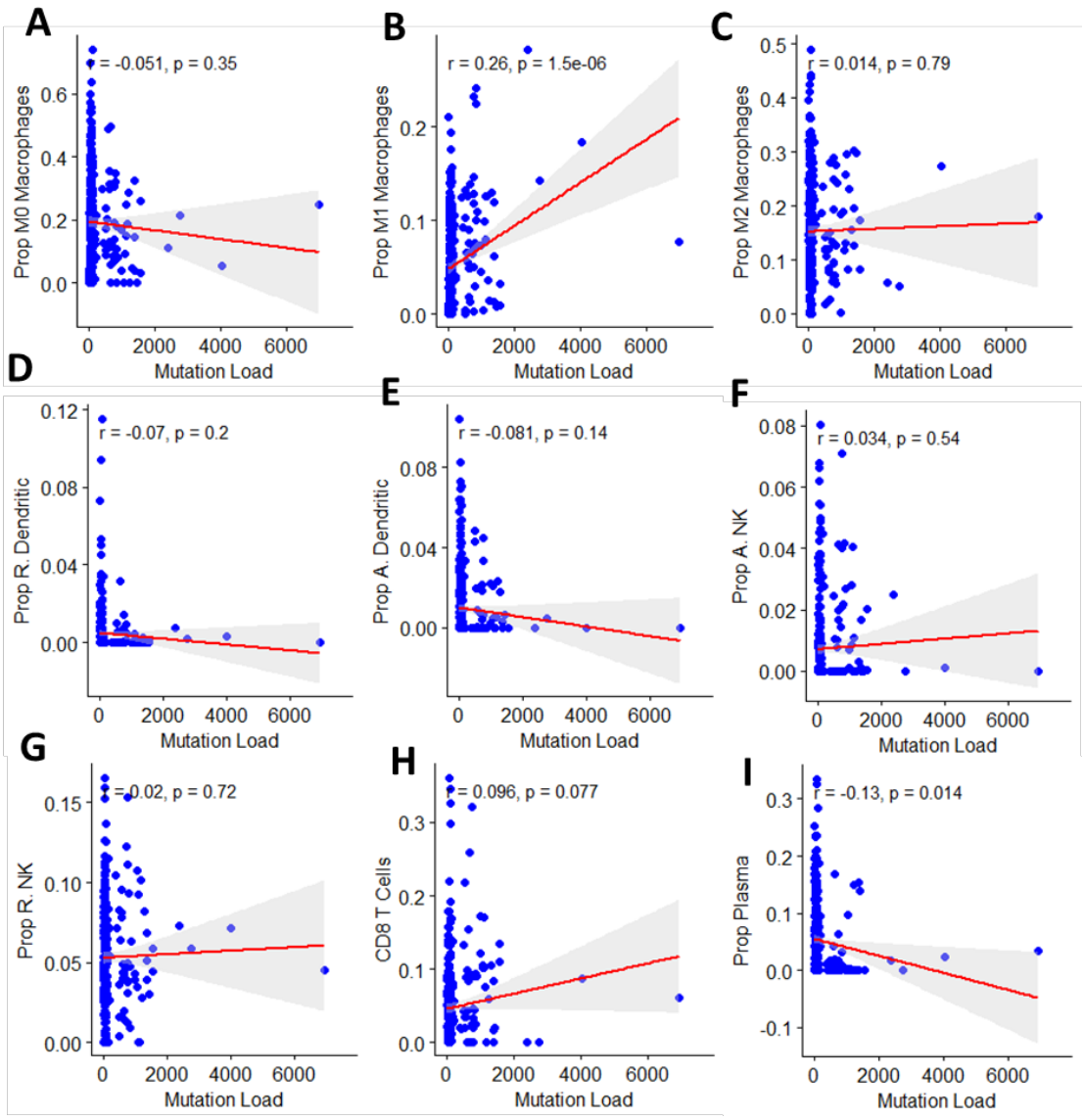
Supplementary Table ST1: Clinical features of the population of 406 COAD patients and 143 READ patients in the CRC cohort as reported in the TCGA clinical feature tables. Stage, MSI status, and tissue location were assessed.



Supplementary Figure SF1: Association of MSI Status with: M0 macrophage polarization (p = 0.30), M1 macrophage polarization (p = 0.00), M2 macrophage polarization (p = 0.82), resting dendritic cells (DC) (p = 0.01), activated DC (p = 0.82), activated natural killer (NK) cells (p = 0.19), resting NK cells (p = 0.64), CD8 t-cells (p = 0.13), plasma cells (p = 0.00).



Supplementary Figure SF2: Association of PD-L1 expression with: M0 macrophage polarization (p = 0.22), M1 macrophage polarization (p = 0), M2 macrophage polarization (p = 0.31), resting dendritic cells (DC) (p = 0.65), activated DC (p = 0.24), activated natural killer (NK) cells (p = 0.70), resting NK cells (p = 0.42), CD8 t-cells (p = 0.10), plasma cells (p = 0.03).

