AG2PI SEED GRANT PROPOSAL

Title of Proposal: *Enabling inter-institutional collaboration in AG2P using federated and transfer learning*

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

Juan Steibel, Professor of Animal Science, J. Lush Endowed Chair of Animal Breeding and Genetics, Iowa State University. jsteibel@iastate.edu

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

James Koltes. Assistant Professor of Animal Science, Iowa State University. jekoltes@iastate.edu.

Robert J. Tempelman. Professor of Animal Science, Michigan State University. tempelma@msu.edu.

Gustavo de los Campos. Professor of Epidemiology & Biostatistics, Michigan State University. gustavoc@msu.edu

Michael VandeHaar. Professor of Animal Science, Michigan State University. mikevh@msu.edu

Collaborator (Name, Title, Affiliation, email):

Grant Administrator:

Aaron Lott, Award Administrator
Phone: (515) 294-5225
Email: grants@iastate.edu

Keywords:

Genomic prediction, genome-wide association, collaborative studies, federated learning.

Project description

1. Objectives/aims

Applications of genomic prediction and genome-wide association (GWA) analyses in plant and animal agricultural species often face the problem of data sharing across multiple private and public institutions. This is particularly true for difficult to measure traits where several institutions are collecting phenotypic and genotypic data, but no single institution possess a dataset with enough individuals to obtain powerful GWA and accurate genomic predictions.

Thus, there is increased interest in methods that allow data integration and sharing while respecting privacy and intellectual property of each individual entity.

Several solutions have been used to circumvent the problem of data sharing in genetic studies. For instance, meta-analysis of GWA studies is commonly used by public-private consortia working on genetic epidemiology (Panagiotou et al., 2013); in this area our group has developed methods to perform GWA using results from multiple GBLUP genetic evaluations (Bernal Rubio et al., 2016). Likewise, meta-genomic-prediction has recently been proposed (Jighly et al., 2022). In meta-analysis each institution performs their own GWA and summary statistics from each of the studies are shared with a core group that performs the integration of results into a more powerful GWA or more accurate genomic prediction. Alternatively, monomorphic encryption (Blatt et al., 2020) has been used for genetic epidemiology to share data while protecting the privacy of each subject in the dataset, and maintaining marker-specific properties. This allows combining data and implementing tests of marker-phenotype association. Although these two approaches are promising and are already being used, there is still the need of methods that allow data integration without sharing data (either individual data or summary statistics) that may be sensitive.

Federated learning (Konečný et al., 2015) has been used for multi-location (site) distributed model fitting without sharing individual-level data between sites. In federated learning, each site holds data and performs some steps of the model training, and statistics are shared with a central site or among sites for implementing other steps of the model training. The process may be repeated iteratively until convergence is obtained or it may be updated on-demand from each site in an asynchronous way. This is different from meta-analysis where information (statistics) flow only from the collaborating institutions to the central node or from one site to another, but there is no feedback once the meta-model parameters have been estimated. Moreover, in most meta-analyses, a common (homogeneous) set of model parameter estimates is produced, while in federated learning, heterogeneous estimates for each site can be obtained while borrowing information across all institutions. Like federated learning, meta-analysis is transfer learning. This is a non-iterative approach to model fitting without sharing data that proceeds like a meta-analysis (only one cycle of updates), but where site-specific model heterogeneity is performed by shrinking the local estimates in each site, towards the common (meta-analytic) estimates.

Federated learning has been used in many applications, specifically for generalized mixed linear models applied to GWA (Chen et al., 2022; Li et al., 2022a) and analysis of medical records (Li et al., 2022b). Federated learning has also been proposed for genotype to phenotype prediction in plants (Danilevicz et al., 2022) but has not been implemented yet.

The overarching goal of this proposal is to contribute to AG2PI through facilitating distributed analysis of plant and animal breeding datasets across multiple competing

stakeholders while respecting privacy and intellectual property rights of each stakeholder. To accomplish this objective, we propose the following specific aims:

1) Develop an R package to implement several algorithms of for federated learning and transfer learning.
2) Evaluate the properties of the proposed methods using animal and plant datasets and compare to transfer learning methods and to meta analytic approaches.
3) Develop a platform for collaborative and decentralized federated learning and transfer learning.

