Master Thesis

The effects of SNP-density and sample size on G-REML estimates of genetic and genic variance.

Tapiwanashe Magwaba

Submitted to
Department of Animal and Aquacultural Sciences
Norwegian University of Life Sciences

In partial fulfilment of the requirements for
European Master Animal Breeding and Genetics

August 2021

Supervisors:

Prof. Dr. Theodorus H.E Meuwissen

Prof. Dr. Henner Simianer

Dr. Jørgen Ødegård

Table of contents

List of figures

List of tables

Abstract

This simulation study investigates the effects of sample size, assortative mating, SNP density, and selection on realized genetic variance (σ_g^2), estimated genetic variance ($\sigma_{\hat{g}}^2$) and genic variance (σ_a^2). Sample size had three levels: n = 300, n = 1500 and n = 3000. SNP density also had three levels: 1K, 50K, and 100K. For the assortative mating scenarios, negative and positive mating were considered. Positive assortative mating (PAM) increased variance whilst negative assortative mating (NAM) and selection reduced the variance. Sample size and SNP density maintained the simulated variances. Both assortative mating methods and selection had an effect on G-REML estimates whilst sample size. SNP density had an effect on estimates genetic variance but not on realised and genic variance.

The realised genetic variance remained unaltered with variations in sample size and SNP density. Estimated genetic variance in these scenarios represents both realised and genic genetic variance hence the actual variance between individuals in a population. Assortative mating and selection, however, alter the realized variance. Subsequently, the estimated genetic variance for these factors is biased and may not represent the realised and genic, and hence the actual population variance.

Keywords: G-REML, genetic variance, genic variance, sample size, SNP density, assortative mating, selection.

1.0 Introduction

The goal of breeding is to improve every successive generation (Atsbeha et al., 2015). This improvement is measured as the phenotypic difference between parent and offspring generation given by the equation $R = h^2S$ [1] (Lush, 1943) where R is the response to selection, h^2 is heritability and S is selection differential. Heritability is therefore important for predicting genetic gain in subsequent generations resulting from the mating of selected parents. It is estimated from variance components, genetic, σ_g^2 and environmental, σ_e^2 using the formula $\hat{h}^2 = \hat{\sigma}_g^2/\left(\hat{\sigma}_g^2 + \hat{\sigma}_e^2\right)$[2].

Estimation of heritability was first done using the path analysis introduced by Wright, (1931). Later, another method based on analysis of variance (ANOVA) which used relative intra-class correlation was developed by Fisher (Visscher & Walsh, 2019). The third method uses a linear model to estimate heritability from pedigree which includes twin studies (Yang et al., 2017). With the advent of large-scale genotyping, efforts to estimates genomic-based heritability have been suggested. These include estimation of variance explained by SNPs discovered in genome-wide association studies (GWAS) termed h_{GWA}^2 and also estimation of variance caused by the entire set of SNP of a genotyping array termed h_{SNP}^2 (de los Campos et al., 2015; Rawlik et al., 2020) among others. Most of the methods are criticised for having inherent flaws (Rawlik et al., 2020; Yang et al., 2017) except for offspring-parent regression method. This sets an obvious demand for understanding variance components which are used for estimating this parameter.

Variance components are often estimated with restricted maximum likelihood (REML) using an animal model where a relationship structure is assumed for the additive genetic animal effects. When genomic data is used, the method is called the genomic restricted maximum likelihood estimation (G-REML). With this method, genomic relationship matrices (GRMs) are used to estimate variance components, fitting REML to estimate the variance explained by all SNPs (Rawlik et al., 2020; Yang et al., 2010).

Efforts have been made to better understand variance components, particularly, genetic variance ($\hat{\sigma}_g^2$). Previous studies focused on its dissection into simpler components (Clo et al., 2020; Lande & Porcher, 2015; Lynch, 2018).

It might be interesting to see how these parameters vary with different scenarios. For quantitative traits, each SNP has a small effect on the phenotype. In human height, for example, some variants have a single effect of 3 mm (Rotwein, 2020).

Previous studies have shown that sample size is inversely proportional to the markers to individuals (M/N) ratio leading to contribution of LD structure to $\sigma_{\hat{g}}^2$. And that with a small sample size, the $\hat{\sigma}_g^2$ is closer to the σ_g^2 due to strong LD contribution and that σ_a^2 is under-estimated or over-estimated when LD contribution is negative or positive respectively. This biasness affects heritability estimates (Rawlik et al., 2020).

The main aim of this study is to investigate the effect of sample size, SNP density, selection, and mating schemes on genetic and σ_a^2 estimates based on genomic estimates of relationships.

2.0 Literature review

2.1 Additive genetic and genic variance

Genotype and environment determine the phenotype of an individual. This can be represented in the form $V_P = V_G + V_E$ [3] where V_P is the phenotypic variance, V_G is the genetic variance and V_E is the environmental variance. The genetic variance can be dissected into three genetic components, $V_G = V_A + V_D + V_I$ [4], V_A being the additive variance, V_D is dominance variance and V_I is the epistatic variance (Huang & Mackay, 2016). One of the main foci in breeding is to single out the V_A from the total genetic components contributing to phenotypic variance (Kolstad, 2005). Additive genetic variance is part of the total genetic variance which occurs due to the aggregate effect of many genes on a quantitative trait with each gene contributing a small effect. For SNPs with effects, the inheritance of a particular SNP causes a deviation from the mean phenotype. This deviation can be used to predict phenotype changes resulting from allelic substitution (Singh & Singh, 2020). The essentiality of V_A is that it is used to calculate heritability (h^2) a key parameter for the determination of genetic gain in response to selection (Huang & Mackay, 2016; Lush, 1943).

Rasch & Mašata (2006) outlined some methods of variance components estimation methods as ANOVA, MINQUE (Rao, 1971), MIVQUE, and REML (Anderson and Bancroft, 1952). With the advent of SNP data, the genomic restricted maximum likelihood (GREML) method uses GRMs to estimates variance components, fitted using REML became the preferred method (Rawlik et al., 2020). Efforts have been made in the past to better understand variance components with a focus on genetic variance. The focus was on its dissection into individual components. Lande and Porcher (2015) showed that genetic variance can be represented in the form $\mathrm{G} = \mathrm{V} + \mathrm{C}$ [5]. G is a variance-covariance matrix which is the total genetic variance, V is a matrix whose diagonals give the genic variance (σ_a^2). Walsh and Lynch (2018) had a similar presentation of total additive variance, $\sigma_A^2 = \sigma_a^2 + d$ [6] . σ_A^2 being the total additive variance, σ_a^2 is the genic variance which is determined by the allele frequencies and when there is no LD, it takes the same value as the additive variance. It is the values of genetic variance expected under random mating. The last part of the equation, d is the LD generated disequilibrium contribution. Other authors also concur with the premise (Clo, Ronfort, and Awad, 2020; Rawlik et al., 2020). Genic variance was first mentioned by Fredeen & Jonsson, (1957) when they studied genetic parameters associated with feed efficiency in pigs. Lande, (1976) envisioned dissecting genetic variance into building components but did not explicitly mention σ_a^2.

Burt, (2015) advanced the premise share by other authors that recent advances in genetics show that the estimation of heritability was flawed from the beginning. Classical heritability has been criticized as being overestimated, whilst GWAS heritability explains a small proportion of the actual heritability. Renewed efforts to revisit σ_a^2 are necessitated by the need to understand better the components which are used in the estimation of heritability.

