Isaac, A., Cheng, H., Li, D., Dekkers, J., Mott, R. (2026) Genotype and phenotype encryption optimised for federated quantitative genetics, *Proceedings 13th World Congress on Genetics Applied to Livestock Production* 23767: https://doi.org/10.31274/wcgalp.23767
©2026 Isaac, et al. CC BY-NC-ND

Genotype and phenotype encryption optimised for federated quantitative genetics

A. Isaac[1], H. Cheng[2], D. Li[2], J.C. Dekkers[3], R. Mott[1*]

[1]University College London, UK; [2]University of California, Davis CA USA; [3]Iowa State University, Ames IA, USA; [*]Corresponding author: r.mott@ucl.ac.uk

Introduction

Animal and plant breeders have long recognized the importance of very large datasets for accurate genetic evaluation and selection, because many loci with individually small effects influence most economically important traits. Continual progress, therefore, depends on increasing sample size. Sharing genetic and phenotypic data across consortia or partners is the most economical way to accomplish this goal. However, there can be technical and commercial impediments to data sharing. Technical requirements include genotyping animals at a common set of markers to an agreed level of accuracy and measuring phenotypes and covariates in a consistent manner. Commercial confidentiality means that it may be illegal or unwise to share raw data because of the risk that confidential intellectual property might be compromised. There is also a need to specify and agree in advance which analyses are to be performed on shared data and what use is to be made of any findings.

The risks associated with data sharing can be reduced by using homomorphic encryption. This means encrypting raw plaintext genotypes and phenotypes in such a way that it is still possible to federate encrypted data from different partners into a single data set that can be analysed to obtain essentially the same results as if all the raw data had been shared and analysed. In particular, it is essential to be able to perform standard quantitative genetic analyses including genome-wide association studies (GWAS) with linear mixed models and with Bayesian Alphabet or Genomic Best Linear Unbiased Prediction (GBLUP) models to estimate marker effects from the federated data. Partners can use these shared estimates for genomic prediction and to map quantitative trait loci (QTL).

In Mott *et al.* (2020) and Cheng *et al.* (2024), we introduced HEGP (Homomorphic Encryption of Genotypes and Phenotypes). The key idea is that, in each of the above analyses, the posterior distribution of the marker effects is modelled as the classical multivariate normal likelihood of the data, multiplied by the prior distribution on the vector of marker effects. Importantly, the multivariate normal likelihood is invariant under an orthogonal transformation simultaneously applied to both phenotypes, covariates, and genotype dosages. Furthermore, if this orthogonal transformation is randomly sampled from the space of all orthogonal transformations, the transformed data have many desirable properties—the transformed phenotypes, covariates, and genotype dosages closely resemble normal random deviates. Transformed data resemble random numbers but produce the same outputs as the plaintext. This observation is the basis of the HEGP scheme of homomorphic encryption.

Suppose a consortium is formed between multiple breeding organisations (commercial or academic). Each organisation prepares its genotypes and phenotypes according to a pre-agreed specification. It then transforms its data using a privately generated key (a random orthogonal matrix) that is not shared. Instead, each partner shares their encrypted data with a trusted broker or with the other partners. Mathematically, the federation of encrypted data entails stacking the encrypted matrices and vectors, which are then input into the relevant Bayesian Alphabet or GBLUP model. It is mathematically equivalent to stacking the original plaintext matrices to federate them, followed by same genetic analysis, or of course by stacking the plaintexts, encrypting the federated data by multiplying by a virtual federated block-diagonal orthogonal

matrix, and performing the analysis. Crucially, the same marker effects are estimated in all cases, whether the plaintext or cyphertext is analysed.
In this paper, we describe a practical and rigorously tested implementation of HEGP suitable for animal and plant breeders and ready for use. We first describe a free open-source Python package, **pyhegp**, which implements the necessary computational steps, paying particular attention to certain technical details needed to ensure the results are the same for plaintext and cyphertext, and second, the workflow necessary to implement a federated analysis, comprising the steps necessary to agree, set up, and implement the federated analysis. Even though the source code is shared, the encryption keys are privately generated by each partner and cannot be deduced from the code, depending on the internal state of each partner's random number generator. The keys can be deleted immediately after encryption as they are not needed for downstream analysis.

