AG2PI SEED GRANT PROPOSAL

Title of Proposal:

Developing a cost-effective method for collecting informative, population-level molecular phenotypes

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

Troy Rowan, Assistant Professor, University of Tennessee -Institute of Agriculture, trowan@utk.edu

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

Dr. Jon Beever, Professor, University of Tennessee Institute of Agriculture, jbeever@utk.edu

Dr. Kurt Lamour, Professor, University of Tennessee Institute of Agriculture, klamour@utk.edu

Dr. Liesel Schneider, Assistant Professor, University of Tennessee Institute of Agriculture, lschneider@utk.edu

Grant Administrator:

Hollie Schreiber

Director, Office of Sponsored Programs

865-974-7123

aggrant@utk.edu

Keywords:

gene expression, cattle, phenotyping, blood

In agricultural species the most easily measured phenotypes are the ones most frequently collected, just as we would expect. For example, in beef cattle, genetic evaluations have 100s of times more birth weight records (an easy-to-measure trait) than feed efficiency records (a more difficult to measure). In many cases, we rely on correlated indicator traits to make genetic progress on economically relevant traits (e.g. sire scrotal circumference & daughter fertility). In the post-genomics era, accelerating genetic progress in agricultural species will rely on the development of novel high-dimension, high-information phenotypes [1]. While sensors and imaging have become effective in phenotyping applications [2,3], measuring intermediate molecular phenotypes (e.g. gene expression, protein abundance) on large cohorts of individuals is now feasible. Intermediate molecular phenotypes have helped uncover high-information functional genetic variation [4,5]. These phenotypes remain difficult and expensive to measure at scale in most agricultural species. The dairy industry has successfully used mid-infrared spectroscopy of routine milk samples to identify metabolites correlated with a variety of economically and environmentally important traits such as disease, feed efficiency, and methane emission [6]. These high-information molecular indicators may eventually serve as input phenotypes for genetic evaluations. Similar phenotypes do not exist in most livestock species. In this proposal, we introduce the idea of using a massively multiplexed targeted sequencing method to inexpensively measure the expression of hundreds of informative genes in easy-to-collect tissues for use as molecular phenotypes in large cohorts of animals or plants.

1. Objectives/aims

This proposal will begin to explore the utility of using targeted gene expression profiling of high-information genes in whole blood as latent phenotypes for a wide range of genotype to phenotype applications. We will refer to this method as targeted gene expression (TGE) throughout this proposal. This seed grant requests funds for a pilot study and resources to facilitate gene target discovery for future implementation at population scales (i.e. herd-level). For this initial work, we plan to focus on beef cattle. That said, we expect that the discovery workflow used to identify the optimal set of high information genes will be broadly applicable to other agricultural species interested in developing high-dimensional/high-throughput molecular phenotyping resources. We believe that TGE has many important applications that can move the scientific community towards connecting agricultural genotypes to phenotypes. This **emerging** proposal contains two main objectives: (1)Leveraging existing gene expression data to identify the informative set(s) of ~500 genes for use in TGE analysis in identifying high-risk stocker calves. and (2)Using this subset of genes to generate preliminary TGE data from ~1,500 commercial stocker calves for use in multiple future projects and grant proposals, including further the **Enabling** and **Establishing** calls in the AG2PI program.

Our TGE workflow will utilize the MosterPlex technology from Floodlight Genomics LLC (Knoxville, TN) to generate relative gene expression counts of ~500 genes in whole blood. Briefly, cDNA will be generated for peripheral blood mononuclear cells (PBMCs) extracted from whole blood samples. Massively multiplexed PCR takes place in a single tube where barcodes are added during amplification. These indexed amplicons are then pooled for Illumina sequencing. This process is more thoroughly described in Mihelic et al. (2020) [7]. This TGE method has typically focused on measuring expression in a handful of genes in a shared pathway. We aim to leverage much larger sets of genes as high-dimensional latent phenotypes. In this project, the genes that we assay will not necessarily be the of the largest effect or be the most differentially expressed. Rather we will focus on identifying the gene set that explains the most variation in overall gene expression

differences between animals and tissues. This will provide us with the most informative set of genes that could act as informative latent phenotypes. We propose three main use cases for the technology: **1)** Using targeted gene expression in whole blood as an easily-measured phenotype correlated with various difficult to measure traits (e.g. fertility, disease risk), **2)** For phenotypic prediction to inform precision management, and **3)** As a method for ultra-high powered eQTL mapping sets of trait relevant genes (i.e. testing SNP associations with expression on subsets of the most differentially-expressed genes).

