Simulating genomic prediction with functional genomic information in pigs

Emmanuel Osei

Independent project in Animal Science, EX0870 • 30 hp
Swedish University of Agricultural Sciences, SLU
Faculty of Veterinary Medicine and Animal Science/ Department of Animal Breeding and Genetics European Master in Animal Breeding and Genetics/ Master's programme in Animal Science
Uppsala, Sweden, 2024

Simulating genomic prediction with functional genomic information in pigs.

Emmanuel Osei

Supervisor:	**Martin Johnsson, PhD, SLU, Department of Animal Breeding and Genetics, Uppsala, Sweden**
Assistant supervisor:	Armin Schmitt, PhD, Georg-August-Universität Göttingen, Division of Breeding Informatics,
Examiner:	Erling Strandberg, PhD, SLU, Department of Animal Breeding and Genetics, Uppsala, Sweden

Credits:	30 credits
Level:	Second cycle
Course title:	Independent project in animal science
Course code:	EXO870
Programme/education:	European Master in Animal Breeding and Genetics/ Master's programme in Animal Science
Course coordinating dept:	Department of Animal Breeding and Genetics
Place of publication:	Uppsala, Sweden
Year of publication:	2024

Keywords: simulation, genomic selection, functional genomics, pig breeding

Swedish University of Agricultural Sciences
Faculty of Veterinary Medicine and Animal Science Department
Department of Animal Breeding and Genetics

Abstract

Genomic selection has been around for more than 10 years in the pig industry. Despite the popularity, accurately predicting complex with genome-wide markers is still challenging. The current limitation in prediction accuracy is, at least partially, due to the use of markers that may not directly influence the traits of interest. Some studies suggest that whole-genome sequencing can improve genomic predictions within and across breeds, but empirical results are inconclusive. Another proposed solution to the problem is the use of functional genomics in genomic selection. Functional genomics studies genes and their interactions using techniques like expression quantitative trait loci (eQTL) and chromatin immunoprecipitation sequencing. Among these techniques, eQTL studies identify genomic regions influencing gene expression levels, which in turn affect conventional phenotypes. Given this complex connection among genomic variants, regulation of gene expression and phenotype formation, there is, therefore, a growing recognition of the potential benefit of integrating functional genomic information, particularly expression quantitative trait loci, into genomic selection models to enhance prediction accuracy. This study was performed in an attempt to find out under what circumstances eQTL markers will improve selection accuracy in genomic selection for pig breeding.

The project considered a simulation of a breeding programme selecting a single polygenic trait affected by 50 genes. Using the R package “AlphaSimR” the pig genome and base population were obtained from a coalescent simulation and used to simulate breeding activities. An eQTL mapping study was performed using a simulated 100k chip. Genomic selection was performed under three scenarios. Two of the scenarios used variants from the eQTL study as markers, and the third used simulated SNP markers represented as 1) eQTL only, 2) eQTL plus SNP, and 3) SNP only. Results did not indicate any big impact of eQTL markers on genomic selection, despite a successful eQTL mapping. The project highlighted some limitations indicating that further investigation of eQTLs in genomic selection is necessary.

Keywords: simulation, genomic selection, functional genomics, pig breeding

Table of contents

List of figures

Introduction

Before using genomics in the pig industry, pig improvement programmes relied on best linear unbiased prediction (BLUP) models with relatedness between animals based on pedigree information (Henderson, 1976; Quaas, 1988). From the 1970s to the 2000s, significant resources were invested in finding the most efficient evaluation model for traits of economic importance. For example, in breeding programmes, estimating genetic trends is important because it makes it possible to evaluate the advantages of the programme and genetic alterations. For commercial breeding, the conventional techniques suggested by Hill (1972a) incorporating control groups were expensive. Alternative strategies were proposed by Blair and Pollak (1984), while Sorensen and Kennedy (1984) showed that mixed models work well in some situations, removing the necessity for a control group. Since then, phenotypic and pedigree information has been the major contributing factors to genetic progress in the pig industry. However, alongside these developments, particularly in the late 1970s and early 1980s, techniques were developed that allowed DNA investigation to become available, leading to the discovery of polymorphic markers in the genome. Soller and Beckmann (1983) hypothesised that markers would be beneficial in constructing more precise genetic relationships for determining parentage and identifying quantitative trait loci (QTL). However, the widespread of this new marker technology would be cut short because of the high cost of genotyping animals.

One of the benefits that came with the development of marker genotyping was the birth of marker-assisted selection (MAS). The promise of MAS was that if a profile appeared to contain genes that directly affected the trait, then potentially great genetic gains could be achieved by selecting animals with the desired marker profile. For instance, in the early 1990s, a marker test known as Hal-1843 became available for commercial pig breeding (Knol et al., 2016). This test allowed the use of MAS against a recessive mutant allele that causes malignant hyperthermia, which is often lethal in stressful conditions and can also result in poor meat quality.

However, although it worked very well for traits affected by one or a couple of genes, MAS did not contribute appreciably to livestock improvement. Simply because

most of the traits of interest were quantitative and complex. Despite the setbacks in using marker-assisted selection, some researchers believed that livestock breeding could still benefit from genomic information to generate accurate breeding values, and this could only work if a dense SNP assay that covers the entire genome were available.

In an attempt to use genomic information again, Meuwissen and colleagues (2001), in a simulation study, extended the idea of incorporating marker information into BLUP, a method introduced by Fernando and Grossman in 1989. Meuwissen et al.'s (2001) extension of Fernando and Grossman's method is what is now widely known as genomic selection. Genomic selection relies on genome-wide markers such as single nucleotide polymorphisms (SNPs) with the idea that each marker contributes to some extent to the trait of interest. In genomic selection, a predictive model is constructed with genotype and phenotype information. The animals' genomic estimated breeding value (GEBV) is then predicted by summing the estimated effects of its segments across the genome (Meuwissen et al., 2001). Genomic selection has been instrumental in livestock breeding, primarily because the breeding values of animals can be estimated as soon as DNA information is available. The advantage of genomic selection over conventional tools lies in its power to a) increase prediction accuracy and b) decrease the generational interval.

Despite the popularity and advantage of genomic selection over conventional tools, accurately predicting traits that are highly polygenic with genome-wide markers is still a challenging task. The current limitation in prediction accuracy is, at least, partially due to the reliance on neutral markers. To solve the problem, Meuwissen & Goddard (2010) suggested that whole-genome sequencing (WGS) could enhance genomic predictions within and across breeds. However, results from empirical studies have been inconclusive thus far. For instance, in empirical work, the accuracies of genomic predictions based on preselected WGS variants were not robust across traits and lines, and the improvements in prediction accuracy that were achieved with WGS compared to standard marker arrays were generally small (Ros-Freixedes et al., 2022). Another suggestive solution for the challenge can be found in the reports of Poland and Rutkoski (2016). They suggested that these challenges may be due to the intricate biological processes between genes and phenotypes (Poland and Rutkoski, 2016). These processes, such as transcription,

translation, metabolism, gene interactions, and epigenetic modifications, could affect the expression of economic traits. However, the details regarding these procedures are largely unknown. It is worth mentioning that economically important traits have a polygenic architecture, and causative variants have small effects, making their detection and quantification difficult. Most of these causal variants, with small effects, are reported to be likely to be located in regulatory regions and affect complex traits through changes in gene expression (Clark et al., 2020). This could be true because several genome-wide association studies in humans have shown that more than 88% of variants associated with disease traits lie outside protein-coding regions (Edwards et al., 2013). These non-protein coding regions have diverse functions that contribute to the complexity and regulation of gene expression, which eventually affects phenotypes such as diseases. Secondly, because humans and pigs have much in common genetically and hence, non-protein coding regions could hold valuable information also for pigs. Given the complex connection among genomic variants, regulation of gene expression and phenotype formation, there is, therefore, a growing recognition of the potential benefit of integrating functional genomic information, particularly expression quantitative trait loci (eQTL), into genomic selection models to enhance prediction accuracy (Clark et al., 2020). The choice of eQTL markers is because eQTLs are those genomic loci associated with regulating gene expression levels, which in turn affect economic traits.

A successful integration of functional genomics into genomic selection could make genomic prediction more accurate for complex traits. Building on this rationale, the objectives of this simulation study are therefore to a) simulate a pig breeding program to model how causative variants affecting expression levels of various genes impact quantitative breeding trait goals, b) to evaluate the potential of genomic prediction with eQTL information, and c) conduct genomic resolution analysis of eQTL mapping to learn how successful eQTL mapping is at in identifying markers that are in close proximity to the causative eQTL variant.

Literature review

2.1 Pig breeding for food security

Successful animal breeding in developed regions involves high-tech evaluation methods, biotechnologies, and organisation. In fast-reproducing species like pigs, the breeding business has been taken over by large companies exploiting the advantage of centralised data management and decision-making to improve genetic gains. For example, compared to other domesticated terrestrial animals, pigs are highly prolific and have relatively short gestation periods, contributing to their success in the breeding subsector. Pigs are useful animals, serving as food, particularly in these times of high food demand due to global population growth. This has put immense pressure on agricultural systems and resources and threatens global food security (Alexandratos & Bruinsma, 2012). For instance, In November 2022, the world's human population reached 8 billion, which is a significant increase from the estimated 2.5 billion in 1950 (Leeson, 2018). According to projections, it is expected to reach 9.7 billion by 2050 and 10.4 billion by 2080 (Ritchie & Roser, 2019). Global population growth is closely linked to global food security. However, addressing this complex issue requires a comprehensive approach that goes beyond simply providing enough food to meet the needs of a growing population. Achieving true sustainability requires careful consideration of various factors, including food production, environmental impact, and resource conservation. In an attempt to ensure a sustainable future, the pig industry plays a role in the establishment of sustainable pig breeding programmes in order to produce enough food for the increasing human population. However, when selecting the best-performing pigs, the industry faces genetic complexities due to the influence of multiple genes on most economically important traits.

