




ARTICLE



<https://doi.org/10.1057/s41599-022-01141-4>

OPEN

Challenges in large-scale bioinformatics projects

Sarah Morrison-Smith ¹✉, Christina Boucher², Aleksandra Sarcevic³, Noelle Noyes⁴, Catherine O'Brien¹, Nazaret Cuadros¹ & Jaime Ruiz²

Biological and biomedical research is increasingly conducted in large, interdisciplinary collaborations to address problems with significant societal impact, such as reducing antibiotic resistance, identifying disease sub-types, and identifying genes that control for drought tolerance in plants. Many of these projects are data driven and involve the collection and analysis of biological data at a large-scale. As a result, life-science projects, which are frequently diverse, large and geographically dispersed, have created unique challenges for collaboration and training. We examine the communication and collaboration challenges in multidisciplinary research through an interview study with 20 life-science researchers. Our results show that both the inclusion of multiple disciplines and differences in work culture influence collaboration in life science. Using these results, we discuss opportunities and implications for designing solutions to better support collaborative tasks and workflows of life scientists. In particular, we show that life science research is increasingly conducted in large, multi-institutional collaborations, and these large groups rely on “mutual respect” and collaboration. However, we found that the interdisciplinary nature of these projects cause technical language barriers and differences in methodology affect trust. We use these findings to guide our recommendations for technology to support life science. We also present recommendations for life science research training programs and note the necessity for incorporating training in project management, multiple language, and discipline culture.

¹Barnard College, New York, NY, USA. ²University of Florida, Gainesville, FL, USA. ³Drexel University, Philadelphia, PA, USA. ⁴University of Minnesota, Minneapolis, MN, USA. ✉email: smorriso@barnard.edu

Introduction

Life-science research—which encapsulates a wide range of biological and biomedical research—is responsible for many major scientific findings in the past decade. Such findings include the completion of a near perfect error-free human genome (Miga, 2020), the rapid sequencing and assembly of the SARS-COVID19 genome (Fernandes, 2020), the identification of nearly 70,000 extant vertebrate species (Rhie, 2021), and the definition of Parkinson's disease sub-types for efficient therapeutic development (Kenneth, 2018)—just to name a few. One of the main drivers of these, and countless other discoveries in life science has been the development of technologies that are capable of identifying the DNA, RNA or amino acid sequence corresponding to a biological sample. One of the first technologies—so called *first generation sequencing technologies*—was capable of sequencing a couple of thousand nucleotides of a DNA sample per day. This technology was fundamental to the sequencing and assembly of the first human genome (International Human Genome Sequencing Consortium, 2001; Venter, 2001). Since then sequencing technologies have increased their throughput (meaning the amount of DNA they can sequence at a given time), decreased their cost, and became highly accurate (Bansal and Boucher, 2019; Lang, 2020). Currently, it is now feasible to sequence an entire human genome in less than day and for less than one thousand dollars (Giani et al., 2020).

Sequencing technologies take as input a biological sample and produce a sequence of signals that are then interpreted to produce the string of DNA, RNA or amino acid. In the case of next generation sequencing, the signal produced is fluorescent light that can be viewed under a microscope that can be translated into a nucleotide (A, C, G, or T) sequence, and in the case of mass spectrometer there is ion/mass charge that is interpreted to produce an amino acid sequence corresponding to a peptide. All of these technologies have been dramatically advanced over the past couple decades, and other laboratory methods (such as optical mapping (Mukherjee, 2018)) have been automated (Giani et al., 2020). Even though one of the drivers of this advancements has been human genetics, very few other life science areas remain untouched by the discovery and advancement of sequencing technologies. Plant biology (Waese, 2017), soil sciences (Bajpai et al., 2021), species evolution (Funk et al., 2018; i5K Consortium, 2013; Luikart et al., 2018; Rhie, 2021), extinction of animal species (Humble, 2020; Shapiro, 2017), and creation of synthetic species have all been advanced due to the creation of these technologies.

What is underlying these technologies is the analysis of data, which frequently requires more time than the generation of the data itself. Pollack (2011) states that “The field of genomics is caught in a data deluge. DNA sequencing is becoming faster and cheaper at a pace far outstripping Moore's law, which describes the rate at which computing gets faster and cheaper. The result is that the ability to determine DNA sequences is starting to outrun the ability of researchers to store, transmit and especially to analyze the data.” Prior research (Morrison-Smith et al., 2015) showed that in order to overcome the challenges of analyzing data, life science researchers collaborate with researchers and trainees in different disciplines, locations, and institutions. Yet earlier research on interdisciplinary science showed that scientific projects that depend on a large number of institutions and disciplines are less successful than those relying on fewer (Cummins and Kiesler, 2005; Kiesler and Cummings, 2002); here, success was defined to include metrics such as graduate and post-graduate supervision, the number of related projects, the frequency of project meetings, and the likelihood of having created a project-related course. These findings are compounded by prior work showing that the development of communication technology has negatively impacted projects by hindering information

sharing (Hinds and Mortensen, 2005), delaying outcomes (Espinosa, 2004), and causing misunderstandings (Cramton, 2001). Because technology has advanced substantially since these prior works, it is natural to ask whether advances in the development of new technology and refinement of existing tools have circumvented these problems of coordination—and if not, what issues remain.

This paper focuses on the challenges of interdisciplinary collaboration in life science. We aim to uncover how researchers from various backgrounds and expertise communicate to perform data analysis from the perspective of life scientists that generate data as a means towards a scientific goal. Follow-up work could reverse this perspective to understand the challenges faced by the researchers in computer science, statistics and data analysis. In light of this focus, we conducted semi-structured interviews with life science researchers from nine research institutions. We performed a bottom-up analysis by constructing an affinity diagram to identify themes related to our goals, which we then correlated to themes from prior work. Our results show that both interdisciplinarity and differences in work culture and practices affect collaboration in life science. We contextualize our result in current research and hence, show that many of the challenges identified in prior work (Olson and Olson, 2000, 2006) persist today in spite of technological advancements. We conclude by offering new perspectives insights for interface and software development, and providing recommendations for training programs to better prepare trainees for collaboration in life science research.