We will use existing public RNA-Seq data, generated by previous studies and initiatives such as Functional Annotation of Animal Genomes (FAANG)[8], and the Farm Animal Genotype-Tissue Expression (FarmGTEx) [9] consortiums to identify the optimal sets of 500 genes use in TGE. We will download raw RNA-Seq data from SRA and process it in a custom Snakemake [10] workflow using the general approach described in Liu et al. (2020) [9]. Using expression counts derived from this dataset, we will employ a "filtering" approach to identify the most informative set of ~500 genes. Briefly, this will include searching for genes expressed in blood, then identifying those with variable expression across individuals, followed by a k-means clustering approach that will identify correlated expression. We will follow this by selecting the most informative genes from each cluster. We may further refine our gene set by identifying genes whose expression in blood is correlated with expression in other tissues. This gene selection process will remain the same regardless of the genetic architecture of the end phenotypes of interest. In the case of our preliminary investigation into stocker cattle, we will be searching for gene expression in the blood that is correlated with expression in other disease-specific tissues or samples pertaining to them, such as nasal swabs, nasopharyngeal swab, or bronchial-alveolar lavage samples. Gene set construction may also be informed by differentially expressed genes from other related RNA-Seq studies [11–13]. It is important to note that this gene discovery process will differ between applications of our TGE approach. For example, using TGE to perform a high-powered eQTL study from a list of differentially expressed genes will use a different set of genes than work aimed at performing phenotypic prediction.

We will use a subset of genes identified by the process above to generate pilot data on a set of ~1500 stocker calves from the Southeastern United States. This subset of ~500 genes will be designed to be predictive of stocker calf disease outcomes and future performance. We choose to use commercial stocker cattle for multiple reasons. First, they represent a unique challenge in the context of genomic prediction, where we are interested in understanding genomic merit, but in the context of predicting an individual's future phenotype from a given time point (intake at backgrounder/feedyard), not only its additive genetic merit. Second, these animals are incredibly heterogeneous with regard to breed makeup and management, making traditional genetic prediction difficult. As a result, we expect that using targeted expression assays could account for non-additive genetics (i.e. heterosis) and epigenetic factors not captured by a breeding value. Upon intake at stocker facilities, we will sample whole blood from each animal. A portion of this blood will be used to generate cDNA for targeted gene expression. The remaining blood will be saved for DNA extraction and genotyping for follow-up studies. Gene expression will then be profiled using the MonsterPlex workflow described above. In addition to genomic phenotypes, animals utilized in the study will be phenotyped in multiple other manners including body weights, disease incidences, antimicrobial treatments, and intake complete blood count (CBC) as part of a separate research effort being led by Co-PI Schneider. Using this as training data, we can demonstrate the efficacy of this method at identifying calves at risk for future disease as well as those likely to exhibit exemplary performance. We will perform cursory analysis of TGE results, including gene

set quality control (did primers work, do we observe gene expression variation across the sampled population) and preliminary analysis of the predictive nature of TGE. Developing complex predictive models is outside the scope of this initial seed proposal.