The remaining part of this chapter will review the literature surrounding the enhancement of genomic selection by integrating functional genomic information. The choice of functional genomic information is also discussed.

2.2 Quantitative traits in pigs

Several genes influence most economic traits in pigs, and the genetic complexities of these traits in pigs become a challenge in advancing breeding strategies. An attempt in the past was to identify major genes that affect the trait of interest (Zak et al., 2017). Although some level of success was achieved, there are situations where a limited number of genes are insufficient to explain all the observed variations in the trait. According to Kearsey and Pooni (2020), this inadequacy in explaining the observed variations is sometimes evident later. Currently, genome-wide association studies (GWAS) and genomic selection are used in quantitative trait studies simply because they are sensitive and can capture significant background variations in other unidentified genes. Quantitative traits are classified into several categories, including morphological traits (e.g., litre size, weight at slaughter, muscle depth), physiological traits (e.g., blood pressure), and behavioural traits (e.g., aggression) (Larzul, 2021). Moreover, molecular phenotypes like gene expression levels and high and low-density cholesterol levels are also considered quantitative traits (Mackay, 2009). An understanding of the type of economic trait is therefore important in order to determine which animals should become parents of the next generation.

2.3 Selective breeding

Selective breeding has evolved. During the 1970s, it was based on phenotypes and pedigree information, lacking genetic or molecular data. Nowadays, selective breeding relies on phenotypic and genotypic information, resulting in more predictable outcomes. Currently, selective breeding is approached in two ways: the conventional approach based

on phenotypes and the modern approach using genetic engineering and advanced technologies to identify genes responsible for desirable traits such as disease resistance or increased yield (Gjedrem & Baranski, 2010; Meuwissen et al., 2016). The conventional approach uses tools like the variance component analysis and best linear unbiased prediction (BLUP) models, while the modern approach uses techniques like marker-assisted selection (for monogenic traits) and genomic selection.

2.4 Conventional approach

2.4.1 Variance component analysis

The genetics of quantitative traits utilises biometrical techniques to understand how different factors influence the phenotype (Kruuk et al., 2014). The genetic model of a phenotype is influenced by genetics and environment and can be expressed as P = f (G, E) (Ritchie et al., 2015). Here, P represents the observed phenotype, while G and E represent the genetics and environment, respectively. The decomposition of this model allows us to focus on a particular component while keeping the other constant. For instance, by assuming no genotype by environment interaction, the phenotypic variance can be divided into genetic and environmental variance. This genetic variance can be further divided into additive, dominance, and epistatic variance. The additive variance is of particular importance because it can be used to predict the performance of the progeny (Plomin & Deary, 2015). The variation in the phenotype due to additive variance is known as heritability (Visscher & Goddard, 2015). In animal breeding, the narrow-sense heritability is particularly significant for traits influenced by additive genetic effects such as body weight or muscle depth. An important aspect of genetics is the inheritance of traits from one generation to the next. Estimated breeding values (EBVs) are a measure of an animal's genetic potential and are influenced by genetic variance. Animals with higher EBVs are more likely to pass on favourable traits to their offspring, contributing to genetic improvement in the population. Therefore, selecting individuals with higher EBVs enhances the genetic quality of a population.

2.4.2 Best Linear Unbiased Prediction

Genetic evaluations guide selection decisions by identifying individuals with superior genetic potential for desired traits. It is important to understand that certain statistical methods used to analyse genetic data are better suited for phenotypic measurements obtained from planned experimental designs and with balanced data sets. However, such situations may only be possible within the laboratory (biosecure animal breeding companies) or greenhouse experimental settings (as in the case of plant breeding). Data collected from agricultural species like pigs are generally highly unbalanced and fragmented due to numerous relationships. Therefore, when we try to fit such data into conventional statistical techniques, we might introduce bias and lose vital information. However, with a mixed model methodology like BLUP, we can effectively estimate genetic parameters like variance components, heritability, and breeding values (Demidenko, 2013). The mixed model methodology is designed to handle extended pedigrees, unequal family sizes, sex-limited traits, and assortative mating. This methodology helps to estimate the genetic covariance structure and improve predictions, even with incomplete or unbalanced data.

2.5 Marker-based approaches

The accuracy of predictions in conventional animal breeding has not always been reliable. This has made marker-based tools popular in pig breeding, particularly genomic selection, as they provide more accurate predictions. Previously, marker-assisted selection was popular; however, it has been largely superseded by genomic selection, with the exception of monogenic traits.

2.5.1 Genomic Selection

Genomic selection is a state-of-the-art tool widely implemented in commercial pig breeding programs. The genomic selection model starts with a simple linear model generally referred to as ordinary least squares or least square regression, which is of the form $Y = 1_n\mu + X\beta + \varepsilon$. (Pohlman & Leitner, 2003). Where Y = n x 1 vector of observed phenotypes, μ is the global mean, β = p x 1 is a vector with marker effects, ε = n x 1 is a vector of random residual effects with $\varepsilon \sim N(0, \sigma_e^2)$. X is a design matrix of dimension n x p, where n is the rows which contains the number of individuals or the genotypes, and p refers to the columns and it contains the markers. This basic model has been modified and improved in a number of ways. Currently, there are models like ridge regression, least absolute shrinkage and selection operators, and Bayesian models, just to mention a few. The implementation of genomic selection in pig breeding has revolutionised the industry by making it more efficient and cost-effective to select choice parents. For instance, it has improved selection of traits such as feed efficiency, growth rate, and meat quality. In recent years, the pig industry has seen significant improvements in feed efficiency, which is a crucial trait in pig breeding. For example, according to Tang et al. (2017), feed accounts for 60-70% of the total cost of pig production. Therefore, selecting pigs with improved feed efficiency can significantly reduce production costs. Furthermore, genomic selection has also helped in selecting meat quality in pigs. The quality of meat is an essential factor in the pig industry, as it affects the price and demand for pork. Pig breeders are able to select and breed animals with the desired meat quality traits, such as marbling, tenderness, and juiciness. This results in higher-quality pork products that meet the demands of consumers (Xu et al., 2019).

Despite recent advances in breeding techniques, accurately predicting traits controlled by multiple genes remains challenging due to the complex interactions and cumulative effects of these genes. To address this, Meuwissen and Goddard (2010) proposed that whole-genome sequencing (WGS) could improve genomic predictions within and across breeds. Early simulations supported this idea and suggested that WGS could be a valuable tool for enhancing breeding outcomes (Meuwissen and Goddard 2010). However, current empirical evidence on the effectiveness of preselected whole-

genome sequencing (WGS) variants in genomic predictions has been inconclusive. For example, Ros-Freixedes and colleagues (2022) found that the accuracies of genomic predictions were not consistent across traits and lines and that the improvements achieved with WGS compared to standard marker arrays were generally small (Ros-Freixedes et al., 2022).

In a different proposal, Poland and Rutkoski (2016) suggest that these challenges may be due to the intricate biological processes between genes and phenotypes. These complex processes include several downstream mechanisms, such as transcription, translation, metabolism, gene interactions, and epigenetic modifications. These factors could influence the expression of the economic trait of interest. Although the details regarding these procedures are mainly unknown, it has been reported that these regions contain regulatory elements such as enhancers, promoters, and silencers. These elements are pivotal in controlling when and to what extent genes are transcribed and translated into proteins. They contribute to the precise regulation of gene expression in different tissues and under various conditions. A different report by Pan et al. (2017) shows that several genome-wide association studies have shown that the majority of variants associated with the traits of interest lie in regulatory regions and could hold information vital to the expression of traits of economic importance. This insight highlights the importance of researching functional genomics in order to gain a better understanding of the function and interaction of genes within an organism.

2.6 Functional genomics

Functional genomics can help prioritise markers likely to affect traits and thus improve the accuracy of genomic prediction. Functional genomics is a branch of genomics that studies genes, gene functions and their interactions (Pevsner, 2015). The primary objective of functional genomics is to better understand how different aspects of an organism interact to produce a specific phenotype. Various techniques, such as expression quantitative trait loci studies and chromatin immunoprecipitation sequencing (ChIP-Seq), are used to better understand how these genes function within the pig genome

(Nonneman & Lents, 2023). Using expression quantitative trait loci (eQTL) studies enables us to explore genetic variants influencing gene expression levels in pig populations (Liu et al., 2020). Researchers are able to uncover regulatory elements controlling RNA abundance and learn about the underlying genetic variants contributing to observed expression variation. On the other hand, ChIP-Seq identifies protein-DNA interactions, revealing the binding sites of transcription factors, while histone modification profiling elucidates the epigenetic modifications associated with active or repressed gene expression (O'Geen et al., 2011). The use of these techniques provides a more complete picture of the complex regulatory mechanisms underlying cellular processes.

Among these techniques, eQTL studies have a more direct impact on economic traits, as they influence gene expression levels, which in turn affect conventional phenotypes.

2.6.1 Expression quantitative trait loci

Standard eQTL analysis involves testing the direct association between markers and gene expression levels. This test is usually conducted on tens or hundreds of individuals. Using the GWAS approach, eQTL mapping allows for the identification of new functional loci without requiring any previous knowledge about specific cis or trans-regulatory regions, which is indicative of how close an eQTL is to the causative variant. Over the past decade, eQTL analysis has been applied to various diseases and complex traits, including cancer, cardiovascular disease, obesity, and psychiatric disorders (Battle et al., 2014). These studies have identified hundreds of genetic variants that are associated with gene expression levels and have provided insights into the biological pathways and mechanisms that underlie disease susceptibility. In addition to identifying genetic variants that regulate gene expression, Battle and colleagues argue that eQTL analysis could be used to identify novel drug targets and biomarkers for disease diagnosis and prognosis (Battle et al., 2014).

In the pig industry, eQTL analysis has been used to identify genetic variants that affect gene expression levels in various pig tissues. For example, in a study by Li et al.

(2019), eQTL studies identify genetic variants associated with differences in backfat thickness in pigs, an important trait for meat quality and production efficiency. Another study by Zhang et al. (2019) employed eQTL analysis to identify genetic variants associated with immune-related gene expression levels in response to PRRS infection in pigs, which could inform disease prevention or treatment strategies.

