ISSN: 1938-7024

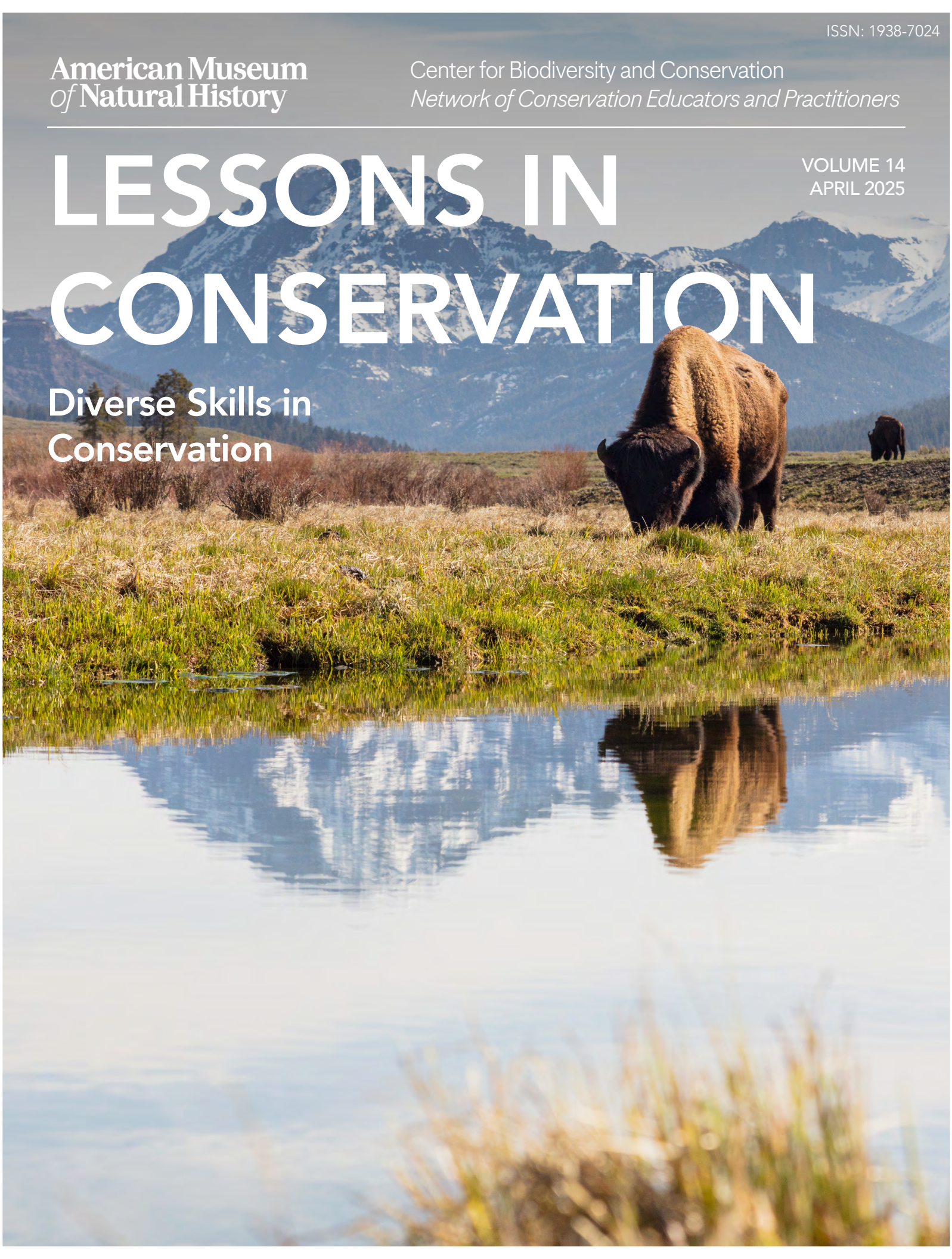
American Museum of Natural History

Center for Biodiversity and Conservation
Network of Conservation Educators and Practitioners

LESSONS IN CONSERVATION

VOLUME 14
APRIL 2025

Diverse Skills in Conservation



American Museum of Natural History

Center for Biodiversity and Conservation
Network of Conservation Educators and Practitioners

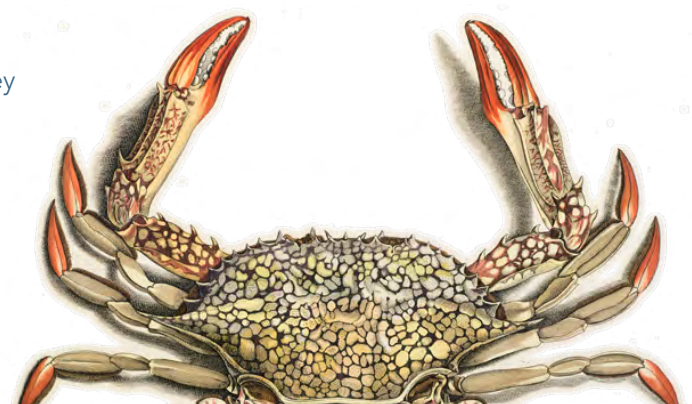
Lessons in Conservation is the official journal of the Network of Conservation Educators and Practitioners (NCEP)—a collaborative project of the Center for Biodiversity and Conservation (CBC) at the American Museum of Natural History—and is published as issues become available. Teaching and learning modules presented here in ***Lessons in Conservation*** are available in modifiable form for teachers on the NCEP website (ncep.amnh.org). All materials are distributed free of charge. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the American Museum of Natural History or the funders of this project. All components of a module (e.g., Syntheses, Exercises, and Case Studies) have been peer-reviewed and approved for publication by NCEP.

Lessons in Conservation
is available online at:
ncep.amnh.org/linc

Cover photo: Bull bison graze along an ephemeral pool in Lamar Valley
Credit: NPS / Jacob W. Frank

Journal illustrations obtained from the American Museum of Natural History's library: images.library.amnh.org/digital/

And the Biodiversity Heritage Library:
www.flickr.com/photos/biodivlibrary/



Letter from the Editors

Dear Reader,

Welcome to Lessons in Conservation, the official journal of the Network of Conservation Educators and Practitioners (NCEP). NCEP is a project of the American Museum of Natural History's Center for Biodiversity and Conservation (CBC), and this journal is designed to introduce NCEP's open-access teaching and learning resources (our "modules") to a broad audience of post-secondary instructors and learners. NCEP modules are designed to be adaptable for each educator's specific classroom or training needs. Modifiable Microsoft Word versions of our modules are available for download through the NCEP module collection at https://ncep.amnh.org, along with any accompanying data files, appendices, presentations, teaching notes, and exercise solutions.

Diverse Skills in Conservation is the theme for this issue of Lessons in Conservation, and within it, you will find, appropriately, a range of different applicable skills in conservation across five topics. The issue begins with a guide to one of the most fundamental skills: direct observations in nature and recording these observations (along with subsequent reflections and summaries) in a journal. The second module introduces an exercise on utilizing data gathered by citizen scientists in an oyster restoration project in New York City, which supports student skill development in summarizing and visualizing data using Microsoft Excel spreadsheets. The third module develops these skills further using the coding software R in a conservation context, showcasing the use of open-access long-term ecological datasets. Students use remote sensing and mapping tools in the fourth module to investigate the interplay between deforestation and the conservation of an endangered frog. The issue concludes with a dive into a rising topic in conservation: the field of museomics, which involves the use of museum specimens to monitor biodiversity or examine the evolution and genetic diversity of species over time. This module includes a synthesis document that provides background theory and practical logistics of conducting conservation museomics projects. The module also includes two exercises that help students critically analyze published research and develop their own research questions.

Additionally, in this issue, you will also find two perspective pieces that we felt help frame the current moment and our potential response to it as educators. The first perspective calls for instructors to better incorporate student voices into classroom and curriculum development, especially in light of personal experiences with climate change and environmental issues. The second perspective calls for balancing the heavy topics discussed in conservation classrooms through a non-traditional assignment that focuses on acts of kindness and the small, yet meaningful, changes we can make to reinforce hope.

We hope you enjoy this issue of Lessons in Conservation and encourage you to visit our website to start using the full collection of NCEP resources in your classroom!

Questions and feedback are welcome at ncep@amnh.org.

Nadav Gazit, Martha Groom, and Suzanne K. Macey

TABLE OF CONTENTS

Climate Change and the Voice from the Classroom

Luiz Caldeira Brant de Tolentino-Neto

Department of Teaching Methodology, The Federal University of Santa Maria, Santa Maria, Rio Grande do Sul, Brazil

DOI: https://doi.org/10.5531/cbc.linc.14.1.1

"To build relationships among generations is to make a commitment to the life that will exist after us."
- Ailton Krenak, Indigenous leader, environmentalist, and author

I started writing this paper in June 2024, during a nightmare that began two months prior, when my home state of Rio Grande do Sul in Brazil suffered a dramatic and historic tragedy due to torrential rains and flooding (Figure 1).

Despite an outpouring of human solidarity to help local people and the Brazilian government's efforts to reduce damages, at least 180 people have died and more than 500,000 people have been displaced from their lands, creating a new wave of environmental refugees and a rural exodus to city borders. Unpredictable event or fate? Is it the result of climate change? Who are guilty and who will bear the brunt of the impacts? What is the role of education (and educators) at times like these?

Figure 1. (A) Streets of Porto Alegre (Rio Grande do Sul capital) during floods in May 2024. On the left, the Guaíba River at a historic record level. (B) Aerial view of Porto Alegre during the floods of May 2024. On the right, the flooded pier buildings. (C) President Lula's delegation visiting the Passo de Estrela neighborhood. Cruzeiro do Sul/RS in June 2024. Image credit: Ricardo Stuckert, CC 2.0.

INTRODUCTION

The motivation for writing this text comes from my experience of the tragedy and recent calls in this journal for "responsive education" (Porzecanski et al., 2023), which motivated me to explore, theoretically and empirically, how science literacy can respond to this emerging demand related to climate change. I propose that if we can improve scientific literacy and make education more relevant to the student then we are more likely to engage students in critical analysis of the current environmental risks, including climate change, environmental challenges, and paths to resilience. The arguments for this are developed based on recent education research I have conducted in the Brazilian context.

THE CASE OF RIO GRANDE DO SUL: CONTEXT AND OPPORTUNITY

Rio Grande do Sul (RS) is the southernmost state in Brazil, a territory once occupied by Atlantic Forest close to the sea and Pampas on the flat plains but continuously transformed into crops and pasture during the last three centuries. Large areas are concentrated in the hands of a few owners (3% of properties occupy 48.6% of the croplands/planted area in the state), making industrial agriculture and cattle farming the state's economic base and exerting a deep cultural influence. It is almost impossible to describe the *gaúcho* people (as those from RS are known) without linking them to the countryside, horse breeding, *churrasco e chimarrão* (barbecue and mate tea), and their traditional forms of music and dance.

A previous catastrophic flood occurred in the state in 1941, leaving a quarter of the population in the state capital of Porto Alegre homeless and causing a dozen deaths. Separated by 83 years, the tragedies have very distant outcomes and contexts in terms of land use. The 2024 rains happened in an area that now had larger and more populated urban areas, with less permeable and vegetated lands, and more bare hills. Additionally, there were now more populated areas on the banks of rivers and lagoons, which had become shallower and more silted as they had had their courses altered to serve industrial agriculture (Lenharo, 2024).

In my experience, this history and current context make environmental issues controversial topics in this society, often avoided and difficult to question in the classroom. In a state where everyone, in some way, is involved in these economic activities, it's not easy to question how the environmental impacts of large monoculture areas or the amount of water used in meat production have become major challenges. Do students engage with and want to learn about these questions? While science education research has provided valuable evidence on the quality of textbooks and impact of curriculum standards, this is something we don't know well. What do the students want to learn, what do they think about environmental issues, about the future, and about science classes and their teachers? With the aim of filling this gap, emerged the Relevance of Science Education (ROSE) project and its *gaúcho* sample (Bordin, 2023).

THE VIEW FROM THE CLASSROOM

The ROSE project is an assessment instrument developed through international cooperation effort to "shed light on the students' voice" (Jenkins, 2006) and bring theoretical and empirical perspectives to bear on how to improve the curriculum, science classes, and how to increase young people's interest in Science and Technology (S&T) (Schreiner & Sjøberg, 2004).

It was translated, adapted, and applied three times in Brazil (in 2007, 2011 and 2014). In 2021, public

financing was approved for the application of ROSE at K-12 levels in RS, led by IDEIA, the Science Education group of the Federal University of Santa Maria. A team made up of professors, researchers in training, and doctoral, master's, and undergraduate students led the adaptation and production of the questionnaire, sample definition (representative of the state), application logistics, data tabulation, and analysis. It was called ROSES-RS.

In light of this recent tragedy I wondered, could the lessons from the ROSES-RS project provide insights about the current moment? The results provide important insights about what young people think about environmental issues and the future and collaborate with educators in addressing these topics. The whole questionnaire has 152 items, many of them directly related to science and technology (S&T), climate change, pollution, and student perceptions of causes of climate change which I bring into this discussion.

In four months, the research team visited 44 public school across 4,000 miles, resulting in a record of the interests and attitudes of 1,892 young *gaúchos*, most of them in early high school. The executive report and some analyses are available, with some sections also in English, on IDEIA's website (www.ufsm.br/ideia).

STUDENTS, SCIENCE, TECHNOLOGY, AND ENVIRONMENTAL ISSUES

Students' Perceptions of Science and Technology and Environmental Problems

We found that the majority of these students view science and technology positively; they believe that science and technology (S&T) will help us find the cure for diseases like AIDS and COVID (~82% agree) and will make our lives better (~82%). Not surprisingly given that finding, most of them (~69%) do not consider that S&T is the cause of environmental problems. This aligns with previous research at the national scale (Massarani, 2021) that found that 70% of young people believe that the main causes of environmental problems are economic and political interests, not S&T.

Student Beliefs About Solutions and the Future

Student opinion is divided between those who agree (~47%) and those who disagree (~53%) that S&T has the potential to solve environmental problems. They also feel responsible for contributing to the threats against nature (~69% agree), and only ~43% of them are optimistic about the future of the planet. Previous studies (Gouw et al., 2016) show young students were previously more optimistic than the young *gaúchos* of 2022. Might the COVID-19 pandemic have influenced current students' projections for the future? These and other factors, such as a worsening planetary crisis, may be at play. While our data cannot speak to this, Franzolin, et al. (2020) have argued that "engagement with environmental issues requires not only knowledge about, interest in, and motivation but also confidence, in the sense that citizens believe that they can contribute to real change acting in individual and/or collective ways" (Franzolin, et al. 2020 p.9).

Student Beliefs About Specific Causes of Environmental Change

Overall, the students' perceptions and attitudes data reflect a belief that they play a role in environmental impacts, can influence their environment, and are even somewhat optimistic about the future of the planet (Figure 2).

When asked about the source of pollution (which was not strictly defined and left for students to interpret), ~74% of young students consider industries the main responsible for environmental

pollution and ~25% attribute these effects to agriculture and livestock (Figure 3).

In contrast, a data review published by Evans (2021) reveals that, in Brazil, deforestation and changes in land use (including agriculture and mining) were responsible for 86% of all greenhouse

Figure 2. Results of Agree/Disagree questions from the ROSES-RS study about students' beliefs on individual responsibilities and environmental impact.

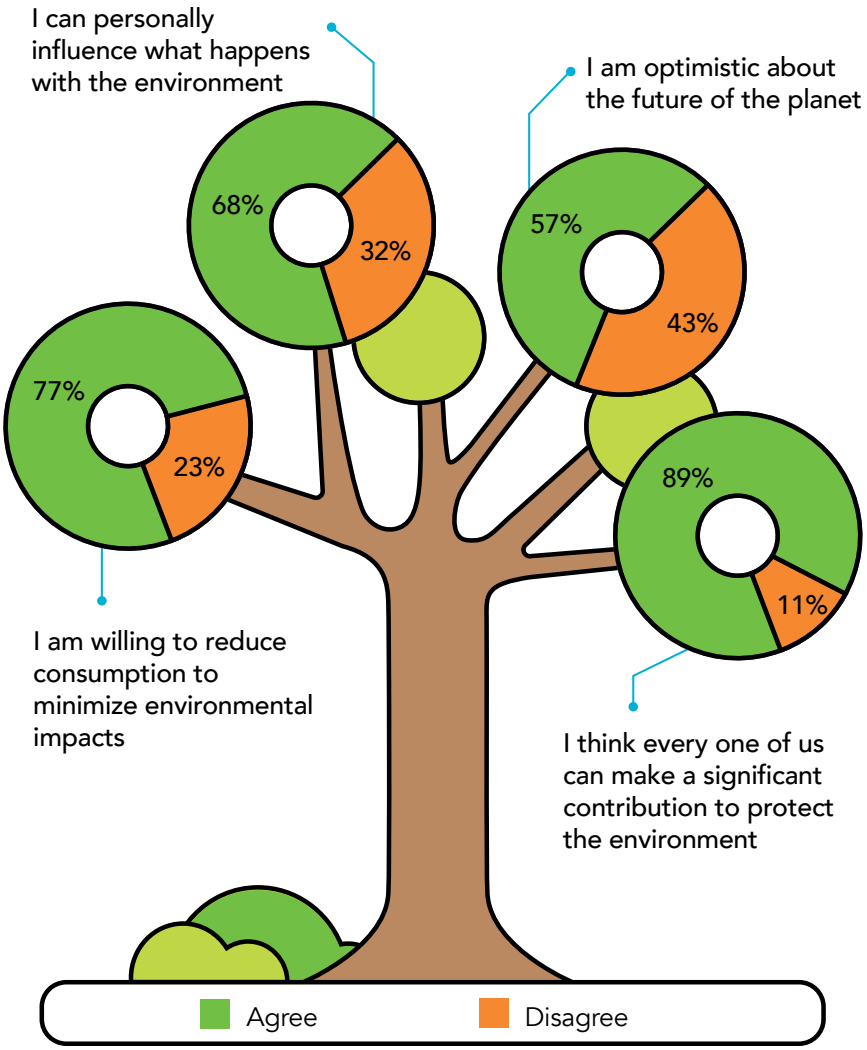
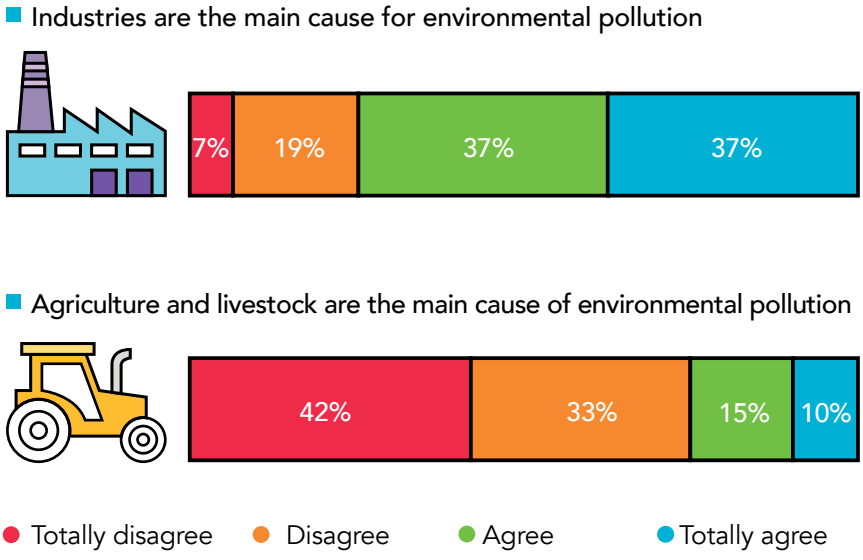


Figure 3. Students were asked to choose, on a scale from "Totally agree to Totally disagree" whether industry and agriculture are the main cause of environmental pollution.



gas emissions between 1850 and 2021. The burning of fossil fuels by industry, transport, and other activities contributes the other 14%. This points to a potentially skewed perception among students in terms of what is driving environmental change. Although we do not know for sure what Brazilian students are referring to when they think of 'environmental pollution', the data is concerning.

I hypothesize that in a state with an economy heavily based on industrial agriculture, it may be difficult, painful, or embarrassing for students to believe that those major economic activities are the main cause of environmental problems. It may also be difficult for teachers. In the context of such a dynamic, potentially also amplified by misinformation, the role of the school gains relevance: the school is the place where, by nature, controversial topics can be discussed with autonomy and respect. The intention should not be to create a clash or a dispute over reason but, through scientific data and sociological analyses, to understand natural phenomena.

What Students Report Wanting to Learn in Science Classes

Most students (~77%) think Science classes are interesting and ~67% considered it useful in daily life. Similar positions were found on a nationwide context (Gouw, 2013; Tonin et al., 2021). Most of the students (~71%) believe that their curiosity is stimulated during Science classes and ~81% of them would like to have more experiments and practical experiences. They affirm that they learn Science outside of school (~82%), when they visit zoos, museums, and planetaries (~67%) and even when they play video games (~48%). Understanding what kinds of topics students themselves wanted to learn about was also a goal of this survey; and considering our current focus on climate and the disaster, we wondered where climate ranked for students in terms of their interests in learning. The top topics they want to learn about are: how to perform first aid, the possibility of life outside Earth, how to protect endangered species, life and death and the human soul, and medicinal use of plants. Climate related topics, such as "changes in climate and how these can be influence by humans," are not among their top interests (it ranked in 47th place).

New samples will show us whether there have been changes in the priorities of these young people, whether environmental issues have gained importance. And looking further ahead, and at teacher preparation, how do those subjects appear in curricula and, more importantly, in classrooms, and how fostering climate change awareness might be connected to these problem-solving interests of the students.

Most public policies ignore students' voices, experienced scholars and teachers have their perceptions, and media and policy makers speak on behalf of young students, often without concrete evidence. Since formal education is an intentional process that involves emotions and affect, it becomes necessary to consider the interests and attitudes of students in balance with the desires and influence of other social groups. One of the paths to narrow the gap between what we need to discuss and what we would like to, is understanding what students wish to learn—and what they currently don't. I am not defending the thesis that schools should only teach what students want to learn, which would be neglecting the role of educators. But ignoring what they want to learn at school is also a form of omission.

AFTER THE TRAGEDY

The unregulated urban growth, the expansion of industrial agricultural borders, the monoculture approach, poor infrastructure, the lack of environmental protections, and misinformation are likely

to have made the damages from the storm more serious, and to potentially make these events more frequent and destructive in the future. It's a complex issue and it deserves a complex analysis but we, educators, must ask and explore what role climate change played within this.

When we consider school as a universal institution—one which the majority of people attend—it is its responsibility to show that this extreme rain event was not an isolated and unpredictable phenomenon, and that Science can be an ally in reducing damage and in the search for solutions. It is tragic that a lack of scientific literacy in Brazil appears not to be a momentary problem or a contemporary crisis, but a deliberate "national project" (Brum & Train, Rob (Translator), 2021; Pereira & De Oliveira, 2024).

At the same time, we must be careful not to advocate a technocracy, in which all decisions are taken exclusively by technicians, scientists, and specialists, naively believing in their neutrality or good and noble intentions. But we should aim to move towards a democratic system that guarantees a voice for scientific evidence in discussions and decision-making about, for instance, how we can deal with future risks and climate-driven migration. In this context, listening to the "voice of the student" is important both in classroom decisions and for public policies.

The ROSES project gives us specific data to guide our decision-making related to curricula and topics and empower us to argue that knowing students' interests and attitudes towards environmental challenges can lead us to an education more relevant to the students. These types of data can contribute to engaging students in critical analysis of the current environmental risks, including climate change challenges and paths to resilience. Given that these data were collected before the extreme rain of 2024, there is potential for longitudinal research to be informative. As research continues, any analysis must be careful and sensitive to the fact that some of these students (and their teachers) probably suffered from the rains and floods, lost their homes, schools, crops, and perhaps relatives and friends.

Yet with the power of evidence, educators can embrace education as a protagonist and active agent of societal transformation, and the school as a democratic space for change. The same logic applies to textbooks and other resources that, fundamental as they are for Brazilian schools, have to encourage and support environmental discussions, informed by science. These textbooks, oftentimes, are teaching materials the teachers need for their own learning: teachers lacking access to professional development resources or courses often update their own knowledge by reading the same books as their students. Hence teacher preparation, pre and in-service, have to consider the study of complex, contemporary phenomena and linked social impacts, such as the relationship between deforestation and extreme weather/climate events.

There is a really complex social-scientific issue here: how can we engage with the cultural value of livestock and meat as a traditional food and their impact on climate change and local landscapes? In the public policy arena of curricula, teaching resources (textbooks), and large scale assessments, changes depend on coordination between conflicting and complex interests. Engaging with these questions within the classroom is more possible.

The university as a teacher training center, as a noble place for innovations and discussions and as a space for political and social organization, plays a fundamental role both in the recovery of history and social relations (Figure 4), as well as in mitigating damage, in the search for alternatives and in overcoming environmental problems.

Figure 4. Team of volunteers working to recover historical archives at the Federal University of Santa Maria, after record flooding in May 2024.

Ailton Krenak says that telling the stories of the abandoned, the excluded, and those diminished by hegemonic history is a way of keeping them alive and "postponing the end of the world," of centering our capacity for resilience, and of sustaining our ability to not give up. The same goes for reflecting on the ideas, interests and attitudes of students, voices silenced by the educational industry. My hope is that science education and scientific literacy can help us engage with this tragedy as a springboard to act on predicting changes (not only natural but also social); to engage with questions of inclusion, accessibility, equity, education for peace, and social justice; and to be responsive to these climate phenomena in terms of education policy by developing support resources for schools, creating teacher preparation courses, analyzing textbooks, and improving.

ACKNOWLEDGMENTS

This study was financed by FAPERGS, CAPES PrInt and National Council for Scientific and Technological Development (CNPq), Brazil

ADDITIONAL RESOURCES

Da Rosa, P. A., & De Tolentino-Neto, L. C. B. (2023). Interesses de estudantes gaúchos em Ciências da Natureza e suas relações com as Unidades Temáticas da BNCC. Revista Insignare Scientia - RIS, 6(3), 103–114. https://doi.org/10.36661/2595-4520.2023v6n3.13665

Gouw, A. M. S. (2013). As opiniões, interesses e atitudes dos jovens brasileiros frente à ciência: Uma avaliação em âmbito nacional [Doutorado em Educação, Universidade de São Paulo]. https://doi.org/10.11606/T.48.2013.tde-08102013-154326

Ocampo, D. M. (2019). As tipologias dos estudantes brasileiros em relação ao interesse em Ciências e Tecnologia: Uma análise baseada nos projetos ROSE e Barômetro Brasil [Universidade Federal de Santa Maria]. https://repositorio.ufsm.br/handle/1/17538

Pinafo, J. (2016). O que os jovens têm a dizer sobre ciência e tecnologia? Opiniões, interesses e atitudes de estudantes em dois países: Brasil e Itália [Doutorado em Educação, Universidade de São Paulo]. https://doi.org/10.11606/T.48.2016.tde-01112016-110406

Silva, C. S. D. S. D., Prochnow, T. R., Pellegrini, G., & Bizzo, N. (2020). Pesquisa de percepções de estudantes do ensino médio sobre os desafios ambientais. Ciência & Educação (Bauru), 26, e20020. https://doi.org/10.1590/1516-731320200020

Tolentino Neto, L. C. B. D. (2008). Os interesses e posturas de jovens alunos frente às ciências: Resultados do Projeto ROSE aplicado no Brasil. [Doutorado em Educação, Universidade de São Paulo]. https://doi.org/10.11606/T.48.2008.tde-16062008-155323

Tonin, K. G. (2022). A preferência dos jovens brasileiros pela disciplina ciências: Interesse absoluto x interesse relativo [Universidade Federal de Santa Maria]. https://repositorio.ufsm.br/handle/1/24509

REFERENCES

Bordin, F. P. R. (with Ribeiro, B. P., Silva, P. S. S. da, Tolentino Neto, L. C. B. de, Gouw, A. M. S., Ocampo, D. M., Hansen, T. R., Piovesan, T. R., Hickmann, M., Amestoy, M. B., Czekalski, R. G., Rosa, P. A. da, & Lopes, A. F.). (2023). Os interesses dos jovens gaúchos em Ciência e Tecnologia: Projeto ROSES-RS 2022. Facos-Ufsm. https://repositorio.ufsm.br/handle/1/29603

Brum, E. & Train, Rob (Translator). (2021, March 9). Bolsonaro has turned Brazil into a global pariah. El País. https://english.elpais.com/usa/2021-03-09/bolsonaro-has-turned-brazil-into-a-global-pariah.html

Evans, S. (2021, October 5). Analysis: Which countries are historically responsible for climate change? Carbon Brief. https://www.carbonbrief.org/analysis-which-countries-are-historically-responsible-for-climate-change/

Franzolin, F., Garcia, P. S., & Bizzo, N. (2020). Amazon conservation and students' interests for biodiversity: The need to boost science education in Brazil. Science Advances, 6(35), eabb0110. https://doi.org/10.1126/sciadv.abb0110

Gouw, A. M. S. (2013). As opiniões, interesses e atitudes dos jovens brasileiros frente à ciência: Uma avaliação em âmbito nacional [Doutorado em Educação, Universidade de São Paulo]. https://doi.org/10.11606/T.48.2013.tde-08102013-154326

Gouw, A. M. S., Mota, H. S., & Bizzo, N. M. V. (2016). Brazilian youth and science: Possible relationship of interest. Revista Brasileira De Pesquisa Em Educação Em Ciências, 16(3), 649–670. https://periodicos.ufmg.br/index.php/rbpec/article/view/4586

Jenkins, E. W. (2006). The student voice and school science education. Studies in Science Education, 42(1), 49–88. https://doi.org/10.1080/03057260608560220

Lenharo, M. (2024). How to recover when a climate disaster destroys your city. Nature, 634(8036), 1032–1036. https://doi.org/10.1038/d41586-024-03472-5

Massarani, L. (with Castelfranchi, Y., Fagundes, V., & Moreira, I.). (2021). O que os jovens brasileiros pensam da ciência e da tecnologia: Pesquisa realizada pelo Instituto Nacional de Ciência e Tecnologia em Comunicação Pública da Ciência e Tecnologia (INCT-CPCT). Casa de Oswaldo Cruz. https://www.inct-cpct.ufpa.br/wp-content/uploads/2021/02/LIVRO_final_web_2pag.pdf

Pereira, F. H., & De Oliveira, R. S. (2024). Journalists and scientists together: The public problem of science disinformation in Brazil. Journal of Science Communication, 23(03). https://doi.org/10.22323/2.23030204

Porzecanski, A. L., Akabas, S., Betley, E., Blair, M., Bynum, N., Ginsburg, J. R., Groom, M. J., Macey, S. K., Moore, A. C., Rivera, C. J., & West, P. (2023). Lessons from a transformative conservation educator and building the future of conservation education. Lessons in Conservation, 13, 5–14. https://doi.org/10.5531/cbc.linc.13.1.1

Schreiner, C., & Sjøberg, S. (2004). Sowing the seeds of ROSE: Background, rationale, questionnaire development and data collection for ROSE (The Relevance of Science Education) – a comparative study of students' views of science and science education. University of Oslo, Faculty of Education, Department of Teacher Education and School Development : Unipub. https://www.observa.it/wp-content/uploads/2021/05/SowingRose.pdf

Tonin, K. G., Tolentino-Neto, L. C. B. D., & Ocampo, D. M. (2021). Young brazilians and their preference for sciences discipline. Research, Society and Development, 10(5), e8210514549. https://doi.org/10.33448/rsd-v10i5.14549

Random Acts of Kindness

Kristel Guimara

Community College of Vermont, Montpelier, VT, USA

DOI: https://doi.org/10.5531/cbc.linc.14.1.2

Kindness is one of the most powerful interpersonal tools that we, as human beings, use to connect with one anocother. When we sense someone's need, we either choose to help in some way... or we choose not to. If we act from empathy, we will offer kindness, and in that moment a surprising, gracious, humanitarian connection is made. This is the positive power that each of us—including children— possess."
- The Random Acts of Kindness Foundation

I have taught courses on biological science, biodiversity, conservation, and climate change for over a decade. These topics carry a heavy weight and can be daunting for students. Early on in my career, I noticed that around the midway point of the semester is when students lose their focus. They become unenthusiastic about the course material. The material I taught online or within the classroom was received with a mediocre response at best. It did not matter how many bells or whistles I utilized to draw the students' attention.

As I struggled to understand why student engagement slips during mid-semester, I would think about the life of a student beyond the classroom. The range of their life experiences is broad and rarely known to the teacher. Yet, this has a large impact on how they engage in any course. Each student has a busy and complex world that can place strain on their academic achievement. Snippets of these experiences are known to the teacher only when a student's grades decrease and assignments accrue a zero. This usually occurs around mid-terms or midway through the semester when they feel overwhelmed and stressed. To break the monotony and regain their participation within the class, I began to include an unorthodox assignment.

Before going into the details about the assignment, let me explain more about the context. The courses I teach have a focus on environmental science, forest ecology, and conservation biology. The material I teach is heavily laden with doom and gloom scenarios paired with concepts that weigh heavily on your heart. The scenarios are often in places beyond the walls of the classroom and the local landscape in which they reside. The messaging in the course materials often places a great deal of obligation on the younger generation to 'save the world.' My students remark on a whirlwind of emotions they experience (ranging from hopelessness to anger), as the feeling of angst and sense of powerlessness can become overwhelming.

One of the main topics discussed in each of the courses is climate change. This topic is a heavy topic in nature and often lacks the integration of empathy. Integrating empathy into each course provides a learned opportunity to connect with others across social divides. This empathy informs an individual of the ways each person responds to another person. Interacting within the classroom (both online and in-person) can be difficult for students, but creating a welcoming space for all students to speak freely provides an opportunity to delve into difficult topics. Creating structured interactions in the classroom, like discussion forums, help provide a space for students to find where their values overlap and simultaneously create an opportunity to understand how someone is harmed by climate change. Creating this space for an open dialogue builds a practice of empathy in the classroom. For example, having empathy for someone who does not have access to clean drinking water can be difficult to

understand if this hardship has not been experienced. Yet, learning of this hardship from a fellow classmate provides an opportunity to discuss this hardship in great detail. It opens the dialogue to apply previous concepts that were discussed during the first half of the semester such as aging water infrastructure (pipes), overdrawn aquifers, environmental policies, agricultural impacts to climate change, local laws, and initiatives.

This emphasis on empathy is why I developed an assignment called, "Random Acts of Kindness." (Many versions can be found in Open Education Resources (https://oercommons.org/) or you can design the assignment to suit a course or a lesson plan of the week.) The assignment entails the following: they are asked to make two 'acts of kindness' within a week. For each of the two, each student writes at least a 200-word paragraph explaining their act, which I have them share in the classroom's discussion forum. They must provide details on what they did, who was involved, how they felt doing it, and why it was important for them to do this act. The student can add more to the initial original post, but these are the basics that are required. After they submit an original post in the online discussion forum, each student must respond to at least three of their fellow classmates' posts. The responses should be a reflection with substance that contains at least 100 words. During the week, I engage with each student, in the discussion forum, to reflect on their acts of kindness and include previous concepts discussed earlier in the semester that would apply to their random acts of kindness.

With this assignment, I hope to encourage each student to do some self-reflection and writing. The description of the assignment underlines the importance in defining what "random acts of kindness" means. Each student is required to go beyond their daily routine to reach out and/or help another person or group of people. Examples are provided such as volunteering in a local animal shelter or helping a sibling/roommate with their homework. Recently, I had a student remark that she enjoys implementing random acts of kindness each day and often refers to it as "Everyday Joys."

My own main takeaways from implementing this assignment are: 1) to remind myself that students are complex, busy, and vary in personal/professional goals and, 2) to continue to learn from each of them, as they are always full of surprises and inspire hope in me.

There is so much that I learn when I implement this assignment from semester to semester. Submitted assignments are mirrored amongst the students based on current events. For instance, at the onset of the COVID-19 pandemic[1], when restrictions were put in place, much of the assignments submitted reflected their fear of the pandemic, reflections on their place of work, and inability to see family and friends. Feelings of disconnect paired with building new routines within their family dynamics left little time to focus on homework. This assignment provided an opportunity for a wellness check to those family members and friends because an assignment was required.

The stories the students shared included, but were not limited to, visiting a parent at work for a surprise lunch break because their parent works two jobs; paying the dinner bill for an unknown veteran a table over; putting together peanut butter and jelly sandwiches to hand out to the houseless people in their community; and helping their roommate with their math homework because their roommate was falling behind in their studies. There are so many stories I wish I could share. Each one has a place on my personal bulletin board that I refer to as 'the wall of kindness.'

[1]Parker, S. W., Hansen M. A., & Bernadowski C. (2021). COVID-19 campus closures in the United States: American student perceptions of forced transition to remote learning. Social sciences 10(2), 62. https://doi.org/10.3390/socsci10020062

Once students complete this assignment and share their learned experiences within the discussion forum, I feel this has created a sense of place and fostered a community within the classroom. I often notice invigorated energy in the classroom that is carried throughout the rest of the semester and learned experiences on the assignment are integrated within their writing. Often students reflect on this assignment and the engagement with their peers during the second half of the semester.
I am mindful that students appreciate and are more likely to achieve when expectations are clearly stated. If you decide to incorporate this activity in your classroom, you can decide what parameters work best for you and your students. Here are some of my suggestions and recommendations:

- I keep the assignment broad and without limitations to allow each student to decide how they want to fulfill this opportunity during the assignment.
- I have found that 200 words for each act is enough for them to think about their responses with a deeper reflection. It pushes them to write a bit more than they would without a word count.
    - But… specialize this assignment to the needs of the students. Sometimes students do not feel that they are not good at writing but better at drawing. This could be an opportunity to include an alternative option to upload a drawing/image to aid the student in describing what they have done for the assignment.
- You might also want to consider other parameters, such as where the acts take place. In my experience, I have found that this assignment is better with a classroom on campus. Some ideas could be showcasing volunteer opportunities on campus or organizations in the community.
- I would also like to emphasize that this assignment is not solely for an environmental course. This assignment is very flexible in design and can be altered to fit the needs of a class.

I always tell my students that I wish they could see themselves how I see them. They can accomplish anything but each of them has a different path to achieve that success. I strongly believe this assignment allows them to see how each of their experiences, in a small portion of a time, can give them hope to achieve success in and out of the classroom.

The design of this assignment provides students with the knowledge that learning through learned and shared experiences goes beyond the written curriculum. They can have a positive change in the world, with a ripple effect, to reach those places far and wide. They can even do it within a curriculum that showcases heavy-laden topics with, in what feels like, unattainable solutions.

In conclusion, I have found that non-traditional assignments have the most impact and provide a memorable experience for each student; especially in times when students have a hard time seeing themselves as we see them. We, as educators, want to foster a thirst for knowledge as our students grow as independent thinkers and pioneers in their craft after graduation. I want my students to walk away from the course knowing more of the course content than when they first arrived. My hope is to awaken that passion for lifelong learning that is carried through, and beyond, the classroom. However, if students leave my class feeling the weight of the world without hope or ambition then did I truly do my job?

Using a Field Journal to Enhance Observation Skills

Kristina M. Hannam

Biology Department, State University of New York at Geneseo, Geneseo, NY, USA

DOI: https://doi.org/10.5531/cbc.linc.14.1.3 | Supplementary: http://doi.org/10.5531/cbc.ncep.0188

ABSTRACT

In this exercise, you will have the opportunity to go outside to practice and refine your observation skills and develop strategies for maintaining a field journal. In the first part of the exercise, a guided practice with your class will lead you from noticing, observing, and recording the natural world, to expanding and practicing new strategies for these skills. Following your guided practice, you will use your newly refined skills to complete independent field journaling practice for one or more 30-minute sessions (as indicated by your instructor). A wrap-up reflection assignment asks you to use your field journal observations to make connections to conservation biology concepts and course themes.

LEARNING OBJECTIVES

After this exercise, students will be able to:

1. Produce field journal entries that demonstrate deep observation and attention to detail through the use of multiple recording strategies including words, images, quantification, and more.
2. Recognize how attentive observation is the basis for understanding species, habitats, and processes in ways that allow for the development of meaningful questions and inferences.
3. Identify if and/or how attentive observation and effective recording of observations can create a deeper connection to place.
4. Reflect on recorded observations and analyze the connections between their observations and key concepts in conservation biology, and other ways of knowing about the natural world.

INTRODUCTION

Field journals and lab notebooks are the primary sources scientists create to document their work and observations of the natural world. Close, careful observation and the documentation of those observations are skills foundational to almost every other activity in science. They provide the basis for critical understanding of similarities and differences (among individuals, species, or habitats), documenting changes over space and time, developing meaningful questions and inferences about phenomena, and testing hypotheses to explain what is observed. The skills of careful observation and systematic documentation need to be practiced and refined to be effective (Nevle & Cina, 2020).

While the specific content and style of lab or field journals may vary by discipline, project, course, or research group, documenting observations typically includes engaging in multiple strategies: using words, numbers and formulas, diagrams, maps, sketches, and more (Rutkoski et al., 2023). Observations should include information we discern from any of our senses, and recording them benefits from using not only words, but a variety of visual techniques. For example, identifying some insects requires distinguishing vein patterns in wings, an observation more effectively recorded with drawings than with words (Landin, 2015). Describing the pattern of notes in a bird song is often more clearly illustrated with a diagram of frequency changes over time, and the layout of a habitat or vegetation structure can be clearly documented in a map or cross-section. Using visual techniques heightens our attention to and ability to record important details (Landin, 2015).

The importance of developing skills for keeping a detailed field journal and developing a practice of using observations to enhance data collection has historically played an important role in recording knowledge about the natural world, especially in the European tradition. Many western scientists, artists, authors, and poets have kept journals including Carl Linnaeus, Charles Darwin, Rachel Carson, and Aldo Leopold. Some of these are still informing conservation science today. For example, phenology observations by Henry David Thoreau and his contemporaries provide a historical comparison for scholars documenting changes in plant leaf-out and flowering as a result of climate change (Ellwood et al., 2022). Charles Darwin wrote about making notes in the field, "Trust nothing to the memory; for the memory becomes a fickle guardian when one interesting object is succeeded by another still more interesting." (Darwin's field journals have been published online, they can be viewed here: http://darwin-online.org.uk/EditorialIntroductions/Chancellor_fieldNotebooks.html. For more examples of field journal pages, see Appendix 1.)

Non-western and Indigenous communities have their own recorded observations, wisdom, and reflections on the natural world in their written and oral traditions. A traditional Japanese calendar from the 1600s, based on an earlier Chinese version, is divided into 72 5-day kō that describe observed and predicted changes in the natural world throughout the year (Medhurst, 2015). Kimmerer (2002) points out that ecological knowledge in the oral and other traditions of Indigenous groups developed from the same basic strategy that western science has long used: systematic observations of nature based on intimate engagement and attentiveness to details of the natural world. Those observations, and the knowledge generated from them represent generations of repeated observations and trial and error experimentation in their local habitats (Schiffman, 2018). Western scientists are just beginning to appreciate the wisdom about the natural world held by native communities around the world, and how that wisdom can inform the questions we ask and the management of native species and natural systems (Robbins, 2018; Schiffman, 2018). We can engage in this shared strategy of learning about the natural world around us by using our field journals to record our deeply attentive, detailed observations.

Many scientific occupations still require documentation in this way, from recording phenology by botanists in the field or in botanical gardens, maintaining a record of specimen collections in museums, daily records of endangered species activity in captive populations, and lab notebooks to document daily research progress. In addition, nurses and doctors make many repeated observations as they care for people, and document those in medical files. Farmers monitor the weather and make observations and documentation of the yield of their crops. Other scientists measure chemicals in our water and air to test for safety and pollution. Bird watchers count eggs in nests and birds visiting feeders. Practice in careful observation, documentation, and reflection will benefit you as a scientist no matter your ultimate career trajectory by developing transferable skills used in future careers (Alkaslassy & O'Day, 2002).

In this exercise, you will be asked to practice a variety of different skills for recording your observations in the field, organizing those recorded observations in regular entries in a field journal, and using your field journal to reflect on your observations and their connections to course themes and conservation concepts. Additionally, you may have the opportunity to reflect on your own preexisiting knowledge and how other cultures develop knowledge about the natural world. Field journals and notebooks are useful not just to the scientists creating them, but for their teams or other scientists now and in the future. For these field journals to be as useful as possible, their contents should be detailed, organized, and document important observations and insights gathered in the process of science.

In your first field journaling experience, you will focus first primarily on noticing and observing the natural world around you, then on recording your observation in words and with drawings. After examining and discussing examples of what full field journals contain, you will be given guidelines and expectations for what your field journal entries should contain for your subsequent observations.

Practice in Recording Observations in the Field

Preparing for Your Observations

For this exercise you will be spending one class or lab period outside with your class at the location of your instructor's choosing. Please make sure you are dressed appropriately, have protection from the weather, and have any other items you may need. You may want a hat, sunscreen, a water bottle, a jacket, bug repellent and clothing you don't mind wearing while sitting on the ground. Make sure you are familiar with poison ivy, poison oak, stinging nettle, or other common irritation-inducing plants or animals from your region so you can avoid them. Depending on where your class goes, you may wish to dress with long sleeves, tuck your pants into socks, and do a quick tick check when you get back home. (For more information on protecting yourself from ticks, see: https://www.lymedisease.org/lyme-basics/ticks/personal-protection/.) If additional accessibility considerations and accommodations are required, be sure to check with your instructor in advance.

For your field notes, at a minimum you will need a notebook and pen or pencil. A dedicated notebook with unlined pages is best. (You can typically find an unlined 5.5 x 8.5 in/14 x 20 cm unlined sketching notebook at local or online retail stores for less than $6.) If you need accommodations such as technological aids for taking notes in class, discuss with your instructor using electronic note-taking aids on a tablet or other device. If you do not need technological accommodations, we recommend minimizing your use of technology to allow you to focus on your observations.

You may also wish to bring with you:
- something to sit on if the ground is wet or damp
- colored pencils or pens
- visual aids for deeper observations (e.g., magnifying lens, binoculars, camera)
- field guides

During your observations, we recommend you turn off notifications on your phone to reduce distractions and only use it as a timekeeping device.

Guided Observation Practice

Step 1: Initial Observations

Your instructor will ask you to find a nearby spot where you can be found and given instructions. You will take 5–7 minutes at your spot to make observations and record them in your field journal. Record observations you make with any of your senses.

After your observation time, the class will reconvene to share what they observed.

Step 2: Focused Specimen Observations & Reflection

Next, your instructor will give you a specific specimen or organism for a "matching game." You will have another 7 minutes to record all your observations. Please use both words and visual representations (sketches, diagrams, etc.), and any quantification you can to capture your

observations. Remember, the goal is not to make "pretty pictures" but to use drawing along with writing to document your observations and capture as many details as possible.

When your 7 minutes are up, each person will place their journal on the ground or on a table with the specimens (scrambled) in the center. By examining the field journal entries and discussing with classmates, you will try to match the specimens to the records in the journals.

Reflect: Looking at others' recorded observations, what strategies for recording observations were especially useful or helpful in identifying individual specimens? What strategies could you include next time to improve your own observations and recorded field notes?

Step 3: Characteristics of Good Observation Journal Entries

Your instructor will share examples of field journal pages from other scientists (Appendix 1; if they haven't already asked you to look these over before the class meets).

With one or two other students, discuss:
- What similarities are there in the ways other people record their observations? What differences are there?
- What was the person trying to record, and how did they do that?
- What do you think are the most effective strategies that different naturalists used in their observations?
- What specific strategies would you use to make the recordings of your observations clearer, more detailed, and complete?

At this point, your instructor will give an overview of the most important components they expect to see in your regular field journal entries. In the next step, you will be making and recording observations for approximately 30 minutes at a time, so your field journal should reflect details that can be captured in this amount of time. These should include the following basic elements, but your instructor may include more requirements beyond these.

Each field journal entry should include at least:
- The date, time (beginning and end), and location of your observations. Typically, this appears in a standard location on the page (e.g., top right corner).
- The weather conditions during your observations.
- Detailed observations reflecting the amount of time you made your observations and demonstrating multiple strategies for recording your observations. This means each field journal entry should include at least two of the following: words (verbal descriptions, labels, etc.), pictures (diagrams, sketches, maps, etc.), and/or numbers (quantification of observations—sizes, numbers, frequencies, etc.). This may also include questions about what you observe or connections you make to knowledge from previous courses or experiences.

Your instructor may also wish to see you recording questions or hypotheses you make about your observations. Keep in mind, you are not expected or required to be able to identify every species and to explain everything that you observe. In fact, the questions that arise as you make observations may be critical to future research and explanation of your observations.

Step 4: Applying Your New Observation Recording Skills

Next, your instructor will give you a longer period of time (20–30 min) to make and record observations following the guidelines they provided for formatting and completing a field journal

entry. Try to incorporate some recording strategies you have not attempted so far. You might:

- Choose a specific plant and spend the full time recording observations on that plant. You could create a simple life-size drawing in your journal. Add in some written details to build a complete description of your plant. Zoom in: pick one plant part—a leaf, a flower or petal, the stem, the fruit—and sketch what you see, then add words and labels, too. Zoom out. Look at everything around your plant. Are there other individuals nearby? Where is the plant growing? Where is it not growing? Is there any evidence of an animal eating or interacting with your plant?
- You might choose to quantify something that catches your attention (e.g., bird or insect sounds, plant or flower types or sizes). Consider how you could measure size, distance, or quantity if you don't have a ruler, field tape, or counter? Can you develop questions about the associations between the quantities you observe and some other environmental variables?
- You might choose to describe things observed through senses other than vision. It may be helpful to use analogies or use sketches/illustrations to describe a sound you hear. Can you describe how something (e.g., a leaf or bark) feels with a lot of detail? Can you describe a smell? How much detail can you include in these descriptions of your observations? Challenge yourself to use multiple different sensory observations on a single specimen.

In the last 2–3 minutes of your time, write down the things that you want to find out more about after your observation session. This can be a good time to use your cellphone to take a picture or use an app (e.g., iNaturalist) to help you identify something unknown.

Independent Observation & Field Journal Practice

Now that you have practiced and refined some different strategies for recording observations in a field journal, you will be assigned some individual, independent practice in adding entries to your field journal. Depending on the assignment given by your instructor, this may be a one-time practice, or an assignment that requires repeated visits to the same location multiple times.

The basic instructions for what to include in a field journal entry given in Steps 3 and 4 above remain the same. Your instructor may add additional requirements or formatting expectations to those instructions. Consider reviewing or even printing and bringing the "Good Practices in Field Journaling" tipsheet in Appendix 2 when you go out for your observation. Additionally, we recommend you review the rubric in Appendix 3 to better understand the expectations of your field journal entry. In addition to reviewing instructions, make sure you have whatever materials you anticipate needing while in the field.

Find a spot in a field, park, woodlot, or other spot where you can safely sit undisturbed. Your instructor may recommend or require you to observe at specific sites. Wherever you observe, make sure it is public access property, or you have explicit permission of the landowner for access. If you feel safest in the outdoors in the presence of others, you can make arrangements with a classmate or trusted friend to spend this time at your observation site with the understanding that your 30-minute observation period will not be a time for interaction with other people, but dedicated to your observations.

When You Arrive at Your Site

Walk quietly to the spot you have chosen to create the least amount of disturbance. Make yourself comfortable (but not too comfortable)—you want to be in a position that you can sit still in for almost 30 minutes. Put down all the materials you have brought with you, and make sure you turn off the volume on your cellphone and any other electronic devices. Your arrival at your sit spot will have caused a disturbance to the biological community you will be observing, so you cannot expect to

make observations right away that will reflect the "normal" conditions of this place. As a result, you should first sit very still for about 3 minutes to allow the community to return to "normal."

Open a page in your notebook and make sure to write the date, time, and a description of the weather. If this is your first time observing at this location, you should spend some time just describing it—are you in a field or a forest, a public park? near a trail, a sidewalk, or a road? You could also create a map or diagram illustrating your surroundings.

Spend the next ~25 minutes sitting very still and just observing everything that is going on around you in the natural environment—animals, plants, sounds, wind... use all your senses—what do you see, hear, smell, feel? Take notes, draw sketches, count or measure, but use most of your time to observe.

You will have to show real evidence of careful, close, detailed observation in your field journal entry. As such, after the first few minutes, you may want to focus your observations on some specific feature or organism. Be sure to use the full 25 minutes. (Note: To get the most out of this time, try to refrain from using phones/electronics if you are not using those to record your observations during this time frame. These electronics provide so many sources of distractions that they might make us miss important details or become less attuned to our environment. Instead, use a watch or your phone only to mark the time).

After your observation time is up, use the next 5–10 minutes to finish up writing notes, perhaps standing up and investigating something closer that you couldn't observe easily from where you were sitting. This is also the time to reflect a little on your observations and make notes about questions that remain for you, connections between what you've observed and this or other courses. You could also take a photo with your phone or use an app (e.g., iNaturalist) to aid in identifying something you observed, or to add to your field journal entry later.

Field Journal Wrap-Up Reflection Assignment

After completing one or more field journal entries (as determined by your instructor), you will use your field journal entries to inform a reflective writing assignment.

Look back over your observations from the past few weeks, think about what you have studied in class or your preexisting knowledge/experiences. Your instructor may assign one (or more) of the reflection prompts, or you may have the opportunity to choose. Below are just a few examples; the exercise notes in the supplementary materials will provide the instructor with additional prompts to consider, or they might create theirown.

Example prompts:

- Identify two things that you learned (about the place you observed and the species there, about yourself, about the role of observation in science, etc.).
- Consider how you now understand a biological concept or conservation differently than you did before.
- What do you now see that you had not seen before (complexities, subtleties, new dimensions, etc.)?
- Was there anything that surprised, excited, inspired, frustrated, or disappointed you about what you learned?

Make sure you make specific reference to observations from your field journal in your reflection. Your

reflection should be 400–700 words (about 2 pages) and will be assessed based on the rubric found in the Appendix 4.

REFERENCES

Alkaslassy, E., & O'Day, T. (2002). Linking art and science with a drawing class. Bioscene, 28(2), 7–14.

Ellwood, E. R., Gallinat, A. S., McDonough MacKenzie, C., Miller, T., Miller-Rushing, A. J., Polgar, C., & Primack, R. B. (2022). Plant and bird phenology and plant occurrence from 1851 to 2020 (non-continuous) in Thoreau's Concord, Massachusetts. Ecology, 103(5), e3646. https://doi.org/10.1002/ecy.3646

Kimmerer, R. W. (2002). Weaving traditional ecological knowledge into biological education: A call to action. BioScience, 52(5), 432–438. https://doi.org/10.1641/0006-3568(2002)052[0432:WTEKIB]2.0.CO;2

Landin, J. (2015). Rediscovering the forgotten benefits of drawing. Scientific American. https://www.scientificamerican.com/blog/symbiartic/rediscovering-the-forgotten-benefits-of-drawing/

Medhurst, R. (2015). Japan's 72 microseasons. Nippon.com. https://www.nippon.com/en/features/h00124/

Nevle, R. J., & Cina, S. (2020). Drawing inspiration in the Eastern Sierra Nevada. Journal of Natural History Education and Experience, 14, 18–21.

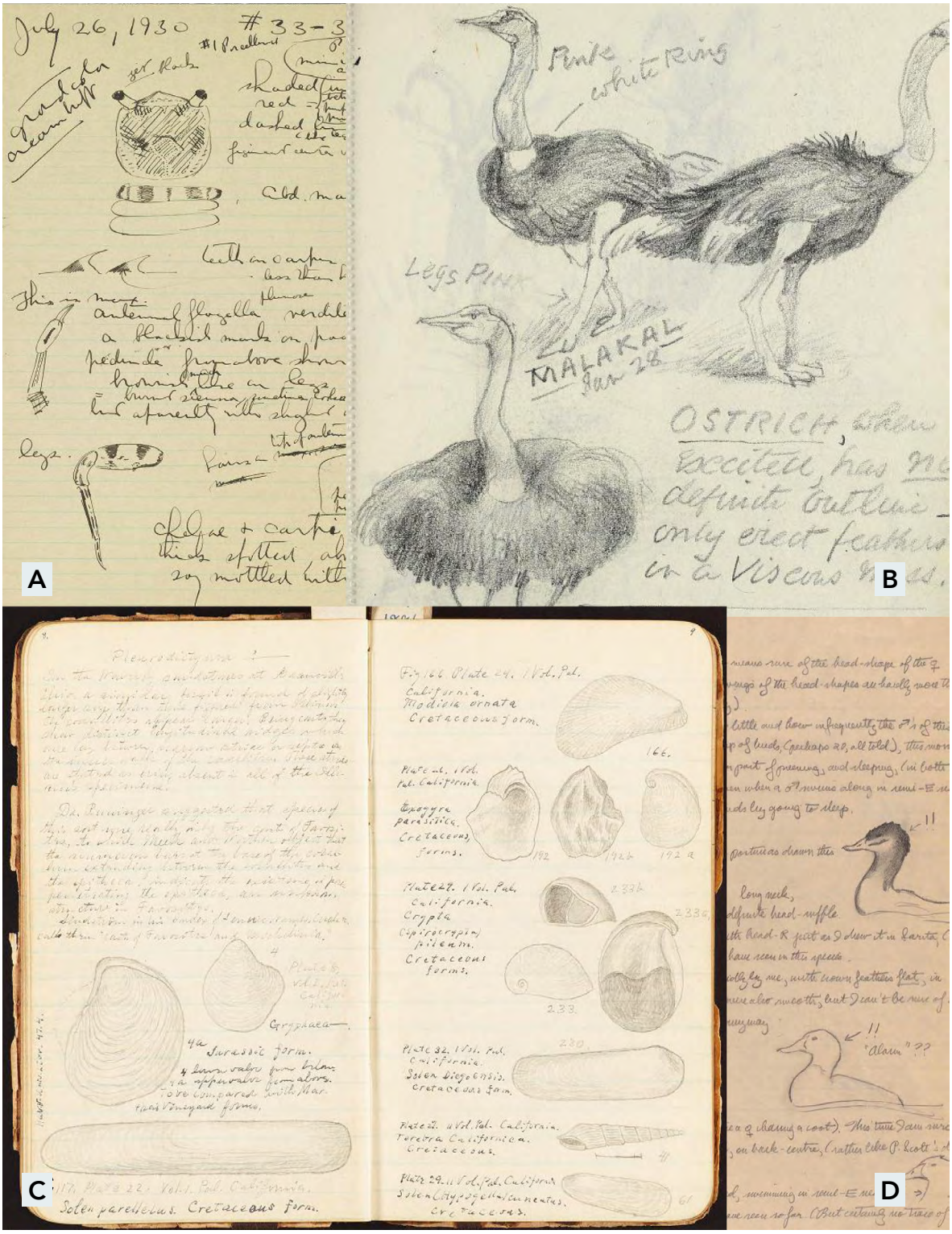
Robbins, J. (2018). Native knowledge: What ecologists are learning from Indigenous people. Yale Environment 360. https://e360.yale.edu/features/native-knowledge-what-ecologists-are-learning-from-indigenous-people

Rutkoski, C., Scamehorn, T., Waterman, R., Klotz, M., & Zettlemoyer, M. A. (2023). Nature journaling: Sharing perspectives between art and science. Plant Science Bulletin, 69(3), 219–220.

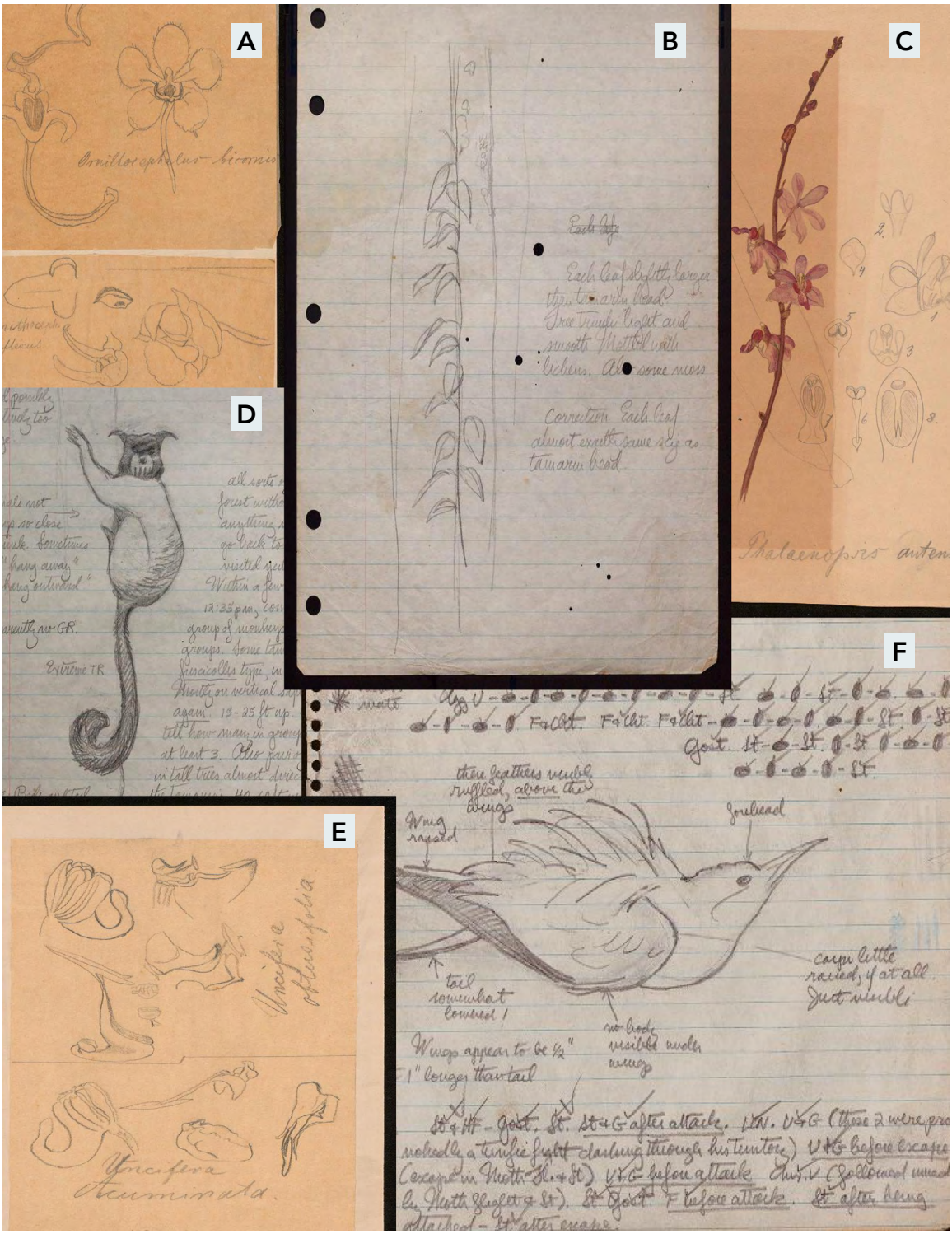
Schiffman, R. (2018). Lessons learned from centuries of Indigenous forest management. Yale Environment 360. https://e360.yale.edu/features/lessons-learned-from-centuries-of-indigenous-forest-management

APPENDIX 1

Sample Field Journal Pages



(A) A page from Waldo L. Schmitt's journals, a carcinologist studying crustacean fauna in Tortugas, Florida. (B) Abel Chapman's observations on birds while on a trip to Sudan. (C) This is a page from August Foerste's specimen notebook. (D) These notes document M. Moynihan's observations on the behavior of ducks in Peru. Image credit: (A) Waldo L. Schmitt Papers, "color notes, Tortugas, July-August 1930" via Smithsonian Institution Archives. (B) Abel Chapman's Sudan sketchbooks via Smithsonian Institution Archives. (C) August Foerste, "Specimen notebook, Ohio, 1887–1888" via Smithsonian Institution Archives. (D) M. Moynihan, "Anatidae (ducks), South America, Peru, Tierra del Fuego" via Smithsonian Institution Archives.



(A, C, E) Observations of different parts of an orchid. (B, D) These illustrations are part of a field journal dedicated to observing monkeys in Panama. (F) These notes document the behavioral observations of gulls in South America. Image credit: (A, C, E) "Descriptions of orchid genera [manuscript], 1880–1908" via Chicago Botanic Garden Lenhardt Library. (B, D) M. Moynihan, "Miscellaneous Monkeys, Barro Colorado Island and Puerto Armuelles, Panama" via Smithsonian Institution Archives. (F) M. Moynihan, "Gull notes (2 of 3)" via Smithsonian Institution Archives.

In addition to those above, many examples of field journal pages can be found online. The Field Book Project of the Smithsonian Institution Archives is a great resource with many digitized field books (https://siarchives.si.edu/about/field-book-project), and Darwin Online has digitized all of Darwin's field books, papers and manuscripts (https://darwin-online.org.uk/EditorialIntroductions/vanWyhe_Manuscripts.html).

APPENDIX 2

Good Practices in Field Journaling

Each field journal entry should include at least:

- The date, time (beginning and end), and location of your observations. Typically, this appears in a standard location on the page (e.g., top right corner).
- The weather conditions during your observations.
- Detailed observations reflecting the amount of time you made your observations, and demonstrating multiple strategies for recording your observations. This means each field journal entry should include at least two of the following: words (verbal descriptions, labels), pictures (diagrams, sketches, maps, etc.), and or numbers (quantification of observations—sizes, numbers, frequencies, etc.).
- Include questions about what you observe or connections you make to your preexisting knowledge from courses or experiences. Keep in mind, you are not expected or required to be able to identify every species and to explain everything that you observe. In fact, the questions that arise as you make observations may be critical to future research and explanation of your observations.

If you get stuck, and can't figure out what to observe and record, whether at the beginning of your observation period, or sometime in the middle, consider adopting one of the following strategies:

- choose one thing and really observe it deeply by zooming in on details, zooming out on context (the habitat/landscape)
- make observations from multiple, different angles
- look for patterns in structure, morphology, or sequences of events
- describe movement
- make connections between observations in the field, or with ideas/concepts
- quantify your observations (e.g., measure with a pencil eraser or count items)
- use more than one of your senses
- use multiple recording strategies (diagramming, mapping, quantifying, describing with analogies).