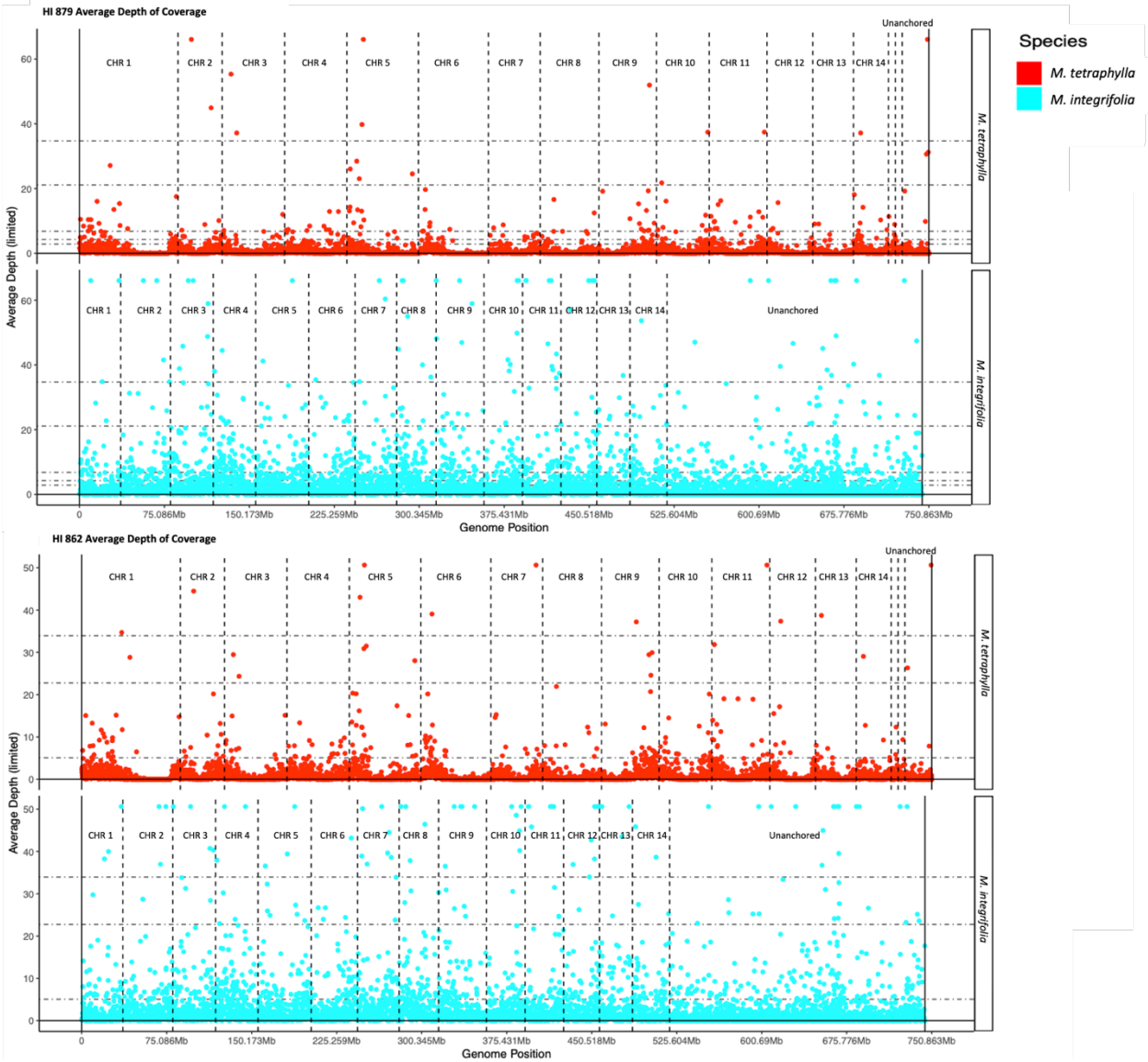


Supplementary Figure SF3: Association of mutation burden with: M0 macrophage polarization (p = 0.35), M1 macrophage polarization (p = 0.00), M2 macrophage polarization (p = 0.79), resting dendritic cells (DC) (p = 0.20), activated DC (p = 0.14), activated natural killer (NK) cells (p = 0.54), resting NK cells (p = 0.72), CD8 t-cells (p = 0.08), plasma cells (p = 0.01).

