University of New England

A Framework for Optimizing Breeding Pairs Using Artificial Intelligence

Zhi Kang Loh

For the award of Doctor of Philosophy

2022

This research has been conducted with the support of the Australian Government Research Training Program Scholarship

Abstract

While selective breeding has played an important role in improving the economic performance of animals, traditional selection methods depend on animal-based data such as phenotypic or Estimated Breeding Values. The advent of novel genotyping technologies have led to genomic data, which directly probed into the genotypic configuration of the animals. This allows the exploitation of non-additive genetic components such as the dominance effects, which previously were not exploitable in selective breeding due to their dependence on the genotypic configurations of the parents, an aspect not made available through animal-based data. The use of such components has been relegated to crossbreeding systems, and rarely in within population mating systems.

For this reason, the aim of this thesis is to explore the optimization of breeding pairs and mating decisions, with emphasis on the use of genomic data. This thesis will explore the use of such data in the exploitation of additive and dominance genetic components while constraining the inbreeding level increment. To cover the large sample space of possible solutions, this project will be conducted using artificial intelligence for the optimization of breeding pairs. The optimization method proposed in this study was validated using a simulated dataset.

It is noted that there could be factors such as genetic architecture and data sizes that would affect the usability of genomic data in the optimization of breeding pairs, which was the reason this project starts by investigating the impact of these factors on the power and false positive rate of detecting quantitative trait loci (QTL) in a Genome-Wide Association Study (GWAS), a tool widely used for the detection of QTL and estimating the effect sizes of genomic regions. This study suggested significant impacts of sample sizes and number of markers, as well as genetic architecture of the traits on the power and false positive rates of the GWAS. This study also explored the performance of GWAS using two commonly used multiple testing correction methods, and also proposed a scoring method that could be used to test the optimality of thresholds between different multiple testing correction methods.

From the findings of this foundational work, techniques that could improve the performance of GWAS experiments have been explored. One such techniques was the calculation of optimal threshold that takes into account the effects of genetic architecture and data size. For this calculation, a method based on Receiver Operating Characteristics was developed to

calculate the optimal threshold of a GWAS. Simulation studies suggested this method performed better in binary classifications and marker selection for genomic predictions, with the use of this optimal threshold resulting in an increment of accuracy of genomic prediction up to 16.8% compared to that of the Bonferroni method, and 7.0% compared to the Benjamini-Hochberg FDR method.

The calculation of optimal threshold requires information on the genetic architecture of the trait, and this has become the basis for the next part of the thesis, where a novel method that estimates the genetic architecture parameters such as number of QTL and shape of the effect size distributions was proposed, while taking into account the impact of various confounding factors such as correlation between markers, heterogeneity in linkage disequilibrium structures, and allele frequency distribution. Using this method, the estimated number of QTL with effect sizes 0.1 σ_e ranged from 69.9% to 167.0% (an average of 109.8%) of the true number of QTL, and for effect size 1.0 σ_e it ranged from 101.6% to 175.8% (an average of 123.6%). The method was developed to be able to estimate the QTL effect size, similar to a GWAS, but taking into account the impact of the confounding factors. This method would also allow the detection of QTL with smaller effect size with more confidence. New statistical tests designed to be powerful at the tail of the QTL distribution were developed, and an observation was made on the preference of utilization of test statistics for optimization of breeding pairs over the estimated effect size of the markers.

For the final chapter, a framework for the optimization of breeding pairs was developed that could optimize both the additive and dominance genetic component while constraining the increment in inbreeding coefficient. For this framework, a genetic algorithm was used. Using the EBVs, this method successfully improved the additive genetic component by up to 87.0% compared to a truncation genomic selection method. Using heterozygosity as a mean of optimizing the dominance component, the genetic lift from the dominance component in offspring is approximately twice the additive genetic gain, although the lift only occurs in the first generation.

This project is important for livestock producers or species conservationists who wished to improve the additive and non-additive genetic components in their breeding herds by using genomic data. It is anticipated that this framework could be further developed into a full-fledged product that could be utilized in a commercial setting.

Certification

I, Zhi Kang Loh, hereby certify that ideas, results, analysis, software and conclusions contained in this thesis are bona-fide works of the candidate, and has not been previously submitted for an award, except where otherwise acknowledged. To the best of candidate's knowledge and belief, works included in this thesis are entirely my own efforts and do not contain materials previously written or published by another person except where due acknowledgements, citations and references are made in the thesis to that work.

Zhi Kang Loh
Doctor of Philosophy Candidate
University of New England
26 October 2022

Acknowledgement

I like to thank the assistance I received from my supervisors Julius H. J. van der Werf and Sam Clark for guiding me through my candidature, reviewing my approaches, ideas and works that I have applied in this project, and giving constructive criticism on any shortcomings in my works. Without their helps, this project would not be possible.

I also like to thank the support from Australian Government Research Training Program Scholarship for providing me financial support to complete this project. Without the support from the scholarship, this project would not be possible to be completed.

I would also like to thank all my family members and friends for providing me emotional and moral supports throughout my journey through the candidature, getting me through challenging times while giving me suggestions and ideas on how to improve my works. This project would also not be possible without their supports.

Thank you all!

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

Selection processes have played a major role in livestock production since the beginning of human civilization, starting from the process of domestication of various wild animals to the formation of specialized breeds tailored for high economic performance. In recent decades, selection has made dramatic increase on the livestock's economic productivity (Bökönyi, 1974; Gill and Harland, 1992).

Previous selection processes utilized animal-based data such as phenotypic data or, derived from that, Estimated Breeding Values (EBVs) of the animals. The advancement of molecular technology and the development of genotyping techniques for high-density genetic marker arrays based on Single Nucleotide Polymorphism (SNP) and Whole Genome Sequencing (WGS) has led to the development of a new class of genetic data utilizable in a selection program: genomic-based data. Genomic-based data, such as a genotype array, can be used in combination with phenotype to estimate the genetic merit contributed by a genomic region toward the phenotype, which can be used to scan for causal variants associated with a trait through Genome-Wide Association Study (GWAS) (Spencer et al., 2009; Visscher et al., 2017). Genomic data can also be used to estimate the additive genetic variance of a trait (Yang et al., 2009) and EBVs of the animals, which is subsequently utilized in genomic selection (VanRaden, 2008). Genomic-based estimates of level of consanguinity between animals have also been developed (VanRaden, 2008), which has been used in Optimal Contribution Selection (OCS) method where selection is done under a constraint of inbreeding coefficient increment (Clark et al., 2013).

Most of the breeding programs have thus far focused on selection using additive genetic effects, with the non-additive effects, such as dominance and epistasis, often being used in crossbreeding program, but rarely for selection and mating within breed. Unlike the additive genetic component, which depends solely on the number of copies of alleles, the non-additive component depends on the exact genotypic configurations of the alleles, which would be scrambled from parent to offspring generations through Mendelian assortment (de Boer et al., 1993). This precludes the use of EBV in optimization of the non-additive genetic component. This is further complicated by the difficulties in estimation of these non-additive genetic components, as that requires multiple observations of the same mating, which occurs mainly in species with larger full sib groups. For this reason, these genetic components were

considered not exploitable through selection and mating designs in most animal-based data (Lynch and Walsh, 1998). Marker-based data directly probes into the genotypic configuration of the parents however, which allows the prediction of offspring genotype, including the non-additive effects which can be estimated from heterozygosity. In theory, genomic-based information would allow the selection of individuals based on both the additive and non-additive genetic components.

Therefore genomic-based data can theoretically also be used to optimize the breeding pairs in a selection program. While this optimization is traditionally done using EBVs of the animals (Kinghorn, 2000), genomic-based data is now available, which can then be used to select animals with high merit and predict offspring merit based on both additive and non-additive effects.

When using genomic data in the optimization of selection and mating, it is imperative to establish which region in the genome is associated with the trait. This could be estimated with several methods, e.g. such as those used in GWAS. As a method, GWAS suffers from several limitations however. Due to the stringent threshold from the large number of markers and low proportion of variance explained by individual QTL, GWAS in general failed to explain a large portion of the additive genetic component (Hall et al., 2016). Studies on the factors that affect the false positive rate of a GWAS, which could produce a misleading result for the optimization, remain lacking. The effects of certain factors, especially those pertaining to the genetic architecture of the traits, on the power and false positive rate of the GWAS also have not been studied widely. There were also problems with how well a threshold balances the power and false positive rate of a GWAS, which previous publications have suggested to be highly dependent on some of these factors (Hoggart et al., 2008; Panagiotou and Ioannidis, 2012). These issues of GWAS could have contributed to the replicability crisis of a GWAS (Heller and Yekutieli, 2014; Wang and Zhu, 2019), which could impede the use of genomic-based information on the optimization of the breeding pairs.

With this in mind, the aim of this project is to design a framework for the optimization of breeding pairs in a selective breeding program, with emphasis placed on optimizing the additive and non-additive genetic components while constraining the increment in level of inbreeding. For this study, only the dominance component was utilized due to the difficulty in obtaining estimates required for the optimization of an epistatic component. Emphasis was placed on the use of additive and non-additive genomic-based information in the optimization

of the breeding pairs, such as estimated effect sizes and test statistics from a GWAS experiment. Due to the vast sample space of the possible mating pairs, methods based on artificial intelligence was used to optimize the mating pairs.

This project starts by investigating the factors that could affect the reliability of the genetic effects estimated at various genomic regions by a GWAS. A comprehensive study on potential confounding factors that could affect such reliability, such as genetic architecture of the trait, data size, allele frequency distribution and correlations between markers, will be detailed in the first experimental chapter (Chapter 3), which serves as a foundation for the subsequent chapters. Using the findings obtained from Chapter 3, techniques that could be utilized to improve the power and false positive rate of the GWAS were developed. This includes the calculation of an optimal threshold that balances power and false positive rate of a GWAS, which is proposed in Chapter 4, and presentation of a method of estimating genetic architecture parameters and QTL effects size while taking into account the effects of aforementioned confounding factors in Chapter 5. In the final chapter (Chapter 6), findings from the previous chapters were incorporated into the development of an optimized selective breeding method that could utilize genomic-based information to optimize the additive and non-additive genetic component under a constraint of increment of inbreeding level. This method was tested using simulated data under varying parameter values.

This project would be important for livestock breeders and producers who aim to improve the genetic merits of the animals while constraining the level of inbreeding and exploiting the non-additive genetic components, as well as breed or species conservationists that aim to preserve the additive and non-additive variation for traits.

Chapter 2. Literature Review

2.1. Abstract

The aim of this chapter is to review previous studies that have been done on optimizing breeding pairs, as well as all the necessary components for such optimization. There are three main sections in this literature review. The first part covers optimal contribution selection for the optimization of contributions of selection candidates to the next generation. The second part of the chapter deals with the components required for the optimization. This includes the estimation of inbreeding coefficient and co-ancestry of the selected animals, as well as previous attempts to detect the additive and non-additive effects of quantitative trait loci (QTL). The third part focuses on factors that would be important in the detection of these QTL, and how to improve their detection. This includes a discussion on the threshold for the multiple testing correction methods in the testing of QTL effect sizes. Findings from this literature review will also be used to establish the most practical approach for designing the breeding pair optimization method, as well as various aspects that need to be taken into account when designing such methods.

2.2. Selection Process and the Beginning of Optimized Selection Program

Selection processes have played a major role in the livestock production since the beginning of human civilization, from the breeding of domestic sheep from mouflons for wool colour and reduced fibre diameter (Ryder, 1973) up to the breeding of cattle for meat and dairy production (Gill and Harland, 1992). The selection process has been imperative in the improvement of the livestock herds to fulfil the ever-increasing needs of the humanity (Bökönyi, 1974).

For most of the history, selective breeding is done based on truncation selection, where the individuals were selected based on their performance alone, with all the substandard individuals being culled and the top animals were chosen to be propagated into the next generation (Akdemir and Sánchez, 2016; Crow and Kimura, 1979). The selection can be done from the traditional phenotype-based selection process up to the most recent genomic selection process (VanRaden, 2008). This breeding system is simple to operate and execute

and is effective in changing the phenotype of the animals (Crow and Kimura, 1979). There are several models being proposed in terms of expected additive genetic gains that could be obtained from a certain breeding strategy, most notably the link between the proportion of sires and dams selected with the expected additive genetic gain per generation (Crow and Kimura, 1979; Falconer, 1989; Robertson, 1970).

The traditional selection method and its associated models have their shortcoming, however. One such shortcomings is the un-optimized contributions and allocation of sires toward the dams. As dams are more restricted in the amount of contributable genetic material to the next generation per animals, they represented a limiting resource toward a breeding program (Wray and Goddard, 1994; Robertson, 1970). Traditional selection methods tend to produce an "equal and randomized contribution" of sires, where each sire has equal chance of contribute to the next generation, and yet their contributions and mating are randomized (i.e. no specific patterns in the sire contributions and sire-dam matching). The equal and randomized contribution of sires might cause less valuable sires to contribute excessively to the gene pool in the offspring while leaving fewer dams for the best sires, impacting the overall performance of the offspring and reducing the efficiency of a breeding program. Randomized contribution also produces un-optimized pairings of sires and dams, which would lead to a failure in optimization of the non-additive genetic components that depend on the exact configuration of alleles of sire and dam (de Boer et al., 1993).

Furthermore, truncation selection method and its associated models assume that the optimal values for some parameters, such as the proportion of selected sires, can be calculated before the commencement of a breeding program, and then remain unchanged for the subsequent generations (Brotherstone and Goddard, 2005; Meuwissen, 1997). Such breeding methods have been noted as the "static" breeding decision, which is less optimal as it fails to exploit unforeseen genetic gains in the subsequent generations (Meuwissen, 1997), while not taking in consideration the actual situations of the breeding herds, such as the consanguinity in the base population. This has become the impetus for a "dynamic" or "tactical" breeding decision, where the optimal sire contributions are calculated using information about the pedigree structure from the breeding herds (Brotherstone and Goddard, 2005; Meuwissen, 1997).

The paradigm of tactical breeding decision allows a more flexible framework for the breeding strategy, which includes controlling the level of inbreeding and the genetic diversity in the

population (Brotherstone and Goddard, 2005; Kinghorn, 2000). While inbreeding can also occur in random and natural selection, the artificial selection process utilized in livestock production systems accelerates the rate of inbreeding as selected animals tend to be more related (Falconer, 1989). Inbreeding has been implicated with a decline in the economic performance and welfare of the animals in a condition known as "inbreeding depression" (Falconer, 1989; Ryder and Wedemeyer, 1982; Schlie, 1967). This led to the development of tactical breeding strategies that aimed to maximize additive genetic gains while restricting the level of inbreeding.

Early attempts of this strategy focused on controlling the proportion of selected sires. The simplest and yet most inflexible method was simply calculating the optimal proportion of top sires to be selected such that the expected increase in inbreeding would be at the predefined level (Toro and Perez-Enciso, 1990). Dempfle (1975) and Dempfle (1990) attempted to achieve this aim by calculating the number of sires contributed by each full-sib family, while finding a balance between a within-family selection, which minimizes the increase in inbreeding coefficient, and between-family selection, which maximize the coefficient. This approach might not applicable if family information is unavailable, limiting its applicability (Toro and Perez-Enciso, 1990; Wray and Goddard, 1994). Using linear programming, Toro and Perez-Enciso (1990) proposed a mate selection method where the best combinations of individual sires and dams were chosen in attempt to maximize the genetic gain under the constraint of inbreeding level increment for one generation. While this method resolves the impact caused by the "randomized contribution", this method assumes a fixed equal number of sires and dams to be propagated into the next generation while disallowing half-sibs in the offspring population, thus with restrictions in its flexibility (Toro and Perez-Enciso, 1990).

Wray and Goddard (1994) argued that all aforementioned methods have made some arbitrary criterion and assumptions to control the increase of inbreeding rate, which may not reflect the true situation that might be encountered in a livestock production system. Therefore, rather than finding the optimal proportion of sires selected from a predefined level of inbreeding coefficient, Wray and Goddard (1994) directly derive a score that dictates the balance between the additive genetic gains and the increase in inbreeding coefficient, which for sire-only selection is defined using the following LaGrange objective function (denoted as f_{obj}) (Wray and Goddard, 1994):

$$f_{obj}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x^T b} - Q * \frac{1}{8} * (\boldsymbol{x^T A x}) \quad [1]$$

Where $\boldsymbol{x}$ is defined as the vector of sire contributions, $\boldsymbol{x^T}$ (with superscript $\boldsymbol{T}$) denotes the transpose of the vector of the sire contributions, $\boldsymbol{b}$ being the sire Estimated Breeding Values (EBVs), $\boldsymbol{A}$ being the numerator relationship matrix between the sires, and Q being the Lagrange multiplier that act as a weighting factor that balances emphasis on the additive genetic gain versus the inbreeding coefficient. The use of LaGrange objective function in balancing the additive genetic gains (i.e. $\boldsymbol{x^T b}$) and the increase in inbreeding coefficient (i.e. $\boldsymbol{x^T A x}$) significantly improves the model's flexibility by allowing a variable number of sires and dams to be selected with variable contributions.

Wray and Goddard (1994) calculated Q using the expected genetic gains for an infinite population size, time domain and observed inbreeding depression. The increment in inbreeding level is not part of the calculation of Q however, due to an assumption of uniform risk of inbreeding depression toward a breeding program, and this has become the subject of criticism by Meuwissen (1997). A high inbreeding coefficient increases the probability of deleterious alleles being drifted toward fixation, increasing its risk toward a breeding program. This is especially problematic if it is used on a trait with little inbreeding depression during the time of calculation of Q, which produces an overly lenient Q. For this reason, Meuwissen (1997) advocated the calculation of Q directly from the targeted increment of level of inbreeding coefficient, with the presumption that the breeders would generally know the acceptable increment of level of inbreeding coefficient. Despite this, the algorithm proposed by Meuwissen (1997) still largely followed of the one by Wray and Goddard (1994), including the use of a LaGrange objective function. Therefore, both Meuwissen (1997) and Wray and Goddard (1994) have been hailed as the pioneering work for the modern-day Optimal Contribution Selection (OCS) method, which has since become the main method for an optimized selective breeding operation (Brothersone and Goddard, 2005; Clark et al., 2013; Nielsen et al., 2011).

Despite the importance of Meuwissen (1997) and Wray and Goddard (1994), both methods are still highly restrictive in terms of its flexibility. This is caused by the fact that both methods have been built using generalized theorems based on isolated studies on the effects of selection on additive genetic gain and inbreeding coefficient. A more unified framework on other aspects pertaining to a breeding operation, such as those related to mate allocations and economic evaluations, remained lacking. For this reason, Kinghorn (2000) employed the "Mate Selection" framework, where indices from various aspects in the breeding operation could be integrated. In essence, given a vector of contributions of chosen and allocated sires

$\boldsymbol{x}$, the integrated index for mate selection, the Mate Selection Index (MSI) has the generalized form as follows:

$$MSI = \sum_{k=1}^{n_{factor}} \lambda_k * f_k(\boldsymbol{x}) \quad [2]$$

Where n_{factor} is the number of factors taken into account by the MSI and λ_k being the weighting factor associated with the components that needed to be optimized, which is defined as $f_k(\boldsymbol{x})$. The $f_k(\boldsymbol{x})$ could include any factors associated with a breeding operation, such as additive genetic gain, impact from inbreeding coefficients, gains from heterosis and cost of mating policy (Kinghorn, 2000).

Maximizing the MSI is not a trivial issue however, as this index comprises of mathematically disparate components. Due to this, Kinghorn (2000) utilized an differential evolution, a form of evolutionary algorithm, in order to find $\boldsymbol{x}$ that maximize the MSI, given the constraints set by the λ_k and $f_k(\boldsymbol{x})$. This allows a more flexible model of optimization to be specified, thus better suited to real life scenarios commonly encountered in a breeding operation.

Since Kinghorn (2000), many other renditions of OCS algorithms have been proposed and published. Many of these new algorithms draw inspirations from Kinghorn's work however, with the most notable aspect being the use of evolutionary algorithm. Due to its flexibility, the evolution algorithm would later be directly used to optimize the OCS's LaGrange objective function as defined in equation [1], allowing the more dynamic and case-by-case breeding strategy that maximizes the additive genetic gains under a constrained inbreeding rate (Gourdine et al., 2012; Sørensen et al., 2008). Evolutionary algorithms have allowed the emergence of new OCS algorithms such as the Look Ahead Mate Selection" (LAMS) algorithm (Shepherd and Kinghorn, 1998) and Differential Evolution based methods (Kinghorn, 2011), and OCS for conflicting breeding objective (Wang et al., 2017a).

Given the flexibility of the evolutionary algorithm, this opened up a possibility for selection that takes into account non-additive genetic components such as dominance effects. While they could have contributed a significant portion of the genetic variance, they are there are difficult to exploit as they depend on the parental genotypic configuration, which is not observable through pedigree data (Crow, 2010; Falconer, 1989; Lynch and Walsh, 1998). However, with the advancement of high-density genomic data such as Single Nucleotide Polymorphism (SNP) markers and whole genome sequence data however, this allowed a

direct observation on the genotypic states of the individuals, therefore allowing the prediction of non-additive genetic component in the progeny. Despite this, OCS and mate allocation algorithms that utilize the non-additive genetic component is currently lacking. While Kinghorn (2000) mentioned the possibility of exploiting the additional genetic gains derived from crossbreeding (i.e., heterosis), the publication did not provide additional information on the calculation and optimization of the non-additive genetic component. González-Diéguez et al. (2019) have proposed a mate allocation method with dominance effect taken into account, but do not restrict the increment in inbreeding level. The lack of methods to increase additive and non-additive while constraining the increase of inbreeding coefficient stands as a missed opportunity of improving the long-term economic yield of a breeding program.

Despite the opportunity, there is a possibility that an OCS that utilizes additive and non-additive genetic components while simultaneously constraining the increase in inbreeding level might not be feasible. For this reason, it is important to establish the feasibility of such OCS, and this can be done by testing the feasibility of obtaining an estimate for each component in the OCS, and the feasibility of combining these estimates into an OCS.

2.3. Aspects of an Optimized Selective Breeding

2.3.1. Inbreeding

Perhaps the most important aspect for an optimized selective breeding program is the control of inbreeding, defined as the breeding of genetically related parents (Falconer, 1989; Griffith et al., 2015). Statistically the coefficient of inbreeding is defined as the probability of two alleles in an individual being inherited from the same copy of allele of a common ancestor (Griffith et al., 2015).

2.3.1.1. Estimating the Changes in Inbreeding Coefficient

To control the level of inbreeding, the changes of inbreeding coefficient per generation of selection (denoted as ΔF) need to be estimated.

There were several methods being put forward to estimate the changes in the level of inbreeding. Some of the earliest methods utilized pedigree data, with the pioneering works being those of Wright (1921) and Fisher (1949). The most important and more practical pedigree-based method came after 1950 however, with Charles R. Henderson developing a matrix that contains the ancestry relationship between each individual (Henderson, 1975).

This matrix, now known as Numerator Relationship Matrix (NRM), has been used to develop efficient algorithms to calculate inbreeding coefficients in very large pedigrees (Quaas, 1976; Meuwissen and Luo, 1992).

In breeding programs inbreeding needs to be managed by controlling the rate of inbreeding, i.e., to limit the increase of the average inbreeding coefficient from generation to generation. Inbreeding coefficients themselves are relative to some based population of animals where ancestry is unknown, e.g., these coefficients would be higher in a population with a known deep pedigree. Using NRM, the expected increase in inbreeding coefficient can be expressed as the co-ancestry among the selected parents (Meuwissen, 1997), which can be calculated as follows:

$$\Delta F = \frac{1}{2}\boldsymbol{x^T A x} \qquad [3]$$

This method requires an accurate and complete pedigree data, with any missing data could lead to underestimation of co-ancestry of selected parents and therefore the rate of inbreeding (VanRaden, 1992). Similarly, inbreeding coefficients can be underestimated with the assumption of unrelated founder population in the NRM (McQuillan et al., 2008).

The availability of high-density genotype array allows alternative methods of estimating the genetic relationship between individuals, such as the "Genomic Relationship Matrix" (GRM) by VanRaden (2008). Mathematically, a GRM is a square symmetric matrix that contains the scaled covariance of genomic states between animals (Gondro, 2015), and the inbreeding coefficient is defined as the diagonal of the GRM subtracted by 1 (Caballero et al., 2022). It can also be defined as the ratio between the variance of a SNP marker with the sum of variances from all SNP markers, subtracted by 1 (Caballero et al., 2022). In the context of optimized selective breeding, the GRM could be used in place of NRM in equation [3] to estimate the inbreeding level changes. GRM has the advantage in its ability in estimating the co-ancestry of apparently unrelated animals (Gondro, 2015). Despite this, the inbreeding coefficient estimated through GRM is not robust against changing allele frequencies (VanRaden, 2008; Zhang et al., 2015), and given the fact that the GRM is the scaled covariance between individuals, negative values in GRM are possible, which might cause some discrepancies in the estimated ΔF with those that utilized NRM.

Besides deriving inbreeding from the genomic relationship matrix as proposed by VanRaden (2008), several other methods have been proposed for the calculation of changes in

inbreeding level. For example, Yang et al. (2010) proposed a similar method for calculation of inbreeding coefficients based on the correlation between uniting gametes, while Li and Horvitz (1953) proposed a method using the expected homozygosity with the assumption of a Hardy-Weinberg Equilibrium (Caballero et al., 2022).

A common feature for these methods is that they define the inbreeding coefficients as the sum of variance explained by each SNP marker scaled by the total variance of the allele frequencies from all the markers. This methodology has been described by Hou and Ochoa (2023) as "ratio of means" (ROM) methods. A close counterpart for these methods is the "means of ratio" (MOR) methods, where the variance of a SNP marker is scaled by variance of the allele frequency from that marker alone. VanRaden (2008), Yang et al. (2010) and Li and Horvitz (1953) have independently proposed the MOR counterpart of their respective methods. Despite this, the MOR methods have been criticized for poorly reflecting the kinship structure as the kinship matrix generated is ill-conditioned (Hou and Ochoa, 2023), and tend to behave poorly with small sample sizes (Caballero et al., 2023). Neither Caballero et al. (2021) and Hou and Ochoa (2023) provided explanations for this observation.

One possible reason for the ill-conditioned matrix from MOR methods lies in the distribution of the kinship estimates, which in turn rooted from the denominators of these estimators. Since the denominator of ROM estimator involves the summations of variances across all the SNP markers, their denominator has a larger magnitude than the denominators of the MOR (which do not involve such summation). The summation of variances increases the magnitude of the denominator of ROM estimators, thus decreases the sample variance and kurtosis of this estimators (compared to MOR estimators). This observation was supported through additional simulations, which suggested MOR estimators of VanRaden (2008) produces a significantly larger variance and kurtosis than the ROM equivalent (Figure 2.1). Furthermore, the expressions for the MOR estimators are closely analogous to the expressions analysed by Pillai and Meng (2016) who proved such expressions produce Cauchy-distributed random variables, a distribution renowned for its "pathological" behaviour of having undefined mean and variance (Mun, 2012). Indeed, the expressions for MOR estimators suggested these estimators would follow a ratio distributions, which often have ill-defined moments (Brody et al., 2002). It is likely that these ill-defined moments are the culprit of ill-conditioning of these kinship matrix, thus the poor estimation of variance component. The observation of ill-conditioned kinship matrix implies the preference of using ROM estimators over MOR estimators in the estimation of inbreeding.

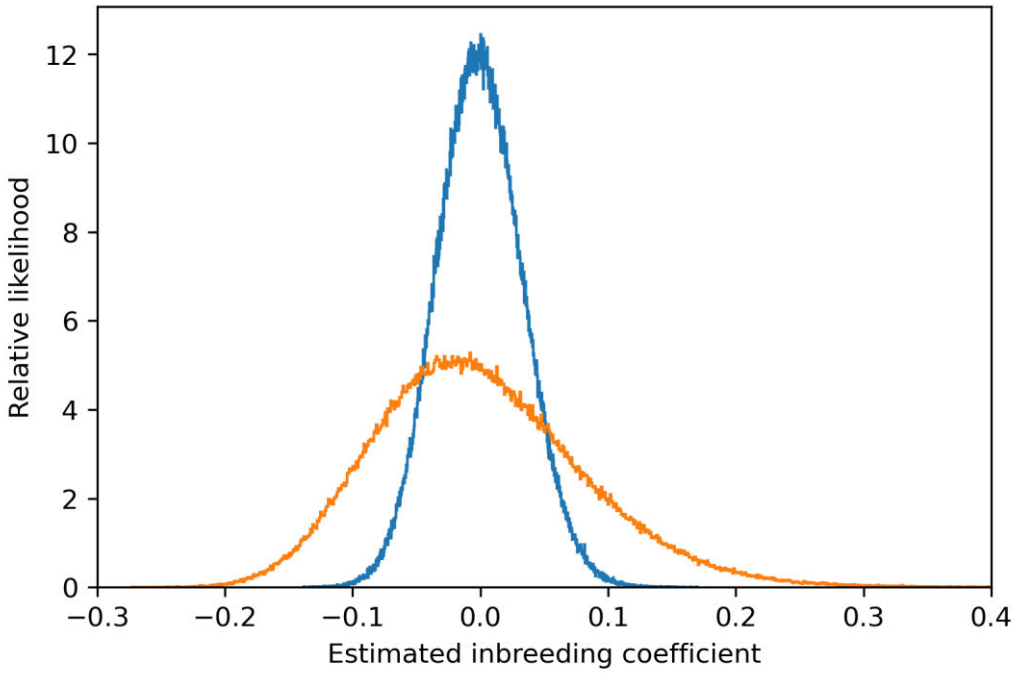

Figure 2.1: Histogram showing the estimated inbreeding coefficient obtained from ratio of mean (ROM) method (blue) and mean of ratio (MOR) methods (orange) from VanRaden (2008). The estimated inbreeding coefficient from MOR method has a sample variance of 0.0066 and kurtosis of 0.491, which are significantly higher than the corresponding values of 0.0011 and 0.027 from that obtained using ROM method. In this simulation, 100 repeats of 5000 unrelated samples with 50k markers with average pairwise linkage disequilibrium of 0.9 were conducted, with allele frequencies following a symmetric beta distribution with shape parameter of 0.5, and minor allele frequency filtering set at 0.01.