With an increase in generational time under random mating, recombination causes LD decay. This LD decay is dependent on the measure of mutation rate (θ) and can be described by the equation; $D_t = D_o(1-c)^t$[7] (Falconer & Mackay, 1996) where D is disequilibrium, c is the recombination frequency and t is the generational time. This affects the differences between genetic and genic variances. The σ_a^2 however, remains unchanged over generational time under the infinitesimal model (Walsh and Lynch, 2018). As generational time increases under the assumption of no migration nor mutation, D decreases asymptotically under the counter influence of drift and mutation. In a finite population, LD and σ_a^2 decreases at a rate of $\frac{1}{(2N_e)}$, N_e being effective population size. This decrease in LD may be due to fixation of alleles by drift and shuffling of markers and QTLs by recombination. For traits influenced by many loci, small and cumulative effects of LD influence genetic variance value. This LD-mediated change in σ_A^2 can be predicted when the value of genic variance is known.

2.2 SNP density

Single nucleotide polymorphisms (SNPs) are a variation of DNA base pairs that occurs when a single nucleotide is altered at a specific genome position (Koopaee & Koshkoiyeh, 2014; Zhang et al., 2020). SNPs genotypes provide genomic data required for genomic selection (GS) of superior animals for breeding and, also genetic maps for genetic variation studies of quantitative traits (Meuwissen et al., 2001).

SNPs have the highest density in the genome of all the markers, they occur in coding and non-coding region, they are more stable and the integration of genomic data is easy (Xia et al., 2019; Zhang et al., 2020). Markers can be classified as low or high density based on the extent of LD, genome size, and species of interest (Xia et al., 2019). Low-density markers are not effective for QTL detection when LD is under rapid breakdown and for genomes of large size (Ibid.). the reasn for their inffectiveness ids that they are sparse and therefore may not be close enough to QTL. Dash et al., (2018) however, found the low-density chip to be efficient for genetic diversity studies in cattle as does high-density chips.

High-density SNP chips are useful for fine-mapping of QTLs and genome evolution studies (Meuwissen & Goddard, 2000; Tortereau et al., 2012). Porto-Neto et al., (2014) reported the discovery of association signals with high-density chips that were not discovered before when lower density was used. LD persistence across populations has been reported to improve with the density of SNPs (Bastiaansen et al., 2014). Whilst high-density chips are more effective in genome studies, their use is constrained by high cost, so low-density chips are used followed by phasing and imputation (Pryce et al., 2014). Imputation is the prediction of unknown genotypes using higher density chips for animals genotyped using a low-density chip (Bolormaa et al., 2015).

This research seeks to investigate and provide empirical evidence on the effect of SNP density has on $\sigma_{\hat{g}}^2$ and σ_a^2.

2.3 Selection, mating, and linkage disequilibrium

Linkage disequilibrium (LD) is defined as the non-random association of alleles at different loci than it would be by chance (Slatkin, 2016). Selection is choosing individuals that have characteristics of interest to be parents of the next generation with the hope that their offspring will inherit those desirable characteristics. Common forms of selection are directional, stabilizing, and disruptive.

LD can be measured as a deviation of gamete frequency from the expected by calculating a coefficient of linkage disequilibrium (D) (Lynch, 2018). Another method of calculating LD is the use of Pearson's LD correlation coefficient (r) which is squared to render all values positive. It is a measure of how independent any two loci are (Lin et al., 2012). The third method is the standardization method, where D is compared to its maximum (Guo, 1997).

Positive assortative mating (PAM) involves mating individuals with similar phenotypes for example, good to good. It alters the additive genetic variance mainly by generating LD and more genetic variance because like alleles, positive or negative, tend to be linked with each other. (Hayashi, 1998). Negative associative mating (NAM) on the other hand reduces realized genetic variance.

Random selection favours recombination and results in higher genetic variance than that produced from directional selection (Sánchez & Woolliams, 2004). The $\sigma_{\hat{g}}^2$ was by regressing offspring genetic values to those of their parents. The approach of Hayashi, (1998) based on an infinitesimal model deliberated on the effect of LD on selection. LD consideration in the selection process is based on the premise that the response to selection on one locus might affect the other locus if the two are in LD thereby influencing the response of haplotypes. Besides selection, genetic drift also creates LD among closely linked loci which in turn reduces the response to selection which McVean & Charlesworth, (2000) referred to as the "Hill–Robertson effect". This effect is weak when few loci are considered but very strong when many closely linked loci are considered. Also, reduction in population size increases LD due to loss of haplotypes and increase in genetic drift, so does inbreeding by augmenting the covariance between alleles at different loci (Slatkin, 2016). Non-random mating may affect genetic variance. Two forms of non-random mating are assortative mating where mating is based on phenotypic resemblance and inbreeding where mating individuals are more related to each other than the population average (Zhang et al., 2020).

(Lande, 1976) postulated that PAM reduces genetic variance since it mates extremes as compared to random mating. This seems to contradict later co-publication (Devaux & Lande, 2008) who suggested that assortative mating creates positive allelic and loci correlations which increase the genetic variance. They also advanced that at equilibrium, the genetic variance, σ_g^2, is bigger under assortative mating than that exhibited under random mating.

NAM generates negative LD, reduces heritability, and therefore reduces the response to selection (Lynch, 2018). In positive assortative mating, individuals with similar phenotypes are mated whilst negative assortative involves dissimilar individuals (Hayashi, 1998). PAM increases genetic variance by creating a positive correlation between pairs of loci (Bulmer, 1971). Disruptive selection was reported to reduce recombination under infinite population size but increase it under finite population (Sorensen & Hill, 1983). In a random mating scheme, breeding males and females are paired by random sampling (Nirea et al., 2012). Under this method, recombination among loci is proportional to LD decay (Sorensen & Hill, 1983).

Bulmer, (2001) reported that under the infinitesimal model, selection induces a temporary correlation between pairs of loci altering genetic variance which reverts when the selection ceases. In the absence of selection, inbreeding leads to a reduction in heterozygosity subsequently reducing the additive genetic covariance within families whilst increasing it among families (Wright, 1969; Crow and Kimura 1970).

The effects of mating, selection, and LD on G-REML estimates will be investigated.

2.4 Sample size

Resource scarcity limits the use of the entire population for an experiment. A proportion of the population considered as a representative sample is used for the experiment. This proportion is used to infer to the real population is called sample size (Faber & Fonseca, 2014). The sample size has to be decided at the beginning of the experiment to avoid extreme (low or high) sizes which can either compromise the results of the experiment (Faber & Fonseca, 2014) or leads to unwarranted experimental costs. The use of an appropriate sample size in an experiment allows the detection of a phenomenon if it does exist in the real population or to confirm its nonexistence if it is not discovered by the experiment (Ibid.).

A sample size reduces the probability of discovering a phenomenon that does exist in the population (error type II). A very large sample size, on the other hand, overestimate statistical differences is overestimated, waste time and resources. The power of the test can therefore be improved by increasing the sample size. Previous studies showed an underestimation of effective population size when a small sample size was used for the investigation (England et al., 2006). Nelson et al., (2015) reported a close to 5 times difference in the estimated population parameter, θ, by using a sample size of 11 000 humans compared to when 23 humans were used. For research like this one, there is a trade-off between the number of individuals and the number of loci per individual. Landguth et al., (2012) suggested there are many benefits in using more markers for relatively fewer individuals.

(Rawlik et al., 2020) reported that sample size affects the ratio of markers (M) to the number of individuals (N). As the sample size increases, the M/N ratio decreases which in turn leads to reduced contribution of the effective LD structure to genetic variance. They also mentioned that, with small sample size, the estimate of σ_g^2 is closer to the true σ_g^2 due to strong LD contribution and that σ_a^2 is under-estimate or over-estimated if LD contribution is negative or positive LD, respectively. This bias translates to heritability estimates (Ibid). Hong and

Park (2012)'s results however, very small sample size is sufficient to detect association when LD is high.

