Project Description: Recent developments in genomics and the implementation of new technologies in agriculture have enabled a new research horizon in the field (Harper et al., 2018, Morota et al., 2018). The quality and quantity of genomic data, including microbiome, gene expression, high density SNP markers, sequence data, among others, have great potential to improve both plant and animal science enterprises. However, with these new developments a new challenge has arisen: practitioners must be able to manage the large data sets that result from these technologies – a skill that has traditionally is not been taught during the training of agricultural scientists (Eisen, 2008).

The goal of the Seed Grant is to catalog the available resources and resource gaps in data science to support the Agricultural Genome to Phenome (AG2P) initiative and to outline solutions to fill the gaps. We will develop surveys to identify how aware students and researchers are of the available resources. Additionally, we will create an online repository linking to available training materials in both plant and animal agricultural data science. In this repository we aim to provide the community a unified access point to information about workshops, seminars, online and in-person classes, and course curricula for different career stages. We will prepare a white paper describing our findings based on the survey and the catalog, focusing on how to advance data science education in AG2P. We will fund a graduate student to carry out this work. This student will work in the laboratory of PI Fragomeni with additional support from our team of investigators. We will use the results of the seed grant as preliminary data to apply for a large educational project to develop solutions to the needs identified in this project.

The problem of training individuals with quantitative skills for agriculture is not new (Misztal, 2007), but the availability of massive high throughput data at ever lower costs has accentuated the issue (Naithani et al., 2019). To deal with the challenge, some universities are proposing novel courses and new minors(e.g., University of Illinois Computer Science + Animal Sciences, BS). Online tutorials, in person seminars and workshops, and online classes are also arising to meet these needs. To the best of our knowledge, there is no curated resource cataloging such courses and resources (Rice et al., 2019). We will identify the resources available for training in data science applied to AG2P, catalog these resources, and identify what is missing and what is necessary to increase engagement of the next generation of scientists in AG2P. Our goal is that, with follow-up funding, this can be an evolving resource that we update with new resources as they become available.

The proposal addresses education in AG2P by **cataloging available educational resources** and **identifying the missing training resources and best practices for sustainable delivery of educational resources to diverse communities.** The specific aims of the proposal are: **a)** develop and organize an online catalog presenting the available resources for data science in animal and plant agricultural genomics; **b)** annotate these resources with information including the target audience, when and how the resource is available, and what specific topics are covered; and **c)** to identify the gaps in training resources and opportunities to advance data science training in AG2P.

Preliminary data/experience: PI Fragomeni and Co-PI Dodd have developed a data science program to train graduate and undergraduate students using the R language. The topics covered are data organization and manipulation, concepts of data analysis, genomic and microbiome analysis applied to animal and food science. This program began in Spring 2019 at the University of Connecticut and has included 6 undergraduate and 4 graduate students each semester. Our activities have identified a need for training in genomic data science. However, the in-person small group approach is not a sustainable way to satisfy the needs of the community.

Co-PI Gondro has developed courses focused on computational genomics. These courses cover programming techniques for analysis of large genomic datasets and Artificial Intelligence applied to genomic data analysis. These courses were delivered at MSU from 2018 onwards and moved to online in 2020 with around 10 grad students in each offering. Gondro is currently developing an Ag AI online course for international students.

The Co-PI Young teaches at a small liberal arts HBCU (historically black colleges and universities). She infuses her plant biology and molecular biology courses with online bioinformatics tools. She is interested in assisting in the development of a comprehensive repository of data science that is applicable to her undergraduate students and researchers. She is also interested in developing stronger collaborations with R1 institutions in plant bioinformatics.

Collaborator Caporaso's lab leads the development of the widely used QIIME 2 microbiome bioinformatics platform, as well as extensive educational resources supporting that platform. The platform itself has been cited over 26,000 times since its initial publication in 2010. The team's most recent educational resources include the QIIME 2 YouTube channel (youtube.com/qiime2) which teaches fundamental concepts in microbiome bioinformatics. This content derives from the team's workshop series (https://workshops.qiime2.org) which has moved online in response to COVID-19. Caporaso has recently become involved in agricultural microbiome research and is interested in bringing tools and training resources to this community.