Method based on effective population sizes in the selected population has also been proposed by Wang et al. (2017a). In this method, the variance effective population size of the selected sires and dam has been utilized, with the equation as defined as such:

$$\Delta F = \frac{1}{8N_m} + \frac{1}{8N_f} \quad [4]$$

Where N_m and N_f are the number of selected breeding sires and dams respectively. This method ignores the inbreeding that might already exist in the base population. This method also does not take into account the impact of varying amount of contributions of sires and dams on the inbreeding coefficient (as an example in the scenario of 100 sires with each contributed to 10 dams, compared to another with 99 sires contributed to one dam, while the one remaining sire contributed to 901 dams). These shortcomings could be easily overcome with the use of relationship matrices, for which equation [3] gives the relationship among the selected parents, thus predicting the inbreeding rate.

2.3.2. Additive Genetic Component

The most important genetic component of the phenotype contributing to long term genetic improvement is the additive genetic component. The loci associated with this component would contribute to the phenotype in an amount proportional to the number of copies of an allele (Falconer, 1989; Lynch and Walsh, 1998). The additive genetic effect at each locus is defined as half of the differences between the homozygotes:

$$a = \frac{\mu_{AA} - \mu_{aa}}{2} \quad [5]$$

Where μ_{AA} and μ_{aa} are the phenotypic mean of the homozygotes with genotype AA and aa respectively. Given n number of additive QTL, the portion of the phenotype explained by the additive loci (denoted as A) is the sum of the individual additive effects of each of the loci (Falconer, 1989; Lynch and Walsh, 1998):

$$A = \sum_{k=1}^{n} g_{a_k} a_k \quad [6]$$

Where g_{a_k} is the number of copies of an allele in locus k, with the values of $\{0,1,2\}$ for genotype aa, Aa and AA, respectively, and a_k being the additive effect size of the additive loci. With the assumption of independence between QTL and Hardy-Weinberg Equilibrium (HWE), the additive genetic variance (denoted as $var(A)$) is defined as the sum of variances contributed by each of the n QTL loci (Falconer, 1989):

$$var(A) = 2 * \sum_{k=1}^{n} p_k(1 - p_k) * a_k^2 \quad [7]$$

Where p_k is the allele frequency of locus k. In reality it is unlikely that loci act independently from each other. It is also infeasible to detect all QTL that contribute to the variance of a trait as there are likely many QTL with very small effects and difficult to detect statistically. Due to this the variances explained by the detected QTL often explained less variances than what is expected from its heritability, causing the missing heritability problem (Maher, 2008), and have prompted studies like Yang et al. (2011) that uses restricted maximum likelihood (REML) method to detect variance from additional QTL. Attempts to detect these additional variance or QTL warrants further studies.

Most of the efforts on QTL detection focused on additive loci, as their effects are heritable, and their frequencies and variances can be easily and reliably altered through a selection process (Falconer, 1989). Many polygenic traits also have significant portion of its variance explainable through additive genetic component, often more so than that explained by non-additive genetic component (Crow, 2010; Visscher et al., 2017), and there are solid groundworks on the methodology of estimation of the effect sizes of the loci.

2.3.2.1. Detection of Additive QTL

The detection of additive QTL is based on regression of the phenotypes on genotypes, which has become the basis of association studies. The advent of high-density markers that span throughout the genome has given rise to Genome-Wide Association Study (GWAS). The additive effects could be estimated one locus at a time (as in Single SNP Regression) or simultaneously (as in SNP Best Linear Unbiased Prediction (SNPBLUP) or Bayesian methods) (Gondro, 2015; Wang et al., 2016).

2.3.2.1.1. Single SNP Linear Regression

The Single SNP Linear Regression method is perhaps the most straightforward method of detecting additive QTL. The basis of this method is to conduct a linear regression of the phenotypes on the genotypes for each marker. Given a locus j, the estimated additive QTL effect sizes (denoted as $\hat{a}$) can be defined as follows (Falconer, 1989):

$$\hat{a_j} = \frac{cov(\boldsymbol{x_j}, \boldsymbol{y})}{2p_j(1-p_j)} \quad [8]$$

With the $cov(\boldsymbol{x_j}, \boldsymbol{y})$ being defined as the covariance between genotype of locus j and the phenotype and the denominator the variance of the genotype

They $\hat{a_j}$ could then be used to test the significance of effects of the loci, with the null hypothesis being a is not significantly different from zero. The test statistic for this hypothesis (denoted as F) could be defined using $\hat{a}$ as follows:

$$F = \frac{2p_j(1-p_j)\hat{a_j}^2(N-2)}{var(\boldsymbol{y}) - 2p_j(1-p_j)\hat{a}^2} \quad [9]$$

Where N is the sample size of the GWAS experiment. Under null hypothesis, the F would follows a Snedecor's F-distribution, which asymptotically approaches a chi-squared distribution with large N (Wang and Xu, 2019), which in turn can be utilized to calculate the

p-values of the marker having its true effect size significantly different from the null hypothesis (Gondro, 2015). A detailed mathematical derivation has been provided in Appendix A.

Compared to SNPBLUP, single SNP linear regression is computationally less demanding, and the method does not shrink the estimated effect sizes of the QTL as strongly, but therefore tends to overestimate the marker effect sizes, especially if the marker has extreme allele frequencies (Gondro, 2015; Wang et al., 2016). The single SNP regression method is the simplest method as it does not need knowledge on the haplotypes. Some studies such as Grapes et al. (2004) suggest that the power for single SNP linear regression is comparable to haplotype-based methods. For sufficiently large sample size and as long as the central limit theorem applies, linear regression does not require normally distributed residuals (Bůžková, 2013). Single SNP regression has the disadvantage of difficulty in defining the true mutation due to the QTL being in linkage disequilibrium with multiple SNPs. This is especially problematic when large numbers of SNPs are being used. Therefore, the method is likely to inflate the number of detected SNPs, which can be solved by having all the SNP fitted simultaneously (Hayes, 2013; Pryce et al., 2010).

2.3.2.1.2. SNP Best Linear Unbiased Prediction (SNPBLUP)

SNP Best Linear Unbiased Prediction (SNPBLUP) is another common method used in GWAS. Unlike the Single SNP Linear Regression, this method fits all the SNP simultaneously, with the genotype fitted as random effect, and is solved through Tikhonov's regularization (Gondro, 2015; Hayes, 2013). For N number of animals and M number of markers, this method estimates the effect sizes by solving the following matrix equation for $\widehat{\boldsymbol{a}}$ (Gondro, 2015):

$$\begin{bmatrix}\widehat{\boldsymbol{\mu}}\\ \widehat{\boldsymbol{a}}\end{bmatrix} = \begin{bmatrix}\mathbf{1}_N^T\mathbf{1}_N & \mathbf{1}_N^T\boldsymbol{X}\\ \boldsymbol{X}^T\mathbf{1}_N & \boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}_M\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{1}_n^T\boldsymbol{y}\\ \boldsymbol{X}^T\boldsymbol{y}\end{bmatrix} \quad [10]$$

Where $\widehat{\boldsymbol{\mu}}$ is a 1×1 vector containing the estimated mean of the phenotype; $\widehat{\boldsymbol{a}}$ is a column vector of length M containing the estimated marker additive effect sizes; $\mathbf{1}_N$ being a column vector of ones with length of N, $\boldsymbol{X}$ being a frequency adjusted genotype array of size $N \times M$, with $\boldsymbol{I}_M$ being an identity matrix of size $M \times M$ and $\boldsymbol{y}$ being a column vector of length N containing the phenotype. The λ is a scalar shrinkage factor that controls the instability of estimated $\widehat{\boldsymbol{a}}$ caused by the near singularity of the $\boldsymbol{X}^T\boldsymbol{X}$ matrix, and is defined as follow (Gondro, 2015; Hoerl and Kennard, 2000):

$$\lambda = \frac{1 - h^2}{h^2} * 2 \sum_{k=1}^{M} p_k(1 - p_k) \quad [11]$$

Where p_k is the allele frequency at the k-th marker locus. It is also noted that as λ could add up to a large number with large number of markers, thus could severely regress the marker genotype contrasts in the estimation of the effects. Matrix $\boldsymbol{X}$ has also been adjusted based on the allele frequency at each locus as follows (Gondro, 2015):

$$\boldsymbol{X} = \boldsymbol{X_{raw}} - 2\boldsymbol{p} \quad [12]$$

Where $\boldsymbol{p}$ is vector of length M containing the marker allele frequencies. In terms of establishing the significance of the markers, the test statistics for the markers, which measures whether the markers jointly have an effect on the phenotypes, are estimated as such (Gondro, 2015):

$$F = \frac{\widehat{\boldsymbol{a}}^2}{var(\widehat{\boldsymbol{a}})} \quad [13]$$

Most of the advantages conferred by this method originated from the simultaneous fitting of all markers and the use of λ, which restrict the overestimation of effect sizes of the QTL, and the SNPs unrelated to QTLs are calculated much closer to zero. Also due to this, phenotype predicted from the estimated effect sizes of this method is also less overestimated and closer to the true additive genetic values (Gondro, 2015).

The main weakness for this method is its computational and memory intensiveness, and relatively slow compared to Single SNP Linear Regression. Due to its simultaneous fitting, this method would also produce strong shrinkage in the estimated effect sizes (Gondro, 2015). This can be partially mitigated my assuming the prior effect of effect sizes followed a non-linear regression, such as Student's t-distribution. Another method was based on estimating the posterior probability of whether the SNP follows a certain model, which assume a prior distribution of SNP effects with large mass at zero, and the remaining SNP in other non-linear distribution such as normal or t-distribution (Hayes, 2013).

A computationally less intensive variant of SNPBLUP utilizes an intermediate vector of animals' relationship matrix and EBVs (Gondro, 2015). With the assumption of $\boldsymbol{y}$ being mean-centred, the EBVs $\widehat{\boldsymbol{b}}$ could be calculated as follow (Gondro, 2015; Hoerl and Kennard, 2000):

$$\widehat{\boldsymbol{b}} = \left[\boldsymbol{I}_N + \left(\frac{1-h^2}{h^2}\right)\boldsymbol{G}^{-1}\right]^{-1} * \boldsymbol{y} \quad [14]$$

Where $\boldsymbol{G}$ is the GRM of the animals. If there are fewer genotyped animals compared to number of SNP markers, the matrix $\boldsymbol{G}$ in equation [14] is smaller than the $\boldsymbol{X^T X}$ in equation [10], thus reducing the computational intensity. Using the $\widehat{\boldsymbol{b}}$, the estimated additive QTL effect sizes can then be backsolved as follows (Gondro, 2015):

$$\widehat{\boldsymbol{a}} = \frac{1}{2\sum_{k=1}^{M} p_k(1-p_k)} * \boldsymbol{X} * \boldsymbol{G^{-1}} * \widehat{\boldsymbol{b}} \quad [15]$$

2.3.3. Non-Additive Genetic Component

Besides the additive genetic variance, there is a significant portion of the genetic variances that did not arise from the number of copies of an allele at a locus, but instead from the interaction between alleles or loci (Falconer, 1989). These non-additive genetic components are more difficult to estimate however, and thus frequently are being omitted in the models, despite its importance in many economically important traits (Lynch and Walsh, 1998).

Unlike the additive genetic component, exploiting the non-additive genetic component is not trivial, as they arise from certain allelic configurations that would be scrambled in the next generation (de Boer et al., 1993). Thus, this component would not be inherited in a predictable manner, and with dependency on the mating configurations of sires and dams. The detection of non-additive effect sizes also provides more challenges; Visscher et al. (2017) commented on the power of detection of an additive locus Q using marker M is in proportion to $R^2_{LD}(Q,M)$, whereas for non-additive genetic loci it is proportional to $R^4_{LD}(Q,M)$. Thus, given an effect size of a QTL, a much larger sample size is needed to detect a non-additive locus compared to additive locus, reducing its feasibility of detection.

There are two major types of non-additive genetic component: the interaction between alleles within a locus, which is known as dominance, and interaction between different loci, known as epistasis (Falconer, 1989). For this study emphasis was placed on optimizing the dominance genetic component. This is due to the difficulty of obtaining an estimate for the epistatic effect sizes, a field that warrants further study (Lynch and Walsh, 1998; Vitezica et al., 2018).

2.3.3.1. Dominance

The term dominance was first defined by Gregor Mendel in his works on plant breeding. He defined an allele being dominant if its effect overcomes the effect of its alternative allele and expressed in the phenotype (Mendel, 1865). Under the current framework of quantitative genetics however, the definition of the dominance has since been reformulated into the interaction between alleles within a locus (Isik et al., 2003). Statistically it is defined as the deviation in the heterozygote genotypic value from that expected from the expected mid-homozygote value (Falconer, 1989).

The prediction of dominance effects can be important as it contributes genetic gains toward the phenotypes (de Almeida Filho et al., 2016). Dominance effects are also thought to play a major role in the heterosis phenomena in crossbred animals, an aspect that has been exploited by breeders to increase the production rate and efficiency by crossing two inbred lines (Goto and Nordskog, 1959; Vitezica et al., 2016; Zeng et al., 2013).

Given a locus, the dominance is defined through the following equation (Zhu et al., 2015):

$$d = \mu_{Aa} - \frac{\mu_{AA} + \mu_{aa}}{2} \quad [16]$$

Assuming the dominance effects across loci are cumulative, given n number of loci with dominance effect sizes, the portion of the phenotype described by the dominance component (denoted as D) is defined as follow (Duenk, 2020):

$$D = \sum_{k=1}^{n} g_{d_k} d_k \quad [17]$$

Where g_{d_k} is the state of heterozygosity of the loci k, with $g_{d_k} = 1$ for heterozygote loci and $g_{d_k} = 0$ otherwise, and d_k being the dominance deviation as defined in equation [16]. Under the assumption of HWE and independence between QTL loci, the variance of dominance genetic component can be calculated through the equation (Zhu et al., 2015):

$$var(D) = 4 \sum_{k=1}^{n} (p_k * (1 - p_k) * d)^2 \quad [18]$$

2.3.3.1.1. Detection of Dominance Genetic Component

The initial models on prediction of breeding qualities often ignores the effect of dominance due to a lack of reliable method of estimating dominance effect (Misztal et al., 1998). Due to massive increase in genotypic and pedigree data, the effect of dominance has been taken into account for some of the models (Aliloo et al., 2017; de Almeida Filho et al., 2016; Sun et al., 2013; Zeng et al., 2013). Lynch and Walsh (1998) have suggested the following methodology of estimating the dominance portion of the phenotypes (denoted as $\widehat{\boldsymbol{d}}$ in this instance):

$$\begin{bmatrix} \widehat{\boldsymbol{\mu}} \\ \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{d}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}_N^T \mathbf{1}_N & \mathbf{1}_N^T & \mathbf{1}_N^T \\ \mathbf{1}_N & \boldsymbol{I}_N + \left(\frac{var(E)}{var(A)}\right)\boldsymbol{A}^{-1} & \boldsymbol{I}_N \\ \mathbf{1}_N & \boldsymbol{I}_N & \boldsymbol{I} + \left(\frac{var(E)}{var(D)}\right)\boldsymbol{D}^{-1} \end{bmatrix}^{-1} * \begin{bmatrix} \mathbf{1}_n^T \boldsymbol{y} \\ \boldsymbol{y} \\ \boldsymbol{y} \end{bmatrix} \quad [19]$$

Where $\boldsymbol{A}$ is the additive genetic relationship matrix, and $\boldsymbol{D}$ is the dominance relationship matrix, which is built using the coefficient of fraternities between individuals. While theoretically feasible, the practicality of this method is impeded by the need of estimating the dominance variance $var(D)$, which itself is not trivial (Lynch and Walsh, 1998).

Unlike additive EBV, $\widehat{\boldsymbol{d}}$ is not directly usable in predicting the dominance component of the offspring. This is due to the ambiguity in the expected offspring dominance component given the values of parent dominance component. This could be illustrated as follows: let h and H be two alleles in a QTL and let a and d be its additive and dominance effect sizes. Let the paternal and maternal additive genetic component be denoted as P_A and M_A, respectively, and for dominance component be P_D and M_D, respectively. The paternal additive genetic component is defined as follows: $\{hh, Hh, HH\} = \{0, a, 2a\}$ and for dominance $\{hh, Hh, HH\} = \{0, d, 0\}$. Using this information, the expected additive genetic component of the offspring can be defined as in Table 2.1.

Note that for all possible offspring additive genetic components, it is always defined as its mid-parent values $\frac{P_A + M_A}{2}$ (Falconer, 1989). Thus, if P_A and M_A are known, it is possible to predict with certainty the expected offspring additive genetic component. This is not the case for dominance component however; with the same parental genotypes, the expected dominance genetic component of the offspring was defined as in Table 2.2.

Table 2.1: The expected offspring additive genetic component given the paternal additive values (P_A) and maternal additive values (M_A).

Additive Genetic Component (Rows: P_A; Columns M_A)	hh $(M_A = 0)$	Hh $(M_A = a)$	HH $(M_A = 2a)$
hh $(P_A = 0)$	0	$0.5a$	a
Hh $(P_A = a)$	$0.5a$	a	$1.5a$
HH $(P_A = 2a)$	a	$1.5a$	$2a$

Table 2.2: The expected offspring dominance genetic component given the paternal additive values (P_D) and maternal additive values (M_D).

Dominance Genetic Component (Rows: P_D; Columns M_D)	hh $(M_D = 0)$	Hh $(M_D = d)$	HH $(M_D = 0)$
hh $(P_D = 0)$	0	$0.5d$	d
Hh $(P_D = d)$	$0.5d$	$0.5d$	$0.5d$
HH $(P_D = 0)$	d	$0.5d$	0

Note that when $P_D = M_D = 0$ the expected offspring dominance coefficient could be either 0 or d. This introduces ambiguities onto the expected offspring dominance, thus making the offspring dominance genetic component unpredictable using P_D and M_D alone. Therefore, the prediction of the offspring dominance component requires the parental genotypes, an inherently genomic information. This entails the requirement for estimation of dominance effect sizes of the markers.

Work on predicting dominance is sparse however; while dominant loci have been detected for some important traits (Billiard et al., 2021) Most of such studies focused on oligogenic traits. Attempts to detect QTL with dominance effects in a polygenic trait remained lacking. Due to this, in term of the practicality of optimization, a proxy that correlates with the dominance effect sizes would be desirable.

2.3.4. Optimizing the Contributions from Each Components

While the mathematical theories behind inbreeding, additive and dominance genetic components were relatively well-established, generalized theories on how to combine these components in an optimized selective breeding remained scarce. This is especially true for

the inclusion of dominance genetic component. In such case, these components can be combined using evolutionary algorithms, such as differential evolution used by Kinghorn (2000).

Despite this, there are different forms of evolutionary algorithms available, each suited for specific types of optimization problems (Gondro and Kinghorn, 2009; Slowik and Kwasnicka, 2020). For example, the differential evolution method is suited for optimization problem in a continuous search space (Storn and Price, 1993). The mating-specific nature of the dominance component means the sire permutations need to be considered during the optimization. As sire permutation is not a continuous quantity, this impedes the usability of differential evolution in the optimization of the dominance component. For this reason, other variants of evolutionary algorithms that can tackle combinatorial problems are needed, and one such variants is the genetic algorithm (Slowik and Kwasnicka, 2020).

Initially developed by Fraser (1957) as a simple simulator for genetic processes, the genetic algorithm has quickly been adopted by computer scientists as a general purpose problem solvers and optimizer (Mitchell and Forrest, 1994). This method has been used successfully to solve complex combinatorial problems with potentially infinitely large sample space such as Sudoku (Gerges et al., 2018), travelling salesman problem (Braun, 1991) and graph partitioning problem (Mühlenbein, 1992). Since optimization of the dominance component involves finding the optimal sire-dam mating configuration, an inherently combinatorial problem, genetic algorithms serves as a promising route for the optimization of this genetic component.

Besides combinatorial problems, genetic algorithms have also been successfully used in problems related to continuous search spaces (Haupt and Werner, 2007). This purpose of genetic algorithm is commonly used in engineering-related problems, such as optimal designs for hypersonic aircraft (Evans and Walton, 2017), microwave absorbing material (Jiang et al., 2009) and antennas for satellite missions (Lohn et al., 2008) and 5G communications (Marasco et al., 2022). The successes of genetic algorithm in optimizing problems with continuous search spaces hinted the possibility of its use in optimizing the combination of animals to be included in the selective breeding process to optimize additive genetic merit and build-up of co-ancestry, similar to the achievements attained by differential evolution in previously published OCS such as that by Kinghorn (2000). Unlike previously suggested mate allocation and sire contribution optimization methods, but similar to that of Toro and

Perez-Enciso (1990), a method that incorporates a genetic algorithm could resolves issues caused by "randomized contributions" of each sire and dams but relaxing the requirements of "equal contributions" from each sire and dams, thus promising a fully dynamic selection tactics for varying genetic architecture of a trait. Overall, genetic algorithms serves as a promising choice for the maximization of offspring's additive and dominance genetic component under a constraint of increment in inbreeding.

In its simplest form, a genetic algorithm starts by generating a population of candidate solutions. In the context of previously published OCS such as that in [1], the candidate solutions would be the vector of contributions of each sire toward the next generation (i.e. $\boldsymbol{x}$). The performance of each solution was evaluated using an objective function (i.e. the $f_{obj}(\boldsymbol{x})$ from equation [1]). From the population of candidate solutions, those with the top performance in term of the $f_{obj}(\boldsymbol{x})$ were selected to be propagated into the next iteration. The top solutions were subjected to the effects of "genetic operators," with the most common being "mutation," where the values in the solutions were replaced or adjusted, and "crossover," where part of the solutions within the population were exchanged. These altered solutions were then fed into the next iteration where the solutions were evaluated and selected again. These processes repeat up to the point of convergence, which for the OCS be defined as the point where subsequent iterations no longer yield a more optimal solution.

Despite the successes of genetic algorithm, there were several shortcomings for this method of optimization. One such shortcomings is its propensity to converge toward a local optimum. There were numerous potential reasons for such convergence, such as a suboptimal hyperparameters for the genetic operators (Heider and Drabe, 1997), a large sample space and a rugged fitness landscape for the optimization (Taherdangkoo et al., 2012).

Several modifications have thus been proposed to mitigate such shortcoming; one such modification is a parallelized genetic algorithm, where the algorithm is ran multiple times in attempt to extract the best solutions from multiple processes. This way, if one of the attempts converges toward a local optimum, there could be other attempts that converges toward the global optimum, thus increasing the chance of finding the latter (Baluja and Caruana, 1995; Mühlenbein, 1992). Another modification is an adaptive genetic algorithm, where the hyperparameters utilized were adjusted according to the performance of the offspring solutions. These adjustments of hyperparameters balance the exploration phase, where the solution space is searched for global optimum (but with the risk of disrupting an optimized

solution), with the exploitation phase, where the optimal solutions are extracted (but with the risk of premature convergence), thus improving the chance of encountering the global optimum (Srinivas and Patnaik, 1994). Heider and Drabe (1997) suggested a genetic algorithm that optimized the hyperparameter values, with these optimized values subsequently fed into another genetic algorithm that solves the actual problem.

Another shortcoming of genetic algorithms is its propensity of disrupting an already optimized solution. As the genetic algorithm approaches convergence, the solution population is on average performing better (thus more optimal) than the starting population, which also suggested these solutions are closer to an optimum. A genetic algorithm in its simplest form tends to disrupt such solutions, thus increasing the risk of missing the global optimum. This can be mitigated through elitist genetic algorithm, where the best parent solutions are propagated unaltered into the next generation among other offspring solutions. This reduces the chance of disrupting an already optimized solutions for the genetic algorithm, thus improving the chance of finding the global optimum (Baluja and Caruana, 1995).

By implementing these modifications onto the genetic algorithm, the chance of finding the global optimum for the optimized sire contribution and sire-dam mating configuration can be greatly improved. These improvements suggested a promising route for using a genetic algorithm in finding the exact configuration of sires and dams that would maximize both additive and non-additive genetic components under a predefined level of inbreeding.

2.4. The Power and False Positive Rate of GWAS

While GWAS has been used in detecting the QTL of a trait, there are several factors that could affect the power and false positive rate of a GWAS. Understanding these aspects of a GWAS could be important for an OCS that utilizes genomic information, such as assigning weights to individuals that have certain genotypic states in a marker. True positives allow correct assignment of QTL for the optimization process, whereas the false positives could mislead the optimization, causing the selection of poorly performing animals to be propagated into the next generation. Therefore, a reliable genomic information maximizes true positives and minimizes false positives, and this quality could be captured using the power and false positive rate of a GWAS. Thus, the effects of these factors on the power and false positive rate needed to be established.

2.4.1. The Ambiguous Definitions

Depending on how the terms are interpreted there are two different approaches in defining the "power" and "false positive rate" of a GWAS.

2.4.1.1. The Two Definitions of Power

The first definition of power was the probability of detecting a QTL given a set of parameters (i.e., true QTL effect size, allele frequency, phenotype variance and sample size). This definition has been used by Wang and Xu (2019), Spencer et al. (2009) and Chapman et al. (2003). Given a critical value for a threshold in a GWAS experiment z, the power of GWAS to detect a marker could then be defined as one minus the cumulative distribution function of the non-central chi-squared distribution with a non-centrality parameter (Wang and Xu, 2019). Wang and Xu (2019) suggested the following equation as a way of calculating a power of GWAS to detect a QTL k, given the estimated effect size $\hat{a}$:

$$power(k) \cong \int_{z}^{\infty} \chi^2_{nct}\left(t; 1, \frac{2p_j(1-p_j)\hat{a_j}^2(N-2)}{var(\mathbf{y}) - 2p_j(1-p_j)\hat{a}^2}\right) dt \qquad [20]$$

Where $\chi^2_{nct}(t; v, x)$ is the probability density function for a non-central chi-squared distribution with argument t with degree of freedom v and non-centrality parameter of x. Despite this, this equation assumes of having the QTL uncorrelated with one another, which for a highly polygenic trait might not be applicable.

The second definition of power, used by Klein (2007), Storey and Tibshirani (2003) and Shen and Carlborg (2013), was the number of QTL being detected by a GWAS out of all the QTL. This definition of power is based on the observation that many of the traits are polygenic in nature, thus a practical and utilizable GWAS would be the one that could detect as many QTL as possible. This is calculated using the following equation:

$$power(GWAS) = \frac{N_{TP}}{N_{TP} + N_{FN}} \qquad [21]$$

Where N_{TP} and N_{FN} are the number of true positives and false negative, respectively. One of the shortcomings for the second definition is that the calculation of N_{TP} and N_{FN} require the number of null and non-null markers, which its estimation is not trivial (some previously published methods, as well as their shortcomings, were detailed in Section 2.5). This definition also assumed independence between tests (i.e. markers), which is broken in an

actual GWAS as the dependency present itself as linkage disequilibrium between markers. The dependency between markers introduces ambiguity in the definitions between true or false positives or negatives (i.e. at what point of linkage disequilibrium between a QTL and a null marker shall become the boundary between "true positive" and "false positive" if the marker is declared as a positive). This ambiguity makes the second definition ill-defined unless such boundary can be defined, which its criteria for definition could warrant further studies.

2.4.1.2. The Two Definitions of False Positive Rate

Similarly, the first definition of false positive rate of a GWAS was defined as the probability of detecting a null marker. Under this definition, given a critical value for a threshold z the false positive rate could be evaluated as follows:

$$false\ positive\ rate(k) = \int_{z}^{\infty} \chi^2(t;1)\ dt \qquad [22]$$

where $\chi^2(t;v)$ is a chi-squared distribution with argument t and degree of freedom v. It can also be thought as the "power (from the first definition)" of GWAS if the true effect size being zero. This equation also assumes the markers being uncorrelated however, thus might not be directly applicable on a GWAS.

The second definition of false positive rate was the proportion of detected markers in a GWAS being a null marker:

$$false\ positive\ rate(GWAS) = \frac{N_{FP}}{N_{FP} + N_{TP}} \qquad [23]$$

Where N_{FP} is the number of false positives. This definition of false positives rate also suffered from the same shortcomings from the second definitions of power, namely the requirements for number of null and non-null markers, and the assumption for independence between markers. Compared to power of GWAS, less attentions have also been placed on factors that affect its false positive rate.

2.4.1.3. The Relationship Between the Two Definitions of Power and False Positive Rate

There are mathematical relationships between the first and second definitions for both terms. In terms of the power in the second definition, Klein (2007) defined this power (denoted as $power(GWAS)$) as the average power of detection for all the non-null SNP markers:

$$power(GWAS) = \frac{1}{N_{qtl}} * \sum_{k=1}^{N_{qtl}} power(k) \quad [24]$$

Where the $power(k)$ being the probability of detecting a QTL as defined in equation [20] and N_{qtl} being the number of non-null SNP markers of a trait.

The false positive rate under the second definition could also be defined in terms of power and false positive rate of the first definition:

$$false\ positive\ rate(GWAS) = \frac{(N_{snp} - N_{qtl}) * false\ positive\ rate(k)}{(N_{snp} - N_{qtl}) * false\ positive\ rate(k) + \sum_{k=1}^{N_{qtl}} power(k)} \quad [25]$$

with $false\ positive\ rate(k)$ being defined as the probability of detecting a null marker.