APPENDIX B: SUPPLEMENTARY INFORMATION CHAPTER 3

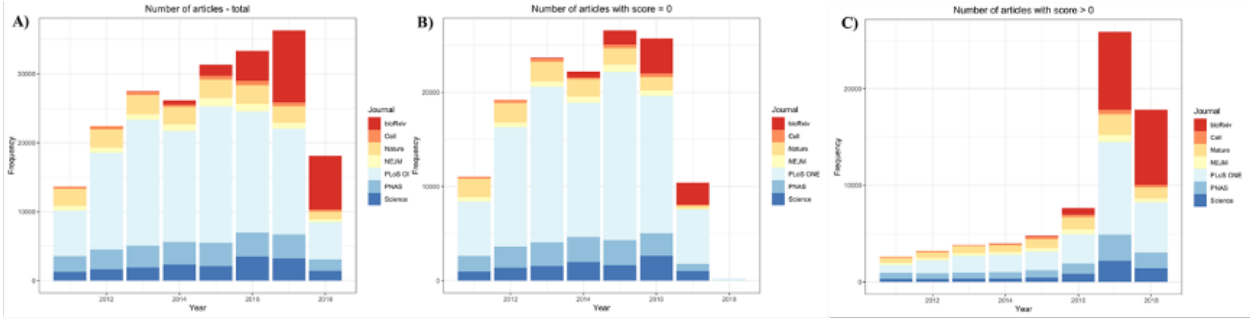
	MT-1 -HI 879			MT-10-HI 862			MT-5-Oahu Tetraphylla			MT-9-Oahu putative hybrid			Control MacInt			Control MacTet		
	Unmapped	MacTet	MacInt	Unmapped	MacTet	MacInt	Unmapped	MacTet	MacInt	Unmapped	MacTet	MacInt	Unmapped	MacTet	MacInt	Unmapped	MacTet	MacInt
Number Reads	54262791			54241998			46657000			43205846			460617384			51773462		
Number Mapped Reads	53873335			54158604			45630572			42479302			459150448			42861472		
Percent Unmapped Reads	48.67%			50.39%			48.59%			49.25%			16.83%			58.02%		
Average MQ	18.80650706			17.59945944			18.44314973			17.78028851			NA			14.92664823		
Median MQ	6			0			6			3			49			0		
Total mapped reads	389456	17512904	36360431	83394	19408531	34750073	1026428	26324256	19306316	726544	24602802	17876500	1466936	78855673	380294775	8911990	27822186	15039286
% of all reads	0.72%	32.27%	67.01%	0.15%	35.78%	64.06%	2.20%	56.42%	41.38%	0.0168	0.5694	0.4138	0.32%	17.12%	82.56%	17.21%	53.74%	29.05%
% Nonzero mapped reads	0%	19.60%	80.40%	0%	24.82%	75.18%	0%	64.82%	35.18%	0	0.6528	0.3472	0%	13.49%	86.51%	0.00%	78.12%	21.88%
All average MQ	0	9.48	23.5	0	10.31	21.71	0	20.71	16.33	0	19.84	15.67	0	26.07	0	0	21.22	12.12
NonZero average MQ	0	30.41	38.16	0	29.95	37.3	0	35.07	37.37	0	34.11	36.79	0	39.79	0	0	34.78	38.34
All Median MQ	0	0	11	0	0	10	0	10	0	0	9	0	0	27	57	0	10	0
Nonzero Median MQ	0	27	40	0	27	40	0	33	40	0	27	40	0	40	60	0	34	40

Supplemental Table ST2. Genomic hybridity statistics for tested samples. Yellow coloring indicates breeding lines, light blue coloring indicates O‘ahu collected putative *M. tetraphylla* and purple indicates controls.



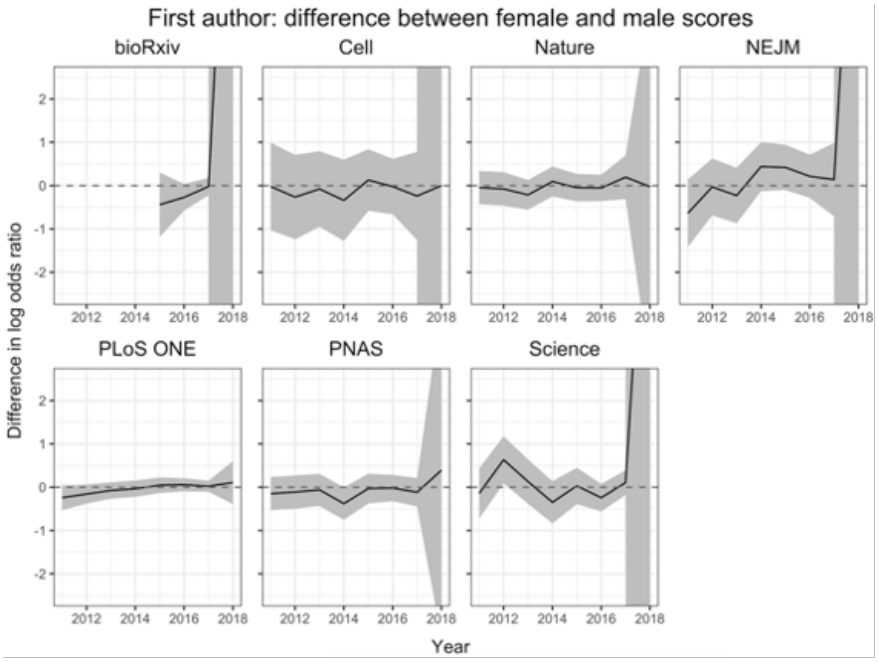
Supplemental Figure SF4: Average depth of coverage across the entire genome for Hawai‘i 879 and Hawai‘i 862. Red indicates *M. tetraphylla* depth and teal indicates *M. integrifolia* depth. Vertical dotted lines represent chromosomes.