2. Furthering the aims of the AG2PI

In this proposal and other future proposals, we aim to demonstrate that whole blood (or other easily collected tissue) TGE results can be leveraged as high-dimensional latent phenotypes that allow us to more easily and precisely identify and classify phenotypes that may be otherwise impossible to measure at scale. This might include disease susceptibility, innate immunity, fertility potential, or metabolic efficiency. Using TGE in tandem with genomic information and traditional genetic predictions can help us achieve phenotypic prediction for individual animals or lines. The relative cost of TGE (~$5/sample) makes population-level collection more feasible than RNA-Seq, proteomics, or molecular assays. Whether measuring latent phenotypes for use in breeding programs, using TGE with machine learning to classify at-risk individuals for precision management, or creating massively powered eQTL studies that identify high-information sequence variants, we believe this technology has the potential to play a part in a variety of genome-to-phenome applications across plant and animal species. While our interest lies in beef cattle, we expect that best practices identified by this, and subsequent work could be immediately utilized in other species. The Monsterplex workflow was initially developed by Co-PI Lamour for targeted sequencing in a fungal system, and has since been applied in mammals, birds, insects, and plants with great success (Clements et al. 2021). This project proposes a novel application of this established technology. We believe that preliminary data generated here may accelerate the adoption of this technology for use across species for molecular phenotyping. As we expand this phenotyping method to other species, we intend to involve additional subject area experts in the process of developing targeted gene sets and implementing this method.

3. Expected outcomes & deliverables

The work proposed here is aimed at **encouraging cross-fertilization of existing or novel AG2P tools, data, or ideas**. Though we will initially work in beef cattle, our aim is to demonstrate the efficacy of targeted gene expression as a method for collecting informative latent phenotypes from easily collected tissues across species. To ensure that this preliminary work serves the greater genome to phenome community, we intend to publish a short communication in an open-access journal detailing our initial work. While this initial work will be far from comprehensive, ensuring that the agricultural genomics community is aware of such a powerful technology is important. Preliminary results will be featured in talks and abstracts at regional and national meetings of interest such as the Plant and Animal Genome (PAG) Conference or the World Congress on Genetics Applied to Livestock Production (WCGALP).

This preliminary work will serve as a proof of concept for targeted gene expression as a powerful latent phenotype for use in genomic studies. All phenotypic data, primers, raw sequence reads, and code will be shared freely upon publication of the project's short communication which will allow the G2P community to access and explore. Future AG2PI proposals will build on the data collected here with genotype data to demonstrate the utility of TGE in performing powerful targeted eQTL studies. Future work will also further explore the predictive ability of TGE in conjunction with genotype data using machine learning approaches. In the future we plan to utilize the approaches described in this proposal in other agricultural species. Ultimately, we are interested in pursuing

genomic technologies that enable on-demand precision management decision support. The use of TGE is an initial step into exploring those possibilities.

4. Qualifications of the project team

Lead PI Rowan has extensive experience in developing computational pipelines to analyze large-scale genomic and phenotypic datasets in livestock. His research program is interested in connecting genomics, animal health, and precision management. His lab will be responsible for processing public data to identify genes for TGE analysis. He will also work with Co-PI Schneider to use results in preliminary predictive models of stocker calf health and performance.

Co-PI Beever is a leader in animal genomics and the director of the UTIA Genomics Center for the Advancement of Agriculture. Dr. Beever's research program is broadly interested in the complex interactions that occur in genomic networks and how they are manifested in the expression of traits in beef cattle. He and PI Rowan will work together on the computational gene set discovery. His well-equipped lab will handle all the project's cDNA preparation.

Co-PI Lamour is a molecular epidemiologist with expertise in fungal plant pathogens. Dr. Lamour is the inventor of the MonsterPlex technology and will assist in primer design and will handle all molecular aspects of generating targeted expression results. He has worked extensively with investigators across species to implement this technology for a variety of use cases. He will be invaluable in helping troubleshoot problems that arise with this work and optimizing it for future projects and industry implementation.

Co-PI Schneider is a biostatistician and livestock epidemiologist. She is interested in applying precision technologies to identify animals susceptible to disease. Dr. Schneider and her graduate students will lead sample collection efforts at multiple large stocker operations across the Southeast in conjunction with her lab's ongoing projects related to beef cattle health.

