OXFORD GENETICS

Using encrypted genotypes and phenotypes for collaborative genomic analyses to maintain data confidentiality

Tianjing Zhao[1,2], Fangyi Wang[3], Richard Mott[4], Jack Dekkers[5] and Hao Cheng[1,*]

[1]Department of Animal Science, University of California, Davis, CA 95616, USA
[2]Department of Animal Science, University of Nebraska-Lincoln, Lincoln, NE 68583, USA
[3]Department of Plant Sciences, University of California, Davis, CA 95616, USA
[4]Genetics Institute, University College London, WC1E 6BT, UK
[5]Department of Animal Science, Iowa State University, IA 50011, USA

[*]Department of Animal Science, University of California, Davis, CA, USA E-mail: qtlcheng@ucdavis.edu.

Abstract

To adhere to and capitalize on the benefits of the FAIR (Findable, Accessible, Interoperable and Reusable) principles in agricultural genome-to-phenome studies, it is crucial to address privacy and intellectual property issues that prevent sharing and reuse of data in research and industry. Direct sharing of genotype and phenotype data is often prohibited due to intellectual property and privacy concerns. Thus there is a pressing need for encryption methods that obscure confidential aspects of the data, without affecting the outcomes of certain statistical analyses. A homomorphic encryption method for genotypes and phenotypes (HEGP) has been proposed for single-marker regression in genome-wide association studies using linear mixed models with Gaussian errors. This methodology permits frequentist likelihood-based parameter estimation and inference. In this paper, we extend HEGP to broader applications in genome-to-phenome analyses. We show that HEGP is suited to commonly used linear mixed models for genetic analyses of quantitative traits including GBLUP and RR-BLUP, as well as Bayesian variable selection methods (e.g., those in Bayesian Alphabet), for genetic parameter estimation, genomic prediction, and genome-wide association studies. By advancing the capabilities of HEGP, we offer researchers and industry professionals a secure and efficient approach for collaborative genomic analyses while preserving data confidentiality.

Keywords: homomorphic encryption; mixed model; genomic prediction; GWAS; joint analysis

Introduction

To conform to and capitalize on the benefits of the FAIR (Findable, Accessible, Interoperable and Reusable) principles in agricultural genome-to-phenome studies, it is necessary to address privacy and intellectual property issues that may prevent sharing and reuse of data in research and industry. First, sharing and reuse of genotypic and phenotypic data enables reproducible research, where researchers can confirm published analyses with minimal effort. Second, for traits that are hard or expensive to measure, a single research group may have limited data for genetic analysis, which may lead to less reliable and underpowered results. This problem may be alleviated by joint analyses that include data from multiple contributors.

Although data sharing and reuse will bring significant benefits to genome-to-phenome studies in both academia and industry, it is often prohibitive to directly share raw genotype and phenotype data due to privacy concerns, commercial interests, and data sharing policies, and because the risks of sharing raw data may not be fully understood by the data owners. For example, although individual identifiers can be anonymized, information about an anonymized individual might still be disclosed by comparing its genotypes to known genotyped relatives. To avoid the concerns about sharing raw data, consortia are often established, and raw data are only shared with members of the consortium or with researchers who are approved for access. In other cases, external researchers may perform analysis on the data owner's computer system without access to the raw data. These approaches, however, still pose risks to privacy and intellectual property, hampering widespread data sharing and reuse.

Homomorphic encryption (HE) refers to a type of encryption of raw data (hereafter referred to as "plaintext") in a manner that obscures confidential aspects of the data, while certain computations on the encrypted data (hereafter referred to as "cyphertext") match the results from the plaintext, when decrypted. While several methods for homomorphic encryption have been proposed for genomic analysis, most limit the types of computations and analyses that can be conducted on the encrypted data (cyphertext). For example, for case-control GWAS, HE schemes were proposed to calculate allelic chi-square test and perform logistic regression (Lu *et al.* 2015; Chen *et al.* 2018; Sim *et al.* 2020; Blatt *et al.* 2020). However, these methods ignore random and fixed effects that account for family and population admixture. Although linear mixed models are widely used in genetic analyses such as genomic prediction and GWAS (Bradbury *et al.* 2007), the use of HE for mixed models is scarce.

Recently, Mott et al. (Mott *et al.* 2020b) proposed an encryption method, called homomorphic encryption for genotypes and phenotypes (HEGP), that is specifically suited to single-marker

regression in GWAS using linear mixed models with Gaussian errors. HEGP is based on high-dimensional random orthogonal transformations of the plaintext that encrypts phenotypes, genotypes, and specified covariates by replacing them with random linear superpositions, such that cyphertext genotypes and phenotypes cannot be linked back to individual identifiers. HEGP preserves linkage disequilibrium between markers but scrambles the genomic relationship between individuals. Moreover, under a linear mixed model with Gaussian errors, the likelihood of the cyphertext is unchanged, such that the encryption does not affect the outcomes of single-marker regression in GWAS analyses. HEGP differs conceptually from other HE methods in that some outputs (particularly the parameter estimates) are unaffected by encryption and do not need to be decrypted.

In this paper, we extend the HEGP scheme for wider application in genome-to-phenome analyses. We demonstrate that HEGP can be effectively applied to many popular mixed models, beyond single-marker regression. These models, including Bayesian variable selection methods such as those in Bayesian Alphabet, are routinely employed in the fields of animal and crop improvement, for genetic analyses of quantitative traits, including genetic parameter estimation, genomic prediction, and GWAS. We show how most of the quantitative genetics toolbox used by animal and plant breeders can be integrated with data-sharing protocols and performed while protecting important types of potentially confidential or commercially sensitive information.

Materials and methods

Homomorphic encryption using high-dimensional random orthogonal matrix

We will use $\mathbf{y}$, a vector of length n, to denote the plaintext phenotypes for n observations, and $\mathbf{M}$ to denote the $n \times p$ plaintext genotype covariate matrix for the n observations across p SNPs. To infer unknowns in mixed models, these quantities will typically be used in the multiplicative forms $\mathbf{M}^T\mathbf{M}$ and $\mathbf{M}^T\mathbf{y}$. Thus, intuitively, any data encryption scheme that leaves the above multiplications unchanged would produce the same GWAS and genomic prediction outcomes.

HEGP uses a high-dimensional random $n \times n$ orthogonal matrix $\mathbf{P}$, such that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ and the determinant $|\mathbf{P}| = 1$. The suitable choices of $\mathbf{P}$ for the purpose of encryption are discussed in a later section. The plaintext genotypes and phenotypes are encrypted as

$$\begin{aligned} \mathbf{M}^* &= \mathbf{PM}, \\ \mathbf{y}^* &= \mathbf{Py}. \end{aligned} \tag{1}$$

because

$$\mathbf{M}^T\mathbf{M} = \mathbf{M}^T\mathbf{P}^T\mathbf{PM} = (\mathbf{PM})^T(\mathbf{PM}) = \mathbf{M}^{*T}\mathbf{M}^* \tag{2}$$

and

$$\mathbf{M}^T\mathbf{y} = \mathbf{M}^T\mathbf{P}^T\mathbf{Py} = (\mathbf{PM})^T\mathbf{y}^* = \mathbf{M}^*\mathbf{y}^* \tag{3}$$

In contrast to other methods of homomorphic encryption, the outputs of HEGP (i.e., marker effect estimates and p-values) are automatically plaintext, regardless of whether the inputs are plaintext or cyphertext. This means that there is no need to decrypt the outputs such as marker effect estimates and hence no decryption key is distributed. In a later section we will show that the HEGP does not affect the inference of marker effects, thus with the plaintext of genotypes, the estimated breeding values (EBV) can be calculated. Otherwise, the EBV calculated from the cybertext of genotypes remains cybertext EBV.

Conceptual overview Figure 1 illustrates HEGP for a small example of 4 individuals (a-d) and 6 SNPs. By multiplying by the random orthogonal matrix $\mathbf{P}$, phenotypes and genotypes in the encrypted data become 'random' linear combinations of phenotypes and genotypes of the original four individuals a-d.

Figure 2 compares plaintext and cyphertext genotypes from a larger pig data set (Cleveland *et al.* 2012). Figure 2(a) and (b) present the heat maps of the plaintext and cyphertext genotypes. For the plaintext, each row represents an individual and each column represents the genotypes for a SNP across individuals. As shown in Figure 2(c) and Figure 2(d), after encryption, the genotypes transform from trimodal values (0/1/2) to continuous values that closely resemble a sample from a normal distribution.

The set of orthogonal $n \times n$ matrices forms a group under multiplication and includes the identity matrix, which is clearly ineffective for encryption. Therefore it is necessary for $\mathbf{P}$ to be randomly generated and independent of the plaintext. A suitable method is derived from the Stiefel manifold, or Haar measure (Hoff 2009; Chikuse and Chikuse 2003), which is measure-preserving, meaning that the measure (loosely speaking, the sampling probability) of any data matrix $\mathbf{M}$ is the same as the measure of $\mathbf{PM}$. For this method, first an $n \times n$ matrix $\mathbf{B}$ is generated, whose entries are sampled independently from a standard normal distribution. Next, an $n \times n$ random orthogonal matrix is generated as $\mathbf{P} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-\frac{1}{2}}$. In detail, $(\mathbf{B}^T\mathbf{B})^{-\frac{1}{2}}$ is computed as $\mathbf{Q}\mathbf{\Lambda}^{-\frac{1}{2}}\mathbf{Q}^T$, where $\mathbf{Q}$ and $\mathbf{\Lambda}$ are obtained from the eigen decomposition of $\mathbf{B}^T\mathbf{B}$, i.e., $\mathbf{B}^T\mathbf{B} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$. Matrix $\mathbf{P}$ is easily seen to be orthogonal (by checking that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$) and furthermore can be shown to be randomly sampled from the Stiefel manifold. The R package rstiefel (Hoff 2012) can be used to generate $\mathbf{P}$.

