Maintaining Data Confidentiality in Collaborative Genomic Analyses Using Encrypted Genotypes and Phenotypes on Disease Resilience in Pigs

By

DONNA LI

THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Animal Biology

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Hao Cheng, Chair

Jack C. M. Dekkers

Anita M. Oberbauer

Committee in Charge

2024

Table of Contents

List of Figures

List of Tables

Abstract

Maintaining Data Confidentiality in Collaborative Genomic Analyses Using Encrypted Genotypes and Phenotypes on Disease Resilience in Pigs

Genome-to-phenome analyses in animal breeding often involves the estimation of genetic marker effects and breeding values, based on individual-level genotype and phenotype information. A genome-wide association study (GWAS) may also used to assess the correlations between single-nucleotide polymorphisms (SNPs) and the phenotype of interest. However, each animal breeder has a relatively small sample size, which could lead to an underpowered statistical analysis and lead to a higher chance of obtaining a false negative result. Using joint analyses by combining individual-level data before performing the analysis can increase statistical power and improve prediction accuracy, but animal breeders may be hesitant to share their animal's information with others, as this can reveal sequences responsible for their animals' economic value. One solution is to implement an encryption scheme to protect individual-level information. Homomorphic encryption for genotypes and phenotypes (HEGP) is a type of encryption that allows encrypted genomic data to be analyzed directly, providing a more secure method of estimating marker effects and breeding values when performing a joint analysis. In this study, HEGP is implemented on a real data set from a disease resilience study in pigs and evaluates the correlation between estimated marker effects and estimated breeding values of the encrypted and unencrypted data, respectively. The estimated percentages of genetic variance for each window obtained from a GWAS using the encrypted data were also compared to the results of the original study from which the data originated. Correlations between estimated marker effects, estimated breeding values, and estimated percentages of genetic variance of each window of the analyses using unencrypted data and encrypted data were all approximately 1, indicating that the implementation of HEGP in GWAS joint analyses produces effectively identical results and does not affect the precision of the obtained results.

Acknowledgements

I would like to express my deepest appreciation to my committee members, Dr. Hao Cheng, Dr. Jack Dekkers, and Dr. Anita Oberbauer. Their expertise and guidance were instrumental in the completion of this thesis. A special thanks to my advisor, Dr. Hao Cheng, for providing me with this incredible opportunity and for his unwavering support and mentorship over the past two years. I am also deeply grateful to Dr. Jack Dekkers for his expertise in quantitative and statistical genetics and for providing the essential files needed to replicate the results of the original study. I extend my sincere thanks to Dr. Anita Oberbauer for her expertise in animal genetics and her thoughtful guidance. I would also like to acknowledge members of the Cheng Lab—Tianjing Zhao, Jiayi Qu, Mark Watson, Olivia Liang, Quazi Abir Hassan Roddur, and Daniel Novoa—for their camaraderie and support during this project. In addition, I would like to thank AG2PI and USDA-NIFA for funding this research and PigGen Canada, who provided the data analyzed in this study. Finally, I'd like to thank my family and friends for supporting me throughout this academic endeavor.

1 Chapter 1

Literature Review

1.1 Genome-to-phenome analysis

As the cost and time to genotype individuals has decreased, the use of genotypic and phenotypic data has become increasingly popular for genome-to-phenome analysis (Tuggle *et al.* 2022). For example, animal breeders may be interested in the loci that are highly associated with certain traits, so they can determine the breeding scheme that is more likely to benefit their population, by decreasing the frequency of unwanted traits and increasing the frequency of desired traits (Tuggle *et al.* 2022). This type of information can be in the form of metrics such as genetic marker effects and breeding values (Meuwissen 2007; Calus 2010).

With advances in the discovery of single nucleotide polymorphisms (SNPs) and more affordable genotyping technologies, genomic selection (GS) was introduced as a method that focuses on estimating the effects of all SNP markers as a way to aid breeders (Meuwissen 2007; Meuwissen *et al.* 2016).

In animal breeding, a commonly used model is the marker effects model (MEM)

(Meuwissen *et al.* 2001; Fernando *et al.* 2014):

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{e}, \quad \boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I_q}\sigma_\alpha^2),\ \mathbf{e} \sim N(\mathbf{0}, \mathbf{I_n}\sigma_e^2) \tag{1}$$

where $\mathbf{y}$ is an $n \times 1$ vector of phenotypic values for n individuals, $\boldsymbol{\mu}$ is an $n \times 1$ vector of the mean μ, $\mathbf{X}$ is an $n \times p$ incidence matrix for p fixed effects, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{M}$ is a $n \times q$ centered matrix of q marker covariates, $\boldsymbol{\alpha}$ is a $q \times 1$ vector of marker effects with variance $\mathbf{I_q}\sigma_\alpha^2$ where $\mathbf{I_q}$ is a $q \times q$ identity matrix and σ_α^2 is the marker effects variance, $\mathbf{e}$ is an $n \times 1$ vector of residuals with variance $\mathbf{I_n}\sigma_e^2$, where $\mathbf{I_n}$ is an $n \times n$ identity matrix and σ_e^2 is the residual variance (Meuwissen *et al.* 2001; Fernando *et al.* 2014).

Rather than assuming all SNPs have an effect on the phenotype, implementing specific prior information into the model may represent the biology better (Meuwissen *et al.* 2016). For example, Bayesian methods assume that some SNPs may have no effect, and the parameter $(1 - \pi)$ assesses how many SNPs have nonzero effects (Fernando *et al.* 2014). BayesCπ utilizes a normal distribution for SNP effects (Habier *et al.* 2011) and BayesB uses a t-distribution for SNP effects (Meuwissen *et al.* 2001). Depending on the underlying biology, these methods may be more suitable for the analysis (Meuwissen *et al.* 2016).