In meat quality traits in pigs, Ballester et al. (2017) identified 92 significant eQTL associated with lipid metabolism across seven chromosomal regions. In a different report, Criado-Mesas et al. (2020) detected 186 significant eQTL associated with six genes regulating lipid metabolism and fatty acid composition. These findings show that genomic variants significantly affect the expression of genes related to muscle metabolism, lipid metabolism, and fatty acid composition, among others, which contribute to the observed variation in meat quality traits. To this end, gene expression can be considered an "intermediate phenotype" since it is expected to be more closely linked to genetic variations than conventional phenotypic characteristics.

2.7 Expression quantitative trait loci for genomic selection

Incorporating eQTLs into genomic selection is a promising concept, and therefore, modelling gene expression data as a predictor in genomic selection is expected to explain more complex biological regulation processes and potentially increase predictive accuracy. For instance, in human medicine, González-Reymúndez et al. (2017) integrated whole-omics data (including whole-genome gene expression profiles) into breast cancer prediction and demonstrated that omics and omic-by-treatment interactions explain a sizable fraction of the variance of survival time, and further suggested that whole-omic profiles could be used to improve prognosis prediction accuracy among breast cancer patients. In insect studies, Li and colleagues investigated the usefulness of transcriptome data for predicting traits in Drosophila melanogaster using 185 inbred lines (Li et al., 2019). They utilised two different combined methods, first, an approach integrating genomic and transcriptome data, GTBLUP (i.e., combining GBLUP and transcriptome

BLUP (TBLUP)). The second also combined GBLUP, and a reproducing kernel Hilbert spaces (RKHS) forming a GRBLUP. Overall, GRBLUP and GBLUP showed similar predictive abilities for most traits, but GRBLUP explained a higher proportion of the phenotypic variance (Li et al., 2019). Notably, only one trait, olfactory perception to Ethyl Butyrate in females, demonstrated a significantly higher predictive ability for GRBLUP (0.23) than GBLUP (0.21).

In plant breeding, Loh and colleagues conducted a comprehensive investigation in 2011, specifically focusing on soybean plants and their resistance to phytophthora sojae. The study compared the effectiveness of using marker genotype data versus gene expression data in predicting the resistance levels of these plants. Based on their findings, it was observed that using gene expression data yielded better accuracy in predicting resistance than using genotype markers alone (Loh et al., 2011). Similarly, in a study conducted by Guo and colleagues 2016, the accuracy of gene expression levels alone was low in maize breeding. However, when these levels were combined with SNP markers, the predictive abilities were significantly increased and were found to be either high or comparable to those achieved with GBLUP alone (Guo et al., 2016).

2.7.1 Concept and method

Incorporating eQTLs in genomic selection requires, first, conducting eQTL mapping studies. The genomic loci identified in this step represent potential eQTLs which can be used directly on chips. However, preselecting sets of variants is very important as currently available methods for genomic prediction are not yet capable of handling very large datasets without exorbitant computational resources (Ros-Freixedes et al., 2022). To this end, the eQTLs markers can be added onto chips as subsets from the pool of eQTLs. The eQTL / SNP chip created at this point is treated exactly as discussed earlier in conventional genomic selection.

Materials and Methods

3.1 Simulation procedure

The AlphaSimR package in R was used in the entire simulation procedure (Gaynor et al., 2021). The general simulation scheme consists of i) simulation of the founder population, ii) simulation of quantitative trait loci, iii) simulating gene expression traits, iv) creating a new population based on the founder population, v) creating traits based on the gene expression traits, vi) simulating breeding programme, vii) performing eQTL mapping, and vii) performing genomic selection based on three different scenarios.

3.2 Simulation founder population and traits

To start the simulation, founder haplotypes were generated using the runMacs function. The runMacs function depends on the Markovian Coalescent Simulator algorithm developed by Chen and colleagues (2009), which efficiently simulates haplotypes under any arbitrary model of population history. The “GENERIC” setting was used as it is meant to be a reasonable all-purpose choice (Gaynor et al., 2021). The generic population history tries to reflect the shrinkage shape of domesticated population histories. 350 pigs were simulated across 18 chromosomes for a generic species since the current version of AlphaSimR does not have a built-in model for pigs. Chromosomes were simulated with the asrhelper package to have different sizes. These were simulated using the pig genome and linkage map reported by Warr et al. (2020) and Tortereau et al. (2012). Of these 350 pigs, sex was set using the "yes_sys" condition in AlphaSimR. The "yes_sys" condition sets the sex ratio to 1:1. Simulation parameters were then set using the SimParam class, and SNP chip containing 5550 markers per chromosome was designated for downstream analysis. This number of SNP per chromosome corresponds to a 100K SNP chip density. Four different SNP chips were simulated in total for this study using the addsnpchip feature in the AlphaSimR package. The densities of the SNP chips were 100K and three 80K chips. The 100K chip was specifically used for the eQTL mapping. The rest of the chips were used in the different genomic selection scenarios

described below. Allele frequency (AF) was calculated by first obtaining the segregation site using the pullSegSiteGeno function in the AlphaSimR package and then calculating the AF based on the matrix data obtained. A threshold of 0.01 was set to eliminate very low frequencies, and a threshold of 0.99 was set to eliminate very high allele frequencies soon to be fixed in the population. The simulation focused on 50 genes (ng) and 11 genomic loci (nl) affecting the expression of these genes and their trait-related parameters. The motivation for this number of genes can be found in Teng et al. (2022) report on the impact of sample size on the detection of eQTLs. The sample size of 350 was determined using power analysis with the genpwr package in R for the major effect (Moore et al., 2020). The sample size estimation parameters in the package were as follows; alpha at 10^-4, Power (1 – β) at 0.8, and minor allele frequency set at p, where p is a vector of minor allele frequencies of the simulated loci. By examining the relationship between sample size and the ability to detect eQTL, the study can assess the reliability and robustness of eQTL findings and make informed decisions about the sample size required for future studies. Additive variances were simulated such that one of the loci had a very high variance of 0.5 contributions, explained, while the 10 had very low variances of 0.05, summing up to one. Additive effects were set based on the additive genetic variance due to each locus as $a = \sqrt{V_a / 2p(1-p)}$ where V_a is the variance, p is the allele frequency. The predicted population mean was also calculated using the additive effects as $M = \sum_{i=1}^{n} a.(p_i - 1 + p_i)$. Heritabilities were set for the gene expression traits based on the parameters of the simulated founder population. They were set at relatively high values of 0.5 for all 50 genes; this is mainly because of the expectation that gene expression is a simpler trait than the downstream quantitative trait. Genetic and environmental components were incorporated into the defined gene expression traits.

3.3 Simulating new populations

A new population, considered generation one, was simulated based on the parameters of the founder population. Phenotypic traits were manually created based on the gene expression traits. The concept is straightforward: a function was created using R

using i) gene expression trait indices, ii) gene expression coefficient, iii) heritability and iv) AlphaSimR "pop class" as an input. The concept was to identify genomic variants affecting gene expression, and the gene expression, in turn affecting a final trait. In mathematical terms, $E_j = \sum_{i=1}^{nl} B_{G_i} . G_i + \epsilon_{G_j}$ where E_j is gene expression for gene j, G_i is the genomic variants (for i = 1,2,3..nl), B_{G_i} is the effects of genomic variants i on gene expression, nl is the total number of genomic variants (i.e., 11), and ϵ_{G_j} is the random error term for gene expression. The effects of gene expression on a trait is mathematically expressed as $T = \sum_{j=1}^{ng} B_{E_j} . E_j + \epsilon_T$ where T is the final trait, E_j is the gene expression for gene j, ng is the total number of genes in the model (i.e., 50), B_{E_j} is the effect size of gene expression j on the final trait.

The combined effect of genomic variants on gene expression and the effects gene expression on the final trait is mathematically expressed as $T = \sum_{j=1}^{m} B_{E_j} . (\sum_{i=1}^{n} B_{G_i} . G_i + \epsilon_{G_j}) + \epsilon_T$. The environmental variance was set based on heritability of 0.5. The phenotypic and genetic value obtained from the trait function as described above was added to the simulation object in AlphaSimR.

3.4 Modelling breeding programs

A first breeding program was designed to include crosses and selections to model successive generations. The breeding programme consisted of 300 dams and 50 sires with 300 crosses and an expected progeny size of 12 per cross, yielding a total number of 3600 young per generation. Through continuous selection and crossing, the program sought to produce a population with genomic variants affecting gene expression that, in turn, also affects the final trait of economic importance. In total, the simulation of this breeding program had 9 generations in all.

3.5 Expression quantitative trait mapping

The AlphaSimR data structures were used to prepare genotype and phenotype data for eQTL analysis from the breeding programme described above. The GWAS function in the rrBLUP package in R was used to perform the expression quantitative trait loci mapping. The analysis was based on the mixed model $y = X\beta + Zg + S\tau + \epsilon$ as discussed in Yu et al. (2006) report, where y is a column vector of phenotype values (n x 1) for n number of pigs, X is the design matrix for fixed effects (n x p),β is a vector of fixed effects that can model both environmental factors and population structure with dimensions (p x 1) (Endelman, 2023). Z is the design matrix for the random effect g (n x m) with g as a column vector of random genetic effects (m x 1)The variable g represents the individual pig as a random effect with Var(g) = $K\sigma^2$, reflecting the additive genetic contribution, where K is an (m x m) genetic relationship matrix based on genomic information, and σ^2 denotes the additive genetic variance. The vector g contains the breeding values for m individuals. Additionally, S is the design matrix for the additive SNP effect τ (n x k) with τ as a column vector of additive SNP effects (k x 1). The ϵ is a column vector of residuals (n x 1) with residual variance of $Var(\varepsilon) = I\sigma_e^2$ as discussed in Endelman's (2023) updated version of the rrBLUP package.