While these equations are theoretically sound, they are difficult to implement in practice. This is because the value of N_{qtl} is unknown. This calculation also relies on a perfect estimation of $power(k)$, which in turn requires a perfect estimation of $\hat{a}$ (i.e. $\hat{a} = a$). This causes difficulties in estimating power and false positive rate under this definition. For this project, focus would be placed on the second definition of power and false positive rate.

2.4.2. Factors that Affect the Power and False Positive Rate of GWAS

2.4.2.1. Sample Size

The factor that was most studied in the literature relates to GWAS sample size, with a general consensus of increased power with sample sizes (Spencer et al., 2009; Visscher et al., 2017). Despite this, a large range of sample sizes have been used, from less than 100 by Ren et al. (2016) up to more than 1 million by Jansen et al. (2019), and the proportion of true and false positives among the significant markers detected by these studies remained unclear.

There were also studies that went into the theoretical aspects of effects of sample sizes of the GWAS. Spencer et al. (2009) stated that in a case control design with purely additive model, the power is directly proportional to the sample N:

$$E(\chi_k^2) \propto N a_k^2 p_k (1 - p_k) R^2(Q, k) \quad [26]$$

Where χ_k^2 is the chi-squared test statistics of the marker k, a_k being the effect sizes of the QTL, p_k being the allele frequency of marker k, and $R(Q, k)$ being the linkage disequilibrium between QTL Q and marker k (Spencer et al., 2009).

Direct studies on the effects of sample size on the false positive rate in the context of GWAS are not as common. The general wisdom for any statistical tests, in which GWAS is a part of, is that increasing the sample sizes reduces the false positive rate (Forstmeier et al., 2017). Despite this, these statistical tests often have assumptions not applicable to GWAS, such as independence between tests (Gondro, 2015). An explicit experimentation of the effects of sample size on the false positive rate in GWAS would be desirable.

2.4.2.2. Genetic Architecture

Very few studies focused on the effect of genetic architecture on the power and false positive rate of GWAS. Most of the previous study, such as Hu et al. (2012) and Daetwyler et al. (2010), aimed at genomic prediction and the architecture's impact on accuracy of phenotypic prediction. Those that mentioned the effect of genetic architecture on the power of GWAS such as Gondro (2015) focus on the effects of polygenicity, with increased polygenicity reduces the power of GWAS. Gondro (2015) stated that the detection of QTL for polygenic traits requires larger sample sizes compared to oligogenic traits, especially for QTL with larger effect sizes.

2.4.2.3. Threshold of GWAS

The large number of markers used in a GWAS experiment constituted an unprecedented level of multiple testing, which increases the false positive rate (Gondro, 2015; Hayes, 2013). As an example, let z be the critical point for the threshold of GWAS. Assuming the markers are uncorrelated, the expected number of false positives can be defined as follows:

$$Number\ of\ False\ Positives = \left(N_{snp} - N_{qtl}\right) * \int_z^\infty \chi^2(t; 1)\ dt \quad [27]$$

If the threshold is set at point such that the Type 1 Error $\alpha_z = 0.05$, the expected number of false positives is $0.05 * (N_{snp} - N_{qtl})$, i.e., 5% of the number of null markers. If for example there are 10,000 null markers this threshold would produce 500 false positives. This phenomenon would further worsen if high density markers were utilized. For this reason, a multiple testing correction method should be employed in a GWAS experiment (Gondro, 2015; Hayes, 2013).

Several types of correction methods have been suggested for GWAS. One such method is the Bonferroni correction, popularized by Dunn (1961). The method calculates the expected Type 1 Error that needed to produce the same number of false positives as in one test. Using the aforementioned example, given $m = 10{,}000$ null markers, this correction attempts to find an α_z that would produce 0.05 false positives, the expected number of false positives for one test, instead of the original 500 false positives. Through proportionality, the expected α_z ($a_{z_{BON}}$) can be calculated as follows:

$$a_{z_{BON}} = \frac{a_z}{m} \quad [28]$$

The Bonferroni correction method is simple to implement and is effective in controlling the false positive rate (Wilson, 2019). It assumes independence between markers however, which increases the threshold stringency and reducing the power of GWAS, especially for a high-density genotype array (Hayes, 2013; Nishino et al., 2018; Wang et al., 2016).

The stringency of threshold from the Bonferroni method has led to the development of alternative methods of controlling multiple testing. One of the most popular class of methods was those that attempted to control the False Discovery Rate (FDR). Pioneered by Simes (1986) before popularized by Benjamini and Hochberg (1995). This method aims at controlling the false discovery rate of the multiple testings.

The method suggested by Simes (1986) is defined as follows: given a level of false discovery rate α, let $pv_1, pv_2, pv_3, \dots, pv_m$ be a list of p-values that have been ordered from the most significant to the least significant that test the following set of null hypothesis $H_0 = \{H_1, H_2, H_3, \dots, H_m\}$, the critical point pv_j is defined as the last p-value that fulfil the following inequality (Simes, 1986):

$$pv_j \leq \frac{j\alpha}{m} \quad [29]$$

for j ranges from 1 to m. This method is known to have a more lenient threshold (i.e., the maximum stringency is equivalent to that of the Bonferroni method, i.e. $j = 1$). Despite this, it also assumes independence between markers in GWAS.

Since Simes, several FDR-based methods have been suggested. Benjamini and Yekutieli (2001) suggested an adjustment being made at the denominator of equation [29] to take into account the effects of dependencies between tests. Storey and Tibshirani (2003) has criticized the original method proposed by Benjamini and Hochberg (1995) as being overly conservative, and thus introduced an adjustment parameter based on the proportion of null compared to all markers. This method also assumes independent or weakly dependent markers. Efron et al. (2001) introduced the concept of "Local False Discovery Rate" (LFDR) which is defined as the expected FDR within a bounded interval of p-values. Extending the LFDR model, Broberg (2005) has introduced a "Pooling of Adjacent Violators" (PAVA) based FDR method, with the assumption that the LFDR is monotonic.

Despite the wealth of multiple testing correction methods, many of these methods do not take into account numerous factors that might affect the optimality (i.e., increase the power of GWAS while constraining the false positive rate) of the threshold. As an example, Pryce et al. (2010) suggested increased stringency of the FDR-based threshold with the use of the single SNP regression method compared to those utilizing haplotype-based methods, reducing the power in the former method. Hong and Park (2012) and Nishino et al. (2018) suggested an increase in genotyping density also increases the sample sizes required to achieve the same power of GWAS. Ioannidis (2007) criticized these methods as they ignored the effects of population stratification and ratio between true and null markers on the threshold. There are also studies that criticized the increased false positive rate of FDR-based methods (Huang et al., 2018; Shen and Carlborg, 2013). None of the previous works have tested the optimality of these thresholds under changing genetic architecture parameters.

2.5. Estimating the Genetic Architecture Parameters

Given the potential impacts of genetic architecture on the power and false positive rate of a GWAS, it might be imperative to estimate the parameters for the genetic architecture of a trait, which includes the number of QTL and the distribution of the QTL effect sizes. Using the estimated genetic architecture, we can calculate the optimal thresholds for a GWAS while taking into account the effects of genetic architecture, from which the number of true

positives can be maximized and that of false positives be minimized, and thus reducing the chance of having the optimization misled by the latter.

Previously proposed methods of estimating the genetic architecture parameters tend to follow a generalized framework of estimating the admixture proportion between null and non-null markers based on a certain presumed parameterized model of QTL effect size distribution (Cheng et al., 2020; Zhang et al., 2018, 2021; Zeng et al., 2017). As an example, Cheng et al. (2020), Lloyd-Jones et al. (2019) and Zhang et al. (2018) utilized Expectation-Maximization (EM) to estimate the admixture proportion, with the assumption the effect sizes distribute according to a normal distribution. Zhang et al. (2021) also proposed a set of similar methods that assumed fixed mixtures of normal distributions and a double exponential distribution. O'Connor (2021) utilizes the characteristic function for a mixture of 13 normal distributions. Park et al. (2010) fitted the previously published GWAS results using exponential and Weibull distributions, and Hall et al. (2016) calculated the number of QTL directly from the proportion of genetic variance explained by the markers and the heritability, with the assumption that the QTL effect sizes follow an exponential distribution. Many Bayesian-based methods such as that by Meuwissen et al. (2001) and Moser et al. (2015) also assumed a normal distribution or a mixture of normal distributions for the modelling of QTL effect sizes.

The use of presumed parameterized models for QTL effect size distribution from all aforementioned approaches have been criticized by Zeng et al. (2017) for having a restrictive shape of the normal or exponential distribution, which could cause failure in capturing the shape of the effect size distribution and led to a reduced accuracy and robustness in estimated effect size distribution. Indeed, both normal (or a fixed number of mixtures of normal) and exponential distributions have fixed kurtosis, which means they might not be able to capture the shape of the tail of the QTL effect size distribution (Mun, 2012). For this reason, Zeng et al. (2017) proposed a nonparametric prior for the variances of an infinite number of mixtures of normal distributions (although in practice still with a fixed number of normal distributions for computational reasons), which produces a more flexible shape of QTL effect size distribution.

Despite its improved flexibility, one assumption for Zeng et al. (2017) is an infinitesimal QTL model where all the SNP markers have nonzero effect sizes, which might not be suitable for an oligogenic trait. Indeed, several authors argued against the infinitesimal model both on

theoretical ground (for example Hill (2010) and Orr (1999)) and on empirical ground (for example Moser et al. (2015) and Orr (1999)). While methods such as Moser et al. (2015) assume a finite QTL model, the use of a fixed number of normal distributions could restrict its flexibility and thus its robustness against changing genetic architectures. Methods that combine flexibility of QTL effect size distribution and assumption for QTL models, be it the infinitesimal model or finite QTL model, remain lacking, and this is an avenue worth further studying.

Besides the inflexibility in distribution and assumptions on QTL models, several other limitations have been identified in the literature. One such limitations is the requirement of arbitrary, user-defined thresholds in these methodologies. For example, the method by Cheng et al. (2020) requires a user-defined input for the null – non-null SNP marker threshold, which can reduce the performance of the methodology if such a threshold is mis-specified. Park et al. (2010) utilized a "trivial effect size" threshold where the QTL with effect sizes smaller than the threshold were excluded from estimation, and Zhang et al. (2018) requires the use of a linkage disequilibrium threshold and a pre-specified linkage disequilibrium window size. The choice of these values could affect the optimality of the genetic architecture parameter estimations. Another limitation in the method by Park et al. (2010) is the requirement for previously published GWAS, which might reduce its usability for a newly studied trait.

Many of the previously published methods were tested on extremely large sample sizes. For example, Cheng et al. (2020) tested methodology on simulated sample sizes of 5000 and 10,000, and Park et al. (2010) used 13,532 human sample for Crohn's disease and 63,000 for height. Methods for constructing a genetic architecture, such as that suggested by Cheng et al. (2020) and Park et al. (2010) relies on the ability to differentiate between noises and signals, and hence requiring the use of significance thresholds, these methods might behave differently for smaller dataset. Indeed O'Connor (2021) acknowledged the vulnerability of their method toward small sample sizes. Given that many GWAS in livestock have been conducted at relatively smaller sample sizes of less than a few thousand phenotypes, the previously published methods might not be as reliable in these cases. Methods that could handle a smaller sample sizes (i.e. sample sizes comparable to GWAS in livestock production) warrant further studies.

Another important aspect worth considering is the effect of linkage disequilibrium structures. Unlike human samples, which have relatively short segments of homozygosity and strong decay in their linkage disequilibrium (Gibson et al., 2006), the strong selection regimes and small effective population size in livestock produce long tracts of homozygosity and extended blocks of linkage disequilibrium in the genome, e.g. as demonstrated in sheep (Al-Mamun et al., 2015a; Kijas et al., 2014), cattle (Porto-Neto et al., 2014; Purfield et al., 2012) and horses (Jasielczuk et al., 2020). Differences in linkage disequilibrium structures could have a large impact on the performance of themethods that estimate the genetic architectures of livestock traits. Indeed, Zhang et al. (2018) assumes the effect sizes are independent on the local linkage disequilibrium structures, and Lloyd-Jones et al. (2019) commented on the negative effects of linkage disequilibrium structures on the convergence of the model. These assumptions could affect the utility of these methods to provide a model for genetic architecture of livestock traits.

Finally, Lloyd-Jones et al. (2019) also commented on the infeasibility of identifying the true underlying mixture distribution of the QTL effect sizes due to linkage disequilibrium. They referred to cases where a significant region captured by a GWAS could either be caused by one causal variant with a large effect, or numerous causal variants with smaller effects that were in linkage disequilibrium. This ambiguity would subsequently affect the estimation of the mixture distribution parameters. Methods to handle such ambiguity warrant further studies.

2.6. Direction for the Project

The aim of this project is to develop a method that optimizes the breeding pairs of sires and dams with the use of evolutionary algorithm such as genetic algorithm. This method would likely be OCS-like method that maximize the additive and dominance genetic component while constraining the level of inbreeding coefficient increment. Emphasis could be placed on the use of genomic data in the optimization of the breeding pairs.

From the literatures, the additive genetic component of the offspring can be calculated using animal-based data such as EBVs and genomic data from a GWAS. The dominance genetic component can only be estimated through genomic data however and given the difficulty of estimating the dominance effect sizes of the markers, a proxy based on genomic data will be used. The objective function for the OCS might also need to be modified from [1] to take into account the mating-specific nature of the dominance component. The optimization of

epistatic component would not be emphasized in this study due to the difficulty of obtaining an estimate for this component (Lynch and Walsh, 1998; Vitezica et al., 2018).

While the effects from some of the factors such as sample size on the power and false positive rate of the GWAS has been widely reported, effects from other factors, such as those pertaining to the genetic architecture, remain largely elusive. This suggests additional work needs to be done to ascertain the effects of various factors on the power and false positive rate of the GWAS. Findings from this investigation could then be incorporated into techniques that could improve the power and false positive rate of the GWAS before incorporating them into the OCS. This could include the establishment of an optimal threshold for the GWAS-based results, and the estimation of genetic architecture parameters such as number of QTL and the distribution of their effect sizes. Preferably, these techniques could be developed in a manner that suits a livestock GWAS-sized dataset while taking into accounts effects from confounding factors such as linkage disequilibrium structures.

This project is important as it can potentially increase the accuracy of selection, concentrating the economically beneficial alleles in the breeding stock, while exploiting the non-additive components such as the dominance effects.

Chapter 3. Effects of Experimental Design, Genetic Architecture and Threshold on Power and False Positive Rate of GWAS

Zhi Loh, Julius H. J. van der Werf, Sam Clark

3.1. Abstract

Genome-Wide Association Studies are an important tool for identifying genetic markers associated with a trait, but it has been plagued by the multiple testing problem, which necessitates a multiple testing correction method. While many multiple testing methods have been suggested, e.g., Bonferroni and Benjamini-Hochberg's False Discovery Rate, the quality of the adjusted threshold based on these methods is not as well investigated. The aim of this study was to evaluate the balance between power and false positive rate of a Genome-Wide Association Studies experiment with the Bonferroni and Benjamini-Hochberg's False Discovery Rate multiple testing correction methods and to test the effects of various experimental design and genetic architecture parameters on this balance. Our results suggest that when the markers are independent the threshold from the Benjamini-Hochberg's False Discovery Rate provides a better balance between power and false positive rate in an experiment. However, with correlations between markers the threshold of the Benjamini-Hochberg's False Discovery Rate becomes too lenient with an excessive number of false positives. Experimental design parameters such as sample size and number of markers used, as well as genetic architecture of a trait affect the balance between power and false positive rate. This experiment provided guidance in selecting an appropriate experimental design and multiple testing correction method when conducting an experiment.

3.2. Introduction

Since high-density genotyping arrays using abundant genetic markers such as Single Nucleotide Polymorphisms (SNPs) have become available, Genome-Wide Association Studies (GWAS) has become an important tool in gene discovery (Wang and Xu, 2019). Hundreds of thousands to several millions of genetic markers can now be used in association studies, where the aim is to estimate and test the effect of genetic variants linked to a

Quantitative Trait Locus (QTL). This has provided a huge research opportunity but the use of large numbers of markers to be tested has also introduced a multiple testing problem of an unprecedented scale. Multiple testing significantly increases the number of false positives when using a standard significance threshold, thus necessitating the use of a correction method to adjust this threshold (Tam et al., 2019).

The most popular method of controlling the number of false positives is the the Bonferroni correction. This multiple testing correction method is based on the joint distribution of all the Student's t-distribution for each individual linear contrast, with the assumption that each of these tests are independent to one another (Dunn, 1961). This method had gained popularity due to its simplicity (Llinares-López et al., 2015; Ionita-Laza, Cho and Laird, 2012), and is considered one of the most effective methods in controlling the number of false positives (Wilson, 2019). However, the Bonferoni method has also been criticized when applied to GWAS as with very large numbers of SNPs tested, it has been perceived as being overconservative, leading to reduced power in identifying causal variants (Gao, Becker, Becker, Starmer and Province, 2010; Huang, Ritchie, Brozynska and Inouye, 2018; Llinares-López et al., 2015; Wilson, 2019). The situation has been further exacerbated by decreasing cost of genotyping, and it now has become common practice to use all genetic variants obtained from Whole Genome Sequence (WGS) information, often exceeding 25 million marker genotypes per sampled individual (Huang et al., 2018; Tam et al., 2019; Visscher et al., 2017).

Alternative multiple testing correction methods have been introduced, many of which have reduced stringency. One class of alternatives is those methods that attempt to control the False Discovery Rate (FDR), with one of the most popular methods being the Benjamini-Hochberg's False Discovery Rate (BH-FDR) method. Initially introduced by Simes (1986), this method aims at testing the ranked p-values against a stepwise threshold that varies based on the rank of the p-values, with the most significant p-value subjected to the most stringent threshold, and other p-values that are less significant are subjected to more lenient threshold. An example of its implementation is provided in Table 3.1.

Table 3.1: An example of implementation of Benjamini-Hochberg's False Discovery Rate (BH-FDR). For this example, 10 SNPs were tested and have their p-values calculated according to their rank j. The point where the p-value of the marker falls below of that calculated from the stepwise threshold is at $j = 4$ is, and this is the point where the threshold of the BH-FDR is set. Note that only the most significant marker (i.e. $j = 1$) had been subjected to the stepwise threshold equivalent to a Bonferroni correction.

Index of ranked p-values (j)	Ranked p-values	Stepwise Threshold $\left(\frac{0.05*j}{nsnp}\right)$	Decision (Accept or Reject Null)
10	0.676	0.050	Accept
9	0.324	0.045	Accept
8	0.213	0.040	Accept
7	0.119	0.035	Accept
6	0.087	0.030	Accept
5	0.034	0.025	Accept
4	0.012	0.020	Reject
3	0.010	0.015	Reject
2	0.006	0.010	Reject
1	0.002	0.005	Reject

Many GWAS have chosen the BH-FDR multiple testing correction method on the grounds of overconservativeness of the Bonferroni correction but appeared to have no consideration on the possibility of increased false positive rate. In the context of gene expression analysis, Huang et al. (2018) considered BH-FDR to have a better balance between power and false positive rate, although they also commented that the use of the BH-FDR resulted in an inflated false positive rate whereas Bonferroni correction had a significantly lower number of false positives. Another consideration in most GWAS is that with the use of dense markers, marker genotypes can be highly correlated. Benjamini and Yekutieli (2001) suggested that in theory this method is valid even when the assumption of independence between tests is violated, as would be the case in GWAS based on dense marker genotypes. An actual study

on the ability of the BH-FDR in controlling the false positive rate in a GWAS and the need to account for the lack of independence between tests is lacking, however.

Several factors could impact the success in detecting QTL associated with a trait while controlling the false positive rate, including parameters related to the genetic architecture of the traits, i.e., the size of the QTL effects, and experiment design, most notably the sample size. Spencer et al. (2009) and Visscher et al. (2017) argued a reduced power of GWAS with small sample size, while Forstmeier et al. (2017) argued an increased false positive rate with small sample size in any statistical test. The low power alongside with increased false positive rate could have contributed to the low replicability of a GWAS experiment where hits from the previous studies failed to be replicated in subsequent studies (Heller and Yekutieli, 2014; Wang and Zhu, 2019). Spencer et al. (2009) and Visscher et al. (2017) argued that increasing the sample size is the most effective way to increase the power of GWAS, and while the number of positives increases with sample size, it is unclear how much of the positives are true positives (a summary of number of positives reported in previous publication is provided in Table 3.2).

The aim of this study is to test the effects of multiple testing correction methods on the power and false positive rate of a GWAS experiment, and subsequently evaluate the effects of experimental design parameters and genetic architecture of a trait on the suitability of the methods. We use simulation to evaluate the power and false positive rate with Bonferroni and BH-FDR correction methods under varying parameter values.

3.3. Method

The effects of the GWAS parameters and multiple testing correction methods were evaluated using simulated genotypes and phenotypes. This simulation is conducted using Python (version 3.7.3).

To simulate a GWAS experiment with independent markers, data from a genotype array with M markers (henceforth denoted as $\boldsymbol{X}$) was generated for N individuals. The distribution of the allele frequency of the markers following a symmetrical Beta distribution (i.e. $Beta(\beta, \beta)$). Values used for the shape parameter β from the beta distribution are provided in Table 3.3. Some of the markers were nominated as QTL, with their effect sizes (in units of σ_e) distributed based on the following gamma distribution:

$$QTL\ Effect\ Size\ (\boldsymbol{a}) \sim gamma(shape\ parameter, scale\ parameter) \quad [1]$$

Table 3.2: Summary of threshold used, sample size, and number of markers and positives used in previous publications. For studies that included multiple traits, data from only one trait was included. BH-FDR stands for the Benjamini-Hochberg False Discovery Rate method, and BON for the Bonferroni method. The publications are ranked based on the sample size used.

Publications	Sample Size	Number of Markers	Correction Method	Alpha	Threshold (-log10(THR))	Number of positives
Ghasemi et al. (2019)	130	41,323	BON	0.05	5.91	7
Zhang et al. (2013)	319	48,198	BON	0.05	5.98	10
Vanvanhossou et al. (2020)	449	32,518	BON	0.05	5.81*	4
Signer-Hasler et al. (2012)	1077	38,124	BON	0.05	5.88	8
Xia et al. (2017)	1141	677,855	BON	1.0	5.83	11
Chang et al. (2018)	1217	671,990	BON	0.05	7.12	11
Al-Mamun et al. (2015b)	1449	48,640	BON	0.01	6.69	39
Weerasinghe et al. (2019)	3454	37974	BON	0.05	5.88	13
Cai et al. (2019)	5373	16,503,508	BON	0.05	8.52*	58535
Dakhlan et al. (2017)	6463	48,599	BON	0.01	6.69	17
Yin and König (2019)	13,827	54,613	BON	0.05	6.04	10
Jiang et al. (2019)	294079	57,067	BON	0.005*	7.00	15215
Smołucha et al. (2021)	155	49,204	BH-FDR	0.05	5.99	1
Wang et al. (2017b)	880	51,727	BH-FDR	0.01	4.00	5
Steri et al. (2019)	946	135,992	BH-FDR	0.08	5.84	5
An et al. (2020)	1,217	67,192	BH-FDR	0.01	6.17*	45
Ibeagha-Awemu et al. (2016)	1,246	76,355	BH-FDR	0.1	4.16*	53
Pegolo et al. (2020)	1,369	23,173	BH-FDR	0.05	4.30	24
Akanno et al. (2018)	5,324	42,536	BH-FDR	0.10	3.16*	294

* Values back-calculated using available data (i.e. alpha, number of SNPs, number of positives)

The $scale\ parameter$ for the gamma distribution in [1] is set at 1.0 for all simulations, and the $shape\ parameter$ is varied based on the $Average\ QTL\ Effect\ Size$, which is provided in Table 3.3. The average QTL effect size is calculated as follows:

$$Average\ QTL\ Effect\ Size = \frac{Shape\ Parameter\ for\ Gamma\ Distribution}{Scale\ Parameter\ for\ Gamma\ Distribution} \quad [2]$$

The $Average\ QTL\ Effect\ Size$ was specified in units of σ_e .With the aforementioned $scale\ parameter$, the $Average\ QTL\ Effect\ Size$ in [2] equates the $Shape\ Parameter\ for\ Gamma\ Distribution$. Markers that were not nominated as QTL would have their effect sizes marked at 0. Using the vector containing the effect sizes for all markers and QTL (denoted as $\boldsymbol{a}$), the additive genetic component of the phenotype (denoted as $\boldsymbol{g}$) is calculated as follows:

$$\boldsymbol{g} = \boldsymbol{Xa} \quad [3]$$

The residual component of the phenotype (denoted as $\boldsymbol{e}$) is then simulated using the variance of vector $\boldsymbol{g}$ and the narrow sense heritability of the trait h^2. The residual component follows a normal distribution with mean of zero and variance as follows:

$$Var(\boldsymbol{e}) = Var(\boldsymbol{g}) * \left(\frac{1 - h^2}{h^2}\right) \quad [4]$$

For all the parameter under study, the heritability was set at 0.3. The vector $\boldsymbol{g}$ and vector $\boldsymbol{e}$ were then summed to obtain the simulated phenotype of the individuals. A GWAS was then conducted using the genotype array and phenotype vector. Single SNP regression was used to estimate the effect sizes of the markers, which would then be used to calculate the p-values for each marker using the Student's t-test.

Using the alpha = 0.05 for type 1 error, the thresholds from both Bonferroni correction and BH-FDR were calculated. The threshold for the Bonferroni correction is defined as alpha divided by number of markers used in the experiment:

$$Threshold\ for\ Bonferroni = -log_{10}\left(\frac{0.05}{Number\ of\ SNPs}\right) \quad [5]$$

For the BH-FDR in this experiment, the threshold is defined as follows (Simes, 1986; Benjamini and Yekutieli, 2001):

$$Stepwise\ Threshold\ for\ BH-FDR = -\log_{10}\left(\frac{0.05 * k}{Number\ of\ SNPs}\right) \quad [6]$$

where k is the point where the k_{th} ranked -log(p-value) of the GWAS becomes larger than the stepwise threshold. The point k is equivalent to the $j = 4$ from the example in Table 3.1. With these thresholds, the power and false positive rate, as well as differences between true and false positives (denoted as $Receiver\ Operating\ Characteristic\ (ROC)\ score$), were calculated. The power is defined as follows:

$$Power = \frac{Number\ of\ True\ Positives}{Total\ Number\ of\ QTL\ used\ in\ Study} \quad [7]$$

For the calculation of power only the QTL with effect size exceeding 0.1 σ_e were taken into account. The false positive rate is as follows:

$$False\ Positive\ Rate = \frac{Number\ of\ False\ Positives}{Number\ of\ True\ Positives + Number\ of\ False\ Positives} \quad [8]$$

And the $ROC\ score$ is defined as follows:

$$ROC\ score = Number\ of\ True\ Positives - Number\ of\ False\ Positives \quad [9]$$

In this study the $ROC\ score$ was used as a measure to test the capability of a threshold in balancing the power and false positive rate of a GWAS. This is equivalent to the weighted Youden's Index as described by Habibzadeh et al. (2016), who have utilized a Receiver Operating Characteristic (ROC) curve to establish the optimal threshold for clinical diagnostic tests.

A multiple testing correction method with its threshold having a high $ROC\ score$ was considered as capable of providing a better balance between power and false positive rates. A threshold with maximum $ROC\ score$ was considered as optimal. This is equivalent to having the point on the ROC curve where the tangent of the curve equals to 1, which has been demonstrated mathematically by Kaivanto (2008). The experiment was then repeated 200 times for each combination of parameter values.

To test the effect of correlations between marker genotypes on the optimal threshold and number of true and false positives, the experiment is repeated with pairwise marker linkage disequilibrium (denoted as r^2) set at 0.8. This is achieved by copying the haplotype state of some of the alleles from one locus to its neighbouring locus while randomizing the haplotype state of other alleles, thus generating a genotype array with a controlled level of pairwise

marker linkage disequilibrium. For correlated markers, besides the true positives (denoted as TP), there were two types of false positives to be identified: (i) correlated false positives (henceforth denoted as FPC), defined as the false positives that its r^2 exceed 0.1 with one or more true QTLs, and (ii) uncorrelated false positives (denoted as FPU), defined as false positives that had its r^2 below 0.1 with any of the QTLs. For the calculation of false positive rate, the number of FPU is used in place of number of false positives in equation [6], and for $ROC\ score$, the number of FPU is used in equation [7].

A list of parameters and value tested is provided in Table 3.3. When a parameter is under study, default values were used of other parameters.

Table 3.3: Parameters tested in this study.

Parameters	Default Value	Alternative Values
Sample Size	2000	200, 800, 1400, 3000, 5000
Shape parameter for Distribution of Allele Frequencies (β)	0.5	0.1, 0.2, 0.3, 0.7, 1.0
Average QTL Effect Sizes (γ)	0.4	0.1, 0.2, 0.3, 0.7, 1.0
Number of Markers	20k	5k, 10k, 40k, 60k, 80k
Number of QTLs	100	20, 50, 300, 600, 1000

The number of QTL was arbitrarily chosen based on the proportion of positives markers out of all the markers in previous studies cited in Table 3.2. The sample sizes used are based on those used by previous studies as cited in Table 3.2.

Besides the parameters listed in Table 3.3, the combined effects of sample size and number of markers on the power and false positive rate of GWAS were also tested. To test the combined effects of both parameters, additional simulations on variable sample sizes have been conducted with number of markers of 5k, 20k and 80k. The sample sizes used in this additional simulation are the same as those provided in Table 3.3. This additional simulation is to test the power, false positive rate and suitability of the correction methods for a GWAS experiment that involves small sample size but large number of markers, as in Steri et al. (2019) that have conducted a GWAS with 946 animals but with 135,992 markers.