The main aim of this study is to investigate the effect of sample size, SNP density, selection, and mating schemes on genetic and σ_a^2 estimates based on genomic estimates of relationships.

3.0 Methods

3.1 Simulation of base populations and breeding schemes.

To investigate the effects of factors hypothesised on genetic and genic variance, scenarios were simulated in the software package "QTL and marker simulator" (QMSim) (Sargolzaei and Schenkel, 2013). Each scenario was simulated with 20 replications, 0.5 heritability, a phenotypic variance of 1.0, and a historical population of 600 individuals for 1200 generations. The base population had the same size as the historical population with 20 generations of breeding, a litter size of 10, and a fixed proportion of male progeny of 0.5.

The genome comprised of 1 chromosome of length 100 cM, 10 000 markers randomly distributed across the genome, 500 random QTLs sampled from a uniform distribution. The SNP and QTL per base pair mutation rate were 1e-7 per generation and a minor allele frequency (MAF) for LD calculation of 0.05 from the first generation was used. Mutations altered the alleles of the bi-allelic markers and QTL back and forth.
The effect of mating on genetic and genic and hence heritabilities were examined by simulating two mating designs: a negative assortative and a positive assortative. Mating individuals were selected at random. The effects of selection were also examined by simulating a selection design based on phenotype. All the males and all females were selected or culled based on their phenotype. The proportion of selected individuals was 0.2 and hence a selection intensity of 1.40.

To investigate the effect of sample size on our parameters of interest, three populations of varying sizes were used. The largest sample size (LSS) of the three had 3 000 individuals, the intermediate (ISS) had 1500 individuals and the smallest (SSS) had 300 individuals. The ISS was established by selecting every second individual from an ordered list of LSS whilst SSS was drawn from LSS by selecting every 10th individual. The LSS included only one individual from every family to reduce the family effect.

To investigate the effect of SNP density, three SNP densities, 1 000 markers, 50 000, and 100,000 marker chips were used. The actual number of SNPs however varied depending on the segregation of the loci.

3.2 Estimation of heritability

QMSim output data were extracted to further steps of the analysis. The marker data and QTL data files with animal identities and parental alleles were converted to SNP genotypes using the Julia software package. The phenotype data from QMSim were edited in the R software package to remove ungenotyped animals and other columns which were deemed not useful for the study. A column of fixed effect with a value of 1 in each row was added to the file as the overall mean which was estimated in the analysis.

The genomic relationship matrix (GRM) and GRM inverse were calculated using the WOMBAT software package (Meyer, 2007). The genomic relationship matrix (GRM) was calculated using the VaRaden1 method (VanRaden, 2008) and centred using allele frequencies. Error and genetic variances were also estimated by WOMBAT. To make the GRM positive definite, 0.01 was added to its diagonals and zeros were added to the first line of the GRM inverse.

3.3 Genetic and genic variances

The genetic variance was estimated in the WOMBAT software package. The true or realized genetic variance was calculated from the QTL allele effect and allele frequency as the product of the square of allele substitution effect and heterozygosity; $\sum[(e_{a1} - e_{a2})^2 * 2 * p_i(1 - p_j)]$ [8], where e_{a1} is the effect of the first allele and e_{a2} is the effect of the second allele, p_i is the freqiency of one allele and $(1 - p_j)$ is the frequency of an another allele.

4.0 Results.

4.1 Segregation of markers and QTLs

QMSim was used to simulate the effect of sample size (n= 300, n= 1500, and n= 3 000), mating (negative assortative and positive assortative), and selection. Three scenarios (1K, 50K, and 100K SNP chips) were simulated to examine the effect of SNP density. On average 55 % of the markers and 54 % QTLs were segregating. Pictorial representations of markers and QTLs are shown in **Figure 1** and **Figure 2,** respectively. **Table 1** shows the number of initial and segregating markers and QTLs from the QMSim output file.

	Initial		Segregating		SE	
Scenario	**Markers**	**QTLs**	**Markers**	**QTLs**	**Markers**	**QTLs**
ss3000	10000	500	5473	273	56	2
NAM	10000	500	5460	274	42	4
PAM	10000	500	5559	280	50	3
SNPd 100K	100 000	500	54699	273	570	4
Selection	10 000	500	5547	278	63	4

Table 1: Initial number of markers and QTLs at generation 0 and segregating markers at generation 20. The simulated scenarios are sample size (ss) with 300, 1500, and 3000 individuals. Mating scenarios were negative assortative mating (NAM), positive assortative mating (PAM). Other scenarios were SNP density (SNPd) and selection. Other scenarios not shown here were not simulated but were extracted as subsets from the simulated.

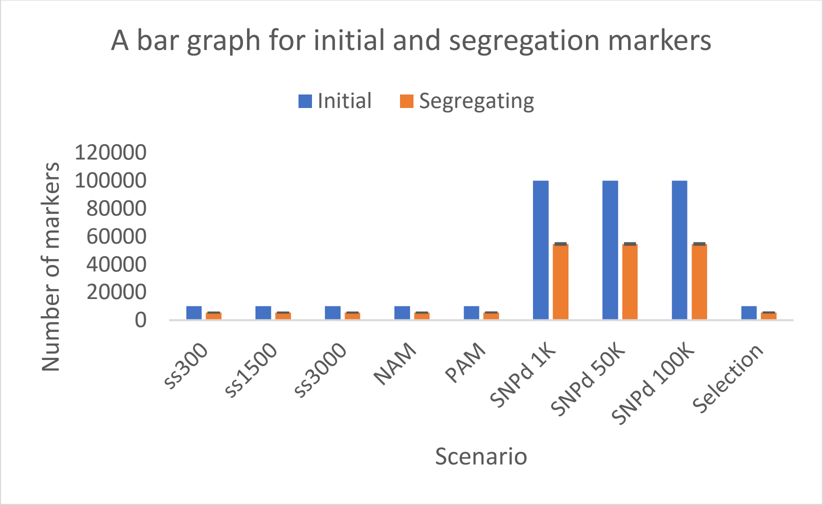
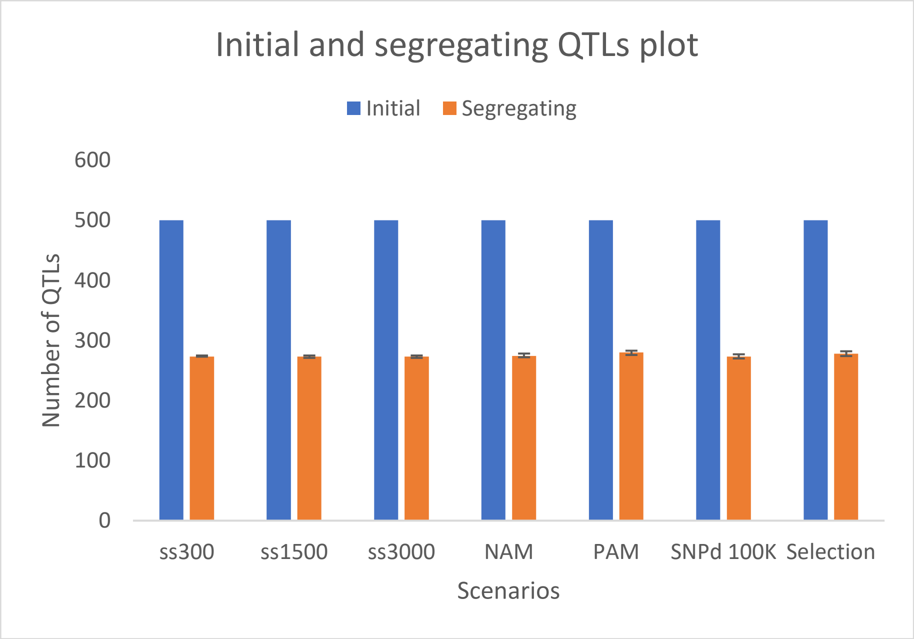


