Project description

1. Objectives/aims

The overall goal of this proposal is to evaluate the ability of a recently proposed homomorphic data encryption method to address privacy or intellectual property issues that prevent data sharing and to enable adherence to and capitalizing on the benefits of the FAIR (Findable, Accessible, Interoperable and Reusable) principles for research data and industry data. A recent review of issues and methods related to safeguarding privacy of genomic data in human genetics is in Wang et al. (2022).

Homomorphic encryption refers to a type of encryption of raw data in a manner that ***obscures confidential aspects of the data without affecting the outcomes of certain computations on the encrypted data***. While several methods for homomorphic encryption have been proposed, most limit the types of computations and analyses that can be conducted on the encrypted data. Recently, a homomorphic encryption method was proposed by Mott et al. (2020) that is specifically suited to genetic analysis of quantitative traits using mixed linear models with Gaussian errors, including genetic parameter estimation, genome-wide association analyses (GWAS), and genomic prediction (GP). The method is based on high-dimensional random orthogonal transformation of the original data that encrypts the relationships of phenotypes and genotypes with individuals in the data set by replacing them with random linear superpositions, such that encrypted phenotypes and genotypes cannot be linked back to individuals. However, the method preserves the relationships of genotypes among SNPs (i.e. linkage disequilibrium) and relationships between SNP genotypes, phenotypes, and specified covariates or fixed effects. As a result, the encryption does not affect the outcomes of analyses that are based on estimation of SNP effects using mixed linear models with Gaussian errors. Another attractive feature of the method is that it allows ***subsets of the joint data to be encrypted with a different random key, without that key having to be shared for joint analysis***. This allows each contributor to the joint data to encrypt the data with a private key prior to sharing. Although mixed linear models with Gaussian errors are common for genetic analysis of quantitative traits, including for GWAS and genomic prediction (e.g. genomic best linear unbiased prediction, GBLUP), other methods used for GWAS and GP do not fall in this category, e.g. the Bayesian variable selection methods for GWAS and GP (Habier et al. 2011; Garrick and Fernando 2013) and it is not clear to what extent the results of these methods are affected by the encryption proposed by Mott et al. (2020).

Against this background, the specific objectives of this proposal are to apply the homomorphic encryption method proposed by Mott et al. (2020) to a unique data set with deep phenotypes and genotypes on pigs from seven private breeding companies, in order to:

i) Validate the encryption method, with separate encryption for the data from each company, for GWAS and GP using standard mixed linear models (GBLUP)

ii) Evaluate the impact of encryption on GWAS and GP using Bayesian variable selection methods.

iii) Demonstrate the encryption methods to the breeding companies and other stakeholders, including the crop and livestock genetics communities, and determine to what extent it addresses their concerns about data sharing and their ability to conduct research using encrypted data.

The data set we will use is described in Cheng et al. (2020), consisting of extensive phenotypes and 650K SNP genotypes on over 3200 pigs from 7 private breeding companies. These breeding companies directly compete with each other 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. For objectives 1 and 2, data from each company will be separately encrypted using the R or Julia scripts developed in Mott et al. (2020) and analyzed using GBLUP and Bayesian variable selection approaches for GWAS and GP, as implemented in the Julia for Whole-genome Analysis Software (JWAS, https://github.com/reworkhow/JWAS.jl) described in Cheng et al. (2018). Several of these analyses have already been applied to the original data, as described in Cheng et al. (2021). Analyses will be repeated for the multiple phenotypes described in Cheng et al. (2020), including phenotypes that violate the assumption of normality (e.g. mortality). Results will be compared to those from analysis of the original data to evaluate the impact of encryption. For objective 3, results will be communicated to the seven breeding companies, requirements for data encryption and standardization of data across companies will be identified, and input will be solicited from the breeding companies on the ability of this approach to address confidentiality concerns and what, if any, other measures are needed. We will also present the method and results in a workshop for the AG2PI communities to solicit their inputs on the effectiveness and limitations of the use of encrypted data for further research.

2. Furthering the aims of the AG2PI

Data sharing is essential for G2P research, as large data sets are needed to answer many G2P questions and for genomic prediction to advance genetic improvement in livestock and crop populations. This includes the sharing of the extensive phenotypic and genetic data that is available in industry, as well as the sharing of data among researchers. In addition, many funding agencies and journals now require research data to be made available publicly following the FAIR principles, not only to allow results and conclusions of the published research to be validated but also to enable the data to be used for other purposes by other researchers. There are, however, several obstacles to data sharing, including the reluctance of private industry to share data because their data are considered to contain intellectual property information or trade secrets and because of concerns that the data could be used to undermine a company's competitive advantage. Some of these concerns can be overcome by secure encryption of the data such that confidential information is protected, while allowing the data to be used for validation and for further research. The proposed work will evaluate a homomorphic encryption approach that has recently been proposed for this purpose on an multi-company industry research data set by testing the effect of encryption on the results of statistical analyses of the data, and to obtain feedback from industry on the suitability of this approach to protect company confidential information. The proposed approach will also enable creation of joint training data for genomic prediction across companies, allowing SNP effects to be estimated based on the combined data and then used for private genomic prediction within each company. Success will be evaluated based on the outcomes of the analyses and the feedback received from industry. The research outcomes and feedback will determine whether additional method development and communication is needed.