Estimated breeding values (EBVs) can be assumed to be represented solely by SNP markers (Fernando *et al.* 2014). Thus, an individual's breeding value is computed as the product of its genotype and the corresponding SNP effects (Fernando *et al.* 2014). Following the MEM equation 1, the EBV for individual j is given by:

$$EBV_j = \mathbf{M}_j\boldsymbol{\alpha} \tag{2}$$

where $\mathbf{M}_j$ is the $1 \times q$ vector of SNP covariates of individual j and $\boldsymbol{\alpha}$ is the $q \times 1$ vector of SNP effects for q SNPs (Fernando *et al.* 2014).

In genome-to-phenome analysis, in addition to genomic selection, genome-wide association studies (GWAS) offer insight into which SNPs are more highly correlated with certain phenotypes (Abdellaoui *et al.* 2023). Since the rise in GWAS studies in the early 2000s (Klein *et al.* 2005; DeWan *et al.* 2006; Burton *et al.* 2007), GWASs and follow-up experiments have been performed for various traits and diseases, and results have shown it is a valuable tool in genomic research (Abdellaoui *et al.* 2023).

1.2 Joint Analysis and Data Privacy Concerns

When an animal breeder is estimating marker effects and breeding values or performing GWASs, one issue that may arise is a relatively small sample size. Without a large enough sample size, the performed statistical analysis may be underpowered and lead to a higher chance of obtaining a false negative result, potentially rendering the result to be unreliable (Yengo *et al.* 2018). The use of joint analyses, where multiple sources, like breeding companies, pool their individual-level data together before performing an analysis, has been shown to increase statistical power and improve prediction accuracy (Yang *et al.* 2012; Yengo *et al.* 2018, 2022; Zhao *et al.* 2023). By participating in a joint analysis, animal breeders can be more confident in the estimated marker effects and breeding values for their animals.

Yet, genotypic and phenotypic data on the individual-level allows for the inference of a given individual's identity, which poses a risk in the case of data breaches (Malin and Sweeney 2004; Zhao *et al.* 2023). One of the main reasons animal breeders prefer to keep their genomic livestock data private may be because sharing such information can reveal sequences responsible for their animals' economic value (Cleveland *et al.* 2012). This could undermine the extensive, proprietary breeding strategies that breeders have been developing for a long time to produce the populations that these animals originate from (Cleveland *et al.* 2012).

Policies regarding the security of databases and access to data are imperative to maintaining a secure environment when working with genotypic and phenotypic data (Malin *et al.* 2011). Though caution should always be taken when handling sensitive data, more can still be done to protect individual-level information in terms of how data is processed and stored.

1.3 Data Encryption

Previous research has considered a variety of methods to protect genetic information, primarily for human data. One such method is anonymization, which involves modifying identifiable information to reduce the likelihood that genomic sequences can be linked to a small number of individuals (Loukides *et al.* 2010). Another method is multiparty computation, which allows analyses to be performed on confidential, shared data, with quality control and population stratification correction (Cho *et al.* 2018). There are also currently encryption schemes that have been considered for protecting genomic data, such as using Advanced Encryption Standard (AES) for encryption and Galois/Counter Mode (GCM) for decryption (Gudodagi and Reddy 2022). However, these strategies have their shortcomings. Individual-level information can still be inferred from genomic sequences (Loukides *et al.* 2010), computation speeds, such as that of multiparty computation (Cho *et al.* 2018), may be slower compared to other methods, like those that implement homomorphic encryption (Blatt *et al.* 2020a), and many encryption schemes require genomic data to be decrypted before performing any analyses (Gudodagi and Reddy 2022).

A particular encryption scheme that has recently received attention is homomorphic encryption, which allows specific computations to be performed on encrypted data the same way that they are performed on unencrypted data (Ogburn *et al.* 2013). This eliminates the need to decrypt the data before performing analyses, greatly reducing the risk of identifiable genomic information being compromised if a data breach occurs (Ogburn

et al. 2013). The results from the data analysis performed on the encrypted data are then decrypted to obtain the same results as the same analysis performed on the unencrypted data (Ogburn *et al.* 2013). Iterations of homomorphic encryption include partially homomorphic encryption, for performing a single operation an unlimited number of times, and somewhat homomorphic encryption, for performing a set of operations a limited number of times (Wood *et al.* 2020). The fully homomorphic encryption scheme, first proposed by Gentry (2009a,b), uses features from both the partially homomorphic encryption scheme and the somewhat homomorphic encryption scheme. Consequently, this allow for an unlimited number of addition and multiplication operations to be performed on encrypted data (Wood *et al.* 2020).

To assess its practicality, the fully homomorphic encryption scheme has been applied to genomic data, such as in GWAS (Lu *et al.* 2015; Blatt *et al.* 2020b,a). For large data sets that may require cloud computation, fully homomorphic encryption allows analyses to be performed securely, as the data is encrypted while in the cloud and during computations (Lu *et al.* 2015). Furthermore, homomorphic encryption has been shown to be compatible with GWAS (Blatt *et al.* 2020a,b) and more than one magnitude faster than multiparty computation, as described in Cho *et al.* (2018).

Mott *et al.* (2020) proposed a method called homomorphic encryption for genotypes and phenotypes (HEGP), where genotypic and phenotypic data are encrypted by random orthogonal transformations to obscure information about individuals and relationships between individuals, while preserving relationships between SNPs. This preserves individual privacy and allows analyses such as GWAS using linear mixed-models (Mott *et al.* 2020). Moreover, unlike traditional homomorphic encryption, HEGP allows marker effects to be estimated directly from HEGP-encrypted data, without the need for decryption after the analysis (Mott *et al.* 2020).

Zhao *et al.* (2023) has previously shown that HEGP is compatible with analyses based on linear mixed-models, such as GBLUP and RR-BLUP, and is suitable for Bayesian

variable selection methods such as BayesCπ. It may be that HEGP is also compatible with other fundamentally similar methods.