3.4. Results

3.4.1. Parameters determining the threshold of multiple testing correction methods

Number of markers and sample size. The threshold from both multiple testing methods is influenced by the number of markers used in GWAS. With an increased number of markers used in GWAS, the threshold increases in stringency. This observation was made in both multiple testing correction methods in both independent and correlated marker system.

When the Bonferroni correction is used, sample size does not have any effect on the threshold of GWAS. This is not the case for BH-FDR however, as the threshold from the BH-FDR is significantly affected by sample size, with larger sample sizes decreasing the threshold. Generally, the threshold calculated by the BH-FDR is less stringent than those calculated by Bonferroni correction (Figure 3.1).

Number of QTL and QTL effects. The number of QTL does not have any influence on the threshold calculated from the Bonferroni correction. The number of QTL has an effect on the threshold of the BH-FDR, however. When the number of QTL is small (e.g., 20) the threshold from the BH-FDR approaches 4.9, and this threshold declines slightly to 4.63 with a number of QTL of 100, but then increases again gradually with larger numbers of QTL (Figure 3.2(a)). A smaller average QTL effect sizes also increases the threshold slightly for the BH-FDR (Figure 3.2(b)). The allele frequency distribution does not have an effect on both multiple testing correction methods (Figure 3.2(c)).

Correlation between marker genotypes: For Bonferroni correction, correlation between markers does not have any effect on the threshold for any of the parameters tested. For BH-FDR however, marker correlation has a significant effect on the threshold. Correlation between markers significantly decreases the GWAS threshold (Figure 3.2). With independent markers, the number of markers and sample size also have significant effects on the threshold of BH-FDR for correlated markers. While the trend is comparable with those in independent markers, the threshold calculated by BH-FDR is lower with correlated markers compared to independent markers for all marker numbers and sample sizes tested. Correlations between markers also caused a similar decline in the BH-FDR threshold for all numbers of QTL, average QTL effect sizes and allele frequency distributions tested in this experiment.

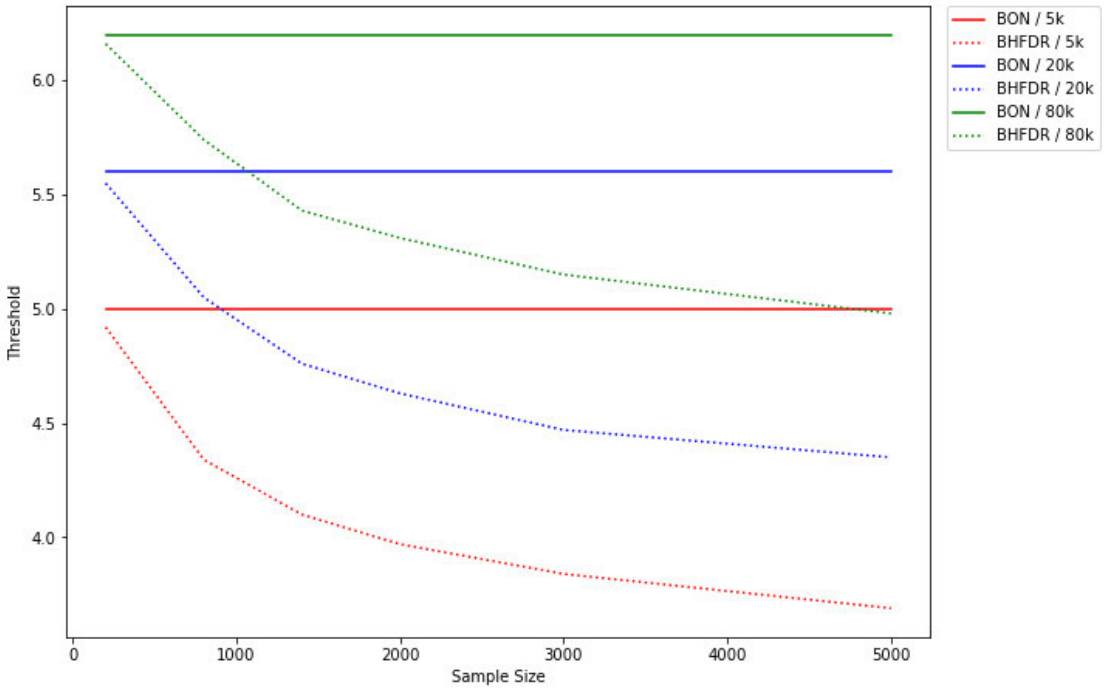

Figure 3.1: Threshold of the Bonferroni correction (in solid lines) and BH-FDR (in dashed lines) under varying sample size and number of markers used in a GWAS experiment. This is the threshold under independent markers. The number of QTL maintained at 100, and the average QTL effect sizes (γ) and allele frequencies (β) are at 0.4 and 0.5, respectively.

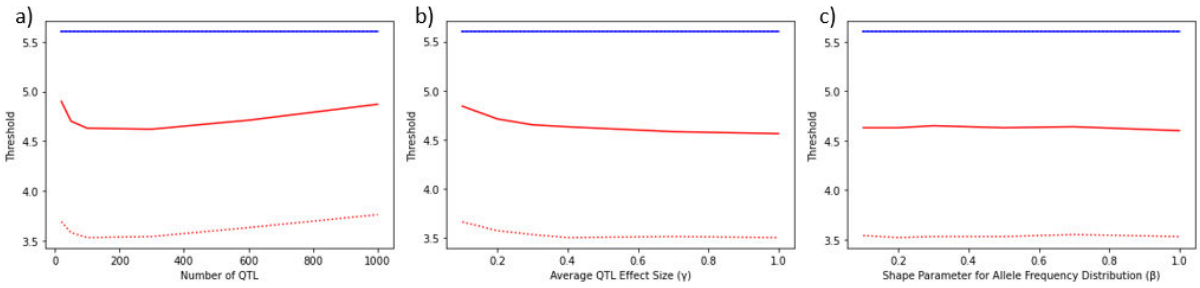

Figure 3.2: The effects of (a) number of QTL, (b) average QTL effect size (γ) and (c) shape parameter for allele frequencies distribution (β) on the threshold for the Bonferroni correction (blue line) and BH-FDR (red lines) for both independent (solid lines) and correlated (dashed lines) markers. The default parameters for each of the plots are as follows: number of QTL at 100, sample size 2000, the number of markers 20k, the average QTL effect size (γ) at 0.4 and shape parameter for allele frequencies distribution (β) at 0.5. The threshold for Bonferroni correction for independent markers fully overlaps with that for correlated markers in this figure, thus indistinguishable from one another.

3.4.2. Parameters determining the power of GWAS

Number of markers and sample size: Due to an increased stringency in threshold from both multiple testing correction methods, the power decreases with an increased number of SNP

markers used in a GWAS experiment. This observation was made for both independent markers and correlated markers. While correlations between markers increased the power for all marker number values tested, such an increase is more significant for experiments with a small number of markers, or when BH-FDR is used in GWAS (Figure 3.3(a)).

Increasing the sample size increases the number of true positives and the power of GWAS, and this increase is more significant when BH-FDR is used. Correlation between markers has no effect on the power of GWAS if Bonferroni correction is used, but significantly increases the power for BH-FDR. This is attributable to an increased leniency in the threshold for the BH-FDR with larger sample size (Figure 3.3(b)).

Number of QTL and QTL effects: The number of QTL that is associated with a trait has a significant effect on the power of detecting the QTL, with the power decreasing when the number of QTL increased, both for independent and correlated markers. This was observed for both multiple testing correction methods, although the power is higher for correlated markers when BH-FDR is used in the GWAS (Figure 3.3(c)).

The average QTL effect sizes (γ) has significant effects on the number of true positives and power of GWAS. With an increased value of γ, the number of true positives increases until it starts to plateau by average QTL effect size of 0.4. In all cases, BH-FDR had a higher number of positives. Correlation between markers also increases the number of true positives for both multiple testing correction methods, and this increment is more significant for BH-FDR (Figure 3.3(d)). The shape parameter for allele frequency distribution (β) has again no effect on the number of true positives and power of GWAS.

3.4.3. Effect of Parameters on False Positive Rate of GWAS

Number of markers and sample size: Despite the increasingly large number of tests needed to be conducted in a GWAS experiment with a larger number of markers, due to the increasingly stringent threshold, the raw number of false positives declines logarithmically. This is observed in both FPU and FPC. Due to a lower number of true positives associated with a more stringent threshold, however, the false positive rate is no longer significantly affected by the number of markers used in a GWAS experiment (Figure 3.4(a)).

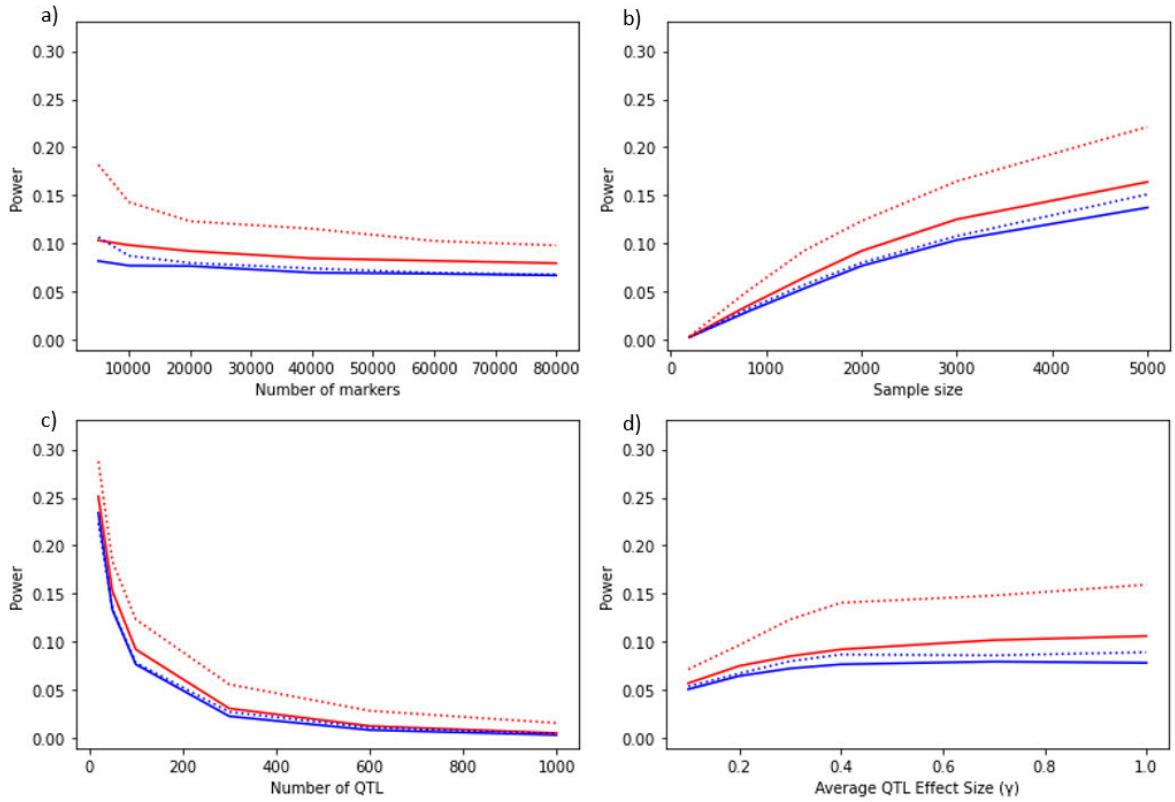

Figure 3.3: The effects of (a) number of markers, (b) sample sizes, (c) number of QTL and (d) average QTL effect size (γ) on the power of GWAS for the Bonferroni correction (blue line) and BH-FDR (red lines) for both independent (solid lines) and correlated (dashed lines) markers. The default parameters for each of the plots are as follows: number of QTL at 100, sample size 2000, the number of markers 20k, the average QTL effect size (γ) at 0.4 and shape parameter for allele frequencies distribution (β) at 0.5.

Unlike marker number, sample size has a significant effect on the false positive rate of a GWAS experiment (Figure 3.4(b)). The false positive rate increased significantly when the sample size is small (i.e., N=200). This trend was observed for both independent and correlated markers, and in both multiple testing correction methods. With larger sample size, the false positive rate remained relatively constant if the markers are independent. This is not the case for correlated markers however; the number of FPU increased significantly with larger sample sizes, and that led to an increase in false positive rate. While this was observed for both multiple testing correction methods, the false positive rate for the BH-FDR is higher for all sample sizes tested in this experiment. While the number of markers does not have an effect on the false positive rate of a GWAS experiment under the default sample size (i.e., N=2000), for small sample size (N=200) a larger number of markers also strongly increased the false positive rate (Figure 3.5).

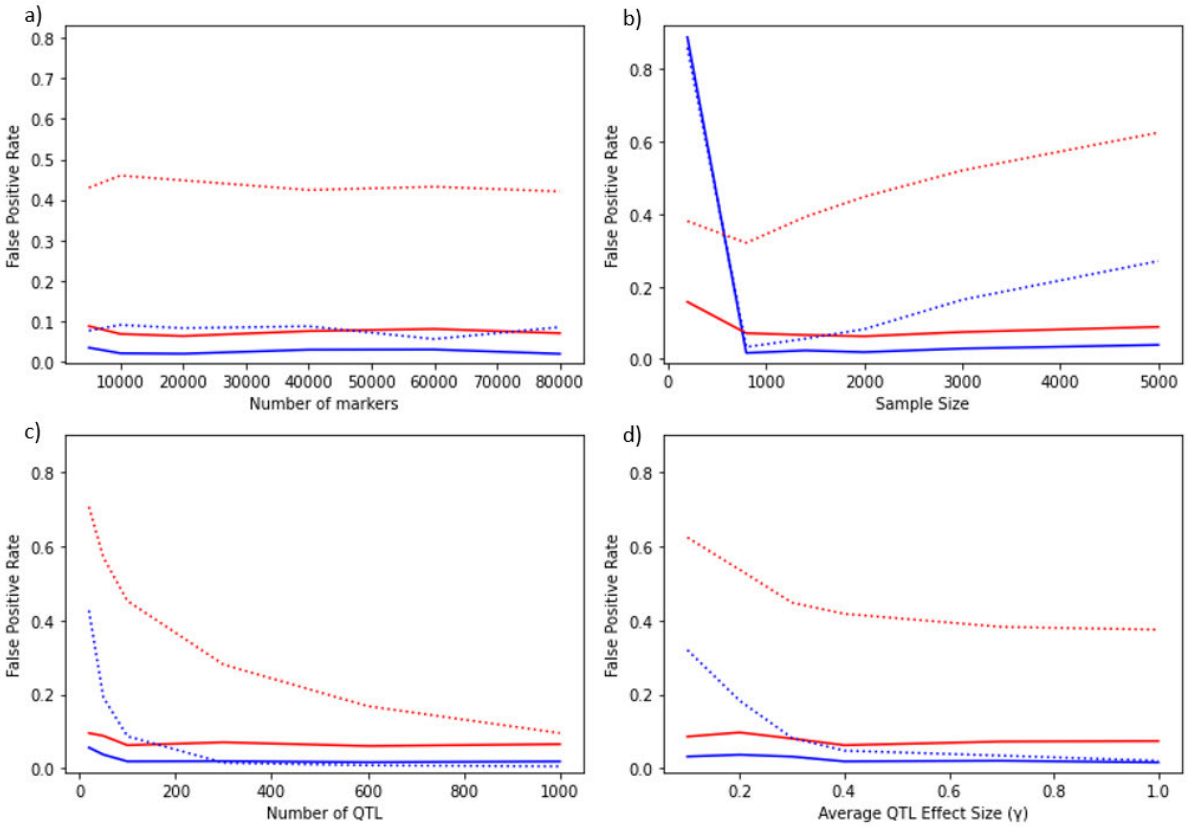

Figure 3.4: The effects of (a) number of markers, (b) sample sizes, (c) number of QTL and (d) average QTL effect size (γ) on the false positive rate of a GWAS for the Bonferroni correction (blue line) and BH-FDR (red lines) for both independent (solid lines) and correlated (dashed lines) markers. The default parameters for each of the plots are as follows: number of QTL at 100, sample size 2000, the number of markers 20k, the average QTL effect size (γ) at 0.4 and shape parameter for allele frequencies distribution (β) at 0.5.

Number of QTL and QTL effects: For independent markers, the false positive rate of a GWAS is not influenced by the number of QTL associated with a trait. This is not the case for correlated markers however; traits with small number of QTL with large effect sizes have a higher false positive rate compared to traits with large number of QTL with small effect sizes, and correlation between markers exacerbated that increment (Figure 3.4(c)). This is caused by an increase in raw number of FPU and a decrease in raw number of true positives (as there is less QTL to be detected in the first place). The increase in the number of false positives with a small number of QTL is due to an increase in significance from the increased proportion of variance explained by those QTL (which also explained the increase in power of GWAS with a small number of QTL). This increase in significance at the QTL also increases the significance of neighbouring null markers, thus increases the false positive rates.

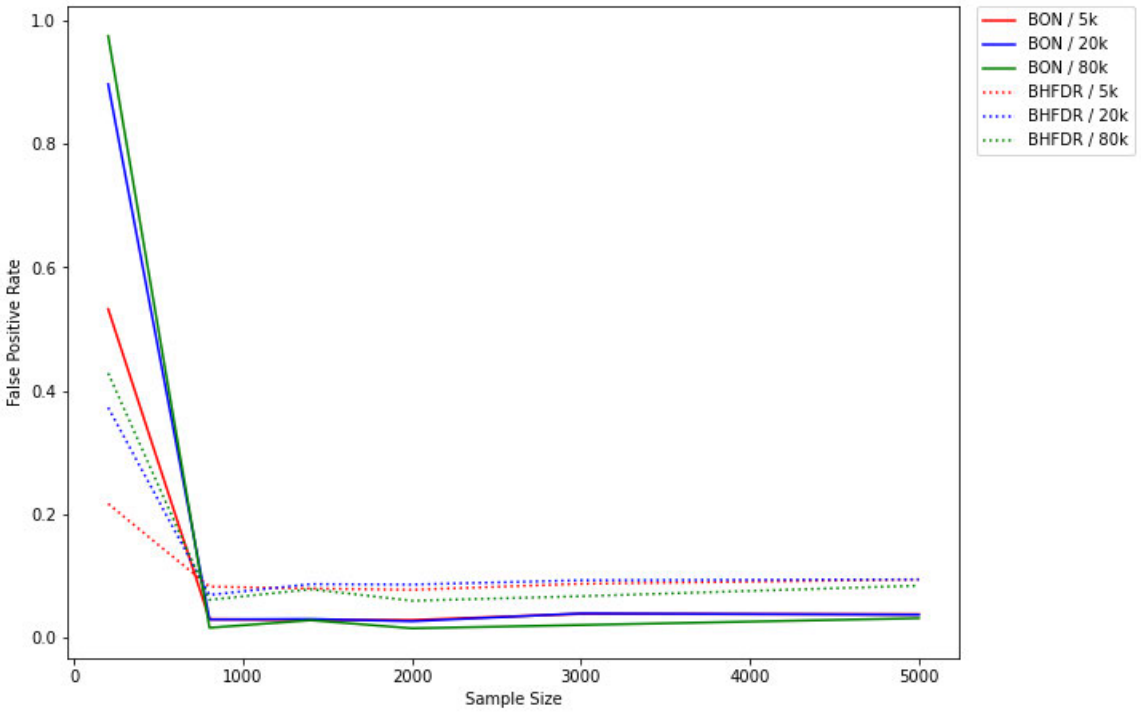

Figure 3.5: The effects of sample size on the false positive rate of GWAS under varying number of independent markers and correction methods. Solid lines represent the number of false positives for the Bonferroni correction whereas dashed lines represent those of BH-FDR. The number of QTL is maintained at 100, and the average QTL effect sizes (γ) and allele frequencies (β) maintained at 0.4 and 0.5 respectively.

The false positive rate is not significantly affected by the average QTL effect size (γ) when the markers are independent. But for correlated markers a lower value for γ significantly increased the number of FPU and false positive rate in both multiple testing correction methods. The number of false positives began to stabilize at an average QTL effect size of 0.4 (Figure 3.4(d)). The number of false positives is not significantly affected by the distribution of the allele frequencies.

Correlation between markers: For all the parameters tested, correlation between markers has a significant effect on the number of false positives detected in a GWAS. The presence of correlation significantly increased the number of false positives, although most of the false positives are correlated to the true QTLs (Table 3.4). This observation can be made in both multiple testing correction method, although the numbers of both correlated and uncorrelated false positives are higher for BH-FDR compared to the Bonferroni correction.

Table 3.4: The number of true positives (TP), correlated false positives (FPC) and uncorrelated false positives (FPU) under varying multiple testing correction methods and dependency between markers. Default parameters had been used in calculating the number of true and false positives for this table.

Multiple Testing Correction Method	Bonferroni correction			BH-FDR		
	TP	FPC	FPU	TP	FPC	FPU
Independent Markers	7.53	NA	0.22	9.13	NA	0.76
Correlated Markers	7.86	51.12	0.80	12.23	98.44	10.18

3.4.4. Effects of Parameters on $ROC\ score$ of Multiple Testing Correction Methods

With increasingly large numbers of markers used, there is a general decline in the difference between number of true and false positives ($ROC\ score$) for both correction methods. For independent markers, the BH-FDR had a higher $ROC\ score$ compared to the Bonferroni correction for all numbers of markers tested in this experiment. This suggests that the threshold for Bonferroni correction provided a less favourable balance between power and false positive rates in a GWAS experiment. The trend changes with the presence of correlations however; when the markers are correlated the BH-FDR had a significantly reduced $ROC\ score$ for all numbers of markers tested. This is attributable to an increased number of FPU when the assumption of independence is violated. With the exception of small number of markers used, which increases the $ROC\ score$, correlation between markers generally do not have a significant effect on the $ROC\ score$ for the Bonferroni correction method (Figure 3.6(a)).

Besides the number of markers used, sample size also has a significant effect on the $ROC\ score$ for both multiple testing correction methods (Figure 3.6(b)). When the markers are independent, compared to Bonferroni correction, BH-FDR has somewhat higher $ROC\ score$ in all sample sizes tested, although this observation is more notable for very small sample size (N=200) or for the larger sample size (N=5000). The presence of correlation changes the trend however; for N=200, the $ROC\ score$ for BH-FDR is higher than for Bonferroni, but with sample size of 800 and larger the $ROC\ score$ of Bonferroni is higher than that of BH-FDR, and with sample size larger than 1400, the $ROC\ score$ actually

decreases with larger sample size for BH-FDR. This is attributable to an increased number of false positives for BH-FDR with large sample sizes.

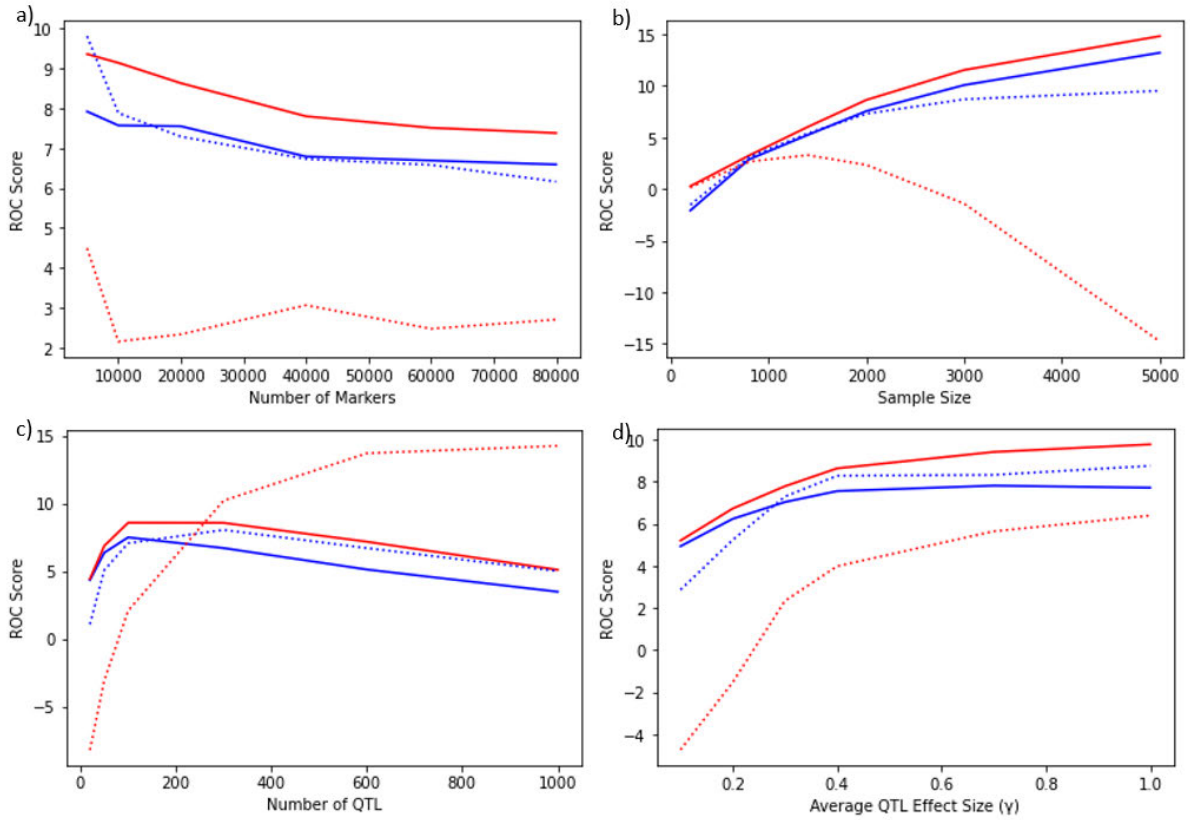

Figure 3.6: The effects of (a) number of markers, (b) sample sizes, (c) number of QTL and (d) average QTL effect size (γ) on the Receiver Operating Characteristics (ROC) Score of a GWAS for the Bonferroni correction (blue line) and BH-FDR (red lines) for both independent (solid lines) and correlated (dashed lines) markers. The default parameters for each of the plots are as follows: number of QTL at 100, sample size 2000, the number of markers 20k, the average QTL effect size (γ) at 0.4 and shape parameter for allele frequencies distribution (β) at 0.5.

For independent markers, the $ROC\ score$ for both BH-FDR and Bonferroni correction would initially increase, but when the number of QTL exceeds 100 there is a slow decline in the $ROC\ score$. This can be attributed to a decline in power with increasingly large number of QTL that have smaller effect. For correlated markers, the $ROC\ score$ of Bonferroni correction followed a similar trend as with independent markers, but the trend is different for BH-FDR. When the number of QTLs is small (less than 200) the $ROC\ score$ went below that of Bonferroni correction, but with further increase in number of QTL, the $ROC\ score$ with BH-FDR became significantly higher than that of Bonferroni (Figure 3.6(c)).

With increased numbers of QTL with large effect sizes (i.e., a high average QTL effect sizes), the $ROC\ score$ for both multiple testing correction method increases (Figure 3.6(d)). This can be attributed to an increase in power with increasingly large average QTL effect sizes. When the markers are independent, the BH-FDR has again a higher $ROC\ score$ than Bonferroni for all parameter values, although the increase is more significant with a larger average QTL effect size. This trend flipped when the markers are correlated however; while BH-FDR has high power in detecting QTLs, the massive increase in number of false positives decreased the $ROC\ score$ to that below of Bonferroni correction. Correlation has a less significant effect on the $ROC\ score$ for Bonferroni correction.

3.5. Discussion

In this experiment the effects of parameters on the threshold of Bonferroni correction and BH-FDR, as well as its associated power and false positive rate, were tested. Unlike BH-FDR, which has its threshold affected by various parameters, none of these parameters have an effect on the threshold of the Bonferroni correction, with the exception of number of SNP markers. This is due to how the threshold is calculated; with the number of SNP markers being the only variable for the calculation of threshold for the Bonferroni correction (equation [5]). The threshold of the BH-FDR also depends on the distribution of the ranked p-values of the markers (i.e., rank "k" inequation [6]). Due to this, any parameters that could affect the distribution of p-values would have an effect on the threshold from the BH-FDR. Parameter values that would increase the -log(p-value) of the markers increase the value of point k and thus decrease the stepwise threshold from [6], thus decreasing the stringency of the threshold. Conversely parameters that decrease that -log(p-value) decrease the value of point k and thus increase the threshold stringency. For example, increasing the sample size of the GWAS increases the test statistic of the marker and thus increases the value of the -log(p-value). This causes an increase in the value of k and thus decreases the stringency of threshold. Conversely a trait with a large number of QTL decrease the proportion of phenotypic variance explained of any given QTL, and this increase the p-values and thus the stringency of the threshold of the BH-FDR. Correlations between markers also causes the "bleeding" of effect sizes from the true markers into the neighbouring null markers, and this produces a peak of true QTL with several neighbouring null markers flanking the peak (Figure 3.7). This increases the p-values of the neighbouring null markers and thus decreases the threshold of the BH-FDR.