Figure 1: *Initial and segregating markers. Initial marker bars do not have error bars. Most scenarios had initial 10 000 markers and approximately half of them were segregated. Scenarios that sought to investigate the effect of SNP density had varied marker initial numbers, 1 000, 50 000, and 100 000.*

Figure 2:Initial and segregating QTLs

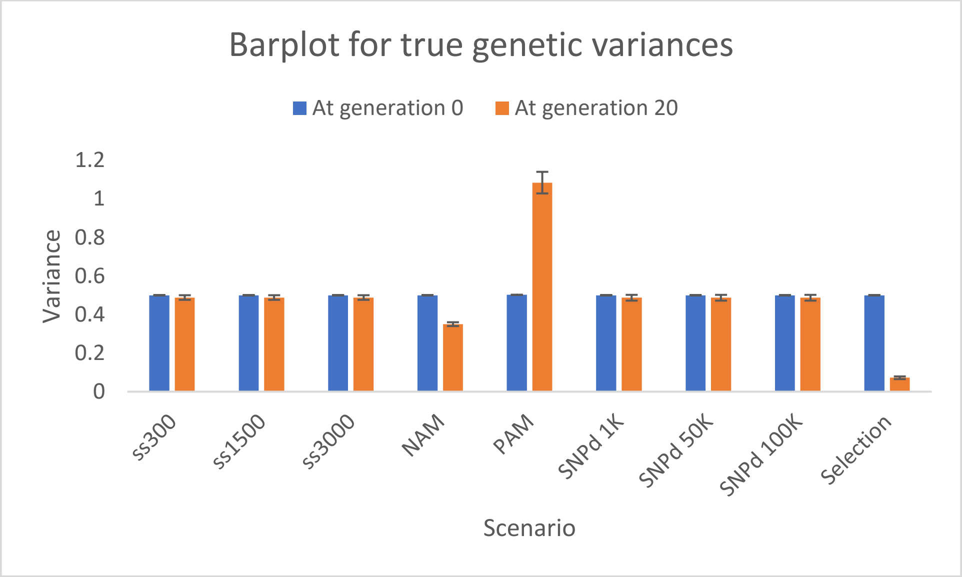


.

4.2 True genetic variance

Realised genetic variance (σ_g^2) which is the variance of true breeding value was calculated for generation 0 and generation 20 from the QMSim output file as the QTL variance. The σ_g^2 at generation 0 was 0.50 across all the scenarios and levels. The same value was maintained for sample size and SNP density after 20 generations, there was therefore little variation due to different levels of sample size and SNP density. Assortative mating and selection showed a deviation from this trend. The σ_g^2 for dropped from 0.50 in generation 0 to 0.35 and 0.07 in generation 20 for negative assortative mating (NAM) and selection respectively. PAM however, showed an increase from 0.50 to 1.08 between the generations. PAM had the highest value whilst selection had the least. Sample size and SNP density did not affect genetic variances. These results are shown in **Figure 3**.

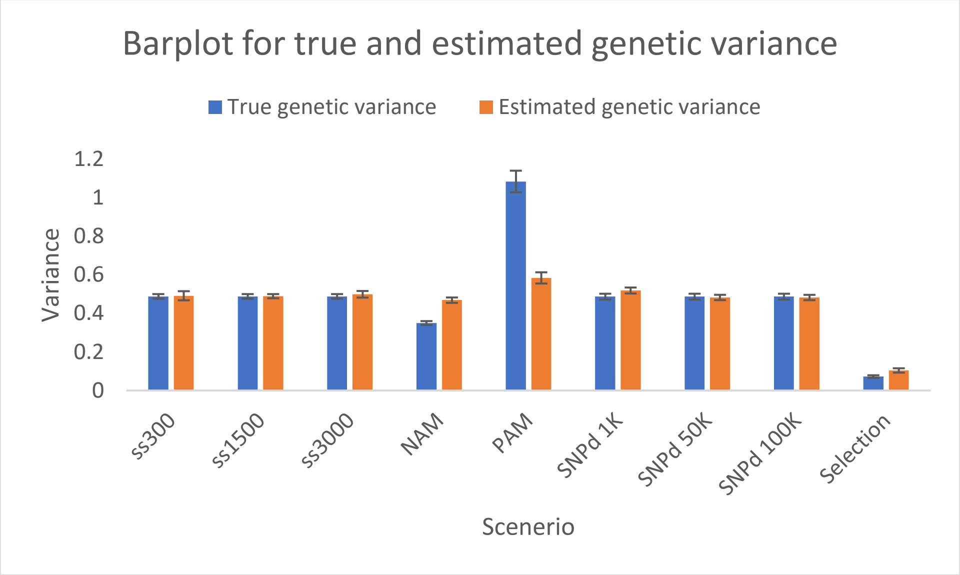
Figure 3: Realised genetic variance for generation 0 and generation 20.



4.3 Realised and estimated genetic variances

There was a small difference between σ_g^2 and $\sigma_{\hat{g}}^2$ for both sample size and SNP density. The difference was significant for assortative mating and selection scenarios. The $\sigma_{\hat{g}}^2$was overestimated for NAM and selection whilst it was underestimated for PAM. The results are shown in **Figure 4**.

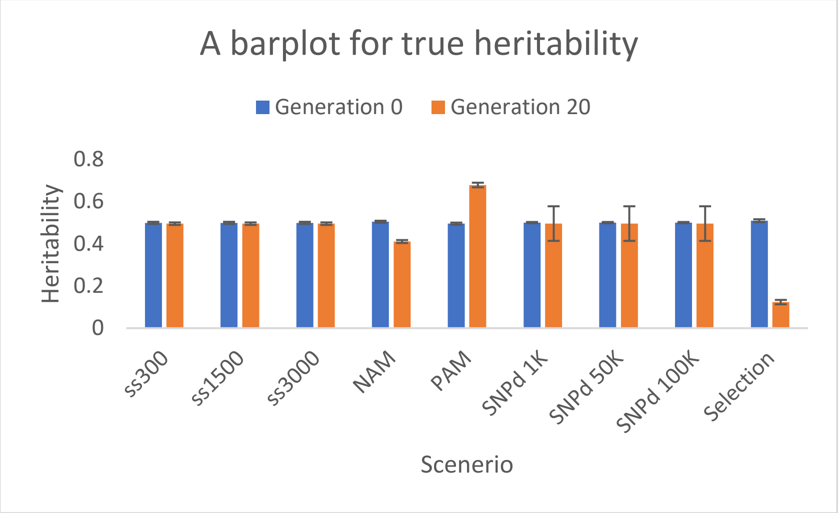
Figure 4: True and estimated genetic variance at generation 20.



4.4 True heritability

The true heritability (h_t^2) values were calculated for generations 0 and generation 20 from the QMSim output file as the ratio of phenotypic and QTL variances. He values were 0.68, 0.50, 0.41 and 0.12 for positive assortative mating, sample size and SNP density, negative assortative mating, and selection, respectively. The calculated value was 0.50 ± 0.005 across all the simulated scenarios for generation 0. Differences were noted for generation 20 for NAM (0.41±0.006), PAM (0.68±0.011) and selection (0.12±0.010), whilst sample size (0.50±0.006) and SNP density (0.50±0.082) showed slight differences across all levels. The THR for NAM and selection in generation 20 were 0.049 and 0.39 lower than at generation 0 respectively. PAM however had a higher value (+0.18) at generation 20 than at generation 0. **Figure 5** below shows the comparison of the two.