2 Chapter 2

Maintaining Data Confidentiality in Collaborative Genomic Analyses Using Encrypted Genotypes and Phenotypes on Disease Resilience in Pigs

2.1 Introduction

Estimated breeding values and SNP marker effects are valuable metrics for genomic selection in animal breeding (Meuwissen 2007; Meuwissen *et al.* 2016). However, one issue that may arise is a relatively small sample size. Without a large enough sample size, the performed statistical analysis may be underpowered and lead to a higher chance of obtaining a false negative result, potentially rendering the result to be unreliable (Yengo *et al.* 2018). The use of joint analyses, where multiple sources, like breeding companies, pool their individual-level data together before performing an analysis, has been shown to increase statistical power and improve prediction accuracy (Yang *et al.* 2012; Yengo

et al. 2018, 2022; Zhao *et al.* 2023). Therefore, joint analyses would likely result in more accurate estimated marker effects and breeding values of the animals.

Even so, one issue that joint analyses of this nature cannot account for is that genotypic and phenotypic data on the individual-level allow for the inference of a given individual's identity, which poses a risk in the case of data breaches (Malin and Sweeney 2004; Zhao *et al.* 2023). Many animal breeders prefer to keep their genomic livestock data private may be because sharing such information can reveal sequences responsible for their animals' economic value (Cleveland *et al.* 2012). This could undermine the extensive, proprietary breeding strategies that breeders have been developing for a long time to produce the populations that these animals originate from (Cleveland *et al.* 2012).

Previous research has considered different methods to protect genetic information, such as anonymization (Loukides *et al.* 2010), multiparty computation (Cho *et al.* 2018), and encryption schemes (Gudodagi and Reddy 2022). However, there are downsides to these strategies. Individual-level information can still be inferred from genomic sequences (Loukides *et al.* 2010), computation speeds, such as that of multiparty computation (Cho *et al.* 2018), may be slower compared to other methods, like those that implement homomorphic encryption (Blatt *et al.* 2020a), and many encryption schemes require genomic data to be decrypted before performing any analyses (Gudodagi and Reddy 2022).

Homomorphic encryption is a type of encryption developed recently that allows encrypted data to be analyzed the same way as unencrypted data, for certain computations (Ogburn *et al.* 2013). This removes the decryption step before analyzing the data, greatly mitigating the risk of compromising identifiable genomic information in the event of a data breach (Ogburn *et al.* 2013). Some data sets may be large enough that they cannot be analyzed locally and require cloud computation, so fully homomorphic encryption can be used to keep the data encrypted while in the cloud and during computations for security (Lu *et al.* 2015). In addition, homomorphic encryption has been shown to be

compatible with genome-wide association studies (GWAS) (Blatt *et al.* 2020a,b).

Mott *et al.* (2020) proposed a method called homomorphic encryption for genotypes and phenotypes (HEGP), where random orthogonal transformations are performed on genotypic and phenotypic data to obscure information about individuals and relationships between individuals, while maintaining relationships between SNPs. This enables the preservation of individual privacy as well as allowing analyses such as GWAS using linear mixed-models, which rely on SNP relationships (Mott *et al.* 2020). HEGP has been previously shown to be compatible with analyses based on linear mixed-models, such as GBLUP and RR-BLUP, and is suitable for Bayesian variable selection methods such as BayesCπ (Zhao *et al.* 2023).

This study implements HEGP on a real data set from a disease resilience study in pigs and evaluates the correlation between estimated marker effects and estimated breeding values of the encrypted and unencrypted data, respectively. In addition, the estimated percentages of genetic variance for each window obtained from a GWAS using the encrypted data were compared to the results of the original study from which the data originated. The current study demonstrates that HEGP provides effectively identical estimated marker effects, estimated breeding values, and percentages of genetic variance obtained from the GWAS run on unecrypted data.

2.2 Methods

2.2.1 Data Set

The raw data set without encryption used in this study originates from Cheng *et al.* (2020), who analyzed disease resilience in wean-to-finish pigs. A total of 3,285 Large White by Landrace barrows were evaluated under the natural challenge protocol outlined in Putz *et al.* (2019), designed to simulate an environment of high disease pressure on a commercial farm. Pigs completed three phases in the following order: quaran-

tine nursery (19 days on average), challenge nursery (27 days on average), and finishing phase (100 days on average), in batches of approximately 60-75 individuals and a total of 50 batches, with a new batch entering the quarantine nursery every three weeks (Putz *et al.* 2019; Cheng *et al.* 2020). As described in Cheng *et al.* (2020), body weight information was collected upon entry of each phase, upon exit of each phase, every three weeks in the finishing phase, and when individuals died or were euthanized. Daily feed intake data was collected in the finishing phase and any missing daily feed intake values were estimated using a rolling average for each animal over the course of five days (Cheng *et al.* 2020). In addition to the body weight and feed intake information, other phenotypic traits recorded includes average daily gain, feed conversion ratio, and carcass qualities (Cheng *et al.* 2020). All pigs were genotyped using a 650K single nucleotide polymorphism (SNP) panel by Delta Genomics (Cheng *et al.* 2020). After quality control (minor allele frequency > 0.05, call rate for marker > 0.10, and call rate for individual > 0.10), a total of 435,172 SNPs for 3,139 pigs were included in the data set and used for analysis (Cheng *et al.* 2020). In a follow-up GWAS study, specifically for the average daily gain in the challenge nursery and using the same SNP array parameters and quality control, a total of 435,172 SNPs and 3,205 pigs were used in the analysis after removing individuals with no batch and pen information (Cheng *et al.* 2021).

2.2.2 Data pre-processing

For the purposes of this study, the unencrypted data set was pre-processed before proceeding with the analysis. Using the same model as Cheng *et al.* (2021), the trait of interest was average daily gain in the challenge nursery and this was the only trait included in the analysis. Further referencing their model, the fixed batch effect, the fixed effect indicating whether the individual died in the challenge nursery or not, the fixed effect of age upon entry to the quarantine nursery, the random effect of the specific batch and pen the individual was in, the random litter effect, random marker effects, and the

residual effect were included in the model.