Relationships between SNPs and between individuals HEGP preserves relationships between genotypes (i.e., linkage disequilibrium, r^2), but scrambles relationships between individuals. Any orthogonal transformation preserves the dot product of two vectors and, geometrically, acts as a rotation of a hypersphere in which SNP genotype vectors and phenotype vectors are represented as points on its surface. The cosine of the angle between any pair of points subtended at the origin equals their Pearson correlation coefficient, or dot product and a rotation merely changes the coordinate system while leaving angles unchanged. In HEGP, all marker genotypes (and phenotypes) are rotated by the same orthogonal matrix as $\mathbf{PM} = [\mathbf{Pm}_1, ..., \mathbf{Pm}_p]$. Thus, the LD between jth and kth marker is preserved since

$$\begin{aligned} (\mathbf{m}_j^*)^T(\mathbf{m}_k^*) &= (\mathbf{Pm}_j)^T(\mathbf{Pm}_k) \\ &= \mathbf{m}_j^T\mathbf{m}_k \end{aligned} \tag{4}$$

where we use the fact that for any orthogonal matrix, $\mathbf{P}^T\mathbf{P} = \mathbf{I}$. For illustration, the LD matrices based on the raw and the encrypted genotypes for 5,000 markers in the pig dataset of Cleveland *et al.* (2012) are shown in Figure 3(a). The LD matrix is calculated as $\frac{1}{n}\mathbf{H}^T\mathbf{H}$, where $\mathbf{H}$ is the normalized genotype matrix. The jth marker of the ith individual is normalized as $H_{i,j} = \frac{M_{i,j}-2p_j}{\sqrt{2p_j(1-p_j)}}$, where p_j is the allele frequency. In Figure 3(a), the two LD matrices are almost identical, and the correlation between elements in the two LD matrices is 1.0.

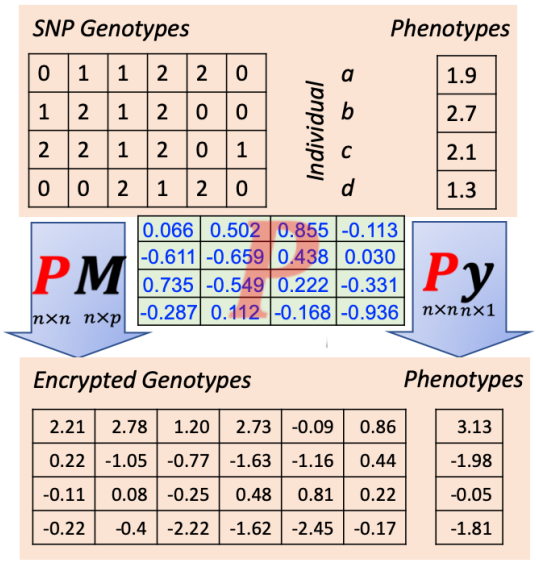

Figure 1 Illustration of homomorphic encryption for genotypes and phenotypes. For 4 individuals (a-d), the raw genotypes for 6 SNPs and the raw phenotypes are provided in the upper part of the figure. The raw data are encrypted by pre-multiplying with a random orthogonal encryption matrix $\mathbf{P}$, which is displayed in the middle of the figure. The encrypted phenotypes and genotypes, shown in the lower part of the figure, are "random" linear combinations of the raw phenotypes and genotypes of the original four individuals a-d.

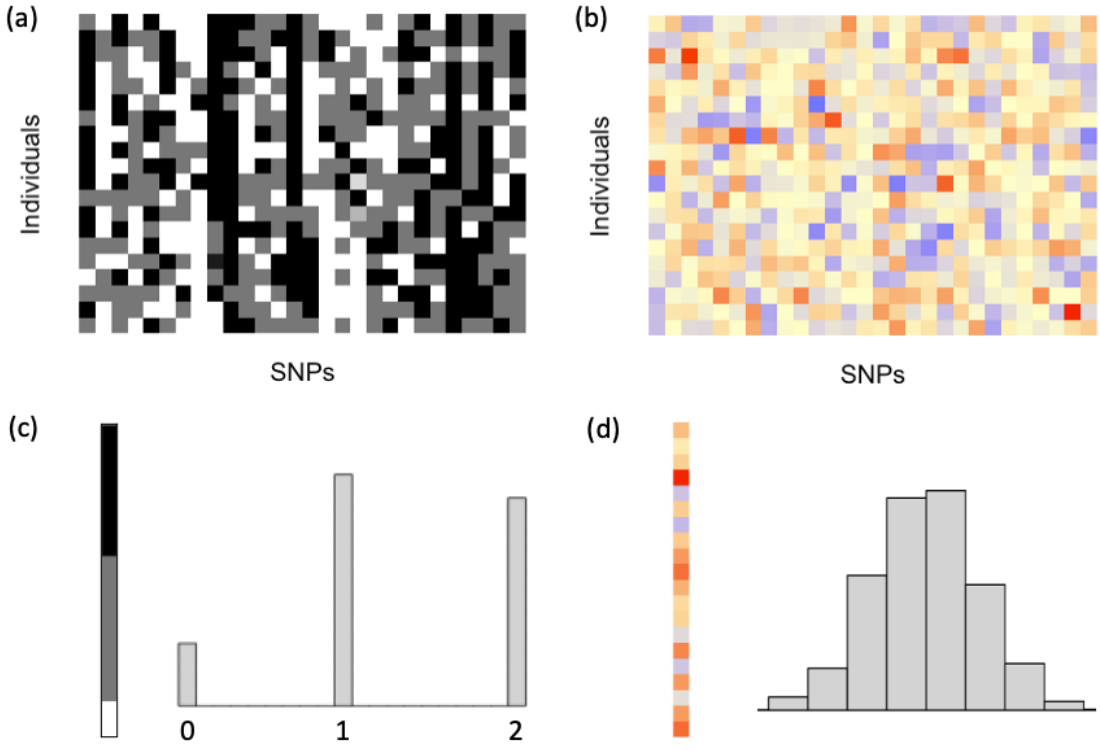

Figure 2 (a) A subset of pig genotypes provided in the data set of Cleveland *et al.* (2012). Genotypes are coded as 0, 1, and 2, which are presented by white, grey, and black colors, respectively. Each row represents one individual, and each column represents one marker. (b) The corresponding encrypted genotypes, encrypted via a high-dimension random orthogonal matrix. The encrypted genotypes are a continuum of real numbers presented by different colors. (c) Sorted genotypes of one marker, coded as 0/1/2 (left), and its trimodal distribution (right). (d) The corresponding encrypted genotypes (same order as in (c)) and their bell-shaped distribution.