Figure 5: True heritability at generation zero and 20.



4.5 Estimated genetic variance ($\sigma^2_{\hat{g}}$)

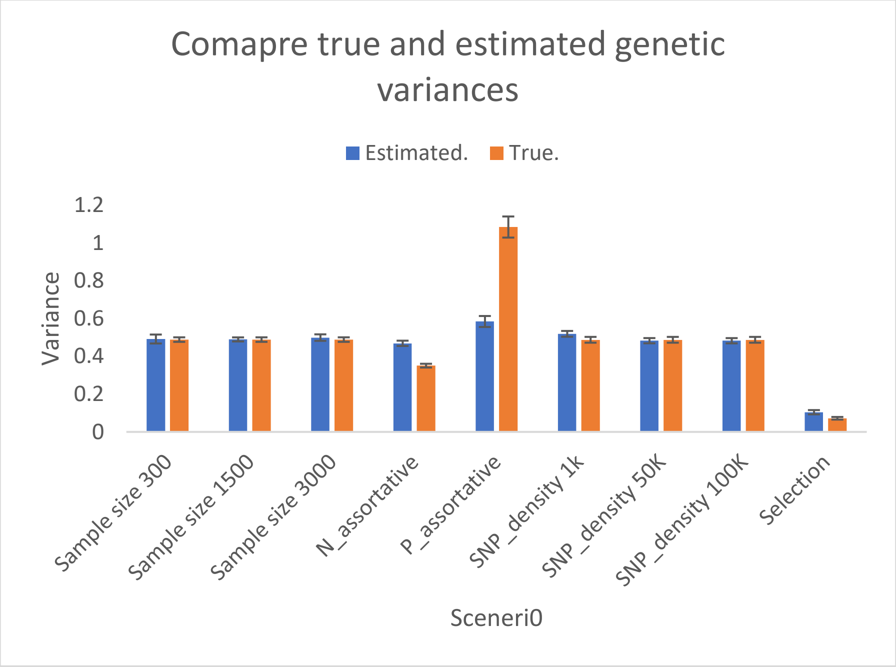
The estimated genetic variance ($\sigma^2_{\hat{g}}$) was computed by the WOMBAT software. There were no significant differences between the three levels of sample size, they were all close to 0.5 with a mean value of 0.49. SNP density however had a significant effect on $\sigma^2_{\hat{g}}$, estimates seemed to decrease as marker density increased. SNP density 1K had a significantly higher estimate (0.52±0.015) than 50K and 100K densities which had estimates of 0.483 (0.014) and 0.482 (0.014) respectively. PAM had a very large estimate whilst selection had the smallest estimate. Negative assortative mating, PAM, and selection had estimates of 0.47, 0.58, and 0.10 respectively.

4.6 True and estimated genetic variances.

Figure 6 shows a comparison plot of true genetic and estimated variances. The $\sigma^2_{\hat{g}}$ is equal to σ^2_g for sample size and SNP density. An over-estimation is observed for NAM and selection by 0.12 and 0.03 respectively whilst there is underestimation for the PAM scenario by 0.5.

Assortative mating and selection had an effect on both components. After 20 generations, σ^2_g mean value for sample size values (0.488 ± 0.0120) was slightly lower than that of $\sigma_{\hat{g}}^2$ (0.494 ± 0.0006) though the difference was not statistically different. For negative assortative mating, the genetic variance had an estimate of 0.47 for which was an overestimate relative to σ_g^2 of 0.35. Positive assortative mating, however, conferred under-estimation of genetic variance as with a value of 0.58±0.03 relative to σ_g^2 of 1.08±0.06. SNP density had an effect on $\sigma_{\hat{g}}^2$. A density of 1K had a genetic variance estimate of 0.52 which was significantly higher than 50K and 100K which had equal estimated values of 0.48.

Figure 6: True and estimated genetic variance after 20 generations.

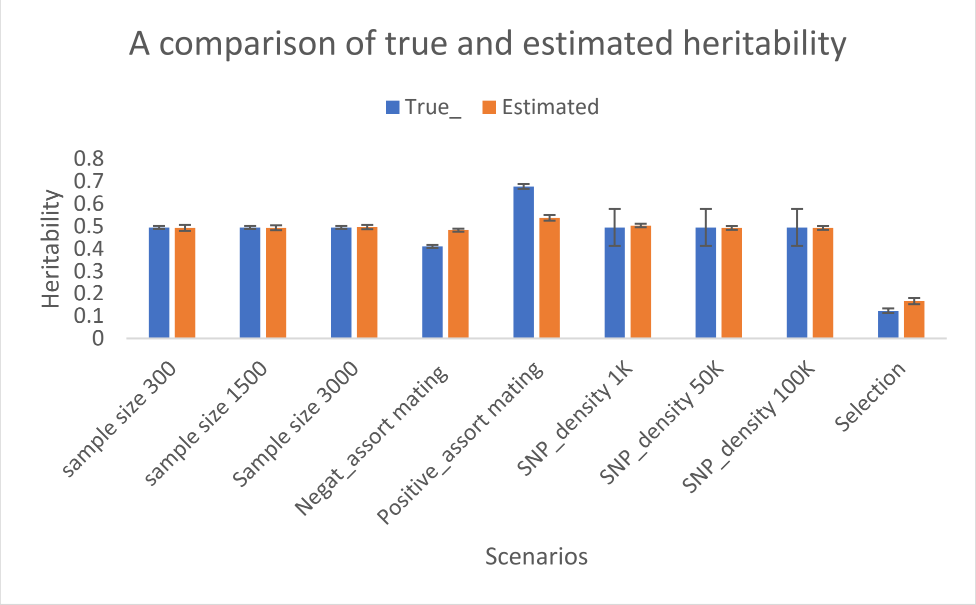


4.7 True and estimated heritabilities

Figure 7 is a bar plot for true and estimated heritability after 20 generations. Though not significantly different, there were slight differences between the three levels of sample size on true (0.495±0.006) and estimated heritability 0.490±0.011 so were SNP density levels with values 0.4955 ± 0.0820 and 0.50±0.008 for true and genetic heritabilities, respectively. Negative assortative mating, PAM, and selection did have an effect on true and estimated heritability. The model overestimated the heritability for NAM and selection by 0.07 and 0.04

respectively whilst it under-estimated the same for the PAM scenario by 0.14. Heritability estimate, 0.17±0.014, for selection was higher than the σ_g^2 , 0.12±0.010.

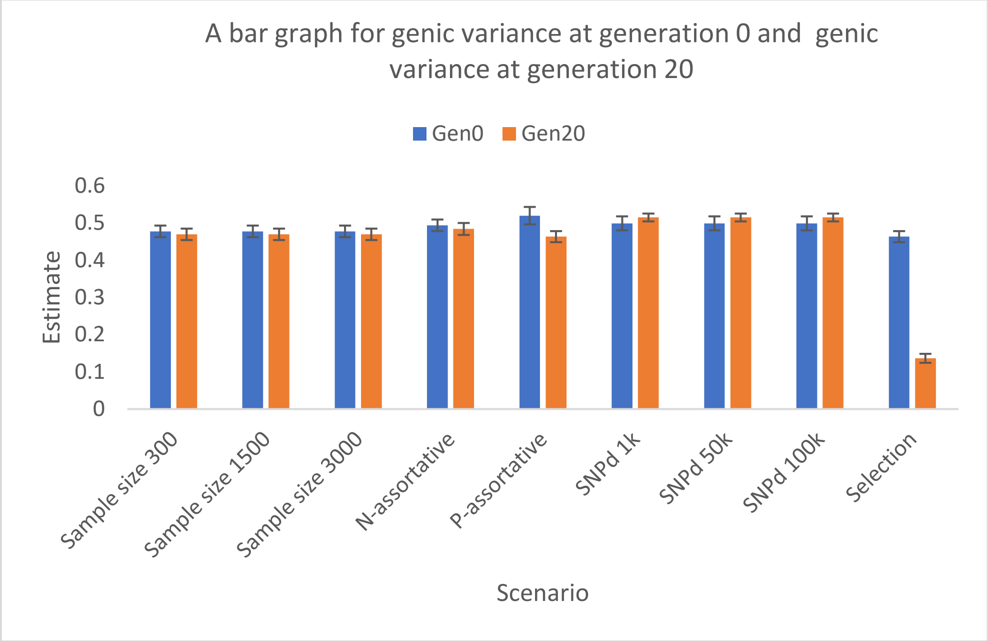
Figure 7: True and estimated heritability after 20 generations.