APPENDIX 3

Assessment Rubric for Field Journal Entries

Criteria	Excellent	Competent	Developing
Includes basic organizational information for each observation	Clearly identifies location, date, time, and weather for each observation. Uniform formatting.	Some ambiguous entries without clear basic information present.	Disorganized and unclear which observations were made when/where, or the conditions.
Includes observations and some interpretation of what is observed	Contains numerous continuous, detailed observations. Notes changes observed over time. Text clearly distinguishes observation from interpretation/ analysis.	Notebook contains some observations, not as detailed or numerous for a ~30 min observation period, and lacks some interpretation.	Multiple weeks of observations missing. Very little (if any) detail. No interpretation of observations.
Includes multiple recording strategies (e.g., written descriptions and sketches of observations, quantification, analogies)	Notebook includes ample examples of observations recorded using multiple strategies (e.g., visual and verbal descriptions). Drawings and diagrams used throughout.	Many entries use only one form of recording.	One mode or another (visual or verbal) largely absent from Notebook.

APPENDIX 4

Assessment Rubric for Reflection Assignments

Criteria	Excellent	Competent	Developing	Needs improvement
Inclusion of Journaling Observations	Observations used in the reflection are specific, carefully chosen, and compelling, consistently showing close observation.	Observations enhance discussion, though not consistently selective or clear.	Observations minimal, with descriptions needing focus or development.	Observations minimal and unconvincing.
Associations: connections to readings/ class/ preexisting knowledge or experiences	Insightful connection of knowledge to observations; descriptions of connections clear and compelling. Evidence of close reading of course materials or thoughtful connections with preexisting knowledge.	Connections of knowledge to observations are effective but not consistently insightful. Descriptions of connections are generally clear but some may need focus or development. Evidence of care and attention for the interpretation of preexisting knowledge.	Connections of knowledge to observations are perfunctory. Descriptions of connections generally need focus or development. Evidence of effort in interpreting preexisting knowledge.	Connections of readings to observations are perfunctory and minimal. Some connections off-point, suggesting limited reading or understanding of concepts.
Reflection: impacts on learning	Exceptionally reflective and thought-provoking analysis of learning or change in thinking as a result of observations & associations visit.	Generally thought-provoking analysis of learning or change in thinking, but sometimes needs focus or development.	Some effort to describe learning or change in understanding, but not explicit, not articulate, or not consistent.	Reflective aspects to writing are minimal or often absent.

Oysters in the City: Analyzing Data to Guide Coastal Restoration in New York City

J. Stephen Gosnell[i,ii] and K. Schreiber[iii]

[i]*Department of Natural Sciences, Baruch College CUNY, New York, NY, USA;* [ii]*Department of Biology, The Graduate Center of the City University of New York, New York, NY, USA;* [iii]*Billion Oyster Project, New York, NY, USA*

DOI: https://doi.org/10.5531/cbc.linc.14.1.4 | Supplementary: http://doi.org/10.5531/cbc.ncep.0186

ABSTRACT

People may picture restoration ecologists working in the field, but in reality, much of their time is spent analyzing collected data in order to communicate progress to stakeholders and guide adaptive management strategies. Data must be summarized numerically and visually to provide information. In this exercise, students will explore these techniques by connecting visual and numerical summaries of oyster restoration progress that Billion Oyster Project provided in their annual monitoring reports to raw data stored in spreadsheets. Students will learn how to summarize data on their own using spreadsheet-style programs.

LEARNING OBJECTIVES

After completing this exercise, students will be able to:

1. Define restoration ecology and discuss the steps needed to restore or supplement populations.
2. Define adaptive management and explain its importance to restoration ecology.
3. Describe the importance of maintaining and storing data for restoration projects.
4. Discuss why oysters are the focus of multiple restoration efforts.
5. Review commonly used summary statistics and use built-in functions to produce numerical and visual summaries of data that has been saved in a spreadsheet-style program.
6. Discuss how summarized data may offer insight on restoration progress and future directions.

INTRODUCTION

In this exercise, you will learn about the field of restoration ecology, and some of the key steps in restoration work: data collection, management, and communication of key results. You will then explore these techniques by connecting visual and numerical summaries of oyster restoration outcomes in New York City.

Ecological restoration aims to recreate, initiate, or accelerate the recovery of an ecosystem that has been disturbed. Disturbances are environmental changes that alter ecosystem structure and function. Common disturbances include logging, damming rivers, intense grazing, hurricanes, floods, and fires. Restoration activities may be designed to replicate a pre-disturbance ecosystem or to create a new ecosystem where it had not previously occurred. Restoration ecology is the scientific study of repairing disturbed ecosystems through human intervention (Vaughn et al., 2010).

Efforts to restore natural populations, communities, or ecosystems require immense amounts of planning. Focal sites must be chosen from multiple candidate options based on the likelihood or necessity of restoration success. For populations that are still present in the wild, methods must be considered to support them through direct supplementation (of food and/or individuals) and/or environmental modification (e.g., invasive species removal, increased habitat protection). If organisms are to be released into the area to promote growth of target populations, they must be selected from

other extant populations or captive breeding facilities. If organisms from captive breeding facilities are used, care must be taken to ensure the released organisms have the skills needed to survive in the wild while also considering impacts of newly introduced genetic variation. Similar concerns arise if the focal organisms are no longer present in the selected area (i.e., they are extirpated there).

After decisions are made, changes in focal populations and their environment must be monitored so that restoration plans can be updated as needed; the use of incoming information to assess and modify existing plans is known as adaptive management. Often these monitoring efforts are undertaken with the involvement of local stakeholders, including government agencies and groups impacted by or involved in the restoration attempt. Data generated from monitoring activities may also be shared among stakeholders; some groups, including funding and government agencies, may require project updates.

From selecting sites and organisms to monitoring changes in each over time, each of these steps requires the collection and analysis of data. Data are collections of information or facts, such as measurements, that are collected and recorded and can be used for reference, comparison, or decision-making. Data are essential to restoration programs because they can be analyzed to discover patterns and trends. This analysis often includes summarizing data visually or numerically to aid in interpretation.

Restoration ecologists collect a wide variety of data to understand the abundance and distribution of organisms at sites and their relationships with each other and the local environments. While data are typically collected from individual organisms (e.g., individual size) or locations (e.g., a sampling site), scientists often can't measure every single organism in an entire population or every spot in a restoration site. As a result, scientists collect and summarize data from a sample of individuals or sites and use it to estimate traits (e.g., average height, average temperature) about the focal population or sites. The science of statistics then allows scientists to determine how robust or accurate their estimates are, and to make inferences about relationships among multiple variables. Here we describe a project in New York City that is addressing these issues as they relate to the restoration of oysters.

OYSTERS IN NEW YORK

Eastern oysters (*Crassostrea virginica*) were once a common species in the waters of New York State, with estimates suggesting oyster reefs covered thousands of acres (Kurlansky, 2006; Zu Ermgassen et al., 2012). Oysters provided a range of benefits to both humans and other organisms. Some of these benefits were direct, such as serving as a valuable food resource to Indigenous communities, European settlements, and eventually growing cities (Zarnoch & Schreibman, 2012). Other benefits were indirect. For example, as oysters filtered particles from the water for feeding, they also improved water quality and stored carbon in shells and soft tissue (Dehon, 2010; Lemasson et al., 2017). The resulting feces (waste products following digestion) and pseudofeces (particles oysters filtered from the water but did not digest) oysters deposited on local sediments provided important resources to sediment microbes that play a key role in denitrification, or the removal of excess nitrogen that can cause eutrophication and other imbalances in an ecosystem (Kellogg et al., 2013; Rose et al., 2021). The dense beds, or reefs, formed by oysters also acted as living breakwaters, reducing the impacts of storms (Dehon, 2010; Brandon et al., 2016; Zarnoch & Schreibman, 2012), and provided habitat for other species, including the juvenile phases of many ecologically and economically important organisms such as blue crabs, shrimp, and fish (Bruno et al., 2003; Levinton & Waldman, 2011).

Both the direct and indirect benefits of oyster activities can be called ecosystem services. Defining and valuing these benefits may help motivate restoration efforts. For example, many of these benefits were lost as over-harvesting, habitat loss, and habitat degradation from pollution and shoreline development led to large declines in oyster populations (Zarnoch & Schreibman, 2012). Harvesting increased to unsustainable levels as the human population of New York City grew, and the ability of the system to support oysters simultaneously decreased due to dredging and other activities to develop shipping lanes and ports, siltation and excess sediment buildup, and loss of water quality associated with population growth (MacKenzie et al., 1997). In fact, the population of oysters in New York's Hudson River Estuary was deemed "ecologically extinct" in 2007 since the remaining population was too small to provide these key benefits (Eastern Oyster Biological Review Team, 2007) and was even considered for listing under the Endangered Species Act in the early 2000's. The loss of oysters in the waters in and around NYC matches global trends, as analyses suggest more than 85% of the world's oyster reefs have been lost (Grabowski et al., 2012).

While oyster populations have been negatively affected by many of these stressors, re-established populations may also help address some of them. For example, the waters of New York City are some of the most nitrogen-loaded in the world, largely due to wastewater input. Although nitrogen input from wastewater treatment plants in Jamaica Bay peaked in 1995 (Benotti et al., 2007; City of New York Department of Environmental Protection, 2020), estimated loads in the bay and Hudson River Estuary remain high (Howarth et al., 2006; Rosenzweig et al., 2018). Nutrient input at high levels may be detrimental to oysters and may impact local fisheries if it is associated with wastewater input. This can lead to a loss of oyster populations and resulting loss of associated ecosystem services such as water filtration and denitrification that improve water quality. However, restored oyster populations can help reduce nitrogen and other water quality problems in local waters (Ayvazian et al., 2021; Parker & Bricker, 2020; Rose et al., 2021) while also reducing storm impacts (Dehon, 2010; Fears, 2016; Scyphers et al., 2011) and providing other lost services (Ayvazian et al., 2020; Smith & Pruett, 2024).

Decreases in wastewater input and other positive changes have improved water quality to the point that oyster restoration is now possible in the area (Zarnoch & Schreibman, 2012). Given these changes and desire to restore ecosystem services associated with oysters, major investments in oyster restoration and habitat monitoring are now underway in New York City. Restoration efforts in the Hudson-Raritan Estuary (HRE) of New York City have been led by Billion Oyster Project (BOP), NY/NJ Baykeeper, New York City Parks, and the Hudson River Foundation along with various federal agencies and local non-profits. Since oysters were largely extirpated from the waters around New York City, these groups have worked to supplement existing populations and develop new ones by releasing oysters into local waters. (For more information on the history and current restoration activities focused on oysters and other habitat-creating organisms, see a BOP story map here: https://storymaps.arcgis.com/stories/0713dad90fe84718add43c5839cade5a.)

In order to restore local oyster populations, these groups need to rear oysters in captivity. Fortunately, oysters have been the focus of many aquaculture efforts, meaning restoration efforts could borrow practices used to produce oysters for food. In natural conditions, male and female oysters release gametes (eggs and sperm) as temperatures rise (Eastern Oyster Biological Review Team, 2007; Kennedy et al., 1996). The release of gametes by one oyster may also trigger other nearby oysters to release gametes. When these gametes meet in the water, fertilization occurs. The resulting larvae are planktonic, meaning their movement is largely directed by tides and currents. These larval oysters feed, grow, and develop for several weeks, including developing an initial shell, before beginning to search for a surface to affix to. Once a suitable location is found (which could be an adult oyster shell

or other hard surface), larval oysters attach to it by secreting an adhesive. Once attached, oysters undergo further metamorphosis, losing larval organs and becoming sedentary organisms; newly attached oysters are referred to as spat. These spat will grow, becoming juveniles and eventually adults that may contribute to future generations.

In restoration, oysters are produced in captivity by emulating these conditions. Adult oysters are collected for use as broodstock in a hatchery. For example, BOP operates a hatchery located on Governor's Island in New York City. Hatcheries may use healthy adult oysters from remnant populations since they may have adaptations to help them survive in local waters (Burford et al., 2014; Eierman & Hare, 2013). These broodstock are exposed to increased temperatures and possibly pre-harvested gametes to induce gamete release. Resulting larvae are transferred to tanks where water flow is managed to keep them in motion and food is continually introduced. Once larvae are ready to settle, they are introduced to setting tanks that contain shell or other material so they can attach and metamorphose into spat. The new oysters are then allowed to grow in protected conditions in a nursery until they are ready for use in restoration projects (Figure 1A). Oyster growth and survival at these nurseries are monitored to ensure oysters are healthy and to determine when they are ready to be transferred to restoration sites.

Depending on the site and needs, oysters are eventually released onto submerged sediments (possibly supplementing local populations and forming reefs) or placed in cages or other devices which are then transferred to restoration sites (Figure 1B). These efforts have helped to raise and set nursery-raised oysters throughout the estuary (Schiff, 2016; Schmidt, 2016).

Figure 1. Examples of structures that hold oysters and provide data via monitoring. (A) Racks or cages provide a protected nursery area where juvenile oysters grow following their time in a hatchery. These oysters are eventually used in reefs (B), where oysters are placed in large quantities in various structures to provide physical structure and habitat for other organisms, much like a natural oyster reef would do. Some oysters are also moved from nurseries to (C) oyster research stations. Image credit: Billion Oyster Project.

Once transplanted, oysters must survive in order to provide ecosystem services, stabilize areas, and provide a starting point for future reefs. Transplanted oysters are thus monitored for growth and survival. Monitoring growth and survival allows ecologists to assess the progress of a cohort or site. Different factors may impact both growth and survival. As noted, different broodstocks, or lineages of oysters, may have different rates of growth or survival due to genetic differences (Burford et al., 2014; Eierman & Hare, 2013). Hatchery practices, such as feeding (McFarland et al., 2020) or exposure to predator cues (Belgrad et al., 2021) may also impact outcomes. Oysters may also have different outcomes depending on the structure used to hold them in restoration sites (Grizzle et al., 2024).

BOP commonly measures the shell length of 25 oysters in each structure type at a site. Why only 25 and not all the oysters? As noted above, measuring all the oysters is typically not possible due to time or access constraints. Thousands of oyster spat may be placed at a given site; for BOP, measuring all the oysters would take a very, very long time. Some structures may also restrict scientists from measuring all transplanted oysters (for example, oysters in interiors of structures such as in image Figure 1B). In other situations (think about measuring wild animals in a population), it's hard to know if you measured every organism. However, scientists can use the sampled oysters to estimate how the population of oysters is faring. A good sample thus balances the ability and time needed to collect the data with ensuring the sample is large enough to actually allow extrapolations to the larger population.

Oysters must also reproduce in order for populations to reach the densities needed to provide ecologically significant effects or become self-sustaining, or able to persist without continual human intervention. Oysters are monitored for reproductive potential, and existing restoration sites and other locations may also be monitored for the presence of newly-settled oysters, or natural recruits.

Salinity and temperature play a key role in impacting oyster outcomes, as does the availability of oxygen (Kennedy et al., 1996). For these reasons, environmental data on these and other parameters (e.g., pH) may be collected as well. This may occur during measurement events or through the use of long-term monitoring devices installed at sites. When environmental data are not directly connected to oyster measurements or sites, it may be determined from other data sources.

Some of this data collection takes place at community reefs, where community scientists play an active role in data collection. For a related module on community science, see the NCEP module "What is community science, and how do I get involved?"(https://doi.org/10.5531/cbc.ncep.0002; Gosnell, 2023). Community scientists also contribute to data collection through deploying and monitoring oyster research stations (Figure 1C). Oyster research stations are small cages of oysters that are placed in local waters by schools or community groups. (For a video of monitoring protocols at oyster research stations, see "Oyster Research Station: Oyster Measurements and Mortality Protocol 2022" here: https://www.youtube.com/watch?v=436RtHVdyew.) Volunteers monitor the oysters at these sites on a regular basis, collecting data on oyster growth and mortality that are then reported to BOP. (You can view an interactive map of these various sites at "Harbor Oysters Map" here: https://www.google.com/maps/d/u/0/viewer?hl=en&ll=40.69341829153447%2C-73.92496718051235&z=11&mid=1EUU_w0PW05qaZxh7y0cb1FD-xd6HSo7u.)

All this monitoring activity provides data that can be used to inform restoration activities. BOP uses this information to determine which hatchery practices are producing successful oysters and which sites are most promising for future expansion and, by considering relationships among environmental factors and oyster outcomes, to identify new locations for restoration work. In order to use these data, however, they must be carefully collected and cataloged. For example, information such as broodstock and set

date (time when oysters are moved to nurseries or reefs) must be able to be matched to oysters that are in the water. Data must then be analyzed to provide insight on restoration activities.

In the following exercise, students consider these issues by reviewing output from BOP's 2018 annual report and the data that led to its findings. They also learn the basics of manipulating data in spreadsheet format and producing summaries of univariate data. In this lesson, we will give examples using Google Sheets, a free online spreadsheet program. Most of the information will easily transfer to other common programs such as Microsoft Excel, though some commands may differ slightly. A more advanced approach to summarizing and analyzing data might include specialized software such as R or Python.

EXERCISE

Data on oyster outcomes and environmental data are collected and analyzed by BOP scientists each year. Information is used to inform restoration activities, but it is also used to update the public and permit-granting agencies on outcomes. These updates play a major role in enabling future work to continue. Permits are required for many restoration activities, and public support may contribute to the approval of permits and also supply resources (e.g., funding, volunteers) to restoration efforts.

Each year BOP produces an annual report for the public concerning its progress. Reports are viewable on the BOP website here: https://www.billionoysterproject.org/research.

For this exercise, we will focus on the report from 2018: "New York City Oyster Monitoring Report 2018," which can be downloaded as a supplementary material with this module under the file name (TNC_BOP_2018_report.pdf). Reading the report is optional, but reviewing the overview and main findings (pages 1–7) is recommended to help you orient yourself to the group and its main findings in the area. Examples of summarized data on oyster outcomes are found on pages 15–42 of the annual report. For this activity we focus on oyster size, or shell height. (Note, Figure 9 from the report is recreated in Figure 2.)

This graph displays the average size over time of oysters at various sites growing in various structures, but what does that actually mean?

From Data to Summaries

These summaries are produced from the raw data collected during monitoring activities. You can view a curated portion of the available data on oyster growth and mortality for each site type (nursery, reef, oyster research station (ORS)), along with more info about these sites, in the zipped file provided as a supplementary material with this module under the file name (Oyster_data.zip). These files are the output from spreadsheet programs. The following instructions will focus on Google Sheets, a free program that may be run in web browsers. To use these files in Google Sheets, you must unzip the folder and upload the files to Google Drive.

Other spreadsheet programs (e.g., Microsoft Excel) can be used with small modifications. Entering the data in these programs enables long-term digital storage of collected information. First, note that each file produced by these programs may contain multiple spreadsheets under various tabs (Figure 3). First, view the data tabs (e.g., "Oyster nursery data," "Oyster reef data," "ORS data"). The data tabs clearly show how the spreadsheet programs allow users to input data. Data are entered in cells that are positioned in rows and columns. Each cell holds a single piece of information, and they can

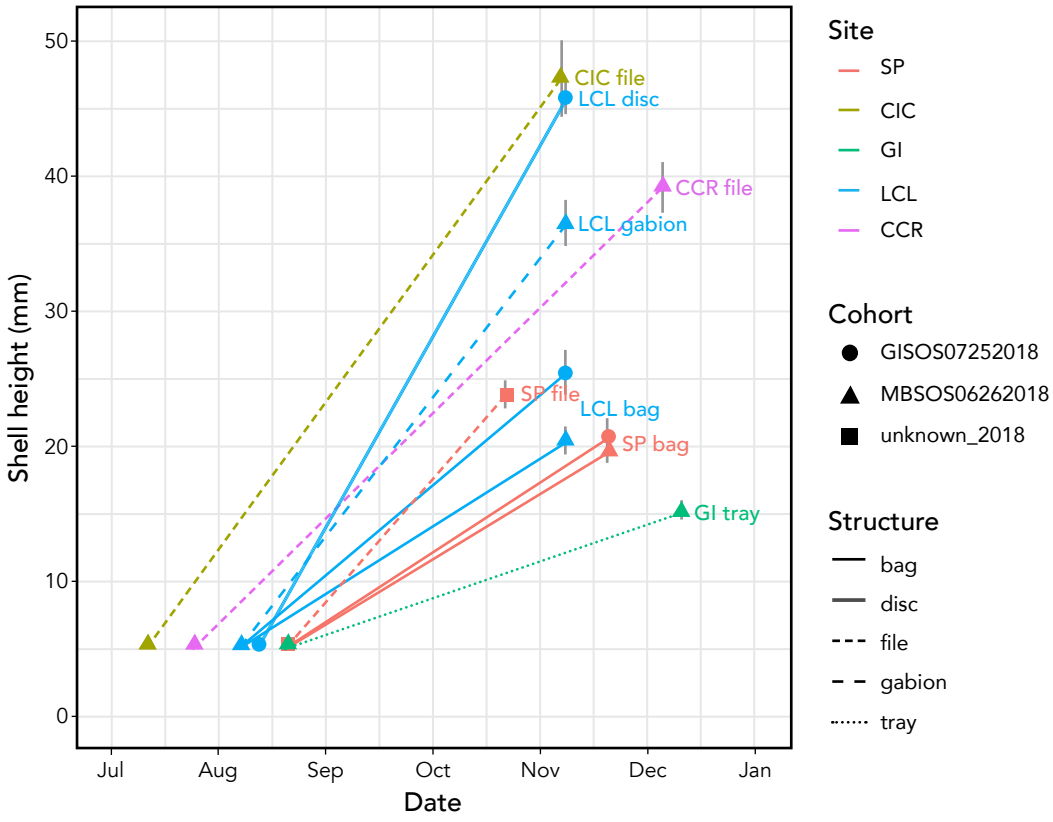


Figure 2. Oyster shell height over time for oysters installed in 2018 at Canarsie Community Reef (CCR), Coney Island Creek (CIC), Governors Island EcoDock (GI), Lemon Creek Lagoon (LCL), and the Sunset Park (SP) Community Reef and bagged shell reef. Error bars indicate 95% confidence intervals. "unknown_2018" cohort is a mix of GISOS07252018 and MBSOS06262018. Image credit: Modified from Burmester & McCann, 2019.

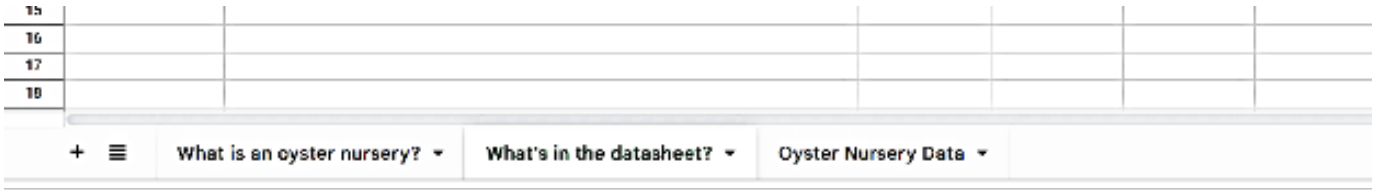


Figure 3. Each workbook focuses on a specific type of restoration site (oyster nursery file shown here) and contains multiple spreadsheets, each in an individual tab. These include background data (What is an oyster nursery), data (Oyster nursery data), and metadata (What's in the datasheet?) tabs.

be arranged so that position has meaning.

For example, for the reef data, the "Oyster reef data" tab each row focuses on a single oyster, and each column contains information on

- Data collection year
- Data collection date
- Site
- Broodstock
- Cohort (which includes broodstock information in addition to history of the oyster in regard to other hatchery protocols)
- Set date
- Installation date
- Structure
- Shell height (mm)
- Survival status

In Google Sheets, rows are identified by number and columns by letters. We will use that convention in this exercise.

Note some of this information is used to identify, or place, the accompanying measured variables. For example, row 2 of the Oyster reef data contains information on the shell height and survival status (measured variables) for an oyster that was measured in 2016 ("data collection column") at LCL ("site column").

1. Why do scientists need to record both identifying and measured variables?

Metadata

Note how many columns are present and the large number of rows in the various data tabs. Column names also differ slightly among sheets. Given the large amounts of provided data, it's critical to let viewers know what they are working with (similar to what was noted above). Metadata fulfills this need. Metadata provides a description of the data set (i.e., data about data). For example, each file contains a tab ("What's in the datasheet?") devoted to metadata. Select one of these metadata tabs (Figure 3). Review the metadata for one of the sheets.

2. Which dataset did you choose?
3. Choose two variables discussed in the metadata. Are these measured variables (data collected during monitoring) or identification (data used to note something about the oyster or site)?
4. Why is metadata important to include alongside actual data?

As noted in the metadata tabs, while data are collected on oyster shell length and survival, they are connected to other pieces of information such as broodstock, location, and structure. This data structure allows long-term analysis to improve future outcomes. For example, analysis of outcomes related to various broodstock may suggest which oyster lineages BOP should focus on reintroducing in the future.

Data

Next, return to one of the tabs containing actual data (e.g., "Oyster nursery data," "Oyster reef data"). These sheets contain entered data. Note, computers are rarely used in the field! This means data that were collected by scientists in the field (Figure 4) must be typed or otherwise entered into these spreadsheets. This step is critical, as mistakes in entering the data will impact the rest of the analysis process.

Once data are entered, spreadsheet programs also allow you to easily manipulate, filter, or summarize the data and potential relationships.

Why do data need to be summarized? As noted above, we want to collect as much data as possible, but even with "only" 25 oysters measured per structure at each site, the amount of data can quickly accumulate. For example, BOP collected data from over 1000 oysters in 2018 from reefs alone. Inferring relationships among or impacts of various factors (e.g., does structure type impact oyster growth or survival) is almost impossible by just viewing the data in a spreadsheet. Summarizing the data can make comparing groups easier and is often followed up by statistical analysis which determines if there are differences among groups or relationships between different variables. The process of analyzing and summarizing data lies at the heart of understanding ecological patterns and relationships.

Scientists and stakeholders involved in oyster restoration efforts may have many questions that they wish to explore. For example, does the structure type or broodstock impact how fast oysters grow? Are

Oyster Measurements: Tagged or untagged shell clumps

This bag of oysters was culled this yea

Site: LCN	Date: 5/10/2018	Cabinet/File/Tray: NA
Recorder: LB	Time: 4:05 PM	

	Each column = 1 shell clump									
Tag #	NA									
Counter Name	LB									
	Measure all live oysters: Shell height (mm)									
1	86									
2	41									
3	50									
4	48									
5	45									
6	49									
7	40									
8	39									
9	61									
10	70									
11	80									
12	49									
13	39									
14	63									
15	72									
16	74									
17	63									
18	61									
19	54									
20	47									
21	61									
22	60									
23										
24										
25										

Figure 4. Example of datasheet completed in the field by BOP scientists. Image credit: Billion Oyster Project.

oysters larger in one site rather than another? We can begin to answer these questions by summarizing the data. Data collected by scientists can be summarized in two main ways: numerical and visual summaries. These summaries may provide information on the distribution of traits in a population.

Numerical Summaries

As an example, let's consider if oysters grow differently at different sites. If we assume oysters that were transplanted to the sites were approximately the same size (which we will for now), we could answer these questions by comparing the current sizes (shell heights) of oysters. Since many oysters may have been measured at any site, we may want to summarize the data.

A common example of a numerical summary, or statistic, is the mean, or average. Given data on oyster shell heights, we could find the mean shell height for each site and compare them. To find the mean shell length for a given structure at a given site, for example, you would sum the 25 individual lengths and divide the result by 25 (the number of oysters sampled) to get the mean.

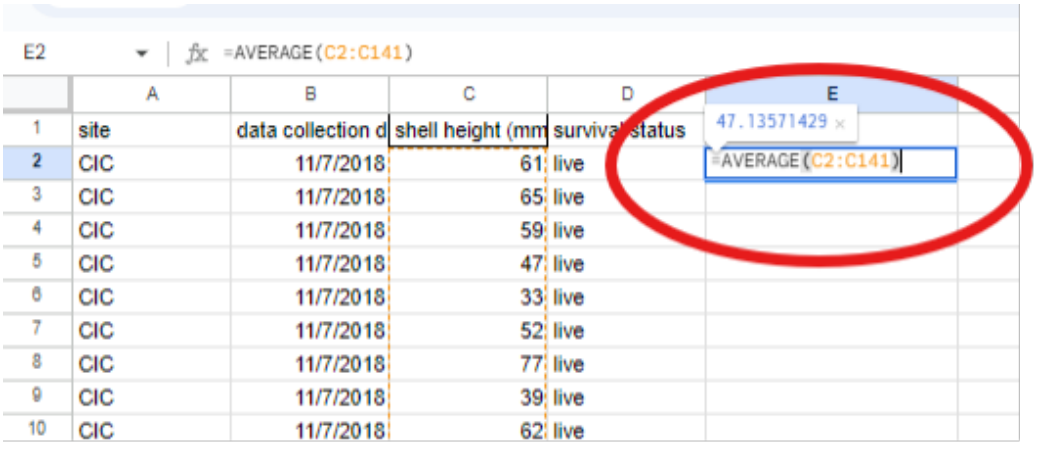
While the mean and other statistics may be calculated by hand, spreadsheets and other computer technology offer a way to quickly carry out large calculations. They do this by using functions to quickly provide a desired output. Functions are also useful as the output will automatically update if the data are changed.

For example, we can calculate the mean of a number of cells in Google Sheets using the average function. To try this, open the Oyster Reefs file and make a copy for your use. You can do this by selecting "File" > "Make a copy" from the top menu. You can also select "File" > "Download" to download a version for use in other programs. Next, go to the "CIC - 11/7/2018 tab". This tab contains the subset of data collected from Coney Island Creek (CIC) on 11/7/2018. Once in the tab, select an empty cell. In the cell, type =AVERAGE(C2:C141) and press enter (Figure 5). What happens?

You just used a function to calculate the average, or mean, shell height for the 140 oysters measured at Coney Island Creek; this data was contained in cells C2 to C141. The "=" sign denotes you are entering a function. The next word is the function you want to use. In this case, the "average" function tells the program to calculate an average. Function names are typically followed by parentheses that contain the arguments, or the input, needed by the function. For example, "C2:C141" indicates the average function is to evaluate the cells found in the C column between rows 2 and 141. When you hit enter, the function will calculate the sum of the cells and the total number of cells. It will then use that information to compute the mean and display that value in the cell. However, if you attempt to edit the cell, you will see it still contains the function. In this case, the cell will display the mean shell height from the CIC site. Note this matches the point associated with the CIC file subset shown in Figure 2 (you can also check Table 3 in the report to verify this). The bars around the point in the figure are confidence intervals, which we will not cover in this lesson.

Other ways to calculate this value include asking for the average of the entire column (input as C:C or by selecting the column header when entering the formula) or by selecting the column and observing summary statistics in the lower right hand of the screen (Figure 6).

5. Now you try it. Select the tab containing data collected from oysters grown on ECOconcrete discs at Lemon Creek Lagoon on 11/8/2018 ("LCL - 11/8/2020 ECOconcrete"). Calculate the mean shell height for oysters collected at this site for this structure by entering the average function in an empty cell in column E (e.g., E2).
6. Imagine there was a mistake made when someone entered the data collected in the field into the spreadsheet. For example, what if cell C2 was entered as 355 instead of 35? Change the value in cell C2 to 355. What happens to the cell containing the average function?



	A	B	C	D	E
1	site	data collection d	shell height (mm	survival status	
2	CIC	11/7/2018	61	live	
3	CIC	11/7/2018	65	live	
4	CIC	11/7/2018	59	live	
5	CIC	11/7/2018	47	live	
6	CIC	11/7/2018	33	live	
7	CIC	11/7/2018	52	live	
8	CIC	11/7/2018	77	live	
9	CIC	11/7/2018	39	live	
10	CIC	11/7/2018	62	live	

Figure 5. Functions may be entered in empty cells.

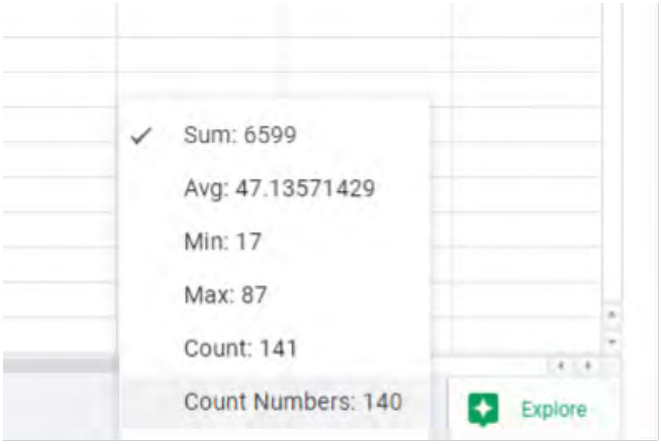


Figure 6. Summary statistics are shown for a selected column or range (group of cells) in the lower right corner of the screen. Note you may need to select the downward pointing arrow icon next to the displayed statistic to see all available statistics.

7. What is the benefit of knowing the average outcome for a trait? What might this value be missing?

Statistics such as the mean are known as measures of central tendency. They help summarize the general response of a group. Other examples of measures of central tendency include the mode and median.

Visual Summaries

Once summaries such as mean shell height are calculated, it may be useful to place them in visual summaries so they can easily be compared among sites (such as in Figure 2). Visual summaries include graphs that use lines and symbols to represent trends in the data. Visual summaries may also include raw (not summarized) data. Just like numerical summaries, these summaries should make general trends and points in the data easier to understand.

Many types of graphs exist. The type of data you are attempting to visualize will often determine the appropriate graph type. While various types of graphs may be used to plot the same data, they each display the data in slightly different ways. For a more thorough review and material on data visualization, which may help you determine the best graph type for your data, we recommend visiting some of the following sites:

- From Data to Viz: https://www.data-to-viz.com/
- VisDepot: An Introductory Resource for Data Visualization: https://necote.github.io/VisDepot/landing.html
- Summarizing data: https://jsgosnell.github.io/cuny_biostats_book/content/chapters/summarizing_data.html

Although different types of graphs exist, all should contain:

- Descriptive title (for stand-alone images) or caption (more common when graphs are used in articles, often at the bottom of the figure)
- Axes labels with units
  - Each axis (side of the graph) should be labeled with what was measured (length, mass, time, etc.) and how it was measured (metric unit)
- Visual representation of the data
  - Points, bars, or other shapes that represent the data

Other options that may be included are legends (useful if you are displaying data from different groups), lines of best fit (trendlines), and confidence intervals.

Here we will first focus on histograms. Histograms are used to visualize variation in a single continuous trait, or outcome variable (like shell height). Variation in measured values is just as important as central tendency. For example, two sites might have the same mean shell height, but one might have oysters that are all approximately the same size, while the other contains many small and many large (relative to the calculated average) oysters. Given oysters and other organisms may need to reach a certain size to be protected from predators or reach sexual maturity, this could be important to restoration efforts. For example, some oysters may become sexually mature as males around 35 mm of age (Harding et al., 2013). Two populations could both have mean sizes around 28 mm, but one may have sexually mature individuals while the other does not.

Distributions of shell height from various sampling points may also be compared to determine if oysters are growing at site. Lack of growth could be due to environmental conditions, such as temperature, or due to the presence of predators. For example, low survival and thus observed growth at Brooklyn Bridge Park has been credited to the presence of oyster drills at the site, which are predators that can consume large numbers of transplanted oysters (McCann, 2018). Standard deviation is another common numerical summary used to consider variation (not discussed here).

To make a histogram, continuous data is first grouped into appropriate bins (in essence, categorizing them). The number of data points in each bin is then calculated. Note bin size or number may be set by the program or user, so the same data may appear slightly different when plotted in different programs or using different bin sizes. The bins are then plotted on the x-(horizontal) axis, and the frequency (or count, sometimes displayed as percentage or proportion of total counts) is plotted on the y-(vertical) axis. Histograms thus provide a visual summary of which outcomes were the most common (highest peaks) and the range of outcomes observed. This can be useful in determining the distribution of a given trait value. For example, histograms are used in the 2018 report to consider variation in shell size at a given location for a given monitoring event (Figure 7).

Histograms also may allow outliers, or points that are extremely large or small relative to the rest of the data set, to be noted. This can be useful in analyzing data for quality and for understanding patterns. For example, if a typo caused a measurement to be entered as 500 mm (instead of 50 mm), this point, and its unusualness compared to other collected data, would be easily noted using a histogram. Similarly, the appearance of smaller oysters in the histogram might suggest oysters are naturally recruiting to the site.

You can create a histogram in Google Sheets by selecting the column with data that you would like to graph and clicking the Insert chart icon (Figure 8). Alternatively, after you select the data, you can select “Insert” > “Chart” from the top menu.

This produces a chart using default options and opens a Chart editor so the chart can be further modified (Figure 9). You may need to select “Histogram” under the “Chart Type” menu. Options under the “Customize” tab allow users to change the title of the graph or axes (“Chart and axis titles” menu; Figure 10).

8. Produce a properly labeled histogram of the data found in the “LCL - 11/8/2018 ECOconcrete” tab. Attach a screenshot of it here.
9. What does the histogram indicate about growth in oysters in Lemon Creek Lagoon (LCL)?
10. How does the visual summary provided by the histogram extend what we can infer about the population compared to information on the mean growth?

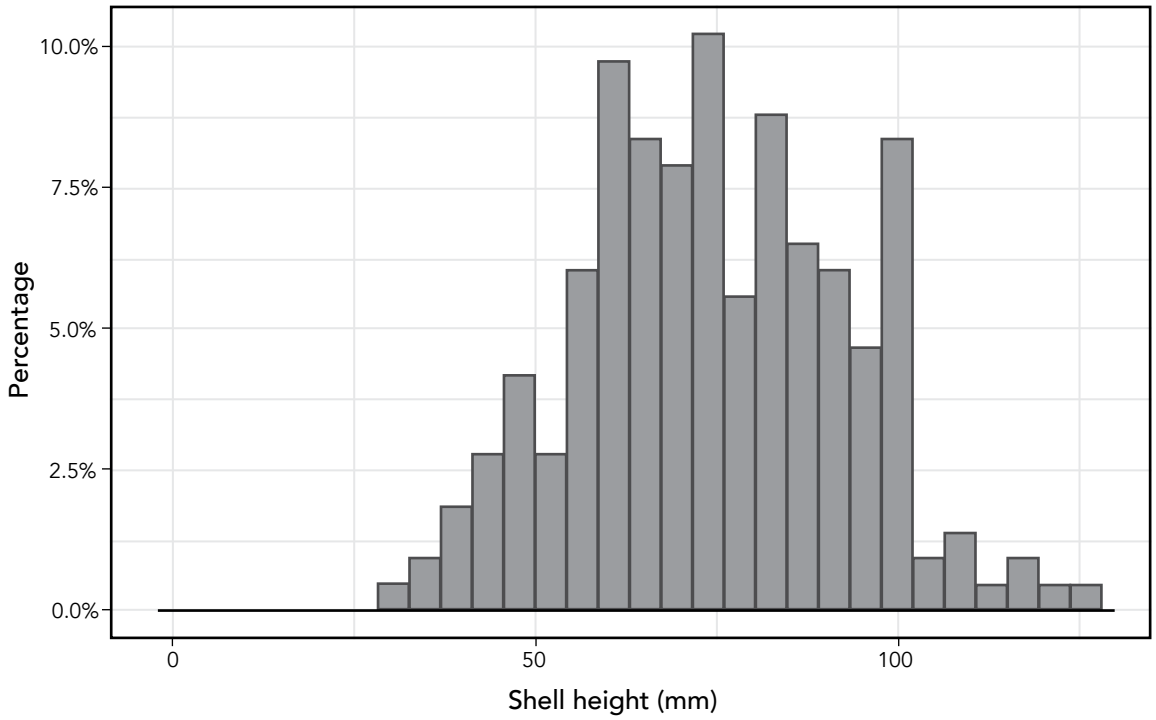


Figure 7. Size frequency distribution of live oyster shell heights at Lemon Creek Nursery in September 2018. Image credit: Modified from Burmester & McCann, 2019.

Figure 8. After selecting data, select the chart icon (red circle) to make a graph.

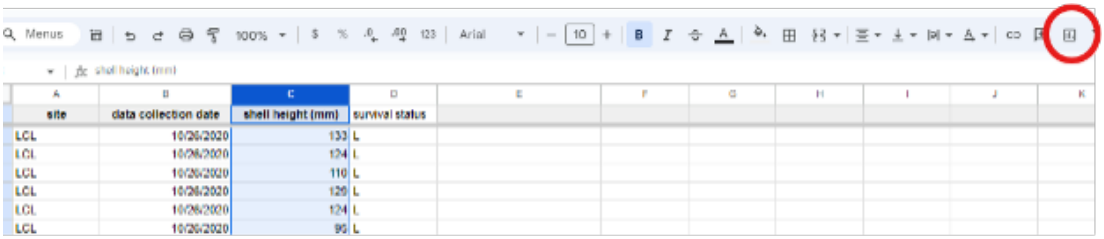


Figure 9. The "Chart editor" allows users to design the chart ("Setup tab") and customize features such as axis labels and position (Customize tab).

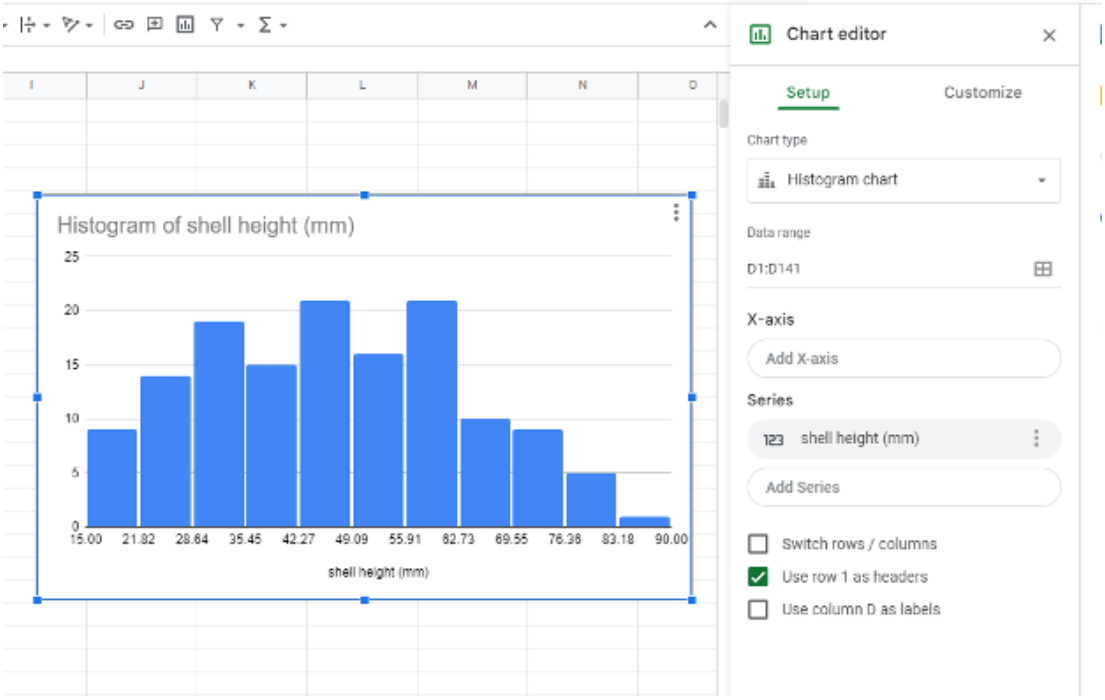
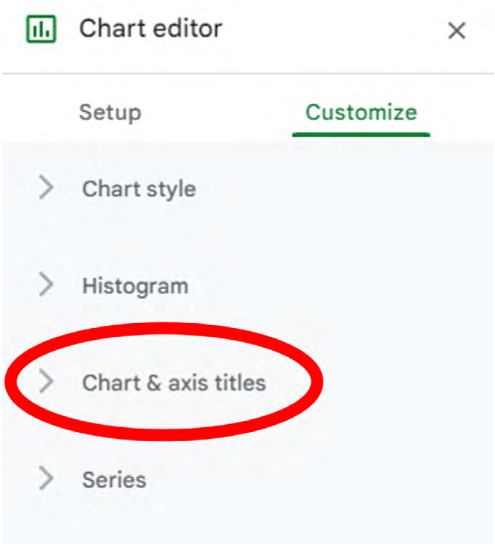


Figure 10. The "Chart & axis titles" menu allows users to change the labels for graphs and axes.



Putting Numerical and Graphical Summaries Together

As previously noted, the histogram gives you more information about the distribution of data than just the mean does. One can also use other numerical summaries to describe how data are distributed. For example, from the histogram of shell heights that we created, we can see which bins the lowest and highest values are found in, but we may not know these actual values due to the binned nature of the data and axis labels (e.g., what is the size of the smallest oyster described in Figure 10?). To find these values, we could sort the data from smallest to largest values. If you do this, the smallest value is the minimum, while the largest is the maximum. The point exactly in the middle is the median. Note that for data sets with an even number of elements (pieces), the median will be the average of the two numbers closest to the middle. The median is another measure of central tendency, like the mean. However, it is less impacted by outliers. Differences between the median and the mean may indicate a dataset has large outliers.

While we could calculate these statistics by hand, we can also calculate them using functions (like we did for the mean). The min function calculates the minimum, the max value calculates the maximum, and the median function calculates the median. To see this, return to the "LCL - 11/8/2018 ECOconcrete" datasheet.

11. Calculate the median for the dataset by using the median function in an empty cell. What is the value?
12. Return the cell you changed to its original value (for example, return cell C2 to 35 from 355). What happens to the cell containing the median function?
13. What are the minimum and maximum values of the dataset? Find them using the min and max functions.

In general, outliers will raise or lower the mean value of the population. This means if a dataset has outliers, the mean value may not be the best measure of central tendency. The median value, however, is not influenced by outliers in the data set.

The minimum, median, and maximum are each components of the common five-number summary for a dataset. The remaining components are the first and third quartile. The first quartile is the data point in the middle between the minimum and median (assuming the data are sorted smallest to

largest); the third quartile is the data point in the middle between the median and the maximum. Subtracting the minimum from the maximum provides the range of a dataset. Subtracting the first quartile from the third quartile produces the interquartile range (sometimes labeled IQR). The IQR is another useful way of describing the spread, or variation, in the measured trait.

After being calculated, the five-number summary may be plotted using a box and whisker plot (also known as a box plot). In Google Sheets, this is known as a candlestick chart. To make a candlestick chart, use functions to complete the five-number summary for the "LCL - 11/8/2018 ECOconcrete" and the "CIC - 11/7/2018" datasets. To compute the first and third quartile, use the quartile function. Unlike the other functions we have learned, it requires two arguments that are separated by a comma. The first is the range of cells to consider. The second is the quartile to compute. So, entering =QUARTILE(C2:C141, 1) in a cell returns the first quartile for the CIC measurements, while entering =QUARTILE(C2:C141, 3) in a cell returns the third quartile. Note the second quartile is the same as the median.

14. Enter the five-number summary for each dataset in the table below.

Location	Minimum	1st quartile	Median	3rd quartile	Max
Coney Island Creek, 11/7/2018					
Lemon Creek Lagoon, 11/8/2018, ECOconcrete					

Next, enter the data in the Five-number summaries tab in the "Five-number summaries" tab of the Oyster Reefs file to produce the box and whisker plot. Note the "Five-number summaries, example" tab offers a template for your analysis as it explores two years of data from Brooklyn Bridge Park (Figure 11). After you enter the data, highlight cells A1:F3 (those containing all the information and column headers) (Figure 11: red circle). Then insert a chart (Figure 11: yellow circle). On the chart editor, select "Candlestick chart" for "Chart type" (Figure 11: blue circle). Ensure the "switch rows/columns" box is not checked and the "Use row 1 as headers" is checked (Figure 11: purple circle). The resulting graph shows box and whisker plots for each location. Note the "example tab" focuses on differences in shell length (mm) at Brooklyn Bridge Park in 2016 and 2017.

15. 15. What does the example graph suggest about how oyster shell length has changed at Brooklyn Bridge Park between the two years?
16. What might explain the fact that smaller oysters were observed at Brooklyn Bridge Park in 2017 than in 2016?
17. What might the presence of smaller organisms in the 2017 sample (compared to the 2016 sample) indicate? Give two possible explanations.
18. 18. Insert a copy of your box and whisker plot comparing the Lemon Creek Lagoon and Coney Island Creek data here. Type in a detailed caption below your figure.
19. What does the graph suggest? What are some possible reasons for the trends you observed?
20. How do your findings compare to those found in the 2018 BOP report for Lemon Creek Lagoon

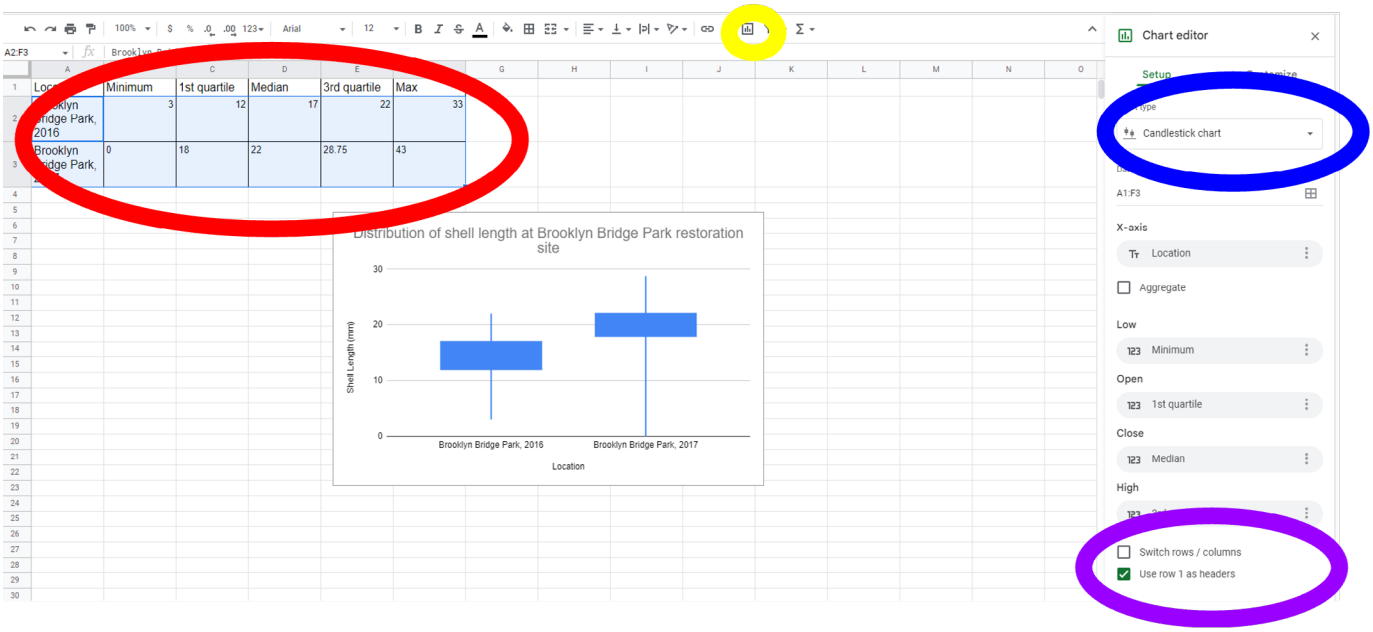


Figure 11. After data are selected (red circle), a chart can be inserted using the "Chart" icon (yellow circle). To construct a box and whisker plot, select "Candlestick chart" under the "Setup tab" (blue circle). Given the noted data structure, ensure the "switch rows/columns" box is not checked and the "Use row 1 as headers" is checked (purple circle).

and Coney Island Creek? You should be able to match your calculated means to those displayed in Table 3 (page 17–18; look for the matching dates and structures as needed and remember ECOconcrete and discs are the same structure type). The histograms you produced should match the shape of those for respective sites, although the BOP report has standardized count to percentage and bin width may slightly alter the graph appearance.

21. The site results for Lemon Creek Lagoon (page 9 of the report) notes impacts of structure on growth. How do the numerical summaries and histograms provided in the report support these concerns? Copy and paste screenshots of the relevant information here and comment on these connections.

CONCLUSION

Assessing how organisms grow at various sites is a major focus of restoration programs. Identifying areas where restored organisms are thriving can encourage future work in the same or similar areas. For the BOP oyster data, trends in growth may also indicate impacts of broodstock, structure, or other factors on oyster survival. Addressing various issues may take different approaches, but observing organism growth is often the first step in noting these limitations, and this requires translating field-collected data into useful summaries. Further work may include gathering observational data on more environmental components or designing and executing experiments to better determine what is limiting or promoting organismal growth.

GLOSSARY

- Adaptive management: the process where restoration or conservation plans are updated given new information about short-term outcomes.
- Broodstock: : refers to the lineage of the oysters. For example, BOP uses oysters they spawn from local broodstock (oysters collected in NY waters) and oysters purchased from other oyster hatcheries. Different lineages may have different traits that impact their survival in restored waters.
- Ecosystem services: are "the benefits people obtain from ecosystems. These include provisioning services such as food and water; regulating services such as regulation of floods, drought,

land degradation, and disease; supporting services such as soil formation and nutrient cycling; and cultural services such as recreational, spiritual, religious, and other nonmaterial benefits" (Millennium Ecosystem Assessment, 2005).

- Hatchery: oysters are spawned in a lab-based hatchery in order to produce offspring, called larvae. Larvae are cared for in the hatchery through their planktonic and settlement phases. During this time oysters float in the water column for a number of weeks before settling onto provided hard substrate. After settlement, juvenile oysters (called spat) are transferred to a nursery.
- Nursery: : an area where spat (small juvenile oysters) can grow in a protected environment prior to being used in restoration or aquaculture activities.
- Outliers: a data point or measurement that is abnormal from other sampled values.
- Restoration ecology: the field focused on repairing damaged or disturbed ecosystems (Vaughn et al., 2010). This may focus on returning some or all function and/or diversity to degraded areas.
- Stakeholder: an individual or group who is connected to or may be impacted by a decision-making process.
- Structure: BOP and other groups use different structures to house oysters when restoring reefs. This can range from placing oysters directly on-bottom to housing them in floating cages, steel cages (gabions), or concrete balls.

ACKNOWLEDGMENTS

A version of this exercise was originally developed with funding provided by Baruch College as part of New York's Open NYS initiative. Module development focused on oyster growth was further supported by a grant from the National Science Foundation (Award # 1839656).

REFERENCES

Ayvazian, S., Gerber-Williams, A., Grabbert, S., Miller, K., Hancock, B., Helt, W., Cobb, D., & Strobel, C. (2020). Habitat benefits of restored oyster reefs and aquaculture to fish and invertebrates in a coastal pond in Rhode Island, United States. Journal of Shellfish Research, 39(3). https://doi.org/10.2983/035.039.0306

Ayvazian, S., Mulvaney, K., Zarnoch, C., Palta, M., Reichert-Nguyen, J., McNally, S., Pilaro, M., Jones, A., Terry, C., & Fulweiler, R. W. (2021). Beyond bioextraction: The role of oyster-mediated denitrification in nutrient management. Environmental Science & Technology, 55(21), 14457–14465. https://doi.org/10.1021/acs.est.1c01901

Belgrad, B. A., Combs, E. M., Walton, W. C., & Smee, D. L. (2021). Use of predator cues to bolster oyster resilience for aquaculture and reef restoration. Aquaculture, 538, 736553. https://doi.org/10.1016/j.aquaculture.2021.736553

Benotti, M. J., Abbene, M., & Terracciano, S. A. (2007). Nitrogen Loading in Jamaica Bay, Long Island, New York: Predevelopment to 2005 (US Geological Survey Scientific Investigations Report Nos. 2007–5051; p. 17). United States Department of the Interior. https://pubs.usgs.gov/sir/2007/5051/SIR2007-5051.pdf

Brandon, C. M., Woodruff, J. D., Orton, P. M., & Donnelly, J. P. (2016). Evidence for elevated coastal vulnerability following large-scale historical oyster bed harvesting. Earth Surface Processes and Landforms, 41(8), 1136–1143. https://doi.org/10.1002/esp.3931

Bruno, J. F., Stachowicz, J. J., & Bertness, M. D. (2003). Inclusion of facilitation into ecological theory. Trends in Ecology & Evolution, 18(3), 119–125. https://doi.org/10.1016/S0169-5347(02)00045-9

Burford, M., Scarpa, J., Cook, B., & Hare, M. (2014). Local adaptation of a marine invertebrate with a high dispersal potential: Evidence from a reciprocal transplant experiment of the eastern oyster *Crassostrea virginica*. Marine Ecology Progress Series, 505, 161–175. https://doi.org/10.3354/meps10796

Burmester, E. M., & McCann, M. (2019). New York City oyster monitoring report: 2018 (p. 69) [Oyster Monitoring Report]. Billion Oyster Project and The Nature Conservancy. https://www.nature.org/content/dam/tnc/nature/en/documents/new-york-city-oyster-monitoring-report-2018.pdf

City of New York Department of Environmental Protection. (2020). Jamaica Bay feasibility study (p. 162). City of New York Department of Environmental Protection. https://www.nyc.gov/assets/dep/downloads/pdf/water/nyc-waterways/jamaica-bay/jamaica-bay-feasibility-study.pdf

Dehon, D. (2010). Investigating the use of bioengineered oyster reefs as a method of shoreline protection and carbon storage [Master of Science in Biological and Agricultural Engineering, Louisiana State University and Agricultural

and Mechanical College]. https://doi.org/10.31390/gradschool_theses.1084
Eastern Oyster Biological Review Team. (2007). Status review of the eastern oyster (*Crassostrea virginica*) (NOAA Tech. Memo. NMFS F/SPO-88, p. 105) [Report to the National Marine Fisheries Service, Northeast Regional Office]. National Oceanic and Atmospheric Administration. https://spo.nmfs.noaa.gov/sites/default/files/TMSPO88.pdf
Eierman, L. E., & Hare, M. P. (2013). Survival of oyster larvae in different salinities depends on source population within an estuary. Journal of Experimental Marine Biology and Ecology, 449, 61–68. https://doi.org/10.1016/j.jembe.2013.08.015
Fears, D. (2016, March 10). This New York storm barrier could have slowed down Sandy. But European settlers ate it. The Washington Post. https://www.washingtonpost.com/news/energy-environment/wp/2016/03/10/how-european-settlers-ate-the-storm-barrier-that-could-have-saved-new-york/
Gosnell, J. S. (2023). What is community science, and how do I get involved? Lessons in Conservation, 13, 38–43. https://doi.org/10.5531/cbc.linc.13.1.3
Grabowski, J. H., Brumbaugh, R. D., Conrad, R. F., Keeler, A. G., Opaluch, J. J., Peterson, C. H., Piehler, M. F., Powers, S. P., & Smyth, A. R. (2012). Economic valuation of ecosystem services provided by oyster reefs. BioScience, 62(10), 900–909. https://doi.org/10.1525/bio.2012.62.10.10
Grizzle, R., Lodge, J., Ward, K., Mosher, K., Jacobs, F., & Krebs, J. (2024). Successful initial restoration of oyster habitat in the lower Hudson River estuary, United States. Restoration Ecology, 32(3), e14077. https://doi.org/10.1111/rec.14077
Harding, J. M., Powell, E. N., Mann, R., & Southworth, M. J. (2013). Variations in eastern oyster (*Crassostrea virginica*) sex-ratios from three Virginia estuaries: Protandry, growth and demographics. Journal of the Marine Biological Association of the United Kingdom, 93(2), 519–531. https://doi.org/10.1017/S002531541200032X
Howarth, R. W., Marino, R., Swaney, D. P., & Boyer, E. W. (2006). Wastewater and watershed influences on primary productivity and oxygen dynamics in the lower Hudson River estuary. In J. S. Levinton & J. R. Waldman (Eds.), The Hudson River Estuary (1st ed., pp. 121–139). Cambridge University Press. https://doi.org/10.1017/CBO9780511550539.012
Kellogg, M., Cornwell, J., Owens, M., & Paynter, K. (2013). Denitrification and nutrient assimilation on a restored oyster reef. Marine Ecology Progress Series, 480, 1–19. https://doi.org/10.3354/meps10331
Kennedy, V. S., Newell, R. I. E., Eble, A. F., & Maryland Sea Grant College (Eds.). (1996). The Eastern oyster: *Crassostrea virginica*. Maryland Sea Grant College.
Kurlansky, M. (2006). The big oyster: History on the half shell. Random House Trade Paperbacks.
Lemasson, A. J., Fletcher, S., Hall-Spencer, J. M., & Knights, A. M. (2017). Linking the biological impacts of ocean acidification on oysters to changes in ecosystem services: A review. Journal of Experimental Marine Biology and Ecology, 492, 49–62. https://doi.org/10.1016/j.jembe.2017.01.019
Levinton, J. S., & Waldman, J. R. (2006). The Hudson River estuary. Cambridge University Press.
MacKenzie Jr., C. L., Burrell Jr., V. G., Rosenfield, A., & Hobart, W. L. (1997). The history, present condition, and future of the molluscan fisheries of North and Central America and Europe Volume 1, Atlantic and Gulf Coasts (NOAA Technical Report No. NMFS 127; A Technical Report of the Fishery Bulletin, p. 185). National Oceanic and Atmospheric Administration. https://spo.nmfs.noaa.gov/sites/default/files/tr127opt.pdf
McCann, M. (2018). New York City oyster monitoring report: 2016–2017 (p. 53) [Oyster Monitoring Report]. The Nature Conservancy. https://static1.squarespace.com/static/5c5604249b8fe80245a0d052/t/5e84fa5335c14c46a304ac48/1585773146958/TNC_BOP_Oyster_Monitoring_Report_2016-2017.pdf
McFarland, K., Plough, L. V., Nguyen, M., & Hare, M. P. (2020). Are bivalves susceptible to domestication selection? Using starvation tolerance to test for potential trait changes in eastern oyster larvae. PLOS ONE, 15(6), e0230222. https://doi.org/10.1371/journal.pone.0230222
Millennium Ecosystem Assessment. (2005). Ecosystems and human well-being: wetlands and water synthesis (A Report of the Millennium Ecosystem Assessment, p. v). World Resources Institute. https://www.millenniumassessment.org/documents/document.358.aspx.pdf
Parker, M., & Bricker, S. (2020). Sustainable oyster aquaculture, water quality improvement, and ecosystem service value potential in Maryland Chesapeake Bay. Journal of Shellfish Research, 39(2), 269. https://doi.org/10.2983/035.039.0208
Rose, J. M., Gosnell, J. S., Bricker, S., Brush, M. J., Colden, A., Harris, L., Karplus, E., Laferriere, A., Merrill, N. H., Murphy, T. B., Reitsma, J., Shockley, J., Stephenson, K., Theuerkauf, S., Ward, D., & Fulweiler, R. W. (2021). Opportunities and challenges for including oyster-mediated denitrification in nitrogen management plans. Estuaries and Coasts, 44(8), 2041–2055. https://doi.org/10.1007/s12237-021-00936-z
Rosenzweig, B. R., Groffman, P. M., Zarnoch, C. B., Branco, B. F., Hartig, E. K., Fitzpatrick, J., Forgione, H. M., & Parris, A. (2018). Nitrogen regulation by natural systems in "unnatural" landscapes: Denitrification in ultra-urban coastal ecosystems. Ecosystem Health and Sustainability, 4(9), 205–224. https://doi.org/10.1080/20964129.2018.1527188
Schiff, Y. (2016, July 28). How a billion oysters are set to change New York's harbor. Observer. https://observer.

com/2016/07/how-a-billion-oysters-are-set-to-change-new-yorks-harbor/
Schmidt, S. (2016, September 4). Oysters are nearly extinct in New York waters. This team is trying to coax them back. The New York Times. https://www.nytimes.com/2016/09/05/nyregion/oyster-project-new-york-harbor.html
Scyphers, S. B., Powers, S. P., Heck, K. L., & Byron, D. (2011). Oyster reefs as natural breakwaters mitigate shoreline loss and facilitate fisheries. PLoS ONE, 6(8), e22396. https://doi.org/10.1371/journal.pone.0022396
Smith, R. S., & Pruett, J. L. (2025). Oyster restoration to recover ecosystem services. Annual Review of Marine Science, 17(1), 83–113. https://doi.org/10.1146/annurev-marine-040423-023007
Vaughn, K. J., Porensky, L., Doshi, M. L., Balachowski, J., Zefferman, E. P., Riginos, C., & Young, T. P. (2010). Restoration ecology. Nature Education Knowledge, 3(10), 66.
Zarnoch, C. B., & Schreibman, M. P. (2012). Growth and reproduction of eastern oysters, *Crassostrea virginica*, in a New York City estuary: Implications for restoration. Urban Habitats, 7. https://www.urbanhabitats.org/v07n01/easternoysters_full.html
Zu Ermgassen, P. S. E., Spalding, M. D., Blake, B., Coen, L. D., Dumbauld, B., Geiger, S., Grabowski, J. H., Grizzle, R., Luckenbach, M., McGraw, K., Rodney, W., Ruesink, J. L., Powers, S. P., & Brumbaugh, R. (2012). Historical ecology with real numbers: Past and present extent and biomass of an imperilled estuarine habitat. Proceedings of the Royal Society B: Biological Sciences, 279(1742), 3393–3400. https://doi.org/10.1098/rspb.2012.0313

Data Analysis in R to Gain Insights for Conservation: Examples from Long-Term Ecological Research

Samantha Sambado and Cheryl J. Briggs

Department of Ecology, Evolution, and Marine Biology, University of California Santa Barbara, Santa Barbara, CA, USA

DOI: https://doi.org/10.5531/cbc.linc.14.1.5 | Supplementary: http://doi.org/10.5531/cbc.ncep.0189

ABSTRACT

The R programming language is a powerful tool for analyzing ecological datasets and gaining valuable insights to inform conservation efforts. This module is designed in two parts to help develop foundational skills for working with ecological data in R. The first part introduces you to R and RStudio, providing a solid foundation for data analysis. The second part focuses on key techniques for data wrangling and visualization. Whether you're new to R or looking to expand your skills from a fresh perspective, this module offers something for those early in their R journey. Throughout, we will use data from long-term ecological research sites to emphasize the critical role of continuous monitoring in understanding the conservation impacts of our changing world.