Related work

Here, we examine studies that have identified the factors influencing scientific collaboration and training.

Collaboration has been extensively studied over the past several decades from a variety of perspectives including the sciences (Armenteras, 2021; Olson and Olson, 2000, 2006) and the humanities (Balestrini et al., 2021; Canfield, 2020; Cooke et al., 2017). This research has uncovered challenges resulting from lack of adequate support for collaboration across geographical, institutional, and disciplinary boundaries (Jirotko et al., 2006). Life-science research is commonly multi-institutional, and therefore may face challenges related to remote work including, but not limited to, the lack of the motivational sense of the presence of others (Olson and Olson, 2006), difficulty establishing and maintaining trust (McDonough et al., 2001; Olson and Olson, 2006; Sarker et al., 2011), increased intra-team conflict due to us-vs-them attitudes (Armstrong and Cole, 2002; Cramton, 2001), and coordination difficulties caused by reduced number of overlapping work hours between collaboration sites (Battin et al., 2001; Casey and Richardson, 2004; Kiel, 2003). In addition, due to the interdisciplinary nature of these collaborations, it is possible that life scientists also face challenges related to group composition due to lack of common ground (Cundill, 2019; Maynard and Gilson, 2014), socio-cultural distance (Hinds and Bailey, 2003; Mortensen and Hinds, 2001; Swigger et al., 2004), and differences in work culture (Cundill, 2019). Each of these challenges has been extensively discussed in recent work exploring the challenges associated with remote work by Morrison-Smith and Ruiz (2020). However, despite decades of research focusing on collaboration challenges, recent work has revealed that life scientists are still encountering problems when collaborating (Morrison-Smith et al., 2015). This indicates that there is a clear need to further investigate the collaboration challenges faced by life science researchers.

A number of prior works have focused on the importance of transdisciplinary and interdisciplinary training in a variety of

contexts including, but not limited to social work (Kemp and Nurius, 2015), medicine (Nash, 2008), and multidisciplinary research in general (Stokols, 2013). However, despite what we know about collaboration, there has been a lack of research focusing on collaboration-based training in the life sciences. There is an abundance of research focusing on supplementing life science education by training life scientists to program (Goodman and Dekhtyar, 2014; Mangul et al., 2017; Mariano et al., 2019; Qin, 2009), use bioinformatics software (Attwood et al., 2017; Miskowski et al., 2007; Ranganathan, 2005), and conduct data science (Emery et al., 2021) in order to carry out their research. However, despite calls for increasing multidisciplinary training in life science (Cech et al., 2000; The National Research Council, 2000) research exploring training life scientists to work and communicate in interdisciplinary collaborations has been much more limited.

In 2008, Stokols et al. (2008) identified that models to guide the development of transdisciplinary training curricula remain to be developed and tested. Since then, Misra et al. (2011) has identified institutional factors that facilitate transdisciplinary training in health research. In the process, they recommended that transdisciplinary training strengthen individuals' communication skills that build and sustain cooperation among team members, management strategies for resolving interpersonal conflict, and foster the ability to reach a consensus regarding research goals and visions to reduce task-related uncertainty. However, these recommendations appear highly general and lack insight into how they could be integrated into training specific to life science. Additionally, Sturmer et al. (2017) recently implemented a one-hour professional development course aimed at undergraduate students participating in the National Institute for Mathematical and Biological Synthesis Summer Research Experience that focused on developing collaboration skills. They identified that among other considerations, communication, setting concrete goals, and developing a shared mental model were the top three most important skills that students identified they needed to foster in order to facilitate teamwork. However, further research is required to identify methods for training these skills in a manner that prepares students for life science research. Thus, the question of how to prepare trainees for collaboration in life science research remains open.

Furthermore despite what we know about collaboration, there has been a lack of research focusing on collaboration-based training in the life sciences. There is an abundance of research focusing on supplementing life science education by training life scientists to program (Mangul et al., 2017; Mariano et al., 2019; Qin, 2009), use bioinformatics software (Attwood et al., 2017; Miskowski et al., 2007; Ranganathan, 2005), and conduct data science (Emery et al., 2021) in order to carry out their research. However, despite calls for increasing multidisciplinary training in life science (Cech et al., 2000; The National Research Council, 2000) research exploring training life scientists to work and communicate in interdisciplinary collaborations has been much more limited.

Methods

In this section, we describe the participant recruitment, their demographics and workflow, the data collection, and the data analysis. All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the University of Florida Institutional Review Board (IRB201602178).

Table 1 Atomized description of the 20 participants. Their identifier, scientific discipline, and academic position are given.

ID	Research area	Title	Location
P1	Epidemiology	Adjunct Faculty	Canada
P2	Animal Sciences	Faculty	United States
P3	Epidemiology	Post-Doc	United States
P4	Microbiology	Faculty	Australia
P5	Plant Biology	Faculty	United States
P6	Epidemiology	Faculty	United States
P7	Biology	Faculty	United States
P8	Bioinformatics	Faculty	United States
P9	Animal Sciences	Faculty	United States
P10	Immunology	Post-Doc	United States
P11	Industrial Hygiene	Post-Doc	United States
P12	Veterinarian	Faculty	United States
P13	Biology	Faculty	United States
P14	Epidemiology	Faculty	United States
P15	Proteomics and Metabolomics	Faculty Lab Director	United States
P16	Agriculture and Food Systems	Faculty	United States
P17	Plant Biology	Faculty	United States
P18	Plant Biology	Post-Doc	China
P19	Animal Sciences	Faculty	United States
P20	Animal Sciences	Faculty/ Chair	United States

Participants. A total of 20 life science researchers aged 28 to 64 (mean = 41.26, standard deviation = 9.64, 10 female) from nine universities and research institutions located in the United States, Canada, Australia, and China participated in this study. We used data saturation (Bonde, 2013) to establish our sample size, i.e., data collection terminated once further sessions resulted in minimal new information. We recruited participants via science discussions on Reddit (Reddit.com, 2017) and email. Each participant was in life science (Table 1). Their current scientific projects varied significantly by size, ranging from two collaborators to over 50. We recruited several participants (P1-9) from an existing collaborative project; whereas the remaining were independent of each other.