4.8 Genic variance (σ_a^2)

Figure 8 shows a plot for σ_a^2 at generation 0 and generation 20. There were no significant differences between σ_a^2 at generation 0and generation 20 for the three levels of sample size though the values were lower after 20 generations than that at generation 0. The values were equal for all levels; 0.48 ± 0.02 and 0.47 ± 0.02 at generation 0 and generation 20 respectively. For negative assortative mating, the σ_a^2 at generation 20 was lower but the difference was not significant. For positive assortative mating, however, σ_a^2 at generation 20 (0.46±0.023) was significantly lower than at generation 0 (0.52±0.015). There were no significant differences between the two variances for different levels of SNP density though σ_a^2 at generation 20 was higher than that at generation 0, the estimates were 0.51(0.019) and 0.50(0.011) respectively. There was a large decline in σ_a^2 for the selection scenario after 20 generations, from 0.46(0.012) at generation 0 to 0.14(0.015) at generation 20.

Figure 8: Genic variance for generation zero and generation 20.



4.9 Comparison of realised genetic variance, estimated genetic variance, and genic variance

PAM increases variance, negative assortative and selection reduce the variance. Sample size and SNP density maintained the simulated variances.

Sample size and SNP density had no effect on true genetic variance andσ_a^2. The $\sigma_a^2$was intermediate in between the estimate and the σ_g^2. The $\sigma_{\hat{g}}^2$ was closer to σ_g^2. NAM had an effect on G-REML estimates. The $\sigma_{\hat{g}}^2$ was overestimated and is between the σ_g^2 and σ_a^2 and, closer to the σ_a^2. PAM had an effect on G-REML estimates. The $\sigma_{\hat{g}}^2$ was under-estimated and lay in between the realised and $\sigma_a^2$and, was closer to the σ_a^2.

SNP density had no effect on σ_g^2 and σ_a^2 but $\sigma_{\hat{g}}^2$. For the 1k scenario, the genetic variance was overestimated, and the σ_a^2 was in between the estimate and σ_g^2. The estimate was closer to the σ_a^2 than it was to the σ_g^2. For higher densities, however, the genetic variance was underestimated and was in between the genic and σ_g^2. Selection had an effect on all the parameters under investigation. The genetic variance was overestimated and was in between the σ_g^2 and σ_a^2. The results are shown in **Figure 9**.

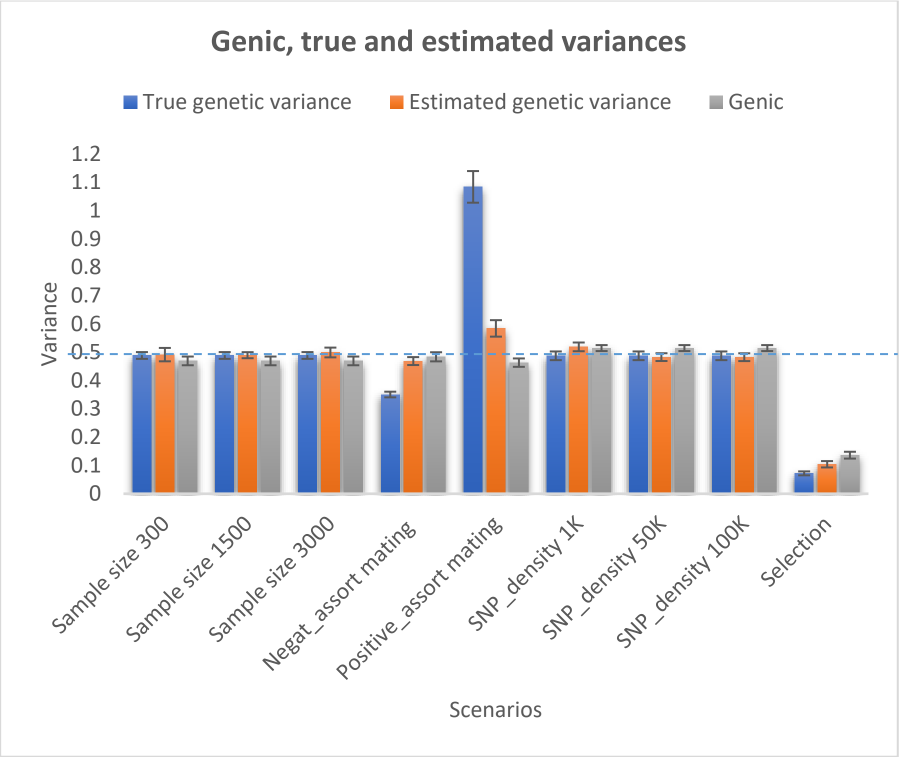


Figure 9: Plot for realised genetic variance, estimated genetic variance, and genic variance after 20 generations. The dotted line shows the simulated variance of 0.5.

For the sample size scenario, the $\sigma_{\hat{g}}^2$ was higher than both σ_g^2 and σ_a^2. The σ_a^2 had the lowest value. The difference was not however statistically significant. The mean values were 0.488, 0.493, and 0.469 for σ_g^2, $\sigma_{\hat{g}}^2$ and σ_a^2 respectively. SNP density scenario showed a non-significant difference between the three variant components. The estimates were in between the σ_a^2 and the σ_g^2 for SNPd 50K and 100K but highest for SNPd 1K. For negative assortative mating, the estimate was in between the other two components but closer to the σ_a^2 than it was to the σ_g^2. The values were 0.35, 0.47 and 0.48 for σ_g^2, $\sigma_{\hat{g}}^2$ and σ_a^2 respectively from the lowest to the highest. For positive assortative mating, however, the values are in reverse order, the σ_g^2 had the highest value, the $\sigma_{\hat{g}}^2$maintained its intermediate position whilst the σ_a^2 had the least value. The σ_g^2 and $\sigma_{\hat{g}}^2$ were highest for all the simulated scenarios. The values were 1.08, 0.58, and 0.46 for σ_g^2, $\sigma_{\hat{g}}^2$ and σ_a^2 respectively. Selection had the least component values of all the scenarios. $\sigma_{\hat{g}}^2$ was intermediate between σ_g^2 and σ_a^2. The values were 0.07, 0.10 and 0.1 for σ_g^2, $\sigma_{\hat{g}}^2$ and σ_a^2 respectively.

5.0 Discussion

Sample size did not have an effect on realized genetic variance, estimated genetic variance and genic variance. Hence there was a lack of effect of sample size on G-REML estimates of genetic variance. This result differs from Rawlik et al., (2020)'s prediction that the estimates diminish as the sample size increases because of the decreased ratio between the number of markers (M) and the number of individuals (N). The latter may be because Rawlik et al.'s data resembled human populations with a very weak family structure and the present results resemble animal breeding data with a strong family structure.