LEARNING OBJECTIVES

After reading through the module and completing the activities students will be able to:

1. Consider the value of long-term ecological research stations by learning about their contributions to monitoring and understanding ecological change.
2. Analyze descriptive data from biological datasets to identify and summarize key trends and patterns in ecological systems.
3. Create effective data visualizations using R to illustrate biological trends and communicate complex ecological data clearly.
4. Demonstrate foundational R skills for statistical analysis and hypothesis testing, preparing students for more advanced data analysis in future conservation studies.

INTRODUCTION

The Importance of Long-Term Ecological Data for Conservation

Rapid changes in climate and land use have resulted in shifts in species distributions, population abundance, and ecosystem functions (Ehrlén & Morris, 2015; Williams & Newbold, 2020). To assess ecological change, we need to have repeated observations collected over long time scales, and often also on large spatial scales (Lindenmayer et al., 2012). Long-term ecological monitoring provides invaluable contributions to conservation science both through analyses that provide actionable insights, and through data collection, which often aids in community engagement (Magurran et al., 2010; Kao et al., 2012).

Populations and ecosystems naturally vary in abundance, distribution, and other attributes, and may respond to short-term environmental changes. But in conservation, we are most concerned about long-term directional change, especially declines in species or ecosystem function. Through analysis of long-term ecological datasets, we can evaluate whether any changes we observe in a population, species, or ecosystem is due to long-term shifts in climate or land use, an episodic event such as major storms or disease outbreaks, or merely part of natural variation (Kao et al., 2012; Jones & Driscoll, 2022). We can also use long-term data to increase the success of a conservation program by gaining insights into strategies that can guide scientifically informed management strategies. The

importance of long-term data goes beyond understanding trends or changes with anthropogenic activity. Long-term data is essential for foundational research on ecological processes and enables the exploration of fundamental biological questions. A comprehensive dataset is invaluable for testing theories and grounding simulations of remote-sensed data with real-world observations.

To understand whether ecological changes are being driven by anthropogenic activities requires a lot of baseline knowledge (Hughes et al., 2017). From time immemorial, all peoples have learned from their collective experiences over centuries, passing knowledge such as oral histories and other practices that help one another thrive in changing environments. With the rise of Western science and quantitative methods for learning from long term phenomena, scientists have a suite of additional tools. The collection and analysis of long-term data can allow conservationists to determine whether and what type of change is occurring in a population or ecosystem, while the use of statistical models can allow them to assess potential correlates or causes of change (i.e., models of causal inference) (Dee et al., 2023).

Several factors influence the usefulness of long-term data and ensure that future researchers can benefit from past collections (Kuebbing et al., 2018; Décima et al., 2024). First, the quality of the data is shaped by the methods used, which can evolve over time. Strong long-term datasets are characterized by clear documentation of data collection methods, often including assessments of how changes in those methods affect the data. Second, the continuity of data collection is critical; interruptions in data collections can hinder the ability to detect trends or other changes. Third, community engagement is essential for successful long-term monitoring. Collecting such data is resource- and time-intensive, and it relies on the efforts of countless individuals over many years. Therefore, training the next generation of scientists within a collaborative community is crucial for the future of ecological research. Finally, learning to summarize and analyze these data is key for advancing conservation efforts and driving meaningful action to protect our planet.

To assess whether changes are meaningful, we first need to identify and understand patterns within the long-term data. For example, if we want to determine whether shifts in oceanic fish populations are driven by an increase in sea-surface temperature, we must effectively summarize the data by comparing fish abundance across different time periods and locations. As conservation biologists, we also need to test hypotheses, such as whether an invasive plant species is outcompeting native species. To do this, we would analyze historical data on the distribution and abundance of both plant species, summarize trends over time, and then use statistical tools to evaluate the significance of observed changes. Tools like the statistical coding software R are essential in this process, as they enable efficient data manipulation, statistical analysis, and visualization—empowering us to draw statistically valid conclusions and make informed conservation decisions.

Analyzing Long-Term Datasets for Conservation Science Questions

Long-term datasets exist in a variety of formats (i.e., museum collections, active field collection surveys) (Magurran et al., 2010). Examples of long-term datasets of use in conservation include the North American breeding bird survey (Sauer et al., 2019), the UK's Environmental Change Network (Morecroft et al., 2009), and the US National Ecological Observatory Network (NEON) (Kao et al., 2012). In this module, we will use data from the United States Long Term Ecological Research (LTER) Network (To learn more about the LTER Network, you can visit their website here: https://lternet.edu/about/.) In 1980, the National Science Foundation founded the first LTER sites to create a legacy of well-designed and documented long-term observations for future generations. As of 2024, there are research programs at 27 LTER sites that are dedicated to synthesizing data to train and educate the broader scientific community and the public.

To analyze and interpret long-term datasets, scientists rely on computational tools that can efficiently handle and process large volumes of data. Popular options include Python, Jupyter, and R, among others, each offering unique strengths for data manipulation, analysis, and visualization. These programming tools empower us to conduct critical analyses that form the foundation for understanding ecological change and guiding conservation action. Whether we're tracking biodiversity trends, assessing the impacts of land use change, or evaluating the effectiveness of conservation interventions, computational tools help us transform raw data into actionable insights. By learning to use computational tools, we are better equipped to analyze, interpret, and communicate the significance of long-term ecological data, ultimately strengthening our ability to make data-driven decisions for conservation. In this module, we will use R, a programming language that facilitates such analysis. Below, we explain what R and RStudio are and how they will be instrumental in your journey through this module.

What is R and RStudio?

R is a programming language whereas RStudio is a user-friendly platform to visualize R tasks. R is a programming language that combines:

- An extensive set of functions for classical and modern statistical data analysis and modeling.
- Graphing functionality for visualizing data and model outputs.
- A free and open-source tool available to a wide internet community.

The following document will walk you through two parts to introduce R programming language while working with long-term ecological data. Module Part 1 focuses on becoming familiar with R and RStudio. Module Part 2 introduces coding activity to illustrate how we can use R to help us explore data to ask and answer questions. Additional goals of this module are for students to increase their data literacy vocabulary (a Key Terms section is included in this material, as well as embedded definitions), find additional resources to guide their self-taught journey, and learn about professional development opportunities through long-term ecological research programs (see Additional Resources).

MODULE PART 1: INTRODUCTION TO R AND RSTUDIO

If you're new to R or want to revisit the basics in a fresh way, start with Module Part 1—it's a great introduction! If you're already comfortable with R and ready to dive into more advanced topics, feel free to jump ahead to Part 1D. We recommend all learners check out the Key Terms section at the end of this document to help with any confusion about R specific language.

Part 1A. Installing R and RStudio

The instructions below are directions for installing R and RStudio locally on a computer. Be aware that minor changes to this process, screenshots, and the versions are expected over time. Additionally, you or your instructor may find it more useful to use web-based versions of these software platforms, such as Posit Cloud which can be found here: https://posit.cloud/. It is also possible to run this exercise on older versions of R if newer versions do not work on your local computer. In general, it is good practice to state the version of R and the version of packages you are using to ensure replicable analysis with software updates.

1. Download R: https://cran.r-project.org/mirrors.html.
    a. Click the link and scroll down to your country and choose one of the CRAN (Comprehensive R Archive Network) mirror. You can technically choose any server you would like; however, it is suggested you use the closest server to your location. This Module uses Oregon State

University CRAN mirror. [Screenshot 1]
   b. At the top of the page, click the appropriate download link depending on your operating system (Mac OS X vs Windows vs Linux). [Screenshot 2]
   c. For Mac users, click the ".pkg" under the latest release that is compatible with your operating system. At the time this module was written, the most up to date R version is R 4.4.2 "Pile of Leaves." For older Macs, it is recommended you download "R-4.42-x86_64.pkg." You may also need to download the recent version of XQuartz. [Screenshot 3]
   d. For Windows users, click "base" and follow the instructions to download the most recent R version (R-4.4.2 at the time of this module). [Screenshot 4]
   e. Open your downloaded packages, agree to the license and continue downloading as usual. Find the R icon (blue R with a grey circle) in your application folder to confirm a successful download.
   f. Note: If you already have R downloaded on your computer with an older version, the new version should automatically replace your older version.

2. Download RStudio for Mac users: https://posit.co/download/rstudio-desktop/ (other users can find more system specific details here: https://rstudio-education.github.io/hopr/starting.html). This website is (and will always be) free to use, and is licensed under the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/).
   a. Click "Download RStudio Desktop" if your operating system is MacOS 12+. If you have an earlier MacOS environment, please download a previous version from here: https://docs.posit.co/previous-versions/. You should be directed to download a .dmg file.
   b. Locate the downloaded .dmg file in your Downloads folder (or wherever your browser saves downloads).
   c. Double-click .dmg file to open it. This will mount a virtual disk with the RStudio installer (white R in a blue circle).
   d. Drag RStudio to your Applications folder.
   e. Go to your Applications folder and double-click RStudio icon.

Part 1B. Navigating the RStudio Interface

The RStudio interface console, known as the RStudio Integrated Development Environment (IDE), is an interactive command-line interface where users can directly interact with the R programming language. It is important to note that RStudio is a GUI (graphical user interface), which provides a user-friendly platform for interacting with the R programming language. This interface simplifies the coding process, allowing users to easily write, execute, and visualize R code. As a result, RStudio will not work unless the R programming language is installed first.

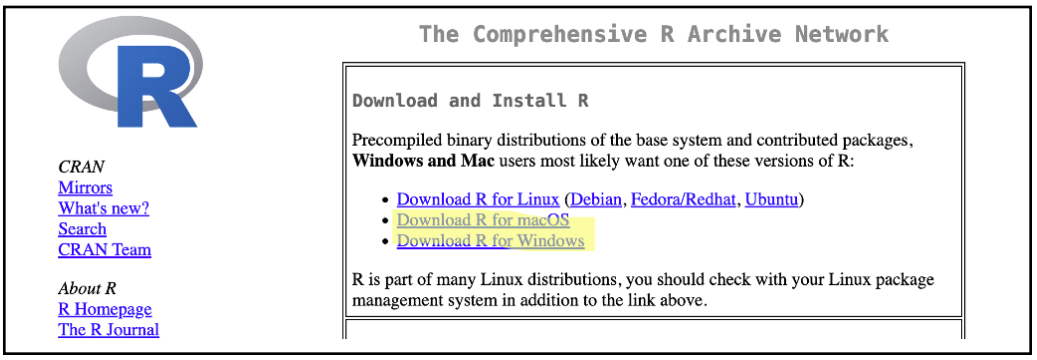
There are 4 "compartments" to be aware of in the RStudio IDE: [Screenshot 5]
1. Source: where you write code—think of this as a Word (or text) document.
2. Console: how R evaluates your code—if you have an error, you will find the message about your error there.
3. Environment/history: where you find saved objects in your working space—if it's not there then R won't run your commands regarding that object.
4. Files/plots/packages/help: where you can find helpful information.

- Note: The orientation of the 4 "compartments" may be different on your screen (for example, the console is bottom left instead of top right), but the function of each compartment remains the same.

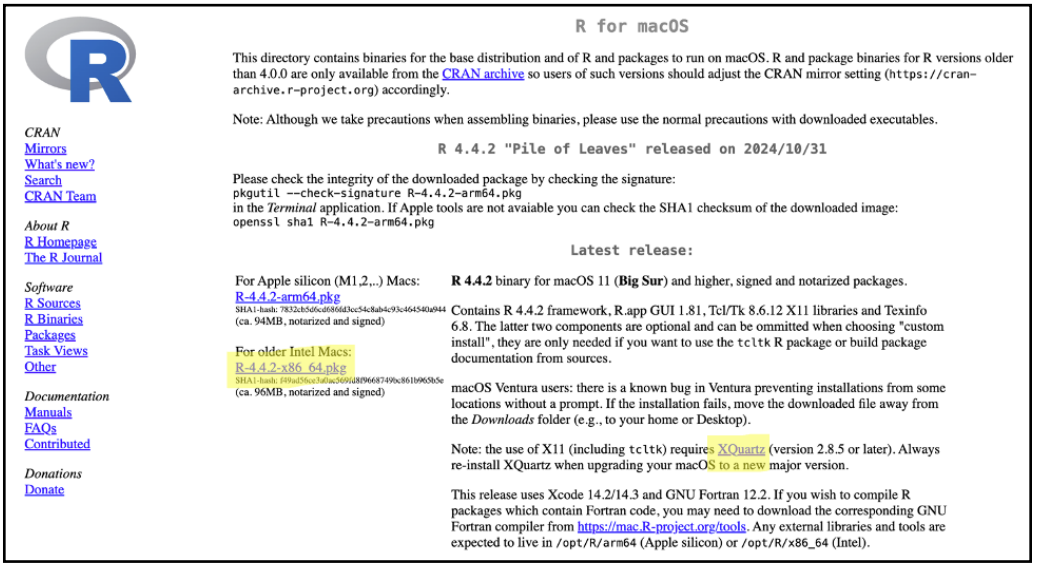
When you first open RStudio, you will see only 3 compartments (left image), but once you open a source file with the top left icon with a green button, then you'll see 4 compartments (right image). [Screenshot 6]

Screenshot 1
Part 1A, 1a:
Click the link and scroll down to your country and choose one of the CRAN (Comprehensive R Archive Network) mirror.

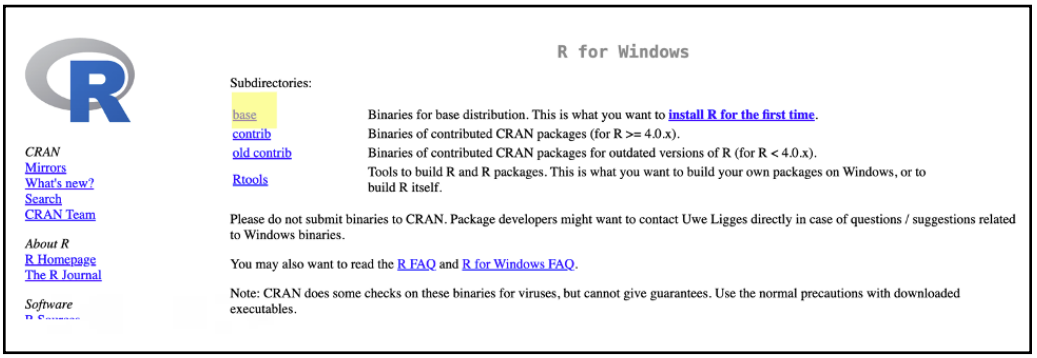
Screenshot 2
Part 1A, 1b:
Click the appropriate download link depending on your operating system



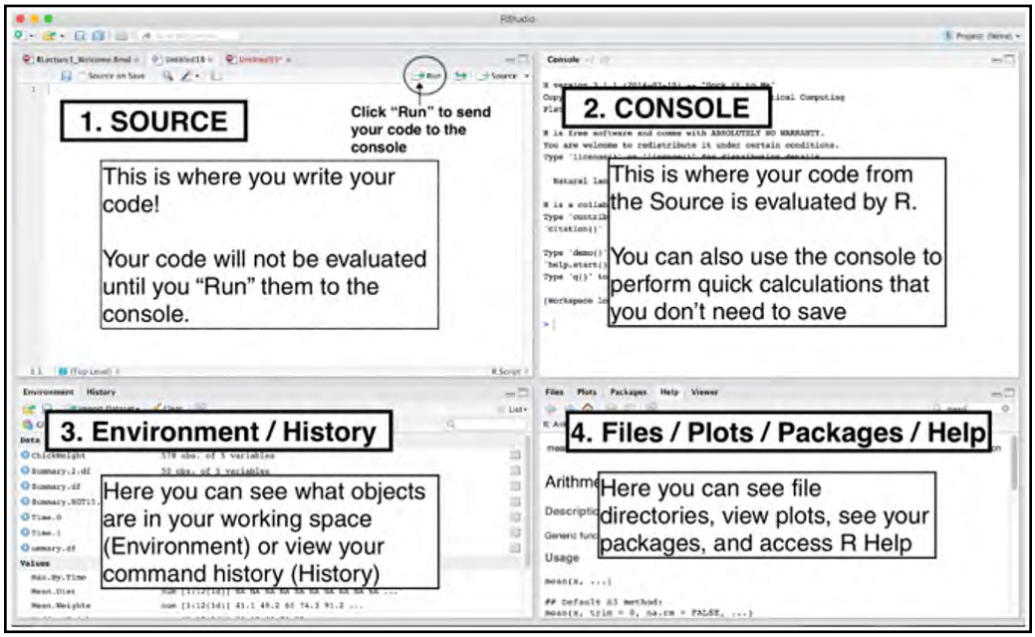
Screenshot 3
Part 1A, 1c:
For Mac users, click the ".pkg" under the latest release that is compatible with your operating system. At the time this module was written, the most up to date R version is R 4.4.2 "Pile of Leaves." For older Macs, it is recommended you download R-4.42-x86_64.pkg. You may also need to download the recent version of XQuartz.



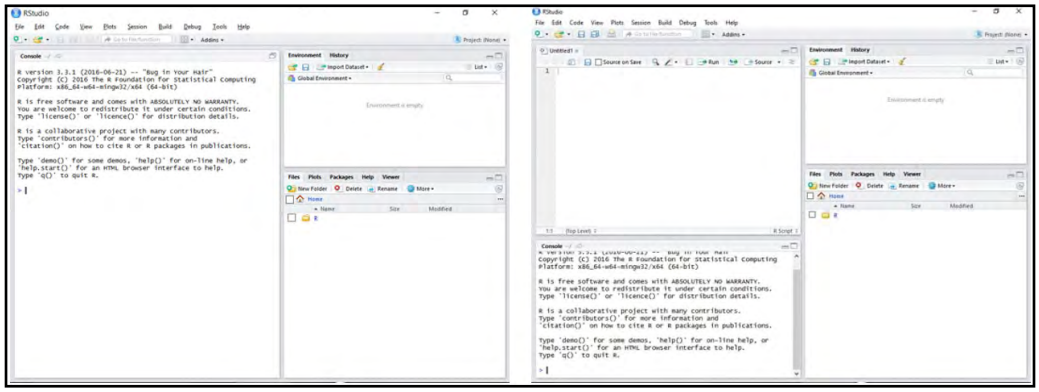
Screenshot 4
Part 1A, 1d:
For Windows users, click "base" and follow the instructions to download the most recent R version (R-4.4.2 at the time of this module).



Screenshot 5
Part 1B:
There are 4 "compartments" to be aware of in the RStudio IDE



Screenshot 6
Part 1B:
When you first open RStudio, you will see only 3 compartments (left image), but once you open a source file with the top left icon with a green button, then you'll see 4 compartments (right image)



Test your understanding: If you wanted to make a note, such as describing the purpose of your .Rmd document, but not have it be evaluated as code, which RStudio console compartment would you write it in?

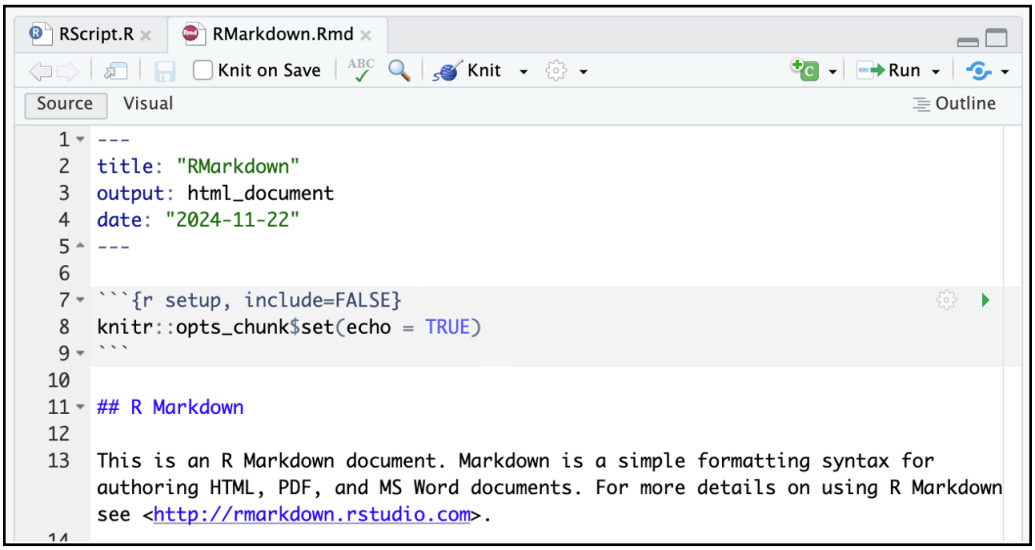
Part 1C. Working with RStudio File Types

Most of conservation science is inherently team-based science due to the large goals of conservation. As part of a team, it is important to communicate effectively and in a reproducible way. The advantage of using R for data analysis and visualization is that there are certain file types that easily integrate into a nice PDF report that can be easily shared with collaborators. Here, we will review the purposes and functions for file types .R, .Rmd, and .html.

- Both .Rmd and .html file types contain the same information (code we need to run and text that describes the code).
- However, the .Rmd file will contain extra code which formats the .html document (e.g., bold words or insert pictures).
- We will typically start in an .Rmd file to write and execute code, however once completed with the code, the .html file may be more visually appealing to read.
- R scripts (.R file) have less functionality as it relates to data visualization, code organization and output generation. [Screenshot 7]
- For this module, examples of code will be in .Rmd files. [Screenshot 8]

Screenshot 7
Part 1C:
R scripts (.R file) have less functionality as it relates to data visualization, code organization and output generation.

Screenshot 8
Part 1C:
For this module, examples of code will be in .Rmd files.



Part 1D. Opening and Editing .Rmd Files In RStudio

.Rmd files are great for writing reports and exporting them as PDF or .html files. For this module we will only be working in .Rmd files. Follow the steps below to open an .Rmd file.

1. Open RStudio.
2. Open a new .Rmd file by selecting File > New File > R Markdown (if you get a message asking you to install packages, you should install them).
3. In the popup window, set your title.
4. Under "Default Output Format" select "HTML" and press "OK."
5. Notice the gray "chunks" that start with "{r}" and end with "". Those chunks contain bits of code and are, in fact, called R chunks.

````
```
This is plain text
```{r cars}
# this is a chunk
```
````

6. There are two important rules of thumb for R chunks:
   a. Any text in an R chunk will be considered R code.
   b. Any text not in an R chunk willl be considered plain text (i.e., no errors).
7. To learn more about formatting your .Rmd document check out R Markdown Cheatsheet at https://rstudio.github.io/cheatsheets/html/rmarkdown.html or https://rstudio.github.io/cheatsheets/rmarkdown.pdf.
````

Note: In R, number signs/hashtags (#) are used to denote comments in the code. Anything following the # on the same line has no effect on the execution of the code. Comments are useful for adding explanations, clarifications, or notes to the code, making it easier for yourself (and others) to understand what the code does.

Test your understanding: Download the R Markdown Cheatsheet and go to the section "Write with Markdown." How would you write the syntax in .Rmd to make a bold word as the output? Hint: the syntax for italics is *word*.

Part 1E. Setting Up Your RStudio Workspace

Now that you have the knowledge on how to maneuver around RStudio let's use RStudio! Every time you create an .Rmd file you will need to make a setup chunk that has the tools and information to complete the rest of your data objectives. The three parts to a well-structured setup chunk are as follows:

1. Setting up the style of your knitted document (i.e., formats your PDF).
2. Install necessary packages (i.e., collections of functions designed to simplify data analysis tasks).
3. Read in your data files.

Action item: Either write out the code from the setup chunk below into your .Rmd file or copy from the provided .Rmd file into your own .Rmd file. If you copy the text directly from this document, make sure that you proofread for changes in spacing (i.e., indentation) and compare and fix it to what is seen here.

```
```{r setup, include=TRUE}

# The first chunk in your Rmarkdown is the setup chunk({r, setup}), this is where you can install/
# load packages and read in data

################################################################################
## part 1. setting up the style of your knitted document
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
# echo = TRUE will have your code show up in the output (good for grading)
# warning = FALSE & message = FALSE will suppress all messages when knitting your document

################################################################################
## part 2. installing packages

# remember, install.packages("PACAKGE") then library(PACKAGE)
  ## Note: quotations for install.packages(), no quotations for library()

# you only need to install a package once to call the library
# after you install.packages(), then hashtag # install.packages()
  ## Note: hashtag (i.e., annotating) is making a comment, it is not seen as code

#install.packages("rmarkdown")
library(rmarkdown) # package for creating .Rmd

#install.packages("knitr")
library(knitr) # package for dynamic report generation

# if you're asked to download other packages, download those as well

################################################################################
## part 3. read in data
# we will use the built in dataset in R called "PlantGrowth"
```
```

```
data("PlantGrowth") # select built in dataset
PlantGrowth # you can manipulate this dataset

```
```

To highlight the components of the lines of code above, we will briefly work through the example above. The hashtags "#" allows you to leave notes or include more explicit organization without causing errors in the code. For example, we break up this single chunk into three parts: setting up the knitting style of the document, installing packages, and reading in the data. If we move to "part 2. installing packages," we can see that **`library()`** is the function and **`rmarkdown`** is the supplied argument. (To read more details on how to structure code within the **`()`**, see "How to Write Functions in R (with 18 Code Examples": https://www.dataquest.io/blog/write-functions-in-r/.)

Test your understanding: In your words, what does "#" do inside the .Rmd chunk? In your .Rmd chunk, "hashtag out" lines of code and see what happens if you run those lines that begin with "#". Do we need to use quotations for **`install.packages()`** or **`library()`** functions?

Part 1F. Conducting a Quick Assessment of Datasets

When you read in your datasets it is important you check that your data was imported correctly and that you have a general sense of the size of your dataset. We will work with the R built in dataset "PlantGrowth". Plant Growth contains data from an experiment examining the effect of different treatment types (control, treatment 1, treatment 2) on the growth of plants. Plants are fundamental to ecosystem function and structure because they form the base of the food chain, provide habitat for wildlife, and play critical roles in carbon sequestration. If you were tasked with restoring degraded habitat by increasing plant biomass, you may be interested in treatments to support accelerated plant growth. Below we will begin with the basic steps of seeing how much data and what type of data we will work with. You should have already read in the data in part 3 of the above code.

1. Check the dimensions of the dataset using the function **`dim()`**. The output (in this case, labeled by a 1 in brackets) returns two numbers, the first will be the number of rows and the second will be the number of columns.

```
```{r}
# i. check the dimensions
dim(PlantGrowth)
```
```

```
## [1] 30 2
```

2. Check the structure of the dataset using the function **`str()`**.

```
```{r}
# ii. check the structure
str(PlantGrowth)
```
```

```
## 'data.frame':    30 obs. of  2 variables:
##  $ weight: num  4.17 5.58 5.18 6.11 4.5 4.61 5.17 4.53 5.33 5.14 ...
##  group : Factor w/ 3 levels "ctrl","trt1",..: 1 1 1 1 1 1 1 1 1 1 ...
```

Test your understanding: In the **`str()`** output, how many levels are there within the column "group"? What are the names of the different levels?
```

CONCLUSION

By this point, you should be familiar with how R looks and how the code is structured. In Module Part 2, you will attempt active coding while learning about long-term ecological research studies.

MODULE PART 2: DATA WRANGLING AND VISUALIZATION

Conservation biology relies on long-term datasets to detect ecological changes and assess the effectiveness of conservation strategies. To help you become familiar with such data, we will use the National Science Foundation's Long-Term Ecological Research (LTER) Program as a foundation. One of the key goals of the LTER Network is to promote education focused on long-term ecological research and the Earth's ecosystems.

Dr. Allison Horst and Dr. Julien Brun have developed an R package,'ltersampler', which transforms a subset of LTER data into an accessible teaching tool for this module. We encourage you to explore the 'ltersampler' project and learn more about the ongoing research at LTER sites here: https://lter.github.io/lterdatasampler/.

Part 2A. Introduction to Data Wrangling with Tidyverse Tools

Long term ecological datasets can be, gently put, a bit messy, for many reasons. Some of the messiness arises from the inherent complexity of ecological data: individual observations may be unique and not fit typical patterns, or changes in data collection protocols over time may complicate comparisons. For example, advancements in technology can make it challenging to relate observations made before and after a protocol change. Human error also plays a role, such as discrepancies in how numbers are recorded (one person might interpret a value differently than other, or data might be recorded in Fahrenheit rather than Celsius). Given that long-term datasets are often collected by many different people over extended periods, errors are inevitable and must be addressed. Data wrangling is the process of cleaning, transforming, and restructuring data into a useable format for analysis. This includes tasks like handling missing data, renaming variables, filtering rows, and creating new columns. As you will discover, data wrangling is a crucial step in any data analysis workflow, and R offers a wide array of powerful tools to make this process more efficient and manageable.

In R, tools such as packages are collections of functions designed to simplify tasks like cleaning, transforming, and visualizing data. The Tidyverse, for example, is not just a single package but a suite of related packages, each with specific functions that work seamlessly together. It's like how a university is made up of various colleges, each specializing in a particular area but contributing to a unified system. The Tidyverse provides a cohesive set of tools that streamline data analysis by offering intuitive, consistent functions across multiple tasks.

To effectively wrangle data, first your dataset must be formatted in a certain way. The core rules of formatting your data for Tidyverse are as follows:

- Each variable has its own column.
- Each observation has its own row.
- Each value has its own cell.

Once your data is in its proper format, you can apply a key Tidyverse tool using the '`dplyr`' package. For large datasets, it's common practice to select or manipulate only a portion of the full dataset to address your specific question. '`dplyr`' provides a suite of functions for data manipulation such as:

- `select()`: selects columns from a data frame.

- `filter()`: filters rows based on conditions.
- `mutate()`: creates new columns or modifies existing ones.
- `arrange()`: sort rows in ascending or descending order.
- `summarize()`: computes summary statistics of data.
- `group_by()`: groups data by one or more variables for aggregation functions.

We will illustrate some of the data wrangling basics in R using datasets built into the package '`lterdatasampler`'.

Test your understanding: True or False, each variable has its own column. True or False, if you want to create a new column using '`dplyr`' functions, you would use the `mutate()` function.

***** OPTIONAL: If you need help installing and maneuvering around RStudio please refer to Part 1 *****

Part 2B. Applying Data Wrangling to Conservation Questions

We will explore two conservation cases that involve species—pikas and fiddler crabs—that are particularly sensitive to climatic changes in their environments. For pikas, we will examine the stress levels of individual pikas and investigate how these may relate to elevation. This analysis suggests that pika populations at higher elevations might be shifting their ranges due to increasing environmental stress, underscoring the need for targeted conservation strategies, such as mitigating environmental stressors or relocating pikas to more suitable habitats (Erb et al., 2011; Galbreath et al., 2009). In the case of fiddler crabs, we will analyze patterns in crab body size in relation to latitude. Since latitude is correlated to water temperature, this may reveal regional variations in environmental conditions, which could influence the management of crab populations and inform policy decisions (Brodie et al., 2017; Darnell & Darnell, 2018). These examples illustrate how species can serve as ecosystem indicators, providing valuable insights into environmental changes and guiding conservation efforts.

Before data wrangling can begin it is important to first develop a potential biological question and then understand what type of data has been collected prior to data visualizations or analysis (see Additional Resources). Entire courses are dedicated to these fundamental concepts, but we briefly summarize the main points below.

Developing Biological Research Questions

There are many ways to decipher patterns in the data, which is a crucial aspect of conservation biology. While developing biological questions is a complex process (see Additional Resources), we provide some general guidelines on how data should be assessed:

- First ask yourself, what pattern are you trying to understand or explain? In biological data analysis, we call this our outcome of interest or our dependent variable 'y'. For example, are you trying to understand whether a population is increasing or decreasing? Your dependent variable might then be population size. Or you might be interested to understand whether species richness is impacted by forest or range management practices, in which case species richness might be your dependent variable.
- Then ponder, what is influencing your outcome of interest? In biological data analysis, it is called our independent variable 'x'. Independent variables are the factors that may impact your variable of interest, or dependent variable.

Some simple conservation biology questions you may have that R could theoretically allow you to

investigate within a dataset:

- Does elevation (independent) influence stress levels (dependent) in pikas?
- Does sea surface temperature (independent) influence crab body size (dependent)?
- Does aquatic habitat quality (independent) influence vertebrate population numbers (dependent)?

Test your understanding: If my question is, "Does soil type influence tree height?", which would be my dependent/Y variable/outcome of interest—soil or tree height?

Understanding Data Types in R

As conservation biologists, you apply your biological expertise to guide important questions. As data scientists, you will need to understand the type of data collected for your biological study system to ensure you can use R to statistically address your questions. In this section, we will introduce the concept of different types of data and explain what they are, and then introduce how they are coded in R.

When you are designing an experiment or collecting data it is good to determine what type of data you are collecting because the type of data also determines what statistical analysis you can complete. For example, if you want to find the correlation between tree canopy cover percentage and total tree saplings found below the canopy, you will need to make sure both are recorded as numeric data. Or if you want to compare the mean count of starfish between a marine protected area and a non-marine protected area, you will need to make sure starfish data are recorded as a numeric and marine protected area as a categorical.

To find a more in-depth description of data types, refer to: https://www.geeksforgeeks.org/data-types-in-statistics/. A non-exhaustive list of data types encountered in this module include:

- Numerical: values that are represented by numbers.
  - Example: 1, 4.03, 7000
- Categorical or Nominal: categories, labels or descriptions that have no intrinsic order or hierarchy.
  - Example: fish species, colors, countries
- Ordinal: discrete categorical units with an order or hierarchy.
  - Example: months of year, life stages of a species

When you call in a new dataset, the first thing to do is to check how R reads in your data. It is good to be aware that in R, data types are called slightly different things. Some ways R may define data variables include:

- Class: the type of data contained in a variable. Usually a character (i.e., text or words), numeric (i.e., number), integer (i.e., whole number), or factor (i.e., grouping values, useful when you have multiple observations for sites or treatments in your data).
- Factor: categorical grouping.
  - Example: fish species, tree species
- Levels: categorical sub-groups of factors.
  - Example: (chinook salmon, sand lance, herring), (oak, madrone, pine)

Now that you have a foundation of how to form questions and how to record data as certain data types, we can open R and RStudio and bring in some data.

Test your understanding: If I had a column labeled bird species and the variations of bird species were blue heron, crow, owl—which grouping is considered a factor (bird species or individual bird names)?

Activity 1. Analyzing Stress Levels of Pikas at High Elevation

Our first example uses the dataset "Pika observations at Niwot Ridge LTER, Colorado, USA". American Pika (*Ochotona princeps*; Figure 1) are small mammals that live on rocky, alpine slopes in western mountains of North America, and are the smallest members of the rabbit family. They eat grasses and wildflowers and store them to feed on through the winter. They are adapted to alpine conditions, and human-induced climate change is posing challenges to their survival. Rising global temperatures, driven by human activities such as the burning of fossil fuels, are reducing the insulating snowpack that traditionally protects the pikas from freezing in their dens. In addition, more frequent and extreme heat waves during the summer put them at increased risk of heat stress and mortality. These temperature shifts and snowpack reductions, directly linked to anthropogenic climate change, are altering the pika's environment and challenging their ability to survive in their natural habitat. Because of their sensitivity to these environmental changes, pikas serve as a crucial indicator species for the health of alpine ecosystems impacted by climate change. One way to measure their response to climate stressors is by assessing their stress hormone metabolites found in pika feces, a non-invasive technique (denoted as `concentration_pg_m` in the dataset, which measures glucocorticoid metabolites). This method allows scientists to monitor pika populations without causing additional harm or stress, providing a more sustainable way to track the impacts of human-driven climate change on these vulnerable species.

- We will use Pika observations at Niwot Ridge LTER in Colorado to illustrate how you can use R to understand basic characteristics of the population in the dataset.
- More information on how data were collected can be found here: https://lter.github.io/lterdatasampler/reference/nwt_pikas.html.

1.1 Set up RStudio working space like the image below (you will need to install and read in packages '`ltersampler`' and '`tidyverse`'). Once the packages are available in your RStudio, then you can read in your data, "`nwt_pikas`" to begin asking your questions.

Reminder: In R, number signs/hashtags "#" are used to denote comments in the code. Anything following the # on the same line has no effect on the execution of the code. Comments are useful for adding explanations, clarifications, or notes to the code, making it easier for yourself (and others) to understand what the code does. Action item: Either write out the code from the setup chunk below into your .Rmd file or copy from the provided .Rmd file into your own .Rmd file. If you copy the text directly from this document, make sure that you proofread for changes in spacing (i.e., indentation) and compare and fix it to what is seen here.

Figure 1. An American pika (*Ochotona princeps*) chewing on a leaf. Image Credit: Marshal Hedin via Wikimedia Commons/CC-BY-2.0.

```r
```{r}
################################################################################
# part 1. installing packages
# remember, install.packages("PACAKGE") then library(PACKAGE)
  ## Note: quotations for install.packages(), no quotations for library()
# you only need to install a package once to call the library
# after you install.packages(), then hashtag # install.packages()
  ## Note: hashtag (i.e., annotating) is making a comment, it is not seen as code

#install.packages("lterdatasampler")
library(lterdatasampler) # package with built in LTER datasets

#install.packages("tidyverse")
library(tidyverse) # package for data wrangling

################################################################################
## part 2. read in data
# we will use the built in dataset from lterdatasampler

nwt_pikas # pika data
```
```

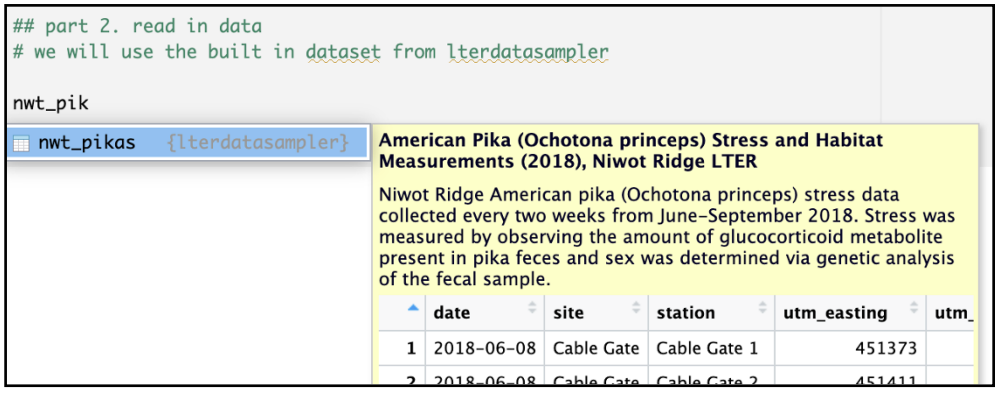
Note: If you uploaded **`library(lterdatasampler)`** correctly then you should see the dataset pop up when typing out "**`nwt_pikas`**" before hitting enter or running the full line of code **`nwt_pikas`**. [Screenshot 9]

Test your understanding: What is our outcome of interest (y variable) in the "**`nwt_pikas`**" dataset for this activity? Hint: it was mentioned in the pika description above and measures glucocorticoid metabolites.

1.2 Exploring a new dataset is always exciting, as it holds a lot of potential. However, if you didn't collect the data yourself, it may not have immediate meaning until you explore its contents. The first step in working with any dataset is to explore it thoroughly. You need to understand the size of the dataset (i.e., the number of rows and columns) and the types of data it contains (e.g., numeric, factors).

To explore the "**`nwt_pikas`**" dataset, start by using the **`dim()`** function to check its dimensions. This will return two numbers after a bracketed line number [1]: the first indicates the number of rows (observations), and the second shows the number of columns (variables). Additionally, use the **`str()`** function to check the structure of the dataset, which provides detailed information about the types of data in each column.

Screenshot 9
Activity 1.1:
Note: If you uploaded **`library(lterdatasampler)`** correctly then you should see the dataset pop up when typing out "**`nwt_pikas`**" before hitting enter or running the full line of code **`nwt_pikas`**.



```{r}
# i. dimensions
dim(nwt_pikas)
```

```
## [1] 109   8
```

```{r}
## ii. structure
str(nwt_pikas)
```

```
## tibble [109 × 8] (S3: tbl_df/tbl/data.frame)
## $ date           : Date[1:109], format: "2018-06-08" "2018-06-08" ...
## $ site           : Factor w/ 3 levels "Cable Gate","Long Lake",..: 1 1 1 3 3 3 3 3 3 3 ...
## $ station        : Factor w/ 20 levels "Cable Gate 1",..: 1 2 3 14 15 16 17 18 19 9 ...
## $ utm_easting    : num [1:109] 451373 451411 451462 449317 449342 ...
## $ utm_northing   : num [1:109] 4432963 4432985 4432991 4434093 4434141 ...
## $ sex                  : Factor w/ 1 level "male": 1 1 1 1 1 NA 1 NA 1 1 ...
## $ concentration_pg_g   : num [1:109] 11563 10629 10924 10414 13531 ...
## $ elev_m         : num [1:109] 3343 3353 3358 3578 3584 ...
```

Test your understanding: How many observations of pikas are there in the dataset "**nwt_pikas**"? On the pika LTER page (https://lter.github.io/lterdatasampler/reference/nwt_pikas.html) look up what "**utm_easting**" and "**utm_northing**" are and report their units. Why would elevation be an important variable to assess with pika population dynamics?

1.3 Certain sites may contribute to stress in pikas, which could be reflected by higher levels of stress metabolites (coded in the column "**concentration_pg_g**"). For example, sites above the tree line, like many high-elevation areas, may have less canopy cover, leading to increased predation on pikas. To explore how "**concentration_pg_g**" levels vary by site, you can group the data by site using the **group_by()** function and then calculate summary statistics, such as the mean concentration, using the **summarize()** function. When you apply the **summarize()** function, you then create a new column, which is called "**mean_concentration**" in the example below.

Note: To link multiple functions, such as **group_by()** and **summarize()**, you can use the pipe operator (%>%). Pipes allow you to chain functions together, essentially saying "do this, and then do that" making your code more readable and efficient.

Note: When applying the pipe operator, it is important to "call in the data set" first by stating which dataset you would like to apply the following functions to. If the dataset "**nwt_pikas**" was not coded before **group_by(site)**, then R would not know where the column site was coming from. You must always state the data set (i.e., **nwt_pikas**) and then column (i.e., site) to make sure you are using the correct data.

Note: As a reminder, R is very sensitive to differences in capitalization, punctuation, and spelling. To call the column "**concentration_pg_g**" from the "**nwt_pikas**" dataset, you must spell it exactly how it is or else the code will not work. See Additional Resources for tips on how to deal with code errors.

```{r}
# use a pipe (%>%) to link series of actions
nwt_pikas %>% # call in dataset
  group_by(site) %>% # group by site
  summarize(mean_concentration = mean(concentration_pg_g)) # calculate mean
```

```
## # A tibble: 3 x 2
##   site     mean_concentration
##   <fct>                 <dbl>
## 1 Cable Gate            5412.
## 2 Long Lake             4730.
## 3 West Knoll            5167.
```

Test your understanding: Which of the three sites have the highest mean concentration of stress metabolite ("`concentration_pg_g`") in pika feces? If you also calculate the mean elevation using the same format as "`mean_concentration`", do you see the highest concentration of stress in the site with the highest mean elevation? What are some biological reasons why pika stress levels would be related to elevation of their sites?

Data wrangling can be a powerful tool with very few lines of code. However, it is important to remember the basics of data wrangling (e.g., rules of '`tidyverse`', correct spelling/punctuation/capitalization, etc.) to ensure you get the intended summaries. These skills can be transferable beyond the LTER data and applied to your own data.

Activity 2. Examining Body Size Patterns of Crabs Across Latitudes

Our second example uses the dataset "Fiddler crabs in salt marshes from Florida to Massachusetts, USA". This data looks at adult fiddler crabs (*Minuca pugnax*; Figure 2) and how their body size relates to mean water temperature. Because the crab's size influences its impact on an ecosystem, and *M. pugnax* are ecosystem engineers that affect marsh functioning, the larger crabs at higher latitudes may have greater per-capita impacts on salt marshes than the smaller crabs at lower latitudes, due to Bergmann's rule (which predicts that organisms at higher latitudes are larger than ones at lower latitude).

- We will use fiddler crab observations at 3 LTER sites along the US east coast to illustrate how you can use R to understand basic characteristics of the population in the dataset.
- More information on how data were collected can be found here: https://lter.github.io/lterdatasampler/reference/pie_crab.html.

Figure 2. Fiddler crabs (*Minuca pugnax*) in the mud at Sippewissett Marsh. Image credit: Taeylenol via Wikimedia Commons/CC-BY-4.0.

2.1 To set up your RStudio working space you will need to install and read in packages '**ltersampler**' and '**tidyverse**'. Once the packages are available in your RStudio, then you can read in your data, "**pie_crab**" to begin asking your questions.

```{r}
# part 1. install necessary packages
library(lterdatasampler) # package with built in LTER
library(tidyverse) # package for data wrangling

## part 2. read in data
# we will use the built in data set from lterdatasampler
pie_crab # crab data
```

Test your understanding: Why did we not use the function **install.packages()** for the second activity? What are the dimensions of the "**pie_crab**" dataset?

2.2 Bergmann's rule predicts that organisms at higher latitudes tend to be larger than those at lower latitudes. This pattern is often linked to cooler temperatures at higher latitudes, which can influence physiological rates and resource availability. Before analyzing how size relates to physiology or resource access, we first need to determine if larger crabs are indeed found at northern latitudes.

To do this, follow a similar approach as in section 1.3. Start by calculating the maximum size (in mm) of each individual fiddler crab, and then arrange the results in descending order using the function **arrange()**. This will help us easily identify any patterns in size across latitudes.

```{r}
pie_crab %>% # call in dataset
  group_by(name) %>% # group by site name
  summarize(max_size = max(size)) %>% # calculate max size
  arrange(desc(max_size)) # arrange in descender size order
```

```
## #  A tibble: 13 x 2
##    name                          max_size
##    <chr>                            <dbl>
##  1 Bare Cove Park                    23.4
##  2 Plum Island Estuary - West Creek  22.1
##  3 Narragansett Bay NERR             21.9
##  4 Virginia Coastal Reserve LTER     21.4
##  5 Jacques Cousteau NERR             20.9
##  6 Delaware Bay NERR                 20.4
##  7 Cape Cod                          19.8
##  8 Sixpenny Island - Connecticut     19.5
##  9 Rachel Carson NERR                17.6
## 10 North Inlet Winyah Bay NERR       17.3
## 11 Zeke's Island NERR                15.7
## 12 Guana Tolomoto Matanzas NERR      14.9
## 13 Sapelo Island NERR                13.6
```

Test your understanding: Which site name has the largest recorded fiddler crab? If you look up on the site on project's data website (https://portal.edirepository.org/nis/metadataviewer?packageid=knb-lter-pie.540.1), does the data support with Bergmann's rule? What would you predict would happen to the salt marsh ecosystems, if surface sea level temperatures continue to rise and it either increased or decreased the average size of fiddler crab populations?

2.3 Although ecological organisms do not recognize political geographic boundaries, management policies aimed at conserving or regulating species are typically implemented at the state or regional level. For crabs, these policies may include measures such as total catch limits, size limits, or seasonal closures of certain areas.

If you were tasked with coordinating state and regional efforts to better manage crab populations, you might choose to divide the data into regions to make more targeted recommendations. In the United States, for example, an arbitrary divide between the northern and southern regions is often set at the 36th parallel of latitude.

First, wrangle your data into two separate objects (denoted by the left side of the arrow <-), "`northern_crabs`" and "`southern_crabs`", based on the latitude associated with each observation. To do this, use the `filter()` function to subset your dataset according to the latitude values. The `filter()` function is powerful and allows for customization through Boolean operators, which are logic-based terms (for example, & is "and," | is "or," and ! is "not.") that help narrow or expand your subset.

It's often useful to provide context for your filtering decision. You can add a new column to your dataset using the `mutate()` function. In this case, create a new column called "`region`" to label each observation as either "`north`" or "`south`" based on latitude. After filtering the dataset into these two regions, calculate the minimum water temperature for each region.

```
```{r}
# i. northern crabs object
northern_crabs <- # create object
  pie_crab %>% # call in data set
  filter(latitude > 36) %>% # filter observations above 36 lat
  mutate(region = "north") %>% # make new column called region
  summarize(min = min(water_temp)) # calculate min water temp
northern_crabs # shows output

# ii. southern crabs object
southern_crabs <- # create object
  pie_crab %>% # call in dataset
  filter(latitude < 36) %>% # filter observations below 36 lat
  mutate(region = "south") %>% # assign observations as south
  summarize(min = min(water_temp)) # calculate min water temp
southern_crabs # shows output
```
```

Think critically: Ectothermic organisms (organisms that are unable to regulate their own body temperature and depend on their external stimuli) have optimal temperatures they can perform better in (i.e., improve foraging and mating outcomes). Although fiddler crabs can survive at lower temperatures, they experience cold stress and prolonged exposure to temperatures approaching freezing can lead to mortality. If the lower threshold for fiddler crabs is closer to 10°C, which region is more likely for fiddler crabs to experience cold stress? Would you recommend both regions carry out mitigation strategies to modify habitat for fiddler crabs to shelter in during cold spells or just one region? What are other conservation concerns that would be different for northern and southern crabs?

Part 2C. Visualizing Biological Data to Form Hypotheses

In this section, we will explore two conservation biology examples—prairie bison and aquatic vertebrates—that are sensitive to environmental changes in their habitats. For bison, we will analyze long-term data on individual mass and herd structure at Konza Prairie Biological Station, which could

later be used to examine how factors like habitat fragmentation, disease outbreaks, and human-wildlife conflict affect their populations. These insights can guide conservation strategies, such as habitat restoration or management to reduce human impacts, ensuring bison populations remain healthy and viable. For aquatic vertebrates, we will study the population dynamics of cutthroat trout and Pacific giant salamanders in the Andrews Experimental Forest. These species are influenced by environmental changes such as pollution, invasive species, and habitat degradation. By examining trends in their population numbers, we can assess the impacts of these stressors and inform management practices aimed at protecting aquatic ecosystems. Together, these examples highlight how monitoring species' responses to environmental changes can provide critical information for targeted conservation efforts.

In this part, we are going to learn how to make figures that tell you and anyone exactly what the data are showing. This tutorial will show you some examples of making figures in R. There is no way a tutorial can tell you everything you need to know about plotting in R, so we are just going to cover the basics. We recommend playing around with some of the code as you work through this tutorial.

If you have a question about how to make certain plots, check out The R gallery websites which can be found here: https://r-graph-gallery.com/. Websites like these give you examples of how to make plots in R with all the code included! If you get really excited about plotting in R, check out the package '`ggplot2`'. This package offers lots of options for customizing graphs and figures (see the ggplot 2 Cheat Sheet here: https://www.datacamp.com/cheat-sheet/ggplot2-cheat-sheet). We will use the '`ggplot2`' package in this tutorial.

To plot in 2 dimensions, the basic format of code in '`ggplot`' is always the same: we start by using the `ggplot()` function to specify what data to use, and which variable to put on which axis. Then, we add on a "`geom`" using the plus sign to specify what kind of plot we want. There are a ton of other functions that can be added from there to customize the plot.

Test your understanding: In your own words, what makes an effective graph? Are there certain things or styles you like or dislike?

Activity 3. Exploring Population Dynamics of Prairie Bison

In this activity we will work with the dataset "Bison masses recorded for the herd at Konza Prairie Biological Station LTER" which tracks long-term changes in the individual mass of bison (*Bison bison*; Figure 3), as well as herd structure, end-of-season weights, and maternal parentage. Once widespread across North America, bison nearly faced extinction in the late 19th century due to factors such as habitat loss and fragmentation, genetic bottlenecks, disease outbreaks (such as Brucellosis), human-wildlife conflict (including competition with livestock), and settler colonialism. (A more in-depth discussion on the impacts of settler colonialism on bison populations can be found in this StoryMap: https://storymaps.arcgis.com/stories/c8d348df1b004fe8a642c8be7d782add.)

- We will use this dataset to demonstrate how to visualize long-term trends in bison populations and explore techniques for creating subplots to focus on specific bison sex. These visualizations will help illustrate how the population is structured (male vs female) and average weights, providing insights into their conservation needs.
- More information on how data were collected can be found here: https://lter.github.io/lterdatasampler/reference/knz_bison.html.

Test your understanding: Based on biological reasoning, would you expect population trends to look

Figure 3. A bison (*Bison bison*) standing in a field. Image credit: Jack Dykinga via USDA.

similar for two populations of buffalo if one population was impacted by disease outbreaks while the other population was impacted by a single human-wildlife conflict event?

3.1 Set up your RStudio working space you will need to install and read in packages '**`ltersampler`**', '**`tidyverse`**', '**`plyr`**', and '**`ggplot2`**'. Once the packages are available in your RStudio, then you can read in your data, "**`knz_bison`**" to begin asking your questions.

```
```{r}
# part 1. install necessary packages
library(lterdatasampler) # package with built in LTER
library(tidyverse) # package for data wrangling
#install.packages("plyr") # if first time using
library(plyr) # to use function revalue() to change factor levels
#install.packages("ggplot2") # if first time using
library(ggplot2) # make nice plots

## part 2. read in data
# we will use the built in dataset from lterdatasampler
knz_bison # bison data
```
```

Test your understanding: What are the dimensions of the "**`knz_bison`**" dataset? Note: You may want to review Part 1F of this module.

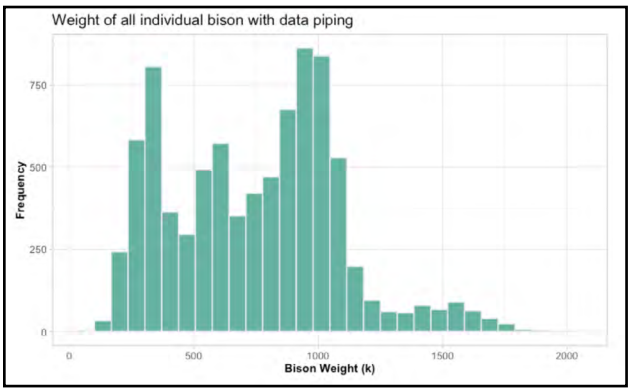
3.2 To first get a sense of the population numbers and how a population may be structured (based on weight), it is good to plot the distribution of all your data. A histogram is a graphical representation of the distribution of a numeric variable. The variable is cut into several bins, and the number of observations per bin is represented by the height. [Screenshot 10]

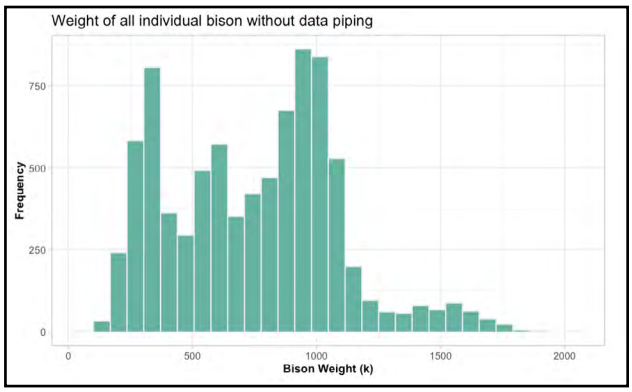
Note: You will see two ways to pass data through the **`ggplot()`** function. Both ways produce the same plot. We prefer piping data into **`ggplot()`** because it allows for better data manipulation without altering the dataset.

- [left graph] pipe data into **`ggplot()`** function > data %>% **`ggplot(aes(…))`**
- [right graph] nest data into the **`ggplot()`** function > **`ggplot(data,…)`**

```{r}
# i. pipping data into ggplot
knz_bison %>% # PIPING IN DATA
  ggplot( aes(x = animal_weight)) +
  geom_histogram(color = "#e9ecef", fill = "#69b3a2", position = "identity") +
  theme_light() +
  labs(x = "Bison Weight (k)", y = "Frequency") +
  ggtitle("Weight of all individual bison with data piping") +
  theme(axis.title = element_text(size = 10, face = "bold"))

# ii. not pipping data into ggplot
ggplot(knz_bison, aes(x = animal_weight)) + # NO PIPING IN DATA
  geom_histogram(color = "#e9ecef", fill = "#69b3a2", position = "identity") +
  theme_light() +
  labs(x = "Bison Weight (k)", y = "Frequency") +
  ggtitle("Weight of all individual bison without data piping") +
  theme(axis.title = element_text(size = 10, face = "bold"))
```





Screenshot 10

Activity 3.2:

To first get a sense of the population numbers and how a population may be structured (based on weight), it is good to plot the distribution of all your data.

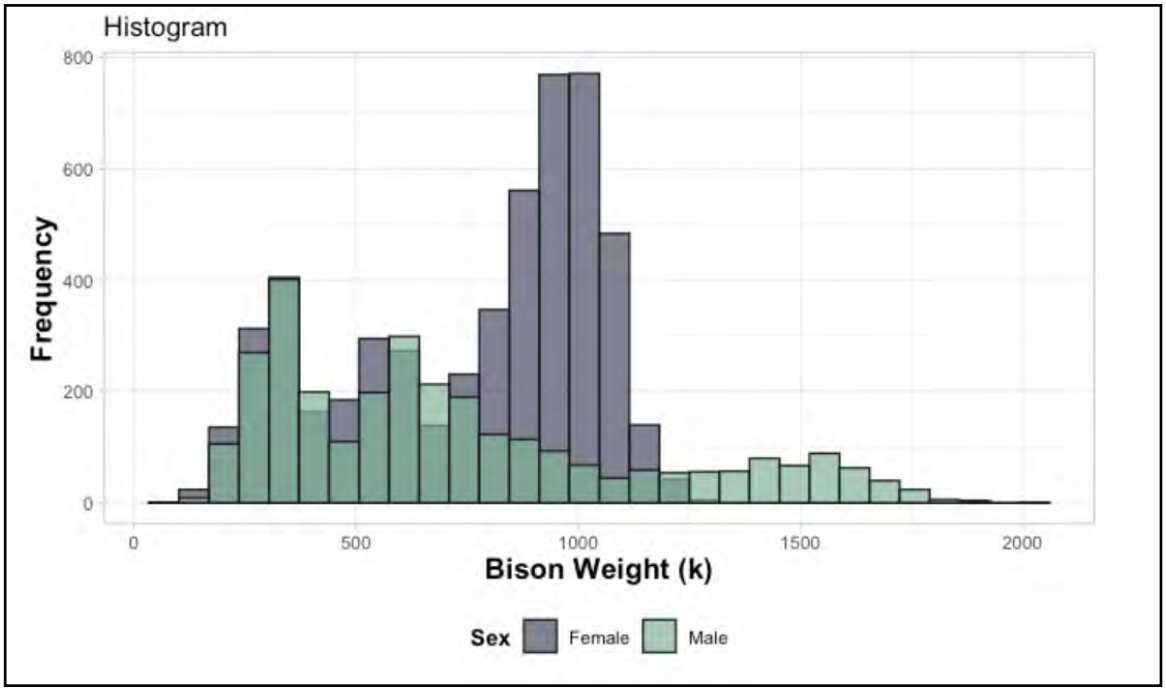
Test your understanding: True or False, "%>%" is the syntax for piping.

3.3 There may be a significant weight difference between male and female bison that should be considered, especially if conservation efforts are focused on addressing genetic inbreeding. Understanding the distribution of males and females in the population is important for these efforts. You can still use a histogram but be sure to differentiate the data by sex using color. In this dataset, males and females are coded as "M" and "F", which may not be as informative as "Male" and "Female". To make the labels clearer, we will use the **`mutate()`** function to transform these codes into more descriptive terms. [Screenshot 11]

Note: It's crucial to make your plot informative and easy to interpret. To improve clarity, use **`ggtitle()`** to add a meaningful title and **`labs()`** to provide informative axis labels. For a more polished appearance, consider using **`theme_light()`** to change the background color and adjust the font size of the text with arguments in the **`theme()`** function.

```{r}
knz_bison %>% # call in data
  plyr::mutate(animal_sex = revalue(animal_sex, # change animal_sex column values
                                    c("F" = "Female", # F to Female
                                      "M" = "Male"))) %>% # M to Male
```

```
  ggplot(aes(x = animal_weight, fill = animal_sex)) +
  geom_histogram(color = "grey10", # make histogram
                 position = "identity", # make it frequency of group
                 alpha = .7) + # increase transparency
  scale_fill_manual(values = c("#3d405b", "#81b29a")) + # color code sex
  theme_light() + # light background color
  labs(x = "Bison Weight (k)", y = "Frequency", # change axis labels
       fill = "Sex") + # change legend label
  ggtitle("Histogram") + # give title
  theme(axis.title = element_text(size = 14, face = "bold"), # change text font
        legend.title = element_text(face = "bold"), # change text font
        legend.position = "bottom") # put legend at bottom
```



Screenshot 11 Activity 3.3: There may be a significant weight difference between male and female bison that should be considered.

Test your understanding: Based on the histogram, do female or male buffalos have higher weight limits? Make another histogram with the same code but change the colors assigned to female and males. You can find a list of colors here: https://r-charts.com/color-palettes/. When you hover over a color you will see it's "hex code" which is a mixed numeric-letter value following a "#". You must replace the previous colors ("#3d405b" and "#81b29a") in the new histogram.

3.4 While histograms are useful for visualizing the full distribution of your data, you might want to focus on summary statistics, such as the median, to identify a target weight for each sex. This can be particularly helpful for conservation biologists when developing diet plans for bison in zoo programs.

To visualize these summary statistics by group (male vs female), we can use a boxplot, which provides a clear representation of the central tendency and variability in the data. [Screenshot 12] A boxplot visually represents the numeric distribution by grouping (i.e., factor). The project, From Data to Viz (found here: https://www.data-to-viz.com/#boxplot), describes the elements of a box plots:

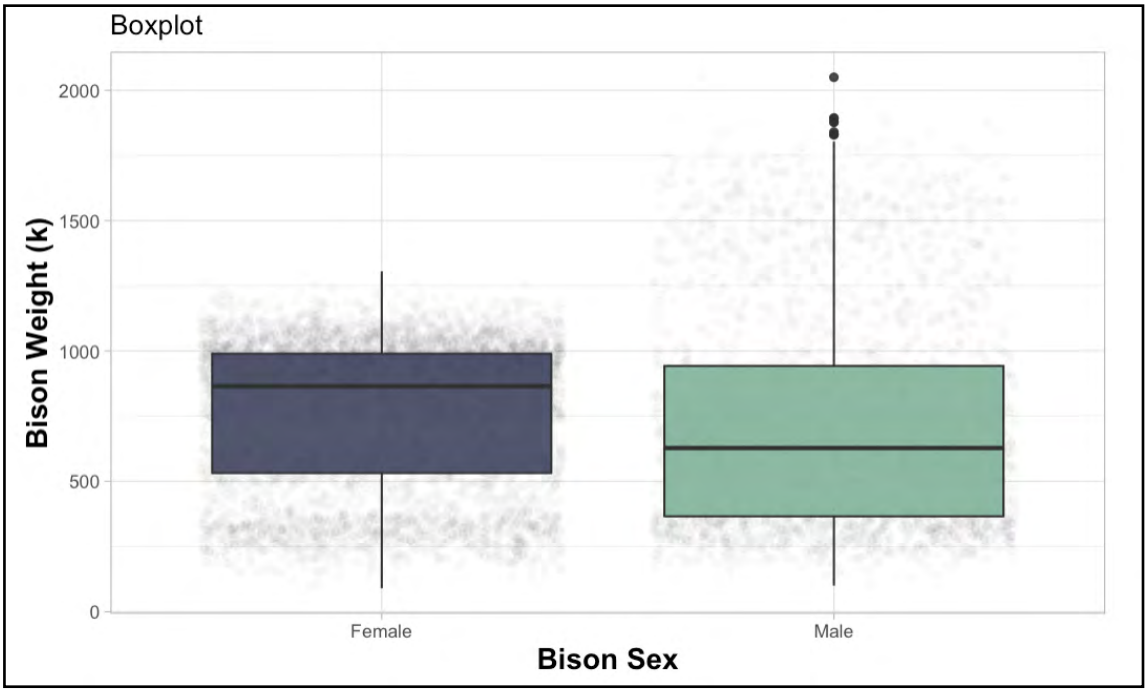
- The line that divides the box into 2 parts represents the median of the data. If the median is 10, it means that there are the same number of data points below and above 10.
- The ends of the box shows the upper (Q3) and lower (Q1) quartiles. If the third quartile is 15, it

means that 75% of the observation are lower than 15.
- The difference between Quartiles 1 and 3 is called the interquartile range (IQR).
- The extreme line shows Q3+1.5xIQR to Q1-1.5xIQR (the highest and lowest value excluding outliers).
- Dots (or other markers) beyond the extreme line shows potential outliers.

To create a boxplot using the similar naming and color scheme for Female and Male Bison in 3.3. we would use this code:

```r
knz_bison %>% # call in data
  plyr::mutate(animal_sex = revalue(animal_sex, # change animal_sex column values
                                    c("F" = "Female", # F to Female
                                      "M" = "Male"))) %>% # M to Male
  ggplot(aes(x = animal_sex, y = animal_weight, # for boxplot need both x&y
             fill = animal_sex)) + # fill in color by animal sex
  geom_jitter(alpha = 0.02, fill = "grey90") + # add points behind boxplot to see raw data
  geom_boxplot(alpha = .9) + # increase transparency
  scale_fill_manual(values = c("#3d405b", "#81b29a")) + # color code sex
  theme_light() + # light background color
  labs(x = "Bison Sex", y = "Bison Weight (k)", # change axis labels
       fill = "Sex") + # change legend label
  ggtitle("Boxplot") + # give title
  theme(axis.title = element_text(size = 14, face = "bold"), # change text font
        legend.title = element_text(face = "bold")) + # change text font
  guides(fill = FALSE) # remove legend for fill color
```

Screenshot 12
Activity 3.4:
To visualize these summary statistics by group (male vs female), we can use a boxplot.



Test your understanding: Based on the boxplot, do female or male buffalos have a higher median weight? Does this make sense based on this previous histogram? Make another boxplot but omit the **`geom_jitter()`** function, you can do this by placing a "#" before **`geom_jitter()`**. Explain which graph, the one with and without **`geom_jitter()`**, is more informative?