APPENDIX C: SUPPLEMENTARY INFORMATION CHAPTER 4

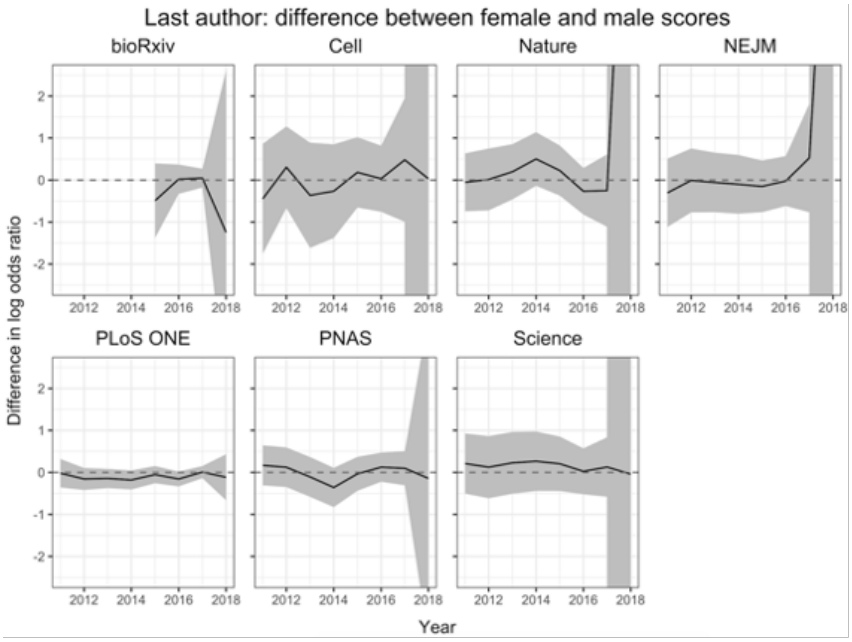


Supplemental Figure SF5. A) Total number of articles by year, by journal. There are fewer total articles in 2018 because we acquired the data from Altmetric.com partway through 2019, but only included articles from 2018 that were old enough to have computed 1-year scores. **B)** Number of articles that received an AAS of 0, by year, by journal. **C)** Number of articles that received an AAS greater than 0, by year, by journal. In later years, almost all articles get a score above 0. This could be because Twitter usership in academia has increased over time, and authors or journals ensure that their articles get tweeted at least once.

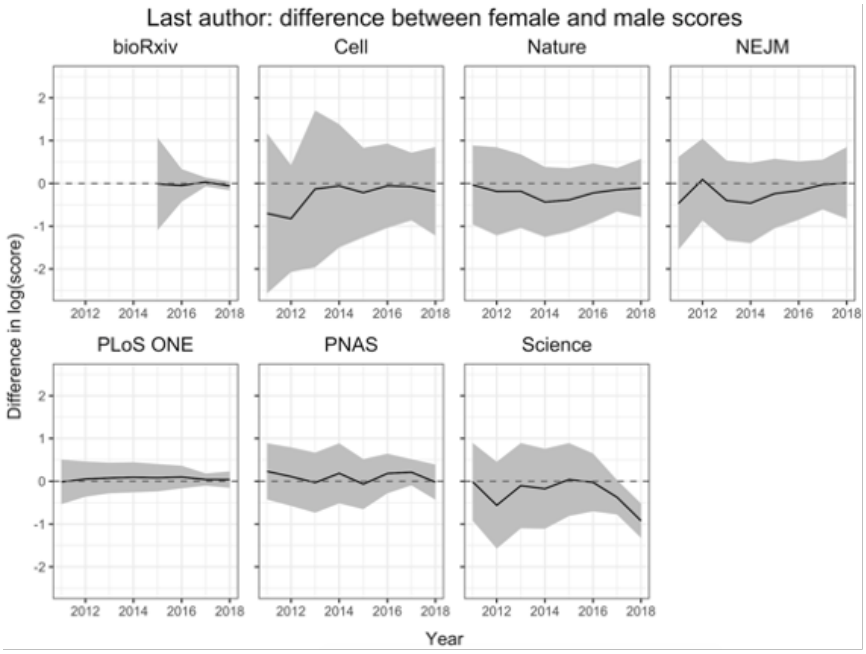
A.


First author: difference between female and male scores
bioRxiv
Cell
Nature
NEJM
PLoS ONE
PNAS
Science
Difference in log odds ratio
Year

B.

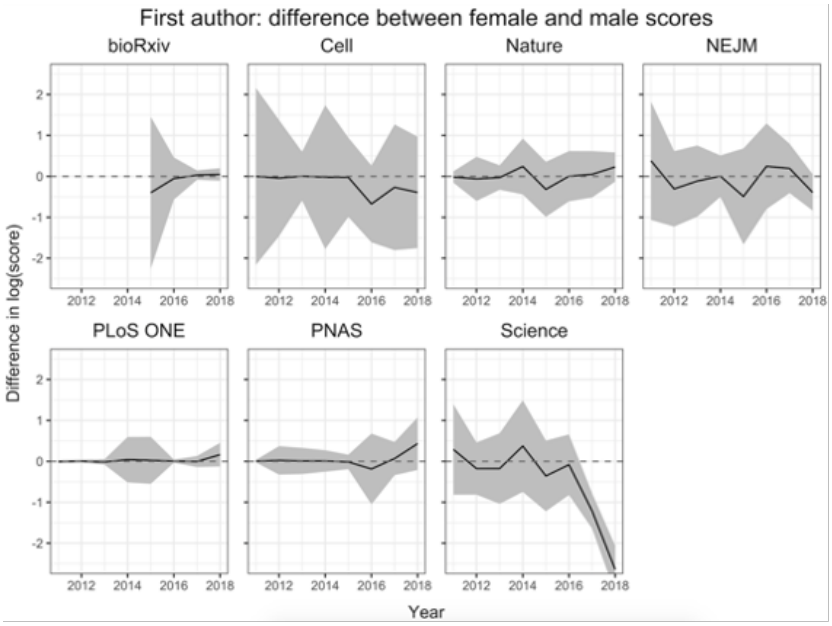

Last author: difference between female and male scores
bioRxiv
Cell
Nature
NEJM
PLoS ONE
PNAS
Science
Difference in log odds ratio
Year

C.