Positive assortative mating had an effect on G-REML estimates whilst negative assortative mating has little effect. The $\sigma_{\hat{g}}^2$ was in between the σ_g^2 and σ_a^2. PAM creates positive allelic and loci correlations which increase the genetic variance. These results concur with Devaux & Lande, (2008). This may be caused by the induced LD which then contributes negatively to $\sigma_{\hat{g}}^2$ leading to underestimation of the σ_a^2. Also, PAM will have even larger effect in the next generation since phenotypes are more likely to be influenced by genetics than the environment. Phenotypic variances under NAM are likely to be influenced more by environmental factors. Since alike individuals are mated in positive assortative mating, there is likely inbreeding with small effective size and hence reduction in variance within the alike mated individuals. On the other hand, the population is split into different strata (high and low) thereby variance increases between the, unlike mating groups. NAM will not give the same results.

Lower SNP density had a higher genetic variance estimate than higher SNP density. This is a deviation from the expected trend as an increase in SNP density increases the LD between QTLs and markers and therefore the $\sigma_{\hat{g}}^2$ is expected to increase. Ogawa et al., (2014) reported that $\sigma_{\hat{g}}^2$ increases with increasing SNP density, which may be expected since a smaller fraction of the genetic variance may be captured by a low-density SNP panel.

Selection showed the lowest genetic variance estimates. This loss of variance can be explained by the fact that selection tend to fix loci. The same trend was reported by Hayashi, (1998). Also, fixation of most QTLs leads to a reduction of $\sigma_{\hat{g}}^2$.

The σ_a^2 is maintained almost constant across scenarios except for the selection scenario. This can be because genic variance is only reduced by genetic drift but for this for this study, the population of 3000 could have been large enough for rapid loss of variance by drift. The phenomenon could also be caused by directional changes of gene-frequencies, which were only present in the selection scheme (Lande & Porcher, 2015).

6.0 Conclusion

Selection has a large effect on the G-REML estimates and the estimated genetic variance takes an intermediate value in between the realized and genic variance. The estimate, therefore, does not reflect the actual population variance. The two assortative mating methods have a larger effect on realized genetic variance than they have on estimated genetic variance. The latter is much closer to the genic variance and therefore largely represents genic variance despite being based on marker rather than QTL relationships. The estimated genetic variance is slightly biased in the direction of the realized genetic variance, which is most evident for positive assortative mating where the difference between realized and genic variance is largest.

Both sample size and SNP density have little effect of the same magnitude on the realized, estimated, and genic variances. The estimated genetic variance for these scenarios seems to be closer to the realised genetic variance than it is to the genic variance. The realised genetic variance remains unaltered with variations in sample size and SNP density. Estimated genetic variance in these scenarios represents both realised and genic genetic variance hence the actual variance between individuals in a population. Assortative mating and selection, however, alter the realized variance. Subsequently, the estimated genetic variance for these factors is biased and may not represent the realised and genic, and hence the actual population variance.

Sample size and SNP density have little effect on G-REML estimates. Heritability estimates obtained using such variance components may be close to the true heritability of a trait in a population. Assortative mating and selection lead to biased estimates which may not represent the actual heritability.

References

Atsbeha, D. M., Kristofersson, D., & Rickertsen, K. (2015). Broad breeding goals and production costs in dairy farming. *Journal of Productivity Analysis*, *43*(3), 403–415. https://doi.org/10.1007/s11123-014-0412-0

Bastiaansen, J. W. M., Bovenhuis, H., Lopes, M. S., Silva, F. F., Megens, H.-J., & Calu, M. P. L. (2014). SNP Effects Depend on Genetic and Environmental Context. *Proceedings of 10th World Congress of Genetics Applied to Livestock Production*, *10*, 356–362.

Bolormaa, S., Gore, K., Van Der Werf, J. H. J., Hayes, B. J., & Daetwyler, H. D. (2015). Design of a low-density SNP chip for the main Australian sheep breeds and its effect on imputation and genomic prediction accuracy. *Animal Genetics*, *46*(5), 544–556. https://doi.org/10.1111/age.12340

Bulmer, M. G. (1971). The effect of selection on genetic variability. *South African Journal of Animal Sciences*, *31*(2), 107–114. https://doi.org/10.4314/sajas.v31i2.3836

Burt, C. H. (2015). Heritability studies: Methodological flaws, invalidated dogmas, and changing paradigms. *Advances in Medical Sociology*, *16*(May), 3–44. https://doi.org/10.1108/S1057-629020150000016002

Clo, J., Ronfort, J., & Abu Awad, D. (2020). Hidden genetic variance contributes to increase the short-term adaptive potential of selfing populations. *Journal of Evolutionary Biology*, *33*(9), 1203–1215. https://doi.org/10.1111/jeb.13660

Dash, S., Singh, A., Bhatia, A. K., Jayakumar, S., Sharma, A., Singh, S., Ganguly, I., & Dixit, S. P. (2018). Evaluation of Bovine High-Density SNP Genotyping Array in Indigenous Dairy Cattle Breeds. *Animal Biotechnology*, *29*(2), 129–135. https://doi.org/10.1080/10495398.2017.1329150

de los Campos, G., Sorensen, D., & Gianola, D. (2015). Genomic Heritability: What Is It? *PLoS Genetics*, *11*(5), 1–21. https://doi.org/10.1371/journal.pgen.1005048

Devaux, C., & Lande, R. (2008). Incipient allochronic speciation due to non-selective assortative mating by flowering time, mutation and genetic drift. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1652), 2723–2732. https://doi.org/10.1098/rspb.2008.0882

England, P. R., Cornuet, J. M., Berthier, P., Tallmon, D. A., & Luikart, G. (2006). Estimating effective population size from linkage disequilibrium: Severe bias in small samples. *Conservation Genetics*, *7*(2), 303–308. https://doi.org/10.1007/s10592-005-9103-8

Faber, J., & Fonseca, L. M. (2014). How sample size influences research outcomes. *Dental Press Journal of Orthodontics*, *19*(4), 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo

Falconer, D. S., & Mackay, T. F. C. (1996). Introduction to Quantitative Genetics (Fourth Edition). In *Trends in Genetics* (Vol. 12).

Fredeen, H. T., & Jonsson, P. (1957). Genic Variance and Covariance in Danish Landrace Swine as Evaluated Under a System of Individual Feeding of Progeny Test Group. *Zeitschrift Für Tierzüchtung Und Züchtungsbiologie*, *70*(4), 348–363. https://doi.org/10.1111/j.1439-0388.1957.tb01056.x

Guo, S. W. (1997). Linkage disequilibrium measures for fine-scale mapping: A comparison. *Human Heredity*, *47*(6), 301–314. https://doi.org/10.1159/000154430

Hayashi, T. (1998). Genetic variance under assortative mating in the infinitesimal model. *Genes and Genetic Systems*, *73*(6), 397–405. https://doi.org/10.1266/ggs.73.397

Huang, W., & Mackay, T. F. C. (2016). The Genetic Architecture of Quantitative Traits Cannot Be Inferred from Variance Component Analysis. *PLoS Genetics*, *12*(11). https://doi.org/10.1371/journal.pgen.1006421

K.E., S., J., R., & D., T. (2013). QMSim User's Guide. *Version 1.10*, *10*(July). http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1527-3350 NS -

Kolstad, K. (2005). Methods for estimating phenotypic and genetic parameters. *Selection and Breeding Programs in Aquaculture*, 121–143. https://doi.org/10.1007/1-4020-3342-7_9

Koopaee, H. K., & Koshkoiyeh, A. E. (2014). SNPs genotyping technologies and their applications in farm animals breeding Programs: Review. *Brazilian Archives of Biology and Technology*, *57*(1), 87–95. https://doi.org/10.1590/S1516-89132014000100013