Plans for achieving aims: We will execute extensive research on the resources available for data science training in animal and plant agriculture. Such resources will include agriculture specific training, and more general resources that can be useful for this training (e.g., resources such as *Practical Computing for Biologists*: https://practicalcomputing.org/). We will tap into our professional networks, use search engine tools, and review pre-existing resources such as Software Carpentry and Coursera to identify relevant content. We will also inventory workshops offered at conferences such as the Plant and Animal Genome Conference (PAG). Finally, we will evaluate curricula and syllabi at major agricultural universities to evaluate what is being covered regularly, and what is still necessary to train the next generation of agricultural scientists.

Once our initial collation of training resources is complete, we will assess the target audience, subject matter, and availability of each identified resource. We will annotate the background necessary for each resource, such as prior training, required courses/disciplines, and the academic level that it is suitable for (i.e., lower/upper undergraduate, graduate). We will include in our database all the annotated activities with a description of recommended prior knowledge. This will be made public through the Internet (e.g., UConn's WordPress service https://aurora.uconn.edu/). We will advertise the product on social media and in relevant discussion forums. We intend to present our findings in PAG [San Diego, CA, 2022], which will increase the reach of our platform.

Concurrent with the research of available resources, we will develop multiple surveys regarding educational resources in AG2P. We will create surveys specific to undergraduate and graduate students, and faculty/professionals. Our goals are to learn what resources are desired and not available, what resources are used, and what level of awareness of resources the community has. We will use the survey's results to understand which training resources are successful and identify the gaps. We will compare the results of the survey with our catalog. We suspect that this comparison will emphasize challenges such as availability or inaccessible delivery methods, and which resources have been harder for agricultural researchers to discover.

We will compile all our results in a white paper to be submitted to a journal such as PLOS Genetics. This paper will include our experiences with the resource research, the survey

results, and what are the consistencies between the two. Based on our findings we will identify the mechanisms towards solutions to fill the gaps in AG2P education. Finally, we will work together on an education grant to be submitted to USDA (e.g., REEU), NSF or another federal research agency, using the preliminary results from this seed grant to justify our application, and enable our team to work towards addressing the identified gaps.

How the project will further the aims of the AG2PI and project evaluation: The ultimate goal of this project is to identify solutions for increasing training in AG2P data science, mainly focusing on multi-omics analysis. We will create an annotated public resource with the identified training resources at different levels. We will identify the audience for each available resource, and we will determine the required expertise necessary for each type of training. Moreover, we will compare different schools' curricula, and identify the gaps in education. Finally, we will use survey results to further identify gaps in AG2P data science training.

We will evaluate the success of our project based both on internal and external evaluations. Internal evaluations will consist of a team effort to evaluate our AG2P genomic data science database and compare it against surveys. External evaluation will be held by incorporating the Center for Excellence in Teaching and Learning from the University of Connecticut We will work together with the center to identify the gaps and mechanisms towards solutions, which will be addressed in a future NSF education plan.

Expected outcomes and deliverables. The results and outcomes from this Seed Grant will be compiled in a report to be submitted to the AG2PI seed grant chair within 90 days of completion of our project. We will create an online public and free repository of the resources available in education of genomic data science in AG2P. We will present our results in the PAG meeting 2022, and we will submit a white paper with our results to an open access journal. We will train a graduate student to perform this work. The student is a pre-existing member of the team, so no new hiring will need to be done to allow us to start work on this project. This student is already working on creating content for education in genomic data analysis.

This Seed Grant will support a diverse and multi-disciplinary team to work together in the diagnostics of the current issues. The research team will use the results of the current proposal to propose a large project to NSF or USDA so that we can fill the gaps. The goals of our future proposal will include mechanisms to broaden participation in AG2P and to develop workforce in the subject.

Qualifications of the project team. Our research team is diverse and cross-disciplinary which aligns with the AG2PI. PI Fragomeni is an animal geneticist with a research focus in genotype-by-environment interaction. He maintains a genomic data science program with undergraduate and graduate students from multiple backgrounds. Co-PI Dodd is a graduate student, with a bachelor's degrees in animal science and in applied mathematics. She has experience with data analysis and is developing tutorials for introducing and teaching data analysis concepts to undergraduate students.

Co-PI Dr. Gondro is a researcher on quantitative genetics and its applications for livestock genetic improvement. With a particular interest in genomic prediction, GWAS, imputation work, sensory panel projects, artificial intelligence applied to animal production, on-site field sequencing and high-performance computing for big data analyses.