The expression and genotypic information were used as input, along with relevant parameters such as a minor allele frequency threshold of 0.05 and the inclusion of population structure correction. The analysis focused on identifying genetic variants associated with gene expression levels. The significant eQTLs from the eQTL mapping were identified and stored as a variable and used in downstream analysis.

3.6 Genomic Resolution

3.6.1 True location of the causative variant

To ensure that the eQTL mapping results were accurate, the distance between the identified eQTLs and the actual location of the causative eQTL variant was calculated, and the absolute values were recorded. The genomic distance from the true location on the chromosome was determined for each significant eQTL. The calculated distances

were then used to evaluate the accuracy of the eQTL mapping in capturing the true location of the causative gene and measured in centiMorgan.

3.7 Simulating genomic selection

Genomic selection was conducted under three different scenarios. Scenario one consisted of only SNPs, scenario two consisted of SNPs and eQTLs, and scenario three only of eQTLs.

3.7.1 SNPs only

Out of the 9 simulated generations, the last six were used in genomic selection. The initial training population for the first generation was modelled using 3600 animals. As the generation progressed, the previous and the current populations were merged and used as the training data set. However, in the fifth (5th) and sixth (6th) generations, only the previous three generations are merged and used as the training data set with an 80K chip with only SNPs. For each generation in the breeding programme, genomic selection was performed using the RR-BLUP method as built in the AlphaSimR package. The selection criteria were the estimated breeding values (EBV) of the animals as predicted by the simulation model, and the top 350 animals were selected and used as parents in the next generation. The data of individuals in various generations were pooled together using the mergePops function in the AlphaSimR, enabling the combination of genetic information from the different generations. This was to mimic real-life scenarios in pig breeding programmes. Prediction accuracy was calculated by assessing the Pearson correlation between the genomic values predicted for the phenotypic trait and the true breeding values. Genetic mean and variance were obtained for each generation using the in-built meanG and VarG, respectively.

3.7.2 eQTL only

In this scenario, the chip used in the genomic selection contained pre-selected eQTLs. These eQTL markers were the significant eQTLs in the eQTL mapping studies. eQTL variants with log-transformed values equal to or greater than 4 were pre-selected. For the preselection, the top 16 significant eQTLs per gene were selected, and any other random eQTLs, making a total of 80,000 eQTLs. The decision for these numbers is due to the fact that economic traits are affected by variants with highly significant associations (usually very small) and others with little to no significant associations (usually many). Therefore, the top 16 variants per gene and the random variants represent the variants with highly significant associations and variants with little significant associations, respectively. These 80,000 eQTLs were added onto a marker chip and used in the genomic selection.

3.7.3 SNPs and eQTL

In this scenario, the chip used in the genomic selection included SNP markers and additional eQTL markers. For the preselection, the top 4 significant eQTLs per gene were selected, and any other random eQTLs as explained in the eQTL only scenario, making a total of 20,000 eQTLs. To test the combined impact of the eQTL variants and conventional SNPs on genomic selection accuracy, we added the 20K eQTL variants to a 60K SNP, making an 80K chip.

Results

The eQTL mapping results were obtained using a simulated gene expression dataset generated to mimic real-world scenarios. The dataset comprised expression profiles for 50 genes across a sample population influencing one conventional phenotype. Results for the genetic gain, selection accuracies, and genetic variance were obtained by observing the progression of the breeding programme over time. There were nine generations in all. The first four were used in the eQTL mapping analysis, and the last six generations were used for genomic selection. Each generation in the simulation is equivalent to a breeding cycle. The results show differences in using different marker scenarios in genomic selection; however, the differences are not robust.

4.1 Expression quantitative trait loci

The first part of the study focused on eQTL analysis, which showed significant associations between genetic variants and gene expression levels. The eQTL analysis detected 3333 significant eQTLs, using a threshold of 4 after log transformation. All 50 genes used in the study had at least one significant eQTL detected. Figure 1, here are randomly selected examples where significant variants were detected.

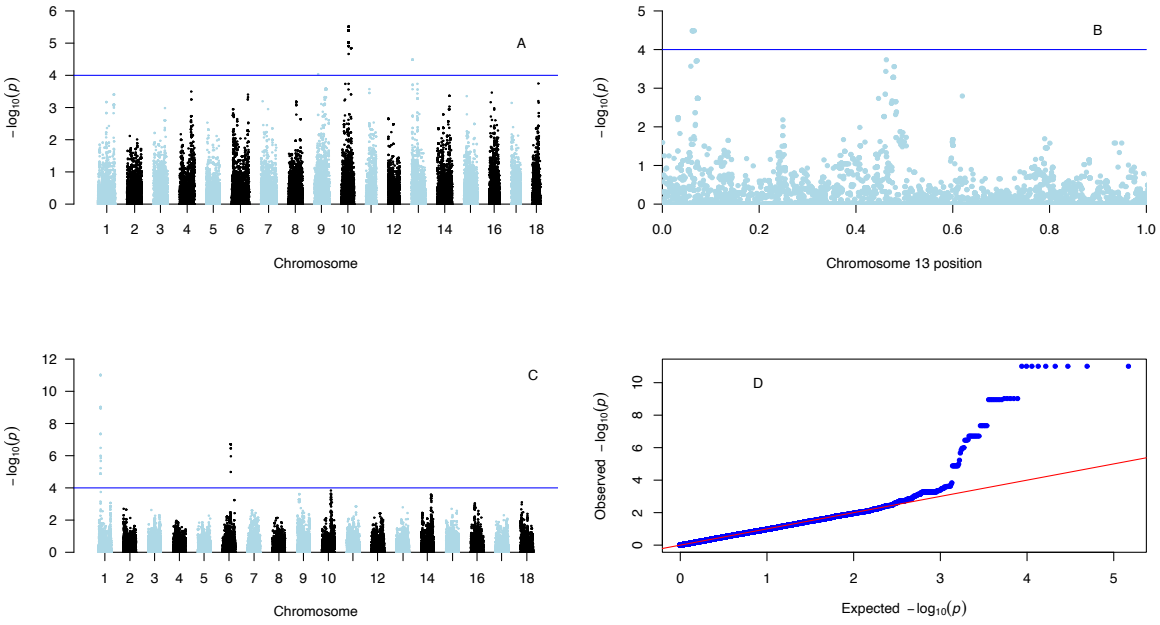

Figure 1: The Manhattan plots A, B, and C show the negative logarithm of the p-value on the vertical axis and chromosome number on the horizontal axis. The blue horizontal line indicates the significant threshold. A variant is significantly associated with gene expression when the variant is on or above the blue line. Plot B represents an up-close view of a single chromosome. Plot D is a Quantile-Quantile plot with the observed negative logarithm of p-values on the y-axis and an expected negative logarithm of p-values on the x-axis.

The highest count of significant eQTL from the analysis was observed in gene 9, with 189 eQTL variants. The lowest eQTL counts were in genes 19 and 40, with counts of 3 eQTL variants each. There was an observed average count of 67 and a standard deviation of 43.68.

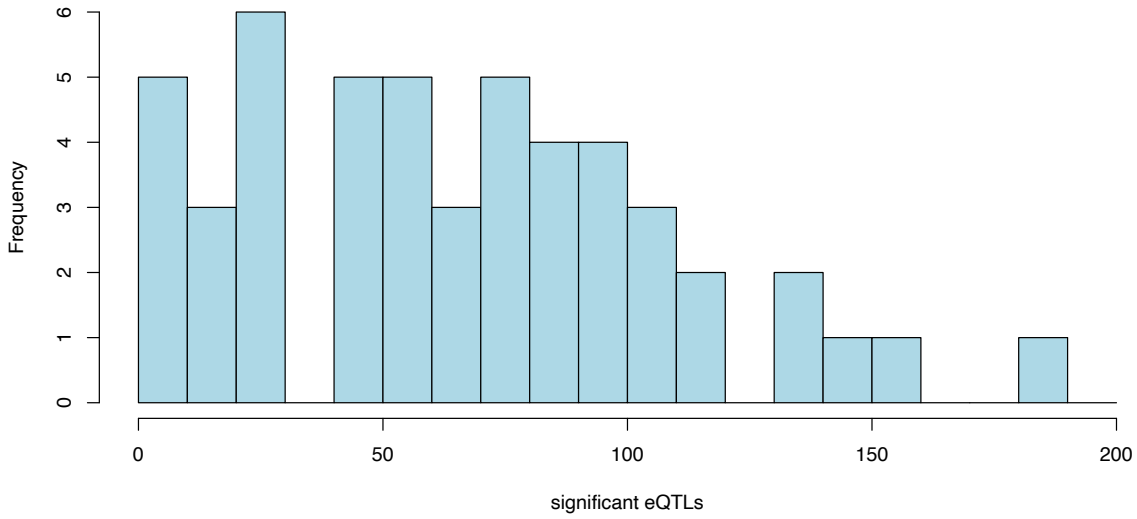

Figure 2: Counts of significant eQTLs within genes

Figure 2. Counts of significant eQTLs within genes. This distribution is not organised in any order. The y-axis shows the frequency of genes. The x-axis represents the range of significant eQTLs.

4.1.1 Genomic Resolution

In the context of eQTL mapping, genomic resolution refers to the ability to pinpoint the specific location of a genetic variant that is associated with a change in gene expression levels. Because this is a simulation study, the location of the causative eQTL variant is known, and therefore, the distances of lead eQTLs from the causative variant can be measured. In the analysis of distances, lead eQTLs from 42 gene expression traits were found between 0 and 10 cM from the causative variant (Figure 3). There were no lead SNPs between 40 and 60 cM from the causative variant.