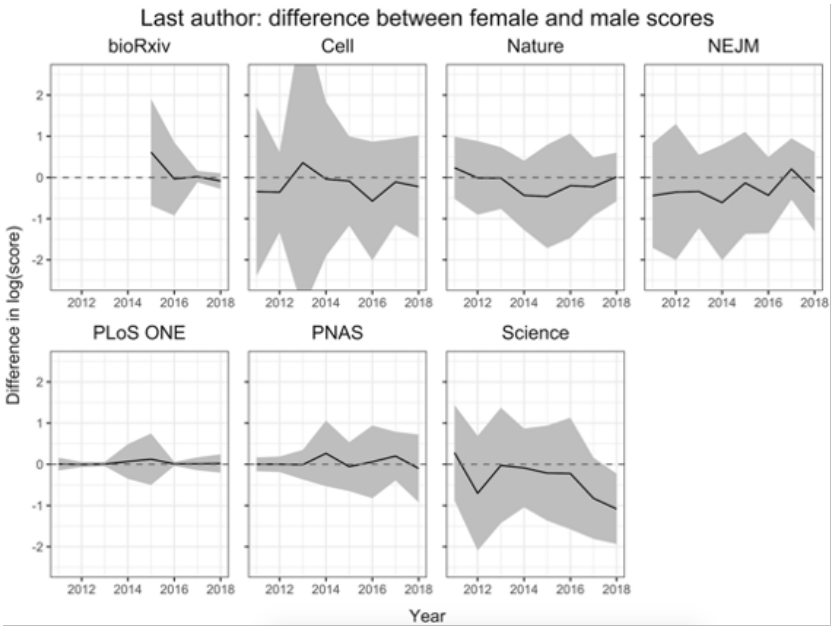


Supplementary Figure SF6. A) *Mean gender bias in seven idiotypic journals for if a journal receives an AAS for first author*s; **B)** *Mean gender bias in seven idiotypic journals for if a journal receives an AAS for last authors;* **C)** *Mean gender bias in seven idiotypic journals for magnitude of AAS for last authors. A) and B) have wide confidence intervals that go beyond the bounds of the figure for 2018 because, by 2018, very few articles have scores of 0 (see Figure S1).*

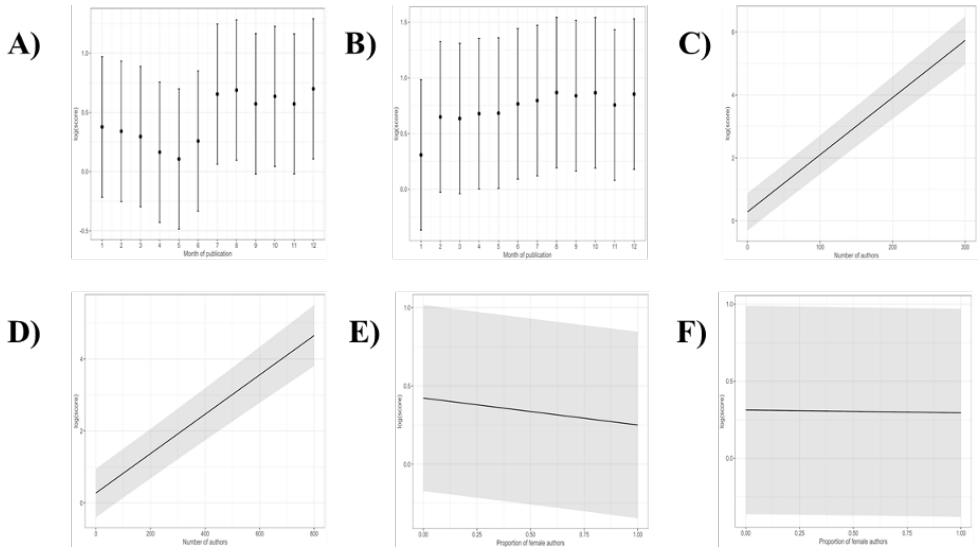
A)



B)



Supplementary SF7. A) *Median gender bias in seven idiotypic journals for the magnitude of AAS for first authors*; **B)** *Median gender bias in seven idiotypic journals for the magnitude of AAS for last authors.*



Supplementary Figure SF8. A) *Mean impact of publication month on log(AAS) for bioRxiv.* **B)** *Mean impact of publication month on log(AAS) for all journals except bioRxiv.* **C)** *Mean impact of number of authors on log(AAS) for bioRxiv.* **D)** *Mean impact of number of authors on log(AAS) for all journals except bioRxiv.* **E)** *Mean impact of proportion of female authors on log(AAS) for bioRxiv.* **F)** *Mean impact of proportion of female authors on log(AAS) for all journals except bioRxiv.*

APPENDIX D: SUPPLEMENTARY INFORMATION CHAPTER 5

Supplementary Document SD1: Seminar Syllabus

INBRE Practicum: Phylogenetic Analysis

This bioinformatics practicum from INBRE is designed for undergraduate students to gain technical training in bioinformatics. The course is focused on applying data science tools in R that are freely available online for genome analysis.