Individuals with missing information for average daily gain in the challenge nursery or any fixed or random effect were removed from the data set. After the pre-processing steps, a total of 435,172 SNPs and 3,172 pigs were used for the data analysis. This is lower than the 3,205 pigs used in Cheng *et al.* (2021), as the encryption method used in this study does not allow for missing information.

2.2.3 Homomorphic Encryption

Homomorphic encryption (HE) is a method of encryption that applies a linear transformation to a matrix of genotypes or phenotypes via a random orthogonal matrix, referred to as the encryption key.

A single random, orthogonal matrix, with size equal to the number of individuals, is sufficient to perform homomorphic encryption. However, this can be impractical when combining data sets from different companies or collaborators who do not wish to share their data. In this case, each individual data set can also be encrypted separately, with its own encryption key, before being shared.

This study specifically utilizes a type of HE known as homomorphic encryption for genotypes and phenotypes (HEGP), as described in Mott *et al.* (2020) and implemented in Zhao *et al.* (2023) using simulated data. The data used in Cheng *et al.* (2021) contains individuals from seven different companies, therefore the genotype and phenotype data were organized by company to be able to encrypt each company's data with its own encryption key.

For each company i, a corresponding orthogonal matrix $\mathbf{P}_i$ was generated using the Stiefel manifold (Hoff 2009, 2021). These matrices were then organized in a diagonal manner as:

$$\mathbf{P}_{all} = \begin{bmatrix} \mathbf{P}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{P}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{P}_7 \end{bmatrix} \tag{3}$$

where $\mathbf{P}_{all}$ is the random, orthogonal matrix used as the key for the entire data set of individuals from all seven companies.

With the transformation by the random, orthogonal matrix $\mathbf{P}_{all}$, the encrypted genotypes $\mathbf{P}_{all}\mathbf{X}_{all}$ and encrypted phenotypes $\mathbf{P}_{all}\mathbf{y}_{all}$ are as follows:

$$\mathbf{P}_{all}\mathbf{X}_{all} = \begin{bmatrix} \mathbf{P}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{P}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{P}_7 \end{bmatrix} \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \vdots \\ \mathbf{X}_7 \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1\mathbf{X}_1 \\ \mathbf{P}_2\mathbf{X}_2 \\ \vdots \\ \mathbf{P}_7\mathbf{X}_7 \end{bmatrix} \tag{4}$$

$$\mathbf{P}_{all}\mathbf{y}_{all} = \begin{bmatrix} \mathbf{P}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{P}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{P}_7 \end{bmatrix} \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_7 \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1\mathbf{y}_1 \\ \mathbf{P}_2\mathbf{y}_2 \\ \vdots \\ \mathbf{P}_7\mathbf{y}_7 \end{bmatrix} \tag{5}$$

where $\mathbf{X}_i$ are the SNP genotypes and $\mathbf{y}_i$ are the phenotypes for the ith company. A visualization for the joint analysis process is in Figure 1.

2.2.4 Validation of the Results Obtained Using Encrypted Data

The analyses were completed using the software JWAS (Julia for Whole-genome Analysis Software) (Cheng *et al.* 2018, 2022) using the BayesB method (Meuwissen *et al.* 2001), with 50,000 MCMC (Markov Chain Monte Carlo) iterations. The GWAS was performed using non-overlapping 0.25-Mb windows. The hyperparameters were determined by referencing those used in Cheng *et al.* (2021).

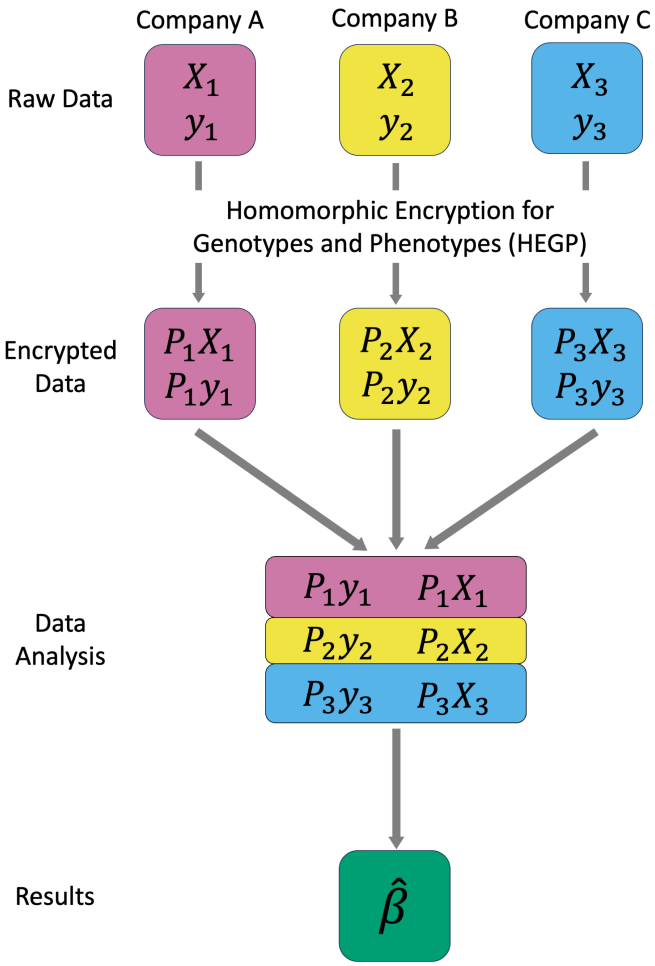


Figure 1: **Example Joint Analysis Flowchart Using HEGP** In this example, three companies (A, B, C) are shown performing a joint analysis using their individual genotype (X_1, X_2, X_3) and phenotype (y_1, y_2, y_3) data. Each company performs HEGP on their data separately with their own encryption key (P_1, P_2, P_3). All the encrypted data is then combined to perform the analysis and obtain the results. In the case of this study, the results are estimated marker effects and breeding values, but the form of the results may be different depending on the intended analysis.