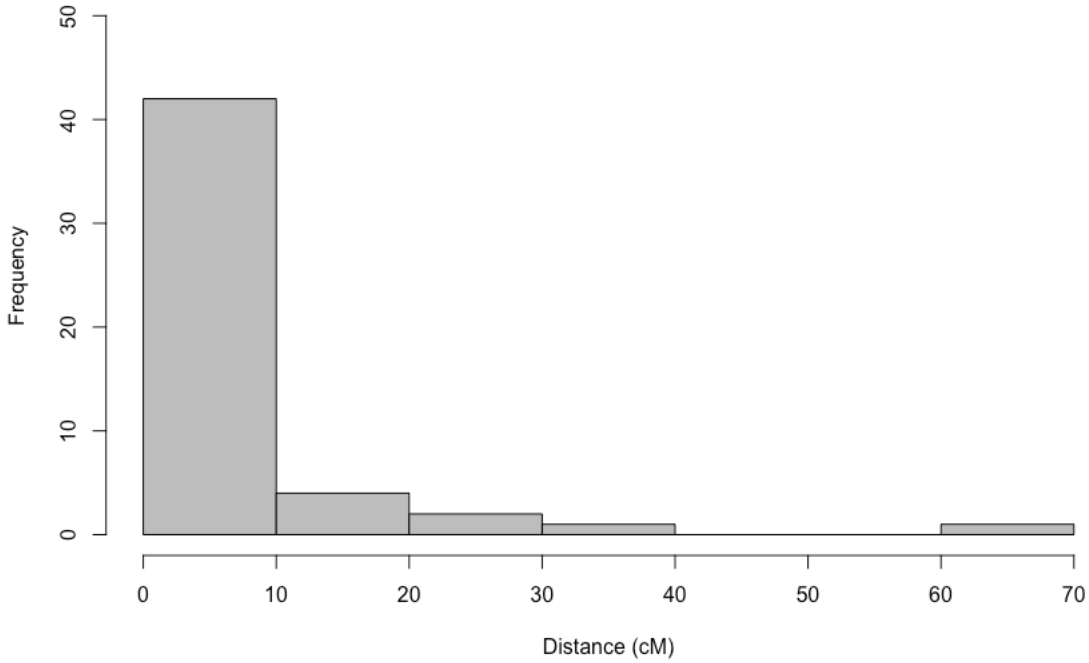

Figure 3: A histogram showing the distribution of distances of lead eQTL SNPs per gene expression trait from the causative variant. The y-axis represents frequency (i.e., the number of observations). The x-axis shows the distance grouped into bins. The vertical bars represent the frequency of observations within each bin.

The average distance between significant eQTLs and the causative variant of corresponding gene expression traits was also analysed. It was observed that eQTLs from 20 gene expression traits were located between 0 and 10 cM from the causative variant (Figure 4). eQTLs from only one gene expression trait were found between 70 and 80 cM from the causative variant (Figure 4). The resolution in this simulated study provides insight into how accurately the eQTL analysis finds the location of the causative variant. After completing the power analysis, we found that our eQTL mapping strategy has performed strongly. In particular, the study was able to successfully identify at least one significant eQTL for each gene we examined. This result provides solid evidence for the effectiveness of the approach in capturing the genetic variations that influence gene expression levels throughout the genome. The study design is comprehensive, and the statistical methods are reliable, as evidenced by the detection of significant eQTLs in all

genes. This suggests that we have well-calibrated our sample size, effect size assumptions, and chosen statistical thresholds to uncover the genetic architecture that governs variations in gene expression.

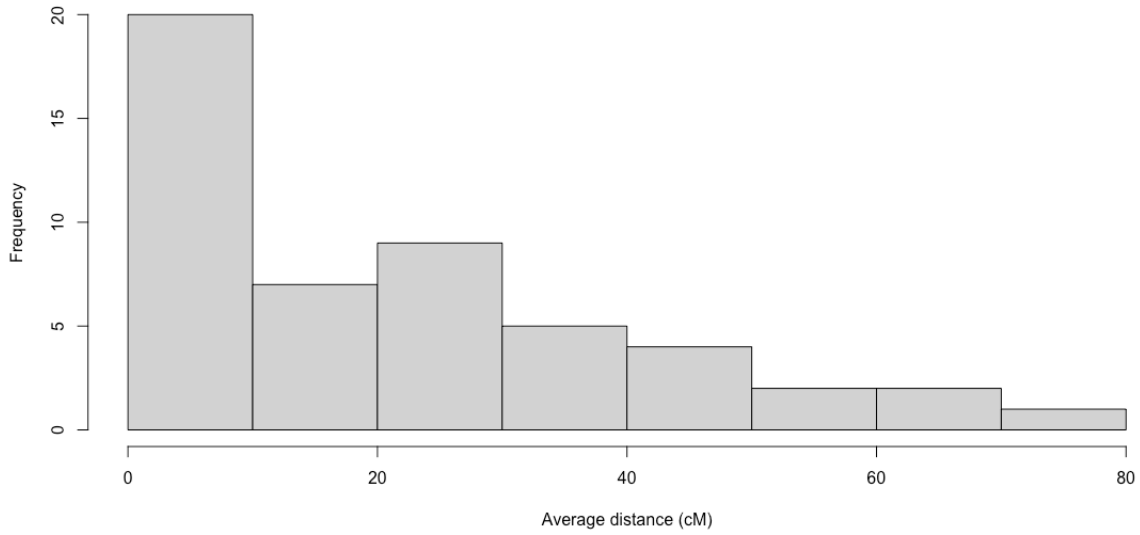

Figure 4: A histogram showing the distribution of average distances of significant eQTL from causal variants per trait. The y-axis represents frequency (i.e., the number of observations). The x-axis shows the distance grouped into bins. The vertical bars represent the frequency of observations within each bin.

4.2 Genomic Selection

The second part of this study analysed genomic selection using conventional SNPs, eQTLs, and a combination of both as markers. The analysis includes mean genetic gains, selection accuracies and genetic variance across the last 6 generations of the breeding programme. For some of the following figures, lines represent different marker scenarios, and the vertical bars in those lines represent standard deviations.

4.2.1 Genetic gain

The estimated genetic improvement was for the entire population in each breeding cycle. For easy interpretation, a baseline was established by subtracting generation one from all other generations. This makes generation one the reference point for all the scenarios under study, as shown in Figure 6. From the analysis, it was observed that the mean genetic gain ranged from 4.1 ± 0.4 to 20.5 ± 1.05 across generations one to six when eQTL alone was used as a selection marker. Similarly, when SNPs were only used as markers in the genomic selection, the mean genetic gain values ranged from 4.1 ± 0.5 to 20.5 ± 0.14 over the same generational span.

When eQTLs and SNPs were combined as selection markers, the analysis showed mean genetic gain values ranging from 4.05 ± 0.5 to 19.8 ± 1.3 across generations one to six, as seen in Figure 5 below.

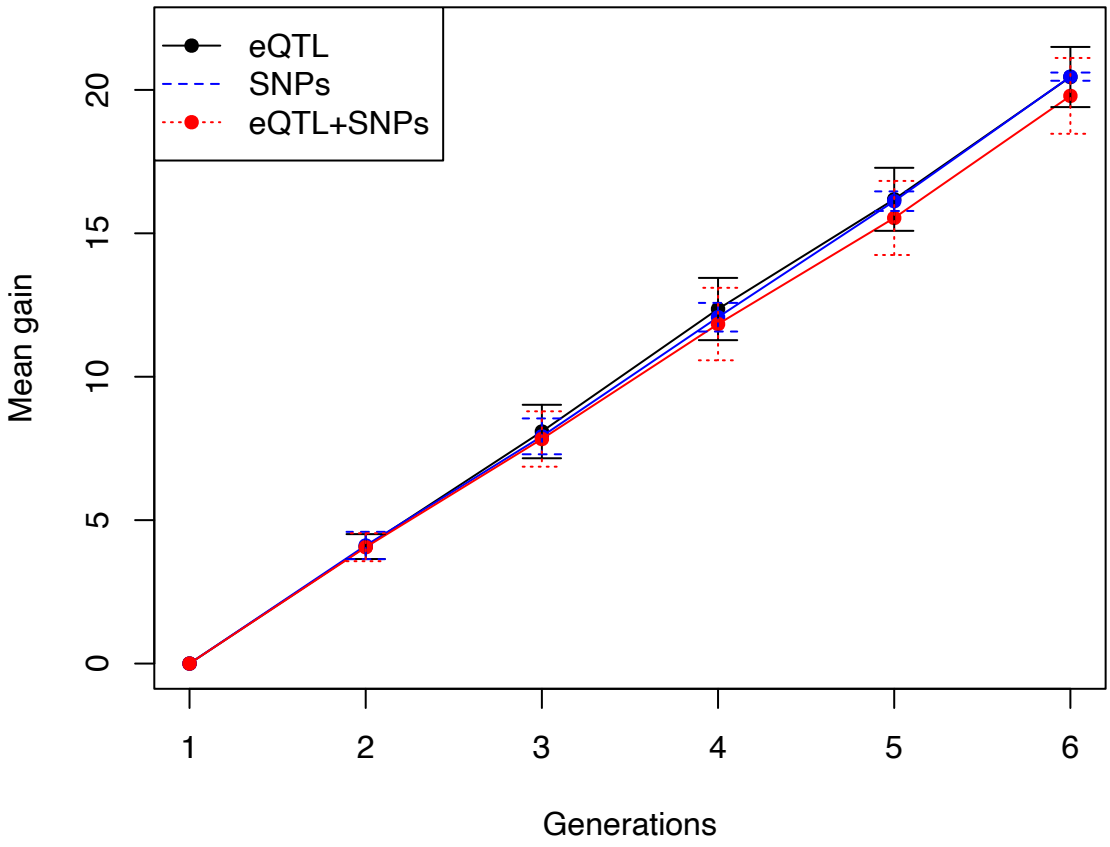

Figure 5: This figure provides a visual representation of genetic gain by different marker scenarios in genomic selection. The y-axis shows the mean gain, and the x-axis represents the generations. The black, blue, and red lines show the trends by eQTL, SNPs, and eQTL+SNPs, respectively.