Graduate Assistant:
Bjarne Bartlett, bjarne@hawaii.edu

INBRE Bioinformatics Objectives:
Gain knowledge of bioinformatic data. Gain knowledge of basic R programming and common tools for processing bioinformatic data.

Course Objectives:
1. Use R to enter and edit expressions and scripts.
2. Read, subset, and reshape, tabular data.
3. Find and install external R packages.
4. Make figures and tables from data.
5. Knowledge of bioinformatic data, including types of data, types of data science, and current challenges in large bioinformatic data sets.
6. Know basic principles of designing a bioinformatic study.

Textbook:
This course follows the recent literature and does not utilize a textbook.

Course Format:

Week 1	**Biology Primer**
June 3/2021 Watch Film: *Cracking the Code of Life*	**Molecular Biology Introduction** ● Central Dogma of Molecular Biology ● Human Genome ● Introns/Exons ● Mutations **Biological Databases Introduction** ● Types of Databases ● NCBI Introduction **Assignments:** Week 1 Homework
Week 2	**R Primer**

June 10/2021 Due: Week 1 Homework: Biological Sequences (on Laulima)*	**R Programming** ● IDE's and R Studio ● Vectors ● Variables ● How to submit homework and R Markdown Documents **R Packages** ● Package Databases ● Installing Packages **Choose Project** ● Hemoglobin ● Cytochrome C ● Histone H1 ● Rheumatoid Factor ● Beta-2 microglobulin ● EGFR **Assignments:** Week 2 Homework: Rentrez **Discussion:** AMP vs. Myriad Genetics
Week 3	**Sequence Alignment in BLAST**
June 17/2021 Due: Week 2 Homework Rentrez R Exercise (on Laulima)	**Methods of Sequence Alignment** ● Algorithms ● Local vs. Global **R Tools for Sequence Alignment** ● Biostrings **Assignments:** Week 3 Homework: Biostrings
Week 4	**Sequence Alignment and Multiple Sequence Alignment in R**

June 24/2021 Due: Week 3 Homework	**Methods of Multiple Sequence Alignment** ● MUSCLE ● TCOFFEE **R Tools for Multiple Sequence Alignment** ● MSA Package ● Biostrings Package **Assignments:** Week 4 Homework: Multiple Sequence Alignment
Week 5	**Phylogenetic Trees**
July 1/2021 Due: Week 4 Homework	**Phylogenetic Tree Construction** ● Clustal Omega **Assignments:** Week 5 Midterm: Present Background on Homologous Sequences
Week 6	**Introduction to Molecular Phylogeny**
July 8/2021 Due: Week 5 Midterm	**Midterm Presentations:** Present a cohort of at least four proteins with homology to your chosen protein. **Assignments:** Week 6 CLUSTAL Phylogenetic Tree Construction
Week 7	**MEGA**
July 15/2021 Due: Week 6 Homework	**Midterm Presentations:** Present a cohort of at least four proteins with homology to your chosen protein. **MEGA: Molecular Evolutionary Genetics Analysis** ● Understand features ● Interpret outputs **Constructing Phylogenetic Trees** **Assignments:** Week 7 Homework: MEGA
Week 8	**Analysis of Phylogenetic Trees**

July 22/2021 Due: Week 7 MEGA Tree Construction Homework	**MAFFT: Multiple Sequence Alignment with Phylogenetic Tools** ● Run MAFFT ● Interpret Results **Phylogenetic Tree Analysis** ● Bootstrapping **Assignments:** Week 8 Homework: Analysis and MAFFT **Discussion:** Identifying Personal Genomes by Surname Inference
Week 9	**Mr. Bayes Workshop**
July 29/2021 Due: Week 8 Homework MAFFT	**Mr. Bayes: Multiple Sequence Alignment with Phylogenetic Tools** ● Run Mr. Bayes ● Interpret output ● Compared to other methods **Assignments:** Group assignment final paper or poster (choose 1) **Extra Credit:** Mr. Bayes analysis for your sequences.
Week 10	**Final Project Due**
August 5/2021 Due: Final Paper/Poster	**Final Paper** ● 1-2 page research paper or a poster on your phylogenetic analysis of a particular protein group. **Course Evaluation Survey**
	Thank you!

* Laulima is the online course management system.

Assignments:

Assignments are for students to measure progress learning the basics of bioinformatics. Assignments are required for a course certificate.

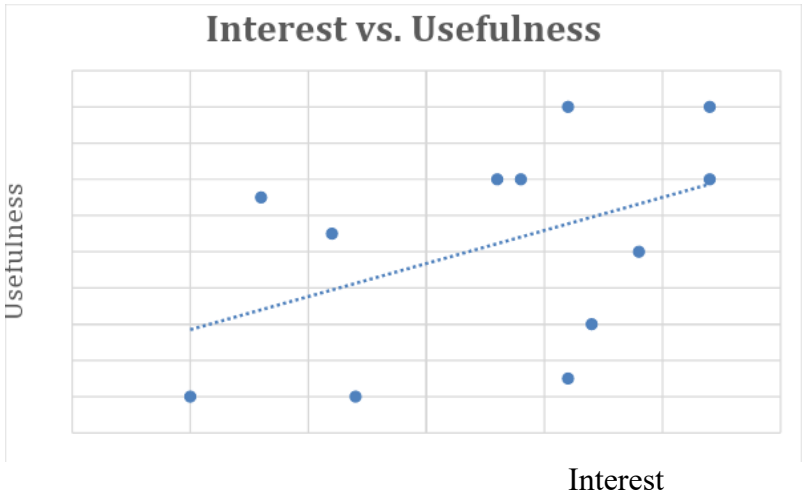
Final Project:

The final project will be chosen from a list of suggested projects and completed in groups. Individual/custom final projects are accepted/encouraged.