Lande, R. (1976). The maintenance of genetic variability by mutation in a polygenic character with linked loci. *Genetics Research*, *89*(5–6), 373–387. https://doi.org/10.1017/S0016672308009555

Lande, R., & Porcher, E. (2015). Maintenance of quantitative genetic variance under partial self-fertilization, with implications for evolution of selfing. *Genetics*, *200*(3), 891–906. https://doi.org/10.1534/genetics.115.176693

Landguth, E. L., Fedy, B. C., Oyler-Mccance, S. J., Garey, A. L., Emel, S. L., Mumma, M., Wagner, H. H., Fortin, M. J., & Cushman, S. A. (2012). Effects of sample size, number of markers, and allelic richness on the detection of spatial genetic pattern. *Molecular Ecology Resources*, *12*(2), 276–284. https://doi.org/10.1111/j.1755-0998.2011.03077.x

Legarra, A., Robert-Granié, C., Manfredi, E., & Elsen, J. M. (2008). Performance of genomic selection in mice. *Genetics*. https://doi.org/10.1534/genetics.108.088575

Lin, C. Y., Xing, G., & Xing, C. (2012). Measuring linkage disequilibrium by the partial correlation coefficient. *Heredity*, *109*(6), 401–402. https://doi.org/10.1038/hdy.2012.54

Lush, J. L. (1943). *Animal Breeding Plans*. The Iowa State College Press Ames, Iowa.

Lynch, B. W. and M. (2018). Evolution and Selection of Quantitative Traits. In *OXFORD UNIVERSITY PRESS* (Vol. 53, Issue 9). https://doi.org/10.1093/oso/9780198830870.001.0001

McVean, G. A. T., & Charlesworth, B. (2000). The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics*, *155*(2), 929–944. https://doi.org/10.1093/genetics/155.2.929

Meuwissen, T. H. E., & Goddard, M. E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics*, *155*(1), 421–430. https://doi.org/10.1093/genetics/155.1.421

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*.

Meyer, K. (2007). WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). *Journal of Zhejiang University SCIENCE B*, *8*(11), 815–821. https://doi.org/10.1631/jzus.2007.b0815

Nelson, M. R., Wegmann, D., Ehm, M. G., Kessner, D., St, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S., Warren, L., Aponte, J., Zawistowski, M., Liu, X., Zhang, H., Zhang, Y., Li, J., Li, Y., Li, L., Woollard, P., … Mooser, V. (2015). *An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people*. *337*(6090), 100–104. https://doi.org/10.1126/science.1217876.An

Nirea, K. G., Sonesson, A. K., Woolliams, J. A., & Meuwissen, T. H. E. (2012). Effect of non-random mating on genomic and BLUP selection schemes. *Genetics Selection Evolution*, *44*(1), 1–7. https://doi.org/10.1186/1297-9686-44-11

Ogawa, S., Matsuda, H., Taniguchi, Y., Watanabe, T., & Nishimura, S. (2014). Effects of single nucleotide polymorphism marker density on degree of genetic variance explained and genomic evaluation for carcass traits in Japanese Black beef cattle. *BMC Genetics*, *15*(1), 1–13. https://doi.org/10.1186/1471-2156-15-15

Perkel, J. (2008). SNP genotyping: Six technologies that keyed a revolution. *Nature Methods*, *5*(5), 447–453. https://doi.org/10.1038/nmeth0508-447

Porto-Neto, L. R., Kijas, J. W., & Reverter, A. (2014). The extent of linkage disequilibrium in beef cattle breeds using high-density SNP genotypes. *Genetics Selection Evolution*, *46*(1), 1–5. https://doi.org/10.1186/1297-9686-46-22

Pryce, J. E., Johnston, J., Hayes, B. J., Sahana, G., Weigel, K. A., McParland, S., Spurlock, D., Krattenmacher, N., Spelman, R. J., Wall, E., & Calus, M. P. L. (2014). Imputation of genotypes from low density (50,000 markers) to high density (700,000 markers) of cows from research herds in Europe, North America, and Australasia using 2 reference populations. *Journal of Dairy Science*, *97*(3), 1799–1811. https://doi.org/10.3168/jds.2013-7368

Rao, C. R. (1971). Estimation of variance and covariance components-MINQUE theory. *Journal of Multivariate Analysis*, *1*(3), 257–275. https://doi.org/10.1016/0047-259X(71)90001-7

Rasch, D., & Mašata, O. (2006). Methods of variance component estimation. *Czech Journal of Animal Science*, *51*(6), 227–235. https://doi.org/10.17221/3933-cjas

Rawlik, K., Canela-Xandri, O., Woolliams, J., & Tenesa, A. (2020). SNP heritability: What are we estimating? *BioRxiv*. https://doi.org/10.1101/2020.09.15.276121

Rotwein, P. (2020). Revisiting the Population Genetics of Human Height. *Journal of the Endocrine Society*, *4*(4), 1–10. https://doi.org/10.1210/jendso/bvaa025

Sánchez, L., & Woolliams, J. A. (2004). Impact of Nonrandom Mating on Genetic Variance and Gene Flow in Populations with Mass Selection. *Genetics*, *166*(1), 527–535. https://doi.org/10.1534/genetics.166.1.527

Singh, V., & Singh, K. (2020). Encyclopedia of Animal Cognition and Behavior. *Encyclopedia of Animal Cognition and Behavior*, *February*, 84–86. https://doi.org/10.1007/978-3-319-47829-6

Slatkin, M. (2016). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, *9*(6), 477–485. https://doi.org/10.1038/nrg2361.Linkage

Sorensen, D. A., & Hill, W. G. (1983). Effects of disruptive selection on genetic variance. *Theoretical and Applied Genetics*, *65*(2), 173–180. https://doi.org/10.1007/BF00264888

Tortereau, F., Servin, B., Frantz, L., Megens, H. J., Milan, D., Rohrer, G., Wiedmann, R., Beever, J., Archibald, A. L., Schook, L. B., & Groenen, M. A. M. (2012). A high density recombination map of the pig reveals a correlation between sex-specific recombination and GC content. *BMC Genomics*, *13*(1). https://doi.org/10.1186/1471-2164-13-586

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423. https://doi.org/10.3168/jds.2007-0980

Visscher, P. M., & Bruce Walsh, J. (2019). Commentary: Fisher 1918: The foundation of the genetics and analysis of complex traits. *International Journal of Epidemiology*, *48*(1), 10–12. https://doi.org/10.1093/ije/dyx129

Waples, R. S. (2006). A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. *Conservation Genetics*, *7*(2), 167–184. https://doi.org/10.1007/s10592-005-9100-y

Wright, S. (1931). *Statistical Methods in Biology*. *26*(173), 155–163.

Xia, W., Luo, T., Zhang, W., Mason, A. S., Huang, D., Huang, X., Tang, W., Dou, Y., Zhang, C., & Xiao, Y. (2019). Development of high-density snp markers and their application in evaluating genetic diversity and population structure in elaeis guineensis. *Frontiers in Plant Science*, *10*(February), 1–11. https://doi.org/10.3389/fpls.2019.00130

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics*, *49*(9), 1304–1310. https://doi.org/10.1038/ng.3941

Yang, J., Zeng, J., Goddard, M. E., Wray, N. R., Visscher, P. M., & Science, A. (2010). *is an important genetic parameter that quantifies the proportion of phenotypic variance in a trait attributable to the additive genetic variation generated by all causal variants. Estimation of*. *2010*(3).

Zhang, J., Yang, J., Zhang, L., Luo, J., Zhao, H., Zhang, J., & Wen, C. (2020). A new SNP genotyping technology Target SNP-seq and its application in genetic analysis of cucumber varieties. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-62518-6