4.2.2 Selection accuracies

The analysis showed that the mean selection accuracies were relatively constant between cases and increased somewhat over generations from 0.309 ± 0.06 to 0.33 ± 0.05 for generations two and six, respectively. For the scenario where SNPs were only used as markers in the genomic selection, the selection accuracies ranged from 0.31 ± 0.064 to 0.343 ± 0.064 across generations two to six. When eQTLs and SNPs were combined, the analysis showed mean selection accuracy values ranging from 0.31 ± 0.06 to 0.35 ± 0.07 across the same generational span, as seen in Figure 6 below.

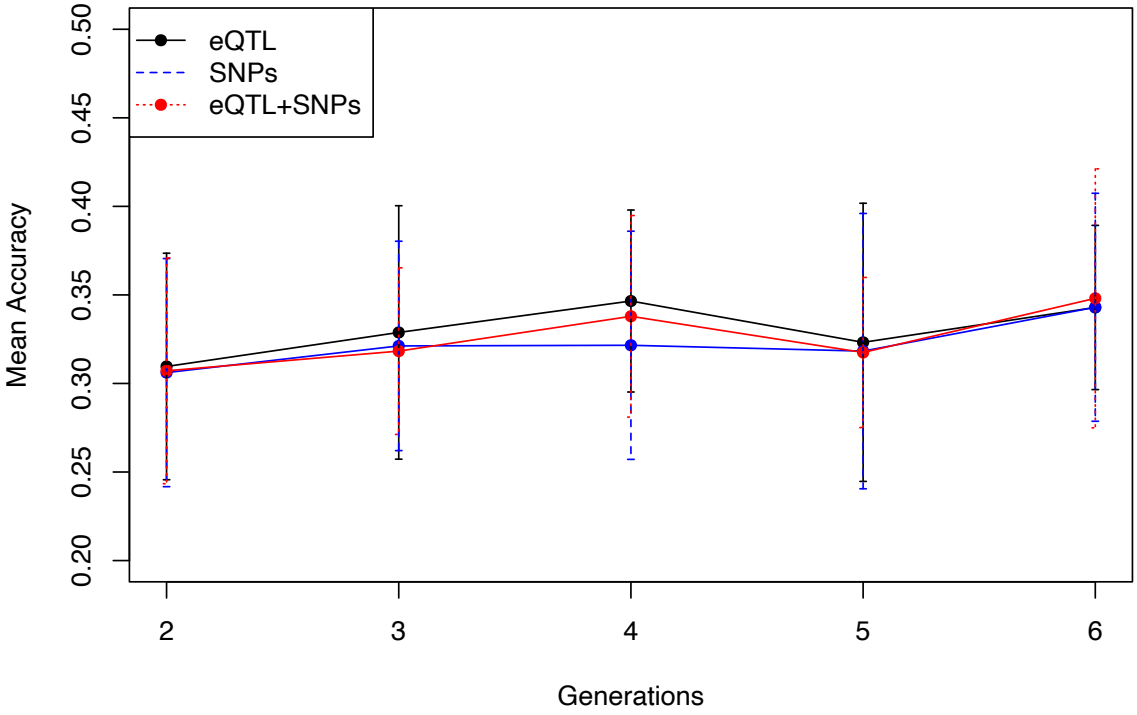

Figure 6: This figure provides a visual representation of selection accuracies of different marker scenarios in genomic selection. The y-axis shows the mean accuracy, and the x-axis represents the generations. The black, blue, and red lines show the trends by eQTL, SNPs, and eQTL+SNPs, respectively.

4.2.3 Genetic variance

The genetic variance was recorded for all generations in the genomic selection programme. From the analysis, it was observed that the genetic variance values ranged from 1.6 ± 1.8 to 1.54 ± 1.28 across generations one to six. Similarly, when SNPs were only used as markers in genomic selection, the genetic variance values ranged from 1.6 ± 1.8 to 1.5 ± 1.30 for generations one to six, respectively, as shown in Figure 7. When eQTLs and SNPs were combined, the analysis showed mean values ranging from 1.644 ± 1.8 to 1.45 ± 1.2 across generations one to six.

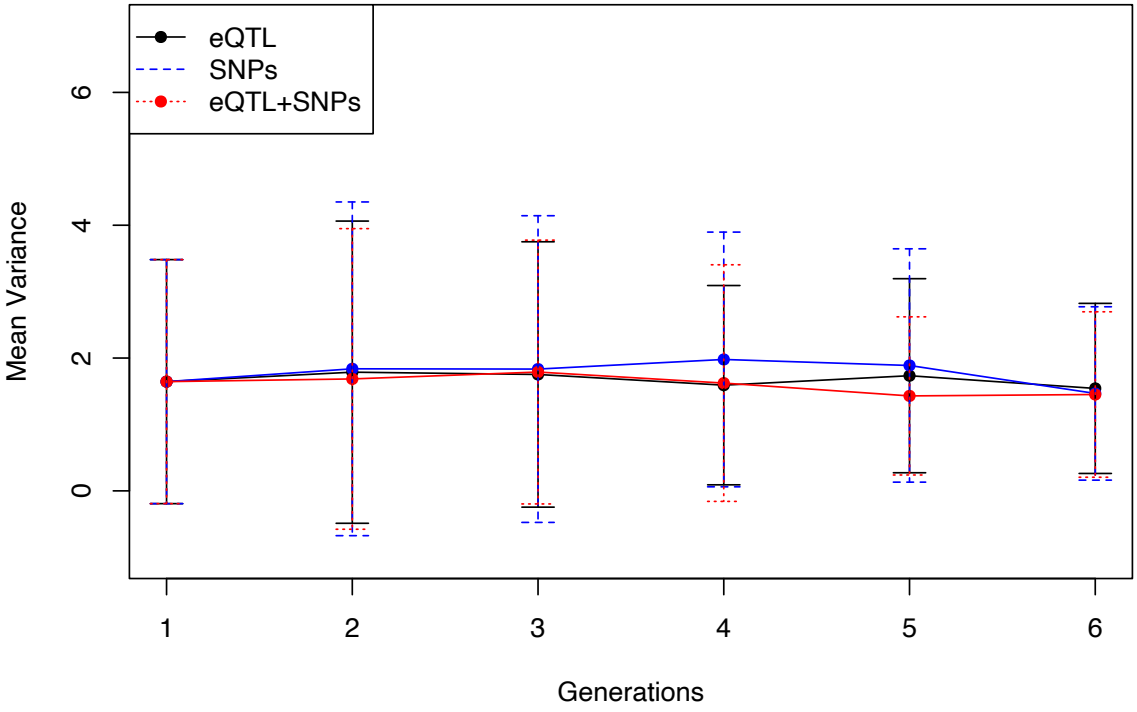

Figure 7: This graph is a representation of the genetic variance of different marker scenarios in genomic selection. The y-axis shows the mean-variance, and the x-axis represents the generations. The black, blue, and red lines show the trends by eQTL, SNPs, and eQTL+SNPs, respectively. The vertical lines in the data points represent the standard deviations.

The study did not find significant improvements in genomic prediction accuracy with the use of eQTLs compared to conventional marker arrays. However, small improvements in selection accuracy were observed.

Discussion

In this study, simulations were used to analyse the response to genomic selection using eQTLs, SNPs, and a combination of eQTLs and SNPs as markers in a pig breeding programme. The study was in two parts; the first focused on the analyses of eQTL mapping, and the second used eQTLs from the mapping results as markers in the genomic selection. The simulated SNP chips in this study were designed with some preselected eQTL variants (with the exception of the scenario with only SNP arrays) such that the density of variants remained similar across sets. The study did not find significant improvements in genomic prediction accuracy with the use of eQTLs compared to conventional marker arrays. Only small improvements in selection accuracy were observed. This section will discuss (1) the effectiveness of eQTL mapping in this study, (2) the selection accuracy that was achieved with the different marker scenarios, and (3) the implications of this study for animal breeding.

5.1 The effectiveness of eQTL mapping

In empirical studies, eQTL mapping often finds a major locus for gene expression that maps back to the same location ("local") of the major locus (i.e., the genetic variant influencing the expression of a particular gene is found near or within the gene's genomic region), as well as other "trans-acting" loci that tend to have smaller effects (Liu et al., 2020). As such, the present study simulated the genetic model in the same way. Here, 11 loci affect the gene expression traits, with one major locus. This major locus explained 0.5 of the additive genetic variance, while the remaining 10 had very low variance of 0.05. Aside from this, heritabilities were set for the gene expression traits based on the parameters of the simulated founder population, as discussed in Chapter 3 above. They were set at relatively high values of 0.5 for all 50 genes; this is because of the expectation that gene expression is a simpler trait than the downstream conventional quantitative trait.

The output of the eQTL mapping was analysed to find how well it performed in locating the causative variant. This was done by determining the distance of the lead

eQTLs from the causative variant expressed in centimorgans. Here, genetic distance in centimorgans has to do with recombination and not any physical positions in base pairs.

Recombination events during gamete production are less likely to occur between two genetic loci if their recombination frequency is low, that is, a low genetic distance. These loci, therefore, have a higher likelihood of being inherited jointly, which means that they typically get passed down as a connected unit from one generation to the next generation. From a biological perspective, 84 % of all the detected lead eQTLs are likely to be inherited with causative variants and passed on to the next generation of animals. However, as this was a simulated study, the eQTL mapping successfully detected variants near the causative variant. What this means is that eQTL mapping is able to detect variants more tightly linked to causative variants. Therefore, in the future, there is a potential use of eQTL mapping to find causative variants for functional studies. From the results, eQTL mapping can partially overcome the limitations of relying on neutral markers in incomplete linkage disequilibrium with causative genetic variants in genomic selections (Clark et al., 2020). For instance, breeding programmes can prioritise functional variants for inclusion in genomic selections to enhance the accuracy of genomic selections.