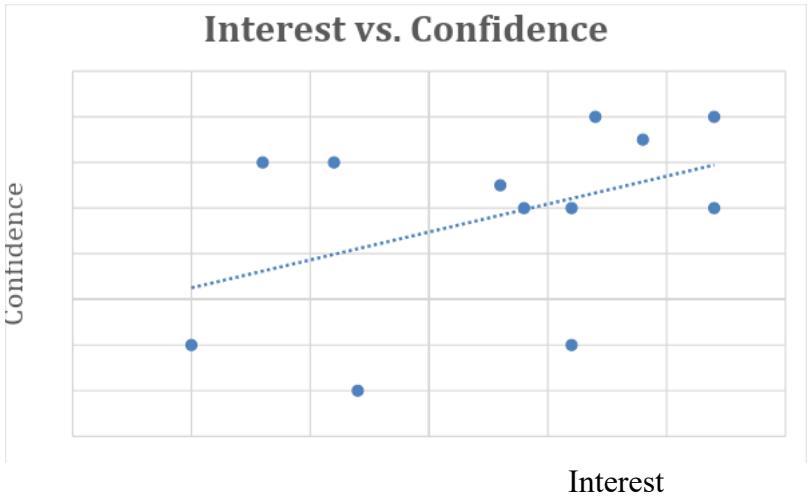
University of Hawai‘i Executive Policy 12.211

Reporting suspected academic, scientific and research misconduct is a shared and serious responsibility of all academic community members. Allegations should not be made capriciously, but indications or evidence of fraud or misconduct must not be ignored. Allegations of unethical conduct are serious and can ruin professional careers.

A.



B.



Supplementary Figure SF9: The plot of scores of students' interest vs. usefulness in topics. The scores are slightly correlated (R^2=0.2). The plot of scores of students' interest vs. confidence in topics. The scores are slightly correlated (R2=0.2).

Supplementary Table ST3: Correlation Matrix

R Values	Interest Score	Confidence Score	Usefulness
Interest Score		0.44	0.57
Confidence Score	0.44		0.80
Usefulness	0.57	0.80	

P Values	Interest Score	Confidence Score	Usefulness
Interest Score		0.1561	0.0505
Confidence Score	0.1561		0.0017
Usefulness	0.0505	0.0017	

n= 12

Supplementary Document 3: Student Comments

- Overall, I think the labs and activities were all valuable regarding the goals of this course. However, my one suggestion would be to just have a more routine date where assignments are posted and due, since some assignments weren't released until two days before the due date, which left me scrambling a bit. Other than that small detail, I enjoyed this program/course and will be able to take these skills into future experiences!
- This course was very helpful, and I wish I had more time to be able to come in and ask questions because of how useful it would be to help out the mentor that I have this semester.
- Personally I don’t have anything to say considering this is a very different course I have ever taken.
- I personally did not like the group project. If you were to do a group project again I think limiting the number of people per group would be better. My group has 13 students yet I feel like lately I have been doing the majority of the work. I think because our group is so big, the other members just wait for someone else to the assignments instead of taking the initiative to contribute.
- I don't believe the bioinformatics course was very helpful. For individuals whose INBRE research does not involve bioinformatics, it should not be required. Despite participating in all of the lectures and assignments, I don't feel like I have a good understanding of the programs we used throughout this course.
- Having group projects was definitely not a good addition to this program. There were many participants who did not help with the assignments and were passively getting credit for the work that 2-3 people were doing. These assignments did not require more than 5 people per group, but we ended up with a fairly large number of individuals in our group.
- I believe the summer INBRE semester should be conducted in the same fashion as the regular fall and spring semesters, and the bioinformatics course should be optional. After speaking with some of my group members and friends in INBRE, I came to realize that we shared a similar view of this course.
- The lectures seemed to have very little connection with the assignments and lacked overall structure.
- As this is an introductory course, which one would assume close to zero participants would have any background in either coding or command line applications, I think dropping R from the course and putting more time into the best application (instead of using various similar ones) would give the students a better foundation. I think focusing on one alignment tool and one phylogenetic tool would be more beneficial in build that foundation.
- Because there was too much content, the lectures glossed over the material without any real depth and without any reference material.
- The assignments were vague, with little to no instruction.

APPENDIX E: PUBLISHED WORK SINCE 2018 (DATE OF JOINING THE PROGRAM)

Georgiadis, A., Durham, J. N., Keefer, L., **Bartlett, B. R.**, Zielonka, M., Murphy, D., ... & Sausen, M. (2018, November). Analysis of cell-free plasma DNA to identify tumors with microsatellite instability and exceptionally high tumor mutation burden in patients treated with PD-1 blockade. In *EUROPEAN JOURNAL OF CANCER* (Vol. 103, pp. E24-E25). THE BOULEVARD, LANGFORD LANE, KIDLINGTON, OXFORD OX5 1GB, OXON, ENGLAND: ELSEVIER SCI LTD.