5. Proposal timeline

- December 1, 2021 - May 31, 2022 - Collection of blood and disease/performance phenotypes
- December 1, 2021 - April 30, 2022 - Computational work identifying genes for use in targeted gene expression analysis. Upon identifying genes and designing appropriate primers, primers will be ordered and analysis can begin shortly after.
- May 15, 2022 - July 31, 2022 - Targeted gene expression data generated
- August 1, 2022 - September 30, 2022 - Analysis of targeted gene expression results
- October 30, 2022 - Submit short communication to appropriate open access journal

6. Engaging AG2P scientific communities & underrepresented groups

We will work to ensure that the results and best practices generated by this work are widely accessible to the greater scientific community. In this and future work on this topic, we will strive to publish solely in open access journals, and make code, primers, and raw data freely and easily available to groups around the world. We will work with relevant industry stakeholders involved with AG2PI to ensure that these methods are developed with endpoint implementation in mind. Finally, we believe that the data generated by targeted gene expression offers a large amount of information for its cost. The potential for widespread use of TGE phenotyping in low-income countries and in under-funded/orphan species could help push forward a wider range of genome-to-phenome research.

Bibliography/References cited

1. Cole JB, Eaglen SAE, Maltecca C, Mulder HA, Pryce JE. The future of phenomics in dairy cattle breeding. Anim Front. 2020;10: 37–44.

2. Silva FF, Morota G, Rosa GJ de M. Editorial: High-Throughput Phenotyping in the Genomic Improvement of Livestock. Front Genet. 2021;12: 707343.

3. Ellen ED, van der Sluis M, Siegford J, Guzhva O, Toscano MJ, Bennewitz J, et al. Review of Sensor Technologies in Animal Breeding: Phenotyping Behaviors of Laying Hens to Select Against Feather Pecking. Animals (Basel). 2019;9. doi:10.3390/ani9030108

4. Xiang R, van den Berg I, MacLeod IM, Hayes BJ, Prowse-Wilkins CP, Wang M, et al. Quantifying the contribution of sequence variants with regulatory and evolutionary significance to 34 bovine complex traits. Proc Natl Acad Sci U S A. 2019;116: 19398–19408.

5. Xiang R, MacLeod IM, Daetwyler HD, de Jong G, O'Connor E, Schrooten C, et al. Genome-wide fine-mapping identifies pleiotropic and functional variants that predict many traits across global cattle populations. Nat Commun. 2021;12: 860.

6. Tiplady KM, Lopdell TJ, Littlejohn MD, Garrick DJ. The evolving role of Fourier-transform mid-infrared spectroscopy in genetic improvement of dairy cattle. J Anim Sci Biotechnol. 2020;11: 39.

7. Mihelic R, Winter H, Powers JB, Das S, Lamour K, Campagna SR, et al. Genes controlling polyunsaturated fatty acid synthesis are developmentally regulated in broiler chicks. Br Poult Sci. 2020;61: 508–517.

8. Clark EL, Archibald AL, Daetwyler HD, Groenen MAM, Harrison PW, Houston RD, et al. From FAANG to fork: application of highly annotated genomes to improve farmed animal production. Genome Biol. 2020;21: 285.

9. Liu S, Gao Y, Canela-Xandri O, Wang S, Yu Y, Cai W, et al. A comprehensive catalogue of regulatory variants in the cattle transcriptome. bioRxiv. 2020. p. 2020.12.01.406280. doi:10.1101/2020.12.01.406280

10. Köster J, Rahmann S. Snakemake--a scalable bioinformatics workflow engine. Bioinformatics. 2012;28: 2520–2522.

11. Scott M, Woolums A, Swiderski C, Perkins A, Nanduri B. Genes and Mechanisms Associated With Experimentally Induced Bovine Respiratory Disease Identified With Supervised Machine Learning Methodology on Integrated Transcriptomic Datasets. doi:10.21203/rs.3.rs-789747/v1

12. Scott M, Woolums A, Swiderski C, Perkins A, Nanduri B, Smith D, et al. Multipopulational Transcriptome Analysis of Post-Weaned Beef Cattle at Arrival Further Validates Candidate Biomarkers for Predicting Clinical Bovine Respiratory Disease. doi:10.21203/rs.3.rs-

600364/v1

13. Scott MA, Woolums AR, Swiderski CE, Perkins AD, Nanduri B, Smith DR, et al. Whole blood transcriptomic analysis of beef cattle at arrival identifies potential predictive molecules and mechanisms that indicate animals that naturally resist bovine respiratory disease. PLoS One. 2020;15: e0227507.