Our participants communicated that they seek to answer a variety of scientific questions that include the following:

- Immunology: prevention, mitigation, and control of infectious diseases affecting humans and animals.
- Animal sciences: improving public food safety associated with consumption of meat products, or consumer demand and satisfaction with these products.
- Plant biology: understanding and mitigating the physiological, genetic and biological effects of environmental stressors on plants.

This is not an exhaustive list but rather gives some specific examples of the breadth of research that our participants aimed to advance. Although our participants are seeking to answer a diverse set of research questions, we found that their projects followed very similar workflows, which can be broadly defined as collection, extraction, preparation, and sequencing of biological material. The resulting data are then transferred to the participants' computer or server and analyzed.

During the collection phase, cells or other biological material of interest are acquired naturally or purchased from a scientific vendor. Upon arrival in the lab, an experimental protocol is performed. After the experiment, the biological material (DNA or RNA) is extracted from the organisms using laboratory methods. These samples are first prepared for sequencing and then sent to

an off-site sequencing facility, where each sample is provided as input to sequencing machines. The output of the sequencing phase is a specialized text file (e.g., fastq), which is then transferred from the sequencing facility back to the participant's computer via SFTP, Windows Remote Desktop, or, when data is too large to easily transmit, a portable hard drive. The sequencing facility can be extremely geographically distant (e.g., the Beijing Genome Institute in China) or near (e.g., at the same institution). Data analysis frequently consists of some combination of bioinformatics and statistics, and may result in scientific conclusions or suggestions for follow-up studies. These steps may be done by multiple scientists or institutions, since specific expertise is needed for each step.

Data collection. We collected data through a series of semi-structured interviews focusing on the identified research questions, allowing us to cover additional topics as they occurred, and thus lessening the probability that important issues would be missed (Lazar et al., 2010). We interviewed local participants at their primary workspace (office or lab), and interviewed the remaining participants via Skype or over the phone. The interviews were ~30 to 60 min in duration and were recorded in audio format, then later transcribed verbatim. Our protocol was approved by our Institutional Review Board (IRB201602178). Participation in interviews was voluntary and participants did not receive compensation. Data is publicly available at <https://anonymous.4open.science/r/collaboration-data-sharing-data-7F0E>.

Data analysis. We performed a bottom-up analysis of participants' responses by constructing an affinity diagram—which is also known as the KJ method (Beyer and Holtzblatt, 1998; Subramonyam et al., 2019)—to expose prevailing themes in the scientists' research goals and work practices. This approach follows methodology for qualitative analysis via coding as outlined by Auerbach and Silverstein (2003). Unlike in qualitative coding, however, instead of each researcher independently organizing data followed by calculating the group's inter-rater reliability, a quantitative measure, the five researchers analyzing the data came to a consensus on all responses. This is appropriate for semi-structured interviews as qualitative coding results in the possibility of applying the same code to different sections of the interview (Jun et al., 2018). We then examined themes from prior work, which enhanced our interpretation of the interview data and allowed us to draw comparisons between our findings and prior knowledge, highlighting new discoveries. This insight facilitated the recommendations for design and training.

Results

We found a clear relationship between the size, distribution, and number of varied disciplines included in a team. Our participants expressed that this was primarily due to the need for expertise and resources—it is unlikely that one researcher has all the necessary expertise to answer a research question with a significant societal or scientific impact.

I have limited abilities, there are some things I know how to do and a bunch of things I don't. A lot of the collaborations have addressed scientific questions that I would have otherwise not have been able to do with my skill set. (P16)

This need for expertise was one of the primary drivers in dictating the size, geography and disciplines needed of a project. Here, we summarize our main findings that largely stem from the interdisciplinary nature of the projects. Table 2 highlights data supporting each key finding.

Communication barriers reduce project efficiency. Team members who have shared experiences (e.g., share a common vocabulary) have fewer difficulties collaborating remotely (Olson and Olson, 2000). This phenomenon is supported by our data in that 17 of our participants with different backgrounds, i.e., less common ground, encountered language barriers that inhibited collaboration by requiring extra effort to successfully discuss project goals and tasks. Hence, we found our participants encountered technical language barriers because they lacked the background necessary to understand all portions of the project. These language barriers affected the ability of researchers to communicate about project goals and individual tasks, making it difficult to understand how their research fit into the whole. The predominance of jargon and the tendency for some words to have different meanings in the context of other scientific fields further exacerbated these difficulties. To address these challenges, our participants attempted to mimic the language of their collaborator(s) by describing their methodology at a high level, which they thought was easier for their collaborator(s) to understand.

Sometimes it's taking something that is complicated and explaining so it's understandable. That's difficult. You have to speak in terms of their language. (P7)

Aside from language, twelve of our participants indicated that interdisciplinary and multi-institutional teams experienced differences in scientific methodology or standards. Moreover, these disparities had a significant impact on the project by requiring that the participants have additional discussions to come to a consensus regarding protocol and, in some cases, redo aspects of the project using different techniques. In some projects, experiments cannot be redone, in which case the participants felt they must reformat the data or "work with what [they] get" (P7). In a worst-case scenario, improper technique can make the data unusable:

She's probably spent 50 to 100 thousand dollars on sequencing and has nothing to show for it simply because proper controls weren't done. (P8)

Lastly, since participants needed to provide a substantial amount of background when discussing almost any aspects of a project, they felt that the processes of initiating and participating in this dialog can be time-consuming and inefficient to achieving the project goals. Thus, the participants often felt that the only solution was to trust that their collaborator(s) knew what they were doing for their part of the project and vice versa.