For Objective 1. We will implement several variants of federated learning and transfer learning for genomic prediction models. We will consider two assumptions: 1) models with homogeneous SNP effects across sites, 2) models with heterogeneous effects across sites, but that borrow information from each other.

For homogeneous effects assumption, we will implement: A) Bayesian weighted average procedure that relies on sharing samples of the posterior distribution of parameters. B) federated models that share summary statistics across sites (e.g: matrices of cross products).

For heterogenous effects assumptions we will implement: A) a variant of the Gauss-Seidel algorithm called Coordinated Descent Gradient (Wu et al., 2021), B) Transfer learning using residual regressions (Zhao et al., 2022), and C) Transfer learning using summary statistics. The basic software tool on which the proposed R package will depend will be the BGLR program (Pérez and de Los Campos, 2014) authored by CO-PI G. De Los Campos.

For objective 2. We will apply the methods developed under objective one to three datasets described below:

1) The dairy data will derive from the US dairy cattle feed efficiency project first funded by USDA-NIFA (2011-2017) and then by FFAR (2019-2024) for which the PI has been Dr. Vandehaar in both cases. This data has provided the basis for national genetic evaluations of Holstein cows for feed efficiency since 2021 It includes lactations on over 6,000 cows with over 90% of the data coming from 5 key partners: Michigan State University, University of Wisconsin-Madison, Iowa State University (including data from coPI Koltes), University of Florida, and the Animal Genomics Improvement Laboratory. Each lactation typically includes at least 6 weeks of continuously recorded daily dry matter intakes and milk yields between 50 and 200 days in milk with weekly recording of milk components (fat, protein, and lactose) and body weights. These phenotypes are used to determine residual feed intake. Genotypes from 78964 SNP markers are QC according to stringent CDCB guidelines.

2) The plant data set will be from the G×E project from the Genomes to Fields (G2F) initiative. This project has collected phenotypic data since 2014 from field trials distributed over the US Maize growing. The currently available data set (2014-2021) includes 77,000 phenotypic records from 4,916 maize hybrids with DNA genotypes derived from the genotypes of the parental inbred lines (~98,000 SNPs after QC by genotyping quality, minor allele frequency, and LD-pruning). The phenotypic measures collected include: grain yield (kg/ha), moisture %, and flowering traits (days to anthesis, days to silking and anthesis-silking interval). The available data were collected in 188 trials conducted in 38 distinct evaluation sites located in the Midwest (n=48,333 phenotypic records), the Northeast (n14,333 records), and the South (n=13,958 records) regions of the US. We have already QC the phenotypic and genotypic data. Additionally, the group led by Dr. de los Campos has generated (and validated) 761

environmental covariates related to radiation, temperature, and water availability. These covariates were generated by running the APSIMx model for each trial in the data set. The G2F offer multiple advantages for this project. This data set is currently being used in a prediction competition launched by the G2F initiative; therefore, we will have multiple benchmarks to compare against.

3) The Natural Disease Challenge swine dataset is described in Cheng et al. (2020), consisting of extensive phenotypes and 650K SNP genotypes on over 5000 pigs from 7 private breeding companies and animals have been phenotyped for several growth and immune disease traits. These breeding companies directly compete in the market but agreed to contribute pigs and data to the work described in Cheng et al. (2020), on the condition that confidentiality of their data would be maintained in all analyses and publications. We will analyze this dataset at Iowa State University. This dataset represents a great example to illustrate the potential of federated learning in genomic prediction and GWA. First, it comes from competing companies that would not share data with each other under any circumstance. Second, it consists of datasets with fewer genetic connections between sites compared, for example, to the dairy feed efficiency dataset, third, the immune response phenotypes are very difficult to collect, such that no single company is able to have enough animals phenotyped. We will use this data to compare the performance of federated learning and transfer learning to those obtained from a join analysis.

To evaluate federated learning algorithms all the described datasets can be partitioned into homogeneous or heterogeneous subsets and the results can be easily compared to a joint analysis model. Moreover, the availability of curated phenotypic and genetic data will minimize time spent in data preparation and will give us the opportunity to concentrate on the evaluation and implementation of federated and transfer learning.