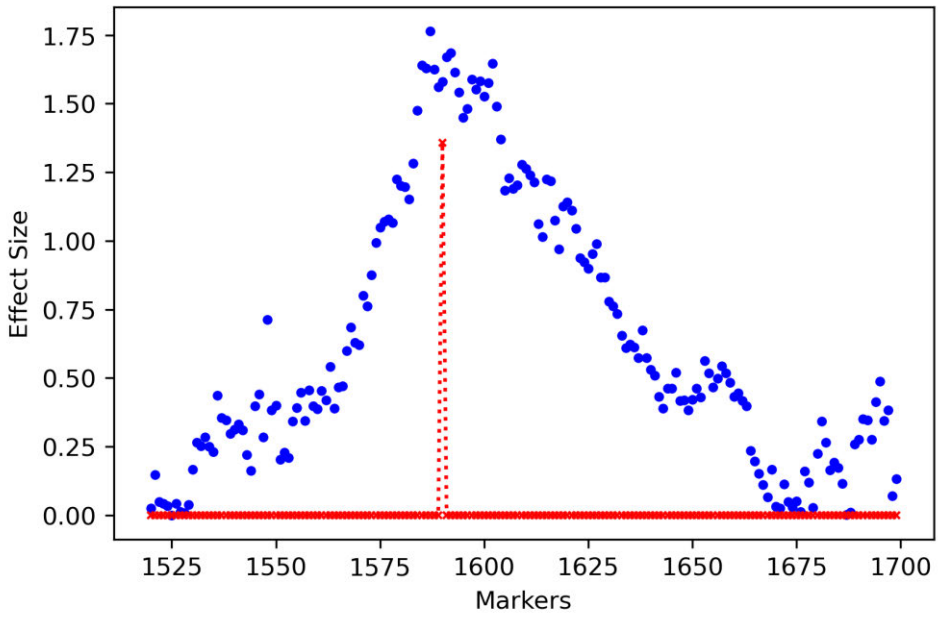

Figure 3.7: The estimated effect sizes (blue dots) of a peak in a correlated marker, showing the effect of correlation on null markers that flanked a QTL (red peak at locus 1589). The marker pairwise correlation is set at $R_{LD} = 0.95$.

The effects of these changes in the threshold of both correction methods would affect the power and false positive rate of a GWAS, with a decreased stringency in threshold increases its power and false positive rate and vice versa. For example, increasing the number of markers caused the threshold to become more stringent as it needs to exclude the additional null markers. This increased stringency however also has the effect of decreasing the number of true positives and thus the power. As the threshold increase in stringency in a logarithmic fashion with an increase in the number of markers, the power also decreases in a similar fashion, approaching zero as none of the true QTL had its p-value exceed the extremely stringent threshold. This could be an issue for Whole Genome Sequencing (WGS) data, where Tam et al (2019) warned the exacerbation of decline in power due to the overconservative threshold, especially when the Bonferroni correction is used. In this situation BH-FDR might serve as a better alternative.

This study also suggested a larger number of markers does not necessarily increase the power, as it might increase the number of null markers utilized in a GWAS experiment and increase the required stringency of a threshold. In fact, we saw a decrease in power with a larger number of markers tested in both correlated and uncorrelated markers. Conversely increasing the sample size increase the -log(p-value) of the true markers, making them more likely to be detected. This increases the number of true positives logarithmically, thus

increasing the power of GWAS. This suggests that increasing the sample size could be more important than increasing the number of markers used in a GWAS experiment.

On the other end of the spectrum, the use of small sample size significantly increased the number of false positives and decreases the number of true positives. This is due to the fact that observations made from a small sample size can often be explained by a larger number of predictors (i.e., SNPs markers), which causes the null markers that have its combination of genotypic values coincided with those of true markers to have an elevated p-value, contributing to the false positive rate (Forstmeier et al., 2017). Combined with the reduced number of true positives, this means a GWAS with small sample size would have low power and high false positive rate. This observation is also supported by the low $ROC\ score$ for small sample sizes, and this elevated the number of false positives, especially with increasingly large number of SNPs. As expected, the results of this experiment casted doubt on the validity of the results obtained from studies with a small sample size, especially for those with high marker density.

Besides the experimental design parameters, the distribution of the QTL effect size of the trait studied also affect the threshold, power and false positive rate of a multiple testing correction method in a GWAS experiment. This agrees with the study of Panagiotou and Ioannidis (2012) which stated that the most suitable threshold used in a GWAS experiment might vary for different populations and genetic architecture of the trait. This could mean that a threshold from one study might not be suitable for another association study. Indeed, this can be observed with the decreased number of true positive as well as increased number of false positives for a trait with a small number of QTL with large effect sizes (small average QTL effect size in this study). While in theory this observation could be used to calculate the threshold optimized for the trait in study, in practice this might not be possible as it requires information on the underlying QTL effect size distribution. Previous works such as those published by Park et al (2011), Hall et al. (2016) and Zhang et al. (2018) have attempted to estimate such distribution using various approaches, although these algorithms assumed the QTL effect sizes followed exponential or normal distribution, and the effect of violation of such assumption (as an example, the QTL effect size followed a gamma distribution) remained untested. Future work should be thereby focus on an algorithm for robust estimation of QTL effect size distribution, and its incorporation into the calculation of optimal threshold for a GWAS experiment.

Panagiotou and Ioannidis (2012) also commented that correlated markers constituted a major source of uncertainty in the suitability of a threshold used in a GWAS experiment. Our results show that this is a valid concern. Due to the "bleeding" effect of the effect sizes, correlation between markers significantly increased the number of false positive in a GWAS experiment, especially with increasingly large number of markers used. As markers become increasingly dense, they become less well separated to one another and they no longer inherited independently, producing linkage disequilibrium between markers (Cheverud, 2001; Falconer, 1989). It is expected that maximal linkage disequilibrium would be observed for WGS data, which made high degree of correlation between markers unavoidable (Pengelly et al., 2015).

While correlations between markers led to an increased number of false positives for BH-FDR and increase in number of QTL also led to a decline in the power of GWAS, such that increase in number of QTLs had led to an increase in the absolute number of true positives and a decrease in number of false positives. This led to an increase in the $ROC\ score$ and a decrease in the false positive rate. This could be attributed to an increase in threshold stringency for the BH-FDR with large number of QTLs associated with a trait, as well as the reduced proportion of variance explained by each of the QTL. Such a decrease in false positive rate and increase in $ROC\ score$ was neither observed for the Bonferroni correction, nor for independent markers. While this initially suggested that BH-FDR might be more suitable correction method compared to those of the Bonferroni for a polygenic trait, especially with correlated marker, it does come with a caveat: the actual false positive rate in BH-FDR is significantly higher than that of the Bonferroni method; the false positive rate from FPU is 0.095 under default conditions for BH-FDR, compared to 0.006 for Bonferroni. Thus, the actual suitability of the correction methods depends on the priority of the experiment; if the priority of the GWAS is to be placed on the explanatory power of the GWAS, then BH-FDR might serve as a better choice, but if the priority is to increase the specificity (i.e. number of false positives out of all null markers) of the GWAS, then the Bonferroni might be preferred.

Rather than choosing one multiple testing correction methods over the other, perhaps a better alternative is to modify the methods so that they could take into account correlation between markers. One such method is a Bonferroni correction that utilized "effective number of independent markers" instead of raw number of SNP markers, similar to those suggested by Cheverud (2001) and Nyholt (2004). This might serve as a promising route for an increased

power. The calculation of effective number of independent markers utilized the variance of eigenvalues obtained from the marker correlation matrix to adjust the "Number of SNPs" in equation [3], thus yielding a less stringent threshold. One downside of this method however is that the calculation of a very large marker correlation matrix is memory and computationally demanding, which might not be feasible with large dataset. Modification of the original methods might thus be required. Studies on the effect of such adjustment of "Number of SNPs" on the false positive rate of GWAS is also lacking as well. Attempts to modify the BH-FDR so that it could take into account correlation between markers had also been done by Benjamini and Yekutieli (2001), and this might serve as a better alternative than the BH-FDR.

While correlation between markers is expected to be at its strongest with WGS data, the effects of correlation on setting the threshold in a multiple testing correction method cannot be ignored in GWAS experiment where WGS data is not used. With pairwise marker correlation as high as 0.8, it has caused such a significant decline in the stringency of the threshold for the BH-FDR that it fails to control the number of false positive and thus the false discovery rate at 0.05 in all parameter value tested. This highlighted the importance of selecting the appropriate multiple testing correction method in a GWAS experiment. The results from this experiment have also run contrary to the claims from Benjamini and Yekutieli (2001) on the validity of the BH-FDR in correlated marker array. This experiment highlighted the unsuitability of the BH-FDR with high density marker arrays, which is to be expected in a real GWAS experiment, especially when the markers are not sufficiently dense. In this situation the Bonferroni correction was shown to be more capable of maintaining the number of false positives.

An important note for this experiment, which would serve as a caveat, is how the number of FPU and FPC is determined. It should be noted that the cut-off point between FPU and FPC (i.e. $r^2 = 0.1$) is arbitrary and changing said cut-off point would affect the number of FPU and FPC. The rationale of using this cut-off point is to differentiate the "false positives" that is caused by correlation with true QTL from those that actually caused by the varying parameter values. This distinction is important in the context of increasing the sample size of the GWAS. While the naïve results of this experiment suggested that increasing the sample size increase the false positive rate of a GWAS, this is more likely the effect of choosing a certain cut-off point for FPC and FPU, as increasing the sample size would not only decreases the required effect size detectable by the GWAS, but also the required correlation

between the marker and the QTL. Thus, given a probability of detection of a QTL, the required correlation between a marker and QTL decreases with increasingly large sample size. This is best illustrated by Spencer et al. (2009) who had provided the following proportionality between the test statistics for the detection of a QTL and correlation between QTL and marker:

$$Test\ Statistics\ \propto N\widehat{a_k}^2 p_k(1-p_k)R^2(Q,k) \qquad [10]$$

Where N is the sample size, p_k is the allele frequency, a_k being the QTL effect size and $R^2(Q,k)$ being the correlation between QTL and marker. This proportionality suggests that for a given a test statistic value, as N approaches infinity, the $R^2(Q,k)$ required for the test statistic to reach said value approaches zero. Thus, as long as a marker has a nonzero correlation with any of the QTL, regardless how small the correlation is, there would be a finite sample size required. Taking this to extreme, this could cause a GWAS to declare excessively large number of null markers to be positive, even if those markers are minimally correlated to the QTL. This could be the situation observed by Jiang et al. (2019), who have utilized 294,079 animals in their GWAS experiment, and with that sample size the experiment declared 27 percent of all markers as positives (15215/57067 = 0.267). It is for this reason that an extremely liberal cut-off point for FPC and FPU of $r^2 = 0.1$ has been chosen for this experiment, to ensure that any FPU detected are "as null as possible" (i.e. as little influenced by a QTL as possible). In the context of multiple testing correction methods, this also suggest that a GWAS with large sample size could afford a more stringent threshold, such as those suggested by the Bonferroni correction.

Given that differing genetic architecture of a trait and experimental designs would affect the suitability of the threshold of a multiple testing method, an algorithm that could test the suitability of such threshold would be desirable. One possible method of testing the appropriateness of the multiple testing correction method is to use the $ROC\ score$. As defined by equations [7], [8] and [9], a low $ROC\ score$ could either be caused by low power, which was associated with an overconservative threshold, or with high false positive rate, which was associated with overly lenient threshold. Only a threshold that could provide a good balance between true and false positives that would have a high $ROC\ score$. Results from this experiment suggested that when the markers are independent, the BH-FDR provided a better balance between power and false positive rate for all parameter values tested when markers are not correlated, but for correlated markers, the Bonferroni correction consistently

provided a better balance between power and false positive rate for all parameter value tested except for highly polygenic trait (i.e. trait with large number of QTL).

Given the relationship between the $ROC\ score$ and the optimality of the threshold, another potential route for further study is to establish an algorithm that could find an optimal threshold that would maximize the $ROC\ score$ based on certain parameters related to experimental design and genetic architecture of the trait in study, similar to what is suggested by Habibzadeh et al. (2016) in finding a threshold that balances the power and false positive rate in a clinical test, and de Smet et al. (2004) had used such an algorithm in balancing true and false positives in a gene expression experiment. Along with suitable modification, such as taking into account the effect of correlation between markers, a similar algorithm could be suggested to be used in a GWAS experiment. A major obstacle for this route is the requirement of prior information on the underlying QTL effect size distribution, which further emphasize the importance of a robust algorithm to estimate it.

In conclusion this experiment suggested that power and false positive rate in a GWAS experiment is affected by the choice of the multiple testing correction method, the experimental parameters such as sample size and number of markers, and the genetic architecture parameters of the trait studied. For independent markers, the BH-FDR provided a better balance between the true and false positives for all parameter values, but for correlated markers, the Bonferroni correction did provide a better balance between true and false positives. The only exception where the BH-FDR provided a better balance between true and false positive with correlated markers is when the trait is highly polygenic, and even so with the caveat of increased false positive rate. This experiment had also suggested that increasing the number of markers used in an experiment would not necessarily increase the power of GWAS but increasing the sample size would increase the power and decrease the false positive rate of GWAS. Our study also showed the importance of having large sample size if large number of markers is to be used in a GWAS experiment, which would be crucial if WGS data is to be used in a GWAS experiment, as a genotype array of such high density would inevitably and excessively increase the stringency of the threshold, necessitating a larger sample size. Future work should focus on a robust algorithm to estimate the QTL effect size distribution and using it to calculate an optimal threshold that could balance the power and false positive rate under arbitrary experimental designs and genetic model of the trait studied in a GWAS experiment.

Chapter 4. A Robust Algorithm for Calculation of an Optimal Threshold in Genome-Wide Association Studies

Zhi Loh, Julius H. J. van der Werf, Sam Clark

4.1. Abstract

While Genome Wide Association Studies have become an important tool in identifying causal loci, the large number of markers utilized has created a severe multiple testing problem which reduces its power. While several methods have been suggested to control false positives, they have their own strengths and shortcomings in terms of balancing power and false positive rates, with none of them taking into account the effects of parameters such as distribution of QTL effect sizes. Using the Receiver Operating Characteristics (ROC), we developed an algorithm for the calculation of an optimal threshold that could balance the power and false positive for a given set of experimental parameters and evaluated its performance against two of the most popular correction methods. Through simulated genotypes and phenotypes, we found that, compared with the frequently used Bonferroni and FDR methods, the optimal threshold performed better in binary classification between significant and non-significant markers, which is important for QTL identification. The optimal threshold leads also to more accuracy in genomic prediction when the threshold was used to as a truncation point when selecting the markers to be used for genomic prediction; the use of optimal threshold led to an increment in accuracy up to 16.8% compared to the Bonferroni method and 7.0% compared to FDR method. This study is important not only within the scope of genomics in term of causal variant identification, but also in signal processing theory for the generalization of ROC algorithm in the context of handling correlated tests and class imbalance.

4.2. Introduction

Since the advent of high-density markers such as Single Nucleotide Polymorphisms (SNPs), Genome-Wide Association Studies (GWAS) have become one of the most important tools in identifying loci associated with a trait. GWAS has found several uses in the field of

genomics, such as identifying causal loci associated with human diseases such as diabetes (Cai et al., 2020) and multiple sclerosis (Cotsapas and Mitrovic, 2018), as well as truncation of null, uninformative markers for genomic prediction (Brøndum et al., 2015). Despite this, the large number of markers used in a GWAS has introduced a severe multiple testing problem which significantly increases its false positive rate. Such issue could be exacerbated by the use of increasingly dense markers or genetic variants derived from Whole Genome Sequence (WGS) data. Such GWAS would necessitate the correct use of a multiple testing correction method.

The most popular multiple testing correction method is the Bonferroni correction, initially introduced by C. E. Bonferroni before being popularized by Dunn (1961). The rationale of this correction method is the observation that the confidence interval of the joint distribution of a number of variates that follow a Student t-distribution can be calculated as the Type 1 Error divided by total number of tests (Dunn, 1961). This correction method had gained popularity in GWAS due to its simplicity of implementation and effectiveness in controlling the false positive rate (Wilson, 2019). In its effort of controlling the false positive rate however, the Bonferroni correction has been widely criticized for its excessively stringent threshold, particularly with the dramatic increase in marker number, as this led to reduced power in GWAS (Huang et al., 2018; Wilson, 2019).

The low power of GWAS with the Bonferroni correction method has prompted numerous other multiple testing correction methods, with the most well-known class of methods being those that attempt to control the false discovery rate (FDR). The first FDR-based correction method is the Benjamini-Hochberg FDR (BH-FDR), first suggested by Simes (1986) before popularized by Benjamini and Hochberg (1995). While previous publications such as Storey (2002) suggested an increased power with the BH-FDR, the effects of violation the assumption of independence of markers required by the BH-FDR was not considered. Indeed, Broberg (2005) had suggested the inability of BH-FDR in controlling FDR when the markers are dependent, while Huang et al. (2018) suggested the failure of BH-FDR in controlling FDR in general.

Since then, a multitude of FDR-based correction methods have been suggested, each with its own strength and weaknesses. For example, Benjamini and Yekutieli (2001) suggested a method that considers the lack of independence between tests. Storey (2002) also suggested another correction method that increases the power by using information from the number of

actual null markers, although with the assumption of independence between tests. Broberg (2005) suggested Pooling of Adjacent Violators FDR (pava-FDR) which enforces monotonicity on the local FDR, defined as the FDR within a range of p-values, with the assumption that local FDR in these models is monotonic (Efron et al., 2001).

Despite the multitude of these algorithms, the optimality of these algorithm in balancing the power and the false positive rates in a GWAS remains unclear, especially with the context of changing parameter values. While Broberg (2005) tested the ability of some of the FDR-based methods in controlling the FDR, the effects of varying parameters associated with a GWAS study, such as those related to experimental design and genetic architecture of the trait, remained unclear. Ioannidis (2007) stated that the ratio between true and null SNPs and population stratification could affect the FDR of a GWAS, while Hoggart et al. (2008) suggested a dependency of the GWAS significance level on the population structure and sample size used in an experiment. None of these studies went into detail on how to take these effects into account when deciding an optimal threshold.

There are also questions on the severity of the impact of false positives in a GWAS. As the results from a GWAS experiment could have multiple uses such as detecting causal loci or marker selection in genomic prediction, one could ask the question of "How severe the impact of false positives is toward a GWAS experiment?", or "Does the severity of impact from these false positives depends on the ultimate purpose of a GWAS results?" Combined with the arbitrariness of the chosen threshold (p-value of 0.05 before correction methods being applied), this also raises the question of suitability of a threshold and its associated multiple testing correction method. While Panagiotou and Ioannidis (2012) commented on the potential impact of such an arbitrary threshold on a GWAS experiment, proper studies on how severe such impact is to the GWAS experiment remain lacking, especially under different ultimate purposes of a GWAS experiment.

Perhaps rather than choosing a threshold arbitrarily or utilizing a one-size-fit-all algorithm in attempts to increase the power of GWAS while controlling its false positive rate, an alternative method could be establishing a threshold that could provide an optimal balance between the power and false positive rate simultaneously. One such method is those based on the Receiver Operating Characteristic (ROC) curve. Initially introduced as a way of distinguishing signals from noise for radar operators in World War II, the ROC had found its use in numerous fields such as medical diagnostic tests (Habibzadeh et al., 2016), psychology

and psychophysics (Streiner and Cairney, 2007) and gene expression analysis (de Smet et al., 2004). Previous studies such as Habibzadeh et al. (2016) and de Smet et al. (2004) demonstrated the possibility of using ROC in identifying the optimal threshold in medical diagnostic tests and gene expression respectively. While in the context of GWAS, the ROC curve has been used to evaluate the sensitivity and specificity of a GWAS experiment (Bossini-Castillo et al., 2021; Patron et al., 2019), or evaluating the performance of a newly developed model (Shafquat et al., 2020). These studies did not test the optimal balance between the sensitivity and specificity of a GWAS however, especially if the GWAS results are to be used for different purposes, or when correlation between markers, genetic architectures and imbalance between number of QTL and number of null markers need to be considered.

With this in mind, the aim for this study is to establish an algorithm for a threshold that provides an optimal balance between power and false positive rates in a GWAS experiment, while taking into account factors that would be relevant in such experiments, such as correlation between markers, effects of genetic architectures and experiment designs, and the imbalance between number of QTL and null markers. This optimality of the threshold would then be tested using simulation under two ultimate uses for a GWAS experiment: in gene discovery and in truncated genomic prediction.

4.3. Theory

4.3.1. Definitions used in this Study

In this study, the optimality or performance of a threshold is defined as its capability in balancing the power and false positive rate. To establish an algorithm that could produce a threshold that could balance the power and false positive rate of the GWAS, concrete definitions for both power and false positive rate were required.

For this study the power of GWAS was defined as follows:

$$power = \frac{Number\ of\ QTL\ detected}{Total\ number\ of\ QTL} \quad [1]$$

The $Number\ of\ QTL\ detected$ can be defined through multitude of ways. It can be defined as the number of QTL with its test statistic exceeding a critical value, its p-value below a threshold, or its negative logarithmically transformed p-value exceeding its correspondingly

transformed threshold. For this study, the negative logarithmically transformed threshold (henceforth defined as THR) was utilized.

The $Number\ of\ QTL\ detected$ depends on the genetic architecture of the trait. For example, a highly polygenic trait would have very large $Number\ of\ QTL\ detected$ but most of which have small effect sizes, with each explaining a very small portion of the additive genetic variance. Whereas an oligogenic trait would have small number of large QTL and each would explain a relatively large portion of additive genetic variance. Methods for the estimation of number of QTL and its effect size distribution will be discussed below. As it is unrealistic to expect a GWAS to detect all the QTL, especially for QTL with very small effect sizes, only true markers with effect sizes greater than the bottom 30% of all QTL were counted under $Total\ number\ of\ QTL$ in this study (i.e., only top 70% of all QTLs were included in the calculation).

Under the same THR, the false positive rate of a GWAS can be defined as follows:

$$false\ positive\ rate = \frac{Number\ of\ null\ marker\ detected}{Number\ of\ positive\ hits} \quad [2]$$

Both $Number\ of\ null\ marker\ detected$ and $Number\ of\ positive\ hits$ can be obtained through the test statistic or the p-values in GWAS, with the assumption that the location of the true QTL can be determined.

4.3.2. Calculation of Power and False Positive Rate in GWAS

The basics of GWAS is to estimate the slope component $\boldsymbol{a}$ of the line of regression such that it would minimize the mean squared deviation of the data points from the line. The phenotypic model assumed by a GWAS experiment is defined as follow (Gondro, 2015):

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} \quad [3]$$

where $\boldsymbol{y}$ being a $N \times 1$ vector containing phenotypic values of N number of animals; $\boldsymbol{X}$ being a $M \times N$ matrix containing genotypic states of M number of markers from N animals; $\boldsymbol{a}$ being a $M \times 1$ vector containing the effect sizes of each marker allele or QTL and $\boldsymbol{e}$ being a $N \times 1$ vector containing the residual component in the phenotype.

For the calculation of power and false positive rate in a GWAS experiment, the negative logarithmically transformed p-value of the markers were required. This transformed p-value can be defined as follows:

$$logpval = -\log_{10}\left(2 * \int_{T_i}^{\infty} t(T_i;\ N-2)\ dx\right) \quad [4]$$

Where $t(x; v)$ is the probability density function (PDF) of Student's t-distribution. The T_i is the test statistics of the marker which, if the Hardy-Weinberg Equilibrium (HWE) is obeyed, were defined as follows:

$$T_i = a_i * \sqrt{\frac{var(\boldsymbol{X}_i) * (N-2)}{var(\boldsymbol{y}) - a_i^2 var(\boldsymbol{X}_i)}}$$

$$= a_i * \sqrt{\frac{2p_i(1-p_i)(N-2)}{Var(\boldsymbol{y}) - 2p_i(1-p_i){a_i}^2}} \quad [5]$$

The mathematics for the derivation of the test statistics and p-values of a marker are provided in Appendix A.

For the calculation of power of GWAS, the number of true positives from a GWAS (denoted as $Number\ of\ TP$) would be needed. In this study the $Number\ of\ TP$ can be defined as the number of true QTL with its $logpval$ exceed the threshold THR:

$$Number\ of\ TP = \left(\#\{logpval_{QTL} \geq THR\}\right) \quad [6]$$

And the power of GWAS calculated as follows:

$$power = \frac{Number\ of\ TP}{nQTL} \quad [7]$$

For this study, the $nQTL$ were defined as number of QTL associated with a trait with effect size larger than the "trivial effect sizes" (denoted as a_{min} in this study). This would be the $power$ that will be used in the subsequent sections.

Several previous publications have attempted to estimate the number of QTL and its associated effect size distribution using GWAS-based statistics. Given a sample size and variance explained by the SNP markers, Park et al. (2010) utilized the previously reported power of a study to estimate the effect size distribution. Cheng et al. (2020) utilized the expectation-maximization algorithm to estimate the proportion of QTL with certain effect size and the variance contributed by said QTL and build a mixture model using these parameters. Zhang et al. (2018) also provided an algorithm for the estimation of number of non-null markers for a disease outcome that could be modelled using logistic regression. Hall

et al. (2016) provided an algorithm for estimating number of QTL using the proportion of variance explained by the QTL and the heritability of the trait.

Despite this, there are many assumptions and limitations in these methods. One such limitations was the inflexibility of the assumed distributions (normal distribution for Cheng et al. (2020) or exponential distributions for Hall et al. (2016)), which might affect the validity of these algorithms on real data, and the effects of varying allele frequency and small sample sizes. For this study the $nQTL$ and its associated distribution, as well as the location of these QTL, are assumed to be known. Further studies should focus on providing a robust algorithm to estimate the $nQTL$, its effect size distribution and its location using a sample size comparable to a GWAS experiment.

For a GWAS experiment, the number of false positives can also be defined as the number of null markers with its $logpval$ that exceed the threshold THR:

$$Number\ of\ FP = (\#\{logpval_{NUL} \geq THR\}) \quad [8]$$

And the false positive rate for a GWAS can be calculated as follows:

$$false\ positive\ rate = \frac{Number\ of\ FP}{Number\ of\ FP + Number\ of\ TP} \quad [9]$$

4.3.3. Balancing the Power and False Positive Rate

4.3.3.1. The Basics of Receiver Operating Characteristics (ROC) Curve

To balance the power and false positive rate in a GWAS, a receiver operating characteristic (ROC) curve can be used. An example of implementation of the ROC in identifying the optimal threshold is provided in Figure 4.1.

The common interpretation for ROC curve is plotting the changes in power under varying probability of false alarm, defined as the total number of false positives over the total number of null cases (Habibzadeh et al., 2016). Under this setting the optimal threshold was defined as the point on the curve where the tangent of the curve equals to one (the red dot on Figure 4.1(b)). This would also be the argument of the maxima for the ROC's Youden's Index, defined as the differences between power and probability of false alarm, as proven by de Smet et al. (2004), Schisterman and Perkins (2005) and Kaivanto (2008) (Figure 4.1(b)).

This interpretation of ROC curve is not directly usable in identification of optimal threshold in GWAS however, as this interpretation placed equal emphasis on false negative and false positives (Chicco and Jurman, 2020; Lobo et al., 2008). Given that the number of null markers generally far exceeds the number of true QTL in a GWAS experiment, placing equal emphasis on false negatives and positives has the ramification of setting the threshold overly lenient, which allows an excessive number of false positives. Indeed, from a test-run example provided in Figure 4.1(c), with 2000 QTL from 50k independent markers, under this definition of optimal threshold, while there are 516 true QTL being detected, the threshold also marked 10,230 null markers as positive, representing a false positive rate of 95.2%. Thus, an alternative interpretation was required.

One such interpretation for the ROC curve, which would be used in this study, is the raw differences between the number of true positives and false positives under varying threshold levels:

$$ROC_{THR} = TP_{THR} - FP_{THR} \quad [10]$$

where the ROC_{THR}, TP_{THR} and FP_{THR} denote the ROC score, number of true positives and number of false positive at threshold THR, respectively. This is equivalent to the Youden's index that had been described in Habibzadeh et al. (2016) weighted by total number of true and null markers, which made up the denominator portions of power and probability of false alarm respectively. The reinterpreted ROC curve has the benefit of its ability in taking into account the massive discrepancy between the number of true and null markers, producing a more applicable threshold (Figure 4.1(d)).

Another benefit for this reinterpretation of the curve is the easiness of obtaining the optimal threshold; as the calculation of $Number\ of\ TP$ and $Number\ of\ FP$ involves counting functions, the ROC curve is not smooth and not differentiable, impeding the discovery of the tangent of the curve. With the reinterpreted ROC curve, the optimal threshold (denoted as THR_{opt}) can easily be obtained as the argument of the maxima of the curve:

$$THR_{opt} = \arg\max(TP_{THR} - FP_{THR}) \quad [11]$$

4.3.3.2. Generalization of the ROC Curve and THR_{opt} calculation

Besides the aforementioned reinterpretation of ROC curve, the formulation can also be generalized to place more emphasis on markers that showed evidence of association with the phenotype (i.e., additional weightage is assigned to an associated marker).

An applicable weightage is weighting the TP_{THR} with the effect size associated to the QTL. The weightage would transform the raw TP_{THR} into the sum of the absolutized effect sizes associated with the true positives:

$$TP_{THR_{wt}} = \sum_{i \in TP} |a_i| \qquad [12]$$

With the index of summation $i \in TP$ denoting a set of QTL being marked as positives.