Mutual respect and trust is necessary for project engagement.

The interdisciplinary projects our participants engaged in often also resulted in different methodological approaches to science, which could influence the participants' ability to trust their collaborator(s). In turn, this sometimes led to feelings of mistrust in a collaborator's competence, impeding progress when previous portions of the project are redone or verified. This is expressed by P7 when they question the methodology, leading to perceptions of quality of work, and translating into mistrust regarding the quality of a collaborator's output or data:

It's hard because if you're receiving samples from them, or you're receiving a protocol, or they're sharing information, the way they would have done it, the method or technique, is much different from what you did. And there might be some disparities there, or I might have even an issue with how it was done and the quality of how it was done. (P7)

Table 2 Overview of major findings with example quotes from each challenge.

Challenge	# Ps	Example quotes
Communication barriers reduce project efficiency	17	"I think it can be difficult to explain ideas, and it can also be difficult to explain technical details just because, you know, either side doesn't have the complete expertize of what the other person is more familiar with. However usually, you know, it gets hashed out through enough communicating"—P5 "There is sometimes differences in vernacular and language and technical language that's used to discuss things. And so you have to have constant communication so that everybody understands"—P7
Disparities in methodology or standards require additional communication	12	"The quality of the data that was generated by the collaborator was really poor and it's not going to be able to be published. So because they specialize in a particular technique to generate that data, I'm not going to be able to go back and redo it or find somebody else to do it."—P5 "If you're receiving samples from them or you're receiving a protocol or they're sharing information the way they would have done it, the method or technique is much different than how you did. And there might be some disparities there, or I might have even an issue with how it was done in the quality of how it was done."—P7
Mutual respect and trust is necessary for project engagement	14	"It's about the respect. It's not the size. It's not it's not how small it is, how big it is. It's how respectful all of them are to each other and the greater group. If you have a thousand people and they all respect the view of the other, it'll go smoothly. If you have two people who can't decide on what day it is, it's going to be a fight the whole way. So it's about mutual respect."—P8 "We work well together and we really respect each other's thoughts and opinions and each other's sort of areas."—P1
Large, distributed teams lead to reduced engagement	8	"I would usually sit there with my phone on mute, doing something else waiting for my ears to perk up for something that was like what's going on with this. So it was probably a fair amount of wasted of time. The fraction of the call that was important to the average individual, the call was quite small."—P16 "So many people involved. So people that can't really contribute, that aren't really engaged."—P19
Perceived priorities are difficult to gauge	8	"One of the one of the biggest difficulties is having collaborators who are really busy and they have multiple other projects going on. And your collaborative project may not be their highest priority or even if it is their highest priority, they still have to sort of split their effort amongst multiple projects."—P13 "It's like I can't directly tell how they put our collaboration project into the priority."—P1

Ultimately, fourteen of our participants felt that the projects where there was a significant amount of *"mutual respect"* (P8,P17) were more successful than those where some team members felt under-appreciated and under-prioritized. Interestingly, we found instances where the work culture shifted from collaborative to competitive—namely when interdisciplinary teams increased expertize in a specific area, this frequently lead to territorial issues. Our participants, such as P3, expressed that the addition of collaborators in a project can pose challenges if the expertize overlaps because there is potential for territorial actions that foster animosity and jeopardize the project's success (e.g., competition for funding sources):

Because you are working in the same field and you are doing the same stuff, there is more potential for territorial actions versus when you are working with people totally outside, they have their own funding sources. (P3)

The outcome of mistrust or lack of mutual respect within the project was frequently interpersonal conflict leading to detachment or disconnection of the project. This detachment, in turn, can cause researchers to drop out of the project, leading to increased costs in terms of time and funding to complete the project. In some cases, interpersonal conflict could interfere with the publication of a completed project.

People creating conflict. If they were sort of ad-hoc projects where there were not a lot of funding in place, then they would tend to fade away you know, evaporate.... If there

was funding on the line and the project was centered here, I end up having to pick up the pieces for what they were doing. And it usually took longer. Sometimes we had to find additional funding because of those failures. Sometimes, and almost inevitably put publication of the research in jeopardy. And some of those projects, well they may have been completed, but they've never been published. (P6)

Large, distributed teams lead to reduced engagement. We found that eight of our participants were more likely to feel disconnected when collaborating in large distributed groups than smaller ones. These participants felt that large group meetings are a *"waste of time"* (P16) due to both inefficiency and a tendency for conversation topics to interest only a few collaborators at a time. This lack of interest is partially due to the structure of the projects consisting of multiple phases, each of which is more interesting to some researchers than others.

There were little bits that were sometimes important to me that I had to share but I would usually sit there with my phone on mute, doing something else waiting for my ears to perk up for something that was like what's going on with this. So it was probably a fair amount of wasted of time. The fraction of the call that was important to the average individual, was quite small. (P16)

In these situations, our participants viewed discussions about a phase that they are not particularly concerned with as going into too much unnecessary detail, and thus lose interest. They

described this issue as being especially prevalent in large, interdisciplinary projects where collaborators are perceived to be apathetic or uninterested in the details of the project that are not directly related to their field.

It is pretty easy for some researchers to drop out of the mix in a large group and that can ultimately be, you know, a herald of death for the whole thing, right? (P12)

One participant (P19) observed that enthusiasm varies throughout the life of the project, depending on how much the researcher cares about that stage. They observed that researchers are typically the most excited at the beginning of the project and later are interested in the results, but lose focus during the middle. This fluctuation in interest levels has important implications for design of training programs to ensure the engagement of the trainees.

Perceived priorities are difficult to gauge. We identified differences in priorities—or perceived priorities—of collaborators as a challenge associated with interpersonal dynamics that especially affected six of our life scientist participants. Researchers are frequently involved with multiple scientific projects, and, thus, they prioritize their efforts.