To demonstrate that the results obtained by analyzing the unencrypted genotypes and phenotypes are essentially identical to those obtained by analyzing the encrypted genotypes and phenotypes, the Pearson correlation coefficient between the estimated marker effects was computed. If the encryption is successful, the correlation between the marker effects from the unencrypted and encrypted data should be approximately 1. The estimated breeding values and the estimated percentages of genetic variance attributed to each window were also compared in this manner. To assess the efficiency of HEGP, the runtimes for estimating marker effects using the unencrypted and encrypted data were compared. The extent of Monte Carlo error was also assessed by comparing the marker effects and breeding values estimated from the unencrypted data using three different random seeds.

For further explanations and derivations, refer to Zhao *et al.* (2023).

2.3 Results

2.3.1 Comparing Estimated Marker Effects and Breeding Values

Scatter plots for the estimated marker effects and breeding values for the encrypted and unencrypted can be found in Figure 2 and Figure 3, respectively. The calculated Pearson correlation coefficients for both the estimated marker effects and breeding values were close to 1, as shown in Table 1.

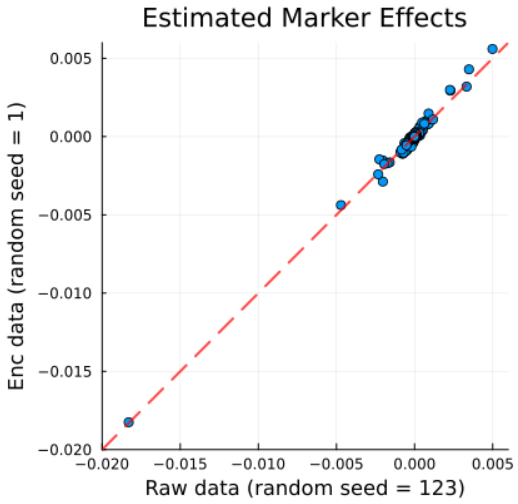


Figure 2: **Scatter Plot of Estimated Marker Effects** calculated using the unencrypted data on the x-axis and calculated using the encrypted data on the y-axis. The dashed red line represents a correlation of 1.

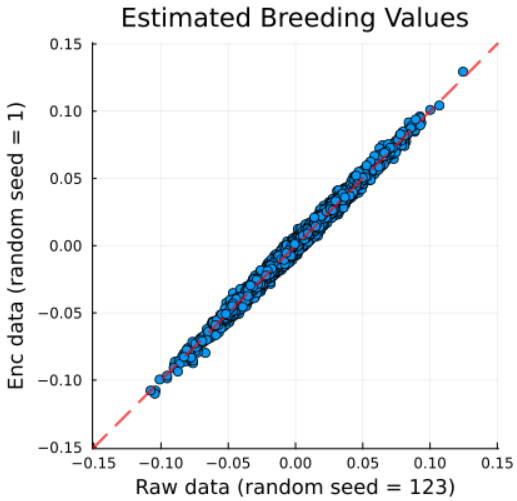


Figure 3: **Scatter Plot of Estimated Breeding Values** calculated using the unencrypted data on the x-axis and calculated using the encrypted data on the y-axis. The dashed red line represents a correlation of 1.

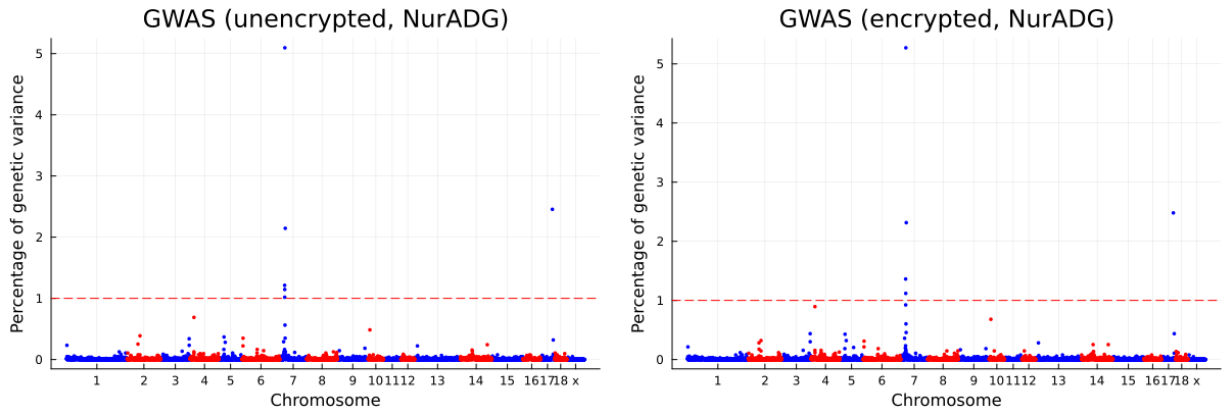
Values	Pearson Correlation Coefficients
Marker Effects	0.9598
Breeding Values	0.9956

Table 1: **Pearson Correlation Coefficients for Estimated Marker Effects and Breeding Values** between unencrypted and encrypted genotypes and phenotypes

2.3.2 Comparing GWAS Results

The percentage of genetic variance attributed to each window, obtained from the unencrypted genotypes and phenotypes, is shown in Figure 4a and those obtained from the encrypted genotypes and phenotypes is shown in Figure 4b.

A scatter plot (Figure 5) of estimated percentage of genetic variance of each window from the analysis using unencrypted data and those from the analysis using encrypted data show that Figure 4a and Figure 4b are essentially the same. The percentage of genetic variance for each window using the encrypted data was also compared to the results of Cheng *et al.* (2021). Pearson correlation coefficients can be found in Table 2.