For Objective 3. We will develop a GitHub that will define principles and standards and provide software developed under objective 1 to enable the AG2PI community implementing federated and transfer learning without sharing data but sharing intermediate model-fitting quantities or sufficient statistics.

The GitHub will include a template and documentation that a collaborative consortium will be able to fork and use to implement their own federated learning. Once a GitHub is forked the collaborating sites will be able to push their own “sufficient statistics” or model fitting quantities and pull those pushed by other groups and perform their own federated/transfer learning. This type of federated learning is called decentralized asynchronous learning. It is decentralized because there is no central server performing computations, but only a central location holding the sufficient statistics. It is asynchronous, because each site can decide when to perform an update of their model using the most up to date results and then pushing up their own results. Note: if a site decides stop pushing updated results, they will not be able to further update their own model estimates.

2. Furthering the aims of the AG2PI

This project further advances the goals of AG2PI by developing methods and tools that will be usable across crops and livestock production systems to implement genome-wide association and genomic prediction. The developed methods will allow handling and integrating disparate data types (collected across different sites) to produce powerful GWA and precise genomic predictions while not requiring data sharing. The success of the project in the short term (within the year of execution) will be assessed through the submission and acceptance of publications and delivery of the software tool (see 3) and in the longer term, the use of the generated tool will be monitored through citations and access to the hosting site.

3. Expected outcomes & deliverables

Anticipated outputs are two peer review publications: 1) methods comparison, 2) tool description and through the delivery of a tool for federated and transfer learning analyses of GWA through GitHub. Also, outputs through participating in AG2PI activities, such as a conference and delivering a webinar, our group will seek to engage potential users of the developed methods and tools. As the tool and method are used by other groups, we will offer collaboration to maintain and extend the tool to cover other cases beyond ridge regression and GBLUP.

Moreover, this seed grant will provide preliminary results for NIFA proposals. To secure further funding, we will directly target the following programs: Animal Breeding, Genetics and Genomics, Data Science for Food and Agricultural Systems, and Plant Breeding for agricultural Production.

The ultimate impact of the methods and tools from generated this proposal will be tested in the long term through assessing the adoption of these federated learning tools by competitive private breeding companies. To increase the likelihood of this happening, our team will disseminate our results in conferences with strong corporate breeding participation such as the poultry breeders round table and national swine improvement federation.

4. Qualifications of the project team

Dr. Steibel is a professor of Animal Science and J. Lush Endowed Chair of Animal Breeding and Genetics at Iowa State University. His area of work is quantitative and computational genetics, genomics and phenomics. And his research focuses in the development, adaptation and application of statistical and computational methods for dissecting the genetic basis of phenotypic variation of production and behavioral traits in livestock species and especially in pigs. His program has been continuously funded through contracts and grants from NIFA, Animal Agriculture community groups and breeders' associations. His current and past projects include developing statistical methods for genomic prediction and genome-wide association, prediction of social genetic effects and implementation of computer vision for livestock phenotyping.

Dr. James Koltes is an Assistant Professor in Animal Science at Iowa State University. He provides expertise in precision dairy farming, data reuse and dairy cattle breeding and genomics. Dr. Koltes participates in several projects related to the acquisition, sharing and mining of big data in animal agriculture including the livestock QTLdb and is also the NRSP8 livestock bioinformatics co-coordinator. Closely aligned to this project, Dr. Koltes is also a co-PI of the dairy feed efficiency project.

Dr. Tempelman is a statistical geneticist associated with the feed efficiency consortium and has been primarily responsible for performing quality control and data editing in combining data from the various members of the consortium. He has already spearheaded various analyses, including GWA, genomic prediction, multiple trait and random regression analysis with this data and will be responsible for forwarding sufficient statistics for the proposed project.

Dr. Michael VandeHaar has been leading efforts in the US for the last 12 years to compile a database of genotypes and phenotypes related to feed efficiency of Holstein cattle. He is recognized as an expert in dairy cattle nutrition and energetics; he will provide guidance on the use of the various phenotypes from the feed efficiency database and on potential applications of the project.