Activity 4. Investigating Species- and Site-Specific Variations in Aquatic Ecosystems

In this activity we will explore the dataset "Records for aquatic vertebrates (trout and salamanders) in Andrews Experimental Forest, Oregon (1987-present)". This study provides one of the longest continuous records of West Slope cutthroat trout (*Onchorhyncus clarki clarki*; Figure 4) populations and includes a monitoring program initiated in 1993 for Pacific Giant Salamanders (*Dicamptodon tenebrosus*). The study focuses on two creek segments: one near a clearcut forest area impacted by deforestation and another upstream in a rare 500-year-old coniferous forest. All captured trout and salamanders are measured and weighed, alongside channel measurements that assess environmental conditions. Aquatic systems face several conservation challenges, including pollution from agricultural runoff (such as pesticides and fertilizers), and industrial waste, all of which have been linked to declines in salamander populations. Invasive species further complicate conservation efforts, as non-native fish, amphibians, and invertebrates can prey on salamander eggs and compete for resources like food, breeding sites, and shelter, which contributes to the decline of native trout and salamander populations. While both species are sensitive to these environmental stressors, each has difference tolerances: trout may be more vulnerable to habitat degradation, while salamanders may be particularly sensitive to pollution. This makes it crucial to monitor their long-term populations trends to understand how each species responds to these challenges and to inform appropriate conservation strategies.

- We will use data on aquatic population monitoring to illustrate how you can visualize long-term trends but also create subplots to visualize species specific responses across time and space.
- More information on how data were collected can be found here: https://lter.github.io/lterdatasampler/reference/and_vertebrates.html.

Think critically: Use biological reasoning to explain why aquatic stream systems may be more sensitive than aquatic marine systems to anthropogenic stressors such as pollution?

```
```{r}
# part 1. install necessary packages
library(lterdatasampler) # package with built in LTER
library(tidyverse) # package for data wrangling
#install.packages("ggplot2") # if first time using
library(ggplot2) # make nice plots


## part 2. read in data
# we will use the built in dataset from lterdatasampler
and_vertebrates # aquatic vertebrate data
```
```

Test your understanding: What are the dimensions of the "`and_vertebrates`" dataset?

Figure 4. An adult cutthroat trout (*Onchorhyncus clarki clarki*) swims in Tryon Creek. Image credit: BES Portland via Flickr/ CC-BY-NC 2.0.

4.1 Long-term trends are crucial for assessing natural variations in populations and understanding how they respond to environmental changes. One way to visualize shifts in populations numbers and connect those changes with conservation concerns, is to use time-series plots, particularly scatterplots. A scatterplot shows the relationship between two numeric variables, with each dot representing a single observation.

To create a scatterplot for the total observations of trout and salamanders by year, begin by wrangling the data. This involves filtering species to your two species of interest using the function **`filter()`**, then group the data by species and year using **`group_by()`** and then summarizing it with **`tally()`**.

Note: When you filter observations in your dataset you will have to have the correct spelling and capitalization or else the filtering won't work. You can check the spelling of each level within a factor by using the **`unique()`** function (i.e., **`unique(and_vertebrates$species)`**. The "**`$`**" calls the column within the dataset. If you want to select both trout and salamander, then the Boolean operator "**`|`**" is needed.

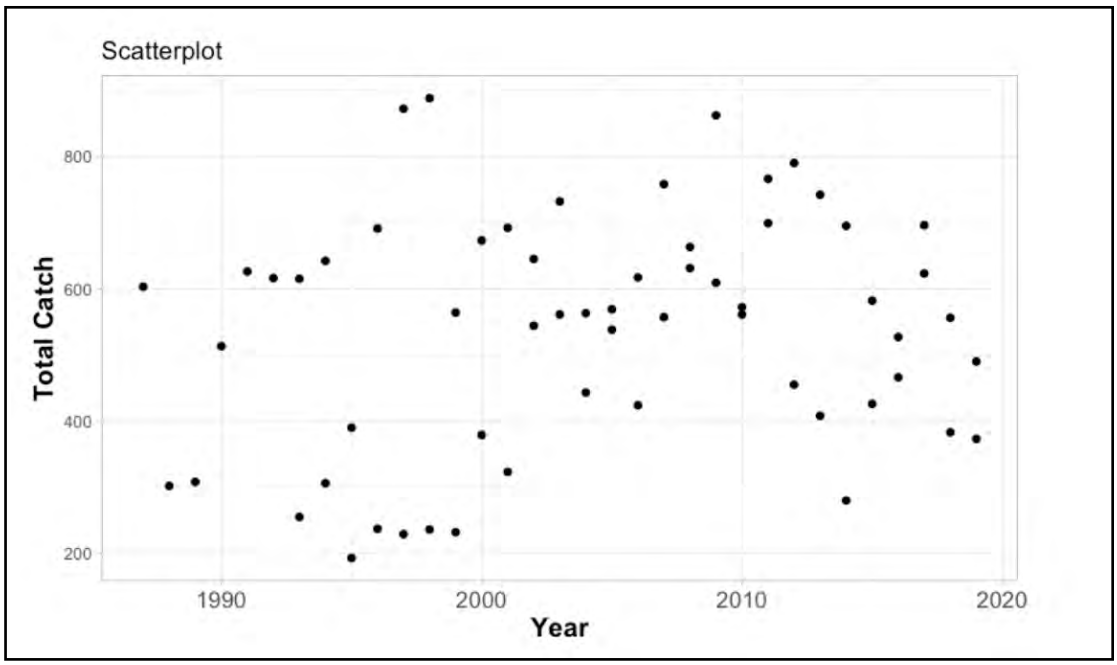
```
```{r}
# i. make data summary
and_vertebrates %>% # call in data
  filter(species == "Cutthroat trout" | # filter for trout OR
           species == "Coastal giant salamander") %>% # filter for salamander
  group_by(species, year) %>% # group by species and year
  tally() # tally observations per grouping
```
```

```
## # A tibble: 60 x 3
## # Groups:   species [2]
##    species                    year     n
##      <chr>                    <dbl> <int>
##  1 Coastal giant salamander          1993   255
##  2 Coastal giant salamander          1994   306
##  3 Coastal giant salamander          1995   193
##  4 Coastal giant salamander          1996   237
##  5 Coastal giant salamander          1997   229
##  6 Coastal giant salamander          1998   236
##  7 Coastal giant salamander          1999   232
##  8 Coastal giant salamander          2000   379
##  9 Coastal giant salamander          2001   323
## 10 Coastal giant salamander          2002   645
## # i 50 more rows
```

Once the data is in the right format to address the question "Do populations of trout and salamanders change over time?", pipe the data into **`ggplot()`** and use **`geom_point()`** to create a scatterplot. [Screenshot 13]

```
```{r}
# ii. pipe data summary into ggplot
and_vertebrates %>% # call in data
  filter(species == "Cutthroat trout" | # filter for trout OR
           species == "Coastal giant salamander") %>% # filter for salamander
  group_by(species, year) %>% # group by species and year
  tally() %>%  # tally observations per grouping
  ggplot(aes(x = year, y = n)) + # x axis is year, y axis is n from tally()
  geom_point() + # observations are points
  theme_light() + # light background color
```
```

```
  ggtitle("Scatterplot") + # give title
  labs(x = "Year", y = "Total Catch") + # change axis labels
  theme(axis.title = element_text(size = 14, face = "bold"), # change title font
        axis.text.x = element_text(size = 12)) # change x axis text font
```



Screenshot 13 Activity 4.1: Once the data is in the right format to address the question "Do populations of trout and salamanders change over time?", pipe the data into **`ggplot()`** and use **`geom_point()`** to create a scatterplot.

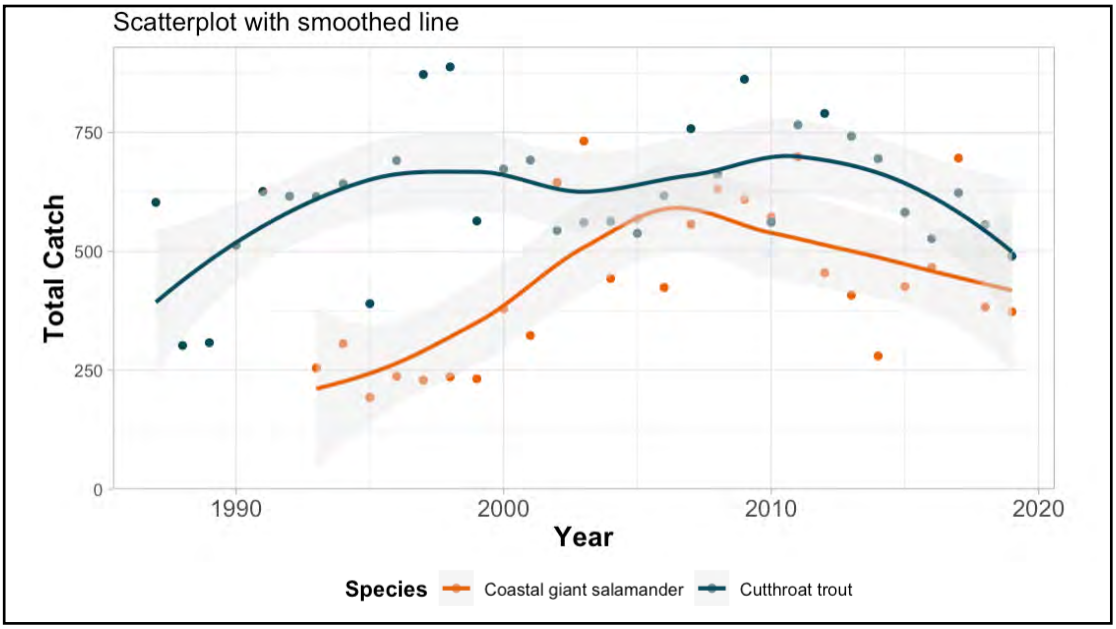
Test your understanding: Based on this scatterplot, what year saw the highest catch of one species? Is there a general trend within this plot? How would you make this plot more informative? What are some questions you have that this plot either doesn't answer or makes you want to dig further?

4.2 While the graph in section 4.1 shows the overall relationship between the total catch of trout and salamanders over time, it doesn't provide insights into species-specific trends. One way to differentiate species in a plot is by assigning each species a distinct color, similar to how we differentiated male and female bison, which is achieved setting "**`color = species`**". However, to make the trends more apparent, it's also useful to add trend lines using **`geom_smooth()`**, which help highlight species-specific responses over time.

In the plot below, you'll see a smoothed line for each species, offering a clearer and more interpretable view of the relationship between the variables. By default, **`geom_smooth()`** fits a loess (locally weighted scatterplot smoothing) curve and includes a 95% confidence interval shaded in grey, which provides additional context for the trend [Screenshot 14]. If you are interested in learning more specifics behind the methods used to generate trend lines (and much, much more), we recommend the Data Visualization course by Andrew Irwin available at https://andrewirwin.github.io/data-visualization/index.html or the reference documentation provided for ggplot2 available at https://ggplot2.tidyverse.org/reference/geom_smooth.html.

```{r}
and_vertebrates %>% # call in data
  filter(species == "Cutthroat trout" | # filter for trout OR
           species == "Coastal giant salamander") %>% # filter for salamander
  group_by(species, year) %>% # group by species and year
  tally() %>% # tally observations per grouping
  ggplot(aes(x = year, y = n)) + # x axis is year, y axis is n from tally()
  geom_point(aes(color = species)) + # color observations by species
```

```
  geom_smooth(aes(color = species), # add a smoothed trend line and color by species
              fill = "grey90",# fill 95% CI with grey shade
              method = "loess") + # use loess method for smoothed line
  scale_color_manual(values = c("#e36414", "#0f4c5c")) + # assign colors for 2 species
  theme_light() + # light background color
  ggtitle("Scatterplot with smoothed line") +  # give title
  labs(x = "Year", y = "Total Catch", # change axis labels
       color = "Species") + # change legend title
  theme(axis.title = element_text(size = 14, face = "bold"),# change title font
        axis.text.x = element_text(size = 12), # change x axis text font
        legend.title = element_text(face = "bold"), # change legend font
        legend.position = "bottom") # change legend positioning
```



Screenshot 14
Activity 4.2:
By default, `geom_smooth()` fits a loess (locally weighted scatterplot smoothing) curve and includes a 95% confidence interval shaded in grey, which provides additional context for the trend.

Test your understanding: Based on this scatterplot, which species has a trend line that is on average higher than the other species? Make another plot but change the `geom_smooth()` argument "loess" to "lm". Does the pattern between the two species still hold? Which species has a steeper slope? Note: Review Chapter 16 (Linear Models) and Chapter 17 (GAM and LOESS Smoothing) in Andrew Irwin's Data Visualization course (https://andrewirwin.github.io/data-visualization/index.html) or the "Methods" section of the reference documentation provided for '`ggplot2`' (https://ggplot2.tidyverse.org/reference/geom_smooth.html), if you'd like to understand these smoothing methods better.

4.3 We've observed species-specific differences in population trends over time between trout and salamanders, but could these variations be more influenced by site-specific environmental factors rather than broader population trends within a single watershed? The vertebrates were collected at two distinct sites, one near a clearcut forest and the other near an old-growth forest. As conservation biologists, we often make assumptions about how habitat quality impacts species, but it's important to consider these site-specific influences.

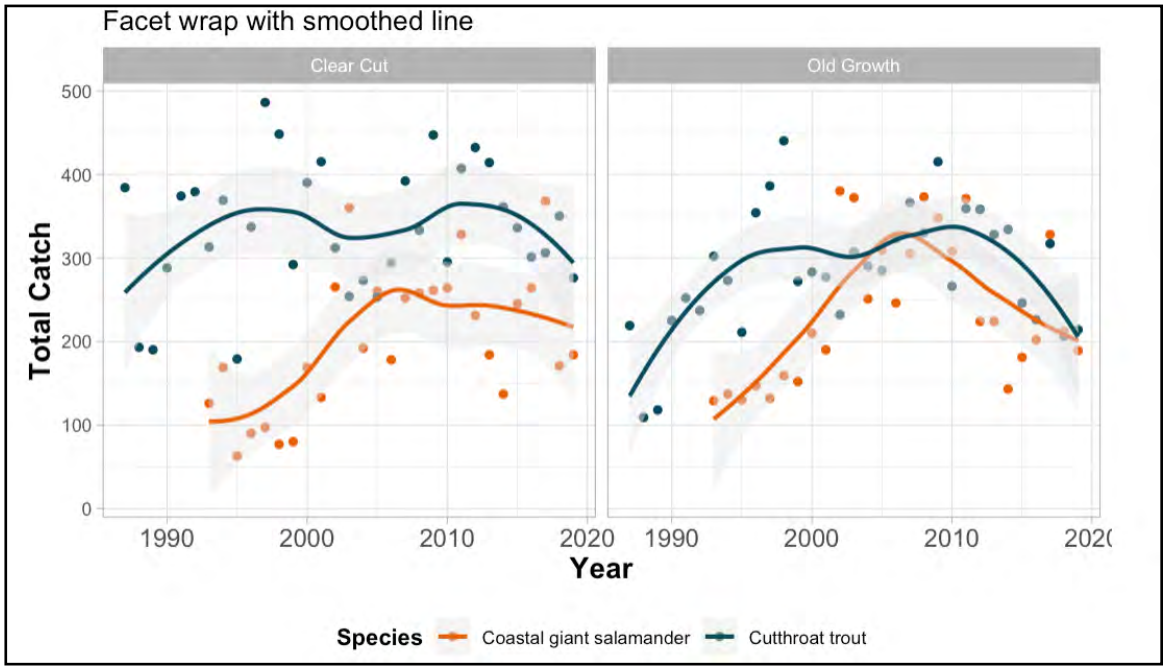
To better understand population dynamics, it would be valuable to visualize both species-specific and site-specific trends. To incorporate the site factor (clearcut vs old-growth forests), we can use `facet_wrap()` to create multiple subplots, one for each site. This function will allow us to separate the data by site, providing a clearer view of how species trends vary across different environmental conditions. [Screenshot 15]

```{r}
# i. make data summary
and_vertebrates %>% # call in data
  filter(species == "Cutthroat trout" | # filter for trout OR
           species == "Coastal giant salamander") %>% # filter for salamander
  plyr::mutate(section = revalue(section, # revalue levels in factor section
                                 c("CC" = "Clear Cut", # change CC to clear cut
                                   "OG" = "Old Growth"))) %>% # change OG to old growth
  group_by(section, species, year) %>% # group by section, species and year
  tally() # tally observations per grouping
```

```
## # A tibble: 120 x 4
## # Groups:   section, species [4]
##    section   species                year     n
##      <chr>   <chr>                  <dbl> <int>
##  1 Clear Cut Coastal giant salamander      1993   126
##  2 Clear Cut Coastal giant salamander      1994   169
##  3 Clear Cut Coastal giant salamander      1995   63
##  4 Clear Cut Coastal giant salamander      1996   90
##  5 Clear Cut Coastal giant salamander      1997   97
##  6 Clear Cut Coastal giant salamander      1998   77
##  7 Clear Cut Coastal giant salamander      1999   80
##  8 Clear Cut Coastal giant salamander      2000   169
##  9 Clear Cut Coastal giant salamander      2001   133
## 10 Clear Cut Coastal giant salamander      2002   265
## # i 110 more rows
```

```{r}
# ii. pipe data summary into ggplot
and_vertebrates %>% # call in data
  filter(species == "Cutthroat trout" | # filter for trout OR
           species == "Coastal giant salamander") %>% # filter for salamander
  plyr::mutate(section = revalue(section, # revalue levels in factor section
                                 c("CC" = "Clear Cut", # change CC to clear cut
                                   "OG" = "Old Growth"))) %>% # change OG to old growth
  group_by(section, species, year) %>% # group by section, species and year
  tally() %>% # tally observations per grouping
  ggplot(aes(x = year, y = n)) + # x axis is year, y axis is n from tally()
  geom_point(aes(color = species)) + # color observations by species
  geom_smooth(aes(color = species), # add a smoothed trend line and color by species
              fill = "grey90", # fill 95% CI with grey shade
              method = "loess") + # use loess method for smoothed line
  scale_color_manual(values = c("#e36414", "#0f4c5c")) +  # assign colors for 2 species
  theme_light() + # light background color
  ggtitle("Facet wrap with smoothed line") + # give title
  labs(x = "Year", y = "Total Catch", # change axis labels
       color = "Species") + # change legend title
  theme(axis.title = element_text(size = 14, face = "bold"), # change title font
        axis.text.x = element_text(size = 12), # change x axis text font
        legend.title = element_text(face = "bold"), # change legend font
        legend.position = "bottom") + # change legend positioning
  facet_wrap(~section) # create subplots based on section
```

Test your understanding: Based on this `facet_wrap` plot, are there differences between clear cut and old growth forest? What are the differences and what are biological reasons for the differences between both species and both locations? What conservation management strategies would you expect to be effective for both species and both locations?



Screenshot 15
Activity 4.3:
We can use `facet_wrap()` to create multiple subplots, one for each site.

Visualizing data is a powerful tool to communicate with scientific and broad audiences. Although there are many ways to visualize data, it is important to remember some common rules to enhance the study results and ensure the correctness of the message (e.g., color-blind friendly colors, clear axis label, etc.). See Additional Resources for how you can use similar graphing tools, like ggplot, to apply to your own data!

CONCLUSION

Our world is rich with data, thanks to the dedication of thousands of researchers who have contributed their time and energy to the success of long-term ecological research stations. By the end of this module, you will have developed foundational skills, such as setting up RStudio on your computer, uploading datasets, exploring data collected by others, and wrangling it to focus on areas of interest. You will also learn how to visualize data to identify trends that can be further explored with statistical techniques learned in other modules. This process will serve as a crucial step in your career as a conservation biologists, empowering you to document and analyze the complexities of a dynamic and ever-changing world.

KEY TERMS

To be honest, this section may seem incredibly dry when you are first learning R. However, mastery of this terminology will make coding and asking for coding help much easier. For example, if you are trying to find out "How to select for trout within your fish species grouping", copying and pasting that question into a web browser may not give you many search hits. But if you reframed your question as "How to select for a level within a factor in R" (where trout is a level within the factor fish species), you will have more responses to look through.

Some important key terms for you to keep and refer to when working in R (starting from need to know when opening RStudio to manipulating data):

- **Working directory**: the folder on your computer linked to your current R session, where you import data from and save files to.

- **CSV file**: a type of file extension commonly used to import data into R, where the values of different variables are compressed together (i.e., a string, or line of values per row) and separated only by commas. R can also accept Excel (.xlsx) files, but we do not recommend it as formatting errors are harder to avoid.

- **Script**: similar to a text editor (e.g., Microsoft Word). This is where you write and save your code for future reference. It contains a mix of code and comments and is saved as a simple text file that you can easily share so that anyone can reproduce your work.

- **Comment**: a bit of text in a script that starts with a hashtag (i.e., #) and is not read as a command. Comments make your code readable to other people. Use them to create sections in your script and to annotate each step of your analysis.

- **Function**: code that performs an action, and really how you do anything in R. Usually takes an input (i.e., argument), does something to it and returns an output (i.e., a test result or plot). There are functions for importing, converting, and manipulating data.

- **Argument**: an element of a function, either essential or optional, that informs or alters how the function works. For instance, it can be a file path where the function should import from or save to `> function(argument) = read.csv(file = "file.csv")`.

- **Command**: a chunk of code that performs an action, typically contains one or more functions. You run a command by pressing "Run" or using a keyboard shortcut like "Cmd+Enter" or "Crtl+R".

- **Package**: a bundle of functions that provide functionality to R. Many packages are automatically built into R, others you have to download for specific needs `> install.packages(package = "tidyverse")`.

- **Object**: the building blocks of R. If R was a spoken language, functions would be verbs (actions) and objects would be nouns (the subject of these actions). Objects are called by typing their name without quotation marks. Objects store data and can take different forms.

- **Variable**: an item that is stored in your Global Environment (top right corner). Variables tend to be on the left-hand side of "`<-`" , whereas the information used to create the variable is on the right-hand side of "`<-`".

- **Data frame**: a type of R object which consists of many rows and columns. Think Excel spreadsheet. Usually the columns are different variables (e.g., age, weight, wingspan), and rows are observations of these variables (e.g., bird1, bird2, bird3).

- **Class**: the type of data contained in a variable. Usually a character (i.e., text or words), numeric (i.e., number), integer (i.e., whole number), or factor (i.e., grouping values, useful when you have multiple observations for sites or treatments in your data).

- **Boolean operators**: tools for filtering data. The syntax behind some important Boolean operators:
  - `>` : greater than
  - `<` : less than
  - `==` : equals to
  - `!=` : not equal to

- >= : greater than or equal to
- <= : less than or equal to

Although activity 3 and 4 are focused on **`ggplot()`**, base-R is still great for plotting. Below is a list of the basic functions for plotting in base-R. By using "**`?`**" in R, you can figure out the details of each of these plotting functions.

- **`plot()`** : plot x and y against each other
- **`hist()`** : make a histogram of your data
- **`scatterplot()`** : make a scatterplot of x and y
- **`abline()`** : add a straight line to your current plot
- **`lines()`** : add more lines to your current plot
- **`curve()`** : plot a function over some range
- **`boxplot()`** : make a boxplot to compare different groups of your data
- **`densityPlot()`** : fit a density kernel to your data
- **`pie()`** : make a pie chart
- **`barplot()`** : make a bar graph

Note: The glossary definitions come from "Stats from Scratch" which can be found here: https://ourcodingclub.github.io/course/stats-scratch/index.html.

ADDITIONAL RESOURCES

Developing Biological Questions

- Indiana University Bloomington library provides flow charts and videos on this topic: https://guides.libraries.indiana.edu/c.php?g=820631&p=7641516.
- George Mason University writing center provides a basic framework on how to write a research question: https://writingcenter.gmu.edu/writing-resources/research-based-writing/how-to-write-a-research-question.

Styling an .Rmd to Make Professional Pdf Reports

- The ultimate RMarkdown guide has everything you will need to know to style your document in a similar way to word processors to make it more aesthetically pleasing: https://bookdown.org/yihui/rmarkdown/installation.html.
- See this quick & interactive guide if you want to learn in a more interactive format instead of static text: https://commonmark.org/help/tutorial/index.html.

Additional Resources on Tidyverse

- Basic tidyverse concepts: https://homerhanumat.github.io/r-notes/tidyverse-concepts.html.
- Tidyverse with animations: https://www.garrickadenbuie.com/project/tidyexplain/.
- Tidyverse cookbook: https://rstudio-education.github.io/tidyverse-cookbook/tidy.html.
- Check out Tidyverse Style Guide for some tips and additional explanations of the different syntax: https://style.tidyverse.org/index.html. Specifically, pipes "%>%" and "|":
  - %>% can be thought of as "and then do this", it links sequences of actions together (i.e. filter and then summarize).
  - "|" can be thought of as "or", it is commonly used within filter function (i.e., select the month of March or April).

Visualization Tips for Effective Plots

- See the Fundamentals of Data Visualization for further discussions on what makes a good graph: https://clauswilke.com/dataviz/.
- The Graphics Principles Cheat Sheet is a reference to find core coding functions quickly: https://graphicsprinciples.github.io/cheatsheet.html.
- With the Color Blindness Simulator you can upload your figure to see how your figure looks with different types of color blindness: https://www.color-blindness.com/coblis-color-blindness-simulator/.

Professional Development Opportunities Within the LTER System

To find the most up-to-date opportunities visit lternet.edu. A potential way to keep track of upcoming professional development opportunities, you can subscribe to the LTER Monthly Newsletter or Community Platform. Some specific opportunities to look for include:

- Research Experience for Undergraduates (REU) which allows for active research participation by undergraduate students across many LTER sites. Visit the "Undergraduate Research Opportunities" page on the LTER site to find potential research experiences: https://lternet.edu/education-and-training/undergraduate/.
- Partnerships for Undergraduate Research (SPUR) Fellowship which is sponsored by SEEDS, a program of the Ecological Society of America, with the goal of broadening participation in ecology.

TROUBLE SHOOTING

Problem 1. Learn How to Spot and Correct Common Coding Errors

R can be annoyingly specific. Spelling, capitalization, and punctuation matters. If your code is throwing errors, we highly recommend checking these three common syntax errors.

1. Spelling
   a. `install.packages('readr')` is correct
   b. `instal.packages('readr')` is incorrect
2. Capitalization
   a. `install.packages('readr')` is correct
   b. `Install.Packages('readr')` is incorrect
3. iii. Punctuation
   a. `install.packages('readr')` is correct
   b. `install_packages('readr')` is incorrect

Problem 2. ggplot is Not Showing the Data You Want

The three main components of a `ggplot` in R are: data, aesthetics (`aes`), and geometries (`geom`). If you don't have all three components your graph won't turn out the way you want it to. Below is an example of what each component of `ggplot` looks like. [Screenshot 16]

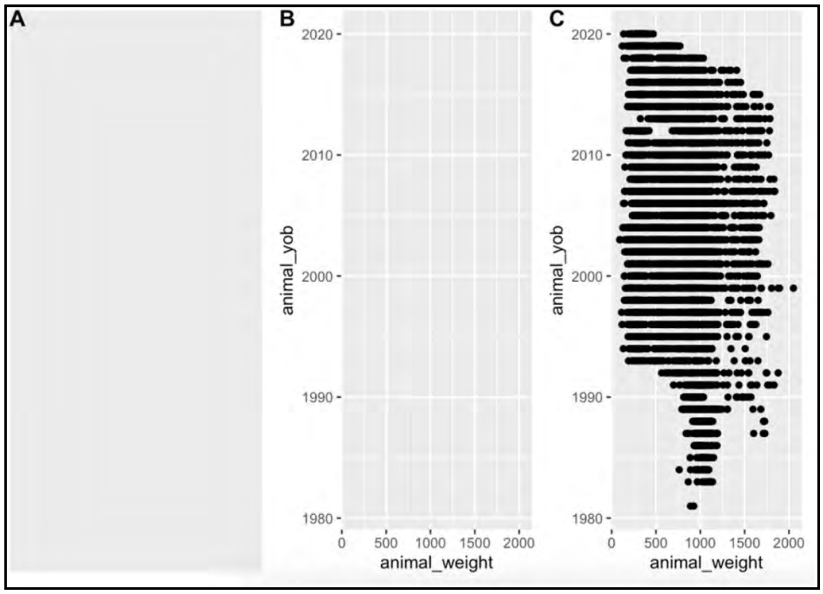
```
ggplot1 <- ggplot() # plot A

ggplot2 <- ggplot(knz_bison, aes(x = animal_weight, y = animal_yob)) # plot B

ggplot3 <- ggplot(knz_bison, aes(x = animal_weight, y = animal_yob)) + # plot C
  geom_point()
```

```
plot_grid(ggplot1, ggplot2, ggplot3, # call all plots
          ncol = 3, # put plots in 3 columns
          labels = "AUTO") # add labels to plots # plot C is best plot because it has labels and
data. The layers necessary to make plot C is ggplot() + aes(data) + geom_point()
```

Screenshot 16
Probelm 2:
The three main components of a **ggplot** in R are: data, aesthetics (**aes**), and geometries (**geom**). If you don't have all three components your graph won't turn out the way you want it to.



Quick Tips for Making Pretty Plots

Use the correct plot for your data
- 1 numeric = histogram
- 1 numeric 1 factor = boxplot
- 2 numeric = scatterplot

Know how to improve graphs with informative labels
- Using `labs()` and `ggtitle()` functions

Learn tricks to make more visually appealing graphs
- `"alpha"` for transparency
- `"fill"` to distinguish groups
- `"theme_*()"` to change plot background
  - `theme_bw()`
  - `theme_light()`
  - `theme_classic()`
- `"theme()"` to change text (axis titles & texts), rectangle (`facet_wrap`) and legend features
  - `theme(axis.text = element_text())`
  - `theme(strip.background = element_rect())`
  - `theme(legend.position = "bottom")`

ACKNOWLEDGMENTS

Material for this module was funded in part by a University of California Santa Barbara (UCSB) teaching grant "Hybrid, Online & Technology-enhanced (HOT) Teaching Mini-grant" awarded to Cheryl J. Briggs and Samantha Sambado in the Fall of 2023. The original material was intended for UCSB undergraduates enrolled in the course "Ecology, Evolution, Marine Biology 146: Biometry".

Contributors to the computer coding material include Dr. Mark Wilbur, Dr. Tatum Katz, Dr. Caroline Owens, Dr. Terra Dressler, Dr. Imani Russel, Jacob Weverka, Kacie Ring, and Caitlin Nordheim-Maestas. SS acknowledges the invaluable resource of 'ltersampler' developed by Dr. Allison Horst and Dr. Julien Brun. SS thanks Dr. Martha Groom, Nadav Gazit, and Dr. Suzanne K. Macey for their invaluable advice and insights throughout the editorial process.

REFERENCES

Brodie, R. J., Roberts, B., Espinosa, J. I., Heilman, K., Borgianini, S. A., Welch, J. M., & Reinsel, K. A. (2017). Seasonal and latitudinal variations in the energy reserves of the mud fiddler crab *Uca pugnax*: Implications for the response to climate change. Aquatic Biology, 26, 113–123. https://doi.org/10.3354/ab00683

Darnell, M. Z., & Darnell, K. M. (2018). Geographic variation in thermal tolerance and morphology in a fiddler crab sister-species pair. Marine Biology, 165(2), 26. https://doi.org/10.1007/s00227-017-3282-y

Décima, M., Steinberg, D., & Sala, L. (2024). Importance and unanticipated use of biological collections in long-term ecological research. DataBits. https://lternet.edu/stories/importance-and-unanticipated-use-of-biological-collections-in-long-term-ecological-research/

Dee, L. E., Ferraro, P. J., Severen, C. N., Kimmel, K. A., Borer, E. T., Byrnes, J. E. K., Clark, A. T., Hautier, Y., Hector, A., Raynaud, X., Reich, P. B., Wright, A. J., Arnillas, C. A., Davies, K. F., MacDougall, A., Mori, A. S., Smith, M. D., Adler, P. B., Bakker, J. D., … Loreau, M. (2023). Clarifying the effect of biodiversity on productivity in natural ecosystems with longitudinal data and methods for causal inference. Nature Communications, 14(1), 2607. https://doi.org/10.1038/s41467-023-37194-5

Ehrlén, J., & Morris, W. F. (2015). Predicting changes in the distribution and abundance of species under environmental change. Ecology Letters, 18(3), 303–314. https://doi.org/10.1111/ele.12410

Erb, L. P., Ray, C., & Guralnick, R. (2011). On the generality of a climate-mediated shift in the distribution of the American pika (*Ochotona princeps*). Ecology, 92(9), 1730–1735. https://doi.org/10.1890/11-0175.1

Galbreath, K. E., Hafner, D. J., & Zamudio, K. R. (2009). When cold is better: Climate-driven elevation shifts yield complex patterns of diversification and demography in an alpine specialist (American Pika, *Ochotona princeps*). Evolution, 63(11), 2848–2863. https://doi.org/10.1111/j.1558-5646.2009.00803.x

Hughes, B. B., Beas-Luna, R., Barner, A. K., Brewitt, K., Brumbaugh, D. R., Cerny-Chipman, E. B., Close, S. L., Coblentz, K. E., De Nesnera, K. L., Drobnitch, S. T., Figurski, J. D., Focht, B., Friedman, M., Freiwald, J., Heady, K. K., Heady, W. N., Hettinger, A., Johnson, A., Karr, K. A., … Carr, M. H. (2017). Long-term studies contribute disproportionately to ecology and policy. BioScience, 67(3), 271–281. https://doi.org/10.1093/biosci/biw185

Jones, J. A., & Driscoll, C. T. (2022). Long-term ecological research on ecosystem responses to climate change. BioScience, 72(9), 814–826. https://doi.org/10.1093/biosci/biac021

Kao, R. H., Gibson, C. M., Gallery, R. E., Meier, C. L., Barnett, D. T., Docherty, K. M., Blevins, K. K., Travers, P. D., Azuaje, E., Springer, Y. P., Thibault, K. M., McKenzie, V. J., Keller, M., Alves, L. F., Hinckley, E.-L. S., Parnell, J., & Schimel, D. (2012). NEON terrestrial field observations: Designing continental-scale, standardized sampling. Ecosphere, 3(12), 1–17. https://doi.org/10.1890/ES12-00196.1

Kuebbing, S. E., Reimer, A. P., Rosenthal, S. A., Feinberg, G., Leiserowitz, A., Lau, J. A., & Bradford, M. A. (2018). Long-term research in ecology and evolution: A survey of challenges and opportunities. Ecological Monographs, 88(2), 245–258. https://doi.org/10.1002/ecm.1289

Lindenmayer, D. B., Likens, G. E., Andersen, A., Bowman, D., Bull, C. M., Burns, E., Dickman, C. R., Hoffmann, A. A., Keith, D. A., Liddell, M. J., Lowe, A. J., Metcalfe, D. J., Phinn, S. R., Russell-Smith, J., Thurgate, N., & Wardle, G. M. (2012). Value of long-term ecological studies. Austral Ecology, 37(7), 745–757. https://doi.org/10.1111/j.1442-9993.2011.02351.x

Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. McP., Elston, D. A., Scott, E. M., Smith, R. I., Somerfield, P. J., & Watt, A. D. (2010). Long-term datasets in biodiversity research and monitoring: Assessing change in ecological communities through time. Trends in Ecology & Evolution, 25(10), 574–582. https://doi.org/10.1016/j.tree.2010.06.016

Morecroft, M. D., Bealey, C. E., Beaumont, D. A., Benham, S., Brooks, D. R., Burt, T. P., Critchley, C. N. R., Dick, J., Littlewood, N. A., Monteith, D. T., Scott, W. A., Smith, R. I., Walmsley, C., & Watson, H. (2009). The UK Environmental Change Network: Emerging trends in the composition of plant and animal communities and the physical environment. Biological Conservation, 142(12), 2814–2832. https://doi.org/10.1016/j.biocon.2009.07.004

Sauer, J. R., Link, W. A., Ziolkowski, D. J., Pardieck, K. L., & Twedt, D. J. (2019). Consistency counts: Modeling the effects of a change in protocol on Breeding Bird Survey counts. The Condor, 121(2), duz009. https://doi.org/10.1093/condor/duz009

Williams, J. J., & Newbold, T. (2020). Local climatic changes affect biodiversity responses to land use: A review. Diversity and Distributions, 26(1), 76–92. https://doi.org/10.1111/ddi.12999

Assessing Land Cover in Forest Reserves Using Remote Sensing Tools

Carlos A. Morales-Ramirez[i], Brian Carroll[ii], Sue Neal[iii], and Benjamin Nein[i]

[i]Department of Geography & Planning, West Chester University, West Chester, PA, USA; [ii]Department of Anthropology, Temple University, Philadelphia, PA, USA; [iii]Department of Political Sciences, Arkansas State University, Jonesboro, AR, USA

DOI: https://doi.org/10.5531/cbc.linc.14.1.6 | Supplementary: http://doi.org/10.5531/cbc.ncep.0187

ABSTRACT

Tropical rainforests are home to many biological species that are currently facing threats due to deforestation and other factors. As human populations grow, institutional and policy frameworks change, new demands for land and natural resources emerge, terrestrial environments continue to be modified impacting the habitats of many species and the environmental services that these habitats provide to their populations. Remote sensing technologies are a great resource that help researchers detect deforestation, make informed decisions, and monitor forest regrowth. Through this case study-based exercise, students use remote sensing imagery and spatial analysis software to identify and map potential deforestation hotspots at a forest reserve in Ghana. Students then consider the patterns in the results, the implications of land cover changes, and then suggest policy recommendations through a discussion and presentation.

LEARNING OBJECTIVES

After this exercise students will be able to:
1. Reflect on the ethical and justice challenges of using remote sensing technologies for conservation.
2. Acquire proficiency in utilizing spatial analysis programs: Collect Earth Online, ArcGIS Online, and QGIS.
3. Develop skills in distinguishing changes in land cover through analysis of satellite imagery.
4. Interpret findings to identify potential deforestation hotspots within a forest reserve.
5. Critically evaluate implications of land cover changes to inform real-world conservation efforts.

INTRODUCTION

Over half of our planet's biodiversity live in tropical rainforests (Wright, 2005). These are important ecosystems for many species that are currently affected by deforestation, which leads to habitat loss (Lewis et al., 2015; Curtis et al., 2018). There are many factors that contribute to the modification of many terrestrial environments, including forests (Ellis et al., 2021). One of the major threats of habitat modification is deforestation (Curtis et al., 2018), which is causing major changes in land cover all over the world (Newbold et al., 2015). This process may negatively impact biodiversity with the amphibian population being one of the most vulnerable groups of species (Blaustein et al., 2011). In tropical rainforests, where amphibians inhabit are particularly susceptible to the negative impacts of deforestation (Barlow et al., 2018; de Oliveira Roque et al., 2018).

As the quality of satellite imagery advances, remote sensing technologies have enhanced our capabilities of monitoring and mapping deforestation and analyzing spatial patterns of change on the landscape (Hansen et al., 2013; Haddad et al., 2015). In recent years, remote sensing has emerged as a leading tool for detecting deforestation because it offers timely and comprehensive data. In large and remote areas, these technologies allow researchers to provide important insight for conservation actions, decision making, sustainable management and practices, and policy formulation that would

otherwise be challenging to do in the field. One key advantage of remote sensing is the capability of many satellite platforms to provide consistent and precise data throughout great spatial extents and prolonged periods of time. There are satellites that are equipped with special sensors that can capture imagery in forests at regular intervals of time, allowing researchers to identify land cover changes, including deforestation (Gibbs et al., 2010).

By analyzing high quality satellite imagery through remote sensing techniques, researchers can identify specific drivers of deforestation such as agricultural expansions, logging, and infrastructure development, allowing them to propose interventions to mitigate them (Asner et al., 2005). Additionally, understanding the remote sensing data obtained, researchers can quantify forest loss due to deforestation and its impact on biodiversity, carbon storage, and even local communities (Hansen et al., 2013).

As is true for other data collection methods, it is crucial to consider limitations and potential drawbacks specific to remote sensing methods and data. For example, we can consider the temporal limitations of remote sensing data and ask: how far do these records go? To what time period can we compare? How has data quality (e.g., resolution) improved? Importantly, there are ethical and human rights concerns when it comes to remote sensing. Governments, as well as other groups and entities such as corporations, use remote sensing technologies in the processes of land dispossession, surveillance, and warfare. When we look at these data, we can consider: who owns the data? Who is involved in the process of generating the data (and how)? Who gets to make decisions about data gathering and usage? How will this data be used? While we might not always be able to answer these (and many more we should consider), taking time to reflect on how we use such technology and methods is an important aspect of ethical scientific work and, unfortunately, not often engaged in in the conservation biology peer-reviewed literature (York et al., 2023).

With these considerations in mind, remote sensing remains a valuable tool for deforestation detection that enables informed decision-making strategies and effective conservation actions. Through this exercise, students explore existing remote sensing data and platforms and analyze deforestation and land cover changes. Each student will be tasked with evaluating a predetermined number of plots within the provided area. Students will categorize subplots within each plot into three distinct categories based on their observations: “Closed Forest”, “Other Vegetated” areas, or “Developed/Barren.” Upon completion of the analyses, students will download the data and utilize ArcGIS Online (or QGIS) to generate a map illustrating their classifications. The exercise ends with group discussions and subsequent presentations reporting back on the group’s results, evaluation of implications, and policy suggestions.

EXERCISE

Discussion: Ethics in Conservation Remote Sensing

To learn more on the ethical considerations in remote sensing and suggested recommendations and resources, we recommend the review by York et al. (2023). After reading the article, what are some ethical and justice challenges of using remote sensing technologies for conservation?

Case Study: Save the Frogs!

In Ghana, a country in Western Africa, forest reserves are under considerable threat from deforestation, largely driven by agricultural activities. Rural communities in Ghana depend extensively on farming

and forest resources to fulfill their daily needs. Studies have highlighted agricultural expansion as a key factor in deforestation, particularly in dense forests. A significant portion of the converted forest land is used for cocoa cultivation, which exacerbates local deforestation (Brobbey et al., 2020; Pendrill et al., 2022). While this issue impacts many species, this discussion will focus specifically on the critically endangered Giant Squeaker Frog (*Arthroleptis krokosua*).

You just started working for Save the Frogs! (https://savethefrogs.com/), an amphibian conservation organization. Your task is to identify potential deforestation in a forest reserve in Ghana that is affecting the habitat of the last giant squeaker frog. Using satellite imagery and mapping tools, you will classify land cover and analyze changes over space. Your efforts to pinpoint significant deforestation will be crucial for targeted conservation strategies aimed at protecting these endangered frogs and their environment. By examining land cover changes, you will gain insights into how human activities like logging and agricultural expansion impact ecosystems. This understanding is essential for developing effective strategies to mitigate negative effects and encourage sustainable land management. Identifying deforestation hotspots will support both immediate conservation efforts and broader environmental protection goals. Additionally, this work project offers you hands-on experience with satellite data analysis, an important skill for your professional development and effective environmental monitoring and management.

Save the Frogs has asked you to complete your first project with the team, you will:
- Analyze satellite imagery in the online platform Collect Earth Online
- Download your results
- Upload and map your results in either ArcGIS or QGIS
- Discuss your results and maps with you peers
- And create outreach materials

- **Reflect:** Before you begin going through the instructions think about the role forests play in supporting biodiversity and maintaining the health of the ecosystems. Consider why forests are crucial for species like the giant squeaker frog, or any other species you are familiar with. Why do you think protecting forested areas is essential for species conservation? Why is this particularly important for endangered species that live in the forest? Are there any benefits or negative impacts of forest protection to humans? Compare the importance of forests between humans and other species.

- **Consider:** As you get ready to analyze the satellite imagery, consider what signs of deforestation you might encounter. How can these indicators assist you in pinpointing key areas that need conservation efforts?

Part 1: Satellite Imagery Analysis Using Collect Earth Online

In this exercise, you will analyze satellite imagery to detect potential land changes on the Sui River Forest Reserve in Ghana. This reserve is home to many species including the threatened Giant Squeaker Frog (*Arthroleptis krokosua*). Check out the Save the Ghana Frogs website (https://www.saveghanafrogs.org/our-work) for more information on the species.

Through the steps provided below, you will answer questions that help you prepare for the final presentation of your results. Write down your answers to these questions and take notes on patterns or any peculiarities you see throughout your analysis.

Step 1: Open Collect Earth Online (CEO)

Navigate to https://collect.earth/ and click "Open CEO." The resulting screen should look like Screenshot 1.

Step 2: Register

Click "Login/Register" on the top right of the webpage and click "Register." Create an account with your institutional email address and create a password. Complete the form and click "Register." Sometimes, the screen will not change after clicking "Register" (this issue occurs frequently when using Google Chrome). If you click "Register" a second time and receive a notice that an account has already been created for your email address, then your account has been created properly. Check your email for your registration confirmation. [Screenshot 2]

- **Note:** Your instructor may have you skip registration, if they have decided to create a project that is open to the public.

Step 3: Confirm registration

Confirm your new CEO account through the link you receive in your email. The link will bring you to the CEO login page. If the link brings you to the CEO homepage instead, click "Login" on the top banner. [Screenshot 3]

Step 4: Log in

At times, logging in will cause the website to send you to the account registration screen from step 2. If this happens, click the "CEO" button in the upper left side of the screen. You should be brought to CEO's Institutions page and be properly logged in. [Screenshot 3]

- **Note:** After students have registered, the instructor will need to manually add the students to the CEO project for this exercise, therefore, you may need to pause your progress here to allow for that administrative step (as described in the instructor's exercise notes, provided as a supplementary material to this module: http://doi.org/10.5531/cbc.ncep.0187).

Step 5: Open the project

After you log in, select your institution name to go to the institution home page determined by your instructor. Select the name of your project to go to the assessment page. [Screenshot 4]

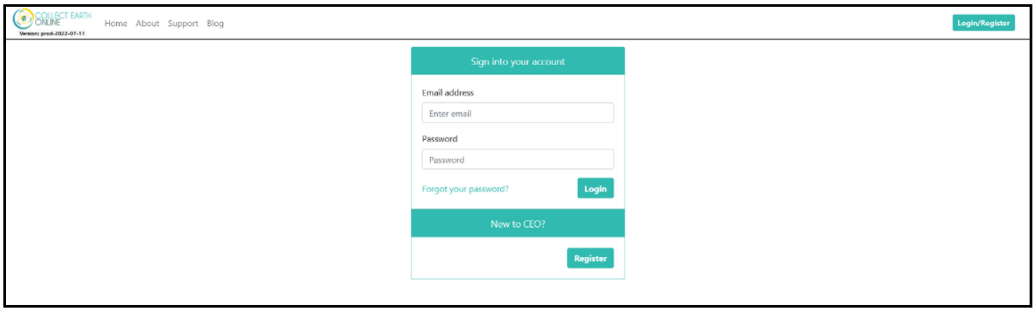
- **Note:** Your instructor will set up your institution's home, create the class project, add class members page prior to running this exercise (via the instructor exercise notes provided as a supplementary material with this module).
- **Note:** The screenshots provided throughout these steps are meant to help you along the way. Your screen will look different at times because you will be analyzing different plots than the ones on the screenshots.

Step 6: Select unanalyzed plots

Make sure the "Unanalyzed plots" Navigate option is selected. Select "Go to First Plot." Select "Planet NICFI Public" from the imagery option drop down list. Select "2021-04" as the imagery date. Ensure that the "Visible" option is selected instead of "Infrared." Imagery selections do not need to be updated for each plot. If you accidentally select another imagery source or date, the correct imagery

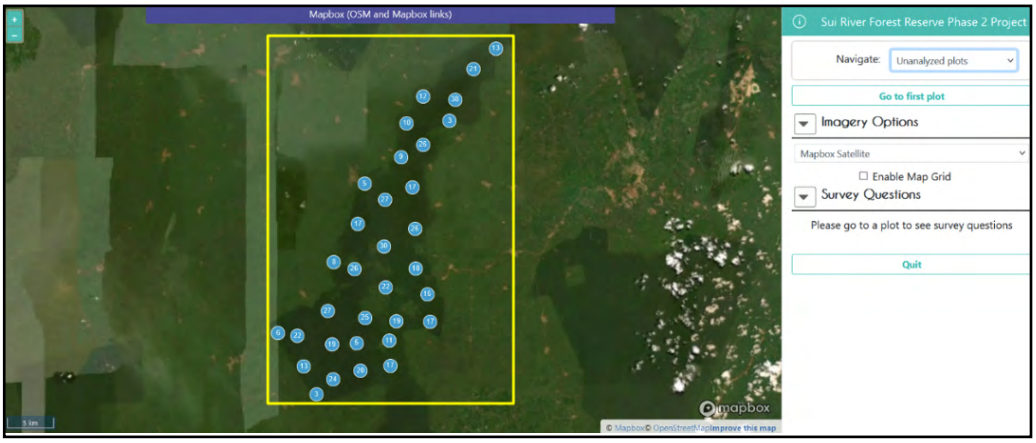
Screenshot 1
Part 1+2, Step 1:
Open Collect Earth Online (CEO).

Screenshot 2
Part 1+2, Step 2:
Register.



Screenshot 3
Part 1, Step 3:
Confirm registration.

Screenshot 4
Part 1, Step 5:
Open the project.



can be re-selected in the method above. [Screenshots 5 + 6]

- **Note:** The use of plots is a form of sample-based monitoring (Lister et al., 2014). This method allows for repeatable, statistically significant monitoring of an area's land cover. Each square plot covers 1 hectare (100 meters x 100 meters), which is a common plot size used in forest coverage studies (Schepaschenko et al., 2019). The plots in this study were originally created to provide statistically significant coverage of Ghana's Sui River Forest Reserve in a project by Carroll et al. (2023) that examined changes in forest cover as it relates to the habitat of the critically endangered giant West African Squeaker frog. The satellite imagery you will be viewing has a resolution of 5 meters and is provided monthly by a private satellite company, Planet, to the Collect Earth Online platform as part of a partnership between the two companies (O'Shea, 2020). All satellite imagery shown in the screenshots in this exercise is © 2017 Planet Labs PBC. A 5-meter resolution image means that each pixel in the image represents a 5-meter area of the ground. In higher resolution images, each pixel would represent a smaller area of the ground, whereas in lower resolution images, each pixel would represent a larger area. The imagery available here is provided in both visible and infrared light spectrums. You will only be analyzing visible light images, but researchers often use infrared light images to detect changes that might not otherwise be easily seen.
- **Note:** "Go to first plot" will take you to the first unanalyzed plot. To view a specific plot, analyzed or unanalyzed, select "Go to first plot." If the plot has been analyzed, select "Analyzed" from the Navigate dropdown menu, otherwise select "Unanalyzed." Then type the desired plot's number into the plot number field and select "Go to plot." Analyzed plots can only be selected by the student who completed the plot or by the teacher/administrator/instructor. If a student has difficulty locating a plot or received an error message when attempting to analyze the plot, the teacher may search for the plot using the "Admin Review" Mode listed below the "Navigate" drop-down menu in the visible in the teacher's account. If you are unable to locate the plot number box, scroll the right-hand box up.
- **Did you know?** When analyzing satellite imagery, it's crucial to use effective methods for interpreting the data. Plots are valuable tools for this purpose because they help clarify and enhance our understanding of the imagery, providing more accurate, detailed, and actionable insights. The use of an appropriate number and scale of randomly selected plots allows for statistically significant analysis of changes over space and time. The selection of plots in this exercise is not associated with data collected on ground. Instead, these are randomly selected to allow for repeatable analysis of forest changes, such land cover, vegetation health, or urban growth. The plots in this exercise are the ones used in Carroll et al. (2023). Even though this exercise does not involve quantitative analysis, plots are useful for measuring and comparing different variables. They offer a clearer perspective on complex data, because they allow us to sample an area and extract specific information from large datasets that aid in interpretation and decision-making. Plots help researchers extract and visualize changes that help them uncover patterns, relationships, and trends that could be overlooked by simply looking at the overall raw image. Using plots appropriately and effectively improves the outcomes of an analysis and supports the ability to draw meaningful conclusions.
- **Reflect**: Given the recent technological advancements, how do you think remote sensing tools impact our ability to monitor and address issues like deforestation?

Step 7: Assign land cover types

Survey Question 1, "Land Cover - April 2021 Planet," will autoload for each plot. There are two methods to complete a plot's land cover analysis. If the plot contains only one land cover type (for example, only forest is contained within the plot), you can click the appropriate land cover type listed under Question

1. All 9 subplot circles will change to the corresponding land cover color: green for "Forest", yellow for "Other Vegetated", and Red for "Developed/Barren." If more than one land cover type is present in a plot, select each subplot circle individually and then select the appropriate land cover type. If you decide to change your assessed land cover type, each subplot must be individually changed in this manner. If analyzing by subplots, choose the land cover type that is most dominant in the square area around each subplot. For example, if a plot appeared to contain a large building and a small vegetated area, assign "Developed/Barren" status to the circles over the building and "Other Vegetated" to the circle(s) over the vegetated area. Try zooming in and out of the satellite image using a mouse wheel or the "+/-" buttons on the top left of the screen.

Important definitions:

1. Closed Forest: refers to dense, contiguous areas of vegetation, typically characterized by a high canopy cover and minimal gaps between trees. This does not include areas containing only a few trees or areas where trees appear to have significant gaps between them. [Screenshot 7]
2. Other Vegetated: includes areas covered by vegetation other than closed forests, such as shrublands, grasslands, woodlands (open forests, with significant gaps between trees), or savannas.
3. Developed/Barren: encompasses urban or built-up areas, as well as barren or sparsely vegetated lands such as deserts, rock outcrops, or areas with little to no vegetation cover. Land cover studies usually list these two categories separately, as developed and built areas are quite different than completely deforested or barren areas. However, in this project, Carroll et al. (2023) were concerned with changes in forested areas vs. non-forested areas so a combined category of developed and barren land cover was used to represent areas obviously devoid of vegetation. [Screenshot 8]

- Tip: Zoom in and out of the image as needed to make your selection easier. Keep in mind that satellite images may not always be of the highest quality, which can make it challenging to accurately identify land use and land cover classes. Factors affecting image quality include atmospheric conditions that can distort or block satellite signals, such as clouds, fog, smoke, or haze. Additionally, limitations of the satellite itself, such as low resolution, calibration issues, or sensitivity problems, can impact image clarity. Technical malfunctions may also contribute to reduced image quality.
- Reflect: What strategies do you think researchers employ to address challenges related to image quality, such as atmospheric interference or low resolution? What about challenges related to how different researchers may classify a plot's land cover, for example, in a case where one researcher may classify an area as "Closed Forest" but another researcher classifies the same area as ""Other Vegetated"? How might these strategies affect the overall analysis? Do you think these challenges create issues with the reliability of the results? Explain (write down your thoughts/answers to these questions).

Step 8: Assign a confidence level

Survey Question 2, "Confidence Level", may be displayed by selecting "2" or pressing the right arrow under "Survey Questions." This question must be answered in order to properly save answers to Question 1. To simplify this activity, provide an overall confidence level for the plot, as we will not assess this in this exercise. You do not need to assess each subplot individually.

- Did you know? To address uncertainties in remotely sensed imagery, researchers often use confidence levels as a quality control measure, similar to those you see on your screen. Including confidence levels allows researchers to review and reassess their analyses, as well as cross-check their work with others. For instance, plots where researchers are only somewhat confident or not

Screenshot 5
Part 1, Step 6:
Select unanalyzed plots.

Screenshot 6
Part 1, Step 6:
Select unanalyzed plots.
(continued)

Screenshot 7
Part 1, Step 7:
Assign land cover types.

Example: Plot is covered in one land cover type, "Forest."

Screenshot 8
Part 1, Step 7:
Assign land cover types.

Example: Plot is covered in more than one land cover type, "Forest" and "Developed/ Barren."

confident will undergo a second round of interpretation. If discrepancies arise in the selected categories, additional analysis by other researchers or external reviewers may be sought to ensure the most accurate classification. Another method for addressing uncertainties is ground truthing, which involves going to the actual site to confirm the classification. You are basically connecting your analysis from the images to the real location on site. This process is ideal but can be challenging, especially in locations that are not easily accessible or that requires special permission/accesses. In addition, acquiring images with finer-resolution will help improve the quality which will make your analysis more precise. However, these types of images are often costly or do not exist for locations such as the one this activity is using in Ghana. While you will not be working with confidence levels in this exercise, understanding this process is valuable. It helps you appreciate how researchers handle low-resolution images or ambiguous classifications, ensuring more reliable and precise results. [Screenshot 9]

- **Reflect**: Do you believe that all the subplots in the provided image should be categorized as "Forest"? Why or why not? Would you classify any of the subplots differently, and if so, how? Please explain your reasoning. (Write down your thoughts/answers to these questions. Also, as you go through your assigned plots, think about those that seem difficult to assess and take notes on what made it difficult.)

Step 9: Save your progress

After both questions are answered for a plot, select "Save." Failure to do so will result in the plot's answers being deleted. Selecting "Save" will automatically move you on to the next plot. For this exercise, you don't need to focus on the confidence level. Just select "Confident" to save your work and proceed to the next plot.

Step 10: Analyze remaining plots

Continue steps 6–9 until students have completed their assigned plots. As a reminder, plot number is visible on the top right side of the page, above "External Tools."

- **Reflect**: Write down your thoughts/answers to these questions.
  - What patterns did you observe in the classifications?
  - How could the different land cover categories ("Forest", "Other Vegetated", "Barren") affect the local environment and community?
  - Are there any noticeable differences or issues related to land cover? Why do you think one area is different than another? Are there any factors that may influence this? (This could include urbanization, agricultural practices, or natural events, such as forest fires or floods, which could help explain how the current land cover developed.)
  - How could the changes from forest to other vegetated areas or barren affect the local species including the Giant Squeaker Frog?

PART 2: DOWNLOADING THE DATA – COLLECT EARTH ONLINE

Step 1: Open Collect Earth Online

Go to Collect Earth Online's (CEO) website here: (https://collect.earth/). [Screenshot 1]

Step 2: Log in [Screenshot 2]

Step 3: Select "Visit" for your project [Screenshot 10]

Step 4: Download your project data

Select "S" in your project's row to download the sample data for your project. This will download a CSV file that can be opened in Microsoft Excel (or another spreadsheet program) containing all your answers to the land cover classification you performed earlier. The completed project will be highlighted in green. Red colored projects are unpublished and yellow indicates a project in progress. The file will be downloaded to your browser's assigned download folder; typically, it can be accessed by clicking the "Downloads" button on the upper right corner of your browser. Open the folder to which the file downloaded. [Screenshot 11]

- Note: the below instructions are specific to Excel, but you may use other spreadsheet software programs and modify the instructions as needed.

Step 5: Open the data

Open the CSV file in Excel and observe the names of the different columns. Select "Do not convert" if Excel asks to convert large number to scientific notations. Note that "lon" and "lat" are the longitude and latitude columns, respectively. Your land classification answers are listed in the column named after the land cover classification question you created earlier. In this case it is "Land Cover – April 2021 Planet." [Screenshot 12]

Step 6: Close the file

Close the Excel file, but leave the folder that stores it open, so you can drag and drop it in Part 3A, Step 5.

PART 3A: MAPPING THE DATA – ARCGIS ONLINE (ALTERNATIVE PART 3B USING QGIS)

Using Map Viewer, an open-source map making tool, you will know map your analyzed data.

Step 1: Open ArcGIS Online

Go to the ArcGIS Online (AGOL) website: https://www.arcgis.com/index.html. [Screenshot 13]

Step 2: Log in

Sign in or create an account if you are a new user. [Screenshot 14]

Step 3: Load data

Select "Content" on the menu bar at the top of the screen. [Screenshot 15]

Step 4: Load data (continued)

Select "New item" on your content homepage. [Screenshot 16]

Step 5: Load data (continued)

Drag and drop the downloaded Excel file into the upload box. You can also search for the Excel file on your computer using the "Your Device" option. [Screenshot 17]

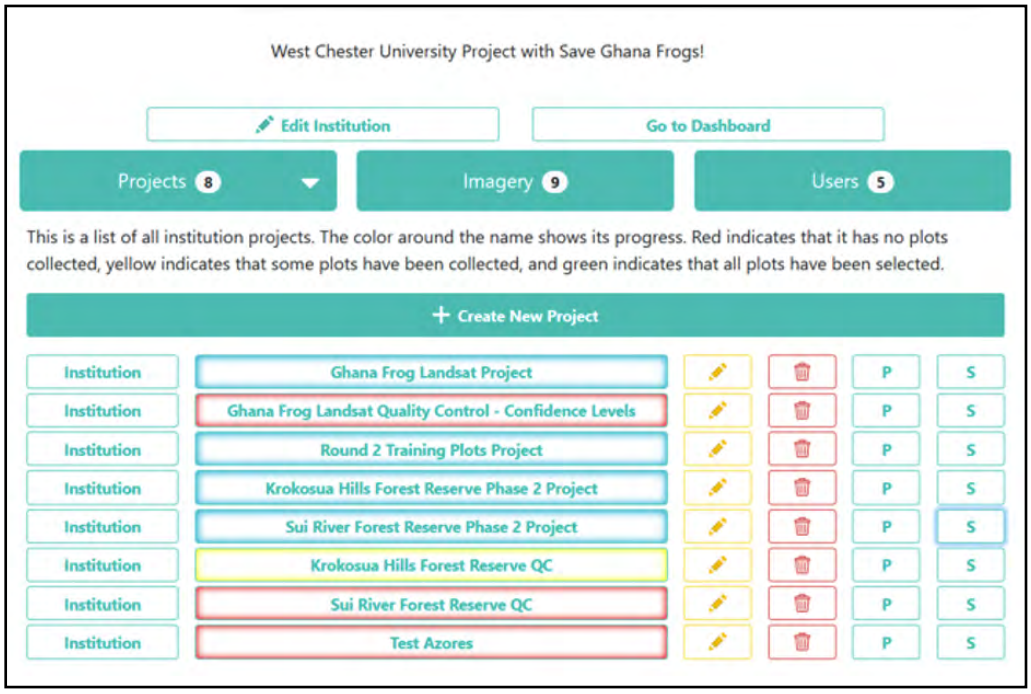
Step 6: Load data (continued)

Select "Add [your file name] and create a hosted feature layer or table" and then click "Next." [Screenshot 18]

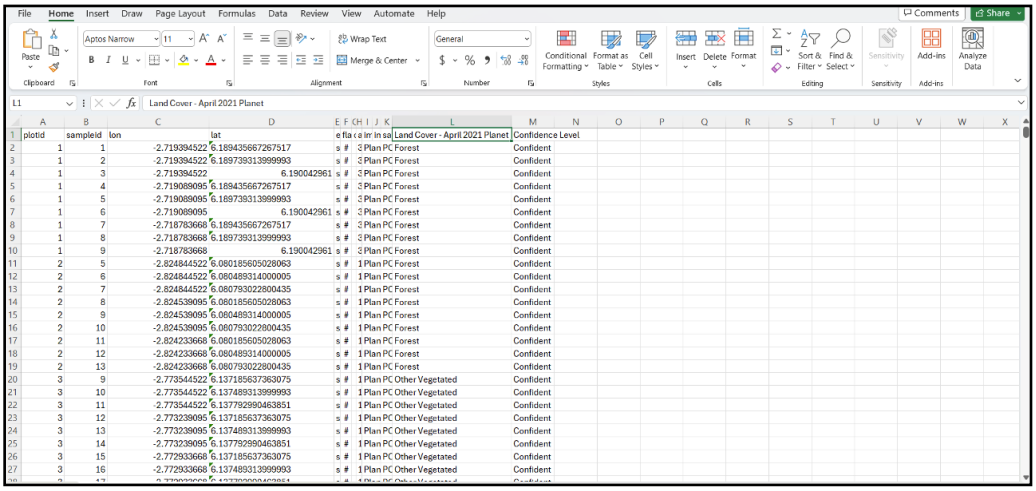
Screenshot 9
Part 1, Step 8:
Assign a confidence level.

Screenshot 10
Part 2, Step 3:
Select "Visit" for your project.

Screenshot 11
Part 2, Step 4:
Download your project data.

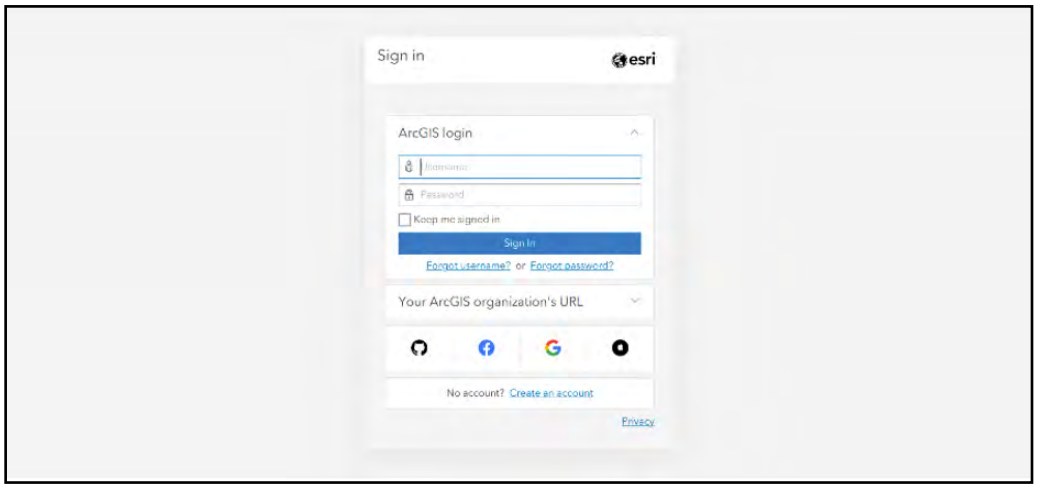


Screenshot 12
Part 2, Step 5:
Open the data.