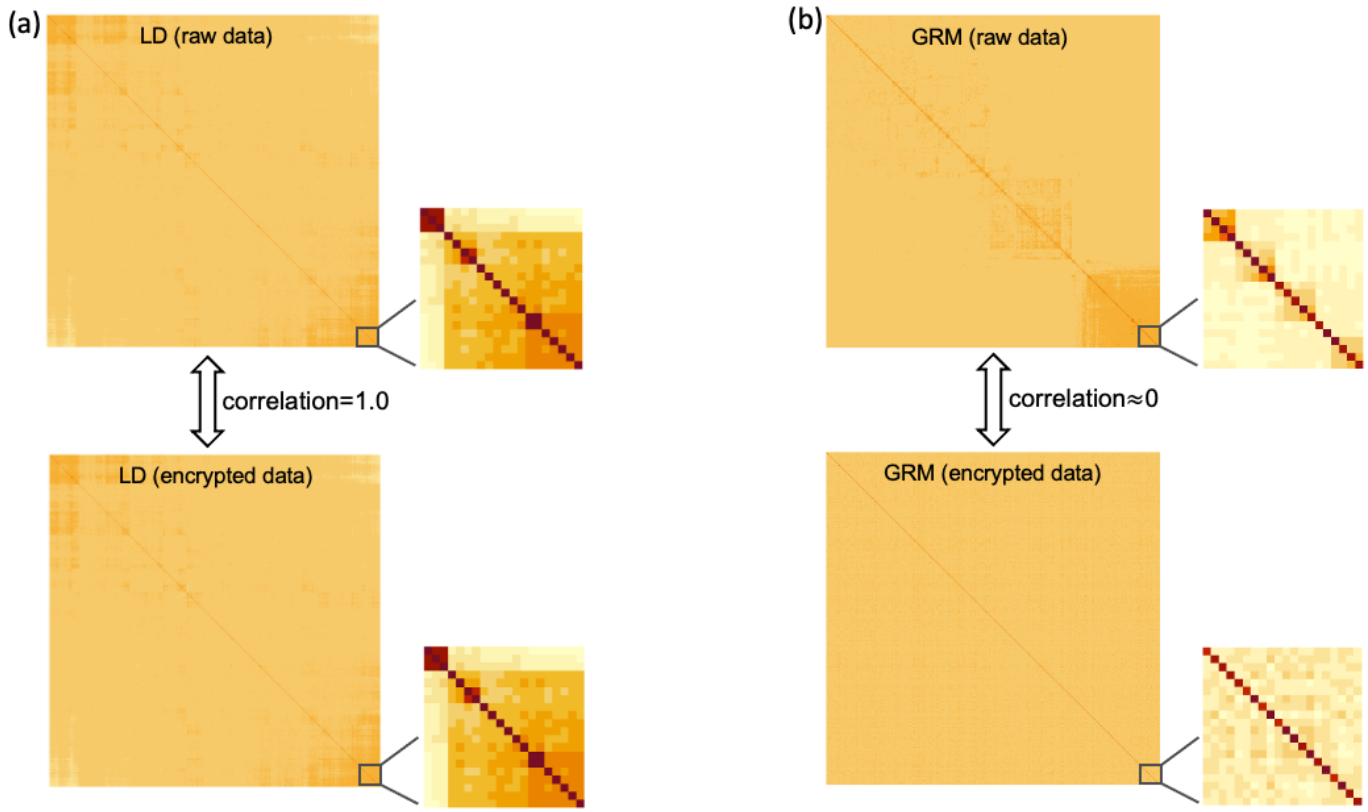

Figure 3 (a) Linkage disequilibrium (LD) matrix calculated using raw genotypes (up), and encrypted genotypes (down). The LD matrix is preserved using encrypted data, and the correlation between the two LD matrices is 1.0. (b) Genomic relationship matrix (GRM) calculated using the raw genotypes (up), and encrypted genotypes (down). The GRM is scrambled using encrypted data, and the correlation between two GRM matrices is close to 0.

In contrast to LD relationships, HEGP scrambles relationships between individuals since $(\mathbf{PM})(\mathbf{PM})^T = \mathbf{P}(\mathbf{MM}^T)\mathbf{P}^T$ and, after transformation, individual records are random linear combinations of the original records. For demonstration, genomic relationship matrices (GRM), calculated as $\frac{1}{p}\mathbf{HH}^T$, based on plaintext and cyphertext genotypes, are shown in Figure 3(b) for a subset of the pig dataset. The elementwise correlation between the two GRM is ~ 0.

Statistical preliminaries

As we will demonstrate, in addition to single-marker regression for GWAS using linear mixed models with Gaussian errors, HEGP is compatible with most genetic analyses that use mixed models, including GBLUP, SNP-BLUP, Bayesian Alphabet, and others.

Mixed models The mixed model is a cornerstone for many quantitative genetic analyses, including genetic parameter estimation, genomic prediction, and GWAS (Meuwissen *et al.* 2001; VanRaden 2008; Hayes *et al.* 2009; Fernando *et al.* 2017; Wang *et al.* 2012, 2016; Moser *et al.* 2015a; Legarra *et al.* 2018). In particular, GBLUP (Habier *et al.* 2007; VanRaden 2008; Hayes *et al.* 2009) is one of the most widely used linear mixed models for genomic prediction. The GBLUP model can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \mathbf{e} \tag{5}$$

where $\mathbf{y}$ is a vector of phenotypes of length n, and $\mathbf{X}$ is the incidence matrix for non-genetic fixed effects, denoted by $\boldsymbol{\beta}$. Vector $\mathbf{u}$ contains additive genetic values for n individuals and follows a multivariate normal distribution $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigma_u^2)$, where $\mathbf{G}$ is the genomic relationship matrix proportional to $\mathbf{MM}^T$, where $\mathbf{M}$ is an $n \times p$ genotype covariate matrix, and σ_u^2 is the genetic variance. Vector $\mathbf{e}$ includes n random residuals and follows a normal distribution, $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where σ_e^2 is the residual variance. Narrow sense heritability is defined to be $h^2 = \frac{\sigma_u^2}{(\sigma_u^2+\sigma_e^2)}$.

The GBLUP model is equivalent to the following marker effects model (hereinafter referred to as SNP-BLUP) in terms of predicting genetic values (Fernando 1998; Habier *et al.* 2007; Strandén and Garrick 2009):

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{M}\boldsymbol{\alpha} + \mathbf{e} \tag{6}$$

where $\boldsymbol{\alpha}$ is a vector of p additive marker effects, with $\boldsymbol{\alpha} \sim N(\mathbf{0}, \mathbf{I}\sigma_\alpha^2)$. The same point estimates of marker effects $\hat{\boldsymbol{\alpha}}$ can be obtained from the estimated genetic values $\hat{\mathbf{u}}$ in GBLUP as $\hat{\boldsymbol{\alpha}} = \mathbf{M}^T(\mathbf{MM}^T)^{-1}\hat{\mathbf{u}}$.

The Gaussian prior distribution of marker effects in SNP-BLUP is just one, analytically tractable, member of the "Bayesian Alphabet", in which a range of prior distributions, reflecting different assumptions about the genetic architecture of the trait, are assigned to the marker effects (Meuwissen *et al.* 2001; Kizilkaya *et al.* 2010; Habier *et al.* 2011; Erbe *et al.* 2012; Moser *et al.* 2015b; Park and Casella 2008; Gianola and Fernando 2019). For example, it is sometimes desirable to model the majority of marker effects as being zero, and to allow occasional markers with large effects. For some traits, such priors are more biologically meaningful than SNP-BLUP and have been widely used in genomic prediction and genome-wide association studies.

In this paper, we demonstrate the effectiveness of HEGP using both SNP-BLUP and BayesCπ (Kizilkaya *et al.* 2010; Habier *et al.* 2011). BayesCπ is a representative of the other "Bayesian Alphabet" models, so the extension of HEGP to other priors for marker effects (Meuwissen *et al.* 2001; Erbe *et al.* 2012; Moser *et al.* 2015b; Park and Casella 2008; Gianola and Fernando 2019) does not present further challenges. BayesCπ (Kizilkaya *et al.* 2010; Habier *et al.* 2011) is typical in that it assigns mixture priors to marker effects, which are multiplied by the Gaussian likelihood of the data to generate the posterior. The BayesCπ model must be fitted using Gibbs sampling, and therefore we must show that HEGP does not perturb the algorithm and produces numerically stable and accurate estimates.

At each step of the Gibbs sampler, a given unknown is sampled from its full conditional posterior distributions given the latest sampled values of all other unknowns. Below we will show that the full conditional posterior distributions of marker effects are identical when using raw or encrypted data, such that the same posterior distributions will be obtained (this also holds for other parameters of interest). Derivations for other parameters of interest in SNP-BLUP and BayesCπ can be found in the Appendix.

Unchanged likelihood using HEGP

In HEGP, the plaintext phenotypes, covariates, and genotype dosages, and the design matrix for fixed effects are encrypted by pre-multiplication by the same random orthogonal matrix, $\mathbf{P}$. The mixed model using cyphertext for both SNP-BLUP and BayesCπ can be written as

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{M}^*\boldsymbol{\alpha} + \mathbf{e}^*, \tag{7}$$

In this model, $\mathbf{y}^* = \mathbf{Py}$ are the encrypted phenotypes, $\mathbf{M}^* = \mathbf{PM} = \mathbf{u}^*$ the encrypted genotypes, and $\mathbf{X}^* = \mathbf{PX}$ the encrypted design matrix for fixed effects. After encryption, the residual variance remains unchanged, represented as $\mathbf{var}(\mathbf{e}^*) = \mathbf{var}(\mathbf{Pe}) = \mathbf{P}^T\mathbf{I}\sigma_e^2\mathbf{P} = \mathbf{I}\sigma_e^2$. The genetic variance becomes $\mathbf{var}(\mathbf{u}^*) = \mathbf{P}^T\mathbf{GP}\sigma_u^2$ after encryption.

We next show that the likelihood of the data is invariant under orthogonal transformation. Define the plaintext variance matrix $\mathbf{V} = \mathbf{G}\sigma_u^2 + \mathbf{I}\sigma_e^2$ and its cyphertext equivalent $\mathbf{V}^* = \mathbf{P}^T(\mathbf{G}\sigma_u^2 + \mathbf{I}\sigma_e^2)\mathbf{P}$. Then the determinant of the variance matrix is invariant because $|\mathbf{V}^*| = |\mathbf{P}^T\mathbf{VP}| = |\mathbf{P}^T||\mathbf{V}||\mathbf{P}| = |\mathbf{V}|$ and hence the Gaussian log-likelihood of the plaintext ($\log L$) equals that of the cyphertext ($\log L^*$):

$$\begin{aligned}
-2\log L(\beta) &= (\mathbf{y}-\mathbf{X}\beta)^T\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\beta) + n\log(|\mathbf{V}|) \\
&= (\mathbf{y}-\mathbf{X}\beta)^T(\mathbf{P}^T\mathbf{P})\mathbf{V}^{-1}(\mathbf{P}^T\mathbf{P})(\mathbf{y}-\mathbf{X}\beta) + n\log(|\mathbf{V}|) \\
&= (\mathbf{P}(\mathbf{y}-\mathbf{X}\beta))^T(\mathbf{P}^T\mathbf{VP})^{-1}(\mathbf{P}(\mathbf{y}-\mathbf{X}\beta)) + n\log(|\mathbf{V}^*|) \\
&= (\mathbf{y}^*-\mathbf{X}^*\beta)^T\mathbf{V}^{*-1}(\mathbf{y}^*-\mathbf{X}^*\beta) + n\log(|\mathbf{V}^*|) \\
&= -2\log L^*(\beta)
\end{aligned} \tag{8}$$

Hence all parameter inference in SNP-BLUP is invariant under orthogonal transformation in the mixed model, resulting in unchanged estimates for β, for the variance components σ_e^2, σ_u^2 and heritability h^2.