Dr. de los Campos has made numerous contributions in quantitative, statistical, and computational genomics. He has developed models and algorithms for parametric and semi-parametric genomic regression. With Dr. Paulino Perez, he developed and maintains BGLR–a very popular R-package for genomic analysis of complex traits. Additionally, he developed BGData–a suite of R-packages that implement methods for analysis of biobank-size data within the R environment. Dr. de los Campos has published extensively on genomic regression with heterogenous effects–a central theme in this application–including genetic-by-environment models, genetic-by-subpopulation, ethnic- and sex-differences (including publications co-authored by Dr. Steibel, the PI of this application). Dr. de los Campos' expertise in models, software, and genomic research aligns very well with the objectives of this proposal.

5. Proposal timeline.

	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb
Objective 1	▨	▨	▨	▨								
Objective 2				▨	▨	▨	▨	▨	▨			
Objective 3							▨	▨	▨	▨	▨	▨

6. Engaging AG2P scientific communities & underrepresented groups

We will interact with the AG2P community through participating in activities such as webinars/field days and conferences. We will engage underrepresented groups through recruiting students from those groups. The ISU interdepartmental graduate programs usually recruit students from diverse backgrounds, and they are afforded the chance to rotate through our laboratories. All students and trainees working in this project will be advised and helped to complete/update an individual development plan considering aspects of their professional and personal development.

Bibliography/References cited

Bernal Rubio, Y. L., J. L. Gualdron Duarte, R. O. Bates, C. W. Ernst, D. Nonneman, G. A. Rohrer, A. King, S. D. Shackelford, T. L. Wheeler, R. J. C. J. C. Cantet, others, J. L. Gualdrõn Duarte, R. O. Bates, C. W. Ernst, D. Nonneman, G. A. Rohrer, A. King, S. D. Shackelford, T. L. Wheeler, R. J. C. J. C. Cantet, and J. P. Steibel. 2016. Meta-analysis of genome-wide association from genomic prediction models. Anim. Genet. 47:36–48. doi:10.1111/age.12378.

Blatt, M., A. Gusev, Y. Polyakov, and S. Goldwasser. 2020. Secure large-scale genome-wide association studies using homomorphic encryption. Proc. Natl. Acad. Sci. 117:11608–11613.

Chen, J., M. Edupalli, B. Berger, and H. Cho. 2022. Secure and federated linear mixed model association tests. bioRxiv.

Danilevicz, M. F., M. Gill, R. Anderson, J. Batley, M. Bennamoun, P. E. Bayer, and D. Edwards. 2022. Plant genotype to phenotype prediction using machine learning. Front. Genet. 13.

Jighly, A., H. Benhajali, Z. Liu, and M. E. Goddard. 2022. MetaGS: an accurate method to impute and combine SNP effects across populations using summary statistics. Genet. Sel. Evol. 54:1–11.

Konečný, J., B. McMahan, and D. Ramage. 2015. Federated optimization: Distributed optimization beyond the datacenter. arXiv Prepr. arXiv1511.03575.

Li, W., H. Chen, X. Jiang, and A. Harmanci. 2022a. Federated Generalized Linear Mixed Models for Collaborative Genome-wide Association Studies. arXiv Prepr. arXiv2210.00395.

Li, W., J. Tong, M. Anjum, N. Mohammed, Y. Chen, and X. Jiang. 2022b. Federated learning algorithms for generalized mixed-effects model (GLMM) on horizontally partitioned data from distributed sources. BMC Med. Inform. Decis. Mak. 22:1–12.

Panagiotou, O. A., C. J. Willer, J. N. Hirschhorn, and J. P. A. Ioannidis. 2013. The power of meta-analysis in genome-wide association studies. Annu. Rev. Genomics Hum. Genet. 14:441–465.

Pérez, P., and G. de Los Campos. 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics. 198:483–495.

Wu, X., H. Zheng, Z. Dou, F. Chen, J. Deng, X. Chen, S. Xu, G. Gao, M. Li, Z. Wang, and others. 2021. A novel privacy-preserving federated genome-wide association study framework and its application in identifying potential risk variants in ankylosing spondylitis. Brief. Bioinform. 22:bbaa090.

Zhao, Z., L. G. Fritsche, J. A. Smith, B. Mukherjee, and S. Lee. 2022. The construction of cross-population polygenic risk scores using transfer learning. Am. J. Hum. Genet. 109:1998–2008.