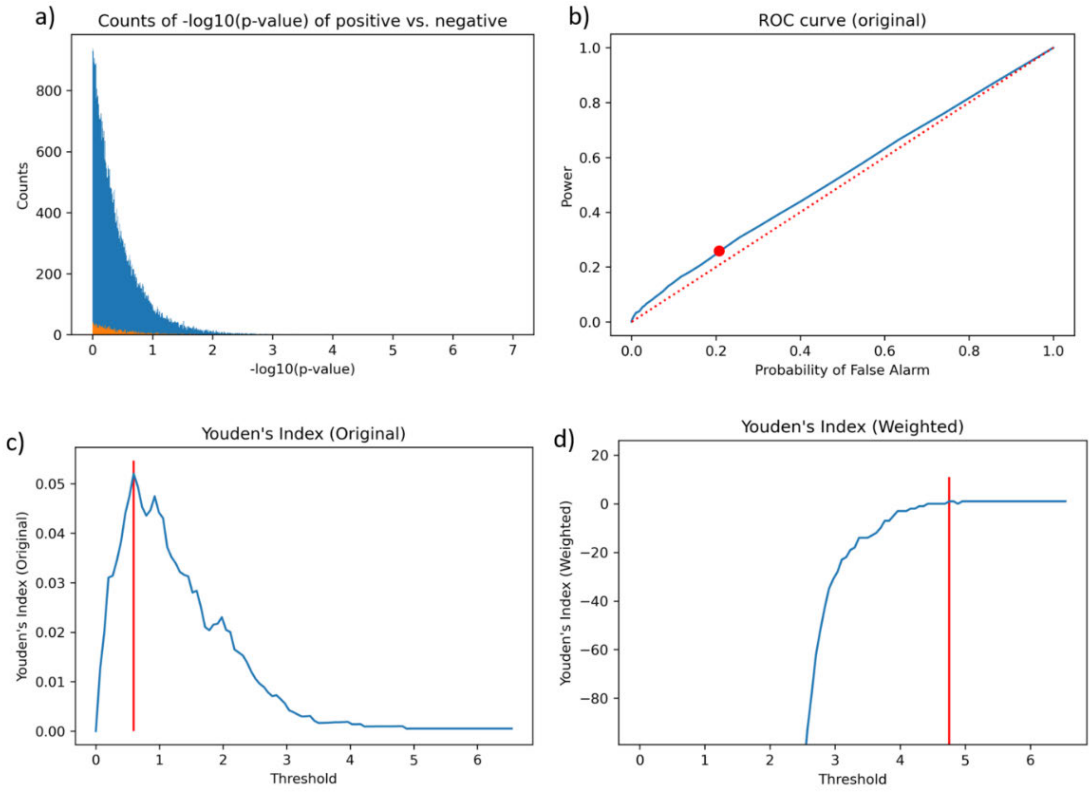

Figure 4.1: The implementation of ROC curve in identifying optimal threshold. Figure (a) illustrated the distribution of –log10(p-value) for both null markers (blue) and true QTL (orange). Figure (b) illustrated the classical ROC curve on the distribution of p-values, with red point indicating power and probability of false alarm for the optimal threshold, defined as the point where the tangent of the curve as a slope of 1. Figure (c) shows the changes in unweighted Youden's Index under varying threshold, with the red line indicating the optimal threshold under this definition. Figure (d) illustrated the changes in weighted Youden's Index under varying threshold, with red line indicating the optimal threshold. The example utilized in generating these figures is conducted using 50k independent markers and sample size of 2000, with number of QTL set at 2000 with effect size distribution of Gamma (0.5, 1) and narrow sense heritability at 0.3.

To ensure the balance of weights between the true and the false positives, the FP_{THR} would also need to be weighted. One applicable weight is the expected value of the QTL effect size distribution, which weighs the number of false positives as follows:

$$FP_{THR_{wt}} = FP_{THR} * \left(\frac{\sum_{i=1}^{nQTL} |a_i|}{nQTL} \right) \quad [13]$$

With this weight, the THR_{opt} was defined as the argument of the maxima in differences between sums of absolute effect sizes associated with the true positives and the number of false positives:

$$THR_{opt} = \arg\max \left(\sum_{i \in TP} |a_i| - FP_{THR} * \left(\frac{\sum_{i=1}^{nQTL} |a_i|}{nQTL} \right) \right) \quad [14]$$

A shortcoming for the weighted false positive as suggested in Equation [13] is its lack of robustness against the distribution of the QTL effect size, most notably a distribution with large number of QTL with small effect sizes, which excessively downweighed the effects of false positives. This is especially problematic if the QTL effect sizes are gamma distributed with small shape parameter value (i.e., a strongly leptokurtic distribution), where the large number of QTL with small effect sizes heavily reduces the weight $\frac{\sum_{i=1}^{nQTL} |a_i|}{nQTL}$ in equation [13].

This can be mitigated by excluding QTL with effect size below certain cut-off point a_{min}. While in theory such exclusion could affect the balance for the $TP_{THR_{wt}}$ in equation [14], in practice such exclusion has minimal effects on the $TP_{THR_{wt}}$, as the true positives are generally overrepresented by detection of QTL with large effect size. This is also in line with the aforementioned impracticality of expectation for a GWAS to detect QTL with small effect sizes. With this in mind, given an effect size cut-off point a_{min}, the optimal threshold from equation [14] was redefined as follows:

$$THR_{opt_{wte}} = \arg\max \left(\sum_{i \in TP, a_i \geq a_{min}} |a_i| - FP_{THR} * \left(\frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} |a_i|}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i} \right) \right) \quad [15]$$

This effect size weighted optimal threshold would henceforth denoted as $THR_{opt_{wte}}$.

Another option for weightage is the additive genetic variances explained by the QTL. With the effect size cut-off point applied, the version of weighted optimal threshold could also be defined as follows:

$$THR_{opt_{wtq}} = \arg\max \left(\sum_{i \in TP, a_i \geq a_{min}} 2p_i(1-p_i)a_i^2 - FP_{THR} * \left(\frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} 2p_i(1-p_i)a_i^2}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i} \right) \right) \quad [16]$$

This version of weighted optimal threshold would henceforth denoted as $THR_{opt_{wtq}}$.

4.3.4. Incorporating the Effects of Correlation Between Markers in ROC Curve

4.3.4.1. The Effects of Correlations on ROC Curve

While the previous algorithm can be used to establish the optimal threshold for a GWAS experiment, it assumed the markers to be independent from one another. This is not realistic in actual GWAS, as some correlation between markers are to be expected. This however has a significant effect on the optimality of a threshold, with examples provided in Figure 4.2. Modification of the original model would thus be required, and with this, the impact of correlation on the ROC curve would need to be established.

One of the main effects of correlation between the markers in a genotype array is the "bleeding" of the effect of true QTLs into its neighbouring null markers, which produces a peak with several null markers flanking a core with true QTL (an example was provided in Figure 3.7). Through additional simulations, the expected amount of effect size received by null marker j from a correlated QTL i can be calculated as follows:

$$a_j = a_i * R_{LD}(i,j) \quad [17]$$

Where a_j is the apparent effect size of a null marker, a_i being the effect size of the QTL, and $R_{LD}(i,j)$ is the linkage disequilibrium between the QTL and null marker. Given two loci i and j, the linkage disequilibrium between the two loci $R_{LD}(i,j)$ can be calculated as follows:

$$R_{LD}(i,j) = \frac{(f_{11}f_{00} - f_{01}f_{10})}{\sqrt{p_i(1-p_i)p_j(1-p_j)}} \quad [18]$$

Where f_{xy} is the haplotype frequency for genotype x in locus i and genotype y in locus j, and p_i and p_j are allele frequency for first and second loci respectively (Mueller, 2004).

Another effect of correlation is the interaction between several correlated QTL. When there were several correlating QTL, they acted synergistically as their apparent effect sizes, scaled by their correlation, combine additively to one another. Indeed, with several correlating QTL, the expected apparent effect size of a QTL of locus j, $\widehat{a_j}$, can be modelled as follows:

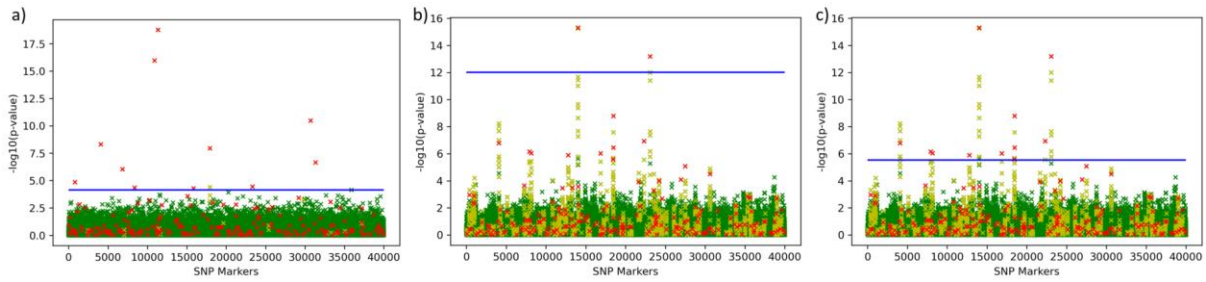

Figure 4.2: An example of the effects of correlation between markers on the thresholds estimated. Figure (a) features a Manhattan plot for GWAS for independent markers, while (b) and (c) featured Manhattan plots with correlated markers with marker pairs correlation set at $R_{LD} = 0.8$. The blue line from each of the figures is the optimal threshold. The optimal threshold presented in figure (a) and (b) is calculated using the original ROC curve, while the threshold in (c) differentiates between the markers correlated with QTL (yellow) from markers uncorrelated with QTL (green). The threshold is calculated using the generalized ROC curve with $R^2_{cut} = 0.05$, $w_t = 1, w_c = 0$ and $w_u = -1$.

$$\widehat{a_j} = \sum_{i=1}^{nQTL} a_i \, R_{LD}(i,j) \qquad [19]$$

Where $R^2{}_{LD}(i,j)$ is the linkage disequilibrium between QTL i and j.

Correlation also altered the distribution of the estimated effect sizes and test statistics of the markers. For independent markers, the test statistics of the true and null markers tend to be rather well-distinguished, with the QTL produces the peaks and null markers formed the base. Thus, one simple threshold can be used to separate the peaks from the base, effectively classifying the null and true markers (an example of such distribution of estimated effect sizes was provided in Figure 4.2(a)). This is not the case for correlated markers however. Due to the "bleeding" effects, the null markers can have their test statistics comparable or even exceed the QTL. This blurs the boundary between QTL and null markers, and if the aforementioned ROC curve is applied, the threshold would be overly stringent as it attempted to exclude the correlated null markers (an example of this overly stringent threshold was provided in Figure 4.2(b)). The original ROC model would thus need to be modified to accommodate the null correlated markers.

One applicable modification is the introduction of an additional threshold which separates the null markers based on their squared level of correlation with the QTL. This additional threshold, denoted as R^2_{cut}, act on the correlation between the QTL and null markers. It separates the markers into three classes: the QTL, the null markers that had correlation

greater than R^2_{cut} with any QTL, and null markers that had its correlation less than R^2_{cut} with all QTL. The number of positives from each class at a threshold THR were denoted as TP_{THR}, FPC_{THR} and FPU_{THR} respectively. Weights can then be assigned to each of the classes, and the ROC score (ROC_{THR}) can then be calculated as follows:

$$ROC_{THR} = w_t TP_{THR} + w_c FPC_{THR} + w_u FPU_{THR} \quad [20]$$

Where w_t, w_c and w_u are defined as the weights for true positives, false but correlated positives and false and uncorrelated positives. The optimal threshold can then be defined as follows:

$$THR_{opt} = \arg\max \left(w_t TP_{THR} + w_c FPC_{THR} + w_u FPU_{THR}\right) \quad [21]$$

This modification can also be seen as a generalization of the original optimal threshold calculation, which is equivalent to the optimal threshold from equation [21] for correlated markers with the term $w_t = 1$, $w_c = 0$ and $w_u = -1$. An example of such threshold using correlated ROC curve is provided in Figure 4.2(c).

4.3.4.2. Generalization of the ROC Curve and THR_{opt} calculation under Correlated Marker System

As in aforementioned generalization of ROC curve and optimal threshold calculation, a similar generalization can also be applied to the ROC curve under a correlated marker system. While modification is not required for the calculation of the weighted number of true positives, it is required for the calculation of the weighted number of false positives.

By classifying the number of false positives into correlated and uncorrelated false positives, it is in effect converting the FP_{THR} from equation [11] into FPC_{THR} and FPU_{THR} in equation [21], and converting the original weight from -1 into a user-definable w_c and w_u respectively. Recognizing these adjustments, the weighted number of false positives can then be calculated by substituting $w_c FPC_{THR}$ and $w_u FPU_{THR}$ in place of FP_{THR} into equation [13]:

$$FP_{THR_{wt}} = (w_c FPC_{THR} + w_u FPU_{THR}) * \left(\frac{\sum_{i=1}^{nQTL} |a_i|}{nQTL}\right) \quad [22]$$

which, when adjusted for increased robustness against the large number of QTL with small effect size, yields the following:

$$FP_{THR_{wt}} = (w_c FPC_{THR} + w_u FPU_{THR}) * \left(\frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} |a_i|}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i}\right) \quad [23]$$

and the corresponding effect size weighted optimal threshold can be calculated as follows:

$$THR_{opt_{wte}} = \arg\max\left(\sum_{i \in TP, a_i \geq a_{min}} |a_i| + \frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} |a_i|}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i} (w_c FPC_{THR} + w_u FPU_{THR}) \right) \quad [24]$$

The optimal threshold weighed by the proportion of additive genetic variance explained can also be generalized as follows:

$$THR_{opt_{wtq}} = \arg\max\left(\sum_{i \in TP, a_i \geq a_{min}} 2p_i(1-p_i)a_i^2 + \frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} 2p_i(1-p_i)a_i^2}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i} (w_c FPC_{THR} + w_u FPU_{THR}) \right) \quad [25]$$

The calculation of the generalized weighted optimal tresholds under a correlated marker system as defined in equation [21], [24] and [25] is provided in Figure 4.3.

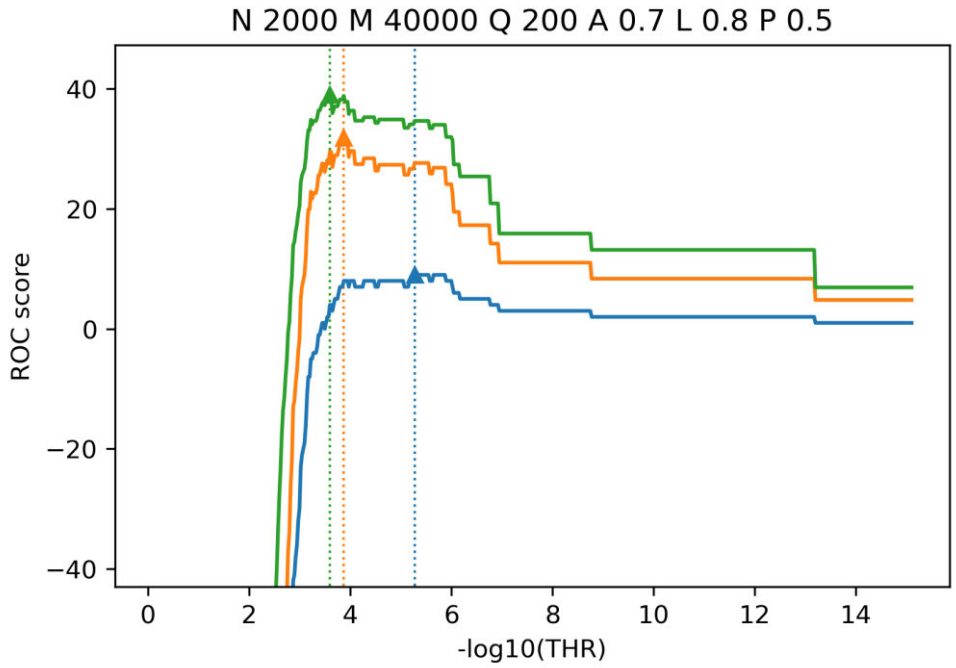

Figure 4.3: The calculation of generalized weighted optimal tresholds featured in section 4.3.4.2. The solid lines depicts the ROC scores for each of the generalization under varying threshold level, and the dotted vertical line depicts the optimal thresholds. The blue lines represents the generalization as shown in equation [21], the orange lines as in equation [24] and green lines as in equation [25]. For all three methods, the sample size was set at 2,500, with number of markers set at 40,000, the pairwise marker correlation set at 0.8. 200 QTL was simulated in this example, with the distribution of the QTL effect size follows a gamma distribution $Gamma(0.7, 1)$, and the "trivial effect size" a_{min} were set at bottom 30% (i.e. only top 70% of all QTL are considered in the calculation of the power). The ROC scores are calculated with $R^2_{cut} = 0.2$, $w_t = 1$, $w_c = 0$ and $w_u = -1$.

4.3.4.3. Calculation of Optimal Threshold for a Highly Polygenic Trait

An important consideration for the classification of false positives is the effect of the proportion of null markers that would be classified as correlated with a QTL. As an example, in a marker array with 100 QTL, if the markers are correlated in such a way that in average 10 markers flanked each side of a QTL has a correlation greater than R^2_{cut}, there were 2*10*100 = 2000 markers being marked as correlated, and the FPC_{THR} would describe the number of positives within these 2000 markers. While this is not so problematic if the trait is oligogenic or if the markers are independent to one another, it might be if both conditions are not met. This could be the case if the trait is strongly polygenic, which might cause most if not all the null markers being marked as correlated with a QTL, producing an overly lenient threshold that fails to exclude the false positives.

Several approaches could be taken to mitigate such an issue. One such approach was to increase the stringency of the effect size threshold a_{min}, or the linkage disequilibrium threshold R^2_{cut}, or by increasing the penalty for w_c (i.e., a negative value for the weight w_c). A less arbitrary method however is by modifying how the weighting factors are defined, and one such way is by assigning weights to each marker based on their correlation with a QTL.

Given a marker $i \in \{1,2,3,\dots,nSNP\}$, let $\boldsymbol{r^2_a}$ be a vector of length $nQTL$ that is defined as follows:

$$\boldsymbol{r^2_a} = \left[R^2_{LD}(i,QTL_1)a_1 \quad R^2_{LD}(i,QTL_2)a_2 \quad R^2_{LD}(i,QTL_3)a_3 \quad \dots \quad R^2_{LD}(i,QTL_{nQTL})a_{nQTL}\right]$$

Where $R^2_{LD}(i,QTL_n)$ is the squared correlation between marker i and QTL of locus n, and a_n being the effect size associated with said QTL. Using the linkage disequilibrium and effect size threshold R^2_{cut} and a_{min}, the vector is trimmed such that any $R^2_{LD}(i,QTL_n)a_n$ that has R^2_{LD} less than R^2_{cut} and a_n less than a_{min} are removed from the calculation. This is to ensure that only null markers that have sufficiently strong correlation with a QTL with significant effect sizes being marked as correlated false positives if the marker is deemed significant in a GWAS. For the remaining $R^2_{LD}(i,QTL_n)a_n$, the squared correlation component $R^2_{LD}(i,QTL_n)$ was extracted, and the maximum of the $R^2_{LD}(i,QTL_n)$, denoted as $R^2_{LD_{max}}$, were obtained for marker i. The $R^2_{LD_{max}}$ would represents the correlation between marker i and the nearest QTL. This procedure is then repeated for all $nSNP$ markers. If the marker tested is a QTL of

significant effect size, it would receive a score $R^2_{LD_{max}} = 1$. If the marker is uncorrelated with any of the QTL it would receive a score of $R^2_{LD_{max}} = 0$.

From the pool of $R^2_{LD_{max}}$ of all markers, a new vector (denoted as $\boldsymbol{r_a^{2*}}$) can be built using the following rules. If the marker is a nontrivial QTL (i.e. $R^2_{LD_{max}} = 1$ and $a_n \geq a_{min}$), then the entry for the new vector was assigned with the value of w_t. If the null marker bears no significant correlation with any significant QTL (i.e. $R^2_{LD}(i, QTL_n)a_n < R^2_{cut}$ or $a_n < a_{min}$) then it was assigned with the value of w_u. For the remaining markers, they were deemed as correlated null markers (i.e. $R^2_{LD}(i, QTL_n)a_n \geq R^2_{cut} > 1$ and $a_n \geq a_{min}$) and were assigned with the median-adjusted $R^2_{LD_{max}}$. In summary, the $\boldsymbol{r_a^{2*}}$ is a vector built with its entries under the following rule:

$$\boldsymbol{r_a^{2*}}(n) = \begin{cases} w_t & , R^2_{LD_{max}} = 1 \wedge a_n \geq a_{min} \\ R^2_{LD_{max}}(n) - median\left(R^2_{LD_m}\right) & , R^2_{LD}(i, QTL_n)a_n \geq R^2_{cut} > 1 \wedge a_n \geq a_{min} \\ w_u & , R^2_{LD}(i, QTL_n)a_n < R^2_{cut} \vee a_n < a_{min} \end{cases} \quad [26]$$

where $median(R^2_{LD_m})$ is the median of all $R^2_{LD_{max}}$ that fulfils the condition of being a correlated null marker (i.e. $R^2_{LD}(i, QTL_n)a_n \geq R^2_{cut} > 1$ and $a_n \geq a_{min}$). The rationale of using the median is to ensure the balance between the number of correlated null markers with positive weights and those with negative weights. By this rule, markers that are closer or more correlated with a QTL were assigned a positive weight, and those that further away with a negative weight.

Under this new definition of weights, the number of correlated positives for each class can then be calculated as the sum of $\boldsymbol{r_a^{2*}}(n)$ of the correlated false positives:

$$FPC_{THR} = \sum_{i \in FPC} \left(R^2_{LD_{max}}(i) - median\left(R^2_{LD_m}\right)\right) \quad [27]$$

with the index of summation $i \in FPC$ denoting a set of correlated null markers being marked as positive by the GWAS. The optimal threshold can then be calculated as the sum of the positives weighted by vector $\boldsymbol{r_a^{2*}}$:

$$THR_{opt_r} = \arg\max\left(w_t \sum_{i \in TP} i + \sum_{i \in FPC} \left(R^2_{LD_{max}}(i) - median\left(R^2_{LD_m}\right)\right) + w_u \sum_{i \in FPU} i\right) \quad [28]$$

which can be simplified in terms of $\boldsymbol{r_a^{2*}}$:

$$THR_{opt_r} = \arg\max\left(\sum_{i=1}^{nSNP} \boldsymbol{r_a^{2*}}(i) * \boldsymbol{\pi_{THR}}(i)\right) \quad [29]$$

where $\boldsymbol{\pi_{THR}}$ is a vector of size $nSNP$ containing the acceptance-rejection status of all SNP markers under a threshold THR. Finally, with the weights vector $\boldsymbol{r_a^{2*}}$, the effect size weighted optimal threshold can be calculated as such:

$$THR_{opt_{wter}} = \sum_{i \in TP} \boldsymbol{r_a^{2*}}(i) * \boldsymbol{a}_i + \frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} |a_i|}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i} * \left(\sum_{i \in \{FPC \cup FPU\}} \boldsymbol{r_a^{2*}}(i)\right) \quad [30]$$

where the index of summation $i \in \{FPC \cup FPU\}$ denoting a set of correlated and uncorrelated false positives. Similarly, the optimal threshold weighted by $\boldsymbol{r_a^{2*}}$ and proportion of additive genetic variance explained can be defined as such:

$$THR_{opt_{wtqr}} = \sum_{i \in TP} \boldsymbol{r_a^{2*}}(i) * \boldsymbol{p}_i(\boldsymbol{1} - \boldsymbol{p}_i)\boldsymbol{a_i^2} + \frac{\sum_{i=1, a_i \geq a_{min}}^{nQTL} 2p_i(1 - p_i)a_i^2}{\sum_{i=1, a_i \geq a_{min}}^{nQTL} i} * \left(\sum_{i \in \{FPC \cup FPU\}} \boldsymbol{r_a^{2*}}(i)\right) \quad [31]$$

An example of calculation of the generalized weighted optimal tresholds under a correlated marker system as defined in equations [25], [29], [30] and [31] is provided in Figure 4.4.

4.4. Simulation Study

The optimality of the threshold was evaluated using simulated genotype and phenotype using Python (version 3.7.3), where the power and false positive rate of the optimal threshold, alongside with the Bonferroni correction and Benjamini-Hochberg False Discovery Rate (BH-FDR), are tested under varying parameters and correlation between markers. The true number of QTL, the distribution of their effect sizes and their locations were assumed to be known for this chapter (i.e. these quantities have been obtained from raw data through methods outside this chapter).

4.4.1. Genome Wide Association Study

To simulate a GWAS experiment, a genotype array (denoted as $\boldsymbol{X}$) with sample size N and number of markers $nSNP$ was generated, with the distribution of the allele frequencies following a Beta distribution. The sample size, number of markers and the shape parameters of the Beta distribution are provided in Table 4.2. The correlation between markers was generated by copying part of the genotypic state from a marker to the adjacent markers while

randomizing the remaining genotypic state for that adjacent marker. The proportion of copying were the targeted level of correlation, which is provided in Table 4.2.

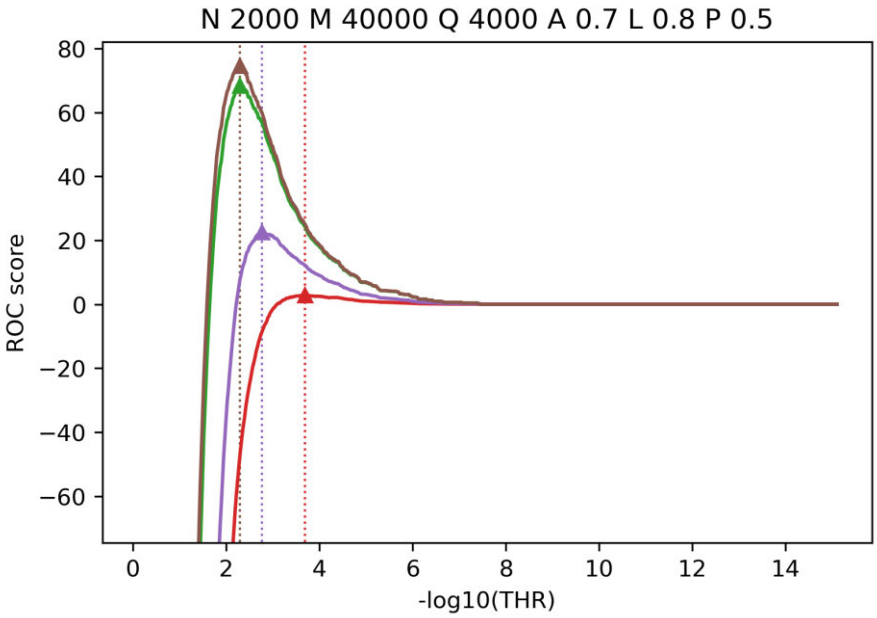

Figure 4.4: The calculation of generalized optimal threshold from section 4.3.4.3 based on varying correlation-based weighting methods. The solid lines depicts the ROC scores for each of the generalization under varying threshold level, and the dotted vertical line depicts the optimal thresholds. The red lines denotes the weighting algorithm as defined in equation [25], the purple lines with equation [29], green lines with equations [30] and brown lines with equation [31]. For all three methods, the sample size was set at 2,500, with number of markers set at 40,000, the pairwise marker correlation set at 0.8, repeated 100 times. 4000 QTL was simulated in this example, with the distribution of the QTL effect size follows a gamma distribution $Gamma(0.7, 1)$, and the "trivial effect size" a_{min} were set at bottom 30% (i.e. only top 70% of all QTL are considered in the calculation of the power). The ROC scores are calculated with $R^2_{cut} = 0.2$, $w_t = 1$, $w_c = 0$ and $w_u = -1$.

Some of the markers were marked as QTL and were associated with an effect size. The QTL effect sizes were generated at random following a gamma distribution, with its shape parameter provided in Table 4.2. The null markers were assigned an effect size of zero, and together with effect sizes of QTL, compiled into vector $\boldsymbol{a}$ containing the effect sizes of all markers. The simulated phenotypes were then generated using the method as defined in equation [3], with the residual vector $\boldsymbol{e}$ simulated with the following normal distribution:

$$\boldsymbol{e} \sim N\left(0, \frac{var(\boldsymbol{Xa}) * (1 - h^2)}{h^2}\right) \quad [32]$$

For this experiment, the narrow sense heritability h^2 was set at 0.3 for all parameter tested. Using the simulated genotype and phenotype, a single SNP regression GWAS was conducted, with their negative logarithmically transformed p-values $logpval$ being recorded.

Based on the QTL effect size and the level of correlation between marker and nearest QTL, the loci were classified into three classes: true QTL, correlated null markers and uncorrelated null markers, with the linkage disequilibrium threshold R^2_{cut} set at 0.2 and the trivial effect size threshold a_{min} set at bottom 30% of all QTL (i.e., only top 70% of all QTLs were included in the calculation). The number of true positives TP_{THR}, correlated false positives FPC_{THR} and uncorrelated false positives FPU_{THR} were also recorded.

4.4.2. Threshold Tested in this Experiment

In this experiment, the performance of various thresholds in maintaining the power and false positive rate were tested. Eight thresholds were tested in this experiment, with six of them being the variants of the optimal thresholds, and the others being based on the Bonferroni method and the Benjamini-Hochberg FDR method. The threshold for the Bonferroni method (THR_{BON}) is defined as follows:

$$THR_{BON} = -\log_{10}\left(\frac{Type\ 1\ Error}{nSNP}\right) \quad [33]$$

Given a set of ordered negative logarithmically transformed p-values $logpval_{sorted}$, the threshold for the Benjamini-Hochberg FDR is defined as the smallest $logpval_{sorted}$ that fulfil the following inequality (Simes, 1986):

$$logpval_{sorted}(j) \leq -\log_{10}\left(\frac{j * Type\ 1\ Error}{nSNP}\right) \quad [34]$$

Where j is the index of the sorted negative logarithmically transformed p-values.

The notations used for the thresholds tested in this experiment, along with their associated equations, are provided in Table 4.1.