(a) (b)

Figure 4: **GWAS Results** (a) using unencrypted genotypes and phenotypes and (b) using encrypted genotypes and phenotypes. In both plots, each point represents a non-overlapping 0.25-Mb window. The x-axes are the chromosome numbers, including both autosomal and sex chromosomes, with the odd chromosomes colored in red and even chromosomes colored in blue. The y-axes are the percentages of genetic variance for each 0.25-Mb window. The dashed red line is the significance threshold set at the percentage of genetic variance of a window equal to 1.

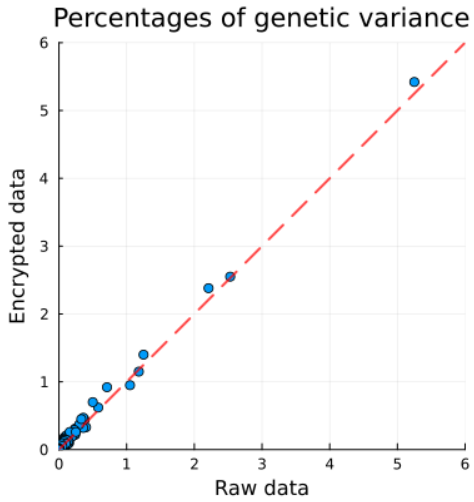


Figure 5: **Scatter Plot of Percentages of Genetic Variances Attributed to Each 0.25-Mb Window** of variances calculated using the unencrypted data on the x-axis and calculated using the encrypted data on the y-axis. The dashed red line represents a correlation of 1.

Values	Pearson Correlation Coefficients
Unencrypted and Encrypted GWAS	0.9844
Original Study and Encrypted GWAS	0.9834

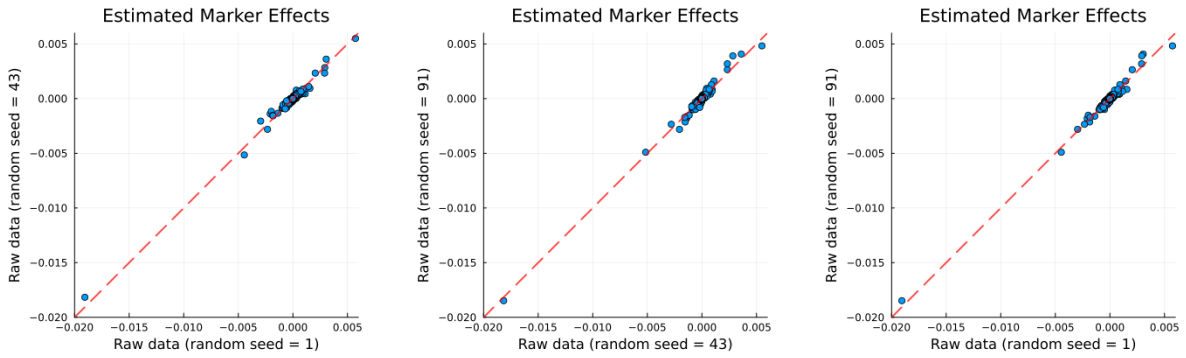
Table 2: **Pearson Correlation Coefficients for GWAS Results** between unencrypted and encrypted genotypes and phenotypes

2.3.3 Runtime

Using 50,000 MCMC iterations, the estimation of marker effects took approximately three times longer to run for the encrypted data than for the unencrypted data (19 hours versus 7 hours).

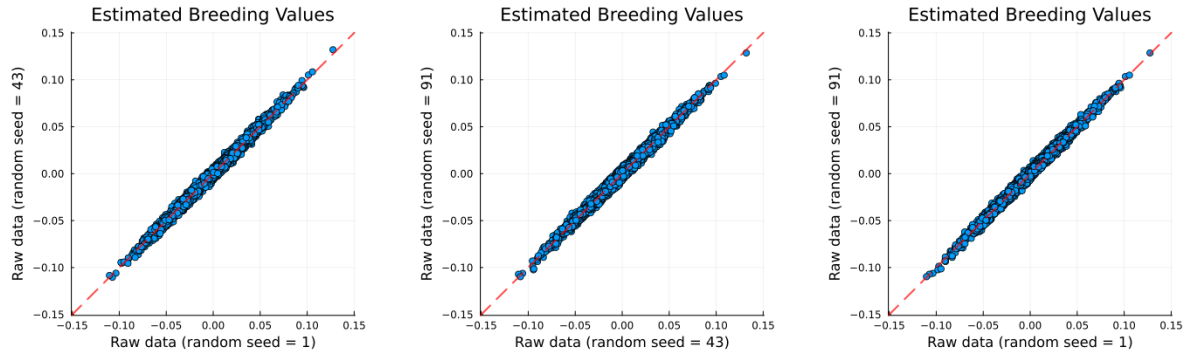
2.3.4 Monte Carlo Error

Figure 6 and Figure 7 present the marker effects and breeding values, respectively, estimated from the unencrypted data using different random seeds, and Table 3 provides the Pearson correlation coefficients corresponding to each subfigure.



(a) (b) (c)

Figure 6: **Scatter Plot of Marker Effects for Assessing Monte Carlo Error** (a) using random seed 1 and random seed 43 (b) using random seed 43 and random seed 91 (c) using random seed 43 and random seed 91. Each point represents a single estimated marker effect from the unencrypted data. The x-axes represent the estimated marker effects using one random seed, and the y-axes represent the estimated marker effects using the random seed to be compared. The dashed red line represents a correlation of 1.



(a) (b) (c)

Figure 7: **Scatter Plot of Breeding Values for Assessing Monte Carlo Error** (a) using random seed 1 and random seed 43 (b) using random seed 43 and random seed 91 (c) using random seed 43 and random seed 91. Each point represents a single estimated breeding value from the unencrypted data. The x-axes represent the estimated breeding values using one random seed, and the y-axes represent the estimated breeding values using the random seed to be compared. The dashed red line represents a correlation of 1.