Inference of unknowns in mixed model

In BayesCπ, the prior for the marker effects is a mixture of a point mass at zero and a univariate normal distribution with a null mean and a common locus variance σ_α^2. The full conditional posterior distribution of the marker effect for locus j when it is

non-zero (i.e., the full conditional posterior distribution of the marker effect for locus j in SNP-BLUP) can be expressed as

$$\left(\alpha_j \mid ELSE\right) \sim N\left(\hat{\alpha}_j, \frac{\sigma_e^2}{\mathbf{m}_j^T\mathbf{m}_j + \frac{\sigma_e^2}{\sigma_\alpha^2}}\right), \tag{9}$$

where $ELSE$ stands for all the other parameters and $\hat{\alpha}_j$ is the solution to

$$\left(\mathbf{m}_j^T\mathbf{m}_j + \frac{\sigma_e^2}{\sigma_\alpha^2}\right)\hat{\alpha}_j = \mathbf{m}_j^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{j'\neq j}\mathbf{m}_{j'}\alpha_{j'}\right). \tag{10}$$

When encrypted genotypic and phenotypic data are used, the full conditional posterior distribution of α_j, when it is non-zero, can be written as

$$\left(\alpha_j \mid ELSE\right) \sim N\left(\hat{\alpha}_j^*, \frac{\sigma_e^2}{(\mathbf{m}_j^*)^T(\mathbf{m}_j^*) + \frac{\sigma_e^2}{\sigma_\alpha^2}}\right), \tag{11}$$

where $ELSE$ stands for all the other parameters, and $\hat{\alpha}_j^*$ is the solution to

$$\begin{aligned}\left((\mathbf{m}_j^*)^T(\mathbf{m}_j^*) + \frac{\sigma_e^2}{\sigma_\alpha^2}\right)\hat{\alpha}_j^* &= (\mathbf{m}_j^*)^T\left(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta} - \sum_{j'\neq j}\mathbf{m}_{j'}^*\alpha_{j'}\right)\\ &= (\mathbf{m}_j^*)^T\mathbf{y}^* - (\mathbf{m}_j^*)^T\mathbf{X}^*\boldsymbol{\beta} - \sum_{j'\neq j}(\mathbf{m}_j^*)^T\mathbf{m}_{j'}^*\alpha_{j'}\end{aligned} \tag{12}$$

We have previously shown that $(\mathbf{m}_j^*)^T(\mathbf{m}_k^*) = \mathbf{m}_j^T\mathbf{m}_k$. Similarly,

$$\begin{aligned}(\mathbf{m}_j^*)^T\mathbf{y}^* &= (\mathbf{P}\mathbf{m}_j)^T\mathbf{P}\mathbf{y}\\ &= \mathbf{m}_j^T\mathbf{y},\end{aligned} \tag{13}$$

and

$$\begin{aligned}(\mathbf{m}_j^*)^T\mathbf{X}^* &= (\mathbf{P}\mathbf{m}_j)^T\mathbf{P}\mathbf{X}\\ &= \mathbf{m}_j^T\mathbf{X}.\end{aligned} \tag{14}$$

Therefore, the full conditional posterior distribution of α_j using cyphertext, as per equations (11) and (12), is identical to that obtained using the plaintext, as shown in equations (9) and (10). Thus, because HEGP does not change the full conditional posterior distributions in Gibbs sampling, the posterior distributions of marker effects are also identical using plaintext or cyphertext. The same conclusion holds for all other parameters of interest (see Appendix). Note that once estimates of marker effects are obtained, the plaintext of genotypes, if available, should be used to calculate the estimated breeding values.

Joint analysis using encrypted data from multiple contributors

A single research study may only contain a limited amount of data that is underpowered for genetic analysis. This issue can be mitigated through joint analyses using data from multiple studies, e.g. Yengo *et al.* (2022a). An attractive feature of HEGP is that it allows each component of the joint data to be encrypted independently. Thus, each contributor generates its own private key and uses it to encrypt its own plaintext prior to sharing it for joint analysis. The keys are never shared. The process for joint analysis then proceeds as described in the following.

For clarity, let's assume there are three contributors. Assuming variance components, such as the marker effect variance and the residual variance, are identical for all parties, the mixed model for the joint analysis of cyphertext can be written as

$$\begin{bmatrix}\mathbf{P}_1\mathbf{y}_1\\ \mathbf{P}_2\mathbf{y}_2\\ \mathbf{P}_3\mathbf{y}_3\end{bmatrix} = \begin{bmatrix}\mathbf{P}_1\mathbf{X}_1\\ \mathbf{P}_2\mathbf{X}_2\\ \mathbf{P}_3\mathbf{X}_3\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{P}_1\mathbf{M}_1\\ \mathbf{P}_2\mathbf{M}_2\\ \mathbf{P}_3\mathbf{M}_3\end{bmatrix}\boldsymbol{\alpha} + \begin{bmatrix}\mathbf{P}_1\mathbf{e}_1\\ \mathbf{P}_2\mathbf{e}_2\\ \mathbf{P}_3\mathbf{e}_3\end{bmatrix}, \tag{15}$$

where the matrices related to the t-th contributor are labeled with subscript "t". This equation can be re-written as

$$\mathbf{P}\begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \mathbf{y}_3\end{bmatrix} = \mathbf{P}\begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\\ \mathbf{X}_3\end{bmatrix}\boldsymbol{\beta} + \mathbf{P}\begin{bmatrix}\mathbf{M}_1\\ \mathbf{M}_2\\ \mathbf{M}_3\end{bmatrix}\boldsymbol{\alpha} + \mathbf{P}\begin{bmatrix}\mathbf{e}_1\\ \mathbf{e}_2\\ \mathbf{e}_3\end{bmatrix}, \tag{16}$$

where $\mathbf{P}$ is a block-diagonal orthogonal matrix

$$\mathbf{P} = \begin{bmatrix}\mathbf{P}_1 & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{P}_2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{P}_3\end{bmatrix}. \tag{17}$$

Thus, conceptually, the stacked cyphertexts are equivalent to the stacked plaintexts after encryption by the block diagonal matrix $\mathbf{P}$, which is the orthogonal matrix assembled from the component random orthogonal matrices $\mathbf{P}_1$, $\mathbf{P}_2$, and $\mathbf{P}_3$. Thus, as shown in the previous section, unknowns inferred from joint cyphertext will be identical to those inferred using the joint plaintext. Note that genomic predictions of the original individuals require the plaintext genotypes and, thus, each contributor can only generate these for their own individuals, but using estimates of marker effects obtained from the combined data for additional accuracy of predictions.

Security of HEGP

The correlation between the centered plaintext genotypes, represented as $\mathbf{m}_j$, and the cyphertext genotypes, represented as $\mathbf{P}\mathbf{m}_j$, is proportional to $\mathbf{m}_j^T\mathbf{P}\mathbf{m}_j$. When $\mathbf{P}$ is "far from" an identity matrix (or scaled identity matrix), such correlations resemble those between two random vectors. For example, using the pig genotypes from Cleveland *et al.* (2012) ($n = 3534$, $p = 50,436$), the empirical distribution of Pearson correlations between the plaintext and cyphertext genotypes for each marker are shown in Figure 4. On average, the correlation between raw and encrypted genotypes is about 0.001, which is very close to 0, and almost all ($\sim 93\%$) correlations are inside the interval $[-0.03, 0.03]$. Thus, without decryption, cybertext genotypes and phenotypes are uninterpretable.

Decryption without knowledge of the key To obtain the raw genotypes from the encrypted genotypes, an orthogonal matrix $\mathbf{Q} \sim \mathbf{P}^T$ should be estimated as the key for decryption. When $\mathbf{Q} = \mathbf{P}^T$, exactly, the plaintext genotypes will be recovered, since then $\mathbf{Q}\mathbf{M}^* = \mathbf{P}^T\mathbf{P}\mathbf{M} = \mathbf{M}$. The distance between $\mathbf{Q}\mathbf{M}^*$ and $\mathbf{M}$ measures how close the decryption is to the raw data. However, because neither $\mathbf{M}$ nor $\mathbf{P}$ are shared, it is difficult to evaluate attempted decryption without a suitable objective function to

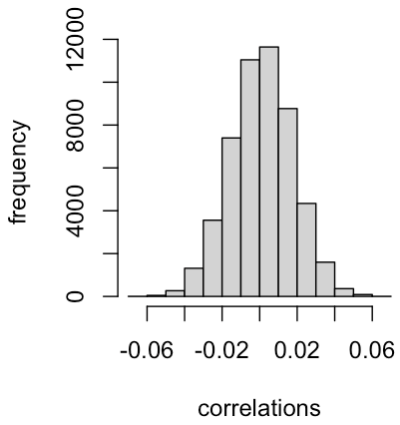

Figure 4 Distribution of correlations between 50,436 pairs of plaintext and cyphertext genotypes using pig dataset in Cleveland *et al.* (2012) ($n = 3534$, $p = 50,436$). The average correlation is less than 0.001.

minimize. Assuming the distance between the attempted decrypted genotype matrix and the plaintext genotype matrix is known (although it is unknown in practice), several strategies to decrypt the genotypes were discussed in Mott *et al.* (2020b). First, in a brute-force approach, numerous random orthogonal matrices (i.e., keys) were generated for decryption. However, massive computing resources would be required to generate and test all possible keys. Mott *et al.* (2020b) reported that even for a dataset with 8 individuals, they could not brute-force the key. A second approach to uncover the decryption key relies on the trimodal distribution of the plaintext genotype frequencies of each marker, assuming all markers are in Hardy-Weinberg equilibrium with publicly-available allele frequencies (such as Figure 2(c)). Mott *et al.* (2020b) attempted to infer the key by maximizing the kernel density estimator of those non-Gaussian distributions. However, the results were unsuccessful. Finally, a decryption challenge for HEGP (Mott *et al.* 2020a), in which attempts were invited to decrypt HEGP-encrypted plaintext genotypes, has so far failed to elicit a successful attack. More discussion can be found in Mott *et al.* (2020b).