For the optimal threshold testing that utilized QTL-null markers correlation as weighting factors (i.e., UWTR, WTER and WTQR), the vector $\boldsymbol{r}_{\boldsymbol{a}}^{2*}$ was built using the rule as defined in equation [26], with the weight $w_t = 1$ and $w_u = -1$. While for optimal thresholds that did not use correlation as a weighting factor (i.e., UWT, WTE and WTQ), the true positives,

correlated false positives and uncorrelated false positives were assigned weights as $w_t = 1$, $w_c = 0$ and $w_u = -1$, respectively.

Besides the thresholds, two additional controls were employed. The first control involves randomly selected markers (henceforth denoted as "RND") and that involves the calculation of the same score for all QTL (denoted as "ALQ"). To ensure the comparability of the controls with the thresholds, the number of markers utilized in both controls equated those obtained from the most lenient threshold (i.e., threshold that yielded the greatest number of positives). In the cases when there are more positives in the most lenient threshold than number of QTL, the ALQ is padded with loci that have the strongest LD with any non-trivial QTL.

Table 4.1: Notations, description and equations of threshold tested in this experiment.

Notations	Description	Equation
UWT	Unweighted ROC optimal threshold	[21]
WTE	ROC optimal threshold weighted by effect size	[24]
WTQ	ROC optimal threshold weighted by additive genetic variance explained	[25]
UWTR	ROC optimal threshold weighted by weightage vector $\boldsymbol{r_a^{2*}}$	[29]
WTER	ROC optimal threshold weighted by effect size and weightage vector $\boldsymbol{r_a^{2*}}$	[30]
WTQR	ROC optimal threshold weighted by additive genetic variance explained and weightage vector $\boldsymbol{r_a^{2*}}$	[31]
BON	Bonferroni correction	[33]
BHF	Benjamini-Hochberg FDR	[34]
RND	Random control	NA
ALQ	All-QTL control, padded with null markers with strongest LD with QTL if needed	NA

4.4.3. Parameter Tested in this Experiment

The list of parameters tested, alongside with their default and alternative values, are provided in Table 4.2. When a parameter is tested under its alternative value, default values were used for other parameters.

Table 4.2: Parameter tested in this experiment.

Parameters	Default Value	Alternative Values
Sample Size (N)	2500	1000, 4000
Number of Markers (M)	40k	400k
Number of QTL (Q)	1000	200, 4000
QTL Effect Size Distribution (A)	Gamma(0.5,1)	Gamma(0.2,1), Gamma(0.9,1)
Pairwise Marker Correlation (L)	0.8	0.1, 0.5
Allele Frequency Distribution (P)	Beta(0.5,0.5)	Beta(0.2,0.2), Beta(0.8,0.8)

For the number of markers of 400k, two levels of pairwise marker correlations were also tested: 0.8 and 0.9779. The latter value was chosen as it is the expected pairwise marker correlation had the 400k marker density is applied onto a genome that would yield a correlation of 0.8 if genotyped on a 40k density. The 400k marker test with pairwise marker correlation of 0.8 will henceforth notated as "400k" and those with correlation of 0.9779 was denoted as "400k*".

4.4.4. Testing the Performance of a Threshold

To test the performance of a threshold, two different measures were utilized.

4.4.4.1. Matthews correlation coefficient (MCC)

The first measure is the Matthews correlation coefficient (MCC), which has been used to test the performance of a threshold as a binary classifier (Boughorbel et al., 2017; Chicco and Jurman, 2020). The rationale of choosing MCC over other measures of performance is its insensitivity toward extreme class imbalance, which is important as the number of null markers in GWAS generally outweighed the number of QTL (Boughorbel et al., 2017).

The MCC is defined as follow (Chicco and Jurman, 2020):

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN) * (TP + FP) * (TN + FP) * (TN + FN)}} \quad [35]$$

Where TP, TN, FP and FN are the number of true positives, true negatives, false positives and false negative respectively. For this experiment, only the uncorrelated null markers were included in the calculation of TN and FP.

The MCC score ranges between -1 and 1, which the extremes represent perfect misclassification and perfect classification respectively. For random classifier the expected score would be 0. If equation [35] is undefined due to the denominator being zero, a score of zero would be assigned as the MCC score (Chicco and Jurman, 2020; Gorodkin, 2004).

It shall be noted that the maximum MCC score attained in this experiment may not necessarily be 1, even for ALQ. This is due to the fact that ALQ comprises not only all the QTL but also the neighbouring markers with the strongest LD with any of the non-trivial QTL. In such a case, the MCC scores for ALQ serves as the maximum score that can be attained by a threshold given a set number of positives. This also makes ALQ a more informative control as it is conditioned by the baseline feasibility of detecting a QTL in a GWAS.

4.4.4.2. Genomic Prediction Accuracy

The second measure utilized to test the performance of the threshold is the accuracy of genomic prediction with marker selection. Previous studies such as Moghaddar et al. (2019) suggested that causal loci made up only a small percentage of the genome, which theoretically allowed markers with low linkage with a QTL to be excluded without significant impact on the accuracy. For this test the thresholds were used to exclude markers with low linkage with a QTL.

For this method, a new, unrelated population of the same sample size and genotyping density as in the test population was simulated. This population was generated using the same algorithm as in the test population for GWAS and threshold calculation. Using the previously generated QTL, the additive genetic component of the phenotype was calculated for the new population as follows:

$$\boldsymbol{y_{new}} = \boldsymbol{X_{new}a} \quad [36]$$

This additive genetic component of the phenotype was treated as the true breeding value (TBV) of the new population.

From each of the thresholds, all the positive loci were obtained. To ensure the full coverage of the genome for the genomic prediction, and to minimize the variability in accuracy due to differing number of positives from each threshold, as well as to remove the confounding effects between the representativeness of selected markers toward kinship between individuals and the accuracy of genomic prediction, the positive loci were padded with a number of random markers up to a density of 10k. This is to test how much better a threshold is in differentiating causal loci from null loci when compared to pure random selection of the markers. Using the padded genotype array, a Genomic Relationship Matrix (GRM) was built using the method suggested by VanRaden (2008). With the assumption of the true h^2 being known, the GRM would then be used to calculate the estimated breeding value (EBV) of the new population, using method as suggested by Gondro (2015).

With the true and estimated breeding values, the performance of the threshold was evaluated as the accuracy of the genomic prediction. The prediction accuracy is defined as the correlation between the true and estimated breeding value of the new population, with a higher accuracy indicating improved optimality of the threshold.

For the two controls (i.e., "RND" and "ALQ"), the former involves the building of GRM and calculating of accuracy using 10k random markers (denoted as "RND"), and the latter involves the building of GRM using all the QTL, padded with their closest neighbouring null markers up to a density of 10k (denoted as "ALQ").

4.4.5. Statistical Test on Effects of Thresholds and Parameters

To ensure the consistency of the results, steps from section 4.4.1 up to section 4.4.4 was repeated 100 times, and the results presented were the mean from all the repeats. To compare the significance of differences between thresholds tested, a pairwise t-test was utilized. To compare the significance of differences between parameter tested, a Welch's independent t-test was utilized. A comparison is deemed significantly different if the negative logarithmically transformed p-value ($logpt = -\log(p - value_{t-test})$) of the t-test is more than 3.

4.5. Results

4.5.1. Threshold Calculated

The thresholds calculated from each of the parameter tested are provided in Figure 4.5.

Generally, the thresholds calculated from the ROC curve tend to be more lenient compared to the Bonferroni method, with the main exception from a trait with small number of QTL (i.e. $Q = 200$) or with small sample size ($N = 1000$), where the threshold from unweighted ROC curve is more stringent than those of Bonferroni correction.

Under the default conditions, the threshold from BHF is comparable to UWT ($logpt = 4.01$) and UWTR ($logpt = 1.07$), but is significantly more stringent than other variants of ROC curve-based thresholds ($logpt$ from 15.67 for WTE to 41.58 for WTQR). This is not the case for other parameter values tested however; the threshold from BHF becomes significantly more lenient than some of the ROC-based thresholds when the markers are strongly correlated to one another (e.g. $L = 0.9779$ at "400k*" dataset) ($logpt$ from 0.03 for WTQ to 21.20 for UWT), with large sample size ($N = 4000$) ($logpt$ from 4.14 for WTE to 25.94 for WTQR), and with more leptokurtic QTL effect size distribution ($A = gamma(0.2,1)$) ($logpt$ from 0.40 for WTQ to 20.56 for UWT), and become more stringent than ROC-based thresholds when the trait is oligogenic ($Q = 200$) ($logpt$ ranges from 9.46 for WTQR to 16.22 for UWT).

The use of weighting factors also has significant effects on the threshold calculated from the ROC curve. For all parameters tested, the thresholds unweighted by effect sizes or additive genetic variance (i.e., UWT, UWTR) are more stringent than their weighted counterparts ($logpt$ from 13.04 between UWTR and WTER to 26.51 between UWT and WTQ under default set of parameters). The use of effect size as weighting factor in the ROC-based threshold (i.e., WTE and WTER) has yielded a more stringent threshold compared to those that utilized additive genetic variance (i.e., WTQ and WTQR) ($logpt = 7.87$ between WTE and WTQ; $logpt = 6.34$ between WTWR and WTQR under default set of parameters). The use of QTL-null marker correlation as weighting factor (i.e., the vector $\boldsymbol{r_a^{2*}}$) has generally yielded a slightly more lenient threshold than those unweighted by this weighting factor ($logpt$ ranges from 0.27 between WTQ and WTQR to 1.22 between UWT and UWTQ).

Besides the type of correction methods and weighting factors utilized, changing parameter values also have profound effects on the threshold calculated from each of the methods. With increased polygenicity of a trait, the thresholds calculated from the ROC decreased in stringency ($logpt$ ranges from 8.66 for UWT to 16.38 for WTQR with the increase in number of QTL from 200 to 4000). The opposite trend is true for BHF however, with the threshold increased in stringency with polygenicity ($logpt = 34.03$). A similar trend is also

observed as the kurtosis of the effect size distribution decreases; by changing the effect size distribution from $A = gamma(0.2,1)$ to $A = gamma(0.9,1)$, which reduces the kurtosis of the distribution by 4.5-folds (Mun, 2012), the ROC-based thresholds significantly decreased in stringency ($logpt$ from 10.54 for WTE to 24.07 for WTQR), whereas the BHF significantly increased in stringency ($logpt = 24.44$).

Increasing the sample size also significantly decreases the stringency for both ROC-based threshold and BHF ($logpt$ from 3.15 for WTQ to 13.18 for UWT with increase of sample sizes from 1000 to 4000, $logpt = 65.82$ for BHF). Increased correlation between markers significantly decreases the stringency of threshold from BHF ($logpt = 83.14$), but no significant changes in ROC-based threshold was observed with changing correlation. The allele frequency distribution does not have a significant effect on the thresholds. With the exception of number of markers, the threshold for BON does not change significantly with varying parameter values. Increasing the number of markers from 40k to 400k increases the stringency of all ROC-based thresholds, as well as thresholds from BHF and BON.

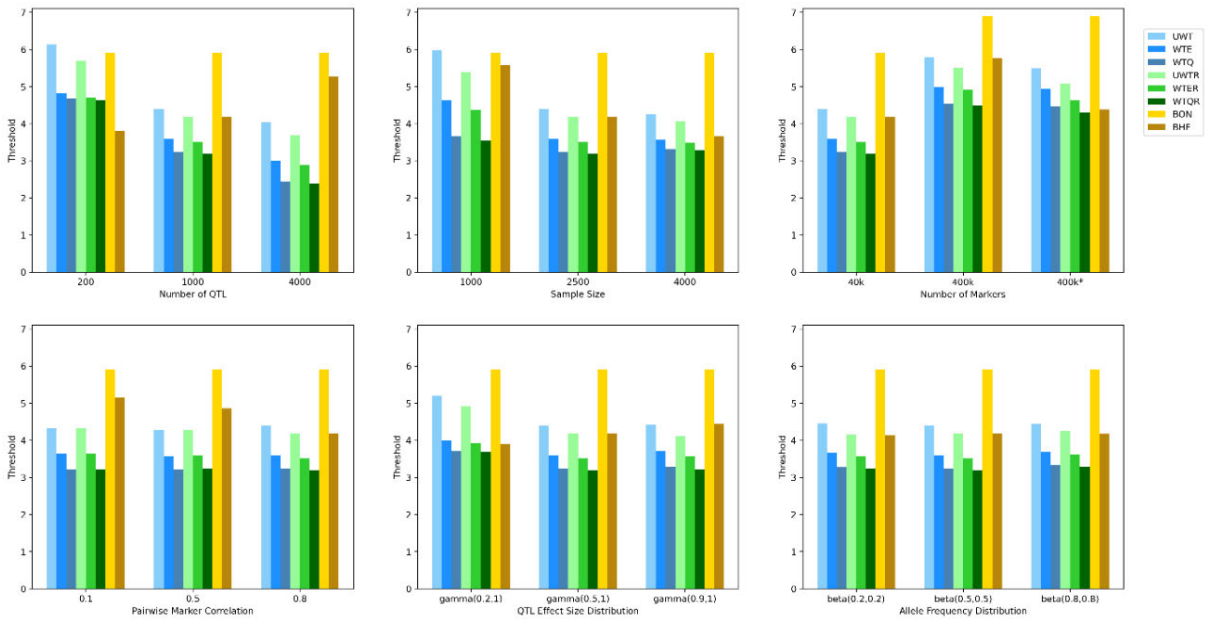

Figure 4.5: Threshold of GWAS obtained from simulation under varying parameter values. Thresholds featured in this figure include: Unweighted ROC-based threshold (UWT), ROC-based threshold weighted by effect size (WTE), ROC-based threshold weighted by additive genetic variance (WTQ), ROC-based threshold weighted by correlation weighting factor $\boldsymbol{r_a^{2*}}$ (UWTR), ROC-based threshold weighted by effect size and correlation weighting factor (WTER) and ROC-based threshold weighted by correlation weighting factor and additive genetic variance (WTQR), alongside with threshold from the Bonferroni correction (BON) and Benjamini-Hochberg False Discovery Rate (BHF).

4.5.2. Matthews correlation coefficient (MCC)

The effects of varying types of thresholds, weighting factors included, and parameter value tested are provided in Figure 4.6.

Consistent with the previously published literature, the MCC score for randomly chosen markers (RND) is effectively zero, whereas maximal MCC scores were observed for all-QTL control (ALQ). Excluding the controls, for all the parameter tested, the MCC score from BON threshold is the lowest, and this is followed by those calculated from BHF, and the significant markers yielded from ROC-based threshold has the highest MCC score. The decline in MCC scores in BON and BHF could be attributed to the increased stringency of its threshold. Under default parameter values, the MCC scores from ROC-based threshold ranges from 48.8% to 54.9% higher than that of BON ($logpt$ from 30.03 for UWT to 35.39 for WTQR), and 7.0% to 11.0% higher than that of BHF ($logpt$ from 1.71 for UWT to 2.77 for WTE). Compared to BON, the MCC score from significant markers from BHF is in general more similar to those from ROC-based threshold. The main exception is when the trait is strongly polygenic or when the sample size is small, where in both cases positives from BHF yielded significantly less MCC score than those of ROC-based threshold ($logpt$ from 16.17 for UWT to 52.73 for WTQR).

In general, the use of weights in the ROC-based threshold increases the MCC score. Unweighted optimal thresholds (i.e., UWT and UWTR) in general yielded the lowest MCC score, and this is followed by those weighted by proportion of variance explained (i.e., WTQ and WTQR) and finally by effect size (i.e., WTE and WTER), although the differences between the latter two are generally not considered to be significant. The only exception is when the trait is polygenic (i.e. $Q = 4000$) where those weighted by effect sizes have lower MCC scores than those weighted by proportion of variance explained ($logpt = 4.48$ between WTE and WTQ; $logpt = 4.67$ between WTER and WTQR). The differences in MCC scores due to correlation weights $\boldsymbol{r_a^{2*}}$ in the ROC-based threshold calculation are not significant.

Increasing the sample size increases the MCC score for all threshold tested ($logpt$ ranges from 80.27 for UWT to 102.73 for BHF with an increase in sample size from 1000 to 4000), while a high polygenicity decreases the score ($logpt$ from 68.36 for WTQR to 75.93 for BHF with an increase in number of QTL from 200 to 4000). The MCC score also decreases significantly as the QTL effect size distribution changes from $gamma(0.2,1)$ to $gamma(0.9,1)$ ($logpt$ from 32.28 for UWT to 69.11 for BON), while increasing the

pairwise marker correlation from $L = 0.1$ to $L = 0.8$ increases the score ($logpt$ from 2.08 for BON to 12.48 for BHF). Interestingly, increasing the genotyping density from 40k to 400k* significantly decreases the MCC score for the positives identified for all the threshold tested ($logpt$ from 11.38 for BON to 31.23 for WTQR). Allele frequency distribution did not have significant effects on the MCC scores.

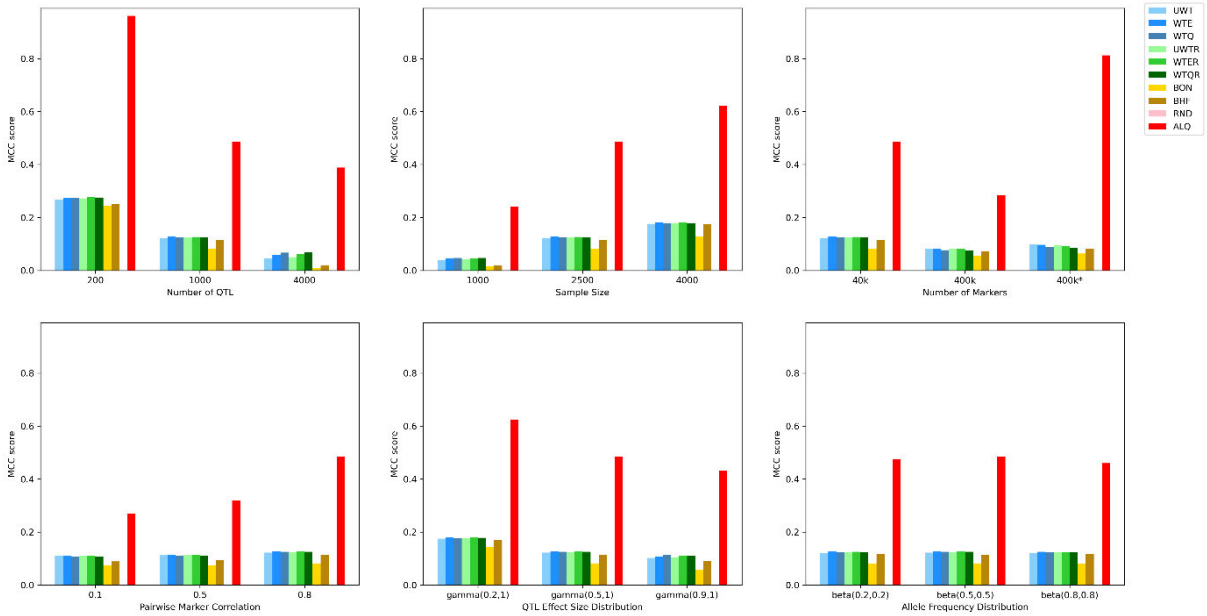

Figure 4.6: MCC scores of the positives obtained through the thresholds under varying parameter values. Thresholds featured in this figure include: Unweighted ROC-based threshold (UWT), ROC-based threshold weighted by effect size (WTE), ROC-based threshold weighted by additive genetic variance (WTQ), ROC-based threshold weighted by correlation weightage factor $\boldsymbol{r_a^{2*}}$ (UWTR), ROC-based threshold weighted by effect size and correlation weightage factor (WTER) and ROC-based threshold weighted by correlation weightage factor and additive genetic variance (WTQR), alongside with threshold from the Bonferroni correction (BON) and Benjamini-Hochberg False Discovery Rate (BHF). These scores were compared against random control (RND) and all-QTL control (ALQ).

4.5.3. Genomic Prediction Accuracy

The accuracies of truncated genomic prediction under varying types of thresholds, threshold weighting factors and parameter values were provided in Figure 4.7.

In general, the accuracies of the truncated genomic prediction from ROC-based thresholds are significantly higher compared to both BON and BHF ($logpt$ up to 28.63 between WTQR and BON; $logpt$ up to 6.35 between WTQR and BHF). The main exception was on an oligogenic trait (i.e. $Q = 200$), where the threshold from BHF yielded significantly higher

accuracy than all ROC-based thresholds tested ($logpt$ up to 8.62 for UWT). Compared to BON, thresholds from BHF have yielded significantly increased accuracy in genomic prediction in all parameters tested, and is often comparable to unweighted ROC-based thresholds. The use of ROC-based threshold increases the accuracy by up to 16.8% higher than that of BON, and up to 7.0% higher than that of BHF. For all parameters tested, truncated genomic prediction from BON has the lowest accuracy.

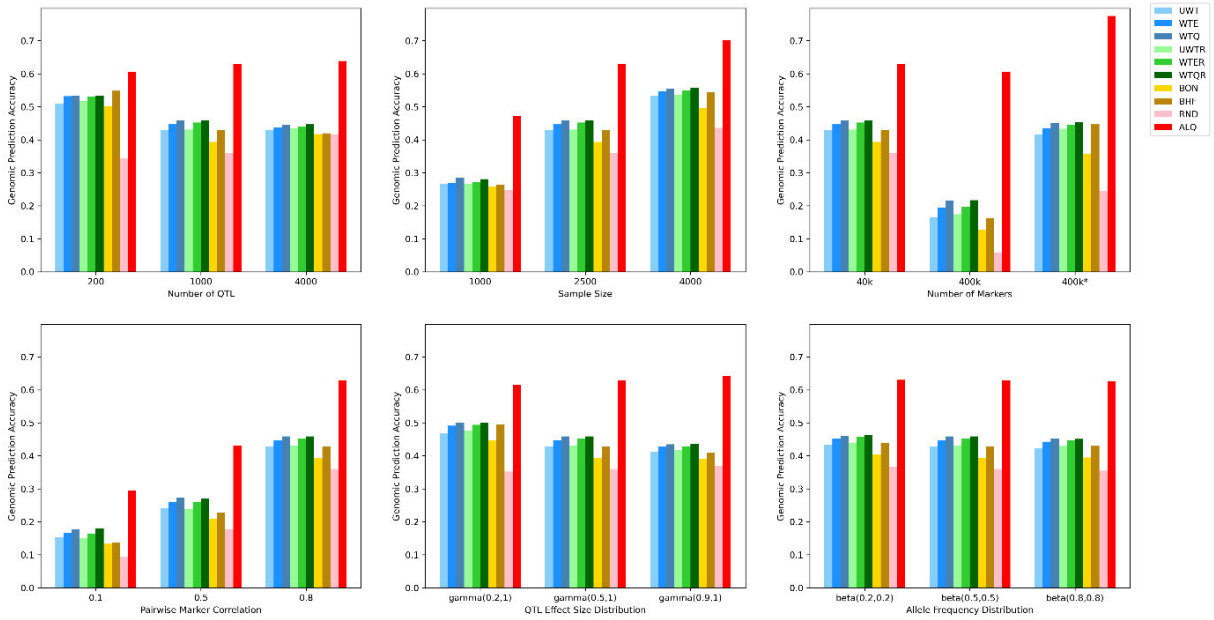

Figure 4.7: Accuracy of truncated genomic prediction calculated from positives obtained from each threshold under varying parameter values. Thresholds featured in this figure include: Unweighted ROC-based threshold (UWT), ROC-based threshold weighted by effect size (WTE), ROC-based threshold weighted by additive genetic variance (WTQ), ROC-based threshold weighted by correlation weightage factor $\boldsymbol{r_a^{2*}}$ (UWTR), ROC-based threshold weighted by effect size and correlation weightage factor (WTER) and ROC-based threshold weighted by correlation weightage factor and additive genetic variance (WTQR), alongside with threshold from the Bonferroni correction (BON) and Benjamini-Hochberg False Discovery Rate (BHF). These scores were compared against random control (RND) and all-QTL control (ALQ).

The use of weights in the calculation of ROC-based thresholds significantly increased the accuracy of the genomic prediction. For all the parameter tested, the accuracies are the lowest for unweighted optimal thresholds (i.e., UWT and UWTR), and this is followed by those weighted by effect size (i.e., WTE and WTER), and finally by proportion of variance explained (i.e., WTQ and WTQR). Compared to unweighted optimal thresholds, the use of thresholds weighted by proportion of variances led to an average of 5.90% increase in the accuracy of the genomic prediction under default set of parameters ($logpt = 7.97$ between

UWT and WTQ; $logpt = 8.36$ between UWTR and WTQR). This trend was associated with a decline in threshold stringency with the use of these weightages. While a slight increase has been detected with the use of correlation weights $\boldsymbol{r_a^{2*}}$, this increment is generally not considered to be significant.

The accuracies of truncated genomic prediction decrease with increasing polygenicity of a trait, especially for BON and BHF which by $Q = 4000$ their accuracies are comparable to those observed in random control ($logpt = 29.58$ between $Q = 200$ and $Q = 4000$ for BON; $logpt = 59.45$ for BHF). A similar observation was made for small sample size ($logpt$ up to 114.96 for WTQ between $N = 1000$ and $N = 4000$). Pairwise marker correlations increase the accuracy of the genomic prediction for all threshold tested ($logpt$ up to 141.86 for WTQR between $L = 0.1$ and $L = 0.8$), while decreases significantly as the QTL effect size distribution changes from $gamma(0.2,1)$ to $gamma(0.9,1)$ ($logpt$ up to 43.06 for BHF).

4.6. Discussion

In this study, the algorithm for the calculation of an optimal threshold was provided, and the performance of this new threshold and its variants were tested in the context of QTL detection in a GWAS and marker selection for genomic prediction. In both tests of threshold performance, the threshold from the Bonferroni method consistently had the highest stringency, and this is associated with the lowest scores for all performance criteria tested. This is in line with previous publications such as Fadista et al. (2016), Kaler and Purcell (2019) and Simes (1986), which stated the overconservativeness of the threshold calculated with the Bonferroni method, which leads to decreased power in GWAS. This is especially problematic for parameter values that reduce the $logpval$, which further reduces the number of QTL detected. This experiment further emphasized the previous notion on the unsuitability of the Bonferroni method as multiple testing correction method, and this called for a more lenient threshold such as those suggested by Benjamini-Hochberg FDR.

While the threshold calculated from Benjamini-Hochberg FDR is less conservative than those by the Bonferroni method, the FDR method also has its own issues. One major issue is that while the threshold varies with the distribution of $logpval$, the way the threshold varies does not always line up with what is required for improved performance for said threshold. This is due to the fact that the calculation of Benjamini-Hochberg FDR does not take into account

the behaviour of the $logpval$ under varying parameter values, thus causing the threshold to blindly follow the distribution of the $logpval$, resulting in a less optimal change. As an example, with a polygenic trait, the reduced $logpval$ of the QTL decrease the value of index j in equation [34], and this would push the right-hand side of the inequality upward, thus resulting in a more stringent threshold. This has effectively decimated its performance in binary classification of the markers in a GWAS and truncation for genomic prediction for a polygenic trait. A similar phenomenon was also observed for the Bonferrroni method, where its insensitivity toward the effects of parameter changes has also resulted in a less optimal threshold. This, compounded with the extreme stringency of the threshold, reduces its performance in binary classification and truncated genomic prediction.

For the calculation of the ROC-based thresholds however, it can and has successfully taken into account the effects of parameters such as genetic architecture and sample sizes on the $logpval$, as the calculation is based on the empirical distribution of the $logpval$ Taking the case of a polygenic trait as an example, as the large number of QTL has significantly reduced the $logpval$ of the QTL, the threshold needs to be more lenient in order to detect the same number of QTL. This has been reflected in the reduced stringency of the threshold from the ROC-based thresholds, thus lessening the negative impact of increased polygenicity on the performance scores; by increasing the number of QTL from 200 to 4000, the MCC score from the ROC-based threshold decreases by 75.27% for WTQR up to 82.84% for UWT, compared to 95.90% for BON and 92.80% for BHF. In genomic prediction, the accuracy declined by 15.52% for UWT up to 17.67% for WTE, compared to 16.93% from BON and 23.36% for BHF. This inclusion of the effects of parameters on the p-values means the ROC-based threshold could make necessary changes to accommodate said effects to maintain its performance.

The use of effect sizes or proportion of variable explained as weights have the effects of decreasing the stringency of the ROC-based threshold, often being more lenient than that suggested by the Bonferroni method or by Benjamini-Hochberg FDR. Despite this, these lenient thresholds were associated with an increase in both MCC score and the accuracy of truncated genomic prediction. This suggested that despite its apparent excessive leniency, the thresholds offered by the ROC algorithm a has higher optimality and performance than both multi-testing correction methods. This is especially true for the WTE threshold when tested using MCC score; while the WTE threshold has an intermediate stringency when compared with UWT and WTQ (threshold stringency $WTQ < WTE < UWT$). In general, WTE

threshold had the highest MCC scores between the three. This suggested that the threshold suggested by WTE is the closest to the optimal threshold as defined by MCC score, thus being the optimal binary classifiers. Similar arguments can also be made for ROC-based thresholds that were weighted by correlation (threshold stringency $WTQR < WTER < UWTR$, but with WTER having the highest MCC score).

The MCC score did not tell the full story however; when the accuracy of truncated genomic prediction is assessed, the WTE and WTER thresholds do not yield the highest accuracy, and instead maximal accuracies were achieved by the most lenient thresholds WTQ and WTQR, indicating the inequality of thresholds that optimize binary classification with those optimize accuracy of truncated genomic prediction. This alludes to the subjective nature of the concept of "optimal threshold" especially when they were asked in the context of differing ulterior purpose of said threshold. The concept of "optimal threshold" is essentially modelled based on certain mathematical functions. For example, just as one can define an optimal threshold based on ROC curve by calculating the point where the tangent of the curve is 1 (Kaivanto, 2008), or maximizing the Youden's Index (Habibzadeh et al., 2016), one can also define an optimal threshold based on MCC by maximizing the MCC score, or optimal threshold based on genomic accuracy by again finding a threshold that maximize such accuracy. In the end, one can define an optimal threshold by maximizing scores from any mathematical function that awards true positives while penalizing false positives, thus indicative to its subjectivity.