Random Seeds Compared	Marker Effects Correlation	Breeding Values Correlation
1 and 43	0.9593	0.9956
43 and 91	0.9579	0.9955
1 and 91	0.9610	0.9963

Table 3: **Pearson Correlation Coefficients for Assessing Monte Carlo Error** of marker effects and breeding values estimated from unencrypted data using random seeds 1, 43, and 91

2.4 Discussion

As the correlation between the estimated marker effects for the unencrypted data and the encrypted data is relatively close to 1, it can be said that the estimated marker effects obtained are effectively identical. The same can be said about the estimated breeding values, which is also shown to be close to 1.

Therefore, the results show that the implementation of HEGP to a linear mixed-model produces virtually identical results and does not substantially affect the estimated marker effects nor the estimated breeding values. The comparison of the GWAS results of the unencrypted data and the encrypted data show that the percentages of genetic variance for each window obtained from both datasets are essentially the same, as their correlation approaches 1.

Based on the correlation between the GWAS results of the encrypted data from the results of Cheng *et al.* (2021), the HEGP method does not affect the precision of the obtained results.

2.4.1 Runtime

In terms of optimization, it would be beneficial to reduce the runtime of estimating marker effects of HEGP encrypted data, as it takes approximately three times longer to run to completion compared to that of the unencrypted data. This is due to limitations in the number of decimals allowed for values in the encryption matrix $\mathbf{P}_{all}$. During the

estimation of marker effects and breeding values, the inverse of $\mathbf{P}_{all}$ ($\mathbf{P}_{all}^{-1}$) multiplied by $\mathbf{P}_{all}$ should yield a sparse identity matrix, by the definition of an orthogonal matrix. However, each value in the encryption matrix is limited to 64-bits, so $\mathbf{P}_{all}^{-1}\mathbf{P}_{all}$ instead yields a matrix with small non-zero values where it should be zero. The non-sparse nature of this matrix causes the completion of the encrypted analysis to take more time. This can be remedied by removing small values when solving for the mixed-model equations. Further testing is needed to determine to appropriate threshold for a "small" value.

2.4.2 Monte Carlo Error

As the correlations between marker effects and breeding values estimated from the unencrypted genotypes and phenotypes using different random seeds are close to 1, the error from MCMC sampling is not large enough to substantially affect the estimations of these values.

2.4.3 Conclusion and Future Studies

The present study demonstrates that HEGP does not affect the estimated marker effects, estimated breeding values, nor the estimated percentages of genetic variance accounted for by each window. These results indicate that HEGP may be a useful tool for animal breeders if they intend to perform joint analyses but would like to ensure their data is protected. Mott *et al.* (2020) and Zhao *et al.* (2023) have discussed the security of HEGP and how various decryption methods would likely be ineffective. Future studies should include further validating HEGP as a robust data encryption method, by analyzing other datasets, such as those of other species or different traits.

There may also be potential in implementing HEGP to other types of analyses besides GWAS and genomic prediction. This study shows that HEGP is compatible with analyses using linear mixed models and BayesB. Zhao *et al.* (2023) demonstrated mathe-

matically that it can also be extended to other Bayesian variable selection methods, such as BayesCπ. Given the derivation of HEGP in Zhao *et al.* (2023), it is likely compatible with generalized linear mixed-models as well, but further studies are needed to validate the robustness of HEGP.

It is also important to highlight the accessibility of HEGP. Currently, there is no software available to easily implement HEGP. For widespread use after HEGP has been shown to be crytologically secure, it may be desirable to create a software that is able to preprocess data and implement HEGP in a streamlined manner.

2.5 Appendix

2.5.1 Funding

This research has been funded by AG2PI Seed Grant 2020-70412-32615 and 2021-70412-35233, and USDA-NIFA AG2PI grant 2023-70412-41054.

2.5.2 Data Availability

This study used genotypes and phenotypes from Cheng *et al.* (2021). The data analyzed in this study were collected on animals that were provided by and are part of the commercial breeding programs of the 7 investing member businesses of PigGen Canada (https://piggencanada.org/, last accessed 12/30/2021). As such, the data and samples generated on these animals are confidential and protected as intellectual property or as trade secrets. As a result, the data analyzed in this study are not publicly available but are stored in a secure data base at the University of Alberta. Data can, however, be made available on reasonable request, as detailed in supplementary file "Data access procedure" in Cheng *et al.* (2021). Cheng *et al.* (2021) was funded by Genome Canada, Genome Alberta, Genome Prairie, PigGen Canada, and USDA-NIFA grant number 2017-67007-26144. Members of PigGen Canada are acknowledged for providing the pigs and

for helpful discussions, including Canadian Centre for Swine Improvement, Fast Genetics, Genesus, Hypor, ALPHAGENE, Topigs Norsvin, DNA Genetics, the Canadian Swine Breeders Association, and Alliance Genetics Canada.

2.5.3 Code Availability

The necessary scripts are available at `https://github.com/dli10/data-encryption`.

References

Abdellaoui, A., L. Yengo, K. J. Verweij, and P. M. Visscher, 2023 15 years of gwas discovery: Realizing the promise. The American Journal of Human Genetics **110**: 179–194.

Blatt, M., A. Gusev, Y. Polyakov, and S. Goldwasser, 2020a Secure large-scale genome-wide association studies using homomorphic encryption. Proceedings of the National Academy of Sciences **117**: 11608–11613.

Blatt, M., A. Gusev, Y. Polyakov, K. Rohloff, and V. Vaikuntanathan, 2020b Optimized homomorphic encryption solution for secure genome-wide association studies. BMC Medical Genomics **13**: 83.

Burton, P. R., D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, *et al.*, 2007 Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447**: 661–678.

Calus, M., 2010 Genomic breeding value prediction: methods and procedures. Animal **4**: 157–164.