Collaborator Dr. Caporaso is a microbiome researcher and software engineer who has led the development of the QIIME microbiome bioinformatics platform, and trained biologists on its use, for over a decade. He has extensive experience in training biologists to use advanced computational tools.

Co-PI Dr. Young has expertise in plant/crop tissue culture and gene editing. She teaches mainly underrepresented minority undergraduate students and utilizes existing bioinformatics database in her courses.

Co-PI Dr. Taxis (Kendrick) is a trained geneticist and educational researcher. She teaches undergraduate courses on genetics and disciplines of animal agriculture as well as a graduate course on teaching methods in animal sciences. Her discipline research program has focused on ways to test and lower bovine leukemia virus in dairy herds, and her educational research has focused on curricular scaffolding or design to aid in enhancing job-relevant knowledge and skills for students interested in agricultural and food sciences.

Proposal timeline. We will work together during a one-year period for this seed grant. We intend to evaluate the resources over a full academic year, starting in the Fall of 2021.

	Fall 2021		Spring 2022	
Activity	Weeks 1 to 7	Weeks 8 to 14	Weeks 1 to 7	Weeks 8 to 14
Identification/evaluation of available resources	X	X	X	
Surveys	X		X	
White paper submission				X
Report to AG2PI				X

How the project will engage the AG2P scientific communities and underrepresented groups. Lack of educational resources and experienced trainers with domain expertise is a major challenge in genomic data science education in AG2P. Researchers and educators promote training and workshops for their students, and some of those resources are freely available online. However, it is not possible or sustainable to train all students in all disciplines, in such a fast growing and multidisciplinary field. Our Seed Grant will identify which resources are needed, which resources are available, and what gaps persist. Our project will engage undergraduate and graduate students as well as researchers and other professionals to effectively generate an online catalog of available resources for each academic level and identify current gaps. In this way, the AG2P community will have a source of the available training, and a list of the content that is lacking. To address the gaps, we will continue to work together, possibly with an even broader group to apply for a multi-PI educational grant.

We will support underrepresented groups by two main approaches. In the first, we will address a broad population in AG2P stakeholders including students from non-traditional ag colleges and departments. We will identify the resources available for all groups and propose projects to fill in the necessary resources that do not exist. We will work to have all the necessary resources available without cost for students. Moreover, we will work on promoting online training, so geographical location will not limit the scope of our work. Finally, we will promote a train the trainer plan, after identifying the needs of the field. We will identify the mechanisms to solve the problems with this grant and use it as preliminary data for continuing the project in the future.

In summary, our project will work with a multi-disciplinary team on identifying which training and educational resources are available for genomics and multi-omics data science in animal and plant agriculture. We will categorize the available resources and identify the gaps. We will promote multiple surveys to identify the needs of educational training in AG2P. We will train a grad student to carry out the project goals under our advising. We will prepare a white paper to distribute our results, and we will prepare an online catalog with the identified resources. We will use findings from this Seed Grant to support a larger educational proposal.

References cited:

Eisen, E.J., 2008. Can we rescue an endangered species?.

Harper, L., Campbell, J., Cannon, E.K., Jung, S., Poelchau, M., Walls, R., Andorf, C., Arnaud, E., Berardini, T.Z., Birkett, C. and Cannon, S., 2018. AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture. *Database*, *2018*.

Misztal, I., 2007. Shortage of quantitative geneticists in animal breeding

Morota, G., Ventura, R.V., Silva, F.F., Koyama, M. and Fernando, S.C., 2018. Big data analytics and precision animal agriculture symposium: Machine learning and data mining advance predictive big data analysis in precision animal agriculture. *Journal of animal science*, *96*(4), pp.1540-1550.

Naithani, S., Gupta, P., Preece, J., Garg, P., Fraser, V., Padgitt-Cobb, L.K., Martin, M., Vining, K. and Jaiswal, P., 2019. Involving community in genes and pathway curation. *Database*, *2019*.

Rice, S., Fryer, E., Ghosh Jha, S., Malkovskiy, A., Meyer, H., Thomas, J., Weizbauer, R., Zhao, K., Birnbaum, K., Ehrhardt, D. and Wang, Z., 2020. First plant cell atlas workshop report. *Plant direct*, *4*(10), p.e00271.