5.2 The selection accuracy

The observed variation in selection accuracies was large, most importantly with overlapping standard deviations in the three scenarios in the study. This makes it difficult to confidently state that one scenario (e.g., the marker set) has a significantly different accuracy. However, the eQTL marker sets have the highest means. This result is consistent with Loh and colleagues' report that showed improved prediction accuracy with eQTLs in genomic prediction. In their findings, it was observed that using gene expression data yielded better accuracy in predicting resistance than using genotype markers alone (Loh et al., 2011). The second-highest mean selection accuracy was observed in the combined eQTL and SNPs marker set. Theoretically, the inclusion of pre-selected variants associated with a trait in the SNP marker array should improve genomic prediction accuracy (Schulz-Streeck et al., 2011). This is because a combination of conventional SNP array and preselected eQTL markers on a chip is expected to provide a wide coverage of the genome and, therefore, hold more information than SNP marker

arrays alone. In biological contexts, this combination encompasses a range of variants (SNPs and functional variants) and, therefore, provides a broad coverage of the genetic structure relevant to the trait. These different variations linked to the trait may be in linkage disequilibrium with other nearby variants and, to some extent, aid in capturing information from neighbouring causative variants.

The mean selection accuracy observed when SNPs were only used as markers in the genomic selection was not different from that of the genomic selection with eQTL plus SNP markers set over a number of generations. To this end, the selection accuracies observed in all the scenarios were variable, and it remains complicated to distinguish them. For instance, particularly in marker sets with only eQTLs and eQTL plus SNP combinations, the accuracies were not constant throughout the breeding programme. Although eQTL mapping was successful, the lack of differences in marker scenarios may have reasons. To begin with, in this study, most of the markers were selected randomly, which means they were not guided by prior knowledge or specific features of the markers. This raises the concern of whether the randomly selected markers capture all relevant information. Specifically, whether the chosen markers effectively encompass the genetic variation and factors influencing the trait under consideration. Lastly, the lack of difference may be due to inadequate representation of genetic diversity and complexities of studied traits by chosen markers. This leads us to a critical analysis of the constraints that our marker selection strategy imposes, shedding light on issues that require careful attention when interpreting our results.

5.3 Limitations of this study

The flexibility of simulation scripts enables easy adaption to various parameter values, offering an outlet for detailed examinations of the resilience and sensitivity of the models. Even though time and computational resources are the only things standing in the way of such research, advances in technology and computer power will pave the way for increasingly comprehensive and intricate simulations.

It is important to acknowledge that the present study has several limitations that could affect the accuracy of the results. To address this, it is necessary to outline these limitations and suggest possible alternatives for the simulation. Therefore, the following points highlight some of the limitations of the study and provide alternative solutions that could be implemented to improve the study's validity and reliability.

1. The breeding programme structure contained relatively small populations compared to the databases commonly available to pig breeders. A larger population usually represents a broader genetic diversity. With more individuals, there is a higher likelihood of capturing a wider range of genetic variants contributing to the trait of interest. The larger sample size provides more information for estimating the effects of individual markers, leading to more precise predictions of genetic merit, and consequently, the differences between scenarios might be revealed with more data.
2. There is a possibility of using a whole genome sequence; however, this study used a 100K chip for the eQTL mapping. The use of a 100K chip might lead to missing relevant variants that could contribute to the trait. Whole-genome sequencing has the potential to identify a broader spectrum of genetic variations, including those located in regulatory regions and non-coding regions that may be crucial for trait regulation.
3. Chromatin sequence data were not taken into consideration in this investigation. To simulate genetic variants linked to chromatin states, the study might have benefited from modelling chromatin quantitative trait loci (cQTLs). To represent how chromatin sequence data affects the simulated characteristic, these cQTLs can be added to the simulation model. For example, by comparing models with solely genomic predictors, only epigenomic predictors, only eQTLs, and a combination of the predictors, the study may have investigated the relative contributions of genetic markers, chromatin sequence data, and eQTLs.
4. Lastly, the current study used a simple RR-BLUP for prediction. RR-BLUP is a linear model that assumes a genetic architecture where the effects of genes are additive in nature. However, there is nothing linear in biological terms. The assumption of the additivity of gene effects may not completely

explain the complexities of genetic interactions in the real world. Therefore, for a more comprehensive understanding and improved predictive accuracy in real-world scenarios, it is necessary to explore advanced Bayesian models that can capture nonlinearity and intricate genetic relationships.

In addition to the points stated above, it is worth mentioning that the complexity of successful prediction phenotype may lie in how various factors are interconnected to produce it. For instance, Boyle and colleagues argue that a large fraction of all genes expressed in relevant tissues can affect a phenotype and that much of the trait variance is mediated through genes not directly involved in the trait in question (Boyle et al., 2017). However, this argument appears at odds with conventional ways of understanding the links from genotype to phenotype. The conventional way assumes a relatively direct molecular pathway from genotype to phenotype. Even so, Boyle and colleagues (2017) provide a clear demarcation of the variance distribution of phenotypes. Although this simulation study followed a similar path by looking at genomics regions that affect gene expression, which in turn affect some final phenotype, the variants captured were unable to fully explain the variabilities in the final trait.

5.4 Implications for breeding practices

The molecular mapping from genotype to phenotype remains unclear for the highly diffuse genetic basis of most economic traits. The results did not indicate a big impact of eQTLs on genomic selection. This can be related to the fact that, in this simulation, eQTL marker sets are less, and their contribution to the selection accuracy was rather small. As no significant impact in selection accuracy was observed with eQTLs, the genetic gain was also not influenced. However, compared to SNPs and eQTL plus SNP marker sets, eQTL was better for a majority of the generations in the simulation. It is worth mentioning that selection accuracy is directly related to genetic gain, and so, therefore, methods that

lead to better estimation of breeding values will lead to higher genetic gain in the population.

Therefore, considering the potential of eQTLs' effect on the estimation of breeding values can positively contribute to a more accurate selection of pigs. For instance, in a breeding company, there are a number of considerations for implementing eQTL mapping in large pig breeding programmes. The core of this lies in understanding the biology of eQTLs and how they influence gene expression. The industry could work with genomics and bioinformatics specialists to build a solid body of knowledge about the function of eQTLs in pig genetics. Integrating eQTL mapping data into currently used genomic prediction models such as Bayesian models is the next stage. To further increase the precision of genomic predictions, the industry could investigate ways to efficiently combine data from eQTLs with those of conventional genetic markers.

In the pig population, the industry could perform eQTL mapping on a subset of the pig population and evaluate the effect on the accuracy of genomic predictions. Verify the outcomes using conventional genomic prediction methods and established breeding values. If the pilot experiments yield encouraging results, the industry could think about expanding the use of eQTL-informed genomic selection. Using the combined data from eQTLs and genetic markers, update breeding values and evaluate the results in terms of genetic gain and selection accuracy.

It is interesting to consider the potential uses of eQTL mapping data in the pig breeding industry beyond genomic prediction. Functional annotation of pig genomes, discovery of novel biomarkers, and understanding of gene regulatory networks are all promising areas that could benefit from this technology.

5.5 Future research

In future research, there is potential to enhance the results by exploring the simulation settings, such as the number of gene expression traits. Many other combinations of variables, such as the number of pigs, SNPs, and genes, can be experimented with by researchers. Additionally, future research attempts may dive into combining real-world eQTL data to enhance the realism of simulations, bringing the virtual models closer to the intricacies observed in biological systems.

In such studies, researchers should consider models such as the ominigenic model (Boyle et al., 2017); this will ensure that genetic contributions to complex traits, which can be partitioned into direct effects from core genes and indirect effects from peripheral genes acting in trans, are captured.

Finally, future research will benefit from large computational resources and time, as this is a computational-intensive task.

Conclusion

Simulations are useful tools for examining different situations, pinpointing areas where knowledge is lacking, and gaining a better understanding of the important factors within a model system. However, the accuracy of the inputs is crucial for simulations to be effective, and they can be difficult to carry out for complex systems. In a study that only considered one productive trait, a simplified model was used to identify gaps using eQTL in genomic selection.

The lack of a big impact of eQTL markers on selection, despite a successful eQTL mapping, could raise arguments against its implementation in pig breeding programmes. The beginnings of big things start small, and although there was no big impact, a number of factors could have been involved in the simulations or done differently. Therefore, an expansion of this study should consider a better combination of the marker densities, statistical models, number of genes, and population size, which are needed to allow drawing more assertive conclusions regarding the impact of eQTLs in genomic predictions in pigs.

References

Alexandratos, N., & Bruinsma, J. (2012). World agriculture towards 2030/2050: the 2012 revision.

Ballester, M., Ramayo-Caldas, Y., Revilla, M., Corominas, J., Castelló, A., Estellé, J., ... & Folch, J. M. (2017). Integration of liver gene co-expression networks and eGWAs analyses highlighted candidate regulators implicated in lipid metabolism in pigs. Scientific Reports, 7(1), 46539.

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., ... & Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome research, 24(1), 14-24.

Blair, H. T., & Pollak, E. J. (1984). Estimation of genetic trend in selected population with and without the use of control population. Journal of Animal Science, 58(4), 878-886.

Boyle, E. A., Li, Y. I., & Pritchard, J. K. (2017). An expanded view of complex traits: from polygenic to omnigenic. Cell, 169(7), 1177-1186.

Chen, G. K., Marjoram, P., & Wall, J. D. (2009). Fast and flexible simulation of DNA sequence data. Genome research, 19(1), 136-142.

Clark, E. L., Archibald, A. L., Daetwyler, H. D., Groenen, M. A., Harrison, P. W., Houston, R. D., ... & Giuffra, E. (2020). From FAANG to fork: application of highly annotated genomes to improve farmed animal production. Genome Biology, 21, 1-9.

Criado-Mesas, L., Ballester, M., Crespo-Piazuelo, D., Castelló, A., Fernández, A. I., & Folch, J. M. (2020). Identification of eQTLs associated with lipid metabolism in Longissimus dorsi muscle of pigs with different genetic backgrounds. Scientific Reports, 10(1), 9845.

Demidenko, E. (2013). Mixed models: theory and applications with R. John Wiley & Sons.

Edwards, S. L., Beesley, J., French, J. D., & Dunning, A. M. (2013). Beyond GWASs: illuminating the dark road from association to function. The American Journal of Human Genetics, 93(5), 779-797.