One of the biggest difficulties is having collaborators who are really busy and they have multiple projects going on and your collaborative project may not be their highest priority. (P13)

Our participants found it frequently difficult to tell whether a collaborator is prioritizing the project and thus, when they are not “*pulling their weight*” (P17). This is frequently manifested when lack of informal interactions led to doubts about their collaborators’ prioritization of the project, creating concern that a vital part of the project would not be completed. Hence, this sense of project involvement is particularly important when researchers are concerned that improper prioritization jeopardizes the project’s timeline. In these situations, our participants sometimes used email to gauge their interest—a collaborator who responds quickly to email is more likely to be prioritizing the project than one who takes weeks or even months to respond.

It can be really hard to tell where they put our collaboration project into priority, but you can tell from email comment—sometimes you can tell they’re working and I get email back really soon, but sometimes it’s like after a couple of weeks may be months then I get a response. (P18)

In some cases, collaborators are even encouraged to drop out of a project, sometimes with a recommendation for a new collaborator, if they are unable to complete their portion in a timely manner. Hence, resolving this issue is complicated and frequently requires additional work on the side of the researcher who is waiting. This issue sometimes caused researchers to make extra work for themselves—such as doing everything in their power to make it more convenient for their collaborators to complete their part of the project.

First I do what I can do on my side and then try to make everything easier and convenient for my collaborators (P18)

Discussion

Over the course of our investigation into the challenges faced by life science researchers, we identified several key issues that

specifically affect life science collaborations: “mutual respect” is important; interdisciplinary causes technical language barriers; differences in methodology affect trust; and perception of a collaborator’s priorities can cause unnecessary work. These key findings demonstrate that collaboration challenges are still impacting life science, despite years of collaboration research. In this section, we discuss the implications of our results for the design of technology and for training in life science.

Implications for design of technology

Support documentation and discussion of scientific knowledge. Our results show that involving various disciplines creates language barriers that delay life science projects. Despite this effect, we also find that multidisciplinary collaborations are likely to continue to be the norm for these types of projects, given their potential to assist in answering broad scientific questions. Therefore, new tools are needed to lower the language barriers and support discussions around scientific knowledge that go beyond current communication tools that simply support communication (e.g., email, Zoom (Zoom Video Communications Inc., 2020), and Slack (Salesforce Inc., 2021c)) over geographical distance. We envision systems that enable teams to both document and discuss activities and methods throughout the project life cycle while providing tools and techniques to minimize language barriers. For example, we envision a specialized word processor similar to Microsoft Word Online (Microsoft Inc., 2021b) or Google Docs (Google Inc., 2021a) that enable remote collaborators to collaboratively work on a document and easily import output of computational programs. To minimize the language barriers, the system could provide easy lookup of specific terminology, method, and datasets through a context menu (i.e., right clicking or hovering over a specific text). This would enable a researcher unfamiliar with a term or method to look-up more information. The context menu can also enable easy linking to shared datasets. These methods not only help minimize language barriers, which will promote more discussion across all members of the team, but also provide more transparency on the approach individual team members are taking so that their is less concern around differences in research methodologies.

Support creation of collaborations with mutual respect and trust. Participants frequently stated that mutual respect and trust were necessary for project engagement and success. However, the need to find expertise in an specific area often results in the creation of a team where trust has not been previously established. To avoid these issues researchers often constrain teams to their current network of collaborators and when forced to reach out rely on the collaborators of trusted collaborators. For example, one participant stated:

So they had collaborated with my advisor before and so they needed some of that specific skill. So they called us for other grants. It’s usually because the people they’re like, first of all, we know them. We know we can work with them. (P3)

This suggest the need of to capitalize systems that focus on social networking to create a space that enable researchers to precisely specify their expertise and specify their past and current collaborators. This would create a social network where researchers who are starting a new project can find new collaborators by searching for a particular expertise, and by determining how the potential collaborators relate to their prior collaborations. The ability to know how the new collaborator fits

within their research network, new teams would have more common collaborators, which would result in higher initial trust and respect. In addition, the network should also enable to see how potential collaborators' expertise may or may not overlap with the current teams expertise. This is important as participants mentioned that too much overlapping of expertise can result in conflict and unhealthy internal competition.

Greater support for management of scientific projects. Participants from our study often commented on that since their collaborators often worked on several projects, they had a difficulty gauging status of their collaborators' tasks, which lead to them questioning their collaborators' priorities. Although software has been developed to assist in creating and running data analysis pipelines (e.g., Galaxy (Afgan et al., 2018)), there is a need for a project management system explicitly tailored to life science pipelines and life science project workflows. We envision a system developed to support a life science project life cycle similar to those that have been developed to support software engineering (e.g., agile development software). These systems are unique in that they not only would require task management, but also enable effective sharing of datasets, analysis pipelines, and project results. To support task management, these systems should enable to collaborators to create milestones and tasks, assign tasks, and specify the status of tasks. The system should also enable researchers to specify tasks that are dependent on one another. In this case, the system should notify members of a task chain notifying them that the dependent task has been completed and the new tasks may now begin. This would enable the researchers to focus on doing science instead of managing the handoff of tasks. Together, these the proposed features would enable a shared sense of ownership of the project while also providing a better sense of each members progress toward the shared goals. We envision that these systems could also provide support for automation. For example, a data analysis pipeline can be set up such that when new datasets are added to the shared project, the pipeline is automatically run with the new dataset and the results are stored and shared. The system would also allow researchers to view and visualize (when appropriate) all results of a pipeline to enable comparisons between pipeline executions and datasets.

Implications for training in life science

Life science requires multidisciplinary training. Despite the issues related to interdisciplinarity in life science collaborations, it is unlikely that future collaborations will consist only of collaborators from the same field due to the need for specific expertise and resources to answer broad research questions with significant societal or scientific impact. Furthermore, researchers appreciate the resulting increase in the range of scientific perspectives and potential for gaining new insight, making it more likely that multidisciplinary collaborations will be pursued. Addressing the challenges associated with the inclusion of multiple disciplines in a collaboration will therefore be crucial in life science training programs. Moreover, the participants were more engaged and had higher perception of their colleagues work when they were knowledgeable of that proponent of the project.