The only identified weakness of HEGP occurs when the data includes variants that are private to an individual. In an extreme case, when each individual has a private variant coded as 1, the plaintext genotype matrix can be written as $\mathbf{M} = \left[\mathbf{I} \quad | \quad \mathbf{M}_{sub}\right]$, where $\mathbf{I}$ represents genotypes of the n private variants, and $\mathbf{M}_{sub}$ represents genotypes of all the other markers. In this situation, $\mathbf{P}$ itself will be included in the encrypted genotypes since $\mathbf{M}^* = \left[\mathbf{P} \quad | \quad \mathbf{M}^*_{sub}\right]$. In practice, this extreme case can be avoided by using common variants. However, it suggests that useful information might be extracted from the encrypted data of lower-frequency variants, suggesting it is best to remove any variant with a frequency under 0.01 or that is private to fewer than about 10 individuals. Since these variants are typically removed during quality control processing, there should be minimal loss of information.

Data analysis

The pig dataset in Cleveland *et al.* (2012) was used to validate the equivalent outcomes from both genomic prediction and GWAS analyses using plaintext and cyphertext. This dataset contains 3534 genotyped individuals and the number of SNP markers is 50,436. We simulated phenotypes based on different values for heritability and numbers of quantitative trait loci (QTL) (i.e., causal variants). In detail, phenotypes with heritability equal 0.1, 0.3, 0.5, and 0.7 were simulated, and 1, 10, 50, and 100% of SNPs were randomly sampled as QTL (16 scenarios). Contemporary group effects were included to simulate phenotypes on individuals from 4 groups. For each simulated scenario, 10 replicates were applied. The genotypes of each marker were centered to have zero mean. The incidence matrix of fixed effects, the genotypes, and the simulated phenotypes were encrypted using a random orthogonal matrix generated as described above. SNP-BLUP and BayesCπ were applied to analyze the plaintext and cyphertext using the JWAS package (Cheng *et al.* 2018, 2022). In all scenarios, 500,000 MCMC iterations were applied to ensure convergence.

We first show that the estimated marker effects ($\hat{\boldsymbol{\alpha}}$) remain unchanged with the cybertext. Using the plaintext of genotypes ($\mathbf{M}$), the estimated breeding values (EBV) are calculated as $\mathbf{M}\hat{\boldsymbol{\alpha}}$, confirming that the EBVs also remain unchanged. Below we only present the results from BayesCπ, and the conclusions drawn were consistent with those from RR-BLUP results.

Results

Estimated marker effects and breeding values Overall, the marker effects estimated from plaintext and cyphertext were very similar, with a Pearson correlation of 0.9929. The results from one replicate in the scenario with $h^2 = 0.3$, $QTL\% = 1\%$ are presented in Figure 5(a), and similar results were observed across all other scenarios. The results for all scenarios are detailed in Table 1, where each value represents the averaged correlation across 10 replicates.

Table 1 Pearson correlations between estimated marker effects from plaintext vs ciphertext in different simulation scenarios. Each value is the averaged correlation from 10 replicates.

		h^2			
		0.1	0.3	0.5	0.7
QTL%	1%	0.9937	0.9950	0.9957	0.9961
	10%	0.9928	0.9927	0.9926	0.9928
	50%	0.9898	0.9937	0.9916	0.9911
	100%	0.9906	0.9932	0.9925	0.9920

The estimated breeding values (EBV) for all individuals with genotypes $\mathbf{M}$ were calculated as $\mathbf{M}\hat{\boldsymbol{\alpha}}_{plaintext}$, using marker effects estimated from plaintext, and as $\mathbf{M}\hat{\boldsymbol{\alpha}}_{cyphertext}$, using marker effects estimated from cyphertext. Overall, the correlation between EBV calculated using the plaintext and those calculated using the cyphertext was about 0.9996. The results of one replicate from the scenario with $h^2 = 0.3$, $QTL\% = 1\%$ are shown in Figure 5(b), and similar results were observed for all the other scenarios. The results for all scenarios are listed in Table 2, where each value is the average correlation from 10 replicates.

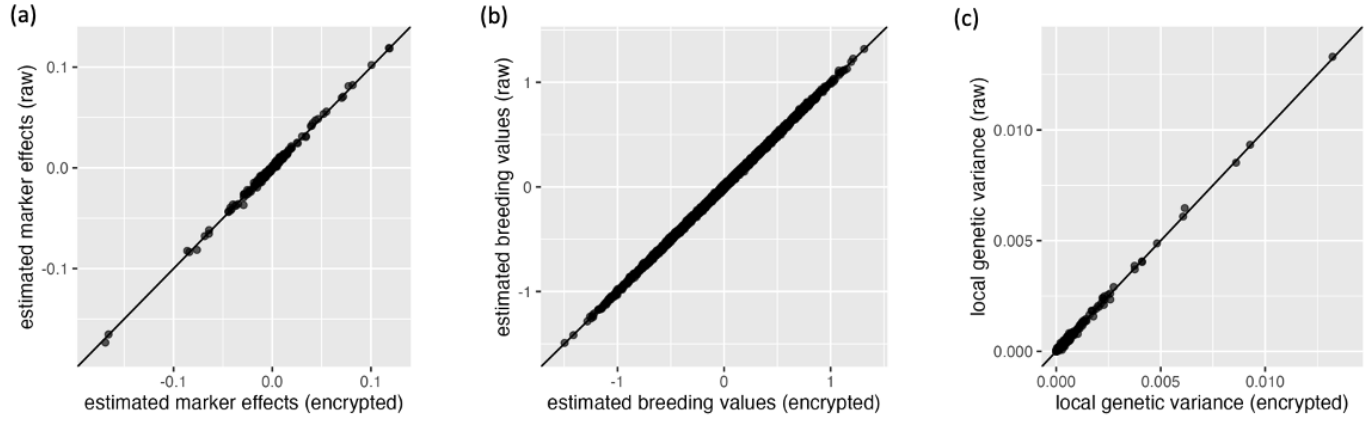

Figure 5 Results from one replicate in the simulation scenario with $h^2 = 0.3$ and $QTL\% = 1\%$. Each dot represents a pair of results calculated from plaintext (x-axis) vs cyphertext (y-axis). The diagonal line indicates when the plaintext and cyphertext result in the same estimates. (a) Comparison between estimated marker effects $\hat{\boldsymbol{\alpha}}_{plaintext}$ and $\hat{\boldsymbol{\alpha}}_{cyphertext}$ (correlation=0.9971). (b) Comparison between estimated breeding values $\mathbf{M}\hat{\boldsymbol{\alpha}}_{plaintext}$ and $\mathbf{M}\hat{\boldsymbol{\alpha}}_{cyphertext}$ (correlation=0.9997). The pig genotypes data was used ($\mathbf{M}$). (c) Comparison between local genetic variances of 2,522 non-overlapped genomic windows (correlation=0.9983).

Table 2 Person correlations between estimated breeding values (EBV) calculated using marker effects estimated from cyphertext ($\mathbf{M}\hat{\boldsymbol{\alpha}}_{cyphertext}$) and EBV calculated using the marker effects estimated from plaintext ($\mathbf{M}\hat{\boldsymbol{\alpha}}_{plaintext}$). Each value represents the average correlation across 10 replicates.

		h^2			
		0.1	0.3	0.5	0.7
QTL%	1%	0.9993	0.9997	0.9998	0.9998
	10%	0.9992	0.9996	0.9997	0.9997
	50%	0.9992	0.9996	0.9997	0.9997
	100%	0.9992	0.9995	0.9997	0.9997

Local genetic variances For GWAS, the genetic variance captured by a genomic window is of interest due to the fact that highly correlated SNPs within a genomic window jointly affect the phenotype, and it is difficult to identify the effect of a single marker (Hayes *et al.* 2010). In GWAS, local genetic variances can be used to estimate window-based posterior probabilities of association (WPPA) (Fernando *et al.* 2017). Here we divided the pig reference genome into 2522 non-overlapping genomic windows, where each window contains about 20 SNPs. The genetic values that are attributed to each genomic window were sampled from their posterior distributions using MCMC.

