AG2PI SEED GRANT PROPOSAL

Title of Proposal:

Developing standardized bioinformatics capacity across multiple agricultural species.

Lead PI (Name, Title, Affiliation(s), email)

Fiona McCarthy, Professor, University of Arizona

fionamcc@arizona.edu

Co-PI (Name, Title, Affiliation(s), email)

Stephanie McKay, Associate Professor, University of Vermont

Stephanie.McKay@uvm.edu

Pankaj Jaiswal, Professor, Oregon State University

jaiswalp@science.oregonstate.edu

Collaborator (Name, Title, Affiliation, email):

none

Grant Administrator:

Rachel Crookston

Manager, Grants and Contracts, University of Arizona

(520) 621-5980

crooksto@arizona.edu

Keywords:

Bioinformatics, workflows, software

Project description

1. Objectives/aims

Technological developments across the field of genomics have resulted in what is commonly referred to as the democratization of genomics [1,2]; that is, the combined decrease in sequencing costs and expansion in applications (transcriptomics, epigenetics, SNP data) have enabled scientists to apply genomics techniques to an increasingly broad range of species and situations. For example, successful genomics programs have produced high-quality genomes for a diverse range of agriculturally important species and associated resources, and utilizing these genomes is supported by USDA National Research Support Programs such as NRSP-8 Animal Genomes, NRSP-10 National Database Resources for Crop Genomics and the related AgBioData project. However, while we can rapidly produce more sequence-based data more cheaply than previously, *the bottleneck in applying these techniques now lies in the application of bioinformatics to make sense of the data produced. For most researchers, it is now easier to generate genomic data than it is to manage and analyze the resulting data.* Several initiatives have been used to develop accessible bioinformatics capacity (e.g., CyVerse [3], Galaxy [4], etc.) but these platforms still require more technical expertise than is commonly available to many scientists and tend to be software agnostic, making open-source software available without recommending or benchmarking specific workflows commonly used for agricultural genomics. Moreover, the agricultural genomics communities continue to amass multiple types of ‘omics and genetics data at an increasing rate *but there is limited re-use of this data*. The agricultural communities’ inability to repurpose existing data sets for gaining insight to novel questions represents a loss of value that is measured in the cost of data collection (and re-collection) and in the time and resources invested in collecting these data sets and to hinder the integration of genomics data sets for future genome to phenome analyses.

Furthermore, the existing siloes between plant and animal sciences adds an additional barrier for genomics analyses. While most bioinformatics workflows are designed to be species-independent, the reality is that plant and animal bioinformaticists frequently use different workflows but do not communicate with each other about these software choices and how they perform for non-model organisms. Providing opportunities for plant and animal bioinformaticists to discuss and compare workflows will help provide standardized workflows that can be used as benchmarks and for education and training. This approach follows the National Institute of Standards and Technology supported ‘Genome in a Bottle’ initiative which seeks to develop technical infrastructure (reference standards, reference methods, and reference data) for use in human genome benchmarking, software development, optimization, and education [5].

The overall goal of this proposal is to develop cross-kingdom collaborations and bioinformatics capacity to support benchmarking for new genomics approaches and integration of agricultural data sets. This overall goal is supported by two specific objectives:

Aim 1. Developing a collaborative network for plant and animal bioinformaticists. Our *rationale* is that a cross kingdom effort to build agricultural bioinformatics capacity will help researchers to more effectively analyze their own data sets and provide a foundation for benchmarking and data integration. As discussed in the AG2PI Thinking Big Conference, convergent or transdisciplinary science provides the opportunity to address complex problems that cannot be solved by research from a single discipline [6]. However, successful convergent science also requires developing a mindful collaboration of scientists who are prepared to dedicate the time and effort required to build an inclusive transdisciplinary team with a common purpose and vernacular.

Approach: We will recruit interested agricultural bioinformaticists who are specifically focused on analyzing genomic data sets by reaching out to the existing NRSP-8 and NRSP-10 groups, although we anticipate additional contacts extending from this initial group to develop an extended network. The initial in-person discussion will focus on developing a common purpose (e.g., discussion of common barriers and identifying commonly used genomics workflows of interest to the agricultural community). Additional discussions will be held via monthly online calls and will focus on specific analyses. We anticipate that as the group focuses on specific analyses, it is likely that additional members will be added to these focus groups. As topics are agreed upon, we will engage with undergraduate students interested in bioinformatics to complete literature searches identifying commonly used workflows used in agricultural sciences and published datasets associated with these workflows. We will also discuss options and interests for continued communication platforms once the funding period for this project is complete (e.g., a dedicated mailing list, slack channel, etc.).

Preliminary data: The personnel listed on this project have already established productive collaborations with both plant and animal science bioinformaticists, including developing, maintaining, and sharing their own analysis workflows and genomics data sets [7-12]. Moreover, Drs. McCarthy, McKay and Jaiswal are experienced at establishing and working collaboratively in interdisciplinary groups to provide resources for agricultural researchers [10-11, 13-14].