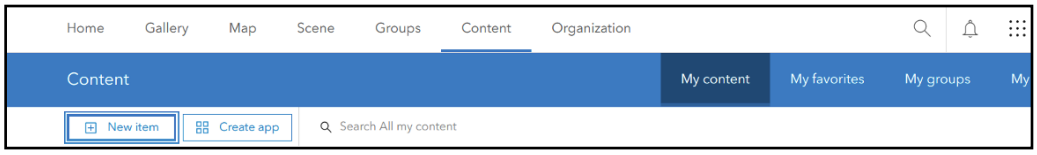
Screenshot 13
Part 3A, Step 1:
Open ArcGIS Online.

Screenshot 14
Part 3A, Step 2:
Log in.



Screenshot 15
Part 3A, Step 3:
Load data.

Screenshot 16
Part 3A, Step 4:
Load data. (continued)



Step 7: Define data columns

AGOL will briefly display a message stating that it is analyzing your file. It will then display a screen with the names of the Excel sheet's data columns and each column's variable type. You don't need to change anything as long as the "lon" and "lat" "Field names" are listed as "Type" "Double" and the land classification "Field name" is listed as "Type" "String." A "Double" variable type allows AGOL to read the data in the "lon" and "lat" fields as real numbers, which is necessary for mapping plot coordinates. A "String" data types allows AGOL to read the land classification data as a string of text. Click "Next." [Screenshot 19]

Step 8: Define spatial coordinate data

This next screen identifies the spatial coordinate data in the Excel file that will allow your plots to be mapped. The "Latitude" location type should be auto-filled with "lat" and the "Longitude" location type should be auto-filled with "lon." If not, select these fields from the drop-down menu for their appropriate location type. Click "Next." [Screenshot 20]

- **Did you know?** Latitude and Longitude coordinates precisely identify locations on Earth. To ensure that your mapping software correctly places each point, it's important to match the "Latitude" data with "lat" and the "Longitude" data with "lon." If your Excel file doesn't automatically label these fields, you'll need to manually select "lat" and "lon" from the drop-down menu or add this prior to uploading your data. This step is crucial for accurate data representation and effective visualization of spatial patterns and relationships. Fortunately, the data downloaded from CEO is already properly labeled, so you won't need to make these adjustments. However, it's good practice to remember this process for any future work with mapping software that requires this type of coordinate data.
- **Reflect**: What challenges do you think you will face if your data does not have latitude and longitude data? How can you address this to make sure your result is accurate? (Write down your thoughts/answers to these questions.)

Step 9: Save upload

This is the final step in the file upload process. No changes are necessary, but if desired you can change the name of the data file, add a summary of what the file contains, or add keyword tags. Select "Save." [Screenshot 21]

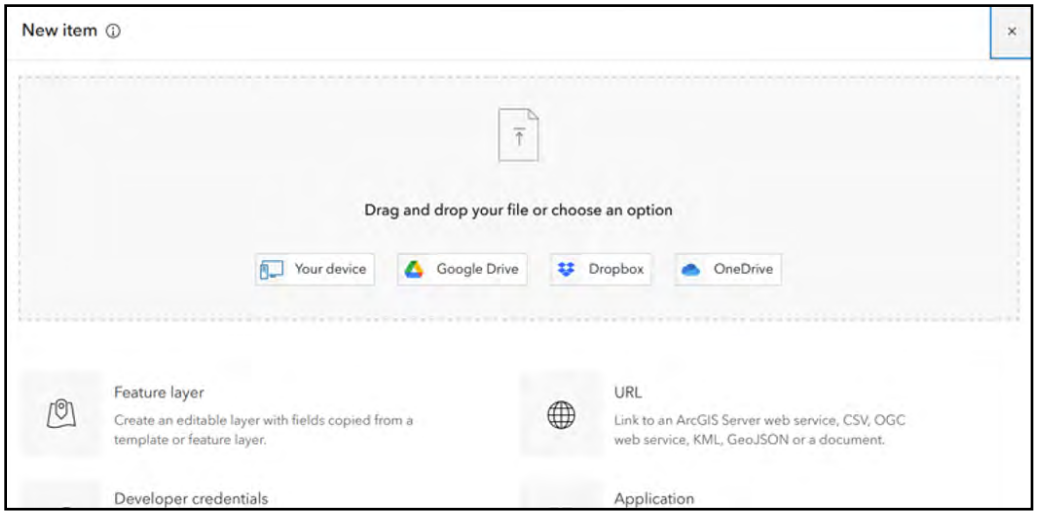
Step 10: Open in Map Viewer

AGOL will display messages regarding the upload and publishing of your data and then display your file's information as seen below. You can add a file description or view its metadata if you want, but for now select "Open in Map Viewer." [Screenshot 22]

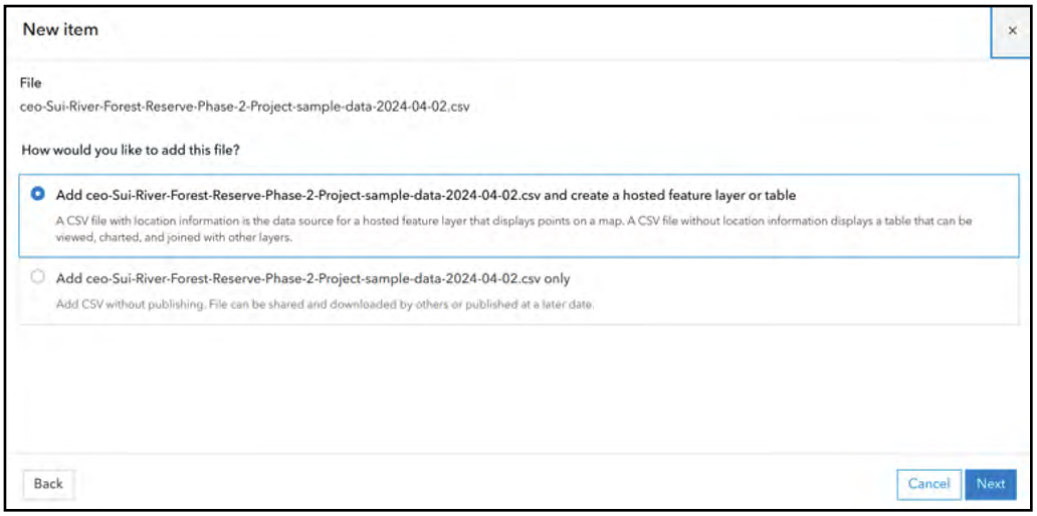
Step 11: Navigate map

Your plots should display, creating a rough outline of the area you selected to study. If you zoom in, using either the mouse wheel or by clicking the "+" button, you should see that each plot is composed of the number of subplots you created earlier. In this example, there are 9 subplots for each plot. We wish to see these subplots color coded by land classification type, but they are all the same color. To fix this, select "Edit layer style" in the "Symbology" box on the right-hand side of the screen. [Screenshots 23 + 24]

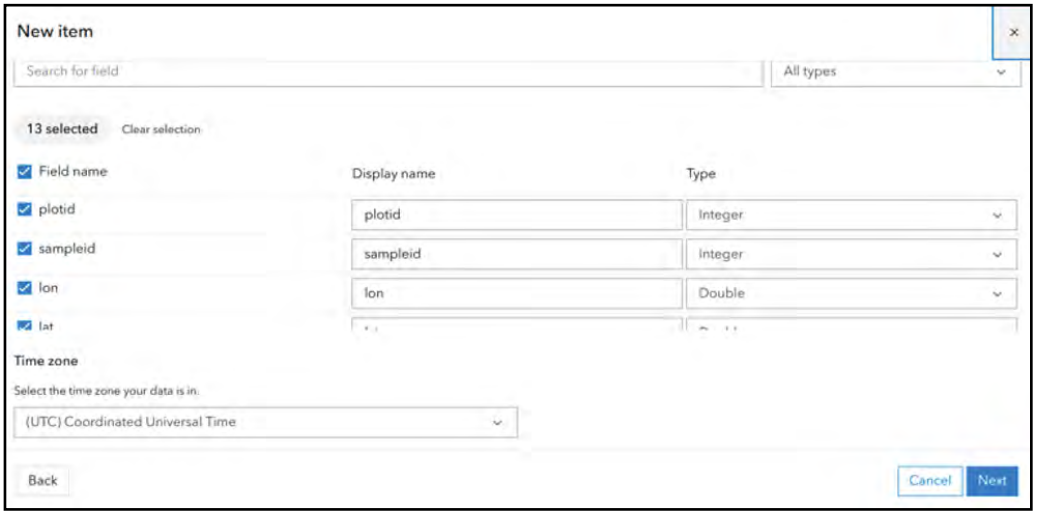
Screenshot 17
Part 3A, Step 5:
Load data. (continued)



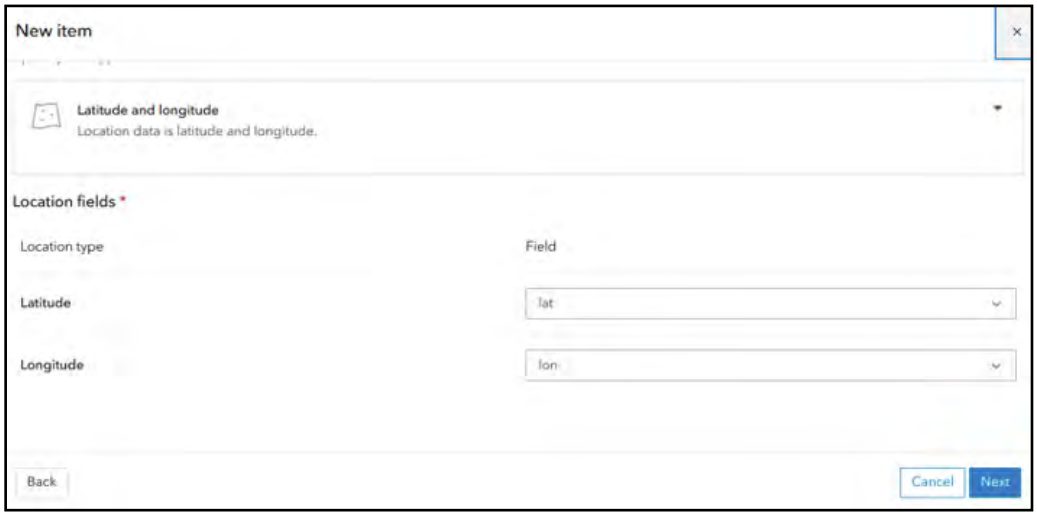
Screenshot 18
Part 3A, Step 6:
Load data. (continued)



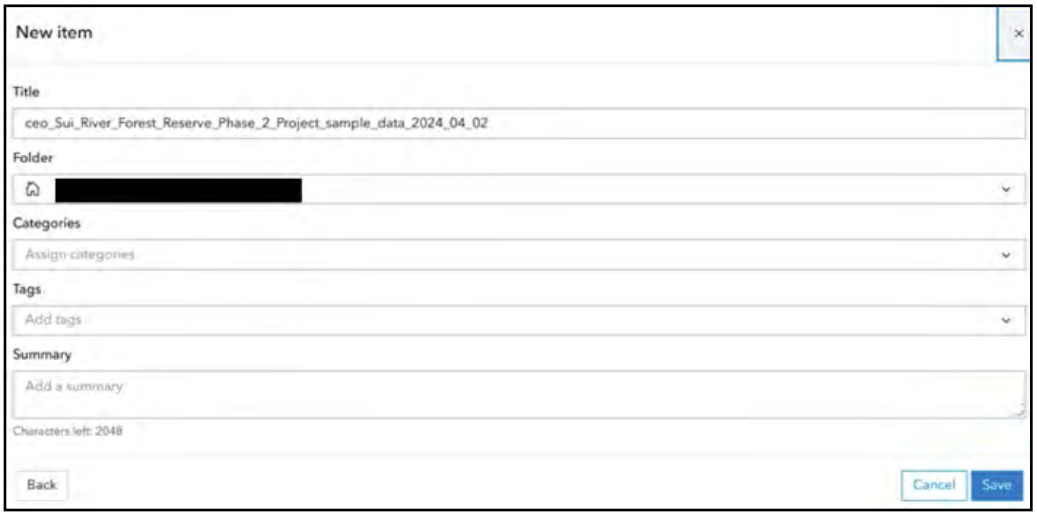
Screenshot 19
Part 3A, Step 7:
Define data columns



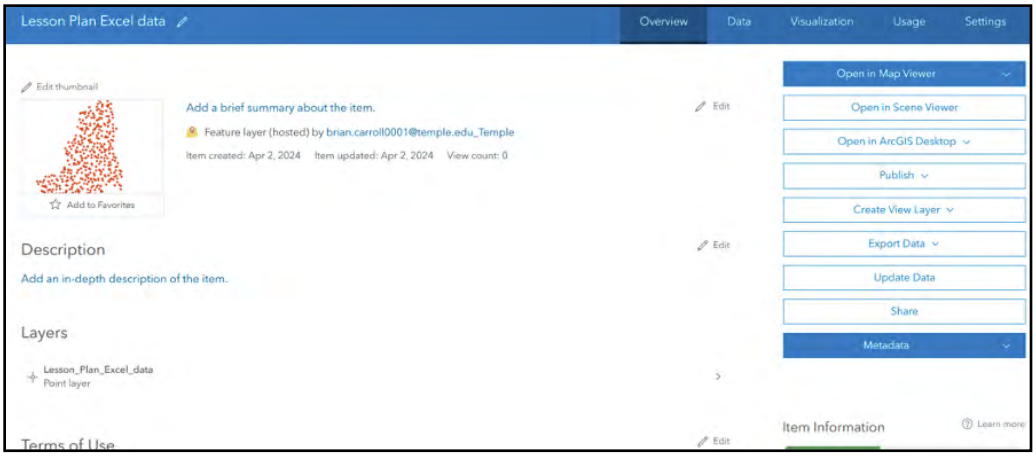
Screenshot 20
Part 3A, Step 8:
Define spatial coordinate data



Screenshot 21
Part 3A, Step 9:
Save upload



Screenshot 22
Part 3A, Step 10:
Open in Map Viewer



Step 12: Choose attributes

Select "+ Field" and then click the check box next to the name of the Excel column that contains your land classification answer. In this example, it is named "Land Cover – April 2021 Planet." Click "Add." [Screenshots 25 + 26]

Step 13: Pick a style

Select "Style Option" under the "Types (unique symbols)" header. [Screenshot 27]

- **Consider**: Now that your plots are classified and mapped, zoom in and out to examine the areas you categorized as "Other Vegetated" and "Barren." Take note of additional features visible on the default base map, such as roads, rivers, and potential human settlements. If you wish to see your results displayed on a different base map, you can select from several options by clicking on the "Basemap" button on the left-hand side of your screen. Reflect on how these features might contribute to the loss of forested areas in those plots. Did you notice any patterns or trends in the locations of the plots categorized as "Other Vegetated" and/or "Barren"? Take notes of your observations for future discussion.

Step 14: Assign colors

AGOL will pre-assign colors to your land classification types. If you like the colors assigned, click "Done." If you want to change the colors, for example to assign green to "Forest", click on the color next to the land classification type, then click the "Fill color" box. [Screenshots 28 + 29]

- **Did you know?** Colors play a crucial role in mapmaking beyond just improving visual appeal. They significantly influence how viewers interpret the information presented. Customizing the colors in your map enhances its effectiveness, ensuring that the information is easily understood with minimal explanation. By assigning specific colors to different data categories or types, you increase the map's clarity and make it easier to recognize and interpret patterns. This approach ensures that your map communicates data clearly and accurately. Be sure to take into account accessibility considerations, for example color blindness.
- **Reflect**: What colors would you choose for each category? Please explain your reasoning. (Write down your answers to these questions.)

Step 15: Save color assignments

Select the color you want from the sliding color bar and then select "Done." Repeat Step 14 until you have assigned your desired colors to each land cover type. Select "Done" on the "Style Options" menu then select "Done" on the "Styles" menu. [Screenshot 30]

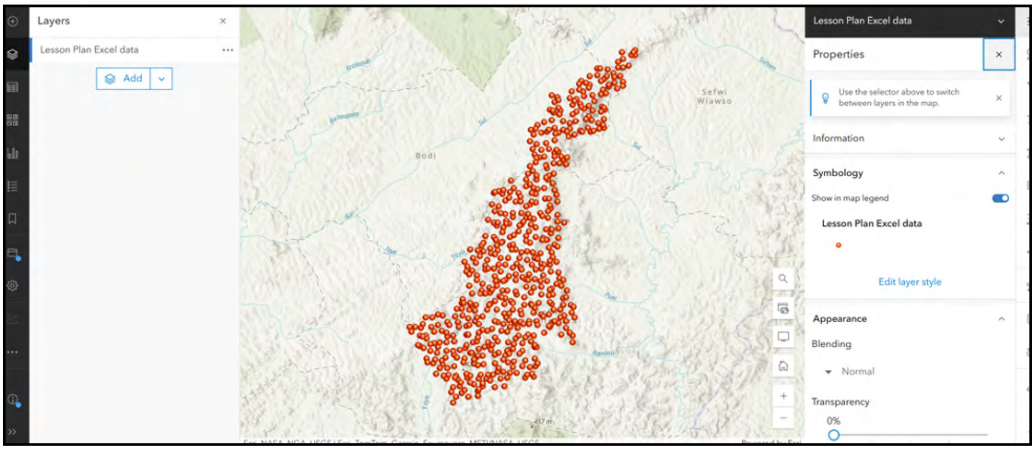
Step 16: Display map symbology

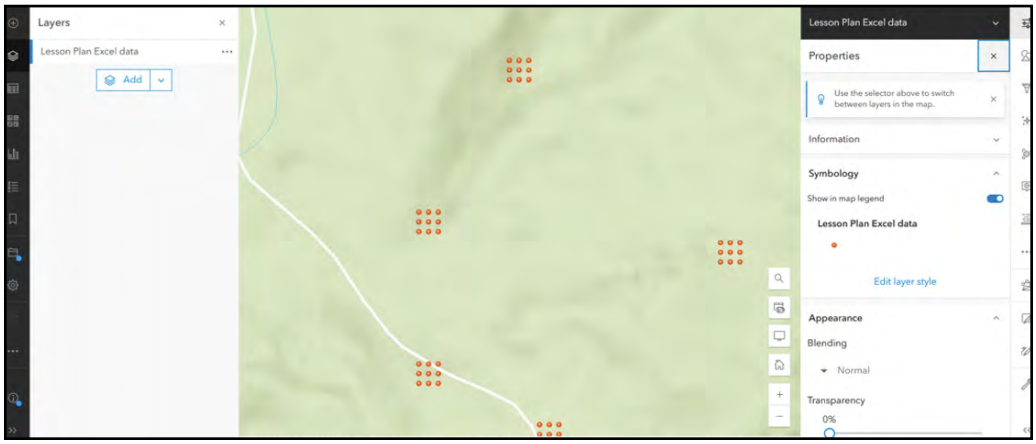
Your plots are now mapped, and color coded to reflect your land cover classification answers. If you forget what each color symbolizes, select the "Properties" button (three horizontal bars icon that will display "Properties" when hovered over) on the top of the right-hand options column to display your map's symbology. [Screenshot 31]

Step 17: Save map

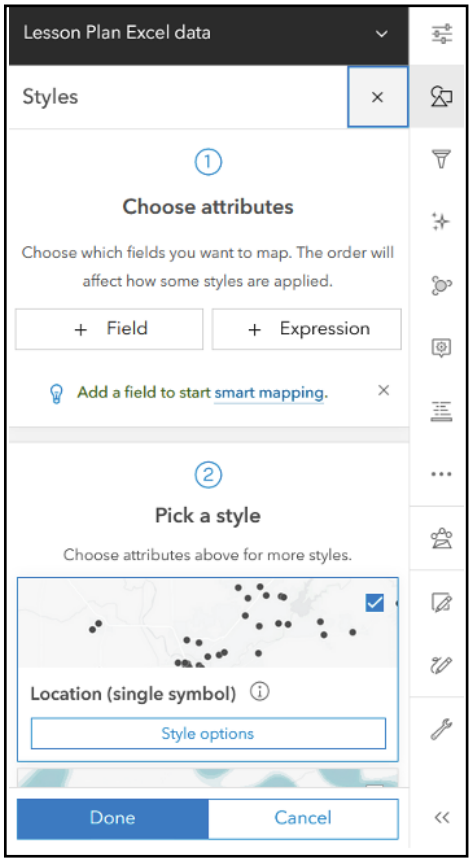
Select "Save and Open" on the left-hand options column and then select "Save As" (folder icon). [Screenshot 32]

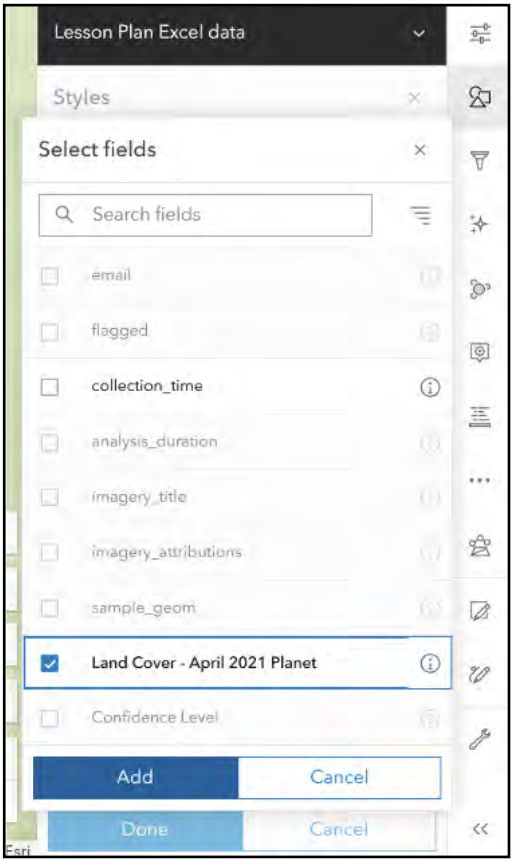
Screenshots 23 + 24
Part 3A, Step 11:
Navigate map



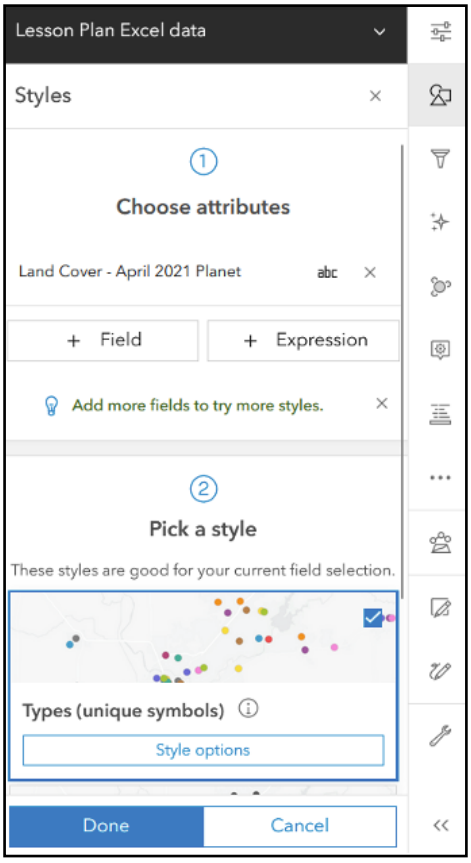


Screenshots 25 + 26
Part 3A, Step 12:
Choose attributes

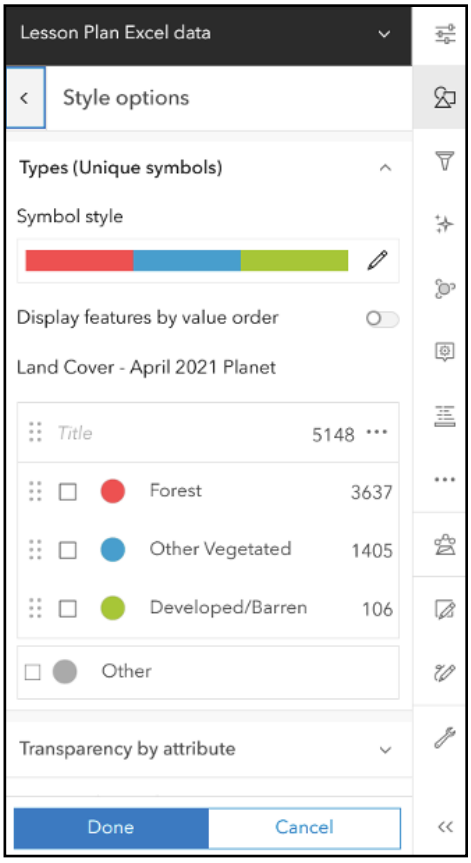




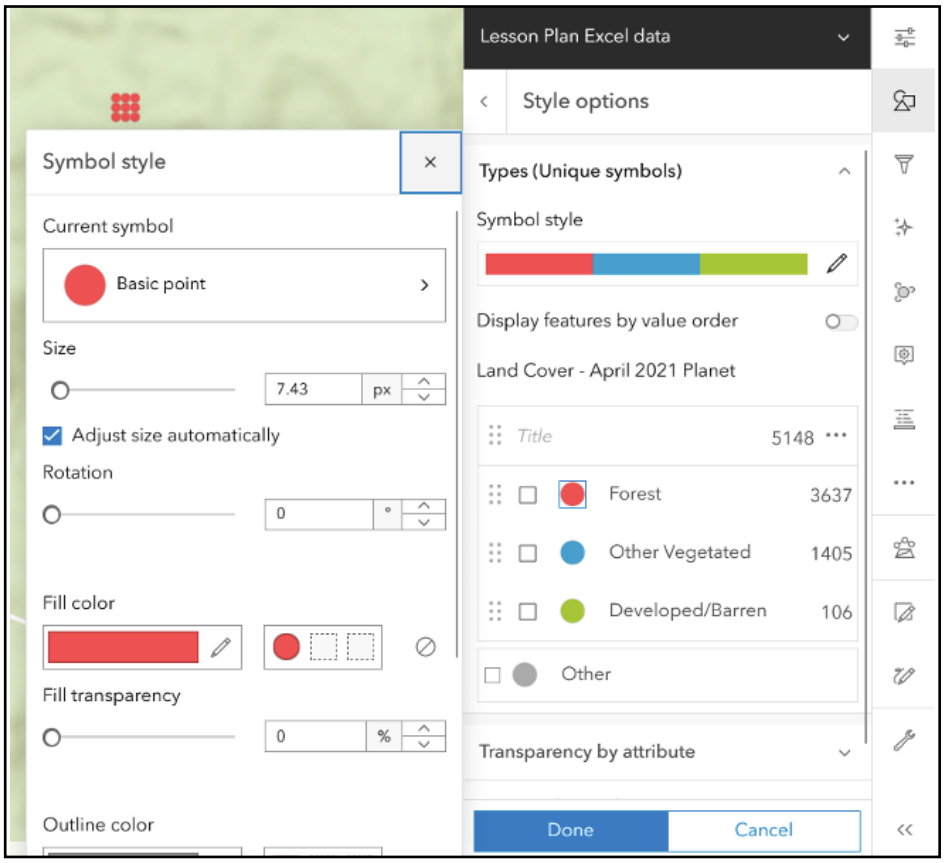
Screenshot 27
Part 3A, Step 13:
Pick a style



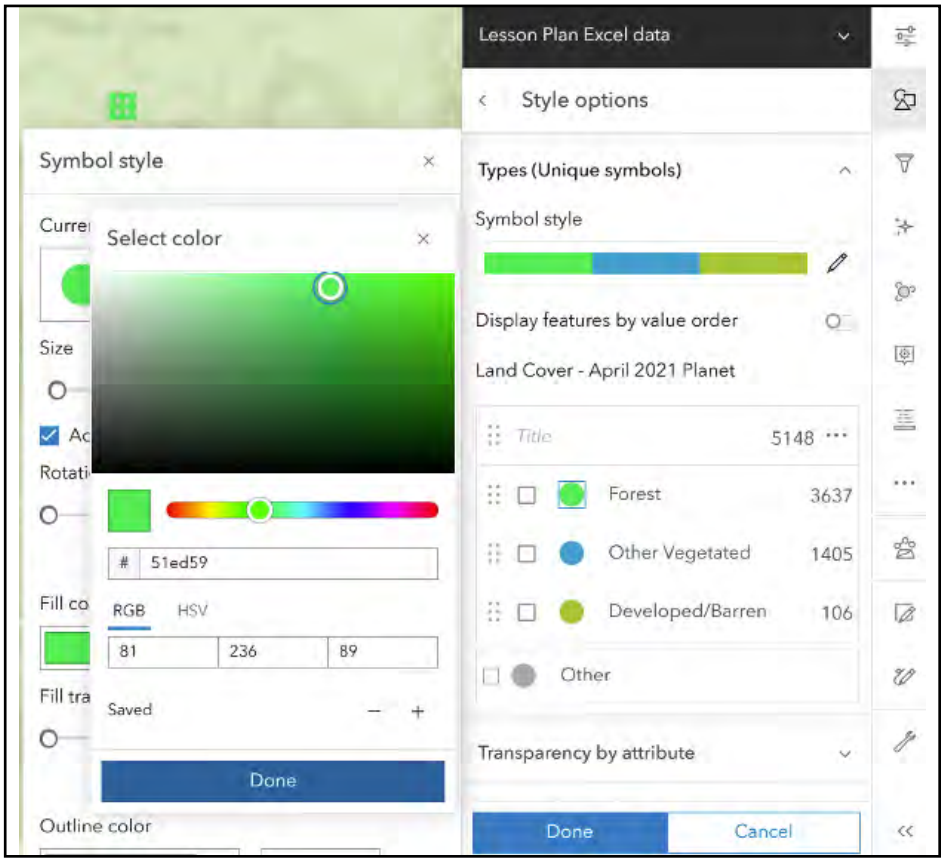
Screenshot 28
Part 3A, Step 14:
Assign colors



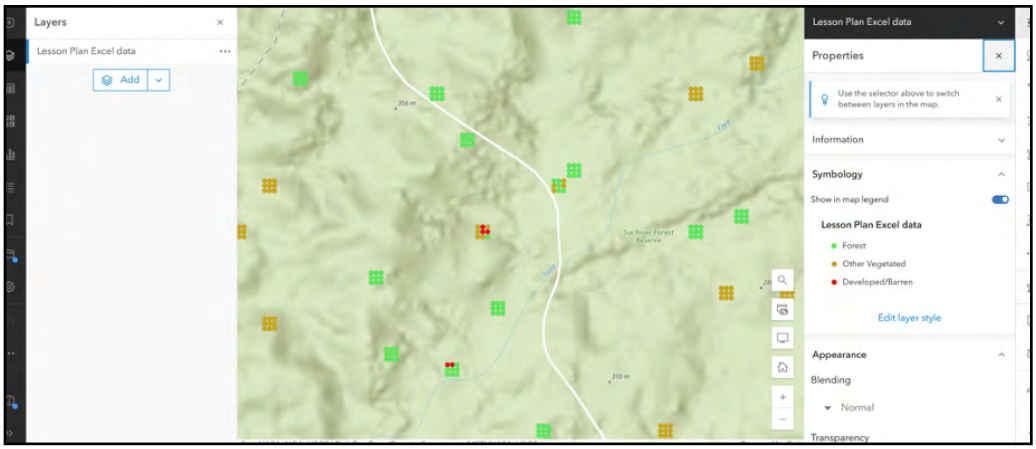
Screenshot 29
Part 3A, Step 14:
Assign colors. (continued)



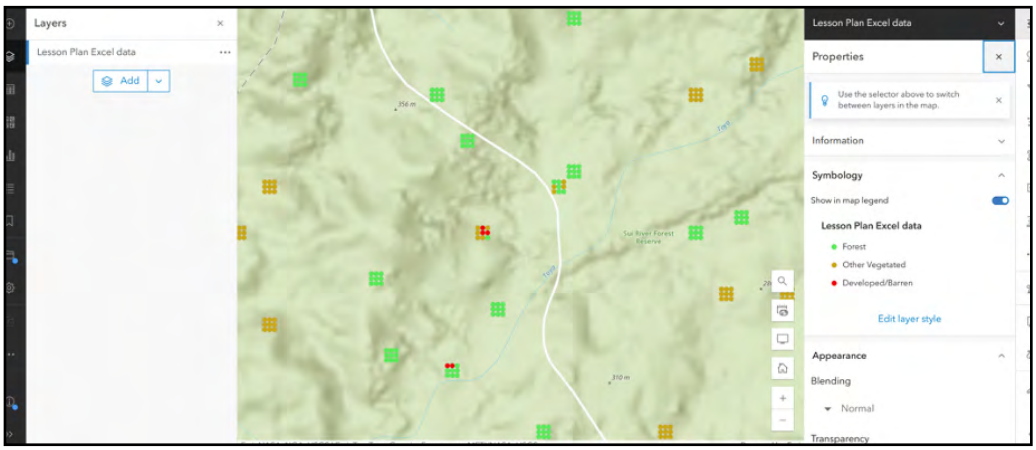
Screenshot 30
Part 3A, Step 15:
Save color assignments



Screenshot 31
Part 3A, Step 16:
Display map symbology



Screenshot 32
Part 3A, Step 17:
Save map



Step 18: Save map (continued)

Give your map a title and then click "Save." As AGOL is an online mapping tool, your map's name needs to be unique. If the name you choose is already in use within AGOL's database, you may receive an error message. In that case, please return and select a different name for your map. [Screenshot 33]

Step 19: Reopen maps

Your map has now been saved to your AGOL Content menu. You can close out of AGOL or continue exploring your map. If you need to access your map again, click on the map's name on your "Content" home page and then select "Open in Map Viewer." [Screenshot 34]

PART 3B: MAPPING THE DATA – QGIS (ALTERNATIVE TO PART 3A)

Using QGIS (v. 3.40), an open-source map making tool, you will now map your analyzed data.

Step 1: Open QGIS

If you have not yet already downloaded the application, go to https://www.qgis.org/download/ to download the software.

- Note: the below instructions were generated with QGIS (v. 3.40), if newer versions of QGIS do not match these instructions, consider downloading v. 3.40.

Step 2: Create new project

Click "New Empty Project" to the right. [Screenshot 35]

Step 3: Load data

"Layer" > "Add Layer" > "Add Delimited Text Layer" [Screenshot 36]

Step 4: Load data (continued)

In the browser bar "File Name" direct yourself towards the downloaded Excel file.

Step 5: Define data columns

The fields will automatically fill based on the fields in the Excel file. However, direct yourself to the section, "Geometry Definition." [Screenshot 37]

Step 6: Define Coordinate Reference System (CRS)

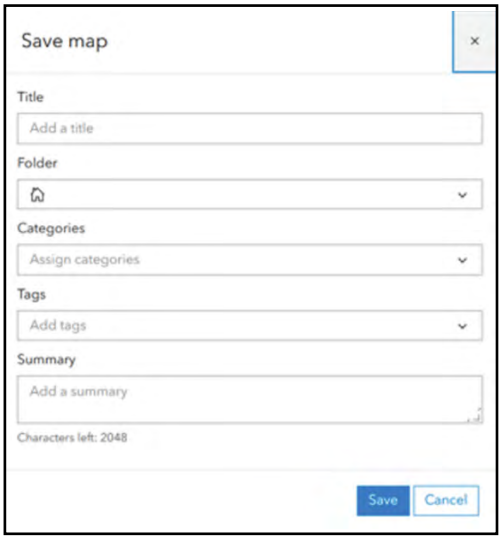
Where it says, "Geometry CRS" select the "Project CRS." This step ensures that the map projects the data in the CRS that matches the data source.

- Note: If you would like to learn more about coordinate reference systems, consider reviewing the QGIS documentation provided online (https://docs.qgis.org/3.40/en/docs/gentle_gis_introduction/coordinate_reference_systems.html).

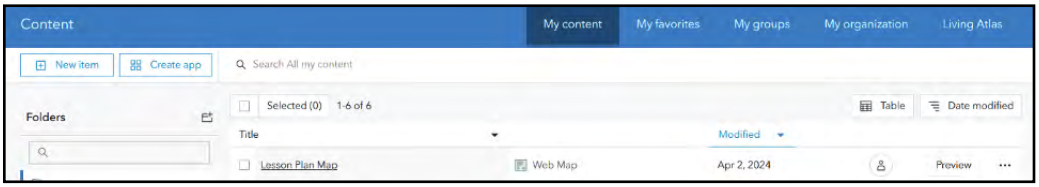
Step 7: Display data

Click "Add" at the bottom and then close the Delimited Layer Window. Your plots should be displayed in the map frame. [Screenshot 38]

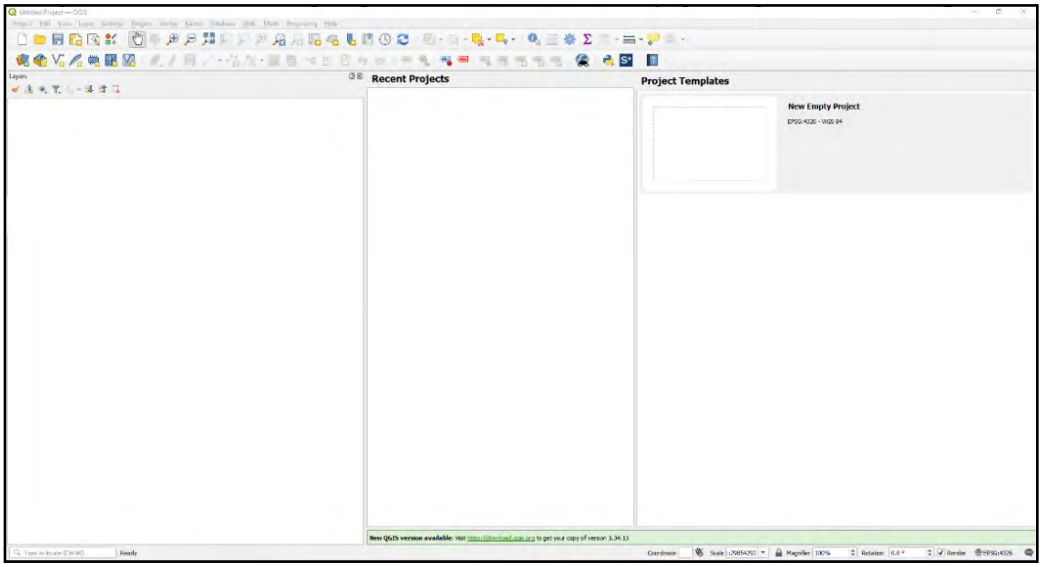
Screenshot 33
Part 3A, Step 18:
Save map (continued)



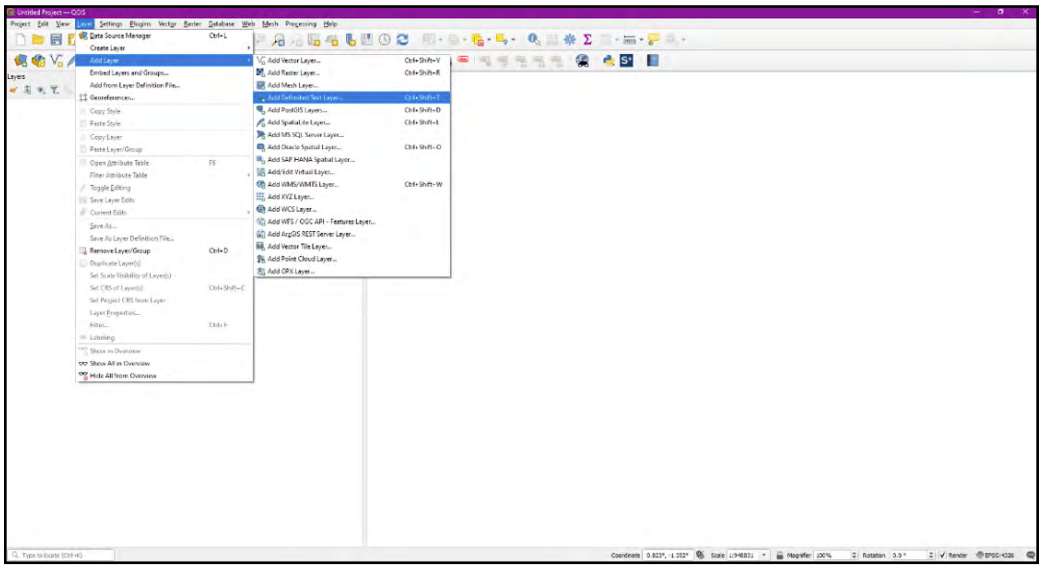
Screenshot 34
Part 3A, Step 19:
Reopen maps



Screenshot 35
Part 3B, Step 2:
Create new project



Screenshot 36
Part 3B, Step 3:
Load data



We wish to see these subplots color coded by land classification type, but they are all the same color. To fix this, we need to symbolize the layer.

Step 8: Symbolize data

Right click the layer in the layers table and select properties. [Screenshot 39]

Step 9: Symbolize data (continued)

Select the Symbology tab to the left. [Screenshot 40]

Step 10: Symbolize data – categorization

At the top of the window where it says "Single Symbol" replace that with Categorized. [Screenshot 41]

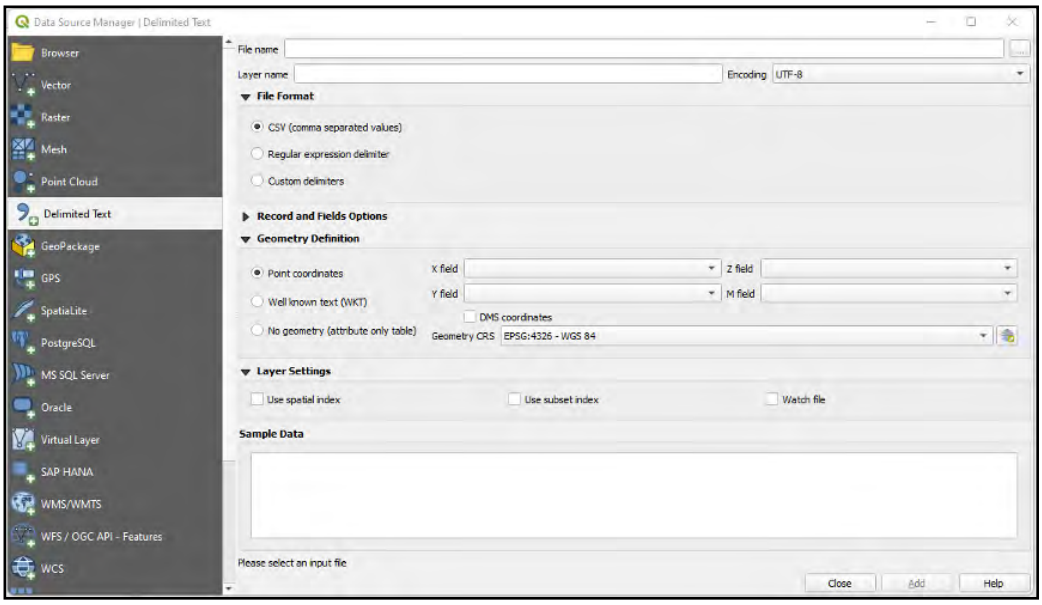
Step 11: Apply classifications

Select "Value" and choose the field, "Land Cover – April 2021 Planet." Select "Classify" and then "Apply." If you wish to change the colors of the different values, click on the dot symbol next to the legend name and select a new color.

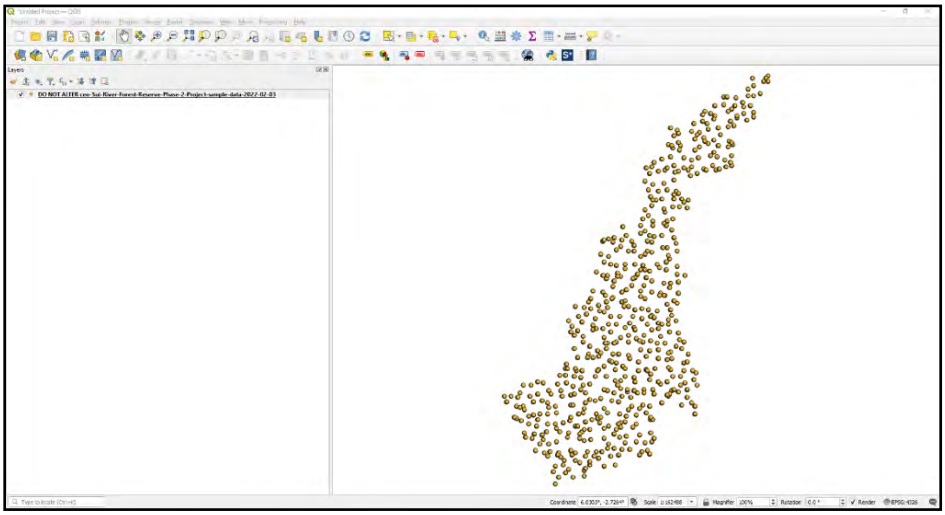
Step 12: Save map

You will now see the layer reflect the symbology change in your map viewer. Don't forget to save the project by going to "Project" and then clicking "Save As." [Screenshot 42]

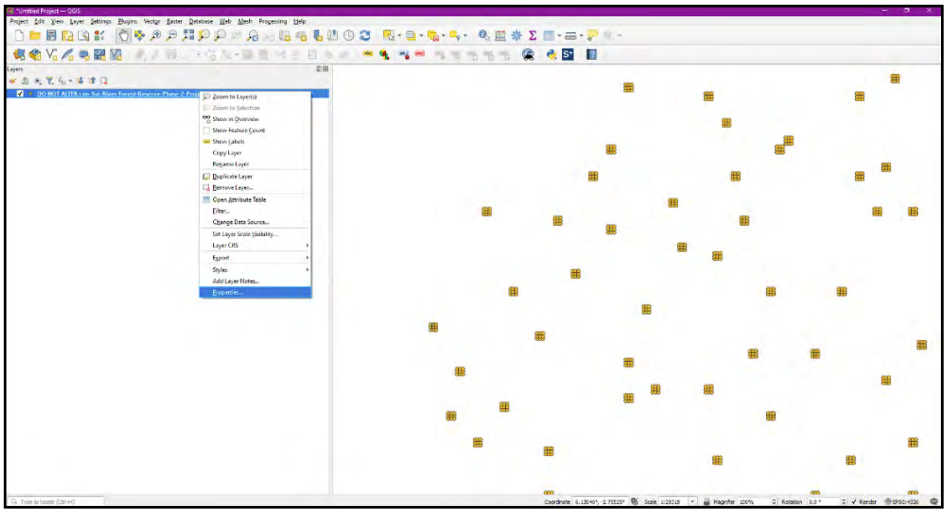
Screenshot 37
Part 3B, Step 5:
Define data columns



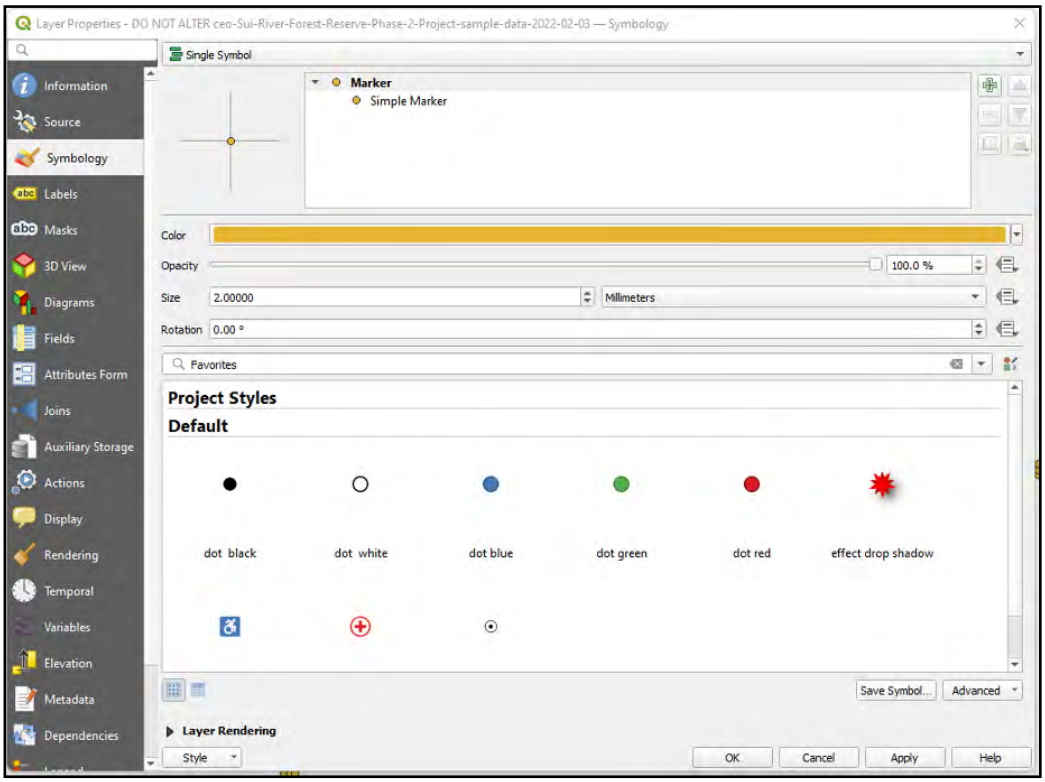
Screenshot 38
Part 3B, Step 7:
Display data



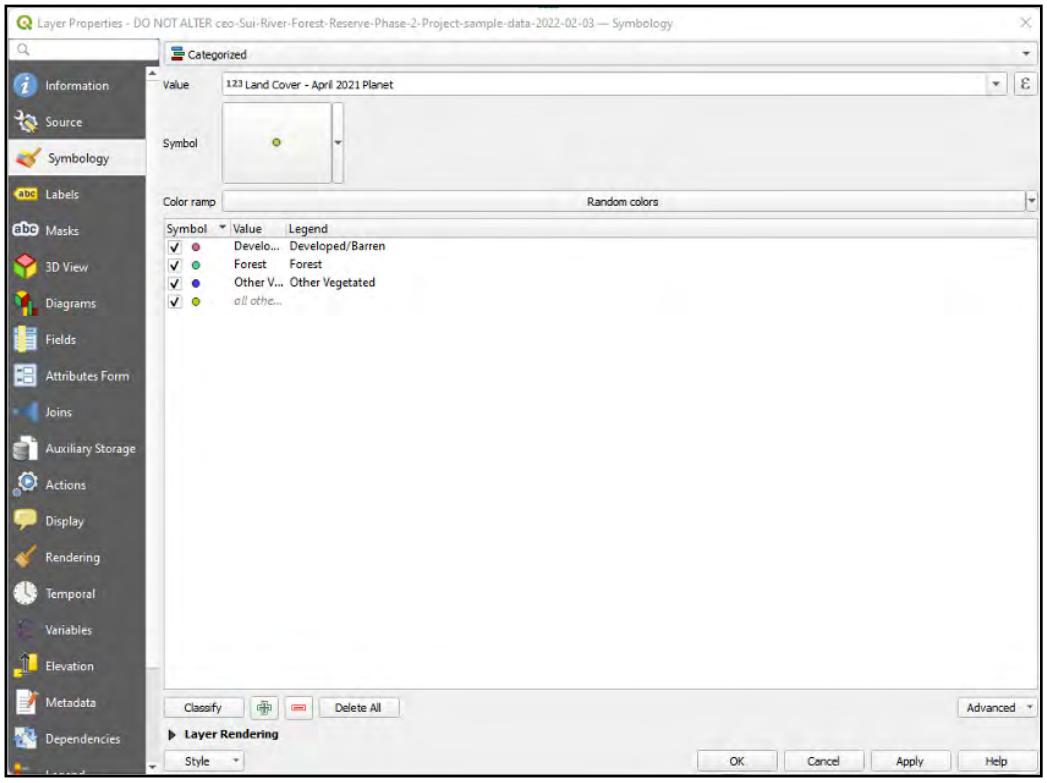
Screenshot 39
Part 3B, Step 8:
Symbolize data



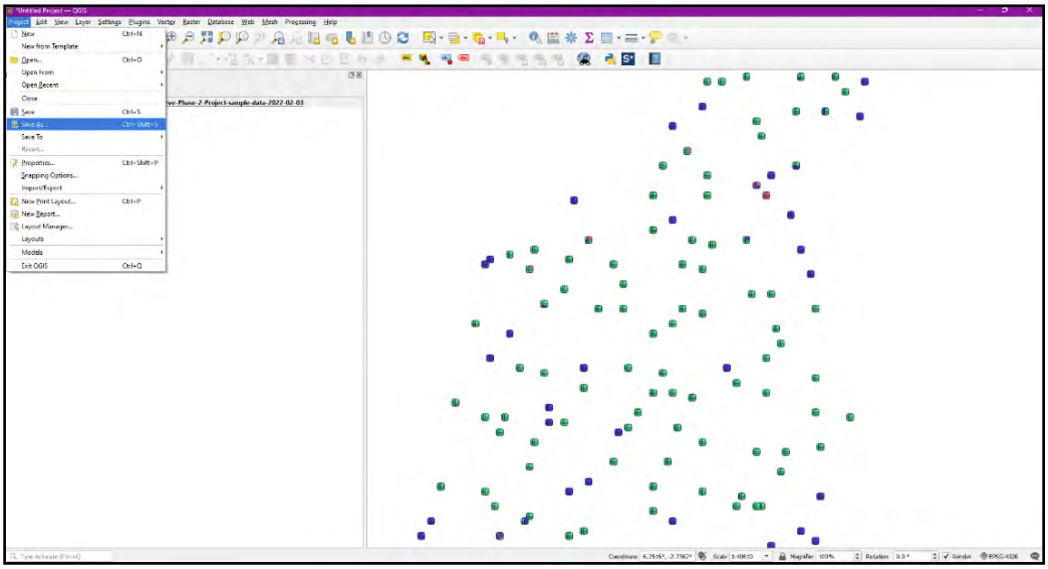
Screenshot 40
Part 3B, Step 9:
Symbolize data (continued)



Screenshot 41
Part 3B, Step 10:
Symbolize data – categorization



Screenshot 42
Part 3B, Step 12:
Save map



PART 4: GROUP DISCUSSION

Congratulations! You've completed your first project as a new employee of Save the Frogs! Now it is time to confer with your colleagues and then create a presentation for the larger Save the Frogs research team.

- Gather all your answers and notes in preparation for your group's discussions.
- Together you will review all the maps created with your analyzed plots. Pay close attention to those plots classified as "Other Vegetated" and "Barren", specifically emphasize the "Barren" areas. These are the areas that will affect the Giant Squeaker Frog the most since it clears out their entire habitat.
- As group, come up with key findings of all your analysis that address the following questions:

- What patterns did you observe in the classifications?
- How could different land cover categories ("Forest", "Other Vegetated", "Barren") affect the local environment and community?
- Are there any noticeable differences or issues related to land cover? Why do you think one area is different than another? Are there any factors that may influence this? (This could include urbanization, agricultural practices, or natural events, such as forest fires or floods, which could help explain how the current land cover developed).
- How could the changes from forest to other vegetated areas or barren affect the local species including the Giant Squeaker Frog?

PART 5: PUTTING TOGETHER A PRESENTATION

- You will have 5–7 minutes to present your project and the implications (unless otherwise specified by your instructor).
- Your presentation should include the following:
  - The final maps.
  - Overall findings highlight the different categories of plots they analyzed.
  - Common themes, trends, or patterns observed. (Highlight any surprises).
  - Discuss potential impacts of habitat loss for the species in the area.
  - Address how having different land cover within a reserve could affect land management strategies.
  - Discuss the benefits of the land cover and what is driving the changes in this reserve (consider the potential ethical issues surrounding the use of technologies such as remote sensing, in monitoring and managing these changes).
  - Offer recommendations to policymakers and decision-makers to implement positive changes that will prevent species like the Giant Squeaker Frog from becoming extinct due to habitat loss.
- The other groups and the instructor will have an opportunity to provide comments and share similar findings.

ACKNOWLEDGMENTS

The screenshots in this exercise contain satellite imagery from Planet Labs © 2017 Planet Labs PBC.

REFERENCES

Asner, G. P., Knapp, D. E., Broadbent, E. N., Oliveira, P. J. C., Keller, M., & Silva, J. N. (2005). Selective logging in the Brazilian Amazon. Science, 310(5747), 480–482. https://doi.org/10.1126/science.1118051

Barlow, J., França, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., Castello, L., Economo, E. P., Ferreira, J., Guénard, B., Gontijo Leal, C., Isaac, V., Lees, A. C., Parr, C. L., Wilson, S. K., Young, P. J., & Graham, N. A. J. (2018). The future of hyperdiverse tropical ecosystems. Nature, 559(7715), 517–526. https://doi.org/10.1038/s41586-018-0301-1

Blaustein, A. R., Han, B. A., Relyea, R. A., Johnson, P. T. J., Buck, J. C., Gervasi, S. S., & Kats, L. B. (2011). The complexity of amphibian population declines: Understanding the role of cofactors in driving amphibian losses. Annals of the New York Academy of Sciences, 1223(1), 108–119. https://doi.org/10.1111/j.1749-6632.2010.05909.x

Brobbey, L. K., Agyei, F. K., & Osei-Tutu, P. (2020). Drivers of cocoa encroachment into protected forests: The case of three forest reserves in Ghana. International Forestry Review, 22(4), 425–437. https://doi.org/10.1505/146554820831255533

Carroll, B., Neal, S., & Morales-Ramirez, C. A. (2023). Establishing a baseline of forest cover in two Ghanaian forest reserves using an open-source remote sensing tool to aid endangered species conservation efforts. The Pennsylvania Geographer, 61(2), 13–31.

Curtis, P. G., Slay, C. M., Harris, N. L., Tyukavina, A., & Hansen, M. C. (2018). Classifying drivers of global forest loss. Science, 361(6407), 1108–1111. https://doi.org/10.1126/science.aau3445

de Oliveira Roque, F., Menezes, J. F. S., Northfield, T., Ochoa-Quintero, J. M., Campbell, M. J., & Laurance, W. F. (2018). Warning signals of biodiversity collapse across gradients of tropical forest loss. Scientific Reports, 8(1), 1622.

https://doi.org/10.1038/s41598-018-19985-9
Ellis, E. C., Gauthier, N., Klein Goldewijk, K., Bliege Bird, R., Boivin, N., Díaz, S., Fuller, D. Q., Gill, J. L., Kaplan, J. O., Kingston, N., Locke, H., McMichael, C. N. H., Ranco, D., Rick, T. C., Shaw, M. R., Stephens, L., Svenning, J.-C., & Watson, J. E. M. (2021). People have shaped most of terrestrial nature for at least 12,000 years. Proceedings of the National Academy of Sciences, 118(17), e2023483118. https://doi.org/10.1073/pnas.2023483118
Gibbs, H. K., Ruesch, A. S., Achard, F., Clayton, M. K., Holmgren, P., Ramankutty, N., & Foley, J. A. (2010). Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. Proceedings of the National Academy of Sciences, 107(38), 16732–16737. https://doi.org/10.1073/pnas.0910275107
Haddad, N. M., Brudvig, L. A., Clobert, J., Davies, K. F., Gonzalez, A., Holt, R. D., Lovejoy, T. E., Sexton, J. O., Austin, M. P., Collins, C. D., Cook, W. M., Damschen, E. I., Ewers, R. M., Foster, B. L., Jenkins, C. N., King, A. J., Laurance, W. F., Levey, D. J., Margules, C. R., … Townshend, J. R. (2015). Habitat fragmentation and its lasting impact on Earth's ecosystems. Science Advances, 1(2), e1500052. https://doi.org/10.1126/sciadv.1500052
Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., & Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342(6160), 850–853. https://doi.org/10.1126/science.1244693
Lewis, S. L., Edwards, D. P., & Galbraith, D. (2015). Increasing human dominance of tropical forests. Science, 349(6250), 827–832. https://doi.org/10.1126/science.aaa9932
Lister, T. W., Lister, A. J., & Alexander, E. (2014). Land use change monitoring in Maryland using a probabilistic sample and rapid photointerpretation. Applied Geography, 51, 1–7. https://doi.org/10.1016/j.apgeog.2014.03.002
Newbold, T., Hudson, L. N., Hill, S. L. L., Contu, S., Lysenko, I., Senior, R. A., Börger, L., Bennett, D. J., Choimes, A., Collen, B., Day, J., De Palma, A., Díaz, S., Echeverria-Londoño, S., Edgar, M. J., Feldman, A., Garon, M., Harrison, M. L. K., Alhusseini, T., … Purvis, A. (2015). Global effects of land use on local terrestrial biodiversity. Nature, 520(7545), 45–50. https://doi.org/10.1038/nature14324
O'Shea, T. (2020, September 22). Planet, KSAT and airbus awarded first-ever global contract to combat deforestation. Planet. https://www.planet.com/pulse/planet-ksat-and-airbus-awarded-first-ever-global-contract-to-combat-deforestation/
Pendrill, F., Gardner, T. A., Meyfroidt, P., Persson, U. M., Adams, J., Azevedo, T., Bastos Lima, M. G., Baumann, M., Curtis, P. G., De Sy, V., Garrett, R., Godar, J., Goldman, E. D., Hansen, M. C., Heilmayr, R., Herold, M., Kuemmerle, T., Lathuillière, M. J., Ribeiro, V., … West, C. (2022). Disentangling the numbers behind agriculture-driven tropical deforestation. Science, 377(6611), eabm9267. https://doi.org/10.1126/science.abm9267
Planet Team PBC. (2017). Planet Application Program Interface: In Space for Life on Earth. https://api.planet.com
Schepaschenko, D., Chave, J., Phillips, O. L., Lewis, S. L., Davies, S. J., Réjou-Méchain, M., Sist, P., Scipal, K., Perger, C., Herault, B., Labrière, N., Hofhansl, F., Affum-Baffoe, K., Aleinikov, A., Alonso, A., Amani, C., Araujo-Murakami, A., Armston, J., Arroyo, L., … Zo-Bi, I. C. (2019). The Forest Observation System, building a global reference dataset for remote sensing of forest biomass. Scientific Data, 6(1), 198. https://doi.org/10.1038/s41597-019-0196-1
Wright, S. J. (2005). Tropical forests in a changing environment. Trends in Ecology & Evolution, 20(10), 553–560. https://doi.org/10.1016/j.tree.2005.07.009
York, N. D. L., Pritchard, R., Sauls, L. A., Enns, C., & Foster, T. (2023). Justice and ethics in conservation remote sensing: Current discourses and research needs. Biological Conservation, 287, 110319. https://doi.org/10.1016/j.biocon.2023.110319

Applications of Museum Collections and Genomics to Biodiversity Conservation

Anna Penna[i,ii,iii], Lauren T. Clark[iv,v], Alexander T. Salis[vi], Suzanne K. Macey[vii], and Mary E. Blair[vii]

[i]*National Museum of Natural History, Smithsonian Institution, Washington, DC, USA;* [ii]*Department of Anthropology, University of Texas at San Antonio, San Antonio, TX, USA;* [iii]*Department of Biology, Lund University, Lund, Sweden;* [iv]*Institute of Comparative Genomics, American Museum of Natural History, New York, NY, USA;* [v]*Centre for Paleogenetics, Stockholm University, Stockholm, Sweden;* [vi]*Department of Herpetology, American Museum of Natural History, New York, NY, USA;* [vii]*Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA*

DOI: https://doi. org/10.5531/cbc.linc.14.1.7 | Supplementary: http://doi.org/10.5531/cbc.ncep.0190

ABSTRACT

To address the challenges of sampling endangered or extinct species in the field, many studies have turned to historically underutilized sources of genetic material: natural history museums. Despite the fact that DNA from specimens collected decades or even hundreds of years ago is often fragmented and degraded, research has shown that historical DNA can still be used effectively to infer phylogenetic relationships and intra-specific patterns of population genetic structure. This synthesis aims to provide students and conservation practitioners with a solid understanding of the methodological strategies needed to apply genetic tools to natural history museum specimens. Specifically, we offer clear definitions and essential considerations for designing a conservation genomics project that includes both modern and historical samples. We recommend that instructors use this synthesis to introduce the foundational knowledge required for two companion exercises: "The application of conservation museomics approaches to the protection of the Iberian Lynx (*Lynx pardinus*)" and "Designing a conservation genomics project incorporating DNA from museum specimens."

INTRODUCTION

Natural history museums and herbaria collections provide a unique spatiotemporal record of life on Earth. These collections house millions of specimens derived from animals, plants, fungi, and other organisms, often consisting of entire preserved organisms or their body parts, commonly referred to as "voucher specimens." Since the sixteenth century, naturalists have preserved these specimens for long-term use in morphology-based taxonomic work and comparative anatomy analyses. Recent advancements in laboratory techniques for studying DNA damage, coupled with the advent of high-throughput sequencing technologies, have transformed these century-old museum specimens into valuable sources of DNA, giving rise to the field of "museomics" (Raxworthy & Smith, 2021; Blair, 2024; Fong et al., 2023). From historical cabinets to modern molecular laboratories, museum collections are now integrating novel technologies that significantly enhance the scientific value of these specimens.

To obtain DNA for genetic analysis, researchers start with a small fragment of biological tissue (typically 50–200 milligrams of bone, teeth, ear punches, hair, etc.) from an organism. The DNA isolated from these samples is categorized based on the time since collection (modern, historical, or ancient) and the preservation process of the tissue source (Table 1). Any DNA recovered from a specimen intentionally preserved for long-term storage in a museum collection is referred to as "archival DNA" (Raxworthy & Smith, 2021). This archival DNA is often exposed to chemicals during preservation, leading to various types of DNA decay beyond the environmental factors that can damage naturally preserved specimens. Additionally, these chemicals can inhibit enzymatic reactions during DNA extraction protocols. In contrast, naturally-preserved specimens, such as those found in

Table 1: DNA can be obtained from samples of various ages (modern, historical, ancient) and collected from different types of biological material that was intentionally prepared for preservation (archival DNA) or has been preserved under certain natural conditions.

	Modern	**Historic**	**Ancient**
Preservation/ age	Typically younger than 30 years old	Between 30* and ~200 years old	Usually >200 years
Archival	Biological tissue preserved for long-term use in genetic research, following a methodology that minimizes the chances of DNA damage (e.g., frozen, or preserved in ethanol or buffer).	Tissue collected from a voucher specimen preserved in natural history collection for long-term morphological research (e.g., dry study skin, formalin-fixed, pinned specimens). Preservation protocols did not account for the risk of DNA damage, often resulting in genetic material that shows some level of damage.	Biological material collected from artifacts intentionally treated to maximize its lifespan but without considering the risk of DNA damage (e.g., the leather from a book cover from medieval times). However, the utility of ancient archival specimens in genetic studies is rarely explored due to the low probability of yielding amplifiable DNA.
Natural preservation	Biological tissue found exposed to the elements (e.g., forensics samples).	Biological material collected from ethnographic collections or archaeological sites.	Biological material collected from archaeological and/or paleontological sites.