Overall, the correlation between local genetic variances estimated using plaintext or cyphertext was about 0.9923. The results of one replicate from the scenario with $h^2 = 0.3$, $QTL\% = 1\%$ are presented in Figure 5(c), and similar results were observed for all other scenarios. The results for each scenario are listed in Table 3, where each value represents the average correlation from 10 replicates.

Table 3 Person correlations between local genetic variances of 2,522 non-overlapping genomic windows calculated from plaintext or cyphertext. Each value represents the average correlation from 10 replicates.

		h^2			
		0.1	0.3	0.5	0.7
QTL%	1%	0.9954	0.9950	0.9966	0.9968
	10%	0.9895	0.9923	0.9929	0.9943
	50%	0.9908	0.9946	0.9932	0.9924
	100%	0.9728	0.9934	0.9936	0.9933

Joint analysis of cyphertext from multiple contributors

To perform joint cyphertext analysis, the 3,534 individuals in the pig dataset were split into two datasets ($n_1 = 500$, $n_2 = 3034$), modelling the scenario of two data contributors. The genotypes were independently centered within each contributor to have zero means. The simulated phenotypes data in the scenario with heritability of 0.3 and 10% QTLs were used. The plaintext phenotypes, genotypes, and covariates were independently encrypted by each contributor. For example, for contributor 1, the encrypted genotype data is $\mathbf{M}_1^* = \mathbf{P}_1\mathbf{M}_1$ with $\mathbf{P}_1$ of size $n_1 \times n_1$, and for contributor 2 the encrypted genotype data is $\mathbf{M}_2^* = \mathbf{P}_2\mathbf{M}_2$ with $\mathbf{P}_2$ of size $n_2 \times n_2$. Only the cyphertexts were shared, not the encryption keys $\mathbf{P}_1$ or $\mathbf{P}_2$. We repeated the previous analyses using the joint cyphertexts and the joint plaintexts. Using the joint cyphertexts yielded results very similar to those using the joint plaintexts.

Moreover, using joint cyphertexts to estimate parameters resulted in significantly higher prediction accuracies than only using the data from a single data contributor. This is to be expected, as a larger sample size improves parameter inference. For the 500 individuals in contributor 1, we calculated their EBV using marker effects estimated from the joint data ($\mathbf{M}_1\hat{\boldsymbol{\alpha}}_{joint}$), as well as using marker effects estimated from only contributor 1's data ($\mathbf{M}_1\hat{\boldsymbol{\alpha}}_{sub1}$). The comparison between the accuracy

of $\mathbf{M}_1\hat{\boldsymbol{\alpha}}_{joint}$ and $\mathbf{M}_1\hat{\boldsymbol{\alpha}}_{sub1}$ is shown in Figure 6, where each dot represents the result of one replicate. The joint data resulted in significantly higher accuracy than using data from a single contributor (pairwise t-test P-value < 0.0005). The same conclusions were drawn for contributor 2.

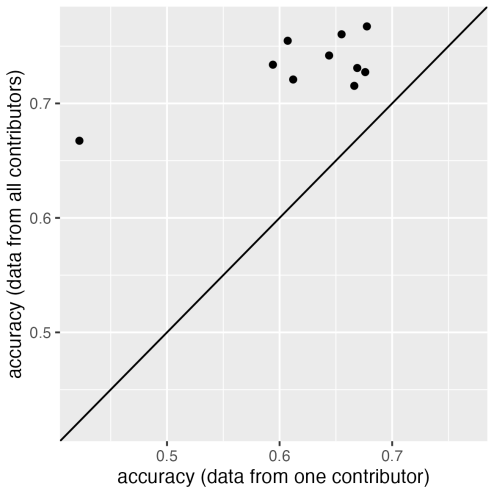

Figure 6 The prediction accuracies of estimated breeding values for 500 individuals in a single data contributor. Each dot represents a pair of results calculated either using only this contributor's data (x-axis) or using the joint data from all contributors (y-axis). Joint analyses had significantly higher accuracies than those using data from a single contributor (pairwise t-test P-value < 0.0005).

Discussion

Overview

In this study we have built on the HEGP methodology introduced in Mott *et al.* (2020b) to show how it can be extended to the wide class of mixed models including Bayesian Alphabet models that are commonly used in animal and crop quantitative genetics. We have also shown how joint analysis of multiple data sets fits into this framework and confirmed the increase in prediction accuracy of breeding values expected from joint analyses with plaintext analyses also holds for cyphertext. Note that the estimated marker effects using encrypted data are the same as those using the raw data.

HEGP enables adherence to, and capitalizes on, the benefits of the FAIR principles in genome-to-phenome studies. In the context of animal and crop breeding, it also addresses many of the privacy, intellectual property, and commercial interest issues that prevent the sharing and reuse of data for both research and industry applications.

Alternatives

The principal alternative strategy to sharing genotype and phenotype data is to share GWAS summary statistics, for example, marker regression coefficient estimates and their standard errors. This approach is most suited to single-marker regression analyses that are typical in human studies (MacArthur *et al.* 2021). To this end, databases have been built to collect GWAS summary statistics (Welter *et al.* 2014; MacArthur *et al.* 2017; Buniello *et al.* 2019), and methods have been proposed to facilitate large meta-analyses needed for the increased power in dissecting the genetic basis of complex traits (Yang *et al.* 2012; Vilhjálmsson *et al.* 2015; Barbeira *et al.* 2018; Lloyd-Jones *et al.* 2019; Privé *et al.* 2020; Yengo *et al.* 2022b; Werme *et al.* 2022). However, meta-analyses based on these summary statistics rely heavily on approximations due to unavailability of the individual-level data. Homomorphic encryption methods such as HEGP do not require such approximations, provided they factorize into a prior distribution of the markers multiplied by the Gaussian likelihood of the data, and can be fitted by MCMC methods such as Gibbs sampling.

Rounding Errors

In HEGP, the individuals' plaintext identities, phenotypes, and genotypes are obscured by pre-multiplying by a high-dimensional random orthogonal matrix. In the resulting cyphertext, the relationships between SNPs, and between SNPs and phenotypes are preserved, but the relationships between individuals are scrambled - in fact, the concept of an individual is nonsensical after encryption, as records in the encrypted data are random linear combinations of the original individuals' records. Theoretically, plaintext and cyphertext should yield identical estimates of marker effects and other parameters, but due to rounding errors, as well as Monte Carlo errors in the case of models using MCMC, the estimated marker effects are not identical but rather very similar, with correlations close to 1.0. In detail, rounding errors occur because the off-diagonal elements of $\mathbf{P}^T\mathbf{P}$ are very small values, $\sim 10^{-13}$. Mott *et al.* (2020b) reported that rounding errors were negligible for $\mathbf{P}$ with dimensions up to $10,000 \times 10,000$. To alleviate the problem of rounding errors for a very large dataset, $\mathbf{P}$ can be constructed as a block-diagonal matrix, where each block is a random orthogonal matrix.

Time complexity

The time taken to generate a random $n \times n$ orthogonal matrix $\mathbf{P}$ from the Stiefel manifold is proportional to n^3, where n is the number of individuals, being dominated by the eigen decomposition. The time taken to multiply the plaintext by $\mathbf{P}$ to produce the cyphertext is proportional to pn^2, where p is the number of markers. In a computer server with five cores, generating $\mathbf{P}$ for the pig dataset ($n = 3,534$) took less than one minute. For a dataset with 10,000 individuals, the time to generate $\mathbf{P}$ was about 5 minutes. However, the time to generate $\mathbf{P}$ for 50,000 individuals was about 8 hours. As shown in Figure 7, running time increased rapidly as sample size increased.

In practice, with hundreds of thousands of individuals, many relatively small random orthogonal matrices (e.g., $50,000 \times 50,000$) could be generated in parallel, and then a large block-diagonal orthogonal matrix could be constructed, with each block being a random orthogonal matrix (i.e, a block-diagonal random orthogonal matrix). This larger block-diagonal orthogonal matrix, as well as any permutation of such a matrix, can be used as the encryption key.

The size of the cyphertext is the same as the plaintext and, therefore, the computational effort required for each iteration in MCMC is comparable. In our analysis of the pig dataset (Cleveland *et al.* 2012), the number of MCMC iterations necessary to ensure the convergence of the MCMC process was also similar between analysis of the plaintext and the cyphertext.

Security

With an appropriately sampled HEGP encryption key, the correlation between raw and encrypted data resembles that between two random vectors. For the pig dataset (Cleveland *et al.* 2012),

the absolute Pearson correlation between raw and encrypted marker genotypes was almost always less than 0.03. To increase the security of the encrypting and lower the risk of discovery of the decryption key, genotypes for SNPs with very low minor allele frequencies should not be shared. Since only cyphertext are shared, the unknown nature of the plaintext genotypes **M** makes it difficult to evaluate decryption attempts. To date, decryption attacks have been proven ineffective Mott *et al.* (2020b), even when **M** was available for evaluation. However, further exploration is still needed to determine whether HEGP is cryptographically secure.