Materials & Methods

The **pyhegp** package is available from https://github.com/encryption4genetics/pyhegp under version 3 or later of the GNU General Public License. It comprises a suite of five programs (listed in Figure 1) that perform the necessary computations to encrypt and share data. QTL mapping used the JWAS package (Cheng *et al*., 2022) https://github.com/reworkhow/JWAS.jl.
Figure 1 shows a workflow in which each partner shares their encrypted data with a broker (hub-and-spoke architecture). It is also possible to avoid the use of a data broker, in which case each partner shares their encrypted data with each other (clique architecture). Box 1 lists the steps involved in a federated analysis. To ensure identical results from plaintext and cyphertext, genotype data must be centred genotype dosages. All data transfers use encrypted data, except for the initial exchange of allele frequencies and sample numbers prior to sharing encrypted data in order that each party can standardise genotype dosages using the overall minor allele frequency (this cannot be deduced from the encrypted genotypes). A Federated Analysis requires a combination of Negotiation, Agreement, Preparation and Analysis, between the central Data Broker and each Party contributing data:

Negotiation. Parties agree to share genetic and phenotypic data to perform a federated quantitative genetics analysis. The agreement specifies: (i) Phenotypes and covariates, including pre-processing (standardisation, normalisation, imputation of missing data). For factor variables the levels must be specified and whether they are random or fixed effects. Deliverable is a complete specification of a numerical tabular text file with each column ordered and detailed. (ii) The choice of reference genome and marker positions and how genotypes are encoded (dosages for additive genetic effects or augmented with additional columns for dominance effects). (iii) Quality control checks on each component dataset pre- and post- encryption, e.g., the estimation of heritability for each phenotype, which should be the same (up to numerical rounding error) in plaintext and cyphertext.

Agreement. The partners then agree on: (i) Data-sharing method (e.g., by FTP of encrypted, compressed tarball). (ii) Specification of the analyses to be performed on the federated data, including the results to be shared with all the participants and their file formats. (iii) Deletion policy of shared ciphertext data. (iv) Publication policy of results.

Preparation. Each party prepares their data using the agreed specification. They run quality control checks (summary statistics) on their plaintext. They encrypt their data using **pyhegp** and repeat the QC checks to ensure they produce the same results.

Data sharing. Encrypted data are packaged (e.g. files with specified names in a compressed tarball). The shared data may be further encrypted for transmission to the central data broker, using public key encryption. Each party sends their quality control statistics to the data broker.

Analysis. (i) Data broker unpacks each shared dataset, reruns quality control checks on each and checks that summary statistics agree. (ii) Data broker combines the shared cypher texts and performs the agreed analyses on the combined cyphertext data. (iii) Results are encrypted using public key encryption and returned to each party (each party receives the same result data, e.g. by downloading from a single repository). (iv) Encrypted data are deleted from the data broker's file system. (v) Distribution and publication of results as agreed.