Therefore, one—perhaps most obvious—finding of our work is need for life science training programs to be multidisciplinary. Our participants felt that although they may not need to actively participate in all aspects of the project, they would be more engaged if they had knowledge of each proponent of the scientific process. Moreover, our finding suggest that this would lead to an increased appreciation of colleagues' contributions. Training programs also need to accommodate the discussion and negotiation process as methodologies and data sharing standards

evolve, rather than just facilitate a decision of which techniques should be used.

Training should break down language barriers via standardization. Cummings and Kiesler (Cummings and Kiesler, 2005) work found that projects incorporating multiple disciplines had as many positive outcomes as projects involving fewer. Our findings, however, indicate that projects with high-field heterogeneity face challenges in the form of language and background barriers, and thus align with the findings of Kiesler and Cummings (2002) that discipline influences project success. This finding is particularly problematic for life science since the bulk of the researchers' work is highly technical and requires conveying specific knowledge of terminology and methodology. While prior work has examined the challenges resulting from differences in project-related terminology (Morrison-Smith et al., 2015). For example, "test procedure" and "phase completion" are not necessarily analogous to the all scientific fields, and terms are frequently specific to a sub-discipline of life science, e.g, "contig" is used to define "a set of overlapping DNA segments that together represent a consensus region of DNA". Furthermore, the transfer of this knowledge is unavoidable since it is impossible for a single researcher to have all the expertise required to complete a project with high societal impact. Participating in additional dialogs to overcome these barriers can further slow the progression of the project. Training programs should provide mechanisms for facilitating, simplifying, and documenting these conversations. Documentation should be done in a manner that allows users to search for abstract representations of information, as discussed by Olson et al. (2008). One way to achieve this is through standardization of terminology, which implies instruction on contextualizing scientific terms.

Life science requires training in project management. The need for explicit management details how collaborations require additional management to overcome the challenges associated with being dispersed geographically or larger in size (Olson and Olson, 2006). We found several instances of this additional work, particularly when coordinating meetings (especially across multiple time zones and in different languages) and ensuring that everyone is up-to-date with the project status. While it is well documented that scientists do not like explicit management (Olson and Olson, 2006)—and our own findings suggest that scientists are not interested in micromanagement and rather prefer ad-hoc collaborations—several participants specifically acknowledged this need for explicit management by describing "a good PI" as someone who dedicates time to ensure that all portions of the project are "moving forward"(P19).

Our findings suggest that mechanisms for training in explicit management should be available, as they are necessary for large-scale project and may not encounter as much resistance as previously thought. For example, funding agencies could mandate or strongly suggest a formal management training program for the Co-PIs, so that they are better able to execute the type of explicit management that is required for large collaborative projects. In addition, our results imply that formal and explicit training in project management should also be available for students in life science. We recommend that universities and other training intuitions offer certificate programs in project management. The curriculum for these certificate programs could include formal pedagogy from the team science research literature, as well as experiential activities.

Work culture of disciplines needs to be incorporated into training programs for life science. Our results provide answers to the research question about the influence of work culture on life science research. While Walsh and Maloney (2007) asserted that

remote collaborations do not experience more challenges associated with culture than co-located teams, results from our study demonstrate that differences in work culture, particularly work practices regarding methodology and data sharing, profoundly affect collaboration in life science. Like McDonough et al. (2001), we found that differences in work practices and culture pose additional challenges in project management and coordination. For life science projects, differences in methodology can call into question the quality of completed work, resulting in delays caused by redoing procedures. Hence, our recommendation is that scientific programs in life science (bioinformatics being inclusive of that definition) take a short, required course in differences in related to discipline culture, which include methodology, data sharing, grant writing procedures, and determination of authorship and co-authorship.

Size of training programs in life sciences needs to be varied. Our study demonstrated that participants felt disconnected to the goals of a project in large groups. Our participants frequently felt this translated into “*mutual respect*” for each other’s contribution, and attributed this to being tied to the size of the research group. In addition, in large groups the participants they sometimes perceived their collaborators as being apathetic or uninterested in the details of the project. Thus, one recommendation that stems from our findings is to have training programs of varied sizes and disciplines; smaller discussion sections or research project groups will allow further engagement and understanding into the scientific progress.

Conclusion

While our study is limited by the sole use of interview data, we were able to elicit many new insights about the collaboration and training faced by life scientists. Future work includes supplementing our results with observations or journaling of work. It is clear that discipline, language, project management, engagement and work culture will vary in future scientific projects and, therefore, the creation of training programs that either overcomes the barriers of all of these factors remains imperative—regardless of their individual impact—or prepares life science researchers with the necessary skills to overcome these barriers.

We presented the results of semi-structured interviews that examined the challenges associated with collaborative life science research. We identified several key issues that specifically affect life science collaborations: “*mutual respect*” is important; inter-disciplinarity causes technical language barriers; differences in methodology affect trust; and perception of a collaborator’s priorities can cause unnecessary work. These key findings demonstrate that collaboration and training challenges are still impacting life science, despite years of collaboration research. Our contributions include (1) new insights into the communication and collaboration challenges that hinder life science research, particularly on the effects of discipline, work practices, and culture; and (2) recommendations for designing training programs to better support life science. Lastly, we note that identifying opportunities for the bioinformatics community to engage with life scientists to design tools that support collaboration and communication in this domain warrants future investigation.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 28 May 2021; Accepted: 21 March 2022;
Published online: 08 April 2022