Protocols for data sharing in HEGP

Finally, we mention some points to consider when sharing HEGP cyphertext. First, it is necessary for all contributors to agree on a common set of markers and covariates to be shared, and on whether phenotypes are to be residualised by removing covariate effects before sharing - in which case the cyphertext versions of covariates need not be shared - or afterwards, during the joint analysis. Second, missing genotypes, covariates and phenotypes must be imputed, either for markers not genotyped in a particular contributor's data, or to fill in sporadic missing values (HEGP does not allow missing data). Third, the sharing topology must be agreed upon: each cyphertext could be shared with all contributors, so that each participant could conduct their own analysis, or instead, it could be shared only with a trusted third party who would perform the agreed-upon analysis. Fourth, although the joint analysis' parameter estimates etc do not need decrypting, the parties may want to agree beforehand on their subsequent use and dissemination.

These considerations would likely require the contributors to set up a protocol for data sharing, reflecting the sensitivity and value of the component data sets, and which will likely vary depending on circumstances and commercial considerations. Notwithstanding the HEGP-specific technical requirements, such a protocol should be simpler to implement than agreements involving the sharing of plaintext data.

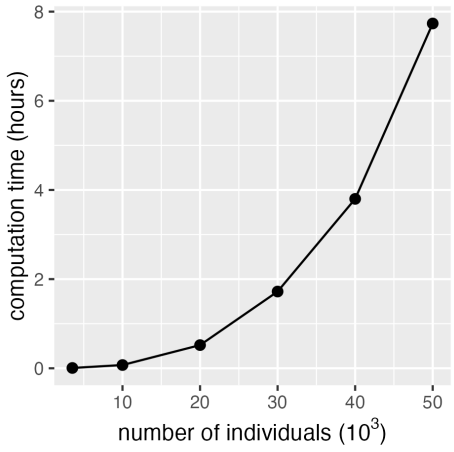

Figure 7 Time to generate a random orthogonal matrix from the Stiefel manifold. The x-axis is the size of **P** matrix (i.e., number of individuals), and the y-axis is the computation time.

Data availability

Pig genotypes used in the analysis are publicly available in Cleveland *et al.* (2012). The simulated phenotypes and all scripts are available at https://github.com/zhaotianjing/encryption. The authors state that all data necessary for confirming the conclusions presented in the article are represented fully within the article.

Funding

This work was supported by a UKRI BBSRC grant BB/V00767X/1 awarded to Richard Mott, by Agricultural Genome to Phenome Initiative (AG2PI) under USDA-NIFA awards 2020-70412-32615 and 2021-70412-35233, and by USDA-NIFA awards 2021-67015-33412, 2023-67015-39564, 2023-70412-41054.

Conflicts of interest

The authors declare that they have no competing interests.

Literature cited

Barbeira AN, Dickinson SP, Bonazzola R, Zheng J, Wheeler HE, Torres JM, Torstenson ES, Shah KP, Garcia T, Edwards TL *et al*. 2018. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from gwas summary statistics. Nature communications. 9:1–20.

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

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. 2007. Tassel: software for association mapping of complex traits in diverse samples. Bioinformatics. 23:2633–2635.

Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E *et al*. 2019. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic acids research. 47:D1005–D1012.

Chen H, Gilad-Bachrach R, Han K, Huang Z, Jalali A, Laine K, Lauter K. 2018. Logistic regression over encrypted data from fully homomorphic encryption. BMC medical genomics. 11:3–12.

Cheng H, Fernando R, Garrick D. 2018. Jwas: Julia implementation of whole-genome analysis software. In: . volume 11. p. 859.

Cheng H, Fernando R, Garrick D, Zhao T, Qu J. 2022. Jwas version 2: leveraging biological information and highthroughput phenotypes into genomic prediction and association. In: . volume 12. pp. 1519–1522.

Chikuse Y, Chikuse Y. 2003. *Statistics on special manifolds*. volume 1. Springer.

Cleveland MA, Hickey JM, Forni S. 2012. A common dataset for genomic analysis of livestock populations. G3: Genes| Genomes| Genetics. 2:429–435.

Erbe M, Hayes BJ, Matukumalli LK, Goswami S, Bowman PJ, Reich CM, Mason BA, Goddard ME. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J Dairy Sci. 95:4114–29.

Fernando R. 1998. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. In: . volume 26. pp. 329–336. Armidale, Australia.

Fernando R, Toosi A, Wolc A, Garrick D, Dekkers J. 2017. Application of whole-genome prediction methods for genome-wide association studies: a bayesian approach. J Agric Biol Environ Stat. 22(2):172–93.
Fernando RL, Garrick D. 2013. Bayesian methods applied to gwas. Genome-wide association studies and genomic prediction. pp. 237–274.
Gianola D, Fernando RL. 2019. A Multiple-Trait Bayesian Lasso for Genome-Enabled Analysis and Prediction of Complex Traits. Genetics. 214:genetics.302934.2019 – 331.
Habier D, Fernando RL, Dekkers JC. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics. 177:2389–2397.
Habier D, Fernando RL, Kizilkaya K, Garrick D. 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics. 12:186.
Hayes BJ, Pryce J, Chamberlain AJ, Bowman PJ, Goddard ME. 2010. Genetic architecture of complex traits and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in holstein cattle as contrasting model traits. PLoS Genet. 6(9):e1001139.
Hayes BJ, Visscher PM, Goddard ME. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. Genetics research. 91:47–60.
Hoff P. 2012. rstiefel: Random orthonormal matrix generation on the stiefel manifold. R package version 0.9, URL http://CRAN. R-project. org/package= rstiefel. .
Hoff PD. 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.
Kizilkaya K, Fernando RL, Garrick DJ. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. J Anim Sci. 88:544–51.
Legarra A, Ricard A, Varona L. 2018. Gwas by gblup: single and multimarker emmax and bayes factors, with an example in detection of a major gene for horse gait. G3: Genes, Genomes, Genetics. 8:2301–2308.
Lloyd-Jones LR, Zeng J, Sidorenko J, Yengo L, Moser G, Kemper KE, Wang H, Zheng Z, Magi R, Esko T *et al*. 2019. Improved polygenic prediction by bayesian multiple regression on summary statistics. Nature communications. 10:1–11.
Lu WJ, Yamada Y, Sakuma J. 2015. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. In: . volume 15. pp. 1–8. Springer.
MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, Junkins H, McMahon A, Milano A, Morales J *et al*. 2017. The new nhgri-ebi catalog of published genome-wide association studies (gwas catalog). Nucleic acids research. 45:D896–D901.
MacArthur JA, Buniello A, Harris LW, Hayhurst J, McMahon A, Sollis E, Cerezo M, Hall P, Lewis E, Whetzel PL *et al*. 2021. Workshop proceedings: Gwas summary statistics standards and sharing. Cell Genomics. 1:100004.
Meuwissen THE, Hayes BJ, Goddard ME. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 157:1819–1829.
Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. 2015a. Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. PLoS Genet. 11:e1004969.
Moser G, Lee SH, Hayes BJ, Goddard ME, Wray NR, Visscher PM. 2015b. Simultaneous discovery, estimation and prediction analysis of complex traits using a Bayesian mixture model. PLOS Genet. 11:e1004969.
Mott R, Fischer C, Prins P, Davies RW. 2020a. Hegp challenge. Available from: https://hegp.genenetwork.org/challenge.
Mott R, Fischer C, Prins P, Davies RW. 2020b. Private Genomes and Public SNPs: Homomorphic Encryption of Genotypes and Phenotypes for Shared Quantitative Genetics. Genetics. 215:359–372.
Park T, Casella G. 2008. The bayesian lasso. Journal of the American Statistical Association. 103:681–686.
Privé F, Arbel J, Vilhjálmsson BJ. 2020. Ldpred2: better, faster, stronger. Bioinformatics. 36:5424–5431.
Sim JJ, Chan FM, Chen S, Meng Tan BH, Mi Aung KM. 2020. Achieving gwas with homomorphic encryption. BMC medical genomics. 13:1–12.
Strandén I, Garrick DJ. 2009. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 92:2971–2975.
VanRaden PM. 2008. Efficient methods to compute genomic predictions. Journal of dairy science. 91:4414–4423.
Vilhjálmsson BJ, Yang J, Finucane HK, Gusev A, Lindström S, Ripke S, Genovese G, Loh PR, Bhatia G, Do R *et al*. 2015. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. The american journal of human genetics. 97:576–592.
Wang H, Misztal I, Aguilar I, Legarra A, Muir W. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genetics Research. 94:73–83.
Wang T, Chen YPP, Bowman PJ, Goddard ME, Hayes BJ. 2016. A hybrid expectation maximisation and mcmc sampling algorithm to implement bayesian mixture model based genomic prediction and qtl mapping. BMC genomics. 17:1–21.
Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L *et al*. 2014. The nhgri gwas catalog, a curated resource of snp-trait associations. Nucleic acids research. 42:D1001–D1006.
Werme J, van der Sluis S, Posthuma D, de Leeuw CA. 2022. An integrated framework for local genetic correlation analysis. Nature genetics. 54:274–282.
Yang J, Ferreira T, Morris AP, Medland SE, Madden PA, Heath AC, Martin NG, Montgomery GW, Weedon MN, Loos RJ *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, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, Graff M *et al*. 2022a. A saturated map of common genetic variants associated with human height. Nature. 610:704–712.
Yengo L, Vedantam S, Marouli E, Sidorenko J, Bartell E, Sakaue S, Graff M, Eliasen AU, Jiang Y, Raghavan S *et al*. 2022b. A saturated map of common genetic variants associated with human height from 5.4 million individuals of diverse ancestries. bioRxiv. .