*Samples younger than that of 30 years may also demonstrate DNA degradation patterns consistent with historic tissues. This is typical of formalin-fixed specimens or soft tissues that have been exposed to sunlight, heat, fluctuations in temperature, or humidity for long periods of time.

archaeological or paleontological sites, are not typically subjected to chemical treatments and are preserved due to specific environmental conditions (e.g., dry and cold environments). These naturally preserved samples, typically dated to be around 200 years old or older, are labeled as "ancient DNA" (Wandeler et al., 2007; Billerman & Walsh, 2019). The field specializing in ancient DNA samples is known as "paleogenomics."

Most specimens currently housed in museum collections and herbaria were collected less than 200 years ago and are classified as "historical DNA" (Wandeler et al., 2007; Billerman & Walsh, 2019). The majority of these specimens were preserved before the rise of genetic research in non-model organisms in the late 1970s. In recent times, genetic research has become central to biological studies, prompting researchers to collect small tissue samples before preparing voucher specimens for long-term preservation ("modern DNA"). This practice helps protect potential DNA sources from various factors that can cause DNA damage.

Museomics, the study of genetic information acquired from museum specimens, leverages thousands of historic and ancient specimens collected worldwide to provide a time series of genetic

diversity and biodiversity monitoring that was unavailable just 20 years ago (Blair, 2024). The rapid advancements in the ability to access genetic information to monitor diversity throughout time from extinct taxa of the past or endangered taxa of the present comes at a crucial time as the warming climate and human activity continues to accelerate the loss of biodiversity. The past few hundred years offer a clear picture of this loss, with species extinctions, shifts in species distributions, and biodiversity declines occurring at rates 100 to 1,000 times greater than the natural background extinction rate throughout geological time (Ceballos et al., 2015; De Vos et al., 2015).

This biodiversity crisis not only impacts ecosystems and species distributions but also affects genetic variation within species and populations. A recent study estimated a six percent loss of genetic variation in wild populations since the Industrial Revolution (Leigh et al., 2019). To better understand these changes in biodiversity, natural history collections provide a unique resource of DNA data for many rare, endangered, and locally or globally extinct species at different time points. By integrating genetic data from field studies and museum collections, museomics allows researchers to include more species in comparative analyses, establish baselines of genetic diversity for declining species, and reduce the cost and time associated with fieldwork. Museum samples have also enabled us to study the emergence of zoonotic diseases like Lyme Disease (Marshall et al., 1994) and avian influenza (Fanning et al., 2002). In unique cases, museomics may also contribute to ongoing conservation efforts (Blair, 2024; Fong et al., 2023; Raxworthy & Smith, 2021). Though the implementation of conservation efforts is not always possible or relevant to all research questions, the generation of genetic information from modern and historic specimens provides an excellent resource for future research and adds to the growing field of conservation genomics.

What is Conservation Genetics?

Conservation genetics is a field of biology that intersects with ecology, taxonomy, molecular biology, and population genetics. It applies genetic principles and tools to the conservation and management of species, populations, or evolutionarily significant units (ESUs), particularly those that are endangered or at risk of extinction, with the goal of studying their health and viability (Desalle & Amato, 2024). Key parameters in population genetics—genetic relationships, genetic diversity, and population size—are crucial for understanding the evolutionary dynamics and long-term survival of species. Genetic relationships reveal how individuals within a population are related, helping to assess inbreeding levels and genetic drift. Genetic diversity, or the variety of traits within a population, is essential for adaptability, as it increases a population's ability to respond to environmental changes and resist diseases. Low diversity can lead to inbreeding depression and a reduction in fitness. Population size influences genetic variation, with small populations more prone to genetic drift and loss of diversity, while larger populations tend to maintain greater genetic variation, supporting long-term resilience. Conservation genetics focuses on maintaining genetic diversity to ensure species' adaptability and survival, helping wildlife managers predict a population's genetic health and mitigate risks such as inbreeding, loss of diversity, and hybridization.

Scientists working in conservation genetics often study endangered or threatened species to understand the genetic factors that influence their survival. Key questions typically include: What has caused the genetic erosion (i.e., loss of genetic diversity) in this population? How can conservation managers mitigate this and improve the population's genetic health? These questions help identify the root causes of reduced genetic diversity, such as habitat loss, loss of connectivity between populations, over-exploitation, or climate change and inform strategies to increase population resilience.

Prior to the revolution of genetics in the early 2000s and genomics in the 2010s, the information

typically available to protect at-risk populations was limited to observations of populations in the field (Box 1). Now, with improved genomic techniques and the ability to access historic and ancient data, conservation geneticists can better understand genetic relationships across time and geographic ranges. This expanded toolkit allows them to more accurately trace the causes of genetic decline and devise more effective conservation strategies.

When examining fluctuations in genetic diversity, it is essential to revisit the fundamental mechanisms of evolution, such as mutation, gene flow, genetic drift, and natural selection. If comfortable with these concepts, proceed to the next section. If you need a more substantial overview of these concepts (and more!), we suggest you read the NCEP module "Conservation genetics"(https://doi.org/10.5531/cbc.ncep.0123).

Mutations serve as the primary source of genetic diversity, arising from stochastic changes that typically occur as molecular 'errors' during cellular replication. Genetic diversity refers to the range of genetic differences among individuals within a population. When these differences approach zero—indicating that individuals within a population, species, or evolutionarily significant units possess nearly identical genetic information across protein-coding regions—the population becomes increasingly vulnerable to genetic diseases and the spread of harmful alleles. This genetic uniformity can severely impair a population's ability to adapt to changing environmental conditions. If a crucial gene associated with fitness is compromised, the overall health, survival, and adaptability of the group can rapidly decline.

Low genetic diversity not only increases the risk of disease but also elevates the likelihood of extinction. Populations with limited genetic variation may also struggle to cope with environmental changes, or invasive species. This lack of adaptability can be particularly detrimental in the face of rapid climate change, habitat destruction, and other anthropogenic pressures, which are now more prevalent than ever. The inability to adapt can lead to population declines and eventual extinction, thereby triggering downstream ecological effects that disrupt entire ecosystems. This can result in extinction cascades, where the decline or extinction of one species negatively impacts others that rely on it for vital ecosystem services, such as pollination, nutrient cycling, or habitat provisioning.

Box 1: Conservation genetics vs. conservation genomics

Conservation genetics and conservation genomics are both key terms to describe how molecular DNA can be applied to questions critical fields in species conservation, and these two terms are used to signal differences in but they differ in scope and technological approach. Conservation genetics focuses on studying genetic diversity within and between populations by analyzing a few specific genes or markers, such as mitochondrial DNA and microsatellites. It helps assess inbreeding, population structure, and gene flow and is useful in managing small or endangered populations. However, it is limited by the small number of markers studied, providing a less comprehensive view of genetic health. Conservation genomics, on the other hand, utilizes advanced sequencing technologies like whole-genome sequencing to analyze the entire genome, offering a more in-depth understanding of genetic variation, including both neutral and adaptive traits. This broader approach allows for better insights into how species adapt to environmental changes and threats like habitat loss or climate change, making it a powerful tool for long-term conservation strategies. While conservation genetics lays the foundation, conservation genomics provides a more detailed, genome-wide perspective that can enhance species management and adaptation efforts.

Additionally, genetic diversity plays a crucial role in maintaining the resilience of ecosystems. Populations with greater genetic variation are more likely to contain individuals with traits that can withstand environmental stresses, ensuring the survival of species during adverse conditions. For instance, a genetically diverse plant population may produce individuals that can survive drought, while a uniform population may not have any individuals with such adaptive traits. This resilience contributes to the stability and functioning of ecosystems, benefiting not just the species involved but also the human communities that depend on these ecosystems
for resources and services.

Once genetic diversity is established, these evolutionary mechanisms interact with it in ways that vary based on the population's size and connectivity. It's important to recognize that these mechanisms do not operate in isolation; they are influenced by demographic and geographical factors. For example, if two populations of the same species become physically isolated from one another, their gene flow—the exchange of genetic material through mating—can be significantly reduced. This decrease in gene flow can lead to the fragmentation of the species into smaller, more isolated populations, where the effects of genetic drift become pronounced. Over time, if these isolated populations remain separate, the resultant loss of genetic diversity is not solely due to diminished gene flow; it is also a consequence of genetic drift and/or inbreeding depression acting more strongly on smaller populations. Consequently, understanding the interplay between these evolutionary mechanisms is crucial for assessing and managing genetic diversity in conservation efforts.

Conservation geneticists use a variety of tools to assess the genetic diversity and health of populations or species. Before collecting samples for genetic or genomic analysis, researchers first define the population and the geographic area of interest, focusing primarily on species that are threatened, vulnerable, or endangered. They may then analyze known phenotypes within the species and identify the species' closest living relatives.

With these foundational questions answered, researchers can develop hypotheses that require biomolecular data, such as genetic, genomic, transcriptomic, or proteomic information. This data is integrated into models along with previously gathered phenotypic and ecological data. These models can provide insights into how adaptable a species or population may be to changing environmental conditions, such as climate and habitat alterations, and help predict future impacts on the species' viability or on specific evolutionarily significant units.

Incorporating Historical and Ancient Samples in Conservation Genetics Studies

A pioneering study by Higuchi et al. (1984) demonstrated the feasibility of recovering authentic DNA from dried muscle tissue obtained from a 140-year-old mounted taxidermy specimen of a quagga (*Equus quagga quagga*; Figure 1). The quagga, a subspecies of the African plains zebra, was hunted to extinction in the 19th century. This groundbreaking study paved the way for researchers to obtain DNA from extinct and rare species housed in museum collections. Since then, numerous studies have incorporated historical and ancient samples into genetic analyses, enriching both the temporal and geographical dimensions of genetic research (Mitchell & Rawlence, 2021; Raxworthy & Smith, 2021).

One common application of museomics, as exemplified by the quagga study, is sequencing DNA from extinct species, such as the iconic dodo (*Raphus cucullatus*; Shapiro et al., 2002), the Tasmanian tiger (*Thylacinus cynocephalus*; Feigin et al., 2018; 2022; Mármol-Sánchez et al., 2023), the wooly mammoth (*Mammuthus primigenius*; van der Valk et al., 2021), and the blue antelope (*Hippotragus leucophaeus*; Plaxton et al., 2023; Hempel et al., 2021). These sequences help confirm the

Figure 1. A museum specimen of the quagga (*Equus quagga quagga*) prepared for display at the Natural Science Museum of Bamberg (Germany) and later sampled for genetic analysis. Image credit: Reinhold Möller via Wikimedia Commons/CC BY-SA 4.0.

taxonomic identity and determine the phylogenetic placement of globally or locally extinct species or populations. Better phylogenies lead to more informed and accepted taxonomies; taxonomic resolution is critical to accurately set species priorities for conservation efforts (Mace, 2004). With the refinement of laboratory protocols to isolate endogenous DNA from preserved specimens, museomics has expanded beyond analyses involving only one or a few individuals, enabling population-level studies.

Natural history collections also offer an alternative to disturbing sensitive ecosystems and sampling protected live organisms, species in remote or inaccessible regions (e.g., areas experiencing conflict or war), or broadly distributed taxa for which collection permits may be difficult to obtain (e.g., Blair et al., 2023; Islam et al., 2024; Penna et al., 2024). By providing researchers access to samples of species and populations that would otherwise be challenging and expensive to collect in the wild, museomics can expand the geographical and taxonomic coverage of genetic studies. These historical samples have been applied to study biogeography, delimit species boundaries, clarify species diversity, and characterize hybrid zones (Raxworthy & Smith, 2021). With the global acceleration of human impacts on biodiversity and habitat loss in recent decades, arguably, the collection of voucher specimens for future research is becoming increasingly important (Rocha et al., 2014). Natural history collections and museomics can offer many benefits to conservation, but it is important to consider the ethics and social/historical context of collections and data derived from collections (Box 2).

The temporal information encoded in ancient and historical specimens allows researchers to address a broader range of questions. By combining modern and museum samples, it is possible to study the dynamics of genetic diversity over time. While historical specimens offer a more limited temporal reach compared to ancient specimens (hundreds of years versus thousands to over a million), the extensive collection of specimens amassed over the last ~200 years provides an unparalleled resource for studying the spatiotemporal records of species and populations globally, spanning multiple generations (Raxworthy & Smith, 2021). Because many historical samples were collected

Box 2: Museomics and ethics

Advances in museomics hold great promise to transform research and practice in biodiversity conservation. At the same time, the increasing use of these technologies and data provokes key ethical questions, some of which relate to general ethical questions in science and in -omics, such as cloning and artificial intelligence, while others relate to the ethics of natural history collections themselves. There are several mechanisms such as formal laws, institutional frameworks and policies, and journal review procedures that are relevant to and that inform ethical decisions made around the curation of and research allowed on natural history collections. Importantly, the ethics and strategies guiding the collection of biological specimens for natural history museums have been renewed and updated as the purpose of museums has evolved (Arengo et al., 2018). Yet, criticism and ethical debates continue to surround biological specimen collection for scientific studies, with some claiming that collection plays a role in species extinction (Minteer et al., 2014), and others arguing that the practice is no longer necessary given new technologies (Byrne, 2023). Collection strategies and institutional policies for collecting do adhere to strict permitting and ethical boundaries, for example by aiming to collect well below levels that would affect demography (Collar, 2000; Winker et al., 2010; ICOM, 2013; Rocha et al., 2014). Others argue that policies and strategies for collecting can and should be grounded in partnerships and agreements with Indigenous guardians and other local resource stewards towards co-stewardship of biodiversity and environmental sustainability (e.g., Blair, 2024).

Thus, conservation museomics, especially as it expands and interacts with a range of disciplines including Indigenous science, anthropology, archaeology, environmental humanities, and museum studies (Blair, 2024), can benefit from broader discussions about data ethics in many fields ranging from remote sensing (e.g., York et al., 2023) to Indigenous data sovereignty. For example, while the incorporation of cultural treasures in museomics is likely to advance the conservation relevance of the field, especially regarding explorations of human relationships with nature and biodiversity, such work must privilege the rights and wishes of Indigenous Peoples and descendant communities to determine how their cultural material and knowledge are used (Kreps, 2008; Simpson, 2009). Further, conservation museomics should consider whether data derived from specimens previously collected from Indigenous lands should be managed in collaboration with descendant communities following the CARE principles for Indigenous Data Governance (Carroll et al., 2020; 2021). Following these principles, Indigenous knowledge holders must be actively involved in any additional collection, as well as in the stewardship and governance of data to ensure ethical reuse of data in specific defined ways. The processes outlined in the CARE principles aim to ensure that Indigenous ethics inform access to data in a way that minimizes harm, maximizes benefits, and allows for future use that promotes the well-being of the descendant community.

when habitat coverage was significantly different, museomics can reveal the impact of climatic events, environmental changes, or human activities. This cross-generational analysis can help disentangle contemporary declines in genetic variation from historically low values driven by ancient population crashes, species-specific traits, and demographic dynamics (Schmid, 2018; Jensen et al., 2022; Brasil et al., 2023). Together, the advantages of population-level and cross-generational analysis allow for a more accurate characterization of changes in historical population size, distribution, connectivity, and relationships, which are crucial for developing conservation and management plans (Leonard, 2008). Genetic data obtained from museum specimens can also potentially help study other biological phenomena, such as biological invasions and species interactions that extend beyond the organisms' DNA, for instance, giving insights into microbiomes, diets, diseases, and parasites.

From the Field to the "Shelf"

According to a 2020 report by the National Academies of Sciences, Engineering, and Medicine, natural history museums in the US alone house between 800 million to 1 billion specimens (National Academies of Sciences, Engineering, and Medicine, 2020). These specimens are organized into dedicated collections based on major biological and taxonomic groups. For example, vertebrate zoology collections include fishes, amphibians, reptiles, birds, and mammals; invertebrate zoology collections encompass insects and related groups (e.g., spiders and crabs), soft-bodied animals (e.g., snails and worms), and various marine organisms (e.g., coral and starfish); and herbaria contain collections of plants, algae, and fungi. Specimens are collected from the field, prepared, preserved, stored, and sampled in various ways for genetic studies.

Collection of Fresh Tissue Samples

In the field, researchers collect blood and/or tissue from various body parts (muscle, spleen, liver, ear-clip), depending on the study question and organism. To collect blood or tissue samples from wild specimens, researchers often use invasive sampling procedures, such as syringes for blood or scissors for internal organs (muscle, liver, or spleen) and external structures (ear, toe, nail). The invasive sampling of internal organs is strictly regulated and requires humane euthanasia, often requiring prior approval from the Institutional Animal Care and Use Committee (IACUC), while the sampling of external structures can often be done without euthanasia. Alternatively, researchers can noninvasively sample biological material left behind by animals in the wild, such as shed hair or feathers, eggshells, and scat.

Including noninvasive sampling for further genetic analysis can help avoid unnecessary destructive sampling of tissue from limited specimens and reduce the bureaucratic burden (e.g., approval of invasive sampling permits) and other operational costs associated with sampling (e.g., setting up multiple traps). However, isolating DNA from noninvasive samples can be challenging due to high levels of DNA damage, lower amounts of DNA, the potential presence of DNA from other organisms (such as environmental microorganisms or microbiota), or the absence of genomic sequences in reference databases, such as GenBank (Locatelli et al., 2020). To minimize these challenges, noninvasive samples should be collected as fresh as possible, and specific laboratory procedures for DNA extraction should be used.

Preparation, Preservation, and Storage of Fresh Tissue

Once fresh tissue samples are collected in the field, they must be preserved and stored individually. Noninvasive samples, such as hair, feathers, or nails, can be kept dry at room temperature or preserved in silica desiccant if they contain moisture (e.g., scat). Fresh tissue, however, is usually preserved in liquid media and kept cold. These samples are often stored in ethanol or another buffer (e.g., RNA later, DNA/RNA Shield) and kept at 4°C, or even -20°C, when possible. If the study requires the preservation of high molecular weight DNA, it is generally argued fresh samples should be flash-frozen immediately after collection and processed for DNA extraction and sequencing as soon as possible (Wong et al., 2012; Dahn et al., 2022). However, flash-freezing in the field can present logistical challenges, especially in remote locations, and can potentially lead to DNA degradation, where putting into ethanol or other buffers may be better (Salis et al., 2025), especially if tissues samples are to be used for multiple projects and stored for future use (for example archived into a "cryo-collection").

Many museums also maintain "cryo-collections" for the long-term preservation of frozen tissue samples collected during modern expeditions, enabling the subsampling of these tissue samples in subsequent genetic research (Radin, 2015). In this case, fresh tissue samples are transferred to -80°C freezers or liquid nitrogen vapor freezers (Figure 2). Storing DNA at such a low temperature and minimizing the

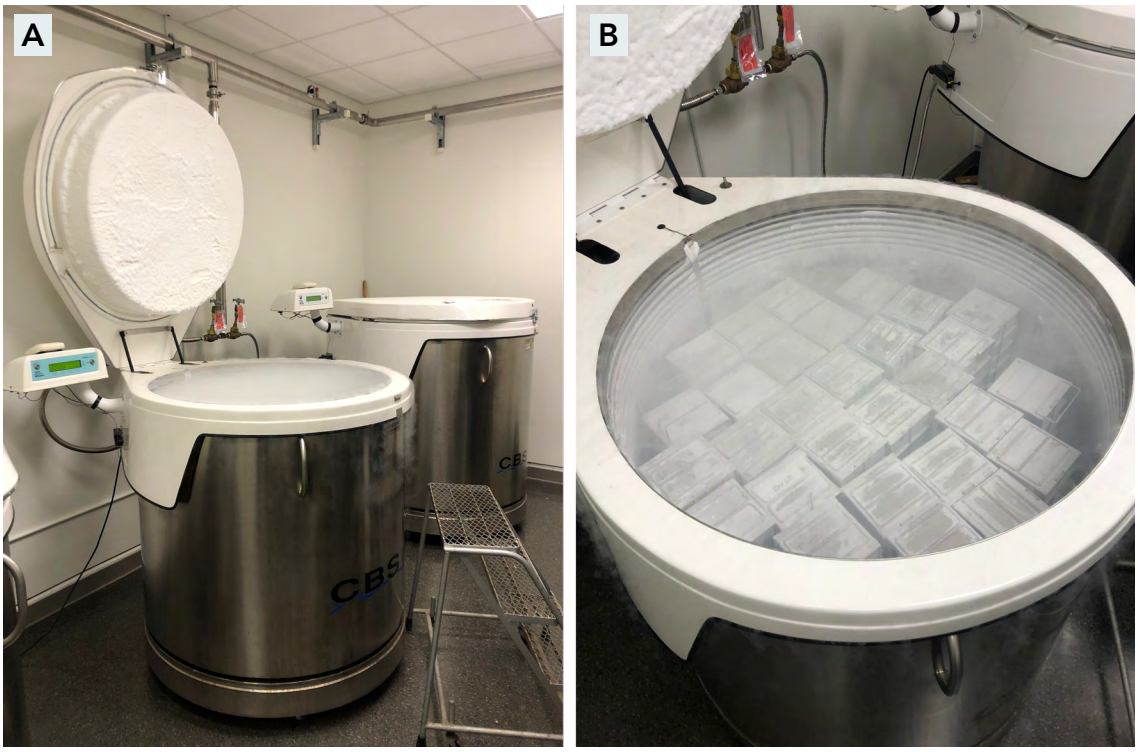


Figure 2. Liquid nitrogen vapor freezers used for long-term storage of biological samples. (A) Shows the tanks from the outside, and (B) shows the interior of a tank filled with 29 vertical stainless steel freezer racks that each fit 9 boxes of 96 samples. Image credit: Anna Penna.

number of thawing cycles helps slow down or halt nucleic acid degradation. These best practices minimize the chances of degrading DNA and other biomolecules for future use. Although some tissue samples stored in cryo-collections were first preserved in the early years of genetic research, they exhibit minimal DNA damage and can still be considered "modern DNA" (Raxworthy & Smith, 2021).

Preservation of Voucher Specimens

The method of preparation used for long-term preservation of voucher specimens varies depending on the organism, tissue composition, and the type of information researchers commonly seek from these specimens (Raxworthy & Smith, 2021). For instance, the color and morphology of bird feathers and mammal hair are crucial for species identification. Therefore, bird and mammal specimens are traditionally preserved as dried skins using a combination of salts and cornstarch. Insects, which have a hard exoskeleton made of chitin, usually maintain their original shape and appearance once dried. However, because these small creatures can be fragile to the touch, they are often preserved using pins inside cardboard boxes.

Aquatic animals, such as marine vertebrates and invertebrates, tend to become too distorted if dried, so they are typically preserved in fluid inside glass jars. Fluid-preserved specimens are also common in herpetological collections (amphibians and reptiles) because these animals have very thin skins that lose their general appearance when dried. Historically, fluid preparations involved an initial step of fixing tissues with formalin to slow down decomposition, followed by transfer to ethanol. More recently, researchers have started preserving specimens in ethanol only, rather than formalin, although this practice varies significantly from country to country (Hahn et al., 2022). For more details on the extensive variety of chemical compounds employed by different institutions across various taxa over the last centuries, you can refer to Simmons (2014) comprehensive catalog.

Table 2 summarizes the most common sources of DNA for museum specimens of different types of organisms, and Figure 3 illustrates some of the different specimen preservation methods.

Table 2: Summary of major types of voucher specimen preservation for different taxonomic groups that can be used as a source of biological tissue for genetic analysis.

Taxonomic group	Preparation type used as a source of tissue
Mammals, birds	Dried skin, bone, teeth, osteocrusts (dried tissue attached to bone, such as muscle or brain), occasionally fluid-preserved specimens (usually stored in ethanol 70%, can contain formalin).
Fishes, amphibians, reptiles	Fluid-preserved specimen (usually stored in ethanol 70%, can contain formalin), bone.
Some invertebrates (e.g., mollusks, arachnids)	Fluid-preserved specimen (usually stored in ethanol 70%, can contain formalin).
Insects, plants, algae, fungi	Dried specimen, pressed in paper, or pinned.

DNA Damage

Routinely obtaining DNA from historical and ancient sources has become possible in the last few decades thanks to recent advances in laboratory protocols for DNA extraction and library preparation (described in more detail in the sections below) and the advent of next-generation sequencing (Chen & Nedoluzhko, 2023). These protocol modifications aim to overcome the challenges of dealing with the differences in size and structural composition of these DNA molecules due to the accumulation of damage.

DNA replication is at the core of the most commonly used sequencing technologies. During replication, a double-stranded DNA molecule is copied to produce two identical molecules. DNA polymerases (a group of enzymes that catalyze DNA synthesis during replication) can duplicate the sequence of nucleotides found in DNA molecules by adding nucleotides at the 3'-OH group of another nucleotide, thus extending the DNA strand being copied. So, if the original DNA molecules present some type of structural or compositional damage, the DNA polymerase will not be able to properly attach to the DNA molecule or extend the nucleotide sequence. Therefore, to better understand the rationale behind the choice of laboratory protocols that should best suit the type of museum specimen available for the project, the researcher must first consider the consequences of these damages to DNA synthesis.

The DNA molecules inside living organisms' cells suffer damage that is repaired by intricate cell machinery. After death, these cellular repair mechanisms stop functioning and the DNA molecules are exposed to numerous factors that threaten its stability. These include digestion by intracellular nucleases and microorganisms, oxidative and hydrolytic damage that results in DNA shearing (that can result in how a nucleotide base is called by a sequencing machine), as well as attachment to proteins and other DNA molecules (crosslink). Consequently, DNA breakages and damages accumulate over time, at rates that can vary depending on the preservation method. That is why researchers usually separate non-modern samples into two categories (i.e., historical and ancient), and differentiate archival from naturally-preserved DNA.

DNA damage can be divided into three main categories (Figure 4): (1) Fragmentation: lesions that lead to a shortening or "fragmentation" of the DNA molecule; (2) Blocking Lesions: lesions that prevent the replication of the DNA molecule by blocking the action of the polymerases; and (3) Miscoding Lesions: lesions that result in the incorrect incorporation of nucleotides during DNA

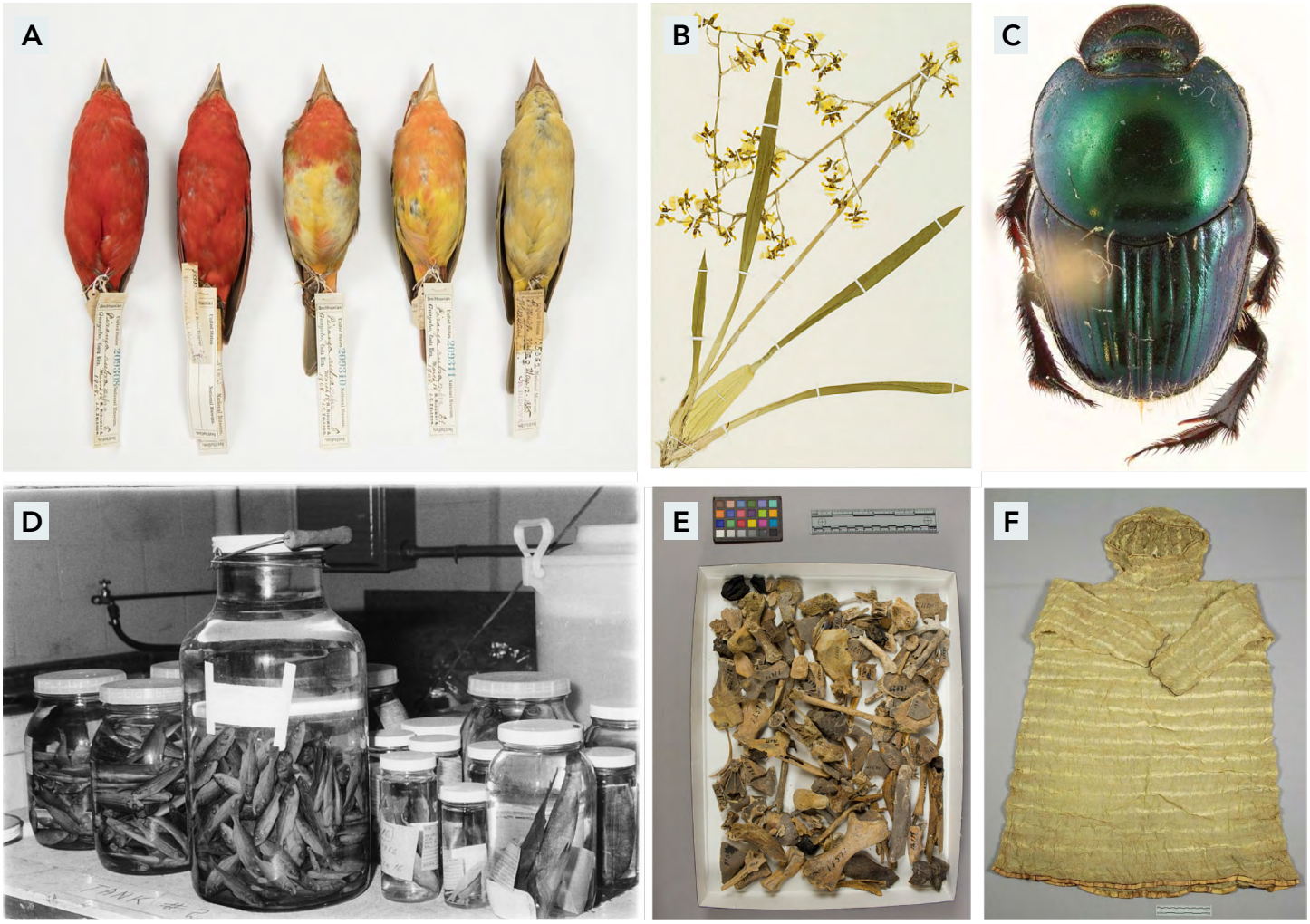


Figure 3. Some examples of specimen preparations often found in natural history collections. (A) Birds and mammal specimens are usually prepared as dry study skins. (B) Plant and algae specimens are dried and pressed on paper. (C) Insects are often preserved as pinned dried specimens. (D) Bats, aquatic organisms (e.g., fish, amphibians, mollusks), and reptiles (e.g., snakes, lizards) are commonly preserved in spirits (e.g., ethanol). Specimens with all internal organs (e.g., for anatomical studies) are also preserved in spirits. Small-sized specimens are most often stored in glass jars, whereas large-bodied specimens are stored in large aluminum cases. (E) Bones are a common find in archaeological collections and can be a great source of DNA. (F) Ethnographic collections often have material objects made of animal and plant parts, such as this waterproof parka made of sea lion intestines by the Aleuts from the North Pacific in the early 1800s. Image credit: (A) "Summer Tanager (*Piranga rubra rubra*), Indiana, USA, 1885" via Smithsonian Institution National Museum of Natural History Division of Birds. (B) "Popcorn Orchid" via Smithsonian Institution National Museum of Natural History. (C) "African elephant dung beetle, Mpala, Kenya, 2001" via Smithsonian Institution National Museum of Natural History Department of Entomology. (D) "Marine specimens in a jar" via Smithsonian Institution National Museum of Natural History Division of Fishes. (E) "Unidentified assemblage of bones collected in an archaeological site in the Dominican Republic in 1872" via Smithsonian Institution National Museum of Natural History Department of Anthropology. (F) "Waterproof parka made of the peritoneal coat of the intestines of sea lions, Aleuts from the Pacific Northwestern Coast, USA, acquired in 1829–1830" via Smithsonian Institution National Museum of Natural History Department of Anthropology.

replication, which are called "missense" nucleotides. Compared to the DNA present inside the cells of living organisms, the DNA of preserved specimens is often sheared into smaller fragments, cytosine bases are deaminated to uracil bases that occur frequently at the end of DNA fragments and degraded DNA molecules often crosslink to other DNA molecules or proteins.

DNA fragmentation leads to shorter DNA fragments either when both strands of the DNA double helix are broken or when only one strand is shortened, such as in single-strand overhangs. Damage

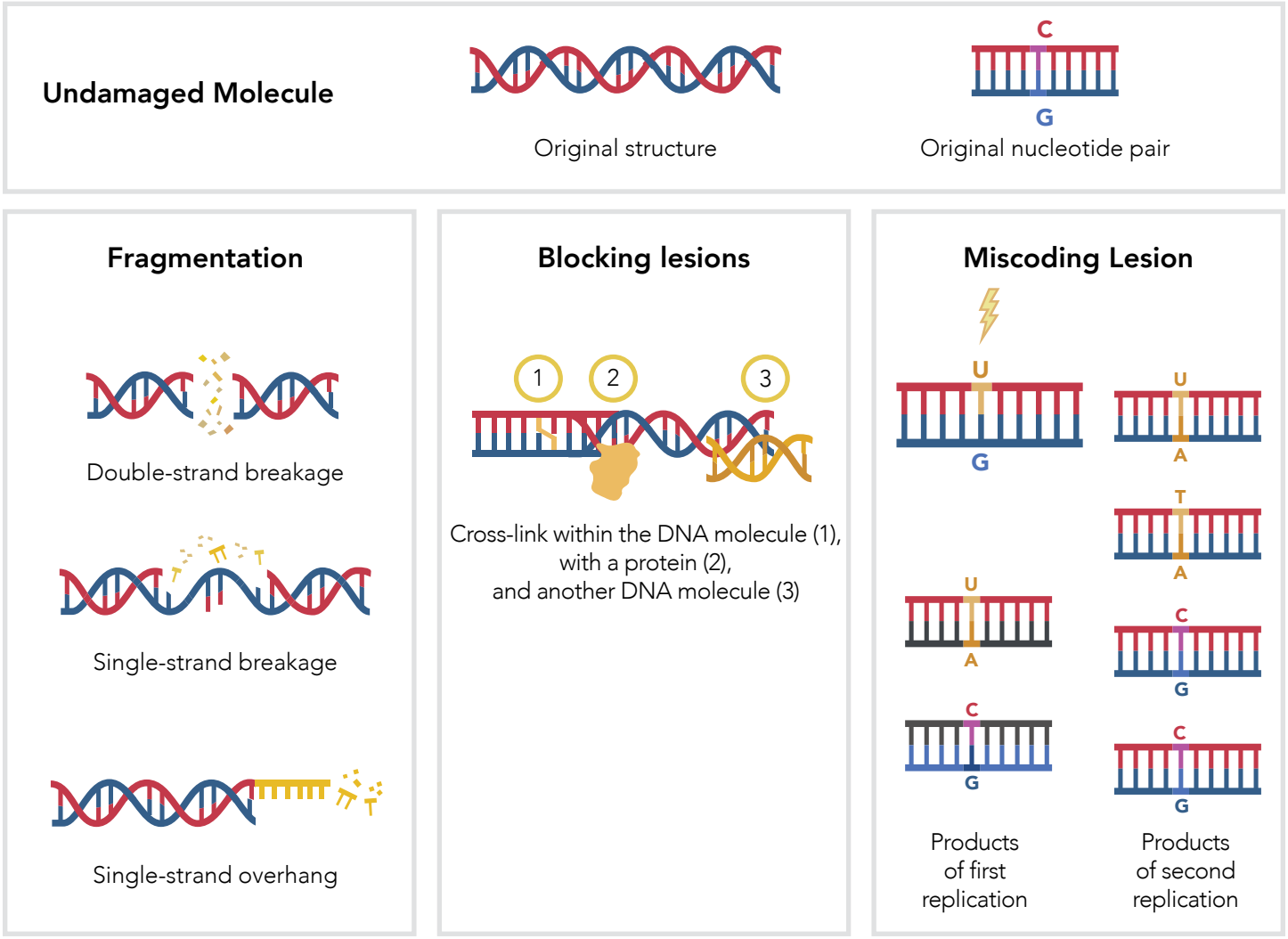


Figure 4. Examples of the three main types of lesions found in DNA molecules: DNA fragmentation, blocking, and miscoding lesions. Image credit: Anna Penna.

to a single strand can also lead to the formation of nicks, where the phosphodiester bond between two adjacent nucleotides is broken. These nicks prevent the extension of the nucleotide chain during replication, potentially resulting in shorter fragments if not repaired. Examples of blocking lesions include the formation of crosslinks within the DNA molecule or between the DNA molecule and proteins or other DNA fragments in the solution. Blocking lesions are common in formalin-fixed specimens and can interfere with DNA replication.

Miscoding lesions lead to the deletion or incorrect incorporation of nucleotides during replication. For instance, one of the most common miscoding lesions found in ancient DNA is the deamination of cytosine to uracil. In this case, DNA polymerases will incorporate an adenine (A) opposite the uracil (U) and, subsequently, a thymine (T) opposite the adenine, causing apparent guanine (G) to adenine (A) and cytosine (C) to thymine (T) substitutions, depending on the strand sequenced. Typically, these miscoding lesions accumulate at the ends of fragments.

As described above, there are multiple sources of ancient and historical tissues from museum specimens that are suitable for genetic analysis. Differences in the types of tissues sampled, the age of the sample, environmental conditions (humidity, temperature, salinity, and pH) and preservation conditions (use of chemicals by museum staff to aid in morphological preservation) can all contribute

to the significant variability observed in the preservation of DNA from museum specimens. It is, thus, important to consider these factors when selecting the appropriate wet lab techniques to ensure that the highest quality endogenous DNA can be isolated from each specimen.

The next section will highlight these key lab techniques in an ancient DNA workflow—from working with museums to destructively sample tissues to extraction and library preparation to sequencing and bioinformatic analysis. These wet lab and bioinformatic techniques have been optimized to isolate the small and damaged DNA fragments typical of historic and ancient tissues from museums. Before sampling, wet lab, or bioinformatics work takes place, researchers need to define their research question. While we do not discuss this process in detail here, we recommend this module's companion exercise "Designing a conservation genomics project incorporating DNA from museum specimens" (https://doi. org/10.5531/cbc.linc.14.1.9), for more guidance on developing a research question.

A PRACTICAL GUIDE FOR INCORPORATING MUSEUM SAMPLES IN GENETIC STUDIES

Designing the Sampling of Genetic Material According to the Project Needs and Resources Already Available

When designing a conservation genomics (or genetics) study, the number of individuals sampled per population or species can vary based on the specific research question. Biological samples are the main source of genetic material. However, acquiring the necessary number and diversity of samples—whether from wild populations, museum collections, or archaeological sites—can be time-consuming, expensive, or logistically complex. Therefore, before starting the tissue sampling phase in the field or from museum collections, it is crucial to carefully consider the study's goals and constraints. Researchers should first assess 1) the temporal or spatial dimensions of the research question; 2) the statistical requirements of the analysis they plan to conduct, and 3) the taxonomy, ecology, and natural history of the study organism. For example, if the goal is to assess decline in genetic diversity over time, it is important to ensure that enough individuals are sampled from multiple historical periods. If the research focuses on population connectivity across the landscape, the sampling should encompass individuals from geographically distinct locations that reflect the species' range, spanning potential barriers to gene flow.

These considerations are directly linked to the broader types of questions asked in conservation genomics. By leveraging genome-wide data, researchers can explore patterns of genetic diversity, detect inbreeding or local adaptation, and model population structure and gene flow—all of which require thoughtful sampling strategies. In addition, sampling design is key for answering phylogenetic questions central to conservation, such as identifying cryptic species, resolving evolutionary relationships among threatened lineages, or determining the closest relatives of recently extinct taxa. Integrating samples from both modern and historical sources, including museum specimens and ancient DNA, enables scientists to uncover evolutionary patterns that may otherwise remain hidden, ultimately providing deeper insights into species boundaries, adaptive potential, and conservation priorities. Careful planning at the sampling stage lays the foundation for robust genetic inferences, guiding effective conservation decisions and enhancing our understanding of biodiversity across space and time.

So how many samples do you need to answer your question? The figures provided here are general guidelines and should be considered as suggestions rather than rigid rules. The aim here is to help you think critically about the challenges involved in sampling and designing a study that will provide the necessary data to answer key conservation questions.

For population-level studies, such as estimating genetic diversity, detecting inbreeding, or identifying signals of local adaptation, researchers typically aim for 20–30 individuals per population to ensure robust allele frequency estimates (Hale et al., 2012; Nazareno et al., 2017). Interestingly, research has been showing that large sample sizes at low sequencing depth are desirable to achieve high accuracy of estimates of genetic variation (Fumagalli et al., 2013; Lou et al., 2021). Studies of population connectivity or landscape genomics require spatially distributed samples, with 10–20 individuals per site across multiple locations, ideally capturing environmental or habitat variation (Meirmans, 2015; Rellstab et al., 2015). In contrast, phylogenomic studies and questions involving cryptic species or evolutionary relationships often prioritize taxonomic breadth over dense population sampling. These projects may only require 1–5 representative genomes per lineage, especially when high-coverage genomes are available (Funk et al., 2012; Lou et al., 2021). However, for species delimitation or defining evolutionarily significant units, sampling 5–10 individuals per suspected lineage can provide the resolution needed to detect subtle genetic differentiation (Carstens et al., 2013). Finally, studies involving temporal sampling—such as using museum specimens or ancient DNA—must balance historical coverage with the technical limitations of degraded DNA, often working with fewer samples but drawing on a wider time span (Fumagalli, 2013). Aligning sample size and sampling design with the specific research objectives ensures efficient use of resources and maximizes the analytical power of conservation genomic studies (McMahon et al., 2014; Shafer et al., 2015).

Before sampling tissue from live specimens in the field, one should start by checking what has already been sampled and stored long-term in museum biobanks for the target species or populations. These biobanks are collections of biological samples and associated information organized systematically for research purposes. Since the proliferation of molecular biology studies, modern cryogenic collections have also become a common practice among natural history museums. These collections house frozen tissues that can be requested as loans for application in genetic research.

You might also want to check what genetic data has already been generated for the focal species/group and is publicly available for download. There are several online repositories that host this type of digital data, such as the National Center of Biotechnology Information (NCBI), American GenBank, European Nucleotide Archive (ENA), Global Genome Biodiversity Network (GGBN), DNA Data Bank of Japan (DDBJ), Chinese Genbank, and Genomes on a Tree, to name a few. These archives contain different data types (raw reads, mapped reads, assemblies, metadata files) or can point to where data were deposited) at various levels of data processing.

Lastly, zoos, botanical gardens, or captive breeding programs often have individuals of rare or endangered animals and plants. These institutions can provide access for sampling the focal species of interest. Sampling from live or recently deceased focal species in these institutions can be a cost-efficient way to obtain high-quality tissue that can proceed directly to library preparation and sequencing. This strategy can be particularly helpful for generating genetic information at high resolution (e.g., annotated chromosome-level reference genomes) and has been successfully implemented by the DNA Zoo consortium (https://www.dnazoo.org/assemblies).

To plan the sampling of historical specimens, the first step is to check museum collections catalogs. Most natural history collections make their catalog databases available online, allowing anyone to search for specimens and download the data. Downloading individual databases can be cumbersome, so using an aggregator that mirrors multiple collections catalogs is the most efficient way to start by simultaneously searching across multiple collections. Examples include the iDigBio portal (https://portal.idigbio.org/) and GBIF (https://www.gbif.org/). There are also clade-

specific aggregators, such as VertNet for vertebrates (http://portal.vertnet.org/search, no longer updated) and JAQ (https://www.jacq.org/) for herbarium specimens. If you are interested in sampling specimens from a particular collection that does not have an online database, directly contact the curators and collection managers to inquire about the specimens registered in their catalog and the sampling procedure. Establishing a relationship with museum specialists can help you determine the best specimens and type of tissues to sample from their collections. Discussing the scientific objectives of your study with the curators and collection managers is an important step in arranging permits for destructive sampling.

Once you have planned your field, zoo, or museum sampling strategy, you will need to arrange all the required permits to collect and transport tissues, and if relevant, handle or observe vertebrate animals. Be aware that tissue collection regulations vary from country to country, and multiple agencies may control these regulations. A good starting point is to check the The Nagoya Protocol on Access and Benefit-sharing (https://www.cbd.int/abs/default.shtml), as well as the degree of protection of your target species under the Convention on International Trade in Endangered Species of Wild Fauna and Flora (CITES), an international agreement aimed at ensuring that international trade in specimens of wild animals and plants does not threaten their survival. For further information specific to US laws, the U.S. Fish and Wildlife Service can provide more details. Additionally, you should check with your local Institutional Animal Care and Use Committee (IACUC) for specific details about your institution's regulations regarding the usage and handling of vertebrate samples.

Some questions to guide this step:
1. Does the sampling cover the temporal/geographic/taxonomic range required to answer your question?
2. How many samples does your study require to have enough statistical power to answer your question?
3. How much of your budget do you need to allocate for sampling?
4. What permits are required to sample in the area and to transport biological material across borders?

Collecting Tissue Samples from Voucher Specimens for Genetic Analysis

Before sampling a museum specimen, it is always recommended to inspect it for signs of mold (in the specimen, as well as in the box where it is stored or in the liquid media in which it is preserved), lumps of salt, and the level of tissue disintegration, as these characteristics are often predictive of low DNA extraction and sequencing success. These external debris and other microorganisms can interfere with enzymatic reactions, or even decrease the proportion of endogenous DNA obtained in the final extraction. Ideally, genetic material should be obtained from tissue sampled directly from the preserved specimen, but practices can vary depending on the study organism. For instance, for large-bodied specimens, it is preferable to avoid relying on debris or fragmented material disconnected from the voucher specimen. For herbarium specimens, which include envelopes of leaf tissue on the specimen sheets, the best practice is to take from these envelopes rather than break more of the specimen on display. It is important to note also that for some types of specimens, it may not be possible to subsample. For example, studying small insects often requires the entire specimen to be used. Pinned specimens are often difficult to subsample as they are so brittle and easily destroyed. Because voucher specimens are invaluable materials in multiple types of biological research (e.g., comparative anatomy, morphology, systematics, ontogeny), it is also important to avoid destroying non-bilateral structures that are relevant in taxonomic research (e.g., sampling a limb is preferred over sampling the thorax).

These precautions are crucial for preserving the integrity of both historical and ancient specimens,

but ancient samples present their own unique challenges. Ancient samples are most often limited to tissue collected from sturdy structures, such as fibers, dry bone, and teeth that have been preserved over long periods of time, with occasional exceptions of preserved soft tissues found in optimal (cold, dry, and anoxic) conditions. Compared to historical-level samples, extractions from ancient samples typically have much lower DNA concentrations and the average size of DNA fragments acquired is shorter (35–250 bp; Pääbo, 1989; Dabney et al., 2013; Dabney et al., 2019), and contain DNA from endogenous and exogenous sources. Once collected, ancient specimens are also stored in museum collections. Since there is no standard of how ancient-level specimens must be stored among museums worldwide, they may be stored alongside historical specimens.

When collecting tissue samples for genetic studies, it is essential to maintain tidy laboratory conditions to avoid contamination from external sources or cross-contamination between samples. Therefore, sterilizing all sampling tools and surface areas is crucial when sampling specimens in the field and museum collections. Given that historical and ancient samples yield much lower DNA concentrations than modern samples, contamination from modern sources of DNA can severely impact the results of a museomics study. Contamination from modern human DNA can be particularly problematic for studies involving humans and other primates. Due to the high levels of genetic similarity from close phylogenetic relationships, modern DNA can go undetected and bias bioinformatics results. Therefore, working with these samples requires extremely stringent laboratory protocols (Llamas et al., 2017). Appropriate personal protective equipment (PPE) including gloves and face masks may also be important to utilize during both sample collection and processing to protect project staff from any potentially hazardous chemicals (e.g., arsenic or pesticides) that may have been used during specimen preparation and conservation.

Switching pre-sterilized disposable blades after collecting each tissue sample is ideal to avoid cross-contamination between samples. Alternatively, metal forceps can be sterilized by washing with RNAaway or DNAaway, followed by an ethanol wash, or heating in a high-temperature clean flame, such as a Bunsen burner or alcohol burner, and finally setting the tool in a UV crosslinker for upwards of 20–30 minutes. The ethanol wash ensures that the DNAaway/RNAaway does not corrode the metal tool. If using a bunsen burner or alcohol burner, wait until the metal surface cools down before touching the tissue.

Once you have collected the tissue, ensure each sample is identified with a unique label and stored in a correctly labeled tube. Tubes must be labeled on the cap and the side of the vial with a resistant marker (i.e., that can stand ethanol, water, and changes in temperature). These are archival best practices that can help in the long-term preservation of the collected tissue. Besides collecting the tissue, it is also essential to record all metadata associated with the sample. Without this ancillary data, the archived tissue loses its scientific value. This ancillary information includes physical features and measurements, the sex and age of the individual, the date and locality it was collected, the type of environment it was found in, who collected it, where it was deposited, and more detailed information such as the presence of parasites and habitat details. In conservation, accurate specimen data is fundamental when data from historical specimens inform conservation translocation decisions (Verry et al., 2019).

While all samples should be kept away from UV light, other storage requirements depend on the type of sample and the type of genetic material required for the study. For instance, fresh blood, muscle, or liver samples must be stored cold, whereas some buffers enable storage at room temperature (e.g., DNA/RNA Shield). Hair, feathers, and blood stored on specialized paper can also tolerate room

temperature for longer. Note, a recent study (Colella et al., 2020) demonstrated that storing tissue in ethanol can decrease the molecular weight of DNA molecules, which can impact the average fragment length of the DNA molecules and reduce the sample's adequacy for specific analyses (e.g., long-read sequencing). Ensuring that your samples are stored properly for the desired study will minimize DNA degradation and contamination.

Some questions to guide this step:
1. What tools do you need to minimize the chances of external and cross-contamination when sampling biological material?
2. How are you labeling your tubes?
3. What are the most appropriate storage conditions for the type of samples you will be collecting?
4. How are you keeping track of any associated metadata?

Isolating DNA from Museum Specimens

Specimen preparation techniques, most of which were developed centuries ago, aim to maximize the preservation of specimen body parts and their utility for traditional anatomy-based research. However, these techniques significantly damage and impair access to chemical components, including DNA molecules. Until recently, obtaining genetic material from museum specimens was challenging and prone to failure because the remaining DNA is in low quantity, highly fragmented, and often attached to proteins. These difficulties have been overcome by recent developments in DNA extraction protocols and the emergence of sequencing technologies that perform well with short DNA fragments (commonly called next-generation DNA sequencing). While these methods provide opportunities to access novel data types from hard-to-access species, they also raise some concerns. Notably, compared to modern tissue, samples obtained from preserved specimens tend to have lower DNA yields. Therefore, they should be handled with care to minimize external contamination from sources other than the target specimen, such as bacteria, fungi, and even humans (Knapp et al., 2012). Non-modern samples should ideally be processed in laboratories dedicated to historical or ancient DNA work, which are physically isolated from modern DNA facilities that follow stringent disinfection routines and enforce the use of personal protective equipment (PPE, such as double layers of gloves, full-body suit, face mask, hair, and beard net). To control for external contamination, every step of laboratory work (DNA extraction, purification, and library preparation) should include negative controls (i.e., blank samples with no DNA present; Knapp et al., 2012). A common concern, especially among museum collection staff, is that sampling tissue from preserved specimens always involves some extent of destructive sampling. To address these challenges, scientists have developed protocols that maximize DNA yields and quality while reducing contamination and damage to the voucher specimen. Different solutions have been attempted to accommodate the diversity of specimen preparation types, with variable success (see Box 3 for more details).

The choice of lab work protocol to isolate genetic material from biological samples depends on the type of tissue at hand. This is not unique to museum specimens and applies to fresh tissue collected in the field as well. Different tissues and organisms require specific DNA extraction protocols. For instance, plants have a cellular wall around the cell membrane and tend to accumulate secondary metabolites that can interfere with downstream enzymatic reactions in the extraction process. Therefore, cell lysis must be preceded by a mechanical step (e.g., grinding in liquid nitrogen with a sterile mortar and pestle) to break down the cell wall. Because plants have different secondary metabolites that can inhibit some enzymatic reactions, extra purification steps may be necessary to ensure these unwanted molecules are removed from the solution.

Over the past decades, scientists have optimized DNA extraction protocols to maximize the yield of genetic data recovered from preserved specimens. Because the amount of DNA that can be recovered varies among parts of the specimen, preservation techniques, and the specimen's age, it is essential to choose the correct protocol that suits your study. Box 3 provides an overview of some of the main challenges to extracting DNA from the most commonly available preparation types and references to help you prepare for this step of the laboratory work. Given that fresh tissue samples tend to be less degraded, numerous well-established protocols and commercial kits have been developed for different tissue types and organisms. Therefore, we won't cover the laboratory protocols for fresh samples here.

Box 3: Overview of advantages and challenges to isolating DNA from different specimen preparation types

Dry study skins

- Examples: Mammals and birds.
- Advantages: Fast and simple to collect samples. Toe pads typically have high amounts of endogenous DNA.
- Challenges: Skins are often treated with chemicals (e.g., arsenic).
- *Examples of protocols*:
  - McDonough, M. M., Parker, L. D., Rotzel McInerney, N., Campana, M. G., & Maldonado, J. E. (2018). Performance of commonly requested destructive museum samples for mammalian genomic studies. Journal of Mammalogy, 99(4), 789–802. https://doi.org/10.1093/jmammal/gyy080
  - Campos, P. F., & Gilbert, M. T. P. (2019). DNA extraction from keratin and chitin. Ancient DNA: Methods and protocols, 57–63. https://doi.org/10.1007/978-1-4939-9176-1_7
  - Penna, A., Blair, M. E., Lui, H. L., Peters, E., Kistler, L., & Pozzi, L. (2024). Overcoming challenges to extracting and sequencing historical DNA to support primate evolutionary research and conservation, with an application to Galagos. International Journal of Primatology, 45, 1375–1403. https://doi.org/10.1007/s10764-024-00429-3
  - Tsai, W. L. E., Schedl, M. E., Maley, J. M., & McCormack, J. E. (2020). More than skin and bones: Comparing extraction methods and alternative sources of DNA from avian museum specimens. Molecular Ecology Resources, 20(5), 1220–1227. https://doi.org/10.1111/1755-0998.13077

Bone and teeth

- Examples: Vertebrates.
- Advantages: Hard tissue that can be found in most ancient specimens, also skeletons are often preserved in historical vertebrate collections.
- Challenges: It is difficult to reach the organic layer, which usually requires more destructive sampling techniques, such as drilling.
- *Examples of protocols*:
  - Dabney, J., & Meyer, M. (2019). Extraction of highly degraded DNA from ancient bones and teeth. Ancient DNA: methods and protocols, 25–29. https://doi.org/10.1007/978-1-4939-9176-1_4
  - Dehasque, M., Pečnerová, P., Kempe Lagerholm, V., Ersmark, E., Danilov, G. K., Mortensen, P., Vartanyan, S., & Dalén, L. (2022). Development and optimization of a silica column-based extraction protocol for ancient DNA. Genes, 13(4), 687. https://doi.org/10.3390/genes13040687
  - Rohland, N., Glocke, I., Aximu-Petri, A., & Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth, and sediments for high-throughput sequencing. Nature Protocols, 13(11), 2447–2461. https://doi.org/10.1038/s41596-018-0050-5

Fluid-preserved specimens

- Examples: Fish, amphibians, reptiles, aquatic invertebrates, larvae, spiders, scorpions, birds, and bats.
- Advantages: Specimens in high numbers or with lots of tissue. The media can be used as a source of DNA but can have high levels of damage and contamination.
- Challenges: High chances of the specimen being fixed in formalin, which creates crosslinks within the DNA molecule, or between protein and other DNA molecules present in the media. Even samples that are preserved only in ethanol, will yield fragmented DNA.
- *Examples of protocols:*
  - Ruane, S., & Austin, C. C. (2017). Phylogenomics using formalin-fixed and 100+-year-old intractable natural history specimens. Molecular Ecology Resources, 17(5), 1003–1008. https://doi.org/10.1111/1755-0998.12655
  - Straube, N., Lyra, M. L., Paijmans, J. L. A., Preick, M., Basler, N., Penner, J., Rödel, M., Westbury, M. V., Haddad, C. F. B., Barlow, A., & Hofreiter, M. (2021). Successful application of ancient DNA extraction and library construction protocols to museum wet collection specimens. Molecular Ecology Resources, 21(7), 2299–2315. https://doi.org/10.1111/1755-0998.13433

Eggshells

- Examples: Birds and reptiles.
- Advantages: Excellent preservation of DNA and other biomolecules, common specimen type in avian and some other collections (e.g., invertebrates and herpetology). Often found in archaeological and paleontological sites. Noninvasive sampling of wild specimens.
- Challenges: Often delicate in nature and prone to species misassignment (species can have eggs that look extremely similar).
- *Examples of protocols*:
  - Oskam, C. L., Haile, J., McLay, E., Rigby, P., Allentoft, M. E., Olsen, M. E., Bengtsson, C., Miller, G. H., Schwenninger, J., Jacomb, C., Walter, R., Baynes, A., Dortch, J., Parker-Pearson, M., Gilbert, M. T. P., Holdaway, R. N., Willerslev, E., & Bunce, M. (2010). Fossil avian eggshell preserves ancient DNA. Proceedings of the Royal Society B: Biological Sciences, 277(1690), 1991–2000. https://doi.org/10.1098/rspb.2009.2019
  - Grealy, A., Langmore, N. E., Joseph, L., & Holleley, C. E. (2021). Genetic barcoding of museum eggshells improves data integrity of avian biological collections. Scientific Reports, 11(1), 1605. https://doi.org/10.1038/s41598-020-79852-4
  - van der Meij, S. E. T., & Nieman, A. M. (2016). Old and new DNA unweave the phylogenetic position of the eastern Atlantic gall crab Detocarcinus balssi (Monod, 1956) (Decapoda: Cryptochiridae). Journal of Zoological Systematics and Evolutionary Research, 54(3), 189–196. https://doi.org/10.1111/jzs.12130

Pinned dried specimens

- Examples: Non-aquatic insects.
- Advantages: One of the more widespread collection types, the number of specimens can often vastly outnumber other collections, making population-level analysis possible.
- Challenges: Destructive sampling can destroy important morphological structures, given that specimens are often very small and collections can have few collected specimens.
- *Examples of protocols*:
  - Lalonde, M. M., & Marcus, J. M. (2020). How old can we go? Evaluating the age limit for effective DNA recovery from historical insect specimens. Systematic Entomology, 45(3), 505–515. https://doi.org/10.1111/syen.12411
  - Orr, R. J., Sannum, M. M., Boessenkool, S., Di Martino, E., Gordon, D. P., Mello, H. L., Obst, M., Ramsfjell, M. H., Smith, A. M. & Liow, L. H. (2021). A molecular phylogeny of historical and contemporary specimens of an under-

studied micro-invertebrate group. Ecology and Evolution, 11(1), 309–320. https://doi.org/10.1002/ece3.7042
- Campos, P.F., Gilbert, M.T.P. (2019). DNA Extraction from Keratin and Chitin. In: Shapiro, B., Barlow, A., Heintzman, P., Hofreiter, M., Paijmans, J., Soares, A. (eds) Ancient DNA. Methods in Molecular Biology, vol 1963. Humana Press, New York, NY. https://doi.org/10.1007/978-1-4939-9176-1_7

Herbaria specimens

- Examples: Plants, fungi, algae, lichen, moss.
- Advantages: One of the more common and widespread collection types. Fast and simple to collect samples.
- Challenges: Extraction requires some type of mechanical lysis step, and depending on the species, secondary metabolites can interfere with enzymatic reactions.
- *Examples of protocols*:
  - Hart, M. L., Forrest, L. L., Nicholls, J. A., & Kidner, C. A. (2016). Retrieval of hundreds of nuclear loci from herbarium specimens. Taxon, 65(5), 1081–1092. https://doi.org/10.12705/655.9
  - Folk, R. A., Kates, H. R., LaFrance, R., Soltis, D. E., Soltis, P. S., & Guralnick, R. P. (2021). High-throughput methods for efficiently building massive phylogenies from natural history collections. Applications in Plant Sciences, 9(2), e11410. https://doi.org/10.1002/aps3.11410
  - Gutaker, R. M., E. Reiter, A. Furtwängler, V. J. Schuenemann, and H. A. Burbano. (2017). Extraction of ultrashort DNA molecules from herbarium specimens. BioTechniques, 62, 76–79. https://doi.org/10.2144/000114517
  - Løken, S. B., Skrede, I., & Schumacher, T. (2020). The *Helvella corium* species aggregate in Nordic countries–phylogeny and species delimitation. Fungal systematics and evolution, 5(1), 169–186. https://doi.org/10.3114/fuse.2020.05.11
  - Cubero, O. F., Crespo, A. N. A., Fatehi, J., & Bridge, P. D. (1999). DNA extraction and PCR amplification method suitable for fresh, herbarium-stored, lichenized, and other fungi. Plant Systematics and Evolution, 216, 243–249. https://doi.org/10.1007/BF01084401

Some questions to guide this step:

1. What preservation method is more commonly used for the organism you are interested in studying?
2. How would you identify unsuitable specimens for which there might be low chances of obtaining genetic material (e.g., presence of mold, salt crystals, level of decomposition of internal organs)?
3. What protocols best suit your target species given the type of preservation method most commonly found in natural history collections?
4. What success rate do you expect for the DNA extraction step?
5. Are there any toxic chemicals you will be exposed to? How should you protect yourself to minimize contact and exposure to them?
6. How much of each laboratory reagent and supplies will you need to obtain genetic material for all the samples you have collected?

From Genetic Material to Sequences

Once the DNA from biological specimens has been isolated, it is time to determine the order of genetic information contained in these molecules through sequencing. Sequencing involves determining the exact order of nucleotide bases (A, T, C, and G) that compose each DNA molecule. Because genetic sequencing requires highly specialized machines and trained personnel, this step is usually outsourced to sequencing companies. Once the DNA molecules are sequenced, they are stored in a digital file and referred to as "reads." In the field of ancient and historical DNA, the ability to go further back in time has improved thanks to the development of laboratory protocols for isolating DNA from old and preserved specimens, as well as advancements in sequencing technologies.

In the era of massively parallel sequencing, it only takes a few hours to simultaneously sequence

billions of reads from multiple individuals. In the final step of DNA extraction, DNA molecules are in solution. Before sequencing, these molecules must undergo a series of laboratory procedures to ensure that all DNA molecules (1) can be traced back to a particular individual and (2) are ready to be sequenced. These are the two major goals of laboratory work, which together are called "library preparation." This library preparation step precedes sequencing on high-throughput sequencing (or "next-generation sequencing" NGS) platforms.

We have already covered the main differences in the DNA molecules obtained from fresh and preserved tissue. To fully understand the rationale behind some of the library preparation options for ancient and historical samples, it is important to understand some basic prerequisites for DNA sequencing on different platforms. In the following sections, we provide a brief overview of the main sequencing technologies currently used and the types of genetic data they can generate. We then discuss the main differences between library preparation methods developed for historical and ancient samples. This practical guide concludes with an overview of the applications of different sequencing strategies to conservation genetics.

Overview of Sequencing Technologies

Before the broad availability of parallel sequencing technologies, Sanger sequencing was one of the most commonly used methods. Developed in the mid-1970s, Sanger sequencing became popular due to its low error rates and straightforward laboratory and bioinformatics processes (and because it really was the only sequencing technique available at the time). This method continues to be effective for characterizing specific and informative genomic regions, such as individual genes or short sequences. Its well-established reliability and affordability make Sanger ideal for small-scale sequencing endeavors that demand precise and accurate results without needing extensive data. Furthermore, Sanger sequencing eliminates the need for intricate library preparation, providing a simple and direct workflow. The only requirement is the development of primers (short nucleotide sequences that bind to the specific genomic region of interest). These primers, combined with a series of enzymatic reactions, are used to produce several copies of the fragment of interest through a process called PCR amplification. This amplification step remains in more recent sequencing technologies, although it does not require the development of specific primers.

Currently, the most popular high-throughput sequencing technologies are Illumina, PacBio, and Oxford Nanopore Technologies (ONT). These technologies use different mechanics and principles, with each company offering a variety of machine models (or "platforms") that can deliver different amounts of data (number of reads and total bases sequenced) per run. Each technology has a unique set of strengths tailored for specific applications, and they can be combined to maximize analytical power. Here, we will expand on the major differences between these technologies in terms of (1) the maximum fragment length sequenced, (2) error rate, and (3) cost (see Table 3 for a summary.) Consequently, the application of these different technologies can vary greatly depending on the project needs and the characteristics of the DNA molecules obtained at the end of the extraction step (such as fragment length and degree of damage).

Fragment Length

In Illumina sequencing, single-end sequencing reads the DNA fragment from one end to the other, providing a single sequence of nucleotides. Paired-end sequencing reads from both ends of the DNA fragment, allowing scientists to obtain two sequences per fragment, one from each strand. Paired-end reads provide more information about the DNA and are helpful for detecting structural variations or overlapping sequences. This extra data makes paired-end sequencing more accurate and useful

for complex DNA analysis. However, when assembling short reads to determine the entire genome sequence, it is common to end up with gaps and incomplete assemblies, known as draft genomes.

More recent long-read strategies produce reads that are tens of kilobases in length, creating overlaps that allow for the generation of complete genome assemblies. PacBio and Oxford Nanopore Technologies (ONT) are the most commonly used long-read sequencing technologies today. PacBio relies on the binding of a slow polymerase to a single molecule of circularized DNA, which is sequenced in real-time multiple times. In ONT platforms, a helicase opens the double strand of the DNA molecule, forcing one of the strands to pass through a very small pore, so the changes in the ionic charge of the different nucleotides can be detected. While ONT can sequence much longer fragments and much faster than PacBio, its accuracy is lower. The requirement of long fragments makes long-read sequencing unrealistic for most museum specimens, and even in archival cryo-collections DNA may exhibit significant degradation that may hinder long-read sequencing success (Salis et al., 2025). Despite the challenges associated with using these technologies on museum specimens successful attempts have been made in rare cases (Quatela et al., 2023; Bein et al., 2025).