References

- Afgan E, Baker D, Batut B, vandenBeek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning B, Guerler A, Hillman-Jackson J, Hiltmann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res* 46(W1 (may)):W537–W544. ISSN 0305-1048
- Armenteras D (2021) Guidelines for healthy global scientific collaborations. *Nat Ecol Evol*. 5(9):1193–1194
- Armstrong DJ, Cole P (2002) Managing distances and differences in geographically distributed work groups. In: *Distributed work*. The MIT Press, Cambridge, MA, USA. pp. 167–186
- Attwood TK, Blackford S, Brazas MD, Davies A, Schneider MV (2017) A global perspective on evolving bioinformatics and data science training needs. *Brief Bioinform* 20(2 (August)):398–404. ISSN 1477-4054
- Auerbach C, Silverstein LB (2003) *Qualitative data: an introduction to coding and analysis*, vol 21. NYU Press
- Bajpai R, Meher J, Rashid MM, Lingayat D (2021) *Metatranscriptomics: a recent advancement to explore and understand rhizosphere*. In: Nath M, Bhatt D, Bhargava P, Choudhary DK (eds.) *Microbial metatranscriptomics below-ground*. Springer
- Balestrini M, Kotsev A, Ponti M, Schade S (2021) Collaboration matters: capacity building, up-scaling, spreading, and sustainability in citizen-generated data projects. *Humanit Soc Sci Commun* 8(1):169
- Bansal V, Boucher C (2019) Sequencing technologies and analyses: Where have we been and where are we going? *iScience* 18:37–41
- Battin RD, Crocker R, Kreidler J, Subramanian K (2001) Leveraging resources in global software development. *IEEE Softw* 18(2):70–77
- Beyer H, Holtzblatt K (1998) *Contextual design: defining customer-centered systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA
- Bonde D (2013) *Qualitative interviews: when enough is enough*. Research by Design
- Canfield KN et al. (2020) Science communication demands a critical approach that centers inclusion, equity, and intersectionality. *Frontiers in Communication* 5:2
- Casey V, Richardson I (2004) Practical experience of virtual team software development. In: *Proc of European Software Process Improvement (Euro SPI)*
- Cech TR, Bond EC, Stevens J (2000) The role of the private sector in training the next generation of biomedical scientists. In: *Proc. of a conference sponsored by the American Cancer Society, the Burroughs Wellcome Fund, and the Howard Hughes Medical Institute*
- Cooke SJ, Gallagher AJ, Sopinka NM, Nguyen VM, Skubel RA, Hammerschlag N, Boon S, Young N, Danylchuk AJ (2017) Considerations for effective science communication. *FACETS* 2:233–248
- Cramton CD (2001) The mutual knowledge problem and its consequences for dispersed collaboration. *Organ Sci* 12(3):346–371
- Cummings JN, Kiesler S (2005) Collaborative research across disciplinary and organizational boundaries. *Soc Stud Sci* 35(5):703–722
- Cundill G et al. (2019) Large-scale transdisciplinary collaboration for adaptation research: Challenges and insights. *Glob Challenge* 3(4):1700132
- Emery N, Crispo E, Supp SR, Kerkhoff AJ, Farrell KJ, Bledsoe EK, O’Donnell KL, McCall AC, Aiello-Lammens M (2021) Training data: how can we best prepare instructors to teach data science in undergraduate biology and environmental science courses? Preprint at bioRxiv <https://doi.org/10.1101/2021.01.25.428169>
- Espinosa JA, Carmel E (2004) The effect of time separation on coordination costs in global software teams: a dyad model. In: *Proc. of the 37th Annual Hawaii International Conference on*, 10–pp
- Fernandes JD et al. (2020) The UCSC SARS-CoV-2 genome browser. *Nat Genet* 52:991–998
- Funk WC, Zamudio KR, Crawford AJ (2018) Advancing understanding of amphibian evolution, ecology, behavior, and conservation with massively parallel sequencing. In: Hohenlohe PA, Rajora OP (eds.) *Population genomics: wildlife. population genomics*. Springer
- Giani AM, Gallo GR, Gianfranceschi L, Formentic G (2020) Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput Struct Biotechnol* 18:9–19
- Goodman AL, Dekhtyar A (2014) Teaching bioinformatics in concert. *PLoS Comput Biol* 10(11):e1003896
- Google Inc. (2021a) Google docs. <https://docs.google.com>
- Hinds PJ, Bailey DE (2003) Out of sight, out of sync: understanding conflict in distributed teams. *Organ Sci* 14(6):615–632. ISSN 1047-7039
- Hinds PJ, Mortensen M (2005) Understanding conflict in geographically distributed teams: the moderating effects of shared identity, shared context, and spontaneous communication. *Organ Sci* 16(3):290–307
- Humble E et al. (2020) Chromosomal-level genome assembly of the scimitar-horned oryx: Insights into diversity and demography of a species extinct in the wild. *Mol Ecol Resour* 20(6):1668–1681

- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J Hered* 104(5):595–600
- Jirotko M, Procter R, Rodden T, Bowker GC (2006) Special issue: collaboration in e-research. *Comput Support Coop Work (CSCW)* 15(4):251–255
- Jun E, Jo BA, Oliveira N, Reinecke K (2018) Digestif: promoting science communication in online experiments. In: *Proc of the ACM on Human-Computer Interaction*, vol 2 (CSCW), pp. 1–26
- Kemp SP, Nurius PS (2015) Preparing emerging doctoral scholars for transdisciplinary research: a developmental approach. *J Teach Soc Work* 35(1-2):131–150
- Kenneth M et al. (2018) The Parkinsons progression markers initiative (PPMI)-establishing a PD biomarker cohort. *Ann Clin Transl Neurol* 5(12):1460–1477
- Kiel L (2003) Experiences in distributed development: a case study. In *Proc. of International Workshop on Global Software Development at ICSE*
- Kiesler S, Cummings JN (2002) What do we know about proximity and distance in work groups? A legacy of research. *Distrib Work* 1:57–80
- Lang D et al. (2020) Comparison of the two up-to-date sequencing technologies for genome assembly: HiFi reads of Pacific Biosciences Sequel II system and ultralong reads of Oxford Nanopore. *GigaScience* 9(12):giaa123
- Lazar J, Feng JH, Hochheiser H (2010) *Research methods in human-computer interaction*. Wiley Publishing
- Luikart G, Kardos M, Hand BK, Rajora OP, Aitken SN, Hohenlohe PA (2018) Population genomics: advancing understanding of nature. In *Population genomics*, Springer, pp. 3–79
- Mangul S, Martin LS, Hoffmann A, Pellegrini M, Eskin E (2017) Addressing the digital divide in contemporary biology: lessons from teaching UNIX. *Trend Biotechnol* 35(10):901–903
- Mariano D, Martins P, HeleneSantos L, de Melo-Minardi RC (2019) Introducing programming skills for life science students. *Biochem Mol Biol Educ* 47(3):288–295
- Maynard MT, Gilson LL (2014) The role of shared mental model development in understanding virtual team effectiveness. *Group Organ Manag* 39(1):3–32
- McDonough EF, Kahn KB, Barczaka G (2001) An investigation of the use of global, virtual, and colocated new product development teams. *J Prod Innov Manag* 18(2):110–120
- Microsoft Inc. (2021b) Microsoft word online. <https://office.live.com>
- Miga KH et al. (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature* 585:79–84
- Miskowski JA, Howard DR, Abler ML, Grunwald SK (2007) Design and implementation of an interdepartmental bioinformatics program across life science curricula. *Biochem Mole Biol Educ* 35(1):9–15
- Misra S, Stokols D, Hall K, Feng A (2011) Transdisciplinary training in health research: distinctive features and future directions. In: *Converging disciplines*. Springer, pp. 133–147
- Morrison-Smith S, Ruiz J (2020) Challenges and barriers in virtual teams: a literature review. *SN Appl Sci* 2(6):1096
- Morrison-Smith S, Boucher C, Bunt A, Ruiz J (2015) Elucidating the role and use of bioinformatics software in life science research. In: *Proceedings of the 2015 British HCI Conference*. ACM, pp. 230–238
- Mortensen M, Hinds PJ (2001) Conflict and shared identity in geographically distributed teams. *Int J Confl Manag* 12(3):212–238. ISSN 1044-4068
- Mukherjee K et al. (2018) Error correcting optical mapping data. *GigaScience* 7(6):giy061
- Nash JM (2008) Transdisciplinary training: key components and prerequisites for success. *Am J Prevent Med* 35(2):S133–S140
- Olson GM, Olson JS (2000) Distance matters. *Hum Comput Interact* 15(2):139–178
- Olson JS, Olson GM (2006) Bridging distance: empirical studies of distributed teams. *Hum Comput Interact Manage Inform Syst* 2:27–30
- Olson GM, Zimmerman A, Bos N (2008) *Scientific collaboration on the Internet*. The MIT Press
- Pollack A (2011) DNA sequencing caught in deluge of data. *New York Times*
- Qin H (2009) Teaching computational thinking through bioinformatics to biology students. In: *Proc. of the 40th ACM Technical Symposium on Computer Science Education (SIGCSE)*. pp. 188–191
- Ranganathan S (2005) Bioinformatics education—perspectives and challenges. *PLOS Comput Biol* 1(6 (nov)):e52
- Reddit.com. (2017) AskScience: Got Questions? Get Answers. <https://www.reddit.com/r/askscience/>
- Rhie A et al. (2021) Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592(7856):737–746. ISSN 1476-4687
- Salesforce Inc. (2021c) Slack. <https://slack.com>
- Sarker S, Ahuja M, Sarker S, Kirkeby S (2011) The role of communication and trust in global virtual teams: a social network perspective. *J Manag Inform Syst* 28(1):273–310
- Shapiro B (2017) Pathways to de-extinction: how close can we get to resurrection of an extinct species? *Funct Ecol* 31(5):996–1002
- Stokols D, Hall KL, Taylor BK, Moser RP (2008) The science of team science: overview of the field and introduction to the supplement. *Am J Prevent Med* 35(2):S77–S89
- Stokols D (2013) Training the next generation of transdisciplinary. In: O'Rourke M, Crowley S, Eigenbrode SD, Wulforst JD (eds.), *Enhancing communication & collaboration in interdisciplinary research*, ch. 4. Sage Publications
- Sturmer KK, Bishop P, Lenhart SM (2017) Developing collaboration skills in team undergraduate research experiences. *Primus* 27(3):370–388
- Subramonyam H, Drucker SM, Adar E (2019) Affinity lens: data-assisted affinity diagramming with augmented reality. In *Proc. of the 2019 CHI Conference on Human Factors in Computing Systems*. pp. 1–13
- Swigger K, Alpaslan F, Brazile R, Monticino M (2004) Effects of culture on computer-supported international collaborations. *Int J Hum Comput Stud* 60(3):365–380
- The National Research Council (2000) Addressing the nation's changing needs for biomedical and behavioral scientists. The National Research Council
- Venter JC et al. (2001) The sequence of the human genome. *Science* 291:1304–1351
- Waese J et al. (2017) ePlant: visualizing and exploring multiple levels of data for hypothesis generation in plant biology. *Plant Cell* 29(8):1806–1821
- Walsh JP, Maloney NG (2007) Collaboration structure, communication media, and problems in scientific work teams. *J Comput Mediat Commun* 12(2):712–732
- Zoom Video Communications Inc. (2020) Zoom for video, conferencing, and phones. <https://zoom.us/>

Acknowledgements

This work is partially supported by National Science Foundation Grant Award #IIS-2013998. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect these agencies' views. The authors would like to thank Isaac Wang, Courtney Sanchez, Megan Hofmann, Julia Chang, Ariel Goldman, Aditi Patil, Dipashreya Sur, Luiza Leschziner, and Christopher Dean for assisting in the data analysis and related works.

Competing interests

The authors declare no competing interests.

Ethical approval

As stated in section "Methods", all procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the University of Florida Institutional Review Board (IRB201602178).

Informed consent

As stated in section "Methods", this study was approved by the University of Florida Institutional Review Board (IRB201602178). As such, all participants provided informed consent before participating.

Additional information

Correspondence and requests for materials should be addressed to Sarah Morrison-Smith.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing,

adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022