Mandal, R., Samstein, R. M., Lee, K. W., Havel, J. J., Wang, H., Krishna, C., Sabio E. Y., Makarov V., Kuo F., Blecua P., Ramaswamy A. T., Durham J. N., **Bartlett B.**, Ma X., Srivastava R., Middha S., Zehir A., Hechtman J. F., Morris L. G. T., Weinhold N., Riaz N., Le D. T., Diaz Jr. L.A., Chan, T. A. (2019). Genetic diversity of tumors with mismatch repair deficiency influences anti–PD-1 immunotherapy response. *Science*, *364*(6439), 485-491.

Llosa, N. J., Luber, B., Tam, A. J., Smith, K. N., Siegel, N., Awan, A. H., Fan, H., Oke, T., Zhang, J., Domingue, J., Engle, E. L., Roberts, C. A., **Bartlett, B. R.**, Aulakh, L. K., Thompson, E. D., Taube, J. M., Durham, J. N., Sears, C. L., Le, D. T., Diaz, L. A. Jr., Pardoll, D. M., Wang, H., Anders, R. A., Housseau, F. (2019). Intratumoral Adaptive Immunosuppression and Type 17 Immunity in Mismatch Repair Proficient Colorectal TumorsMSS Colorectal Tumor Immune Microenvironment. *Clinical Cancer Research*, *25*(17), 5250-5259.

Llosa, N. J., Luber, B., Siegel, N., Awan, A. H., Oke, T., Zhu, Q., **Bartlett, B. R.**, Aulakh, L. K., Thompson, E. D., Jaffee, E. M., Durham, J. N., Sears, C. L., Le, D. T., Diaz Jr. L. A., Pardoll, D. M., Wang, H., Housseau, F., Anders, R. A. (2019). Immunopathologic Stratification of Colorectal Cancer for Checkpoint Blockade ImmunotherapyImmunopathologic Stratification of Colorectal Cancer. *Cancer immunology research*, *7*(10), 1574-1579.

Smith, K. N., Llosa, N. J., Cottrell, T. R., Siegel, N., Fan, H., Suri, P., Chan, H. Y., Guo, H., Oke, T., Awan, A. H., Verde, F., Danilova, L., Anagnostou, V., Tam, A. J., Luber, B. S., **Bartlett, B. R.**, Aulakh, L. K., Sidhom, J-W., Zhu, Q., Sears, C. L., Cope, L., Sharfman, W. H., Thompson, E. D., Riemer, J., Marrone, K. A., Naidoo, J., Velculescu, V.E., Forde, P. M., Vogelstein, B., Kinzler, K. W., Papadopoulos, N., Durham, J. N., Wang, H., Le, D. T., Justesen, S., Taube, J. M., Diaz L. A., Brahmer, J. R., Pardoll, D. M., Anders, R. A., Housseau, F. (2019). Persistent mutant oncogene specific T cells in two patients benefitting from anti-PD-1. *Journal for immunotherapy of cancer*, *7*(1), 1-8.

Georgiadis, A., Durham, J. N., Keefer, L. A., **Bartlett, B.R.**, Zielonka, M., Murphy, D., White, J. R., Lu, S., Verner, E. L., Ruan, F., Riley, D., Anders, R. A., Gedvilaite, E., Angiuoli, S., Jones, S., Velculescu, V. E., Le, D. T., Diaz Jr., L. A., Sausen, M. (2019). Noninvasive Detection of Microsatellite Instability and High Tumor Mutation Burden in Cancer Patients Treated with PD-1 BlockadeMSI and High TMB in cfDNA Predict Immunotherapy Response. *Clinical Cancer Research*, *25*(23), 7024-7034.

Bartlett, B., Zhu, Y., Menor, M., Khadka, V. S., Zhang, J., Zheng, J., Zheng, J., Jiang, B., Deng, Y. (2020). Development of a RNA-Seq based prognostic signature for colon cancer. *International Journal of Computational Biology and Drug Design*, (5-6), 488-503.

APPENDIX F: WORK IN REVIEW OR PREPARATION

Paudel, R*, **Bartlett, BR***, Zamora, CM, Keach, J, Coarite-Gutierrez, R, Hawkins, J, Ahmad, A, Motomura-Wages, S, Kirk, ER, Kantar, MB, Lamour, KH, Shintaku, M, Miyasaka, S. Exploration of Taro breeding populations that show improved performance against diseases in Hawaii (**in Review People Plants Planets**)

*Contributed equally to this manuscript

Bartlett, BR, Kantar, MB, Stitt-Bergh, M, Bingham, JP. 2022. A Data Science Primer to Engage Undergraduate Students in Research. (**In Review, Biochemistry and Molecular Biology Education**)

Tamrazi A, Sundaresan S, Gulati A, Witwer KW, Frederick TJ, Wadhwa V, **Bartlett BR**, Diaz Jr. LA. "Endovascular image-guided sampling of tumor-draining veins provides an enriched source of oncological biomarkers." (**In Review, Frontiers in Oncology)**

Bartlett, BR, Cho, A, Presting, G, Laspisa, D, Gore, MA, Kantar, MB. 2022. Genomic Resources for Macadamia tetraphylla and an examination of its historic use as a crop resource in Hawaii. (**In preparation, BMC Res Notes**)

Bartlett BR, Zamora CM, Bingham JP, Hubbard A, Runck B, Kantar MB. "Journal influence on the short- and long-term impact of science." (**In Preparation**)