Appendix

Homomorphic encryption for single markers analysis in GWAS

Mott *et al.* (2020b) proposed the homomorphic encryption for genotypes and phenotypes (HEGP) method for single-marker regression in genome-wide association studies (GWAS) using linear mixed models with Gaussian errors, where the raw geno-

types and phenotypes data were pre-multiplied by a high-dimensional random orthogonal matrix. Mott *et al.* (2020b) showed that such encryption does not change the likelihood of the quantitative trait in the GWAS model, a point we will illustrate further below. For more details, refer to Mott *et al.* (2020b).

In detail, for n individuals genotyped with p markers, the raw genotypes and phenotypes matrix were pre-multiplied by the same random orthogonal matrix as

$$\begin{aligned} \mathbf{M}^* &= \mathbf{PM} \\ \mathbf{y}^* &= \mathbf{Py} \end{aligned} \tag{18}$$

where $\mathbf{M}$ is an $n \times p$ genotype matrix, $\mathbf{y}$ is the vector of phenotypes of length n, $\mathbf{P}$ is an $n \times n$ random orthogonal matrix whose columns and rows are orthonormal vectors (i.e, $\mathbf{P}^T = \mathbf{P}^{-1}$), and $\mathbf{M}^*$ and $\mathbf{y}^*$ are encrypted genotypes and phenotypes, respectively. The covariate matrix $\mathbf{X}$ should be encrypted as well, producing $\mathbf{X}^* = \mathbf{PX}$.

In GWAS, the linear model used to test the significance of jth SNP ($j = 1, ..., p$) is

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{m}_j\alpha_j + \mathbf{u} + \mathbf{e} \\ &= \mathbf{X}\boldsymbol{\beta} + \mathbf{m}_j\alpha_j + \boldsymbol{\epsilon} \end{aligned} \tag{19}$$

where $\mathbf{y}$ is the phenotype, $\mathbf{X}$ is the covariate matrix, $\boldsymbol{\beta}$ is the fixed effects of covariates, $\mathbf{m}_j$ is the (centered and scaled) genotypes of jth SNP, α_j is the regression coefficient of jth SNP. $\mathbf{u}$ is a random vector for polygenic effects with $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G}\sigi_u^2)$, and $\mathbf{e}$ is a random vector of residuals with $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$. Thus, the variance of $\mathbf{y}$ is $var(\mathbf{y}) = \mathbf{G}\sigma_u^2 + \mathbf{I}\sigma_e^2 = \mathbf{V}$.

Applying orthogonal encryption, the above GWAS model becomes

$$\begin{aligned} \mathbf{Py} &= \mathbf{PX}\boldsymbol{\beta} + \mathbf{Pm}_j\alpha_j + \mathbf{P}\boldsymbol{\epsilon} \\ \mathbf{y}^* &= \mathbf{X}^*\boldsymbol{\beta} + \mathbf{m}_j^*\alpha_j + \boldsymbol{\epsilon}^* \end{aligned} \tag{20}$$

where the variance of encrypted phenotypes becomes $var(\mathbf{Py}) = \mathbf{PVP}^T$. We showed in Equation 8 that the likelihood is invariant under orthogonal transformation Thus, HEGP leaves likelihood-based inferences for GWAS model - used to test the significance of a single marker - unaffected. This includes the maximum likelihood parameter estimates and P-values for likelihood-based tests of significance.

Gibbs sampler for the linear mixed model

The full conditional posterior distributions of parameters of interest in SNP-BLUP and BayesCπ are shown below. More details can be found in Fernando and Garrick (2013).

Residual variance The full conditional posterior distribution of residual variance σ_e^2 follows a scaled inverse chi-square distribution with $n + \nu_e$ degrees of freedom and scale parameter $\frac{\mathbf{e}^T\mathbf{e} + \nu_e S_e^2}{n+\nu_e}$. That is,

$$f(\sigma_e^2|ELSE) \propto (\sigma_e^2)^{-\frac{n+\nu_e+2}{2}} exp[-\frac{1}{2\sigma_e^2}(\mathbf{e}^T\mathbf{e} + \nu_e S_e^2)], \tag{21}$$

where $\mathbf{e}$ is the residuals. Since $(\mathbf{e}^*)^T(\mathbf{e}^*) = \mathbf{e}^T\mathbf{P}^T\mathbf{Pe} = \mathbf{e}^T\mathbf{e}$, the full conditional posterior distribution of σ_e^2 is unchanged with the encrypted data.

Marker effect variance The full conditional posterior distribution of σ_α^2 follows a scaled inverse chi-square distribution with $k + \nu_\alpha$ degrees of freedom and scale parameter $\frac{\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \nu_\alpha S_\alpha^2}{k+\nu_\alpha}$, where $k = \sum \delta_j$ is the number of markers included in the model. In detail,

$$f(\sigma_\alpha^2|ELSE) \propto (\sigma_\alpha^2)^{-\frac{k+\nu_\alpha+2}{2}} exp[-\frac{1}{2\sigma_\alpha^2}(\boldsymbol{\alpha}^T\boldsymbol{\alpha} + \nu_\alpha S_\alpha^2)] \tag{22}$$

We have proven that $\boldsymbol{\alpha}$ is unchanged with the encrypted data, thus, the full conditional posterior distribution of σ_α^2 is also unchanged.

Fixed effects The full conditional posterior distribution of jth fixed effects β_j follows a univariate normal distribution with mean $\frac{\mathbf{x}_j^T(\mathbf{y} - \mathbf{M}\boldsymbol{\alpha} - \sum_{j' \neq j}\mathbf{x}_{j'}\beta_{j'})}{\mathbf{x}_j^T\mathbf{x}_j}$ and variance $\frac{\sigma_e^2}{\mathbf{x}_j^T\mathbf{x}_j}$.

Given $\mathbf{X}^* = \mathbf{PX}$, we have

$$\begin{aligned} [\mathbf{x}_1^*, ..., \mathbf{x}_p^*] &= \mathbf{P}[\mathbf{x}_1, ..., \mathbf{x}_p] \\ &= [\mathbf{Px}_1, ..., \mathbf{Px}_p] \end{aligned} \tag{23}$$

Thus, the $\mathbf{x}_j^T\mathbf{x}_j$ is unchanged using encrypted data since

$$\begin{aligned} (\mathbf{x}_j^*)^T(\mathbf{x}_j^*) &= (\mathbf{Px}_j)^T(\mathbf{Px}_j) \\ &= \mathbf{x}_j^T\mathbf{x}_j \end{aligned} \tag{24}$$

The $\mathbf{x}_j^T\mathbf{y}$ is also unchanged using encrypted data since

$$\begin{aligned} (\mathbf{x}_j^*)^T(\mathbf{y}^*) &= (\mathbf{Px}_j)^T\mathbf{Py} \\ &= \mathbf{x}_j^T\mathbf{y} \end{aligned} \tag{25}$$

Similarly, $\mathbf{x}_j^T\mathbf{M}$ and $\mathbf{x}_j^T\mathbf{x}_{j'}$ are also unchanged using encrypted data. Thus, the full conditional posterior distribution of β_j is unchanged.

Indicator variables In BayesCπ, an indicator Bernoulli variable δ_j is introduced for locus j that is 1 with probability $1 - \pi$ and 0 with probability π. The full conditional posterior distribution of indicator variable δ_j is:

$$\begin{aligned} &f\left(\delta_j = 1 \mid ELSE\right) \\ &= \frac{f_1\left(r_j \mid \sigma_\alpha^2, \sigma_e^2\right) f\left(\delta_j = 1\right)}{f_0\left(r_j \mid \sigma_e^2\right) f\left(\delta_j = 0\right) + f_1\left(r_j \mid \sigma_\alpha^2, \sigma_e^2\right) f\left(\delta_j = 1\right)}, \end{aligned} \tag{26}$$

where $f_1\left(r_j \mid \sigma_\alpha^2, \sigma_e^2\right)$ is a univariate normal distribution with

$$E\left(r_j \mid \sigma_\alpha^2, \sigma_e^2\right) = 0, Var\left(r_j \mid \sigma_\alpha^2, \sigma_e^2\right) = \left(\mathbf{m}_j^T\mathbf{m}_j\right)^2\sigma_\alpha^2 + \mathbf{m}_j^T\mathbf{m}_j\sigma_e^2,$$

and $f_0\left(r_j \mid \sigma_e^2\right)$ is a univariate normal distribution with

$$E\left(r_j \mid \sigma_e^2\right) = 0, Var\left(r_j \mid \sigma_e^2\right) = \mathbf{m}_j^T\mathbf{m}_j\sigma_e^2,$$

and

$$r_j = \mathbf{m}_j^T\left(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \sum_{j' \neq j}\mathbf{m}_{j'}\alpha_{j'}\delta_{j'}\right).$$

We have showed that r_j and $\mathbf{m}_j^T\mathbf{m}_j$ are unchanged with encrypted data. Thus, the full conditional posterior distribution of δ_j is unchanged.

Inclusion probabilities In BayesCπ, the full conditional posterior distribution of inclusion probability π follows a Beta distribution with shape parameter $p-k+1$ and $k+1$. That is,

$$f(\pi|ELSE) \propto \pi^{(p-k)}(1-\pi)^{k} \tag{27}$$

Using encrypted data will not affect the full conditional posterior distribution of π.