Error Rates and Accuracy

Determining the nucleotide order in DNA molecules is the main goal of sequencing. Yet, how can we be sure if the nucleotides being represented are, in fact, correct? There are two types of accuracy: consensus accuracy (is there enough support to determine which nucleotide is present in this position?) and read accuracy (has the machine detected the correct nucleotide in this position?). Understanding some aspects of the sequencing technologies is important to ensure you can distinguish actual biological information from sequencing errors.

The higher the number of unique molecules sequenced for a given sample, the higher the chances of assembling the entire genetic information present in the organism, which is commonly described as high "library complexity." Here, the term library refers to the entire pool of molecules sequenced (the reads) for each processed sample (the libraries). The number of unique reads that overlap in a given genomic position determines the coverage or "sequencing depth" for that position. Coverage of a genomic region can be increased with high library complexity, a higher number of reads sequenced for each unique sample, and by sequencing both strands of the DNA molecule (paired-end sequencing). The higher the coverage, the higher the confidence in determining the exact nucleotide present in that specific genetic position, a bioinformatic step usually referred to as "consensus calling." Consensus calculation will get more computationally demanding with increase in sequencing effort, but it cannot account for errors emerging from the sequencing machine.

The sequencing technologies available today have different inherent error rates for individual measurements. Furthermore, accuracy differs not only between technologies but also across genomic regions, as some stretches of the genome are inherently more difficult to read (e.g., palindromic sequences, AT- and CG-rich regions, or very repetitive regions such as telomeres and centromeres). In such cases, it doesn't matter how many times you sequence a particular region; if there are systematic errors and the sequencing platforms consistently make the same mistake, it will not be improved with higher coverage or more computationally demanding bioinformatics analysis. This is particularly true if the fragments sequenced are short and map to ambiguous positions in the reference genome (see the glossary and "Sequencing strategy" section for more information on reference genomes), such as repetitive regions. One strategy to overcome this problem is to combine short-read with long-read sequencing, which continuously read long DNA fragments multiple times.

In conclusion, our level of confidence in determining the nucleotide order from our sequencing data comes from both isolating potential sequencing errors that the machine makes and ensuring that the reads used to determine the sequence for that particular position can be confidently placed at a given genomic region. These confidence steps are performed during quality control steps in the bioinformatics pipeline, a series of computer-based command-line steps that precede the genetic analyses.

Cost

There is considerable variation in sequencing costs per country, given the availability of trained personnel and sequencing equipment, and the cost of importing reagents can change dramatically. Therefore, before planning a conservation genetics project, it is vital to first obtain quotes from laboratory and sequencing companies and evaluate the available budget.

A major concern when deciding the sequencing effort (how many reads you need to reach the desired coverage) is whether the library constructed for that particular sample has enough endogenous DNA (number of reads that map to a reference genome) and a sufficient number of unique DNA molecules (library complexity). For taxa with large genomes and samples with sufficient DNA yield, if a library lacks complexity and has a high number of duplicate DNA molecules (clonality), focusing on hybridization capture protocols (see Box 4) can be considered a strategy to reduce the cost of sequencing non-desired or highly duplicated molecules.

The cost of sequencing is rapidly changing with advances in sequencing platforms, especially due to the increase in output data (number of reads) that can be generated in a single sequencing cycle. To fully take advantage of the sequencing outputs, a commonly used strategy to minimize costs is to combine the DNA molecules from multiple individuals into a single sequencing cycle, referred to as multiplexed sequencing. This strategy adds minimal labor and costs during the laboratory and bioinformatics steps. Briefly, it requires the addition of unique identifiers in the library preparation step (see next section), which will later be used to bioinformatically sort out the provenance of each molecule ("de-multiplexing" step in the bioinformatics pipeline).

When preparing a budget, it is also crucial to consider the computational resources necessary for data storage and analysis. This includes evaluating the costs associated with acquiring and maintaining hardware, such as servers and storage devices, as well as the expenses related to software licenses and cloud services. As the number of raw reads increases, the size of the files storing these reads also grows significantly. Additionally, the size of intermediate files generated during data processing can escalate quickly, further increasing storage requirements. The budget should also account for the memory needed to store and run analyses efficiently. This involves ensuring that there is sufficient RAM and processing power to handle large datasets and complex computations. Furthermore, specialized personnel may be required to manage and analyze the data, ensuring that the infrastructure can handle the volume and complexity of the data efficiently.

Some questions to guide this step:

1. Given the quality of the DNA obtained, would you be able to use long-read sequencing technologies?
2. How can you minimize the sequencing costs for your project?
3. How many samples can you combine in a single lane if you are multiplex sequencing? Think about the output of different sequencing platforms and the desired coverage per sample.
4. Is there a reference genome that you can use in your study?
5. What is the genome size of your focal species?

Table 3: Summary of the main differences, advantages, and challenges of the different sequencing technologies most commonly used in conservation genetics studies. Units: bp = base pair, Kbp = kilo-base pair, Mbp = megabase pair. (* "2x" denotes pair-end sequencing)

		High Throughput Sequencing	
	Sanger	**Short-read sequencing (Illumina)**	**Long-read sequencing (PacBio; ONT)**
Fragment length	500–1000 bp	2x150 bp or 2x300 bp (*)	10–25 Kbp (PacBio); 1–3 Mbp (ONT)
Error rate	0.01%	0.1–0.5%	As of 2024, 0.1% for PacBio HiFi sequencing, and errors are a bit higher for ONT
Advantages	More affordable and extremely efficient if you are focusing on a single or few genetic regions of interest. Because it has the lowest error rate, it manages repetitive sequences well, but can be challenging to determine the number of bases in long tails. Very easy to analyze the data, given that there is well-established software and community support.	Extremely cost-effective, especially when sequencing multiple individuals. Sequencing at higher coverage allows high sensitivity and confidence in determining the presence of rare and low-frequency loci. Because it has been used by the community for so long, it has extremely well-established laboratory and bioinformatics pipelines.	Longer fragments have a greater chance of containing enough unique information to anchor them properly (position and direction) in the genome. Can resolve repetitive regions. Some ONT platforms are extremely fast and portable, allowing sequencing in the field.
Challenges and disadvantages	Requires developing multiple primers and performing multiple PCR amplifications. Lowest scalability when compared to high-throughput sequencing strategies. Consequently, has a much higher price per base pair sequenced if you need to sequence a lot.	The fragment size is a limiting factor in assembling complete reference genomes. Given that most historical/ancient molecules have some level of breakage, this is not such an issue. Struggles to determine the precise number of identical consecutive nucleotides.	It requires extremely high-quality DNA (unfragmented), and the cost/sample can be extremely expensive. Pipelines are not well-established yet. Bioinformatics is computationally demanding and can be laborious.

Choosing a Library Preparation Protocol that Suits Your Samples

Compared to modern samples, the DNA molecules recovered from historical and ancient samples tend to be much shorter. Consequently, short-read technologies (e.g., Illumina) are a more appropriate choice than long-read sequencing. Given that Illumina sequencing is the most commonly used technology for short reads, this section will focus on some details of the library preparation protocol options available for sequencing on this platform (see Table 4 for a summary.)

One of the major goals of library preparation is to prepare the DNA molecules to be read by the sequencing machines. In Illumina platforms, this preparation involves adding adaptors to each DNA molecule, allowing them to bind to a glass flow cell where bridging amplification occurs. Most projects maximize the sequencing output of Illumina machines by opting for multiplexed sequencing of libraries from various individuals. In this case, the reads from each individual library must also be associated with its sample identity. To achieve this, a unique combination of oligos (6–8 nucleotide

sequences, sometimes referred to as unique "barcodes") are added to the end of every DNA molecule of each individual. These barcodes are sequenced together with the DNA molecules they are bound to, and thanks to their unique combination of nucleotides, every read from a particular individual can be isolated bioinformatically (in a process called "demultiplexing"). Combining multiple individual libraries in multiplex sequencing is of course constrained by the desired coverage, the size of the genome, DNA extraction yield, and sequencing strategy. For instance, DNA extraction from a single small-sized insect tends to yield much less DNA than that from tissue of a large mammal. In this case, the library preparation must account for the low DNA yield.

When working with historical and ancient samples, the DNA yield, size of fragments, and the level of degradation present in the DNA molecules will determine the best choice for library preparation. Most library preparation methods optimized for historical and ancient samples have modifications that account for the fact that these samples usually result in (1) much lower input DNA concentrations, (2) shorter and damaged fragments, and in many cases, (3) DNA that is either single or double-stranded. To overcome the challenge of dealing with low-input DNA, most modified library preparation protocols include a few rounds of PCR amplification after the library preparation to increase the number of molecules that can be sequenced. However, this strategy comes with the cost of dramatically increasing the degree of clonality in the library, meaning that most DNA molecules are exact copies of other molecules (or duplicates). While this strategy can help you proceed from library preparation to sequencing, it does not increase the power of completing the entire genomic sequence of an organism. For recommendations on the number of cycles, see Meyer & Kircher (2010).

As for fragment size, the DNA molecules found in high-quality, modern samples are almost intact. If researchers seek to sequence them on a short-read sequencing platform, shearing to the ideal fragment size (~150 or 300 bp) by physical fragmentation (via sonication, or sound waves) or enzymatic fragmentation is the first step of the library preparation. Because the DNA recovered from ancient and historical samples is already extremely fragmented (35–250 bp), they do not require fragmentation prior to library preparation. Moreover, the two strands of shorter DNA molecules separate from each other (a process called "denaturation") at lower temperatures than longer DNA molecules. Therefore, the enzymes used in these protocols have activation temperatures different from those of traditional library preparation protocols. Lastly, many DNA fragments can be lost during laboratory preparation. This can be addressed with protocol optimizations such as using lo-bind tubes (which minimize the affinity of DNA molecules to bind to the inside walls of the tubes), minimizing the number of tube changes (i.e., preparing multiple reactions inside the same tube), adapting the purification step to increase the ratio of binding buffer, and using silica columns that retain shorter fragments (e.g., Qiagen QiaQuick and Qiagen MinElute columns).

Unlike DNA recovered from modern tissues, ancient and historical DNA contains a series of damage signatures, such as the deamination of cytosine that is read by the sequencing machine as uracil, and the formation of single-strand overhangs. Some library protocols can remove the most commonly found miscoding damage (i.e., removing the deoxy uracils using uracil-DNA glycosylase (UDG) and endonuclease VIII (Briggs et al., 2009)); while others preserve these damage patterns using full or partial UDG methods (Rohland et al., 2018), which can be extremely useful to authenticate DNA sequences from human remains, for instance. Lastly, some DNA extraction protocols involve a hot-treatment step (e.g., Hot Alkali Treatment, specifically designed to remove blocking lesions from formalin-preserved specimens). This treatment leads to the separation of the two strands of DNA (a process called "denaturation"), so in this case, a single-strand DNA library preparation method may be required. In general, historical samples from other types of preservation methods tend to respond

Table 4: Comparison of double-stranded (ds) and single-stranded (ss) DNA library protocols currently available for Illumina sequencing of historical and ancient libraries.

	dsDNA libraries	ssDNA libraries
Fragment size	40–500 bp	30–120 bp
Initial amount of input DNA	Some protocols can work with low-input	Ideal for very low input
Allows input of denatured DNA?	no	yes
Keep the information in overhangs?	some protocols	yes
Cost and time consuming	moderate	high
Protocol intricacy	moderate	high

well to double-strand DNA library preparations, but more fragmented samples perform better with a single-stranded library preparation protocol.

Here are some of the most commonly used library preparation protocols for ancient and historical samples:

- Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, 2010(6), pdb-prot5448. https://doi.org/10.1101/pdb.prot5448
- Carøe, C., Gopalakrishnan, S., Vinner, L., Mak, S. S., Sinding, M. H. S., Samaniego, J. A., Wales, N., Sicheritz-Pontén, T., & Gilbert, M. T. P. (2018). Single-tube library preparation for degraded DNA. Methods in Ecology and Evolution, 9(2), 410–419. https://doi.org/10.1111/2041-210X.12871
- Gansauge, M. T., & Meyer, M. (2013). Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nature protocols, 8(4), 737–748. https://doi.org/10.1038/nprot.2013.038
- Kapp, J. D., Green, R. E., & Shapiro, B. (2021). A fast and efficient single-stranded genomic library preparation method optimized for ancient DNA. Journal of Heredity, 112(3), 241–249. https://doi.org/10.1093/jhered/esab012

Here are some of the common uracil-DNA glycosylase (UDG) treatment methods:

- Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., & Pääbo, S. (2009). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research, 38(6), e87. https://doi.org/10.1093/nar/gkp1163
- Rohland, N., Harney, E., Mallick, S., Nordenfelt, S., & Reich, D. (2015). Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. Philosophical Transactions of the Royal Society B: Biological Sciences, 370(1660), 20130624. https://doi.org/10.1098/rstb.2013.0624

Some questions to guide this step:

1. What is the average size of fragments obtained after the DNA extraction?
2. After the DNA extraction, are the DNA molecules denatured?
3. Would your project benefit from the information found in the original information from the unpaired nucleotides often found in degraded molecules?

Sequencing Strategy

In any project design, it is vital to ensure that the research question can be answered with the data in hand. When choosing a sequencing strategy for a conservation genomics project, researchers must consider four major factors: (1) genomic resolution, (2) coverage level, (3) number of individuals/samples required to answer the research question, and (4) cost of sequencing. Research projects are heavily constrained by budget limitations, making the cost of sequencing a critical factor that can limit the possibilities of a genomic study. The differences between the main sequencing strategies in terms of genomic resolution achieved, number of individual samples usually employed, and cost of sequencing are summarized in Figure 5.

The highest level of genomic resolution available today is provided by reference genomes. The quality of a reference genome increases with the level of assembly and annotation. A reference genome can be generated using single or multiple tissues from one individual or by combining data from multiple individuals. Most questions addressed using reference genomes focus on understanding gene functions and the evolution of genomic structure at a comparative level, which relies on well-assembled and annotated references. To increase the accuracy of these references

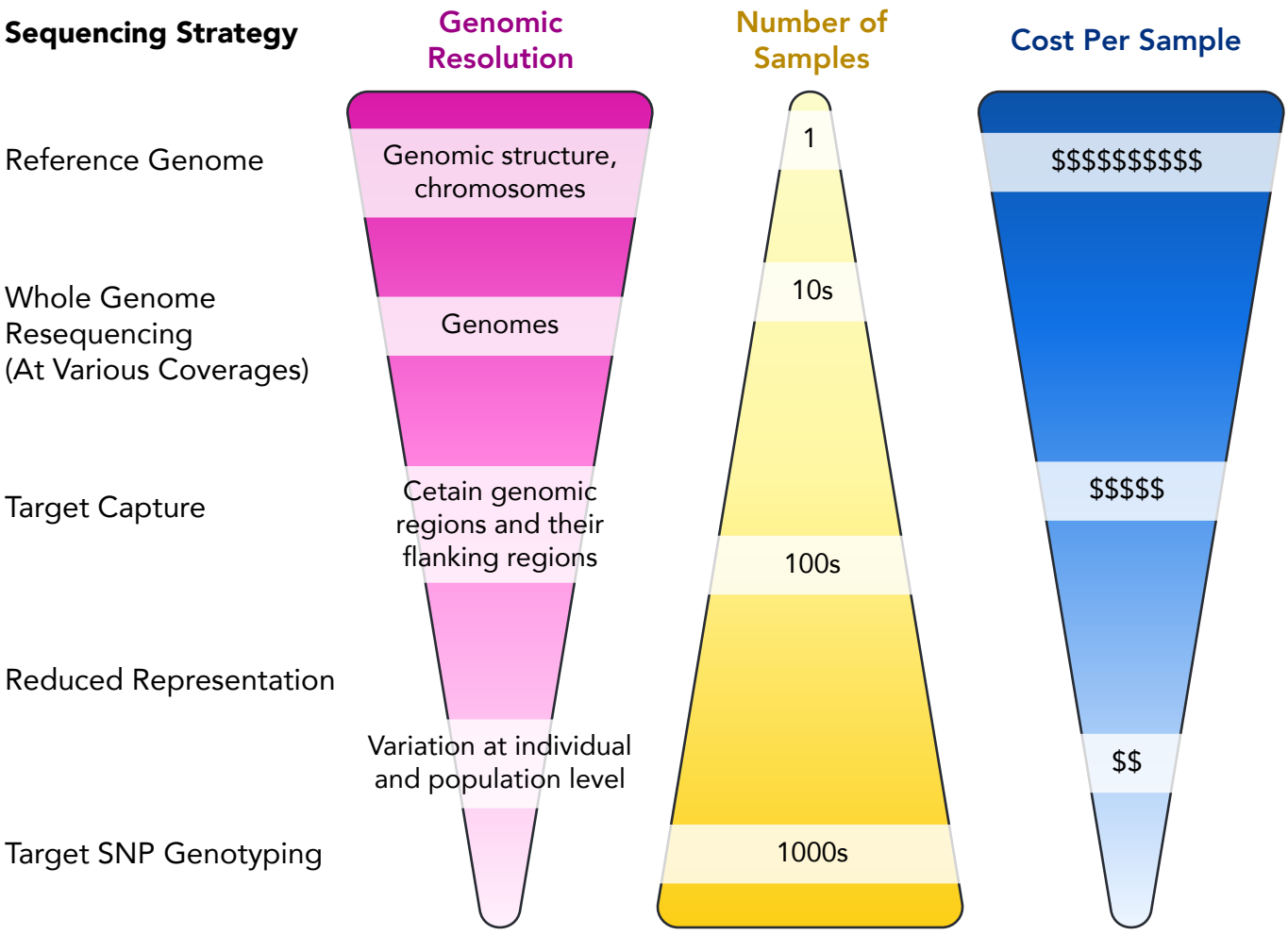


Figure 5. Sequencing strategies can have drastically different costs per sample. This simplified diagram indicates the differences among the sequencing strategies most commonly used in conservation genetic studies when it comes to genomic resolution, number of individual samples, and cost per sample. Image credit: Anna Penna. Figure modified from Threatened Species Initiative Training Module 5.1 (available at https://threatenedspeciesinitiative.com/training-modules/).

(determining sequences, presence of variation, number of copies, relative position, and directionality of challenging genomic regions), researchers usually combine different sequencing strategies (i.e., short and long reads). Consequently, generating high-quality reference genomes can be extremely costly, laborious, and computationally demanding. Very few conservation genetics projects have the necessary funds to generate a well-annotated reference genome. However, some conservation and population genetic analyses indirectly rely on the availability of a good reference genome, given that other sequencing strategies rely on these references for mapping. One way to minimize the costs and labor of generating a reference genome is to incorporate publicly available data, such as the resources available on NCBI and DNAZoo. International collaborative efforts, such as sequencing consortia, are also improving the applications of genetic tools beyond basic research by combining expertise and resources to generate high-quality sequencing data.

As mentioned above, the choice of sequencing strategy depends on the question the project seeks to answer, the required number of samples, and budget availability (Figure 5). The level of DNA degradation in the samples available is another factor that can limit the applicability of some sequencing strategies. For instance, if DNA fragments are too short, restriction-enzyme approaches to generate reduced representations of the genome might result in shorter fragments that won't pass size selection. Even moderately fragmented data can jeopardize the success of long-read sequencing approaches. DNA degradation is strongly related to sample age, but also to the preservation method (what type of chemicals it was exposed, at what buffer and temperature it was stored in, how many times it was thawed, and so on).

In conservation genetics, most questions revolve around the levels of genetic diversity at a population level. Therefore, the required level of genomic resolution is much lower, whereas the number of individuals sequenced needs to be much greater. In this case, approaches such as whole genome resequencing (including at lower coverage), reduced representation or target SNP genotyping are more appropriate. Most reduced representation approaches (e.g., restriction site-associated DNA sequencing and target capture) do not require a reference genome in the bioinformatics pipeline, whereas whole genome resequencing approaches rely on de novo assembly or read mapping to a reference, especially when not sequencing at higher coverage. Therefore, if you opt for whole genome resequencing, it is important to first check the availability of same-species (or phylogenetically close-enough) reference genomes in public repositories (e.g., NCBI).

Another cost-effective approach that remains useful in conservation studies is the sequencing of organellar genomes (mitochondrial DNA or chloroplast DNA) as well as genotyping microsatellites (2-6 base-pair long tandemly repeated DNA sequences found across the genome). These markers can be amplified using traditional PCR methods and are relatively inexpensive compared to large-scale genomic representations. Organelle DNA in particular is found in abundance in the cells, so the protocols require smaller amounts of DNA, making them suitable for studies with limited funding or poor-quality samples. Organelles can be used to identify maternal lineages and infer historical demographic events, whereas microsatellites offer high variability for detecting fine-scale population structure, kinship, and inbreeding. However, both organelle and microsatellites represent only a small fraction of the genome and may not capture genome-wide patterns of variation or selection, which can limit their resolution in some evolutionary or conservation questions. Another disadvantage of microsatellites is that the development of new markers for non-model species can be costly and time-consuming and the amplification success is considerably reduced for degraded material.

Population-level questions (e.g., population structure, gene flow, inbreeding, population size, etc.) often

require good measures of allele frequencies in the populations analyzed. For such a detailed analysis, genetic information from dozens to hundreds of individuals is required. Traditional studies have used a combination of PCR-amplified markers (from nuclear DNA or organelles, such as the mitochondrion and chloroplast) to assess the levels of genetic diversity of the population. With the advent of methods that can track the genealogical history of thousands of loci (e.g., coalescent-based methods), some of the challenges of sequencing numerous individuals can be overcome by using information from thousands of loci from one or a few individuals. Recent and rapid advancements in sequencing strategies have allowed researchers to characterize the frequency and history of genetic diversity by obtaining genealogical histories from more genomic regions by using reduced representation or whole genome resequencing approaches. Conversely, other projects seeking to resolve phylogenetic relationships between species or genera, for example, can use genetic data from much fewer samples (e.g., as little as one sample per taxonomic group). A common strategy in phylogenomic projects that rely on museum specimens to characterize the genetic information of rare or hard-to-sample taxa is to apply hybridization capture to amplify genetic data from conserved regions of the genome and their flanking regions (see Box 4). This capture strategy has been a successful and cost-effective method of answering both population genetic and evolutionary questions.

Box 4 provides a summary of what sequencing strategies of various genomic outputs and resolutions can add to a conservation genomic study, including a few advantages, disadvantages, and the technologies most commonly used.

Box 4: Sequencing strategies commonly used in conservation projects

Reference genome

- A fully assembled genome that can be used as a representation of an individual, population, species, or other taxonomic group. The assembly level of a reference genome can vary a lot depending on the sequencing technology used and the annotation quality. To improve the quality of the reference genome, researchers combine short and long-read sequencing and employ a series of bioinformatics analyses that can predict the function of genomic regions (i.e., functional and structural annotation of assemblies).
- Technologies used: Often requires a combination of short read with deep sequencing (coverage >60x), long-read (PacBio, Hi-C), followed by genome annotation.
- What can be done with it?
  - Identify genes or specific mutations associated with specific traits or diseases.
  - When analyzed at a comparative level (i.e., between reference genomes), it can be useful to determine the genetic basis of certain traits/adaptations.
  - Compare the evolution of genomic structure (e.g., structural variants, number, and direction of gene copies, relative positioning of genes in chromosomes) across the tree of life.
  - A useful resource for reference-based mapping (e.g., whole-genome resequencing)
- Disadvantages:
  - Still extremely expensive and laborious in the laboratory further and bioinformatic analyses.

Whole genome resequencing (WGrS; also known as genome skimming or WGS)

- In this approach, the entire genome of an individual is sequenced at a low (>5X), medium (5–30X), or high coverage (30X or higher), depending on the question and budget available. The reads are then mapped against and compared to sequences of a known reference genome.
- Technologies used: short-read sequencing.
- What can be done with it?

  - Can provide a good resolution of genetic diversity among individuals at various genomic regions.
  - Can be used to obtain SNPs, insertions, deletions, gene rearrangements, and calculate important conservation genetics parameters, such as heterozygosity (genetic diversity), runs of homozygosity (levels of inbreeding), gene flow (connectivity between populations), and genetic load (levels of genetic mutations).
- Disadvantages:
  - Analyzing whole genome data for multiple individuals can be computationally demanding (time of analysis and memory required to store and analyze the data). It can yield many continuous stretches of DNA sequence that are ordered and oriented using paired-end or long-read data, but may still contain gaps. Can cause difficulties and not be cost effective for samples that have low endogenous content of the target species (e.g., samples that are highly contaminated in the case of ancient DNA and historical DNA, or samples where multiple species may be present such as fecal samples and other non-invasive samples), where the majority of sequenced DNA will be from off target species.
- Warning:
  - Assembly can be done *de novo* if the sequences are not fragmented, have extremely high quality, and are sequenced at high coverage. Reads can also be mapped to a reference genome. If the divergence between the reference and the reads is too high (i.e., due to phylogenetic distance, for instance), mapping biases may be present in downstream analyses.

Reduced-representation sequencing approaches

- The sequencing of a subset of the genome by using a combination of restriction enzymes that cut the DNA molecules at specific sequence combinations, followed by a size selection step to get more uniform fragments in the whole genome. Because only digested fragments are sequenced, this approach greatly reduces the complexity of the genome.
- Technologies used: various combinations of restriction site enzymes followed by short-read sequencing of multiple libraries with unique identifiers combined ("multiplexed sequencing") usually in Illumina platforms.
- What can be done with it?
  - A cost-effective way to generate population screens (SNPs) for hundreds of specimens.
  - Can be used to characterize genetic structure, kinship, relatedness and inbreeding levels, and gene flow, all useful in delimiting conservation units and assessing population connectivity across the landscape.
- Disadvantages:
  - Does not sequence variants across the entire genome, and coding regions might be sequenced only partially. Mapping to a reference genome can increase accuracy, but it is also possible to use de novo approaches.
- Warning:
  - Combining reads obtained from different digestion experiments (i.e., that used different enzymes or fragment selection) can lead to excessive missing data that might bias phylogenetic and population clustering results.

Hybridization capture approaches

- The sequencing of a subset of the genome using artificially designed RNA baits. These baits are small oligos that hybridize with specific regions of the genome and are then captured using magnets
- Technologies used: qPCR, bait hybridization followed by capture.

- What can be done with it?
  - Baits can either be designed for a specific project (e.g., target-SNP genotyping) to detect specific genetic variants used in disease screening, detection of pedigrees and hybrids; target specific genetic regions that are transcribed into proteins and some associated regulatory regions (e.g., whole exome sequencing); or target highly conserved regions of the genome that allow comparisons across higher phylogenetic scales (e.g., Ultra Conserved Elements, UCE).
  - Can also be applied in noninvasive and museum samples to amplify specific genomic regions at a desired coverage, reducing the chances of sequencing DNA molecules from exogenous sources.
  - Can be used on highly contaminated samples and samples with multiple species present to specifically target and enrich the DNA of the target species.
- Disadvantages:
  - Requires pre-existing knowledge of the genome to bioinformatically develop the baits.
- Warning:
  - Indirectly, a well-annotated reference genome will be required to use bioinformatics to design the baits. Designing and acquiring the baits can be expensive, which might limit the number of samples sequenced.

Decision Tree to Guide a Museomics Study Design

Now that we covered some of the main steps between the planning of the tissue sampling and the generation of genetic sequences, Figure 6 provides a decision tree that can help the logical reasoning process required to design a museomics study. This decision tree should make the overall planning process more organized and transparent, for instance, when selecting potential source of data and appropriate methods. Use this decision tree to identify potential challenges and alternatives at each step, anticipate problems and plan solutions.

Bioinformatic Considerations

The analysis of sequencing data from ancient and historical samples shares many commonalities with the analysis of data from modern samples, including processing of raw sequencing data (quality control, demultiplexing, adapter trimming, and quality filtering), mapping of reads against reference genomes or databases, post-processing (removal of duplicates, quality filtering of map reads), and variant calling. However, some extra steps and modifications in the bioinformatics pipelines must be included due to challenges arising from the lower sequencing depth, and higher levels of fragmentation, degradation, and contamination found in the sequencing reads obtained from DNA of non-modern samples (Orlando et al., 2021).

In the bioinformatic analysis of ancient and historical DNA, degradation, fragmentation, and contamination are critical factors that must be meticulously addressed. Degradation results in highly fragmented DNA, which can impact the performance of some mapping algorithms. Specialized alignment tools and protocols can more accurately map these shorter reads to the reference. Aligning reads from historical and ancient DNA to a reference genome can introduce reference bias, where reads containing reference alleles are preferentially mapped over those with alternative alleles. These degraded DNA reads also exhibit specific post-mortem damage patterns, such as cytosine deamination, which must be accounted for during analysis. Tools like MapDamage (Jónsson

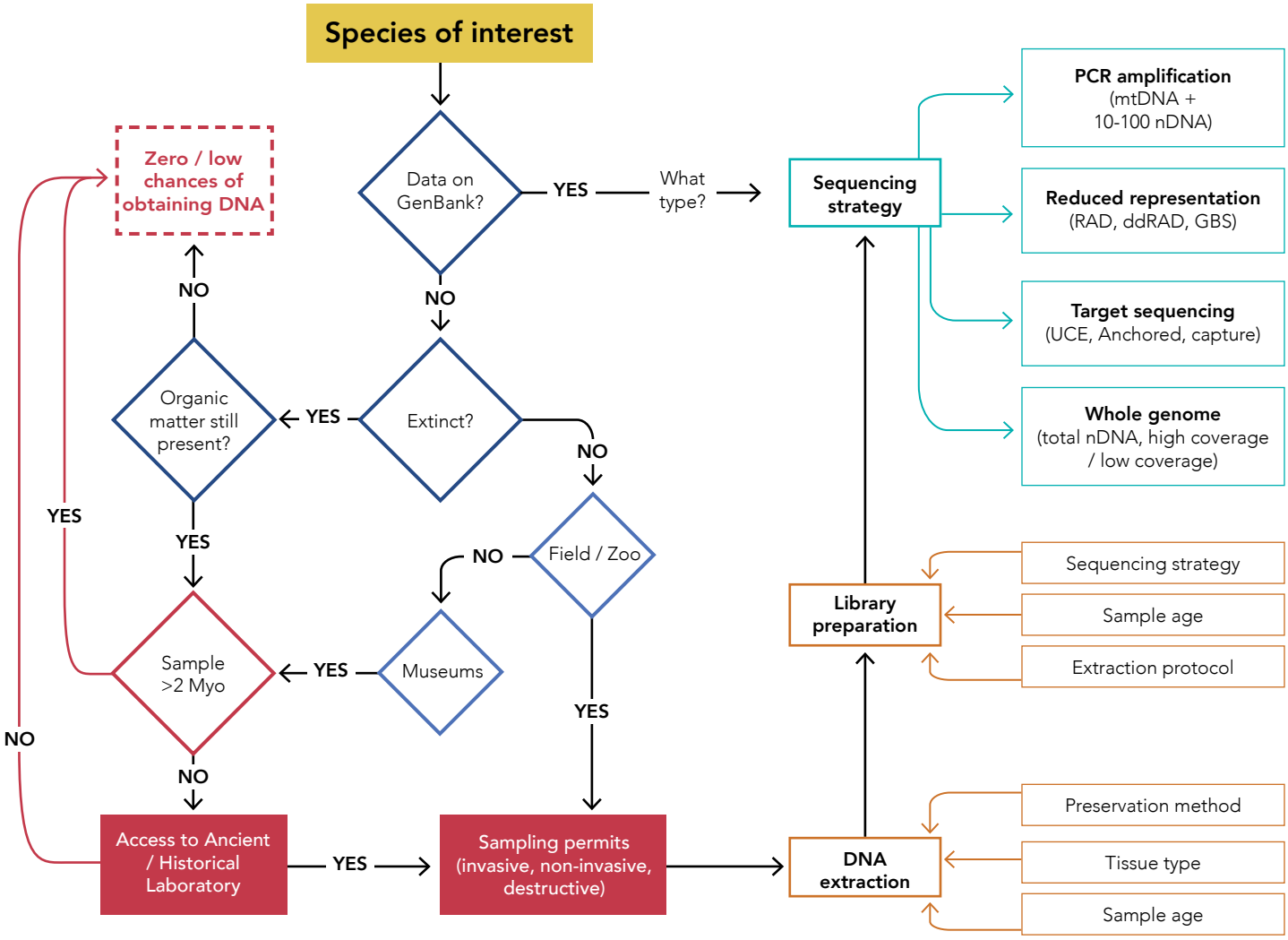


Figure 6. Decision tree to assist in the experimental design of genetic study that incorporates tissue from museum specimens. Image credit: Anna Penna.

et al., 2013) and DamageProfiler (Neukamm et al., 2021) can be used to identify and correct for characteristic post-mortem damage patterns, such as cytosine deamination, which can also serve as a marker to authenticate ancient DNA by distinguishing it from modern contaminants. Contamination from environmental sources and modern DNA requires stringent control measures and computational methods to ensure the authenticity of the ancient DNA. These steps are essential to ensure the reliability and accuracy of genetic data derived from ancient and historical samples, allowing researchers to draw valid conclusions about past populations and species.

The main specialized bioinformatics pipelines currently available for the analysis of genetic data from degraded DNA include EAGER (Peltzer et al., 2016) and PALEOMIX (Schubert et al., 2014), which are specifically tailored for ancient DNA analysis and offer a series of tools to perform the main steps of the bioinformatics data quality control and downstream analysis, including mapping reads to a reference genome, removing contaminants, damage pattern analysis, and various other downstream analyses. Often, the amount of DNA recovered from museum and ancient samples is insufficient to generate high-depth sequencing, either due to the low initial volume of DNA template or due to contamination from exogenous sources. Therefore, the analysis of non-modern samples can also benefit from pipelines designed to maximize the utility of low-coverage, such as ANGSD (Korneliussen et al., 2014) and loco-pipe (Zhou et al., 2024) which incorporate probabilistic frameworks

and best practices for handling genotype uncertainty and other challenges (Lou et al., 2021). These pipelines also have a versatile tool-set for population genetics analysis used to investigate population structure. Another pipeline useful for genetic diversity analysis is GenErode (Kutschera et al., 2022), which is designed to investigate genome erosion in endangered and extinct species and can process reads from modern and ancient samples.

The downstream results generated from these bioinformatics pipelines—such as estimates of genetic diversity, gene flow, and relatedness—are essential for answering key research questions in conservation genomics, evolutionary biology, and phylogenetics. For example, measures of genetic diversity can reveal whether a population is at risk of inbreeding or loss of adaptive potential, while gene flow analyses can help determine whether isolated populations are genetically connected or fragmented. Patterns of relatedness between specimens—especially when integrated with spatial and temporal data—can provide insight into historical population structure, range shifts, or even undocumented extinctions. These insights are critical for understanding how species and populations have responded to past environmental changes, informing conservation strategies for threatened taxa today. Moreover, in phylogenetics and biodiversity studies, historical and ancient genomes can help resolve the evolutionary relationships of extinct or elusive species, refine species boundaries, and provide a deeper temporal perspective on diversification processes.

In Summary

Museomics leverages genetic material from museum specimens, allowing researchers to obtain DNA from a wide range of organisms, including those that are rare, endangered, or extinct. Advances in DNA extraction and sequencing technologies have enabled the recovery of genetic information from historical and ancient samples, despite challenges like low DNA concentrations and fragmentation. By combining different sequencing platforms and samples from modern and historical sources, researchers can obtain valuable insights into the genetic diversity, population structure, and evolutionary history of various species. These insights may then be applied to species conservation by identifying at-risk populations, informing evidence-based management strategies, tracking genetic health, mitigating inbreeding, and enhancing the resilience of species to environmental changes. Consider reading the NCEP Module, "Conservation genetics" (https://doi.org/10.5531/cbc.ncep.0123), for more on the foundational concepts and applications of conservation genetics.

GLOSSARY

- Admixture: The cross between two individuals of a separate species or population whose offspring could result in a new hybrid population. This is not to be confused with introgression/hybridization in which backcrosses among individuals of separate populations are continuous. Admixture, on the other hand, occurs as a one-time event or through multiple generations and increases genetic diversity through the acquisition of new mutations.
- Allele(s): Alternative forms of a gene. These typically refer to changes in protein-coding regions (genes) of the genome that differ and can thus change the phenotype of the individual.
- Carrying capacity: The maximum number of individuals within a population that an environment can sustain.
- Bioinformatics pipeline: A series of computer instructions (scripts) that are deployed to execute several tasks by different programs. In Genomics, these computer programs are specifically designed to implement a set of complex algorithms to process sequence data.
- Chromosome: A single molecule of DNA and associated proteins (e.g., histones that coordinate the condensation of the chromatin, and other proteins responsible for the binding of transcription

factors). Convenient packaging of genetic information. The DNA molecules are really long, so they must be compacted in a way to fit inside the cell. This level of organization is crucial during cell division (ensures the genetic material will be divided in a proper way), and to minimize the exposure of the DNA molecules, unpacking it only in regions needed for cellular processes.

- Coverage: The number of unique reads that overlap in a given genomic position will determine the "coverage" or "sequencing depth" for that position. The higher the coverage, the higher the confidence to determine the exact nucleotide present in that specific genetic position or the "identity" of that position, in a step that is usually referred to as "consensus calling."
- Deleterious mutations/alleles: Alleles, or alternate forms of a gene, that are typically recessive. An accumulation of deleterious alleles in the genome can decrease the fitness of individuals within the population and, thus, can reduce the effective population size within a population or species over time.
- Deamination/deaminated sequences: The loss of an amine group that typically results in the misidentification of a cytosine base as a uracil base. This is a common source of ancient DNA damage that can also be used to authenticate genetic information produced.
- De novo approaches: A technique used to sequence and assemble the complete genome of an organism without the use of a reference genome. This approach is particularly useful for studying species with no prior genomic information, as it reconstructs the genome sequence by iteratively aligning the reads in the dataset to each other.
- Endogenous DNA (Exogenous DNA): The term is used to differentiate the genetic material of the target species (endogenous) from that belonging to parasites, fungi, bacteria, other microorganisms, and contaminant DNA from sources such as researchers and laboratory reagents (exogenous). Determining the identity of the DNA molecules obtained after extraction requires sequencing and bioinformatic analysis.
- Evolutionarily significant units (ESU): A population or group of organisms that is defined as distinct for the purposes of conservation. This term is typically used to more broadly classify a distinct group that cannot be classified as a subspecies or population, but still garners the need for conservation protection.
- Exome: The portion of the genome that consists of the coding regions of genes responsible for producing proteins. Whole exome sequencing is a technique used to sequence all the exons in a genome, allowing the identification of genetic variations that may be linked to rare genetic disorders and understanding the genetic basis of complex traits. It provides an efficient way to focus on the most functionally important parts of the genome.
- Fitness (biological): The ability of an individual (or sometimes in reference to a population or species) to produce viable offspring and thus pass on its genetic material. This can be referred to either with respect to an individual's genotype or phenotype.
- Fixed mutations/alleles: A gene variant that has reached a frequency of 100% in a population has become 'fixed.' Whether the gene variant will be present in all members of the population (again, fixation) or lost entirely, meaning no members of the population possess this allele, depends on how that mutation is selected.
- Gene: A region of the genome that is transcribed into RNA plus the associated regulatory elements required for that transcription. Some examples are protein-coding genes, regulatory genes, tRNA, mRNA, or rRNA genes.
- Genetic load: A decrease in the fitness of individuals in a gene pool on average due to the accumulation of deleterious alleles.
- Genetic rescue: A conservation strategy aimed at increasing the genetic diversity within a population by introducing new genetic variation (individuals with unique genomes).
- Genetic differentiation/genetic distance: A measure of the difference in the frequency of alleles

between different populations due to structure. Often measured using the fixation index (Fst), which ranges from 0, which means complete sharing of genetic material between populations, and 1, no sharing of genetic material.

- Genome: One complete set of chromosomes (sex chromosomes and autosomes), including the sequences of all genes, introns (non-coding), exons (coding regions), and organelle genetic material (mitochondria and chloroplasts). In humans, the entire genome has around 3.3 billion base pairs (bp), and the mitochondria has around 16 thousand bp.
- Genome assembly: The process of putting together the entire order of nucleotides that constitute the genome of a given individual, Ideally, all the reads that were sequenced but do not belong to the organism's own genome should be discarded. This process can be done using a reference genome, though mapping the reads sequenced against a reference sequence (see below); or de novo, in which the entire genome is assembled by finding overlaps between the multiple reads sequenced. De novo approaches are good to detect structural variants, but require very high-quality raw read data and can be very computationally demanding. The level of assembly of the raw sequences varies from contigs (continuous overlapping regions) to chromosomes (when the reads can be spatially attributed to different DNA molecules); and can include annotation information (see below).
- Genome annotation: Obtaining quantitative and qualitative information from the whole genome sequenced. Some of the quantitative metrics include the proportion of GC content, the contiguity between fragments, the average coverage, number of genes or gaps identified. The qualitative information are specificities about what the genomic regions represent (genes, regulatory regions, telomeres, repetitive elements, and so on).
- Genomic/genetic erosion: The collective damage to a species' genome or gene pool due to lack of genetic diversity that are characterized by an increase in maladaptation, genetic introgression, and/or increased genetic load.
- Heterozygosity: Likelihood that a pair of randomly selected alleles will be different (i.e., the proportion of individuals that are heterozygotes in a population at a particular gene or region of the genome.)
- Inbreeding depression: A reduction in the biological fitness of an organism due to an accumulation of (mostly recessive) deleterious mutations that results from the offspring of parents that are too closely related to one another. This term is commonly used in reference to a population or species in which there are so few total individuals in which inbreeding and, thus, loss of genetic diversity presents itself. The term can be used in contrast with outbreeding depression, which is a reduction in biological fitness when individuals from separate species, strains of locally adapted populations cross.
- Introgression/introgressive hybridization: Typically referred to as introgression, introgressive hybridization, or hybridization. This refers to backcrosses between populations that are distinct but gene flow between populations is present.
- Mapping (reads): A bioinformatics step that consists of aligning short DNA sequences (called "reads") obtained from sequencing to a known reference genome. This step helps identify the placement of each read in the genome, allowing researchers to detect variations like mutations or structural differences.
- Microsatellites: Repetitive regions throughout the genome that are well-characterized and used to study the genetic variation of closely related species.
- Multiplexed sequencing: When DNA libraries from multiple samples are combined and sequenced together in a single run, using unique molecular "barcodes" to distinguish between the samples. This approach saves time and costs, allowing researchers to analyze many samples simultaneously without losing track of which data belongs to which sample.

- Nucleotide diversity: Average number of pairwise differences between all different pairs of DNA sequences in the population or sample.
- Reference genome: An entire assembled genome that can be used as a representation of the genome of an individual or species. The references can have different levels of assembly and annotation quality. Used to map raw reads against.
- Shearing (DNA): Breaking down long strands of DNA into smaller fragments. This step is necessary for short-fragment sequencing technologies (e.g., Illumina) of fresh tissue samples. Often, DNA shearing is done by sonication, a method that uses high-frequency sound waves to create vibrations that disrupt the DNA molecules, resulting in random shearing into fragments of varying sizes. This step must be completed before library preparation.
- Size selection: A lab method in which a specific DNA size fragment (e.g., 250 bp) is isolated from genomic DNA and other unwanted fragment sizes. This is typically performed as a part of library preparation, a process by which samples are fragmented and uniquely "labeled" in order to be successfully processed by NGS technologies and high throughput sequencing. Size selection is typically achieved using magnetic beads, instruments like the BluePippin, and gel extraction.
- SNPs: SNPs (or single nucleotide polymorphisms) are variations among individuals in a single nucleotide at a specific position within a genome. These are used to study phylogenetic relationships between individuals of different populations or species.
- Target SNP genotyping: A technique used to identify and analyze specific single nucleotide polymorphisms (SNPs) in a genome. Instead of sequencing the entire genome, this method focuses on particular SNPs of interest, which are often associated with traits, diseases, or evolutionary markers, allowing for a more cost-effective and efficient genetic analysis.
- Ultra Conserved Elements (UCE): Regions of the genome that show little to no variation across different species. UCEs are often found in noncoding regions of the genome but can play important roles in gene regulation and development. Although their exact role is not fully understood, these elements and their flanking regions (which are more variable) are useful for studying evolutionary relationships and tracing the genetic history of organisms.
- Whole genome sequencing: The process of generating the sequences (or raw reads) that will be used to assemble a given genome. Through the combination of different sequencing technologies (short and long reads), it is possible to better characterize the different parts of the genome. The higher the uniqueness of the DNA molecules and the number of reads used to assemble the genome, the higher the confidence in determining the sequence and relative positioning of the DNA molecules that compose the genome.
- Whole genome resequencing: When the entire genome of an organism is sequenced, typically to identify genetic variations such as single nucleotide polymorphisms (SNPs), insertions, deletions, or structural variants. Unlike de novo sequencing, where the genome is sequenced for the first time, resequencing compares the sequenced genome to a reference genome to detect differences.

BIBLIOGRAPHIC RESOURCES

The following readings provide more background on the field of museomics and applications of genetic tools to conservation.

General Introduction to Museomics

- Barnett, R. (2018). Zooarchaeology and ancient DNA. In The Encyclopedia of Archaeological Sciences. https://doi.org/10.1002/9781119188230.saseas0623
- Blair, M. E. (2024). Conservation museomics. Conservation Biology, 38(3) e14234. https://doi.org/10.1111/cobi.14234
- Buerki, S., & Baker, W. J. (2016). Collections-based research in the genomic era. Biological Journal of the Linnean Society, 117(1), 5–10. https://doi.org/10.1111/bij.12721

- Card, D. C., Shapiro, B., Giribet, G., Moritz, C., & Edwards, S. V. (2021). Museum genomics. Annual Review of Genetics, 55(1), 633–659. https://doi.org/10.1146/annurev-genet-071719-020506
- Nakahama, N. (2021). Museum specimens: An overlooked and valuable material for conservation genetics. Ecological Research, 36(1), 13–23. https://doi.org/10.1111/1440-1703.12181
- Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: Advances and challenges in museomics. Trends in Ecology & Evolution, 36(11), 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

Applications of Genetics and Genomics to Conservation

- Hogg, C. J. (2024). Translating genomic advances into biodiversity conservation. Nature Reviews Genetics, 25(5), 362–373. https://doi.org/10.1038/s41576-023-00671-0
- McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? Evolutionary applications, 7(9), 999–1007. https://doi.org/10.1111/eva.12193
- Supple, M. A., & Shapiro, B. (2018). Conservation of biodiversity in the genomics era. Genome Biology, 19, 1–12. https://doi.org/10.1186/s13059-018-1520-3
- Theissinger, K., Fernandes, C., Formenti, G., Bista, I., Berg, P. R., Bleidorn, C., Bombarely, A., Crottini, A., Gallo, G. R., Godoy, J. A., Jentoft, S., Malukeiwickz, J., Mouton, A., Oomen, R. A., Paez, S., Palsbøll, P. J., Pampoulie, C., Ruiz-López, M. J., Secomandi, S., … & Zammit, G. (2023). How genomics can help biodiversity conservation. Trends in Genetics, 39(7), 545–559. https://doi.org/10.1016/j.tig.2023.01.005
- Willi, Y., Kristensen, T. N., Sgrò, C. M., Weeks, A. R., Ørsted, M., & Hoffmann, A. A. (2022). Conservation genetics as a management tool: The five best-supported paradigms to assist the management of threatened species. Proceedings of the National Academy of Sciences, 119(1), e2105076119. https://doi.org/10.1073/pnas.2105076119

Classic Examples Studies Using a Museomics Approach

- Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. Nature, 312(5991), 282–284. https://doi.org/10.1038/312282a0
- Kjaer, K. H., Pedersen, M. W., De Sanctis, B., De Cahsan, B., Korneliussen, T. S., Michelsen, C. S., Sand, K. K., Jelavic, S., Ruter, A. H., Schmidt, A. M. A., Kjeldsen, K. K., Tesakov, A. S., Snowball, I., Gosse, J. C., Alsos, I. G., Wang, Y., Dockter, C., Rasmussen, M., Jorgensen, M. E., … Willerslev, E. (2022). A 2-million-year-old ecosystem in Greenland uncovered by environmental DNA. Nature (London), 612(7939), 283–291. https://doi.org/10.1038/s41586-022-05453-y
- Mármol-Sánchez, E., Fromm, B., Oskolkov, N., Pochon, Z., Kalogeropoulos, P., Eriksson, E., Biryukova, I., Sekar, V., Ersmark, E., Andersson, B., Dalén, L., & Friedländer, M. R. (2023). Historical RNA expression profiles from the extinct Tasmanian tiger. Genome Research, 33(8), 1299–1316. https://doi.org/10.1101/gr.277663.123
- Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., Schubert, M., Cappellini, E., Petersen, B., Moltke, I., Johnson, P. L. F., Fumagalli, M., Vilstrup, J. T., Raghavan, M., Korneliussen, T., Malaspinas, A.-S., Vogt, J., Szklarczyk, D., Kelstrup, C. D., … Willerslev, E. (2013). Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. Nature, 499(7456), 74–78. https://doi.org/10.1038/nature12323
- van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., … Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. Nature, 591(7849), 265–269. https://doi.org/10.1038/s41586-021-03224-9

ACKNOWLEDGMENTS

We would like to thank Mckenna Santiago Coyle, Melina Giakoumis, and Stephen Gaughran. A.P. was funded by the American Society of Mammalogy (ASM), Society for Systematic Biology (SSB), Systematics Association, Linnean Society of London, and National Science Foundation of the United States (NSF; Award No. BSC-2120691); A.P. was also funded by the Smithsonian Institution Graham Funds in Human Origins and the University of Texas at San Antonio (UTSA). A.T.S was funded by the U.S. National Science Foundation (DBI 2029955). L.T.C. was funded, in part, by the U.S. National Science Foundation (DBI 2029955). M.E.B. and S.K.M. were funded by the National Science Foundation of the United States (Awards No. BCS-1926105/1926215). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

REFERENCES

Amato, G., DeSalle, R., Ryder, O. A., & Rosenbaum, H. C. (Eds.). (2009). Conservation genetics in the age of genomics. Columbia University Press. https://doi.org/10.7312/amat12832

Arengo, F., Porzecanski, A. L., Blair, M. E., Amato, G., Filardi, C., & Sterling, E. J. (2018). The future of natural history museums. In E. Dorfman (Ed.), The future of natural history museums (1st ed., pp. 82–100). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315531892-6/essential-role-museums-biodiversity-conservation-felicity-arengo-ana-porzecanski-mary-blair-george-amato-christopher-filardi-eleanor-sterling

Bein, B., Chrysostomakis, I., Arantes, L. S., Brown, T., Gerheim, C., Schell, T., Schneider, C., Leushkin, E., Chen, Z., Sigwart, J., Gonzalez, V., Wong, N. L. W. S., Santos, F. R., Blom, M. P. K., Mayer, F., Mazzoni, C. J., Böhne, A., Winkler, S., Greve, C., & Hiller, M. (2025). Long-read sequencing and genome assembly of natural history collection samples and challenging specimens. Genome Biology, 26(1), 25. https://doi.org/10.1186/s13059-025-03487-9

Billerman, S. M., & Walsh, J. (2019). Historical DNA as a tool to address key questions in avian biology and evolution: A review of methods, challenges, applications, and future directions. Molecular Ecology Resources, 19(5), 1115–1130. https://doi.org/10.1111/1755-0998.13066

Blair, M., Cao, G., López-Nandam, E., Veronese-Paniagua, D., Birchette, M., Kenyon, M., Md-Zain, B., Munds, R., Nekaris, K., Nijman, V., Roos, C., Thach, H., Sterling, E., & Le, M. (2023). Molecular phylogenetic relationships and unveiling novel genetic diversity among slow and pygmy lorises, including resurrection of *Xanthonycticebus intermedius*. Genes, 14(3), 643. https://doi.org/10.3390/genes14030643

Blair, M. E. (2024). Conservation museomics. Conservation Biology, 38(3) e14234. https://doi.org/10.1111/cobi.14234

Brasil, S. N. R., Kelemen, E. P., & Rehan, S. M. (2023). Historic DNA uncovers genetic effects of climate change and landscape alteration in two wild bee species. Conservation Genetics, 24(1), 85–98. https://doi.org/10.1007/s10592-022-01488-w

Briggs, A. W., Stenzel, U., Meyer, M., Krause, J., Kircher, M., & Pääbo, S. (2009). Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Research, 38(6), e87. https://doi.org/10.1093/nar/gkp1163

Byrne, A. Q. (2023). Reimagining the future of natural history museums with compassionate collection. PLOS Biology, 21(5), e3002101. https://doi.org/10.1371/journal.pbio.3002101

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., & Hudson, M. (2020). The CARE Principles for indigenous data governance. Data Science Journal, 19, 43. https://doi.org/10.5334/dsj-2020-043

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., & Stall, S. (2021). Operationalizing the CARE and FAIR Principles for Indigenous data futures. Scientific Data, 8(1), 108. https://doi.org/10.1038/s41597-021-00892-0

Carstens, B. C., Pelletier, T. A., Reid, N. M., & Satler, J. D. (2013). How to fail at species delimitation. Molecular Ecology, 22(17), 4369–4383. https://doi.org/10.1111/mec.12413

Ceballos, G., Ehrlich, P. R., Barnosky, A. D., García, A., Pringle, R. M., & Palmer, T. M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. Science Advances, 1(5), e1400253. https://doi.org/10.1126/sciadv.1400253

Chen, N., & Nedoluzhko, A. (2023). Ancient DNA: The past for the future. BMC Genomics, 24(1), 309. https://doi.org/10.1186/s12864-023-09396-0

Colella, J. P., Tigano, A., & MacManes, M. D. (2020). A linked-read approach to museomics: Higher quality de novo genome assemblies from degraded tissues. Molecular Ecology Resources, 20(4), 856–870. https://doi.org/10.1111/1755-0998.13155

Collar, N. J. (2000). Collecting and conservation: Cause and effect. Bird Conservation International, 10(1), 1–15. https://doi.org/10.1017/S0959270900000010

Dabney, J., Meyer, M., & Paabo, S. (2013). Ancient DNA damage. Cold Spring Harbor Perspectives in Biology, 5(7), a012567. https://doi.org/10.1101/cshperspect.a012567

Dabney, J., & Meyer, M. (2019). Extraction of highly degraded DNA from ancient bones and teeth. Ancient DNA: Methods and Protocols, 25–29. https://doi.org/10.1007/978-1-4939-9176-1_4

Dahn, H. A., Mountcastle, J., Balacco, J., Winkler, S., Bista, I., Schmitt, A. D., Pettersson, O. V., Formenti, G., Oliver, K., Smith, M., Tan, W., Kraus, A., Mac, S., Komoroske, L. M., Lama, T., Crawford, A. J., Murphy, R. W., Brown, S., Scott, A. F., … Fedrigo, O. (2022). Benchmarking ultra-high molecular weight DNA preservation methods for long-read and long-range sequencing. GigaScience, 11. https://doi.org/10.1093/gigascience/giac068

De Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R., & Pimm, S. L. (2015). Estimating the normal background rate of species extinction. Conservation Biology, 29(2), 452–462. https://doi.org/10.1111/cobi.12380

Fanning, T. G., Slemons, R. D., Reid, A. H., Janczewski, T. A., Dean, J., & Taubenberger, J. K. (2002). 1917 avian influenza virus sequences suggest that the 1918 pandemic virus did not acquire Its hemagglutinin directly from birds. Journal of Virology, 76(15), 7860–7862. https://doi.org/10.1128/JVI.76.15.7860-7862.2002

Feigin, C. Y., Newton, A. H., Doronina, L., Schmitz, J., Hipsley, C. A., Mitchell, K. J., Gower, G., Llamas, B., Soubrier, J., Heider, T. N., Menzies, B. R., Cooper, A., O'Neill, R. J., & Pask, A. J. (2017). Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. Nature Ecology & Evolution, 2(1), 182–192. https://doi.org/10.1038/s41559-017-0417-y

Feigin, C., Frankenberg, S., & Pask, A. (2022). A chromosome-scale hybrid genome assembly of the extinct Tasmanian tiger (*Thylacinus cynocephalus*). Genome Biology and Evolution, 14(4), evac048. https://doi.org/10.1093/gbe/evac048

Fong, J. J., Blom, M. P. K., Aowphol, A., McGuire, J. A., Sutcharit, C., & Soltis, P. S. (2023). Editorial: Recent advances in museomics: revolutionizing biodiversity research. Frontiers in Ecology and Evolution, 11. https://doi.org/10.3389/fevo.2023.1188172

Fumagalli, M. (2013). Assessing the effect of sequencing depth and sample size in population genetics inferences. PLOS ONE, 8(11), e79667. https://doi.org/10.1371/journal.pone.0079667

Funk, W. C., McKay, J. K., Hohenlohe, P. A., & Allendorf, F. W. (2012). Harnessing genomics for delineating conservation units. Trends in Ecology & Evolution, 27(9), 489–496. https://doi.org/10.1016/j.tree.2012.05.012

Hahn, E. E., Alexander, M. R., Grealy, A., Stiller, J., Gardiner, D. M., & Holleley, C. E. (2022). Unlocking inaccessible historical genomes preserved in formalin. Molecular Ecology Resources, 22(6), 2130–2147. https://doi.org/10.1111/1755-0998.13505

Hale, M. L., Burg, T. M., & Steeves, T. E. (2012). Sampling for microsatellite-based population genetic studies: 25 to 30 individuals per population is enough to accurately estimate allele frequencies. PLOS ONE, 7(9), e45170. https://doi.org/10.1371/journal.pone.0045170

Hempel, E., Bibi, F., Faith, J. T., Brink, J. S., Kalthoff, D. C., Kamminga, P., Paijmans, J. L. A., Westbury, M. V., Hofreiter, M., & Zachos, F. E. (2021). Identifying the true number of specimens of the extinct blue antelope (*Hippotragus leucophaeus*). Scientific Reports, 11(1), 2100. https://doi.org/10.1038/s41598-020-80142-2

Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A., & Wilson, A. C. (1984). DNA sequences from the quagga, an extinct member of the horse family. Nature, 312(5991), 282–284. https://doi.org/10.1038/312282a0

International Council of Museums. (2013). ICOM code of ethics for natural history museums. ICOM. https://icom.museum/wp-content/uploads/2018/07/nathcode_ethics_en.pdf

Islam, S., Peart, C., Kehlmaier, C., Sun, Y.-H., Lei, F., Dahl, A., Klemroth, S., Alexopoulou, D., Del Mar Delgado, M., Laiolo, P., Carlos Illera, J., Dirren, S., Hille, S., Lkhagvasuren, D., Töpfer, T., Kaiser, M., Gebauer, A., Martens, J., Paetzold, C., & Päckert, M. (2024). Museomics help resolving the phylogeny of snowfinches (Aves, Passeridae, *Montifringilla* and allies). Molecular Phylogenetics and Evolution, 198, 108135. https://doi.org/10.1016/j.ympev.2024.108135

Jensen, E. L., Díez-del-Molino, D., Gilbert, M. T. P., Bertola, L. D., Borges, F., Cubric-Curik, V., De Navascués, M., Frandsen, P., Heuertz, M., Hvilsom, C., Jiménez-Mena, B., Miettinen, A., Moest, M., Pečnerová, P., Barnes, I., & Vernesi, C. (2022). Ancient and historical DNA in conservation policy. Trends in Ecology & Evolution, 37(5), 420–429. https://doi.org/10.1016/j.tree.2021.12.010

Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F., & Orlando, L. (2013). mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics, 29(13), 1682–1684. https://doi.org/10.1093/bioinformatics/btt193

Knapp, M., Clarke, A. C., Horsburgh, K. A., & Matisoo-Smith, E. A. (2012). Setting the stage – Building and working in an ancient DNA laboratory. Annals of Anatomy - Anatomischer Anzeiger, 194(1), 3–6. https://doi.org/10.1016/j.aanat.2011.03.008

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: Analysis of next generation sequencing data. BMC Bioinformatics, 15(1), 356. https://doi.org/10.1186/s12859-014-0356-4

Kreps, C. (2008). Indigenous curation, museums, and intangible cultural heritage. In L. Smith & N. Akagawa (Eds.), Intangible heritage (1st ed., p. 16). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9780203884973-16/indigenous-curation-museums-intangible-cultural-heritage-christina-kreps

Kutschera, V. E., Kierczak, M., Van Der Valk, T., Von Seth, J., Dussex, N., Lord, E., Dehasque, M., Stanton, D. W. G., Khoonsari, P. E., Nystedt, B., Dalén, L., & Díez-del-Molino, D. (2022). GenErode: A bioinformatics pipeline to investigate genome erosion in endangered and extinct species. BMC Bioinformatics, 23(1), 228. https://doi.org/10.1186/s12859-022-04757-0

Leigh, D. M., Hendry, A. P., Vázquez-Domínguez, E., & Friesen, V. L. (2019). Estimated six per cent loss of genetic variation in wild populations since the industrial revolution. Evolutionary Applications, 12(8), 1505–1512. https://doi.org/10.1111/eva.12810

Leonard, J. A. (2008). Ancient DNA applications for wildlife conservation. Molecular Ecology, 17(19), 4186–4196. https://doi.org/10.1111/j.1365-294X.2008.03891.x

Llamas, B., Valverde, G., Fehren-Schmitz, L., Weyrich, L. S., Cooper, A., & Haak, W. (2017). From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR: Science & Technology of Archaeological Research, 3(1), 1–14. https://doi.org/10.1080/20548923.2016.1258824

Locatelli, N. S., McIntyre, P. B., Therkildsen, N. O., & Baetscher, D. S. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. Proceedings of the National Academy of Sciences, 117(51), 32211–32212. https://doi.org/10.1073/pnas.2007421117

Lou, R. N., Jacobs, A., Wilder, A. P., & Therkildsen, N. O. (2021). A beginner's guide to low-coverage whole genome sequencing for population genomics. Molecular Ecology, 30(23), 5966–5993. https://doi.org/10.1111/mec.16077

Mace, G. M. (2004). The role of taxonomy in species conservation. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 359(1444), 711–719. https://doi.org/10.1098/rstb.2003.1454

Mármol-Sánchez, E., Fromm, B., Oskolkov, N., Pochon, Z., Kalogeropoulos, P., Eriksson, E., Biryukova, I., Sekar, V., Ersmark, E., Andersson, B., Dalén, L., & Friedländer, M. R. (2023). Historical RNA expression profiles from the extinct Tasmanian tiger. Genome Research, 33(8), 1299–1316. https://doi.org/10.1101/gr.277663.123

Marshall, W. F., Telford, S. R., Rys, P. N., Rutledge, B. J., Mathiesen, D., Malawista, S. E., Spielman, A., & Persing, D. H. (1994). Detection of *Borrelia burgdorferi* DNA in museum specimens of *Peromyscus leucopus*. Journal of Infectious Diseases, 170(4), 1027–1032. https://doi.org/10.1093/infdis/170.4.1027

McMahon, B. J., Teeling, E. C., & Höglund, J. (2014). How and why should we implement genomics into conservation? Evolutionary applications, 7(9), 999–1007. https://doi.org/10.1111/eva.12193

Meirmans, P. G. (2015). Seven common mistakes in population genetics and how to avoid them. Molecular Ecology, 24(13), 3223–3231. https://doi.org/10.1111/mec.13243

Meyer, M., & Kircher, M. (2010). Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protocols, 2010(6), pdb.prot5448. https://doi.org/10.1101/pdb.prot5448

Minteer, B. A., Collins, J. P., Love, K. E., & Puschendorf, R. (2014). Avoiding (re)extinction. Science, 344(6181), 260–261. https://doi.org/10.1126/science.1250953

Mitchell, K. J., & Rawlence, N. J. (2021). Examining natural history through the lens of palaeogenomics. Trends in Ecology & Evolution, 36(3), 258–267. https://doi.org/10.1016/j.tree.2020.10.005

National Academies of Sciences, Engineering, and Medicine. (2020). Biological collections: Ensuring critical research and education for the 21st century (p. 25592). National Academies Press. https://doi.org/10.17226/25592

Nazareno, A. G., Bemmels, J. B., Dick, C. W., & Lohmann, L. G. (2017). Minimum sample sizes for population genomics: An empirical study from an Amazonian plant species. Molecular Ecology Resources, 17(6), 1136–1147. https://doi.org/10.1111/1755-0998.12654

Neukamm, J., Peltzer, A., & Nieselt, K. (2021). DamageProfiler: Fast damage pattern calculation for ancient DNA. Bioinformatics, 37(20), 3652–3653. https://doi.org/10.1093/bioinformatics/btab190

Ochoa, A., Onorato, D. P., Roelke-Parker, M. E., Culver, M., & Fitak, R. R. (2022). Give and take: Effects of genetic admixture on mutation load in endangered Florida panthers. Journal of Heredity, 113(5), 491–499. https://doi.org/10.1093/jhered/esac037

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., & Warinner, C. (2021). Ancient DNA analysis. Nature Reviews Methods Primers, 1(1), 14. https://doi.org/10.1038/s43586-020-00011-0

Pääbo, S. (1989). Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. Proceedings of the National Academy of Sciences, 86(6), 1939–1943. https://doi.org/10.1073/pnas.86.6.1939

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., & Nieselt, K. (2016). EAGER: Efficient ancient genome reconstruction. Genome Biology, 17(1), 60. https://doi.org/10.1186/s13059-016-0918-z

Penna, A., Blair, M. E., Lui, H.-L., Peters, E., Kistler, L., & Pozzi, L. (2024). Overcoming challenges to extracting and sequencing historical DNA to support primate evolutionary research and conservation, with an application to galagos. International Journal of Primatology, 45(6), 1375–1403. https://doi.org/10.1007/s10764-024-00429-3

Plaxton, L., Hempel, E., Marsh, W. A., Portela Miguez, R., Waurick, I., Kitchener, A. C., Hofreiter, M., Lister, A. M., Zachos, F. E., & Brace, S. (2023). Assessing the identity of rare historical museum specimens of the extinct blue antelope (*Hippotragus leucophaeus*) using an ancient DNA approach. Mammalian Biology, 103(6), 549–560. https://doi.org/10.1007/s42991-023-00373-4

Quatela, A.-S., Cangren, P., Jafari, F., Michel, T., De Boer, H. J., & Oxelman, B. (2023). Retrieval of long DNA reads from herbarium specimens. AoB PLANTS, 15(6), plad074. https://doi.org/10.1093/aobpla/plad074

Radin, J. (2015). Planned Hindsight: The vital valuations of frozen tissue at the zoo and the natural history museum. Journal of Cultural Economy, 8(3), 361–378. https://doi.org/10.1080/17530350.2015.1039458

Raxworthy, C. J., & Smith, B. T. (2021). Mining museums for historical DNA: advances and challenges in museomics. Trends in Ecology & Evolution, 36(11), 1049–1060. https://doi.org/10.1016/j.tree.2021.07.009

Rellstab, C., Gugerli, F., Eckert, A. J., Hancock, A. M., & Holderegger, R. (2015). A practical guide to environmental association analysis in landscape genomics. Molecular Ecology, 24(17), 4348–4370. https://doi.org/10.1111/mec.13322

Rocha, L. A., Aleixo, A., Allen, G., Almeda, F., Baldwin, C. C., Barclay, M. V. L., Bates, J. M., Bauer, A. M., Benzoni, F., Berns, C. M., Berumen, M. L., Blackburn, D. C., Blum, S., Bolaños, F., Bowie, R. C. K., Britz, R., Brown, R. M., Cadena,

C. D., Carpenter, K., … Witt, C. C. (2014). Specimen collection: An essential tool. Science, 344(6186), 814–815. https://doi.org/10.1126/science.344.6186.814

Rohland, N., Glocke, I., Aximu-Petri, A., & Meyer, M. (2018). Extraction of highly degraded DNA from ancient bones, teeth, and sediments for high-throughput sequencing. Nature protocols, 13(11), 2447–2461. https://doi.org/10.1038/s41596-018-0050-5

Salis, A.T., Watson, Z., Forcellati, M., Ali, N., Karmakar, S., Kaur, J., Rabibisoa, N., Raxworthy, C.J. and Smith, B.T. (2025), Unrecognised DNA Degradation in Flash-Frozen Genetic Samples in Natural History Collections. Molecular Ecology Resources e14121. https://doi.org/10.1111/1755-0998.14121

Schmid, S., Neuenschwander, S., Pitteloud, C., Heckel, G., Pajkovic, M., Arlettaz, R., & Alvarez, N. (2018). Spatial and temporal genetic dynamics of the grasshopper *Oedaleus decorus* revealed by museum genomics. Ecology and Evolution, 8(3), 1480–1495. https://doi.org/10.1002/ece3.3699

Schubert, M., Ermini, L., Sarkissian, C. D., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., & Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nature Protocols, 9(5), 1056–1082. https://doi.org/10.1038/nprot.2014.063

Shafer, A. B. A., Wolf, J. B. W., Alves, P. C., Bergström, L., Bruford, M. W., Brännström, I., Colling, G., Dalén, L., De Meester, L., Ekblom, R., Fawcett, K. D., Fior, S., Hajibabaei, M., Hill, J. A., Hoezel, A. R., Höglund, J., Jensen, E. L., Krause, J., Kristensen, T. N., … Zieliński, P. (2015). Genomics and the challenging translation into conservation practice. Trends in Ecology & Evolution, 30(2), 78–87. https://doi.org/10.1016/j.tree.2014.11.009

Shapiro, B., Sibthorpe, D., Rambaut, A., Austin, J., Wragg, G. M., Bininda-Emonds, O. R. P., Lee, P. L. M., & Cooper, A. (2002). Flight of the dodo. Science, 295(5560), 1683–1683. https://doi.org/10.1126/science.295.5560.1683

Simmons, J. E. (2014). Fluid preservation: A comprehensive reference. Rowman & Littlefield Publishers. https://doi.org/10.5771/9781442229662

Simpson, M. (2009). Museums and restorative justice: Heritage, repatriation and cultural education. Museum International, 61(1–2), 121–129. https://doi.org/10.1111/j.1468-0033.2009.01669.x

Van Der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., … Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. Nature, 591(7849), 265–269. https://doi.org/10.1038/s41586-021-03224-9

Verry, A. J. F., Scarsbrook, L., Scofield, R. P., Tennyson, A. J. D., Weston, K. A., Robertson, B. C., & Rawlence, N. J. (2019). Who, where, what, wren? Using ancient DNA to examine the veracity of museum specimen data: a case study of the New Zealand rock wren (*Xenicus gilviventris*). Frontiers in Ecology and Evolution, 7, 496. https://doi.org/10.3389/fevo.2019.00496

Wandeler, P., Hoeck, P. E. A., & Keller, L. F. (2007). Back to the future: Museum specimens in population genetics. Trends in Ecology & Evolution, 22(12), 634–642. https://doi.org/10.1016/j.tree.2007.08.017

Winker, K., Michael Reed, J., Escalante, P., Askins, R. A., Cicero, C., Hough, G. E., & Bates, J. (2010). The Importance, effects, and ethics of bird collecting. The Auk, 127(3), 690–695. https://doi.org/10.1525/auk.2010.09199

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., Koepfli, K.-P., Houck, M. L., Perelman, P., Mastromonaco, G., Bentley, A. C., Venkatesh, B., Zhang, Y., Murphy, R. W., & G10KCOS. (2012). Tissue sampling methods and standards for vertebrate genomics. GigaScience, 1(1), 2047-217X-1–8. https://doi.org/10.1186/2047-217X-1-8

York, N. D. L., Pritchard, R., Sauls, L. A., Enns, C., & Foster, T. (2023). Justice and ethics in conservation remote sensing: Current discourses and research needs. Biological Conservation, 287, 110319. https://doi.org/10.1016/j.biocon.2023.110319

Zhou, Z. T., Owens, G. L., Larson, W. A., Lou, R. N., & Sudmant, P. H. (2024). loco-pipe: An automated pipeline for population genomics with low-coverage whole-genome sequencing. Bioinformatics Advances, 4(1), vbae098. https://doi.org/10.1093/bioadv/vbae098

The Application of Conservation Museomics Approaches to the Protection of the Iberian Lynx (*Lynx pardinus*)

Lauren T. Clark[i,ii], Alexander T. Salis[iii], Anna Penna[iv,v,vi], Megan Wallace[i], Lochlan Sife Krupa[vii], and Mary E. Blair[viii]

[i]*Institute of Comparative Genomics, American Museum of Natural History, New York, NY, USA;* [ii]*Centre for Paleogenetics, Stockholm University, Stockholm, Sweden;* [iii]*Department of Herpetology, American Museum of Natural History, New York, NY, USA;* [iv]*National Museum of Natural History, Smithsonian Institution, Washington, DC, USA;* [v]*Department of Anthropology, University of Texas at San Antonio, San Antonio, TX, USA;* [vi]*Department of Biology, Lund University, Lund, Sweden;* [vii]*Ecology, Evolution and Environmental Biology Department, Columbia University, New York, NY, USA;* [viii]*Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA*

DOI: https://doi. org/10.5531/cbc.linc.14.1.8 | Supplementary: http://doi.org/10.5531/cbc.ncep.0190

ABSTRACT

This exercise is intended to provide students with a real-world example of how museum specimens can provide additional context to challenges and considerations in the broader field of applied conservation genomics. After a brief introduction to the study system—the conservation status of Iberian lynx (*Lynx pardinus*)—students will be asked to review an open-access, peer-reviewed publication that features samples from varied museum, archaeological, and paleontological contexts. Through reading the guide and discussion questions, students will reflect upon study design, concepts, and challenges presented at the intersection between the fields of conservation genomics and museum-based studies. The exercise ends with students breaking down the research into the main components (e.g., research question, independent and dependent variables, hypotheses, predictions) to set a structure for critically reading other scientific studies or designing their own research question.