Aim 2. Developing accessible, scalable, bioinformatics workflows for agricultural researchers. Our *rationale* is that while there are many software workflows for analyzing genomics data sets, researchers are often selecting workflows based upon what they can get to work and have limited or no time for testing or comparisons of multiple software options. While we appreciate that we cannot mandate a single software option for any analysis, our goal is to provide detailed worked examples and protocols based upon publicly available data sets.

Approach: After discussion with plant and animal bioinformaticians we will establish a set of common genomic analyses (e.g., RNASeq against a reference genome, epigenetics, SNP identification) and – supported by review of recent agricultural literature - software most commonly used for doing these analyses. Members of the plant and animal bioinformatic network will be asked to identify senior undergraduate or graduate students who will be supported over the summer to develop a detailed protocol using public data sets chosen from both plant and animal agricultural species. Students and their mentors will be selected based upon a competitive application process, and we anticipate expanding this opportunity to the broader community to recruit students and mentors, as required. Review criteria will include the student's interest in agricultural research, the impact of hands-on training for their research and diversity of students and their institutes. Students will work in pairs to prepare detailed published documentation using either a plant or animal data set for each software workflow and will also make the workflow available as a container that can be installed by individuals or as bioinformatics platforms (e.g., CyVerse). Having students work in pairs promotes professional development and collaboration amongst students and their mentors, and builds cross-kingdom collaborations.

During the Fall semester additional students will be recruited to alpha-test these protocols with additional data sets, and we anticipate that the mentors will develop publications which outline the protocols and provide information about cross-comparisons of software used to analyze data sets. The workflows will be described in detail using protocols.io and where possible these protocols will be linked to micropublications that ensure datasets are publicly available and citable. In each case, students and mentors will be surveyed to provide information about their research experience.

Preliminary data: The senior personnel on this proposal are experienced at analyzing multiple types of genomics data [7-12] and routinely mentor undergraduate and graduate students. Moreover, Drs McCarthy, McKay and Jaiswal are experienced at establishing and working collaboratively in interdisciplinary groups to provide resources for agricultural researchers [10-11, 13-14]. We are experienced at developing biocontainers, installing workflows on CyVerse and using them in Nextflow applications to support scalable and reproducible bioinformatic analyses.

2. Furthering the aims of the AG2PI

The overall goal of this proposal is to develop cross-kingdom collaborations and bioinformatics capacity to support benchmarking for new genomics approaches and integration of agricultural data sets. Therefore, this proposal directly addresses the AG2PI overall goal of building cross-kingdom research communities to address the challenges of Genome to Phenome (G2P) research and specifically AG2PI Project Goal 2: *Identify research needs, opportunities, and gaps in methods, technologies, physical infrastructure, and data management*. This project also directly addresses the Seed Grants – Coconut Program Scope 1: *Develop tools and datasets that can be used across multiple crop species to advance genome engineering tools for integrated optimization of crop yield and livestock feed for improved animal reproduction and nutrition*. Our project is specifically designed to develop a network of agricultural bioinformaticians from both crops and livestock species and to develop bioinformatic resources focused on supporting agricultural research projects. The *expected products* from this proposal will also support the integration of genomics data types and enable advances in applying genome to phenome approaches across agricultural systems.

The basis for evaluating the success of this project will be: (1) engagement of agricultural bioinformaticians; (2) the public release of bioinformatic workflows which can be used by researchers and for benchmarking; and (3) opportunities provided to undergraduate and graduate students to develop practical training in bioinformatic analysis.

3. Expected outcomes & deliverables

The immediate *deliverables* for this project are collaborations between plant and animal bioinformaticians, the development and public release of bioinformatic workflows to support common genomic analyses and training opportunities for undergraduate and graduate students interested in agricultural careers. This translates to *anticipated outcomes* which include the closer collaboration of plant and animal bioinformaticians to support agricultural genomics; publicly available genomic analysis workflows to support researchers who wish to analyze their own data sets; a foundation for software benchmarking to support reproducibility, software development and integration of genomics data; and opportunities for agricultural students to get practical training in commonly used genomic analysis techniques. These outcomes will *support AG2PI* by supporting a cross-kingdom bioinformatics research community which will support genomics research and enable genomics data integration. Our *next steps* will be to leverage this project to seek funding for (1) developing educational opportunities for students to develop additional workflows, benchmarks, and cross-comparisons and (2) expanding the network to include data scientists and engineers working with collecting phenotype data and developing phenotype devices.