3. Expected outcomes & deliverables

We expect to demonstrate that encryption does not affect GWAS and GP results when using GBLUP methods for traits that follow normality assumptions. We also expect to be able to quantify the impact of encryption on GWAS and GP results from GBLUP approaches for traits that do not follow normality assumptions and for Bayesian variable selection approaches. This knowledge will be important to determine whether additional developments in the encryption methodology are needed. We will also obtain feedback from breeding companies on the effectiveness of the approach in terms of alleviating data sharing concerns. This will also inform whether additional methodology development and/or additional outreach is needed. The proposal uses an existing multi-company industry research data set as a case study. However, results will apply to other data sets across the livestock and plant communities and industries.

4. Qualifications of the project team

Dr. Dekkers has has over 30 years of experience as a faculty member and active researcher in quantitative genetics and genomics. He has extensive experience in developing and leading multi-disciplinary collaborative programs.

Dr. Cheng has expertise in statistical genetics and software development. He has developed multiple novel algorithms and statistical models for quantitative genetics of big data. The development of a software tool called "JWAS", for which Cheng is the primary developer, was funded by two USDA-NIFA-AFRI projects.

Dr. Tuggle has more than 30 years of experience in analyzing porcine genomes, from genetic and physical gene mapping to functional and bioinformatic analyses of gene networks. He has extensive expertise in applying transcriptomics to understanding immunogenetics and genomics.

Dr. Mott is an expert in bioinformatics and statistical genetics of animal and plant populations
He recently developed and published a method for homomorphic genotype encryption (Mott et al 2020 Genetics) based on random orthogonal matrices as encryption keys.

Dr. Fang has extensive expertise and experience in quantitative genetics/genomics, bioinformatics and comparative genomics. He leads the FarmGTx project, which is an international consortium (over 30 universities and institutes involved) aiming to build a comprehensive catalog of regulatory variants (e.g., eQTLs) across farm animal species.

5. Proposal timeline

Objective	Milestone	9/22-12/22	1/23-5/23	6/23-7/23	7/23-8/23
i)	Complete GBLUP analyses	■			
ii)	Complete Bayesian analyses		■		
iii)	Conduct Industry outreach			■	
	Conduct AG2PI workshop				■

6. Engaging AG2P scientific communities & underrepresented groups

Geneticists from the seven breeding companies will be engaged right from the start, as we have a monthly virtual meeting with them. Other industry and AG2P scientific communities will be engaged by providing a workshop in the summer of 2023 to presents the results of the work and generate discussion on whether this addresses the needs of the community and/or what its limitations are, requiring further work. Results will also be communicated to the scientific community and industry through a scientific publication. The proposed methodology will allow smaller organizations to team up to create the large data sets that are needed for G2P research without compromising their intellectual property and confidentiality.

Bibliography/References cited

Cheng, H., Fernando, R. L., and Garrick, D. J. 2018 JWAS: Julia implementation of whole-genome analysis software. Proceedings of the World Congress on Genetics Applied to Livestock Production,11.859. Auckland, New Zealand. http://www.wcgalp.org/proceedings/2018/jwas-julia-implementation-whole-genome-analyses-software

Cheng, J., Putz, A.M., Harding, J.C., Dyck, M.K., Fortin, F., Plastow, G.S., Canada, P. and Dekkers, J.C., 2020. Genetic analysis of disease resilience in wean-to-finish pigs from a natural disease challenge model. *Journal of Animal Science*, *98*(8), p.skaa244. https://doi.org/10.1093/jas/skaa244

Cheng, J., Fernando, R., Cheng, H., Kachman, S.D., Lim, K., Harding, J., Dyck, M.K., Fortin, F., Plastow, G.S. and Dekkers, J., 2021. Genome-wide association study of disease resilience traits from a natural polymicrobial disease challenge model in pigs identifies the importance of the major histocompatibility complex region. *G3 Genes| Genomes| Genetics*. https://doi.org/10.1093/g3journal/jkab441

Garrick D.J., Fernando R.L. (2013) Implementing a QTL Detection Study (GWAS) Using Genomic Prediction Methodology. In: Gondro C., van der Werf J., Hayes B. (eds) Genome-Wide Association Studies and Genomic Prediction. Methods in Molecular Biology (Methods and Protocols), vol 1019. Humana Press, Totowa, NJ. https://doi.org/10.1007/978-1-62703-447-0_11

Habier, D., Fernando, R.L., Kizilkaya, K. and Garrick, D.J., 2011. Extension of the Bayesian alphabet for genomic selection. *BMC bioinformatics*, *12*(1), pp.1-12. https://doi.org/10.1186/1471-2105-12-186

Mott, R., Fischer, C., Prins, P. and Davies, R.W., 2020. Private Genomes and Public SNPs: Homomorphic encryption of genotypes and phenotypes for shared quantitative genetics. *Genetics*, *215*(2), pp.359-372. https://doi.org/10.1534/genetics.120.303153

Wan, Z., Hazel, J.W., Clayton, E.W., Vorobeychik, Y., Kantarcioglu, M., and Malin, B.A., Sociotechnical safeguards for genomic data privacy. *Nat Rev Genet* (2022). https://doi.org/10.1038/s41576-022-00455-y