Endelman, J. B. (2023). Fully efficient, two-stage analysis of multi-environment trials with directional dominance and multi-trait genomic selection. Theoretical and Applied Genetics, 136(4), 65.

Fernando, R. L., & Grossman, M. (1989). Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution, 21(4), 467-477.

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: an R package for breeding program simulations. G3, 11(2), jkaa017.

Gjedrem, T., & Baranski, M. (2010). Selective breeding in aquaculture: an introduction (Vol. 10). Springer Science & Business Media.

González-Reymúndez, A., de Los Campos, G., Gutiérrez, L., Lunt, S. Y., & Vazquez, A. I. (2017). Prediction of years of life after diagnosis of breast cancer using omics and omic-by-treatment interactions. European Journal of Human Genetics, 25(5), 538-544.

Guo, Z., Magwire, M. M., Basten, C. J., Xu, Z., & Wang, D. (2016). Evaluation of the utility of gene expression and metabolic information for genomic prediction in maize. Theoretical and applied genetics, 129, 2413-2427.

Henderson, C. R. (1976). A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics, 69-83.
Hill, W. G. (1972). Effective size of populations with overlapping generations. Theoretical population biology, 3(3), 278-289.
Kearsey, M. J., & Pooni, H. S. C. N. (2020). Genetical analysis of quantitative traits. Garland Science.
Knol, E. F., Nielsen, B., & Knap, P. W. (2016). Genomic selection in commercial pig breeding. Animal Frontiers, 6(1), 15-22.
Kruuk, L. E., Charmantier, A., & Garant, D. (2014). The study of quantitative genetics in wild populations. Quantitative genetics in the wild, 1-15.
Larzul, C. (2021). How to improve meat quality and welfare in entire male pigs by genetics. Animals, 11(3), 699.
Leeson, G. W. (2018). The growth, ageing and urbanisation of our world. Journal of Population Ageing, 11, 107-115.
Li, Z., Gao, N., Martini, J. W., & Simianer, H. (2019). Integrating gene expression data into genomic prediction. Frontiers in genetics, 10, 126.
Liu, Y., Liu, X., Zheng, Z., Ma, T., Liu, Y., Long, H., ... & Xu, X. (2020). Genome-wide analysis of expression QTL (eQTL) and allele-specific expression (ASE) in pig muscle identifies candidate genes for meat quality traits. Genetics Selection Evolution, 52, 1-11.
Loh, P. R., Tucker, G., & Berger, B. (2011). Phenotype prediction using regularized regression on genetic data in the DREAM5 Systems Genetics B Challenge. PloS one, 6(12), e29095.
Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics, 10(8), 565-577.
Meuwissen, T. H., Hayes, B. J., & Goddard, M. (2001). Prediction of total genetic value using genome-wide dense marker maps. genetics, 157(4), 1819-1829.
Meuwissen, T., & Goddard, M. (2010). Accurate prediction of genetic values for complex traits by whole-genome resequencing. Genetics, 185(2), 623-631.
Meuwissen, T., Hayes, B., & Goddard, M. (2016). Genomic selection: A paradigm shift in animal breeding. Animal frontiers, 6(1), 6-14.
Nonneman, D. J., & Lents, C. A. (2023). Functional genomics of reproduction in pigs: Are we there yet?. Molecular Reproduction and Development, 90(7), 436-444.
O'Geen, H., Echipare, L., & Farnham, P. J. (2011). Using ChIP-seq technology to generate high-resolution profiles of histone modifications. Epigenetics Protocols, 265-286.
Pan, W., Wu, C., Su, Z., Duan, Z., Li, L., Mi, F., & Li, C. (2017). Genetic polymorphisms of non-coding RNAs associated with increased head and neck cancer susceptibility: a systematic review and meta-analysis. Oncotarget, 8(37), 62508.
Pérez-Enciso, M., Rincón, J. C., & Legarra, A. (2015). Sequence-vs. chip-assisted genomic selection: accurate biological information is advised. Genetics Selection Evolution, 47(1), 1-14.
Pevsner, J. (2015). Bioinformatics and functional genomics. John Wiley & Sons.
Plomin, R., & Deary, I. J. (2015). Genetics and intelligence differences: five special findings. Molecular psychiatry, 20(1), 98-108.
Pohlman, J. T., & Leitner, D. W. (2003). A comparison of ordinary least squares and logistic regression.

Poland, J., & Rutkoski, J. (2016). Advances and challenges in genomic selection for disease resistance. Annual review of phytopathology, 54, 79-98.
Quaas, R. L. (1988). Additive genetic model with groups and relationships. Journal of Dairy Science, 71(5), 1338-1345.
Ritchie, H., & Roser, M. (2019). Gender ratio. Our world in data.
Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. Nature Reviews Genetics, 16(2), 85-97.
Ros-Freixedes, R., Johnsson, M., Whalen, A., Chen, C. Y., Valente, B. D., Herring, W. O., ... & Hickey, J. M. (2022). Genomic prediction with whole-genome sequence data in intensely selected pig lines. Genetics Selection Evolution, 54(1), 1-18.
Schulz-Streeck, T., Ogutu, J. O., & Piepho, H. P. (2011, December). Pre-selection of markers for genomic selection. In *BMC proceedings* (Vol. 5, No. 3, pp. 1-4). BioMed Central.
Soller, M., & Beckmann, J. S. (1983). Genetic polymorphism in varietal identification and genetic improvement. Theoretical and Applied Genetics, 67, 25-33.
Sorensen, D. A., & Kennedy, B. W. (1984). Estimation of response to selection using least-squares and mixed model methodology. Journal of Animal Science, 58(5), 1097-1106.
Tang, K. L., Caffrey, N. P., Nóbrega, D. B., Cork, S. C., Ronksley, P. E., Barkema, H. W., ... & Ghali, W. A. (2017). Restricting the use of antibiotics in food-producing animals and its associations with antibiotic resistance in food-producing animals and human beings: a systematic review and meta-analysis. The Lancet Planetary Health, 1(8), e316-e327.
Teng, J., Gao, Y., Yin, H., Bai, Z., Liu, S., Zeng, H., ... & Fang, L. (2024). A compendium of genetic regulatory effects across pig tissues. Nature Genetics, 1-12.
Tortereau, F., Servin, B., Frantz, L., Megens, H. J., Milan, D., Rohrer, G., ... & Groenen, M. A. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. BMC genomics, 13(1), 1-12.
Visscher, P. M., & Goddard, M. E. (2015). A general unified framework to assess the sampling variance of heritability estimates using pedigree or marker-based relationships. Genetics, 199(1), 223-232.
Warr, A., Affara, N., Aken, B., Beiki, H., Bickhart, D. M., Billis, K., ... & Archibald, A. L. (2020). An improved pig reference genome sequence to enable pig genetics and genomics research. Gigascience, 9(6), giaa051.
Xu, Y., Liu, X., Fu, J., Wang, H., Wang, J., Huang, C., ... & Zhang, A. (2020). Enhancing genetic gain through genomic selection: from livestock to plants. Plant Communications, 1(1).
Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., ... & Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature genetics, 38(2), 203-208.
Zak, L. J., Gaustad, A. H., Bolarin, A., Broekhuijse, M. L., Walling, G. A., & Knol, E. F. (2017). Genetic control of complex traits, with a focus on reproduction in pigs. Molecular Reproduction and Development, 84(9), 1004-1011.

Zhang, H., Yin, L., Wang, M., Yuan, X., & Liu, X. (2019). Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Frontiers in genetics, 10, 189.

Popular science summary

This study aimed to improve the accuracy of genomic selection for pig breeding by addressing the challenge of accurately predicting highly polygenic traits using genome-wide markers. Although genomic selection is popular, the current accuracy limitations are due, in part, to relying markers not specifically selected for their functional relevance. In this case the markers might not capture the underlying genetic variations responsible for the traits we aim to predict.

To address this, the study explored genes and their expressions within genomic selection, specifically using markers that affect gene expression levels. The study simulated a breeding program targeting a single trait influenced by 50 genes to investigate the circumstances under which eQTL markers could enhance selection accuracy in pig breeding. I simulated a pig breeding program in a computer, and three genomic selection scenarios were explored. Two of the scenarios used variants from the markers from gene expression, while the third used simulated SNP markers. These scenarios aimed to evaluate the impact of markers on selection accuracies.

The results did not reveal a substantial impact of eQTL markers on genomic selection, contrary to expectations. The study's findings underscored certain limitations and emphasized the necessity for further investigation into the role of gene expression markers in optimising genomic selection strategies.

Acknowledgements

I want to express my deepest gratitude to God Almighty for providing me with the grace to complete this work successfully. I also want to thank my supervisors, Martin Johnsson and Armin Schmitt, for their immense intellectual support and guidance throughout the study.

I am also extremely grateful to my mother and sister, Mary and Christiana, respectively, for their immeasurable moral support and encouragement that helped me sail through the study smoothly. I also want to express my deepest gratitude to my beloved, Deborah. Deborah, your patience, love, and positive energy have made the challenges more manageable and the successes more joyful.

Finally, I thank the EMABG scholarship board for funding my studies. A big thank you to all my colleagues in the EMABG program for their contributions and support. May God richly bless you all.

Publishing and archiving

Approved students' theses at SLU are published electronically. As a student, you have the copyright to your own work and need to approve the electronic publishing. If you check the box for **YES**, the full text (pdf file) and metadata will be visible and searchable online. If you check the box for **NO**, only the metadata and the abstract will be visible and searchable online. Nevertheless, when the document is uploaded it will still be archived as a digital file. If you are more than one author, the checked box will be applied to all authors. You will find a link to SLU's publishing agreement here:

- https://libanswers.slu.se/en/faq/228318.

☒ YES, I/we hereby give permission to publish the present thesis in accordance with the SLU agreement regarding the transfer of the right to publish a work.

☐ NO, I/we do not give permission to publish the present work. The work will still be archived and its metadata and abstract will be visible and searchable.