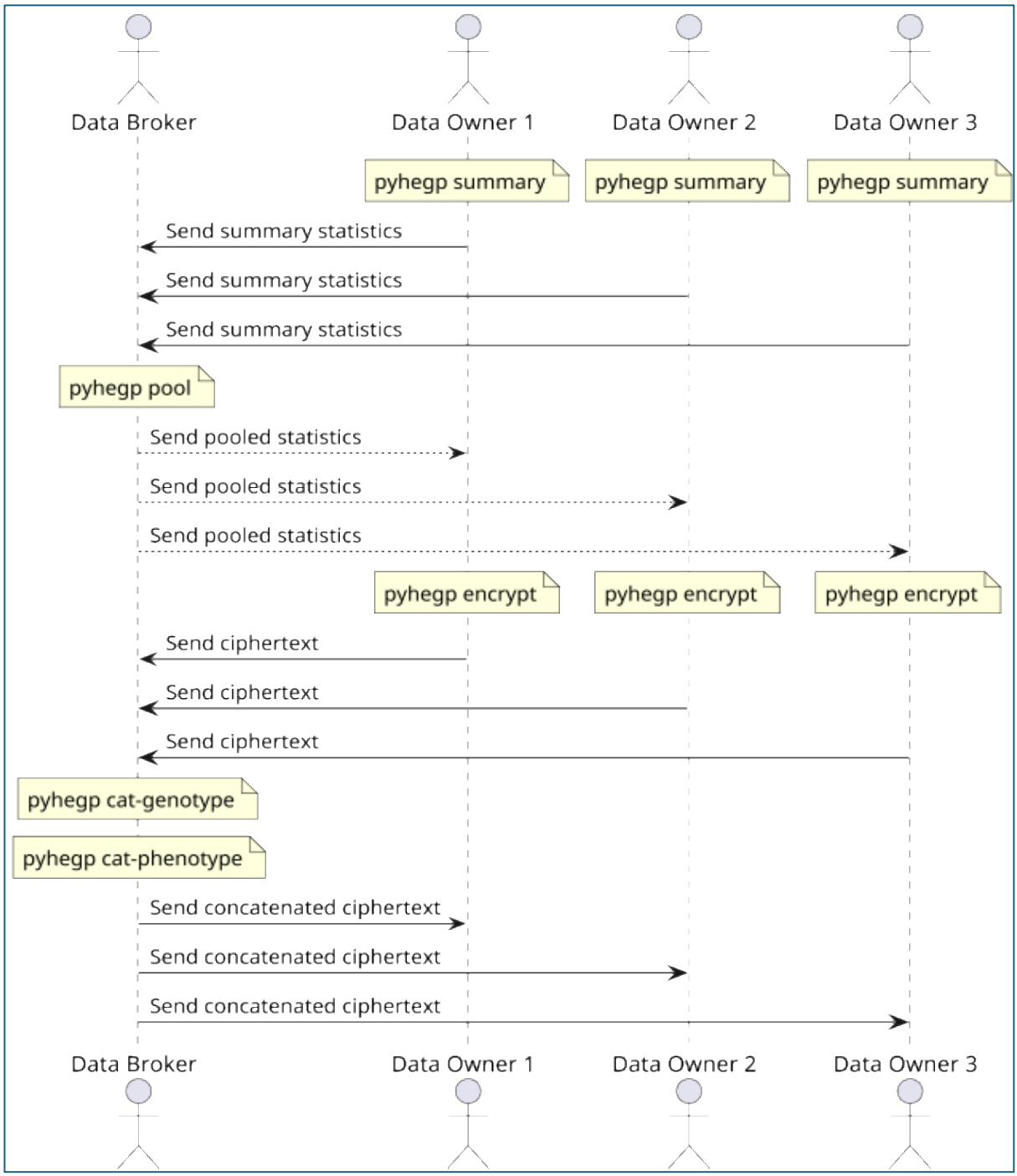

Figure 1. Schematic of a federated HEGP analysis involving three partners (data owners) and one broker. Yellow boxes show the names of the pyhegp functions involved in each step.

Results

We demonstrate **pyhegp** by analysing a federated pig dataset (Cheng *et al*., 2022) using a Markov Chain Monte Carlo step, which is a rigorous test of cyphertext versus plaintext analysis, since the posterior distribution is sampled using a random number generator. The federated analysis used 3,176 pigs from 7 partner companies and 433,422 SNPs. We removed SNPs that were fixed across all individuals. We standardized the genotypes and included the intercept as a fixed effect for both the plaintext and cyphertext analyses. Both analyses used the same random seed. The correlations were 0.99175 between plaintext and cyphertext marker effects were, 0.9717 between model frequencies, 0.9992 between estimated breeding values (EBV), and 0.9984 between GWAS prGenVar (proportion of genetic variance attributed to each 1 Mb window). Figure 2 shows scatter plots corresponding to these correlations. In addition, we have verified, in a federated **pyhegp** of mouse data, that the log p-values of SNP effects estimated in a linear mixed model are consistent between plaintext and cyphertext (R=0.9998).