4. Qualifications of the project team

Dr. McCarthy (PI) is experienced with functional and comparative genomics analyses, including development of bioinformatic workflows, and supporting researchers working with plant and animal genomes. She currently serves as the Bioinformatics Co-coordinator of the NRSP-8 Animal Genomes project and a member of the AgBioData Sustainability Working Group. Dr. McKay (coPI) is experienced in epigenetic and comparative genomics analyses for multiple livestock species and is a member of the NRSP-8 Animal Genomes project. She is the PI on the Cattle Genome to Herd Phenotyping for Precision Ag (CG2HP) initiative funded by AG2PI. Dr. Jaiswal (coPI) has developed bioinformatic resources for multiple crop species and routinely curates and integrates genomic and phenotype data for crops. The senior personnel on this project are experienced in establishing collaborative networks, developing publicly accessible resources to support agricultural research, and providing bioinformatics training and education for researchers and students. Drs McCarthy and McKay are both members of the NRSP-8 Animal Genomes project and have recently collaborated to develop capacity funding for this project. Drs McCarthy and Jaiswal have also established a productive collaborative relationship and have worked together within the Gene Ontology Consortium to provide functional annotations for agricultural species [13] and to provide complementary bioinformatic analyses for newly sequenced genomes [12].

5. Proposal timeline

Milestones	**2023**										**2024**	
	M	**A**	**M**	**J**	**J**	**A**	**S**	**O**	**N**	**D**	**J**	**F**
develop bioinformatic network	■											
identify workflows		■										
identify data sets		■	■									
establish sustainable communication				■								
identify & select students		■	■									
develop workflow & protocols				■	■	■						
alpha testing and cross comparison							■	■	■			
prepare publications										■	■	
review and workflow/protocols release											■	■
final report to AG2PI												■

6. Engaging AG2P scientific communities & underrepresented groups

Per USDA and AG2PI rules, any "content, data, resources or tools generated from this proposal using AG2PI support will be made available to the broader stakeholder community." The PIs are committed to working with the AG2PI team to make seed project activities and outcomes visible through AG2PI communication channels. The PIs will make every effort to engage underrepresented groups, including new faculty, universities from EPSCOR states and minority serving institutions. Moreover, students working on this project will be selected to include groups traditionally underrepresented in STEM; we note University of Arizona is a Hispanic serving Institute and the PI is experienced at mentoring diverse students. Finally, as AG2PI is committed to Diversity & Inclusion, so too are the PIs and their respective institutions. Evidence of commitment to Diversity & Inclusion by the PIs can be found in the graduate students that they are currently advising who are of different races, ages, sex, nationalities, citizenships, sexual orientation, genetic disposition, neurodiversity, and disabilities.

Bibliography/References cited

1. Jackson, S. A., Iwata, A., Lee, S. H., Schmutz, J., & Shoemaker, R. (2011). Sequencing crop genomes: approaches and applications. *New Phytologist*, *191*(4), 915-925.
2. Shendure, J., & Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, *26*(10), 1135-1145.
3. Merchant, N., Lyons, E., Goff, S., Vaughn, M., Ware, D., Micklos, D., & Antin, P. (2016). The iPlant collaborative: cyberinfrastructure for enabling data to discovery for the life sciences. *PLoS biology*, *14*(1), e1002342.
4. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Research*, Volume 50, Issue W1, 5 July 2022, Pages W345–W351, doi:10.1093/nar/gkac247
5. Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., ... & Salit, M. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Scientific data*, *3*(1), 1-26.
6. Roco, M. C. (2020). Principles of convergence in nature and society and their application: from nanoscale, digits, and logic steps to global progress. *Journal of Nanoparticle Research*, *22*(11), 1-27.
7. Saha, S., Cooksey, A. M., Childers, A. K., Poelchau, M. F., & McCarthy, F. M. (2021). Workflows for rapid functional annotation of diverse arthropod genomes. *Insects*, *12*(8), 748.
8. McCarthy, F. M., Pendarvis, K., Cooksey, A. M., Gresham, C. R., Bomhoff, M., Davey, S., ... & Burgess, S. C. (2019). Chickspress: a resource for chicken gene expression. *Database*, *2019*.
9. Ibeagha-Awemu, E. M., Bissonnette, N., Bhattarai, S., Wang, M., Dudemaine, P. L., McKay, S., & Zhao, X. (2021). Whole Genome Methylation Analysis Reveals Role of DNA Methylation in Cow's Ileal and Ileal Lymph Node Responses to Mycobacterium avium subsp. paratuberculosis Infection. *Frontiers in Genetics*, *12*.
10. Ibeagha-Awemu, E. M., Kiefer, H., McKay, S., & Liu, G. E. (2022). Epigenetic Variation Influences on Livestock Production and Disease Traits. *Frontiers in Genetics*, *13*.
11. Tello-Ruiz, M. K., Stein, J., Wei, S., Youens-Clark, K., Jaiswal, P., & Ware, D. (2016). Gramene: a resource for comparative analysis of plants genomes and pathways. In *Plant Bioinformatics* (pp. 141-163). Humana Press, New York, NY.
12. Gao, Y., Liu, X., Jin, Y., Wu, J., Li, S., Li, Y., ... & Gu, L. (2022). Drought induces epitranscriptome and proteome changes in stem-differentiating xylem of Populus trichocarpa. *Plant Physiology*, *190*(1), 459-479.
13. Gene Ontology Consortium. (2010). The Gene Ontology in 2010: extensions and refinements. *Nucleic acids research*, *38*(suppl_1), D331-D335.