Cheng, H., R. Fernando, D. Garrick, *et al.*, 2018 Jwas: Julia implementation of whole-genome analysis software. In *Proceedings of the world congress on genetics applied to livestock production*, volume 11, p. 859, World Congress on Genetics Applied to Livestock Production.

Cheng, H., R. Fernando, D. Garrick, T. Zhao, and J. Qu, 2022 Jwas version 2: leveraging biological information and highthroughput phenotypes into genomic prediction and association. In *Proceedings of 12th World Congress on Genetics Applied to Livestock Production (WCGALP) Technical and species orientated innovations in animal breeding, and contribution of genetics to solving societal challenges*, pp. 1519–1522, Wageningen Academic Publishers.

Cheng, J., R. Fernando, H. Cheng, S. D. Kachman, and K. Lim, 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 **12**: jkab441.

Cheng, J., A. M. Putz, J. C. S. Harding, M. K. Dyck, F. Fortin, *et al.*, 2020 Genetic analysis of disease resilience in wean-to-finish pigs from a natural disease challenge model. Journal of Animal Science **98**: skaa244.

Cho, H., D. J. Wu, and B. Berger, 2018 Secure genome-wide association analysis using multiparty computation. Nature Biotechnology **36**: 547–551.

Cleveland, M. A., J. M. Hickey, and S. Forni, 2012 A common dataset for genomic analysis of livestock populations. G3: Genes—Genomes—Genetics **2**: 429–435.

DeWan, A., M. Liu, S. Hartman, S. S.-M. Zhang, D. T. L. Liu, *et al.*, 2006 Htra1 promoter polymorphism in wet age-related macular degeneration. Science **314**: 989–992.

Fernando, R. L., J. C. Dekkers, and D. J. Garrick, 2014 A class of bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genetics Selection Evolution **46**: 50.

Gentry, C., 2009a A fully homomorphic encryption scheme .

Gentry, C., 2009b Fully homomorphic encryption using ideal lattices. Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC'09) pp. 169–178.

Gudodagi, R. and R. V. S. Reddy, 2022 Encryption and decryption of secure data for diverse genomes. Lecture Notes in Electrical Engineering **836**: 505–514.

Habier, D., R. L. Fernando, K. Kizilkaya, and D. J. Garrick, 2011 Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics **12**: 186.

Hoff, P., 2021 rstiefel: Random orthonormal matrix generation and optimization on the stiefel manifold. R package version 1.0.1, URL https://CRAN.R-project.org/package=rstiefel .

Hoff, P. D., 2009 Simulation of the matrix bingham–von mises–fisher distribution, with applications to multivariate and relational data. Journal of Computational and Graphical Statistics **18**: 438–456.

Klein, R. J., C. Zeiss, E. Y. Chew, J.-Y. Tsai, R. S. Sackler, *et al.*, 2005 Complement factor h

polymorphism in age-related macular degeneration. Science **308**: 385–389.

Loukides, G., A. Gkoulalas-Divanis, and B. Malin, 2010 Anonymization of electronic medical records for validating genome-wide association studies. Proceedings of the National Academy of Sciences **107**: 7898–7903.

Lu, W.-J., Y. Yamada, and J. Sakuma, 2015 Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. BMC Medical Informatics and Decision Making **15**: S1.

Malin, B., G. Loukides, K. Benitez, and E. W. Clayton, 2011 Identifiability in biobanks: models, measures, and mitigation strategies. Human Genetics **130**: 383.

Malin, B. and L. Sweeney, 2004 How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. Journal of Biomedical Informatics **37**: 179–192.

Meuwissen, T., 2007 Genomic selection: marker assisted selection on a genome wide scale. Journal of Animal Breeding and Genetics **124**: 321–322.

Meuwissen, T., B. Hayes, and M. Goddard, 2016 Genomic selection: A paradigm shift in animal breeding. Animal Frontiers **6**: 6–14.

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819–1829.

Mott, R., C. Fischer, P. Prins, and R. W. Davies, 2020 Private genomes and public snps: homomorphic encryption of genotypes and phenotypes for shared quantitative genetics. Genetics **215**: 359–372.

Ogburn, M., C. Turner, and P. Dahal, 2013 Homomorphic encryption. Procedia Computer Science **20**: 502–509.

Putz, A. M., J. C. S. Harding, M. K. Dyck, F. Fortin, G. S. Plastow, *et al.*, 2019 Novel resilience phenotypes using feed intake data from a natural disease challenge model in wean-to-finish pigs. Frontiers in Genetics **9**: 660.

Tuggle, C. K., J. Clarke, J. C. M. Dekkers, D. Ertl, C. J. Lawrence-Dill, *et al.*, 2022 The

agricultural genome to phenome initiative (ag2pi): creating a shared vision across crop and livestock research communities. Genome Biology **23**: 3.

Wood, A., K. Najarian, and D. Kahrobaei, 2020 Homomorphic encryption for machine learning in medicine and bioinformatics. ACM Computing Surveys (CSUR) **53**: 1–35.

Yang, J., T. Ferreira, A. P. Morris, S. E. Medland, G. I. o. A. T. G. Consortium, *et al.*, 2012 Conditional and joint multiple-snp analysis of gwas summary statistics identifies additional variants influencing complex traits. Nature Genetics **44**: 369–375.

Yengo, L., J. Sidorenko, K. E. Kemper, Z. Zheng, A. R. Wood, *et al.*, 2018 Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. Human Molecular Genetics **27**: 3641–3649.

Yengo, L., S. Vedantam, E. Marouli, J. Sidorenko, E. Bartell, *et al.*, 2022 A saturated map of common genetic variants associated with human height. Nature **610**: 704–712.

Zhao, T., F. Wang, R. Mott, J. Dekkers, and H. Cheng, 2023 Using encrypted genotypes and phenotypes for collaborative genomic analyses to maintain data confidentiality. Genetics **226**: iyad210.