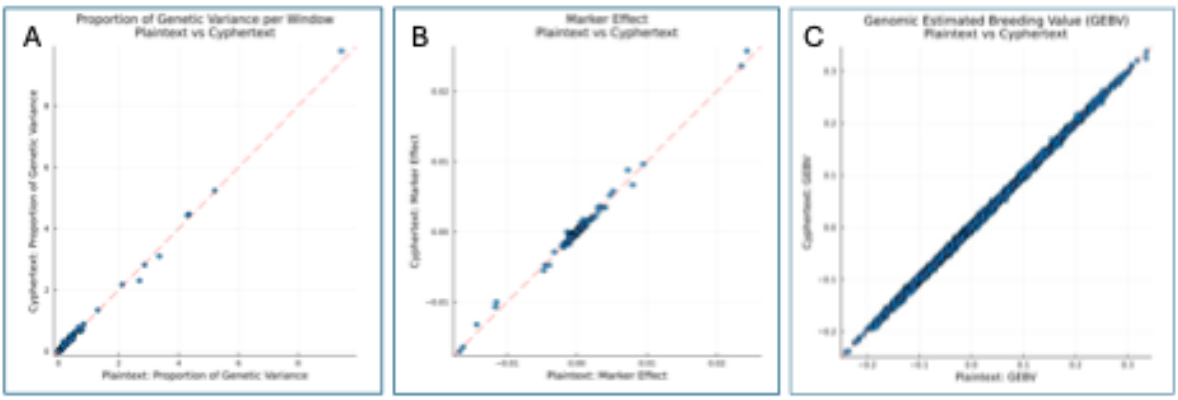

Figure 2. Scatter plots for estimates of (A) GWAS prGenVar, (B) marker effects and (C) estimated breeding values (C) for analyses of federated pig data using plaintext (x-axis) and cyphertext (y-axis). The dashed lines represent perfect equality.

Discussion

The example federated analysis in Figure 2 is numerically very close to the plaintext analysis. It identifies the same QTLs (here defined as markers with prGenVar>1), thereby confirming the utility of HEGP. Additionally, 13,617 of the SNPs analysed were fixed in at least one of the seven component data sets but nevertheless segregated in the federated data, permitting the estimation of effects also for these markers, which would be impossible in an analysis of a single partner's data, and thereby potentially revealing novel effects.

We have emphasised here both the statistical and procedural aspects of successfully conducting a federated genetic analysis. The **pyhegp** package has been designed to streamline both aspects. It is agnostic to the software to be used for genetic analysis, as long as the software accepts standard tabular files as inputs. In conclusion, HEGP provides a route for commercial and non-commercial breeders to share their data for mutual benefit. The **pyhegp** package is freely available, and we would be pleased to work with any interested groups to help them use it.

References

Cheng, H., R. Fernando, D. Garrick, T. *et al.* (2022). In Proc. 12th WCGALP (pp. 1519–1522), BRILL Publishers. https://brill.com/edcollbook-oa/title/68845

Cheng J, Fernando R, Cheng H, *et al.* (2022). G3, 12, jkab441, https://doi.org/10.1093/g3journal/jkab441

Mott R, Fischer C, Prins P, *et al.* (2020) Genet. 215:359-372. https://doi.org/10.1534/genetics.120.303153

Zhao T, Wang F, Mott R, *et al.* (2024). Genet. 226:iyad210. https://doi.org/10.1093/genetics/iyad210