LEARNING OBJECTIVES

Through this case-study based exercise, students will critically analyze a conservation genomics/museomics/paleogenomics project, including understanding study methodologies, interpreting results, and evaluating the significance of findings.

Specifically, students will:

1. Apply conservation genomics principles: students will interpret the case study and apply their understanding of conservation museum genomics to answer comprehension and critical thinking questions.
2. Formulate and interpret scientific hypotheses: students will identify the hypotheses outlined in the paper to reconstruct the research questions and workflow from the case study.
3. Develop experimental design skills: students will identify relevant research questions, define variables and key parameters for measurement, and outline an experimental design encompassing the genetic sampling of both modern and historical specimens.

PREREQUISITES

Please read this module's synthesis, "Applications of museum collections and genomics to biodiversity conservation" (https://doi. org/10.5531/cbc.linc.14.1.7), prior to completing this exercise.

This exercise assumes that students are familiar with the fundamental principles of conservation genetics and population genetics analyses. For example, students should be comfortable with major principles used in conservation genetics, such as genetic diversity, genetic drift, gene flow,

evolutionary systematics, geographic isolation, and inbreeding.

If you would like to review background material on these concepts, we suggest you read the NCEP module "Conservation genetics" (https://doi.org/10.5531/cbc.ncep.0123).

This module also has a companion exercise: "Designing a conservation genetics project incorporating DNA from museum specimens" (https://doi. org/10.5531/cbc.linc.14.1.9), which builds off Part 2 of this exercise and applies the skills to a new hypothetical case study.

INTRODUCTION: THE CONSERVATION STATUS OF THE IBERIAN LYNX (*LYNX PARDINUS*)

When the number of mature Iberian lynx (*Lynx pardinus*; Figure 1) reached a record-low 62 mature individuals, the species was added to the IUCN Red List (https://www.iucnredlist.org/) as Critically Endangered in 2002 (Nowell, 2002). At this time, *Lynx pardinus* was subject to a combination of multiple threats that included losing over 80 percent of its habitat between 1960 and 1990 (Rodríguez & Delibes, 2002; Palomares et al., 2011). The modification and fragmentation of the mixed grassland and forest habitat likely contributed to the existing scarcity of one of the lynx's main prey species: the European rabbit (*Oryctolagus cuniculus*). This species was already in decline due to an outbreak of the myxomatosis virus (Delibes-Mateos et al., 2009). Additionally, habitat loss also may have led to the contraction of the original nine populations of *Lynx pardinus* in the 1990s to two isolated, remnant populations of the Iberian lynx—one located near Dõnana National Park near the southwestern coast of Spain and the other Andújar-Cardeña population about 250 miles (~400 km) east in the Sierra Morena mountain (Palomares et al., 2011).

In 2003, after several surveys of the lynx by national and international bodies, it was concluded that the Iberian lynx was at risk of extinction. In response, the Spanish Environmental Ministry and the Andalusian Environmental Council in southern Spain signed an agreement to move forward in a captive breeding program for the species (Vargas et al., 2008).

The intended goals of the conservation program were first directed towards *in situ* conservation by

Figure 1: Kittens of *Lynx pardinus*, the Iberian lynx, photographed drinking water in a protected area in Southern Spain. Image credit: Angel Enrique Diaz Martinez.

monitoring and expanding both the Iberian lynx and its prey, the wild rabbit. To accomplish this first goal, the team set out to genotype founders of the lynx captive breeding program to determine which individuals from either Dõnana or Sierra Morena would be most suitable to be the founders of the admixed population. The second goal of the program was to look to the future of *ex situ* conservation that would look to the future by preparing habitat for the free-ranging populations as well as to consider the genetic management of both wild and captive breeding populations as a single metapopulation (Vargas et al., 2008).

In 2005, Saliega and Garfio, two captive parents—one from the Sierra population and the other from the Dõnana population—gave birth to the first three captively-bred, admixed Iberian lynx cubs at the El Acebuche captive breeding center in Dõnana National Park. Once captive-bred individuals reached maturity, it was imperative to establish a wild population of admixed individuals given how vulnerable the two remnant populations were currently (Vargas et al., 2008). Releases of captive-bred individuals began in 2009 and 2010 in two areas of the Sierra Morena mountains in the historic range of Iberian lynx. These individuals had offspring with individuals from the Andújar-Cardeña population, demonstrating that reintroductions would aid in the goal of forming a meta-population dynamic. Five years later, in 2014 and 2015, reintroductions of hybrids started in four areas outside Andalusia (in southern Spain and Portugal) (Pérez de Ayala, 2019). Around this same time, the IUCN re-classified the Iberian lynx species from Critically Endangered to Endangered (Rodríguez & Calzada, 2014).

In 2020, there were an estimated 1,111 Iberian lynx individuals, though in order to achieve Favourable Conservation Status (FCS), a classification from the European Union that ensures protection from extinction, at least eight more subpopulations needed to be formed based on historical population data (Pérez de Ayala, 2019). Due to the significant fragmentation of the habitat of the Iberian lynx over its history and decline in the 20th century, it was necessary to rely on both historic census and paleogenomic data from museum and paleontological samples, for the Iberian lynx to be reclassified once more by the IUCN, as Vulnerable in 2024. This change in distinction finally allowed the *Lynx pardinus* to reach 'green status' according to the IUCN. For more information about the conservation history and current status of Iberian lynx, see the resources provided in Appendix 2.

OVERVIEW: CONSERVATION MUSEOMICS AND THE IBERIAN LYNX

Museum, as well as archaeological/paleontological (or otherwise 'ancient'), samples can be incredible gateways and "timestamps" of genetic diversity at various time points in the past. In comparing DNA produced from these samples to contemporary specimens, researchers can also get a much clearer picture of fluctuations in past demography (e.g., loss of heterozygosity, increased genetic differentiation between populations, bottlenecks) in the past and how that may contribute to current population structure.

The research case study presented here used both historic (collected between 1920 and 1990) and archaeological (prior to 1920) Iberian lynx samples to understand questions about what was once a critically endangered species with fewer than 100 living individuals (Gil-Sánchez & McCain, 2011).

To complete the exercise below (in Part 1 and Part 2), read the research article by Casas-Marques et al. (2017) published by Molecular Biology and Evolution titled "Spatiotemporal dynamics of genetic variation in the Iberian lynx along its path to extinction reconstructed with ancient DNA" (https://doi.org/10.1093/molbev/msx222). Below are guiding questions as you read through.

Appendix 1 is a worksheet for your answers to Part 1 and Part 2.

PART 1: CASE STUDY QUESTIONS

Reading Guide Questions

We recommend you start by reading through all of the questions, actively reading the article with these questions in mind, and then answering the questions on your own.

1. Why do the authors suspect genetic drift played such a large role in the genetic history of Iberian lynx? It may be helpful to refer to Figure 6 in your answer.
2. What is the likely reason for higher ancestral genetic variation observed among the oldest, ancient samples used in this study?
3. How were the authors able to conclude that inbreeding vs outbreeding depression was likely the cause of the current genetic structure? What implications does this have for potential genetic management strategies ("to mix or not to mix")?
4. What conservation strategies are suggested as a result of the paper's findings?
5. Why were museums essential to this study?

Group Discussion Questions

Discuss the questions below with a partner, a small group, or as a class.

6. What parts of the genome were sequenced and analyzed among the modern, historic, and ancient samples, as compared to only the modern and historic samples used in this study? Why do you think the collection of data differed between the sample groups? Consider the role of DNA degradation and how that may influence different genetic markers (mitochondrial or nuclear DNA) examined in this study.
7. What did the analysis of DNA from historic and ancient samples add to this study? Do you think the authors would have reached the same conclusion if they had only analyzed modern Iberian lynx samples?
8. Extensive metadata (e.g., dates and locations) were available for all of the samples used in this study. In your opinion, if this data were not available, should the researchers have continued with their same research plan? Keep in mind that DNA extraction is an inherently destructive process.
9. The authors mention that low genetic diversity across the species may have made Iberian lynx more resistant to the impacts of genomic erosion and potential extinction. How might this skew public perception of the need to conserve this species? What arguments provided from the paper would you share with the skeptics to protect the Iberian lynx populations mentioned?

PART 2: RESEARCH FOUNDATION AND PROJECT DESIGN

Now that you've critically read and interpreted the details of a conservation museomics study (Casas-Marce et al., 2017), we are going to go back to basics and think about the main components of this research. Part 2 includes guiding questions to help build the skills needed to identify, develop, and answer a research question. These guiding questions are general and may be applied to your review of any research project or used as a jumping-off point for designing your own research project (see this module's companion exercise "Designing a conservation genomics project incorporating DNA from museum specimens" (https://doi. org/10.5531/cbc.linc.14.1.9).

Guiding Questions to Interpret the Research Design of Casas-Marce et al. (2017):

1. What were the authors trying to answer in this study? Formulate a research question from the study.

2. In your view, is genetic diversity an independent or dependent variable as it's presented in the paper? Explain. Recall that an independent variable is a parameter that, when altered, affects the dependent variables but is not affected by those variables itself (e.g., time, total area, temperature, flow rate, etc.). The dependent variables are all other components influenced by the dependent variable. What are other independent and dependent variables from the text?
3. Consider the null and alternative hypotheses from the role of genetic diversity you describe in your answer to question 2. Formulate both null and alternative hypotheses with genetic diversity in mind from the publication using "If… then…" statements to connect your research question and experimental variables. For example: "If the hypothesis is correct, then I expect an increase in the independent variable will lead to an increase in a dependent variable." This positive relationship can then be tested through a formal statistical framework. Hypotheses serve as the starting point for any experiment. Therefore, any flaw in their formulation may lead to flaws in the design of the entire experiment.
   Let's define the null hypothesis (H_0) and the alternative hypothesis (H_1). In the scientific method, we can only falsify hypotheses. Usually, H_0 assumes there is no change or relationship between the observations, which researchers try to disprove. H_1 is accepted if there is sufficient evidence to reject the null.
4. Now, let's create a graphical representation of our expectations for genetic diversity over time, based on the results described in the paper, connecting the independent and dependent variables you identified in the first part of question #2 above. Contrast the expectations for each hypothesis using a pair of illustrative graphs depicting the projected outcomes for H_0 and H_1. These could be a line graph, scatter plot, boxplot, etc. Don't forget to label your axes!
5. Write a concise paragraph explaining how these contrasting results, as indicated in the graph drawn above, can be interpreted. In your explanation, be sure to mention a research question, your null and alternative hypotheses, and conclusions the researcher might make depending on the observed relationship between genetic diversity and time in your graph.

ACKNOWLEDGMENTS

We would like to thank Mckenna Santiago Coyle, Melina Giakoumis, and Stephen Gaughran. A.P. was funded by the American Society of Mammalogy (ASM), Society for Systematic Biology (SSB), Systematics Association, Linnean Society of London, and National Science Foundation of the United States (NSF; Award No. BSC-2120691); A.P. was also funded by the Smithsonian Institution Graham Funds in Human Origins and the University of Texas at San Antonio (UTSA). A.T.S was funded by the U.S. National Science Foundation (DBI 2029955). L.T.C. was funded, in part, by the U.S. National Science Foundation (DBI 2029955). M.E.B. was funded by the National Science Foundation of the United States (Awards No. BCS-1926105/1926215). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

REFERENCES

Abascal, F., Corvelo, A., Cruz, F., Villanueva-Cañas, J. L., Vlasova, A., Marcet-Houben, M., Martínez-Cruz, B., Cheng, J. Y., Prieto, P., Quesada, V., Quilez, J., Li, G., García, F., Rubio-Camarillo, M., Frias, L., Ribeca, P., Capella-Gutiérrez, S., Rodríguez, J. M., Câmara, F., … Godoy, J. A. (2016). Extreme genomic erosion after recurrent demographic bottlenecks in the highly endangered Iberian lynx. Genome Biology, 17(1), 251. https://doi.org/10.1186/s13059-016-1090-1

Delibes-Mateos, M., Ferreras, P., & Villafuerte, R. (2009). European rabbit population trends and associated factors: A review of the situation in the Iberian Peninsula. Mammal Review, 39(2), 124–140. https://doi.org/10.1111/j.1365-2907.2009.00140.x

Fordham, D. A., Akçakaya, H. R., Brook, B. W., Rodríguez, A., Alves, P. C., Civantos, E., Triviño, M., Watts, M. J., & Araújo,

M. B. (2013). Adapted conservation measures are required to save the Iberian lynx in a changing climate. Nature Climate Change, 3(10), 899–903. https://doi.org/10.1038/nclimate1954

Gil-Sánchez, J. M., & McCain, E. B. (2011). Former range and decline of the Iberian lynx (*Lynx pardinus*) reconstructed using verified records. Journal of Mammalogy, 92(5), 1081–1090. https://doi.org/10.1644/10-MAMM-A-381.1

Nowell, K. (2002). Revision of the Felidae red list of threatened species (No. 37; CAT News, pp. 4–7). IUCN Cat Specialist Group. https://www.felidae.org/KNOWELLPUBL/nowell2002redlistcn.pdf

Palomares, F., Rodríguez, A., Revilla, E., López-Bao, J. V., & Calzada, J. (2011). Assessment of the conservation efforts to prevent extinction of the Iberian lynx. Conservation Biology, 25(1), 4–8. https://doi.org/10.1111/j.1523-1739.2010.01607.x

Rodríguez, A. (2024). *Lynx pardinus*. The IUCN Red List of Threatened Species 2024: E.T12520A218695618. https://dx.doi.org/10.2305/IUCN.UK.2024-1.RLTS.T12520A218695618.en

Rodríguez, A., & Calzada, J. (2014). *Lynx pardinus* (errata version published in 2020). The IUCN Red List of Threatened Species 2015: E.T12520A174111773. https://dx.doi.org/10.2305/IUCN.UK.2015-2.RLTS.T12520A174111773.en

Rodríguez, A., & Delibes, M. (2002). Internal structure and patterns of contraction in the geographic range of the Iberian lynx. Ecography, 25(3), 314–328. https://doi.org/10.1034/j.1600-0587.2002.250308.x

Vargas, A., Sánchez, I., Martínez, F., Rivas, A., Godoy, J. A., Roldán, E., Simón, M. A., Serra, R., Pérez, MaJ., Enseñat, C., Delibes, M., Aymerich, M., Sliwa, A., & Breitenmoser, U. (2008). The Iberian lynx *Lynx pardinus* Conservation Breeding Program. International Zoo Yearbook, 42(1), 190–198. https://doi.org/10.1111/j.1748-1090.2007.00036.x

APPENDIX 1

Applications of Conservation Museomics to the Protection of the Iberian Lynx (*Lynx pardinus*)

Please read this module's synthesis, "Applications of museum collections and genomics to biodiversity conservation" (https://doi. org/10.5531/cbc.linc.14.1.7) prior to completing this exercise.

This module also has a companion exercise: "Designing a conservation genetics project incorporating DNA from museum specimens," (https://doi. org/10.5531/cbc.linc.14.1.9) which may be a helpful resource.

Date: _____ Name/Group: _____

Part 1: Case Study Questions

Reading Guide Questions: we recommend you start by reading through all of the questions, actively reading the article with these questions in mind, and then answering the questions on your own.

1. Why do the authors suspect genetic drift played such a large role in the genetic history of Iberian lynx? It may be helpful to refer to Figure 6 in your answer.

2. What is the likely reason for the ancestral genetic variation observed among the oldest ancient samples used in this study?

3. How were the authors able to conclude that inbreeding vs outbreeding depression was likely the cause of the current genetic structure? What implications does this have for potential genetic management strategies ("to mix or not to mix")?

4. What conservation strategies are suggested as a result of the paper's findings?

5. Why were museums essential to this study?

Group Discussion Questions: discuss the questions below with either a partner or group.

6. What parts of the genome were sequenced and analyzed among the modern, historic, and ancient samples, as compared to only the modern and historic samples used in this study? Why do you think the collection of data differed between the sample groups? Consider the role of DNA degradation and how that may influence different genetic markers (mitochondrial or nuclear DNA) examined in this study.

7. What did the analysis of DNA from historic and ancient samples add to this study? Do you think the authors would have reached the same conclusion if they had only analyzed modern Iberian lynx samples?

8. Extensive metadata (e.g., dates and locations) were available for all of the samples used in this study. In your opinion, if this data were not available, should the researchers have continued with their same research plan? Keep in mind that DNA extraction is an inherently destructive process.

9. The authors mention that low genetic diversity across the species may have made Iberian lynx more resistant to the impacts of genomic erosion and potential extinction. How might this skew public perception of the need to conserve this species? What arguments provided from the paper would you share with the skeptics to protect the Iberian lynx populations mentioned?

Part 2: Research Foundation and Project Design

1. What were the authors trying to answer in this study? Formulate a research question from the study.

2. In your view, is genetic diversity an independent or dependent variable as it's presented in the paper? Explain. Recall that an independent variable is a parameter that, when altered, affects the dependent variables but is not affected by those variables itself (e.g., time, total area, temperature, flow rate, etc.). The dependent variables are all other components influenced by the dependent variable. What are other independent and dependent variables from the text?

Independent variables	Dependent variables
•	•
•	•

3. Consider the null and alternative hypotheses from the role of genetic diversity you describe in your answer to question 2. Formulate both null and alternative hypotheses with genetic diversity in mind from the publication using "If… then…" statements to connect your research question and experimental variables. For example: "If the hypothesis is correct, then I expect an increase in the independent variable will lead to an increase in a dependent variable." This positive relationship can then be tested through some formal statistical framework. Hypotheses serve as the starting point for any experiment. Therefore, any flaw in their formulation may cause a flaw in the design of an entire experiment.

   So, let's define the null (H0) and alternative hypotheses (H1). In the scientific method, we can only falsify hypotheses. Usually, H0 assumes there is no change or relationship between the observations, which researchers try to disprove. H1 is accepted if there is sufficient evidence to reject the null.

H0:

H1:

4. Now let's make a graphical representation of our expectations of genetic diversity over time, based on the results described in the paper, connecting the independent and dependent variables you identified in the first part of question #2 above. Contrast the expectations for each hypothesis using a pair of illustrative graphs depicting the projected outcomes for H0 and H1. These could be a line graph, scatter plot, boxplot, etc. Don't forget to label your axes!

5. Write a concise paragraph explaining how these contrasting results, as indicated in the graph drawn above, can be interpreted. In your explanation, be sure to mention a research question, your null and alternative hypotheses, and conclusions the researcher might make depending on the observed relationship between genetic diversity and time in your graph.

APPENDIX 2

Additional Resources on Iberian Lynx:

- Lynxconnect project: https://lifelynxconnect.eu/en/

Popular Media Articles:

- Mongabay: Dasgupta, S. (2024). Incredibly rare Iberian lynx making a "dramatic recovery." Mongabay. https://news.mongabay.com/short-article/incredibly-rare-iberian-lynx-making-a-dramatic-recovery/
- IUCN: IUCN. (2024). Iberian lynx rebounding thanks to conservation action—IUCN Red List. IUCN. https://iucn.org/press-release/202406/iberian-lynx-rebounding-thanks-conservation-action-iucn-red-list
- Washington Post: Melnick, K. (2024). A European wild cat was nearly extinct. Now, it is making a comeback! The Washington Post. https://www.washingtonpost.com/world/2024/06/22/iberian-lynx-peninsula-endangered-population/
- WWF: Iberian lynx: Threats. WWF. https://wwf.panda.org/discover/our_focus/wildlife_practice/profiles/mammals/iberian_lynx/ibelynx_threats/#:~:text=Under%20future%20climate%20change%20conditions,its%20resilience%20to%20climate%20change
- SAPIENS magazine: Chang, L. (2024). Bringing back the world's most endangered cat. SAPIENS. https://www.sapiens.org/biology/iberian-lynx-conservation-efforts/

Designing a Conservation Genomics Project Incorporating DNA from Museum Specimens

Anna Penna[i,ii,iii], Megan Wallace[iv], Lauren T. Clark[iv,v], Lochlan Sife Krupa[vi], Suzanne K. Macey[vii], Luca Pozzi[iii], and Mary E. Blair[vii]

[i]National Museum of Natural History, Smithsonian Institution, Washington, DC, USA; [ii]Department of Anthropology, University of Texas at San Antonio, San Antonio, TX, USA; [iii]Department of Biology, Lund University, Lund, Sweden; Department of Herpetology, American Museum of Natural History, New York, NY, USA; [iv]Institute of Comparative Genomics, American Museum of Natural History, New York, NY, USA; [v]Centre for Paleogenetics, Stockholm University, Stockholm, Sweden; [vi]Ecology, Evolution and Environmental Biology Department, Columbia University, New York, NY, USA; [vii]Center for Biodiversity and Conservation, American Museum of Natural History, New York, NY, USA

DOI: https://doi. org/10.5531/cbc.linc.14.1.9 | Supplementary: http://doi.org/10.5531/cbc.ncep.0190

ABSTRACT

In this exercise, students reflect critically on the research process and understand the different steps required to implement a research project. Specifically, students will develop a conservation museomics project using a set of hypothetical case studies that include the sampling of historical and modern specimens. Students will 1) use fundamental principles of conservation genetics to formulate research questions and hypotheses, 2) map out the components of their hypothetical experiment, and 3) create contingency and budget optimization plans to overcome common research challenges and a visual aid of their choosing to communicate conservation management strategies for hypothetical taxa.

LEARNING OBJECTIVES

In this exercise, students will apply conservation genetics concepts to develop and explore a hypothetical research project.

Specifically, students will:

1. Apply conservation genomics principles: students will apply principles of conservation genetics and museum genomics to design a comprehensive conservation project that addresses biodiversity conservation issues.
2. Formulate and interpret scientific hypotheses: students will formulate hypotheses based on their research questions and expected outcomes, contrasting their expectations and interpreting results.
3. Develop experimental design skills: students will identify key parameters for measurement and outline a study design encompassing the genetic sampling, processing, sequencing, and analysis of both modern and historical specimens.
4. Reflect on common research challenges: students will contemplate potential issues that can arise throughout the research process, including budget constraints. Based on these obstacles, students will create contingency plans and budget optimization plans.
5. Communicate conservation management strategies: students will brainstorm management strategies, practice communicating their findings, and express solutions and the importance of conservation to the public, a critical skill to develop.

PREREQUISITES

Please read this module's synthesis, "Applications of museum collections and genomics to biodiversity conservation" (https://doi. org/10.5531/cbc.linc.14.1.7) prior to completing this exercise.

This exercise assumes that students are familiar with the fundamental principles of conservation

genetics and population genetics analyses. For example, students should be comfortable with major principles used in conservation genetics, such as genetic diversity, genetic drift, gene flow, evolutionary systematics, geographic isolation, and inbreeding.

If you would like to review background material on these concepts, we suggest you read the NCEP module "Conservation genetics" (https://doi.org/10.5531/cbc.ncep.0123).

This module also has a companion exercise: "The application of conservation museomics approaches to the protection of the Iberian lynx (*Lynx pardinus*)" (https://doi. org/10.5531/cbc.linc.14.1.8), which may be helpful for setting the foundations of designing a new conservation genetics project.

The authors of this exercise and their colleages have also developed a hands-on bioinformatics exercise of a population genetic study that combines historical and modern samples that may be a helpful extension to this exercise (https://github.com/sjgaughran/amnh-museumomics).

OVERVIEW

Natural history collections are like time capsules. In the US alone, they house almost a billion specimens preserved during different periods, including prehistoric times! With the recent advancements in DNA sequencing technologies and molecular biology protocols for the isolation of genetic material, it is now possible to obtain genetic data from these preserved specimens—a field of research often referred to as "museomics." Thanks to the collective dedication of various naturalists, field biologists, curators, and collection staff, natural history collections provide records that can be used as a temporal comparison to investigate population trends, including specimens that serve as testimony of populations or species that are now extinct in the wild.

In this exercise, you will be presented with a few hypothetical scenarios in which principles of conservation genetics can be combined with museomics to better understand population trends, connectivity, and evolutionary relationships. Upon completion of the exercise, you will have developed a conservation genetics study design and risk analysis of a project incorporating historical samples.

First: carefully read the hypothetical cases and the respective researchers' interests found in Appendix 1. Pick one example to focus on in this exercise.

PART 1: RESEARCH FOUNDATION

Before starting an experiment, it is important to have a clear research question in mind. This question will serve as a guide through testing hypotheses and determining what data to collect. Thus, the research question must be aligned with the researcher's interests (the project goal) and make clear connections with measurable or quantifiable variables. These connections are formalized as hypotheses that will then be tested using statistical methods.

If you have not already, read the hypothetical case studies in Appendix 1 and select one to be the focus of the following prompts. Alternatively, your instructor may encourage you to come up with your own research question/case study.

1. First things first. What is this study aiming to answer? Formulate a research question that aligns with the researcher's interest. A research question must have a clear goal that connects the

research interest and the elements of the study (e.g., species, populations, regions, time). *Write your research question in the worksheet (Appendix 2).*

2. Now, identify the independent and the dependent variables in this study. An independent variable is a parameter whose change affects the dependent variables but is not affected by those variables itself (e.g., time, total area, temperature, flow rate). The dependent variables are all other components influenced by the independent variable. In conservation genetics, the dependent variables will often be quantitative (i.e., a numerical data). Here, we suggest recalling the fundamental principles of conservation genetics and the main parameters used in population genetics research [genetic diversity, gene flow, runs of homozygosity, inbreeding coefficient, amount of deleterious genetic variation (i.e., genetic load), effective population size, allele frequency]. We provide an overview of these concepts in this module's synthesis document (https://doi. org/10.5531/cbc.linc.14.1.7).
Identify one independent variable and at least two dependent variables that the researcher will need to measure to answer the research question you formulated in #1 above and write them on the worksheet.

3. Formulate your hypotheses connecting your research question and the experimental variables (what you will be quantifying). One strategy is to use "If… then…" statements. For example: "If x (independent variable) increases, I expect an increase in y (a dependent variable)." By formulating these statements, you can start making predictions about the relationship between variables that can then be tested through some formal statistical framework. Hypotheses serve as the starting point for any experiment. Therefore, any flaw in their formulation may cause a flaw in the design of an entire experiment. So, let's define the null (H_0) and alternative hypotheses (H_1). In the scientific method, we can only falsify hypotheses. Usually, H_0 assumes there is no change or relationship between the observations, which researchers try to disprove. H_1 is accepted if there is sufficient evidence to reject the null.
Write your H_0 *and* H_1 *in the worksheet.*

4. Now let's make a graphical representation of your expectations, connecting the independent and dependent variables you identified in #2 above. Contrast the expectations for each hypothesis using a pair of illustrative graphs, one depicting the projected outcomes for H_0, and the other the projected outcomes for H_1. These could be a line graph, scatter plot, boxplot, etc. Don't forget to label your axes!
Draw graphs contrasting the expectations for H_0 *and* H_1 *in the empty boxes provided on the worksheet. You may wish to draw multiple graphs depending on the number of dependent variables you identified in #2 above.*

5. Write a concise paragraph explaining how these contrasting results can be interpreted. In your explanation, be sure to mention the research question and conclusions the researcher might make depending on the observed relationship between the variables identified in #2 above.
On the worksheet, write the conclusions that can be made based on the expectations depicted in #4 above. Refer back to the hypotheses formed in #3 above.

6. Before moving on to Part 2, pause and think critically: Have you considered the assumptions you are making about relationships or data? What covariates might be interesting to consider? Does correlation mean causation? How can you control for variation? Write a few sentences about your thought process on these questions (or others that you have)—while you might not have the answer, the critical thinking may inform how you design your experiment.

PART 2: EXPERIMENTAL DESIGN

Now that we have a clear idea of what our research goal is, what we must quantify, and what we will be testing, it is time to consider how we will accept or reject the null hypothesis. To this end, you will outline an experimental design that includes the genetic sampling of modern and historical samples to answer the research question and test the hypotheses you formulated in Part 1.

7. What would the ideal sample size be to test this hypothesis and answer the researcher's question and why? What are some challenges you might encounter when trying to acquire and sequence this number of samples?
   Complete the table provided in the worksheet (Appendix 2) and answer the questions above. In your answer, consider sample acquisition, sample quality, and budget.

8. What would be the ideal source of tissue and preparation type (e.g., bone, organ tissue, skins, dried whole specimen) most commonly available for your study organism? What lab work protocols are more appropriate in this case? What are some challenges you might encounter when working with said type of preparation?
   Complete the table provided in the worksheet (Appendix 2) and answer the questions above. In your answer, consider sample acquisition, sample quality, and preservation method (e.g., ethanol, formalin, liquid nitrogen.)

9. Organize the study design in a workflow. When preparing your workflow, think about the different phases of a conservation museomics project (e.g., sample collection, lab work, data analysis) and how to progress from one step to the next. More importantly, think about the relevant intermediate steps to take within each phase. Here are some questions to help you prepare the workflow:
   - o A research experiment starts with gathering the data. Will you need to sample different localities or temporal windows?
   - o What type of genetic data will you need to generate for this investigation? Do they require different laboratory steps?
   - o How will these steps allow you to measure the dependent and independent variables needed to test the hypothesis you proposed?

   Below are some example workflows (Figures 1–4) to inspire the development of your own.
   Complete the flowchart containing the main steps required for the completion of the proposed research. You can add as much detail as you want, but be sure to cover the three major research phases (sample collection, lab work, and data analysis).

Specific details on data analysis go beyond the content of this lesson. The authors of this exercise have developed a bioinformatics activity that may be a helpful resource (https://github.com/sjgaughran/amnh-museumomics). Below are other resources that can be useful to guide the bioinformatics part:

- Schubert, M., Ermini, L., Sarkissian, C. D., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Martin, K., McCue, M., Willerslev, E., & Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nature Protocols, 9(5), 1056–1082. https://doi.org/10.1038/nprot.2014.063
- Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., & Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. Genome biology, 17, 60. https://doi.org/10.1186/s13059-016-0918-z
- Yates, J. A. F., Lamnidis, T. C., Borry, M., Valtueña, A. A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., & Peltzer, A. (2021). Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. PeerJ, 9, e10947. https://doi.org/10.7717/peerj.10947

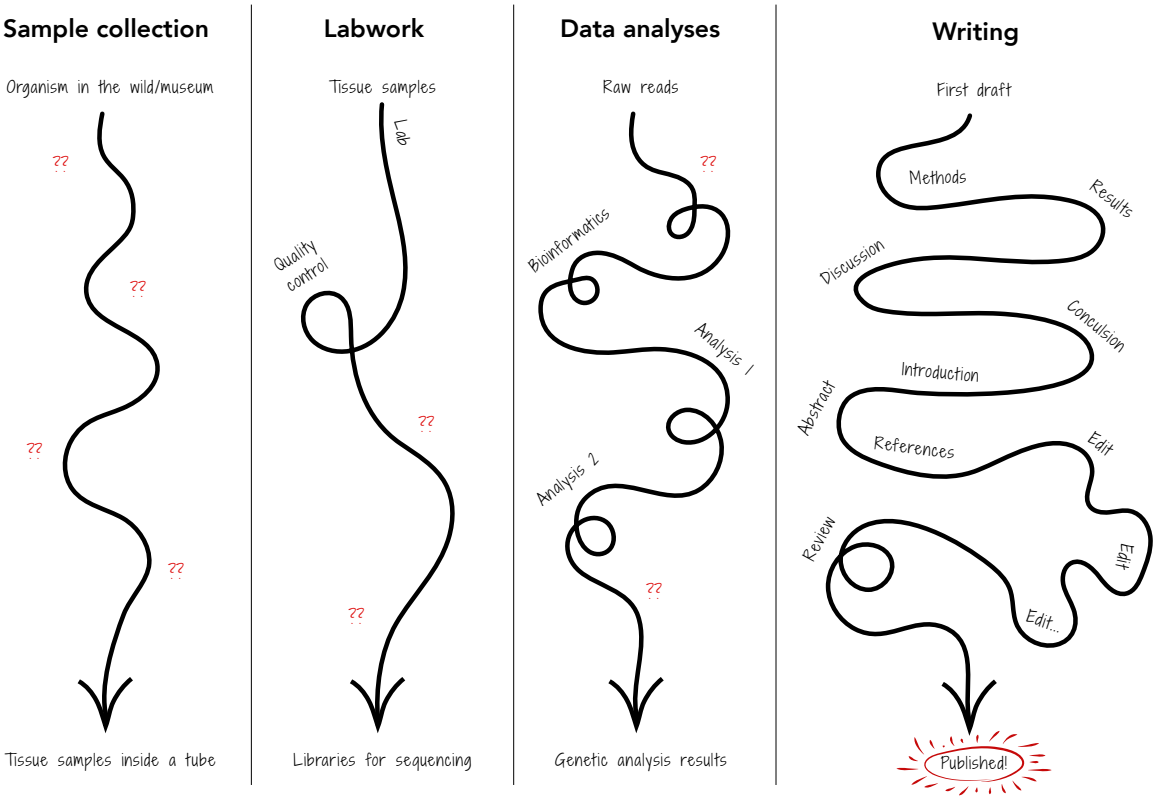


Figure 1: Each phase of the project will have many intermediate steps. Using a workflow can help you plan the study design and make sure you are prepared for each phase of the study. Image credit: Anna Penna.

Figure 2: Example workflow depicting the intermediate steps of specimen collection, sampling, and storage. Image credit: Anna Penna.

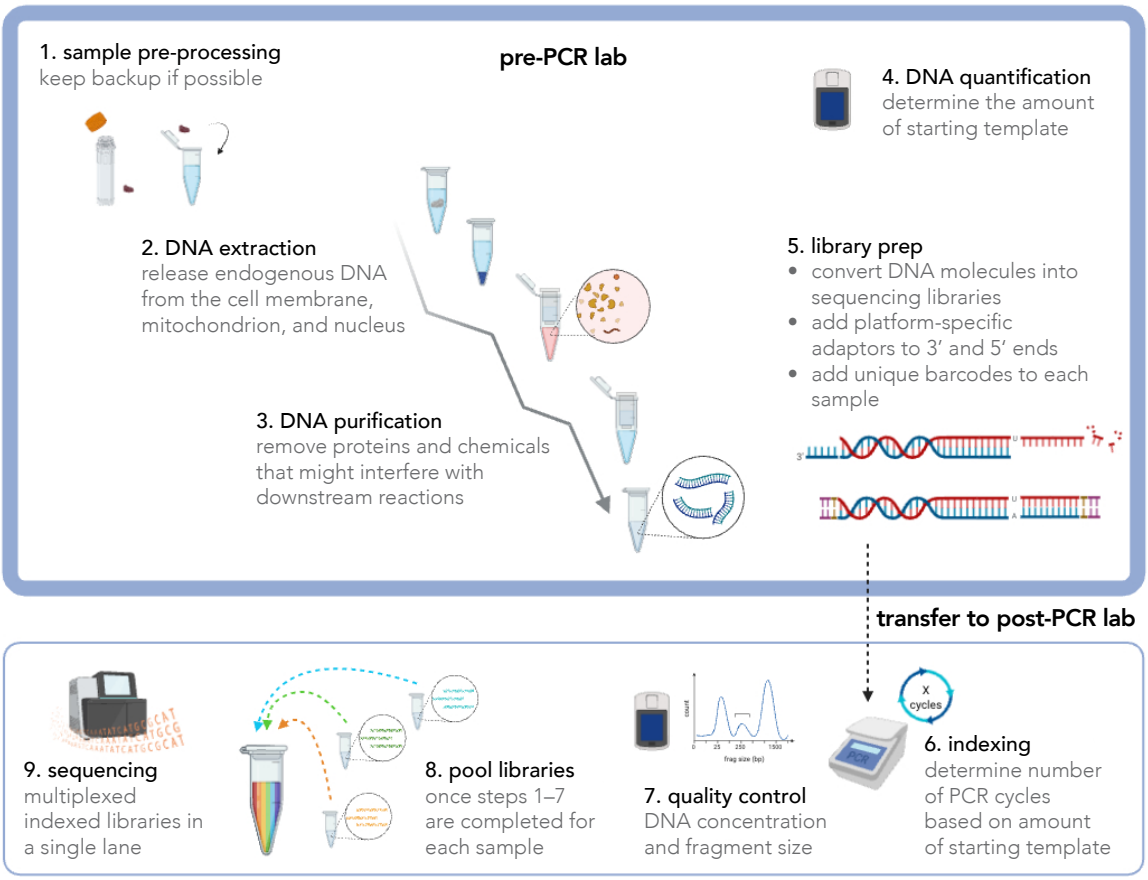


Figure 3: Example workflow depicting the intermediate steps of laboratory work and sequencing. Image credit: Modified from Penna et al., 2024.

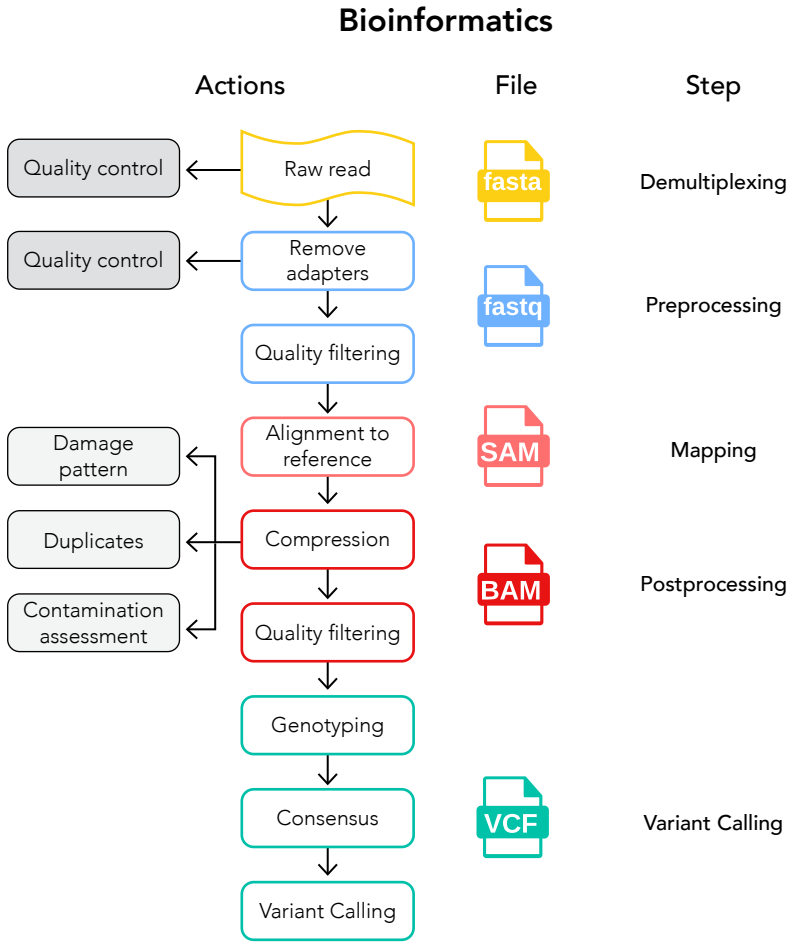


Figure 4: Example workflow depicting general steps taken in the bioinformatics pipeline. Image credit: Anna Penna.

PART 3: CRITICAL THINKING AND COMMUNICATION

Now that you have thought about all the main steps required to implement and complete this research project, it is time to reflect on some of the challenges you may encounter, as well as the potential applications of the data from your study towards conservation efforts. The considerations that emerge from this reflection can be used to create a contingency plan for research hiccups, a project budget optimization plan, and a conservation management plan.

10. What obstacles could arise when working with museum specimens compared to those recently collected in the field? Why do museum specimens pose a challenge? What is the likelihood of these problems occurring? List some strategies to overcome or avoid such obstacles.
    Indicate one potential obstacle for each major research phase, briefly explain why this can be challenging, approximate the likelihood of this occurring, and list risk evasion/avoidance strategies in the table provided in the worksheet.

11. Which steps of this study design do you expect to require the most funding? Think about how you would acquire the samples and determine the sequencing strategy and desired sequencing coverage. For instance, while capture sequencing can help you increase the coverage of a target genomic region, the cost of developing and purchasing baits can be very expensive. Consequently, although the capture strategy can increase the coverage of certain regions, it might restrict the number of samples you can sequence. Fixed monetary sources for your research can impact how many samples you can collect, process, and sequence, so how would you optimize your budget to best accomplish each major phase of your study design?
    Indicate one potential budget constraint for each major research phase, briefly explain why and how lack of funds would impact this stage of the study design, approximate the level of expense, and list cost-saving strategies in the table provided in the worksheet.

12. If the data from your study suggests significant change between historical and modern populations, think about how your findings could be applied to conservation management. How would you best communicate your results if you wanted to share them with management authorities? What are some tools and strategies that you could suggest implementing to aid modern populations in your hypothetical case study? An example conservation strategy is wildlife crossings, which can connect two disparate populations that have been separated by the construction of a highway.
    In a separate document, create a pamphlet or poster aimed at communicating to both management authorities and the public the conservation issue that has affected your hypothetical taxa, the findings from your study, and proposed methods for rehabilitating these hypothetical populations or species. This final exercise should focus on applying science and science communication, so think outside the box and have fun with it!

ACKNOWLEDGMENTS

We would like to thank Mckenna Santiago Coyle, Melina Giakoumis, and Stephen Gaughran. A.P. was funded by the American Society of Mammalogy (ASM), Society for Systematic Biology (SSB), Systematics Association, Linnean Society of London, and National Science Foundation of the United States (NSF; Award No. BSC-2120691); A.P. was also funded by the Smithsonian Institution Graham Funds in Human Origins and the University of Texas at San Antonio (UTSA). A.T.S was funded by the U.S. National Science Foundation (DBI 2029955). L.T.C. was funded, in part, by the U.S. National Science Foundation (DBI 2029955). M.E.B. and S.K.M. were funded by the National Science Foundation of the United States (Awards No. BCS-1926105/1926215). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.

APPENDIX 1

Designing a Conservation Genomics Project Incorporating DNA from Museum Specimens

Hypothetical Case Studies

Example 1

A recent study suggested that the distribution of a small alpine chipmunk species in the Sierra Nevada Mountains in California has shifted to higher altitudes every year, most likely due to increases in surface temperatures driven by climate change. Unfortunately, the mountains are not getting taller, so the populations that occupy the high-elevation areas are getting more and more confined to the very highest portions of the mountain range. If temperatures continue to rise, these populations might not have enough time to adapt and could go extinct. In the early 1900's, naturalists from a nearby museum collected a series of specimens of this species throughout its entire geographic and altitudinal range. You have a permit to sample 10 specimens per locality in this region for genetic studies.

Researcher's interest: Investigate the impact of this recent shift in distribution on the species' extinction risk.

Example 2

An archipelago on the coast of Aotearoa New Zealand has high levels of bird endemism, including an entire radiation of endemic species that evolved into different body sizes, feeding ecologies, and diverse niches. Unfortunately, ever since the accidental introduction of rats in the late 1800s by Europeans, the natural populations found on the islands have shown signs of drastic demographic changes. The rats feed on the birds' eggs and even prey on some adults in their breeding grounds. Consequently, some populations have gone through a rapid decline in population size over the past few decades. The most extreme case is that two out of the three species of flightless birds reported from the island have gone extinct, as they have not been seen for over a century. Lucky for you, a naturalist collected and preserved over 500 bird specimens from this archipelago before the introduction of rats, including the two now-extinct species of flightless birds and the other 20 species currently recognized in this endemic radiation.

Researcher's interest: Determine if the loss of flight evolved once or multiple times in this endemic bird radiation.

Example 3

The construction of a dam in the early 1930s completely changed the hydrography of a region. Streams that once were connected into a complex network flowing into the river have now turned into a big, flooded area. Another major consequence of such change was the discontinuity of the river flow. This river is home to an endemic salamander, which only reproduces in these waters. The interruption of the river flow is potentially impeding individuals from moving freely up and down the river, which might isolate the headwater populations from those in the river delta. Lucky for you, a famous herpetologist from the early 20th century collected over 300 individuals of this species of salamanders from this river before the dam's construction.

Researcher's interest: Evaluate the impact of the dam on the connectivity between populations from up and down the river.

Example 4

A species of butterfly only lays their eggs on the flowers of a particular plant species in the small (fictional) European country of Genovia. Unfortunately, the queen ruling Genovia in the 1960s was extremely allergic to the pollen produced by this particular flower, so she called for its eradication. A Genovian entomologist, foreseeing the consequences of this change, collected a set of male and female adults, juveniles, and pupae, and then added the preserved specimens to his private collection. Meanwhile, in the fields of Genovia, most of the butterflies die out, but those that survive shift to lay their eggs on the flowers of another plant species. The petals of this new host plant emit a similar fluorescence and conceal the butterfly eggs fairly well. However, there's a larger contrast between the colors of the petals and those of the butterflies, making the adult butterflies more susceptible to predators. Over time, the butterflies' coloration and wing patterns begin to change, and the butterflies become more susceptible to disease.

Researcher's interest: Determine the source(s) of increased disease susceptibility in the present-day butterfly population.

APPENDIX 1

Applications of Conservation Museomics to the Protection of the Iberian Lynx (*Lynx pardinus*)

Please read this module's synthesis, "Applications of museum collections and Genomics to biodiversity conservation" (https://doi. org/10.5531/cbc.linc.14.1.7), prior to completing this exercise.

This module also has a companion exercise: "The application of conservation museomics approaches to the protection of the Iberian lynx (*Lynx pardinus*)" (https://doi. org/10.5531/cbc.linc.14.1.8), which may be a helpful resource.

Date: _____ Name/Group: _____

Part 1: Research Foundation Worksheet

Hypothetical Example: ___

1. First things first. What is this study aiming to answer? Formulate a research question that aligns with the researcher's interest. A research question must have a clear goal that connects the research interest and the elements of the study (e.g., species, populations, regions, time). Research Question:

_____

_____

_____

2. Now, identify the independent and the dependent variables in this study. An independent variable is a parameter whose change affects the dependent variables but is not affected by those variables itself (e.g., time, total area, temperature, flow rate, etc.). The dependent variables are all other components influenced by the independent variable. In conservation genetics, the dependent variables will often be quantitative (i.e., a numerical data). Here, we suggest recalling the fundamental principles of conservation genetics and the main parameters used in population genetics research [genetic diversity, gene flow, runs of homozygosity, inbreeding coefficient, amount of deleterious genetic variation (i.e., genetic load), effective population size, allele frequency]. We provide an overview of these concepts in this module's synthesis document (01-ConservationMuseomics_SYN.docx).

Independent variable	Dependent variables
•	• •

3. Formulate your hypotheses connecting your research question and the experimental variables (what you will be quantifying). One strategy is to use "If… then…" statements. For example: "If x (independent variable) increases, I expect an increase in y (a dependent variable)." By formulating these statements, you can start making predictions about the relationship between variables that can then be tested through some formal statistical framework. Hypotheses serve as the starting point for any experiment. Therefore, any flaw in their formulation may cause a flaw in the design of an entire experiment. So, let's define the null (H0) and alternative hypotheses (H1). In the scientific method, we can only falsify hypotheses. Usually, H0 assumes there is no change or relationship between the observations, which researchers try to disprove. H1 is accepted if there is sufficient evidence to reject the null.

H0:

H1:

4. Now let's make a graphical representation of your expectations, connecting the independent and dependent variables you identified in #2 above. Contrast the expectations for each hypothesis using a pair of illustrative graphs, one depicting the projected outcomes for H0, and the other the projected outcomes for H1. These could be a line graph, scatter plot, boxplot, etc. Don't forget to label your axes!

5. Write a concise paragraph explaining how these contrasting results can be interpreted. In your explanation, be sure to mention the research question and conclusions the researcher might make depending on the observed relationship between the variables identified in #2 above.

Expectations and Conclusions:

6. Before moving on to Part 2, pause and think critically: Have you considered the assumptions you are making about relationships or data? What covariates might be interesting to consider? Does correlation mean causation? How can you control for variation? Write a few sentences about your thought process on these questions (or others that you have)—while you might not have the answer, the critical thinking may inform how you design your experiment.

Part 2: Experimental Design

7. What would the ideal sample size be to test this hypothesis and answer the researcher's question and why? What are some challenges you might encounter when trying to acquire and sequence this number of samples?

Specimen age	Ideal sample size	Species or population	Locality

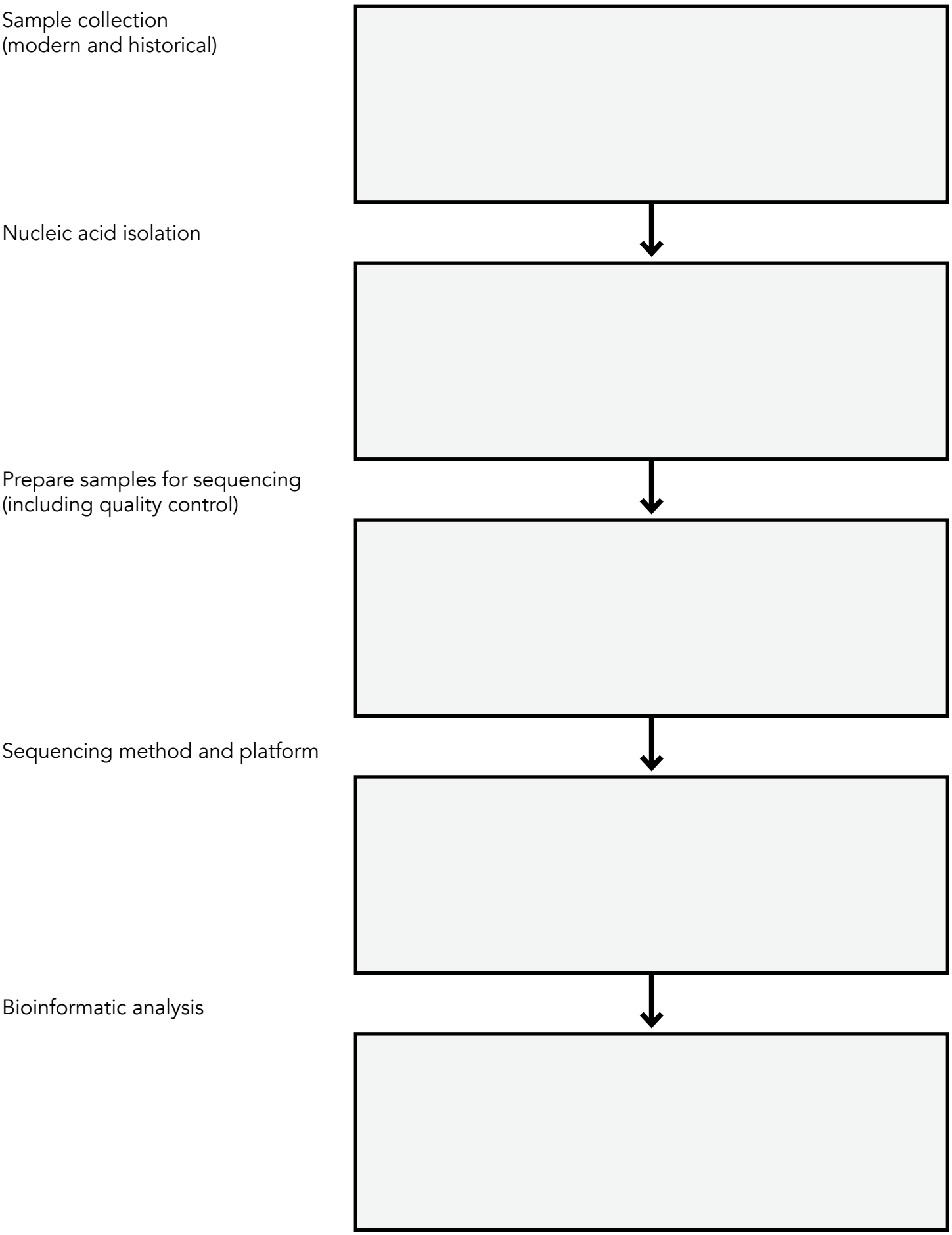
Justification and challenges of determining an acceptable sample size:

8. What would be the ideal source of tissue and preparation type (e.g., bone, organ tissue, skins, dried whole specimen, etc.) most commonly available for your study organism? What lab work protocols are more appropriate in this case? What are some challenges you might encounter when working with said type of preparation?

Specimen age	Ideal sample size	Species or population	Locality

Justification and challenges of selecting specimens for your study:

9. Organize the study design in a workflow. When preparing your workflow, think about the different phases of a conservation museomics project (e.g., sample collection, lab work, data analysis) and how to progress from one step to the next. More importantly, think about the relevant intermediate steps to take within each phase. Here are some questions to help you prepare the workflow:
    - o A research experiment starts with gathering the data. Will you need to sample different localities or temporal windows?
    - o What type of genetic data will you need to generate for this investigation? Do they require different laboratory steps?
    - o How will these steps allow you to measure the dependent and independent variables needed to test the hypothesis you proposed?


Sample collection
(modern and historical)
Nucleic acid isolation
Prepare samples for sequencing
(including quality control)
Sequencing method and platform
Bioinformatic analysis

Part 3: Risk Analysis and Budget Planning

10. What obstacles could arise when working with museum specimens compared to those recently collected in the field? Why do museum specimens pose a challenge? What is the likelihood of these problems occurring? List some strategies to overcome or avoid such obstacles.

Risk Analysis

Research phase	**Obstacle**	**Impact of obstacle**	**Likelihood of occurring** (low, medium, or high)	**Strategies to avoid or overcome obstacle**
Sample Collection				
Lab Work				
Data Analysis				

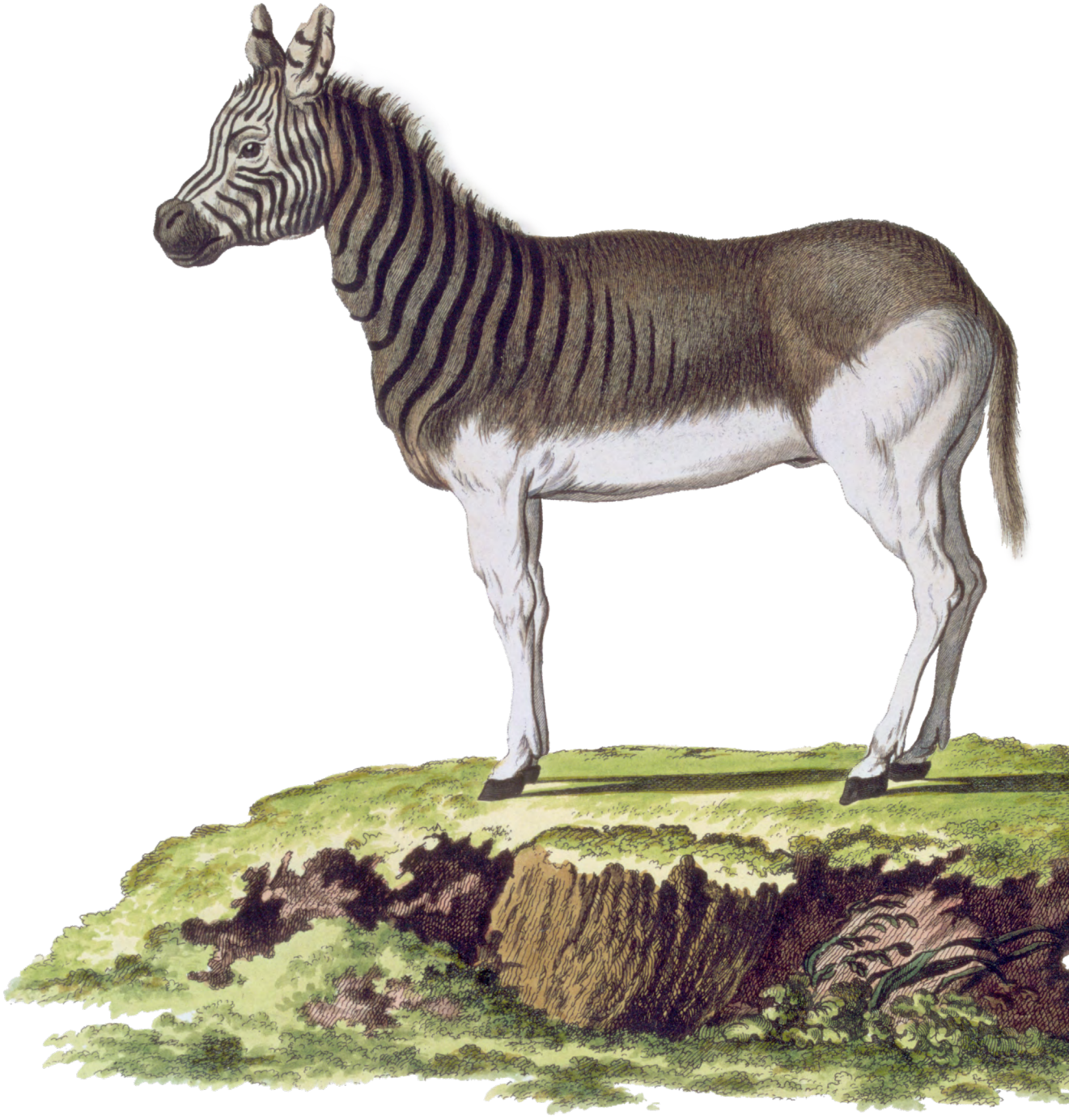
11. Which steps of this study design do you expect to require the most funding? Think about how you would acquire the samples and determine the sequencing strategy and desired sequencing coverage. For instance, while capture sequencing can help you increase the coverage of a target genomic region, the cost of developing and purchasing baits can be very expensive. Consequently, although the capture strategy can increase the coverage of certain regions, it might restrict the number of samples you can sequence. Fixed monetary sources for your research can impact how many samples you can collect, process, and sequence, so how would you optimize your budget to best accomplish each major phase of your study design?

Budget Planning

Research phase	**Obstacle**	**Impact of obstacle**	**Likelihood of occurring** (low, medium, or high)	**Strategies to avoid or overcome obstacle**
Sample Collection				
Lab Work				
Data Analysis				

12. If the data from your study suggests significant change between historical and modern populations, think about how your findings could be applied to conservation management. How would you best communicate your results if you wanted to share them with management authorities? What are some tools and strategies that you could suggest implementing to aid modern populations in your hypothetical case study? An example conservation strategy is wildlife crossings, which can connect two disparate populations that have been separated by the construction of a highway.

We kindly thank the authors and reviewers for their time, effort, and dedication to the materials provided in this issue of Lessons in Conservation.



We welcome your comments and feedback. To write to NCEP or for more information, contact the Network of Conservation Educators and Practitioners at:

American Museum of Natural History
Center for Biodiversity and Conservation
200 Central Park West
New York, New York 10024
ncep@amnh.org

Lessons in Conservation
is available electronically at
ncep.amnh.org/linc

Adaptable Microsoft Word versions of these modules are available for download from the NCEP module collection at ncep.amnh.org along with accompanying presentations, teaching notes, exercise solutions, supplementary files, and links to other relevant open educational resources.