Perhaps a more objective question that could be asked is "how much a false positive can be tolerated?" In the context of binary classification such as those of MCC score, as false positives can easily lead to misidentification, they were penalized in a more severe manner than in truncated genomic prediction, which could accept a level of false positives as long as the true positives remained the majority in the pool of positives included into the prediction. In the context of GWAS, if the aim is to increase the proportion of additive genetic variance explained by the positive markers, then a lenient threshold might not be deleterious after all. But if the aim is to maximize the accuracy of loci identified, then a more stringent threshold might be required (Chicco and Jurman, 2020). For the thresholds suggested in this study, the use of weighting factors w_t, w_c and w_u allow the users to set prioritizations for the true and false positives, with a larger w_t promotes the detection of true positives through a more lenient threshold, and a larger w_u increases the penalization of false positives by increasing the threshold's stringency.

While this study has suggested an algorithm for calculation of an optimal threshold, one limitation for this study is that it assumed the homogeneity of the linkage disequilibrium structure, which might not be the case for a strongly inbred population with small effective population sizes (Gondro, 2015). This shortcoming could be easily overcome by identifying the effect of the QTL and their location however, and we believed this is an aspect worth further studying. Another aspect worth further studying is the estimation of genetic architecture parameters, such as number of QTL, distribution of QTL effect sizes and their location from real data, which can then be fed into this method and obtaining an optimal threshold.

Another limitation for this methodology is the arbitrariness of the trivial effect size cut-off point and linkage disequilibrium threshold. This is unavoidable as there are no objective methods of determining these quantities, and these methods involve asking questions that have no objective answers. Determining the trivial effect size cut-off point involves the question of "should I consider a locus with effect size X be a detectable QTL?" which is an ill-defined question as it depends on how much should the detection of the QTL be prioritized, besides the numerous other factors such as genetic architecture of the trait, experiment designs, allele frequency of the marker and the QTL and linkage disequilibrium between QTL and marker. Whereas determining the linkage disequilibrium threshold involves asking the question "how strong the linkage disequilibrium a null marker should have with a QTL such that a positive on that marker be counted as a true positive?" which the only non-arbitrary answer are either exclusion of all null markers regardless how close the marker is to a QTL (i.e. linkage disequilibrium threshold of 1.0) or any markers with zero linkage disequilibrium with any of the QTL (i.e. linkage disequilibrium threshold of 0), both of which are impractical in a GWAS. Due to the lack of objectivity in these questions, the use of arbitrary trivial effect sizes cut-off point and linkage disequilibrium threshold is unavoidable.

In conclusion, an algorithm for the calculation of optimal threshold based on ROC curve that could take into account the effects of genetic architecture and experiment design has been developed. Using the algorithm as well as its various generalizations, the calculated threshold has achieved increased performance in binary classification for identification of causal variants, and increased accuracy for truncated genomic prediction, when compared to the Bonferroni and Benjamini-Hochberg FDR under varying genetic architecture and experiment designs. By showing the inequality in optimal threshold in binary classification and genomic

prediction, this experiment had also revealed the arbitrary nature of the concept of optimal threshold, especially in the context of different use of such threshold. Despite this, the full application of such threshold requires information on the distribution of the QTL effect sizes, and while previous publications for its estimation are available, they suffered from numerous shortcomings. Further studies on the robust estimation of QTL effect size distribution would thus be desirable.

Chapter 5. A Flexible, Semi-Parametric Algorithm for Estimation of Genetic Architecture Parameter

Zhi Loh, Julius H. J. van der Werf, Sam Clark

5.1. Abstract

While Genome-Wide Association Study has been used to identify the location and effect size of the QTL, it failed to detect large portion of the QTL and thus additive genetic variance. For this reason, alternative approaches that attempted to estimate the genetic architecture parameters have emerged. Previous methods that attempted such estimation exist, but they failed to take into account many of the assumptions and phenomenon that would be encountered if such approach is to be taken, such as reliance on previously published Genome-Wide Association Study results and effects of confounding factors such as allele frequency distributions and linkage disequilibrium structures. Thus, the aim of this study was to develop a method that could estimate the parameters of genetic architecture such as number of QTL with certain effect sizes, and the shape of QTL effect size distribution, while taking into account the effects from the aforementioned phenomenon. Using this method, the estimated number of QTL with effect size 0.1 σ_e ranges from 69.9% to 167.0% (an average of 109.8%) of the true number of QTL, and for effect size 1.0 σ_e the range was from 101.6% to 175.8% (an average of 123.6%). This method could also provide an estimate of marker effect sizes, but with consideration from the confounding factors such as allele frequency distributions, correlation between markers and heterogeneity in linkage disequilibrium structures. The algorithm would be important for gene discovery and estimation of location and effect size of the causal loci.

5.2. Introduction

Genome-Wide Association Study (GWAS) has successfully been used in identifying the causal loci for diseases in human (Pearson and Manolio, 2008; Tam et al., 2019) or production traits in livestock (Bedhane et al., 2019; Hay and Roberts, 2018). Despite this, GWAS is generally underpowered in detecting the large number of QTL with small effect sizes, which led to an underestimation of additive genetic variance explained by the QTL detected by GWAS compared to the genetic variance estimated from classical analysis of

variance (Hall et al., 2016). For individual QTL, the estimated effect size obtained through GWAS is generally overestimated, especially when the QTL has small effect size (Hall et al., 2016; Xu et al., 2003). The power of GWAS is further burdened by the severe multiple testing from the sheer number of markers to be tested (Pearson and Manolio, 2008). The stringent threshold required to exclude the false positives would also mean that the signals from the QTL with small effect sizes would be buried in the sea of noise from the null markers, making them effectively undetectable (Tam et al., 2019; Zhang et al., 2018). This is especially true for a trait with low heritability, where the increased residual variance contributes into excessive noises for the null markers (Tam et al., 2019).

Perhaps rather than relying on an arbitrary threshold to statistically test the association of the loci with the trait, an alternative method was estimating the distribution of the QTL effect sizes. Several previous publications have attempted this; Park et al. (2010) published an algorithm to estimate QTL effect size distribution by calculating the power of detected QTL from previously published GWAS. Cheng et al. (2020) and Zhang et al. (2018) utilized an expectation-maximization (EM) algorithm on summary statistics of GWAS to estimate the parameters for the mixture model of the proportion of null and non-null markers, while using an empirical Bayesian approach to estimate the threshold that classifies the markers as null or non-null. Hall et al. (2016) utilized the proportion of additive genetic variance explained and detection threshold as a method to estimate the number of QTL associated with a trait. In an attempt to improve the flexibility of the model used, Zeng and Zhou (2017) specify a nonparametric prior on the variance in the normal distributions, which in turn used to model the QTL effect size distribution.

Despite these attempts, there were numerous assumptions being utilized in these methods, many of which could impact their reliability of the estimation. As an example, the algorithm suggested by Cheng et al. (2020) relies on user-defined cut-off points between null and non-null markers, which might not be optimal for varying genetic architectures. Park et al. (2010) relies on previously published GWAS results, which might not be available. There are also many aspects and issues worth considering during the estimation of the QTL effect size distribution, such as the confounding effects of linkage disequilibrium structures that alters the distribution of estimated marker effect sizes and test statistics of a GWAS, and the changing allele frequency distributions which affects the error in estimated effect sizes of GWAS, and these were generally not discussed in previous work. An algorithm that could take these aspects into account is currently lacking.

With this in mind, the aim of this study is to propose a method for estimation of parameters for a genetic architecture, such as number of QTL and the shape and scale parameter for the QTL effect size distribution. The method was tested using simulated dataset with varying underlying genetic architectures. It is anticipated this algorithm could provide an estimate of genetic architecture parameters, which could then be used to estimate the QTL effect sizes of a marker, a genomic region, or an animal. The strengths, assumptions and weaknesses of this algorithm would also be evaluated and discussed in this study.

5.3. Preliminary Concepts and Notations

While GWAS is usually used in identifying causal loci, their statistical properties often revealed more information than just the location and strength of the loci. One such piece of information is the distribution of the output of the GWAS, which can provide tell-tale signatures on the underlying distribution of the QTL effect sizes. Indeed, if there are no errors in the estimation of the effect sizes of each marker, and these markers can accurately reflect the QTL effect size, then the expected distribution of estimated effect size from a GWAS would correspond with the underlying QTL effect size distribution. Due to various confounding factors, such as allele frequency distribution and correlation between markers, such idealized situation would almost certainly never be achieved. The effects of the underlying QTL effect size distribution, as well as the effects from these confounding factors, are discussed in Appendix B.

In many instances in this chapter, there would be discussions on the properties of distribution, and in a loose sense, treating a sequence of distributions as if it is a sequence of random variables. Thus, one could discuss concepts such as the distributions of the distributions (i.e., how the distributions distribute along the axes) and the test statistics for their equality. For this chapter, the level of distribution was denoted using the notation $\mathbb{d}^n$, where the superscript denotes the level of distribution. For example, given a sequence of random variables, the distribution of the random variables was denoted as $\mathbb{d}^1$, while the distribution from a sequence of $\mathbb{d}^1$ distributions be denoted as $\mathbb{d}^2$, and so forth. An illustration of levels of distribution is provided in Figure 5.1.

There are parallelisms between the concepts from a sequence of distributions with those from a sequence of random variables; just like a large sequence of random variables allows the determination of average of their distribution, one can also discuss the concept of "expected distribution" (denoted as $E(\mathbb{d})$). The "dispersion of distribution" (denoted as $V(\mathbb{d})$) can be

thought of as the distribution version of "variance" in random variables and is defined as the variability in the form or shape of the distribution around the expected distribution. As in test statistics in $\mathbb{d}^1$ can be used in testing the significance of differences of an observed random variable compared to expected $\mathbb{d}^1$ (as in test statistics for t-test for hypothesis testing), test statistics can also be applied to $\mathbb{d}^2$ which can then be used to test the significance of differences of an observed distribution $\mathbb{d}^1$ compared to expected $\mathbb{d}^2$.

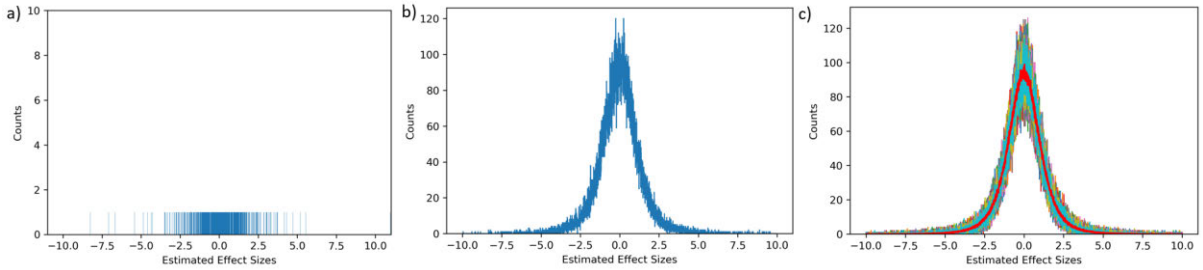

Figure 5.1: The different levels of distributions, with (a) showing the individual random variables. The density of the random variable can then be used to build a histogram that shows the density of the random variable as featured in Figure (b), which in this paper be denoted as $\mathbb{d}^1$. With a sequence of $\mathbb{d}^1$s, they can be used to build a distribution of $\mathbb{d}^1$s as featured in Figure (c), which be denoted as $\mathbb{d}^2$ in this study. The red line in Figure (c) is the expected distribution across the sequence of $\mathbb{d}^1$s, and the dispersion of distributions across the sequence of $\mathbb{d}^1$s be manifested as the multi-coloured edge around the red line.

There are several emergent properties in distributions that do not present in singular random variables. One such examples is the shape of $\mathbb{d}^1$, which provides insights on the moments of the random variables that build the $\mathbb{d}^1$. While it might not be meaningful in discussing the mean and variance of each individual value in the random variables that make up a $\mathbb{d}^1$, it is meaningful to discuss these quantities for each $\mathbb{d}^1$ that make up a $\mathbb{d}^2$. Unlike random variables, one can also perform calculus operations in each element in $\mathbb{d}^2$; as an example, one can integrate the probability density function of $\mathbb{d}^1$ to yield its corresponding cumulative distribution function (CDF), which was denoted as $\mathbb{D}^1$. Finally, while there is only one operation for testing the equality of two random variables (i.e., subtracting the value of two random variables), there are multiple operations that could be done in testing the equality between two distributions, such as maximal distance between two distribution or area between curves between them (Dowd, 2020). Examples of these tests are provided in Appendix C.

Perhaps the most familiar analogy for these levels of distributions can be found in the context of two-sample Kolmogorov-Smirnov test, which aimed to test the equality of distributions

from two sets of random variables (i.e., two $\mathbb{d}^1$s). The $\mathbb{d}^1$s can be integrated into their empirical cumulative distribution function (ECDF), both of which will be denoted as $\mathbb{D}^1$. The test statistic for the Kolmogorov-Smirnov test (t_{KS}) is defined as the supremum distance between the ECDF of the test distribution ($\mathbb{D}_A^1$) and the ECDF of theoretical distribution ($\mathbb{D}_{H_0}^1$) (Naaman, 2021; Simard and L'Ecuyer, 2011):

$$t_{KS} = \sup\left|\mathbb{D}_A^1 - \mathbb{D}_{H_0}^1\right| \quad [1]$$

Hypothetically if we can sample the ECDF of the theoretical distribution $\mathbb{D}_{H_0}^1$ m number of times, we can get a sequence of length m containing the theoretical distributions, which was denoted as $\mathbb{D}_{H_0}^2$. Similarly, $\mathbb{D}_A^1$ can also be sampled n number of times, with the resulting sequence denoted as $\mathbb{D}_A^2$. The Kolmogorov-Smirnov test can then be applied to each of the $\mathbb{D}_{H_0}^1$ in $\mathbb{D}_{H_0}^2$ with each of the $\mathbb{D}_A^1$ in $\mathbb{D}_A^2$, which produces an $m \times n$ array containing the t_{KS} from each combination of $\mathbb{D}_{H_0}^1$ and $\mathbb{D}_A^1$. This 2-dimensional array was denoted under the notation $t_{\mathbb{D}^2}$ in this chapter. Additional subscripts might be appended to indicate the test from which $t_{\mathbb{D}^2}$ originated from, such as in this example $t_{\mathbb{D}_{KS}^2}$ where the subscript KS denotes this is a $t_{\mathbb{D}^2}$ from a Kolmogorov-Smirnov test between two $\mathbb{D}^2$s.

This chapter will utilize and perform mathematical operations on multidimensional arrays (i.e., arrays with more than two dimensions). Thus, the following notations will be used: scalar values will be denoted using regular scripts, which could be in capital or small letters (as an example, x); 1-dimensional vectors were denoted as bold small letter (as example, $\boldsymbol{x}$) or with one bolded subscript (as example, $\boldsymbol{x_a}$); for a two dimensional matrices they were denoted with a bold capital letter (as example, $\boldsymbol{X}$) or with one bolded subscript (as example, $\boldsymbol{X_a}$); and for arrays with three or more dimensions, bolded subscripts were appended to the bold capital letter to denote the nature of the axes. As an example, the $\boldsymbol{X_{a,b,c,d,e}}$ would be a 5-dimensional array, with the $\boldsymbol{a}$ in the subscript be denoting the first axis of the array, $\boldsymbol{b}$ be the second axis of the array, and so on.

5.4. Phenotype Model Assumed in this Method

For this study, the phenotype is assumed to follow a purely additive polygenic model. Given N number of animals and $\mathbb{k}$ number of QTL, the phenotype is defined as follows:

$$\boldsymbol{y} = \boldsymbol{Xa} + \boldsymbol{e} \quad [2]$$

Where $\boldsymbol{y}$ is a column vector of length N containing the phenotypes of the animals; $\boldsymbol{X}$ being a matrix of size $N \times \mathbb{k}$ containing the genotypic states of the QTL; $\boldsymbol{a}$ being column vector of length $\mathbb{k}$ containing the QTL effect sizes; and $\boldsymbol{e}$ being a column vector of length N containing the residual component of the phenotype. The $\boldsymbol{a}$ in this study is assumed to follow a gamma distribution, and is parametrized as follows:

$$\boldsymbol{a} \sim \Gamma(\mathbb{a}, \mathbb{b}) \quad [3]$$

Where $\mathbb{a}$ and $\mathbb{b}$ are the shape and scale parameters of the gamma distribution. The distribution of QTL effect size will be denoted as $\mathbb{d}_{QTL}$ in this study.

The residual component will be modelled using a normal distribution, and is defined as:

$$\boldsymbol{e} \sim \mathcal{N}\left(0, \frac{(1-h^2) * var(\boldsymbol{Xa})}{h^2}\right) \quad [4]$$

Where h^2 is the narrow sense heritability of the phenotype. For this chapter, the value of $\mathbb{k}$, $\mathbb{a}$ and $\mathbb{b}$ will be the target of estimation, and the true genetic architecture will be denoted as follows: $Q(\mathbb{k}, \mathbb{a}, \mathbb{b})$.

5.5. The Method

The method of estimation of genetic architecture parameters proposed in this study relies on the internal consistency of the distributions of the marker test statistics from a GWAS. This internal consistency means, given a set of genotypic data, if there are two sets of phenotypes with the same underlying genetic architectures, both sets would produce similar marker test statistics distributions when GWAS is conducted on them (more information provided in Appendix B). Therefore, if we could propose a set of genetic architecture parameters, one could use said parameters to simulate a set of QTL and phenotypes, and from which conducting a GWAS on the simulated phenotypes. If there is another set of observed phenotypes (with unknown underlying genetic architectures) that when GWAS-tested produces a set of marker test statistics with a similar distribution with that obtained from the simulated phenotypes, one could infer that the observed phenotypes have the same genetic architectures as that from the simulated phenotypes. This observation would become the basis for the genetic architecture estimation method proposed in this study.

This method uses two sequences of distributions: a sequence of ECDF test statistics derived from multiple GWAS on observed phenotypes (henceforth denoted as $\mathbb{D}^2_{FT_{obs}}$), and a

sequence of ECDF of test statistics derived from multiple GWAS with simulated phenotypes (denoted as $\mathbb{D}^2_{FT_{sim}}$). The simulated phenotypes will be generated using a set of proposed genetic architecture models (denoted with a square bracket $[\mathbb{k}, \mathbb{a}, \mathbb{b}]$). The aim is to minimize the differences between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$, with the $[\mathbb{k}, \mathbb{a}, \mathbb{b}]$ that minimizes said differences will be denoted as $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$. The objective function associated with the minimization will be the $t_{\mathbb{D}^2}$ test statistics between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$, which is described in Appendix C.

5.5.1. The Layout of the Method

The algorithm requires three inputs to estimate the $Q(\mathbb{k}, \mathbb{a}, \mathbb{b})$: the $N \times M$ genotype array of size (denoted as $\boldsymbol{X_{full}}$), where N is the sample size and M is the number of SNP markers, encoded in {0,1,2} or (-1,0,1} form, a $N \times 1$ phenotype array (denoted as $\boldsymbol{y_{full}}$) and the additive genetic variance or narrow sense heritability of the trait. An overview schematic for this method was provided in Figure 5.2.

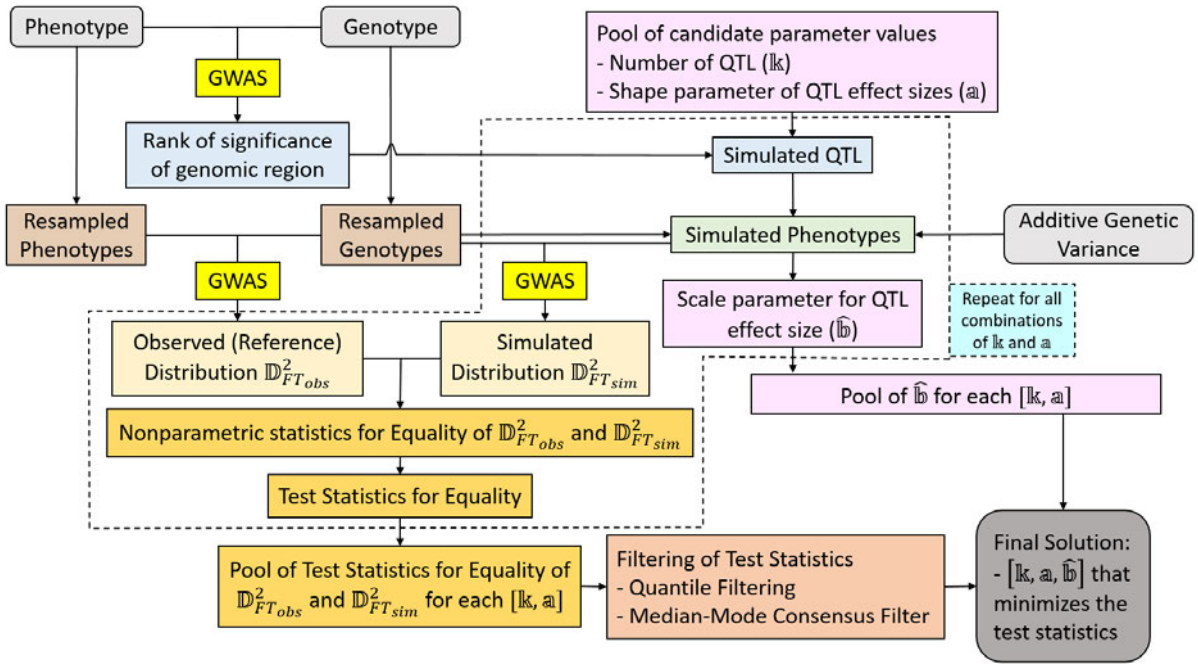

Figure 5.2: An overview schematic for the algorithm. The inputs for this algorithm are presented as light grey rounded rectangles, and the output as the dark grey rounded rectangle at bottom right corner.

5.5.1.1. Estimation the Rank of Significance of Association of a Genomic Region with Phenotype

The first step in this algorithm is to estimate the rankings for the significance of association between a genomic region and the phenotype (i.e., how strong a region is associated with the

phenotype). The ranking would be used to assign the location of the QTL generated from the proposed model [𝕜, 𝕒, 𝕓] (further details in section 5.5.1.4). This QTL location assignment step is done to improve the reliability of the algorithm toward a genotype array with heterogeneous linkage disequilibrium structures (as an example, in an inbred population (Gondro, 2015)).

The estimation of rank of significance of association of genomic regions starts obtaining marker-wise significance of association, which is done by conducting a single SNP linear regression using the full genotype array $\boldsymbol{X_{full}}$ and full phenotype vector $\boldsymbol{y_{full}}$. As extreme allele frequencies reduce the reliability of estimated effect sizes, markers with minor allele frequency (MAF) less than 0.05 will be excluded from the GWAS, with the number of retained markers denoted as M_{maf}. The estimated effect sizes es_i and test statistics ft_i of locus i defined as follows (Gondro, 2015; Kremelberg, 2011):

$$es_i = \frac{cov\left(\boldsymbol{x_{full_{*,i}}},\ \boldsymbol{y_{full}}\right)}{var\left(\boldsymbol{x_{full_{*,i}}}\right)} \quad [5]$$

$$ft_i = \frac{{es_i}^2 * var\left(\boldsymbol{x_{full_{*,i}}}\right) * (N-2)}{var\left(\boldsymbol{y_{full}}\right) - {es_i}^2 * var\left(\boldsymbol{x_{full_{*,i}}}\right)} \quad [6]$$

Where $var\left(\boldsymbol{x_{full_{*,i}}}\right)$ is the genotypic variance for locus i from the full genotype array, $var\left(\boldsymbol{y_{full}}\right)$ is the phenotypic variance from the full phenotype vector, and $cov\left(\boldsymbol{x_{full_{*,i}}}, \boldsymbol{y_{full}}\right)$ is the genotypic-phenotypic covariance from both full genotype array and phenotype vector.

From this operation, two vectors of length M_{maf} containing the estimated effect sizes and test statistics from all filtered markers across full sample size, denoted as $\boldsymbol{es_{full}}$ and $\boldsymbol{ft_{full}}$, will be generated. The vectors that will be used rank the significance of association of a region of a genome with the phenotype.

To take into account the bleeding effects from the correlation on the ranking of significance of genomic regions, the vectors $\boldsymbol{es_{full}}$ and $\boldsymbol{ft_{full}}$ were deconvolved using an iterative method. Although developed independently, this deconvolution method is similar to Högbom's CLEAN algorithm (Högbom, 1974), but with modifications to take into account the effects of various phenomenon commonly encountered in a GWAS experiment, such as

the effects of extreme allele frequencies on increasing the error of the estimated marker effect sizes.

The deconvolution starts by calculating the pairwise marker correlation between each of the marker pairs, with the resulting array of size $M_{maf} \times M_{maf}$ denoted as $\boldsymbol{R}$, which each of the element calculated as follows:

$$r_{i,j} = \frac{cov\left(\boldsymbol{x_{full}}_{*,i}, \boldsymbol{x_{full}}_{*,j}\right)}{\sqrt{var\left(\boldsymbol{x_{full}}_{*,i}\right) * var\left(\boldsymbol{x_{full}}_{*,j}\right)}} \quad [7]$$

The top marker within $\boldsymbol{ft_{full}}$ were the identified, from which the locus of the peak, i, was identified. The estimated marker effect size at locus i, $\tilde{a}_i$, were then kept in a new vector denoted as $\boldsymbol{es_{deconvolved}}$.

Using the matrix $\boldsymbol{R}$, the contribution of $\tilde{a}_i$ onto the estimated effect sizes for all markers, denoted as $\tilde{\boldsymbol{a}}_{\boldsymbol{r_i}}$ were calculated as follows:

$$\tilde{\boldsymbol{a}}_{\boldsymbol{r_i}} = \tilde{a}_i * \boldsymbol{R}_{i,*} \quad [8]$$

Where $\boldsymbol{R}_{i,*}$ denotes row i of matrix $\boldsymbol{R}$. The $\boldsymbol{es_{full}}$ would then be adjusted with $\tilde{\boldsymbol{a}}_{\boldsymbol{r_i}}$ as follows:

$$\boldsymbol{es_{full}} = \boldsymbol{es_{full}} - \tilde{\boldsymbol{a}}_{\boldsymbol{r_i}} \quad [9]$$

while the locus i in $\boldsymbol{es_{full}}$ be "muted" by assigning it as "NAN" and would no longer involved in further calculations. This is to prevent overcorrection of estimated effect size for the locus, which could lead to numerical instability of solution in the deconvolved estimated effect sizes.

Using the adjusted $\boldsymbol{es_{full}}$, the corresponding adjusted $\boldsymbol{ft_{full}}$ was calculated as follows:

$$\boldsymbol{ft_{full}} = \frac{\left(\boldsymbol{es_{full}}\right)^2 * var_N\left(\boldsymbol{X_{full}}\right) * (N-2)}{var\left(\boldsymbol{y_{full}}\right) - \left(\boldsymbol{es_{full}}\right)^2 * var_N\left(\boldsymbol{X_{full}}\right)} \quad [10]$$

Where $var_N\left(\boldsymbol{X_{full}}\right)$ is defined as taking the variance of $\boldsymbol{X_{full}}$ along N (i.e., column-wise) axis.

The new top marker in this $\boldsymbol{ft_{full}}$ and the corresponding $\tilde{a}_i$ was identified. The process was then iterated until all the markers in $\boldsymbol{es_{full}}$ were assigned as "NAN," and all the deconvolved estimated effect sizes were allocated to $\boldsymbol{es_{deconvolved}}$. The deconvolved test statistics for the

markers, denoted as $\boldsymbol{ft_{deconvolved}}$, were calculated using equation [8], with $\boldsymbol{es_{deconvolved}}$ being used in place of $\boldsymbol{es_{full}}$. Loci that have their MAF less than 0.05 were assigned with "NAN" in $\boldsymbol{es_{deconvolved}}$ and $\boldsymbol{ft_{deconvolved}}$.

An illustrative pseudocode for the deconvolution process was provided as follows:

```
## Input data
X_full # full genotypes
y_full # full phenotypes
M = ncol(X_full)

X_full = MAF_filter(X_full, 0.05) ## remove markers with MAF < 0.05

N, Mmaf = nrow(X_full), ncol(X_full)
var_X = var(X_full, axis=1)                # column-wise variance (i.e. genotypic variance)
var_y = var(y_full)                        # phenotypic variance

## Single SNP Linear Regression (SSR) GWAS to obtain marker-wise significance
es_full, ft_full = SSR_GWAS(X_full, y_full) # estimated effect sizes, F-test statistics

#### deconvolute (de-correlate) the marker effect sizes
## Obtaining pairwise marker correlation matrix (R)

R = matrix(shape=(Mmaf, Mmaf))
for m1 in range(Mmaf):
        for m2 in range(Mmaf):
                r_ij = cov(X_full[:,m1], X_full[:,m2]) / sqrt(var(X_full[:,m1])*var(X_full[:,m2]))# eqn[7]
                R[m1,m2] = r_ij

## de-correlate the markers using matrix R
es_deconvolved = vector(length=Mmaf)
for i in range(Mmaf):
        i = which(ft_full == max(ft_full))   # which loci (i) has the highest F-stats
        a_i_tilde = es_full[i]
        es_deconvolved[i] = a_i_tilde
        es_full[i] = NAN # mute the locus to prevent instability in decorrelated solution

        R_istar = R[i,:]
        a_ri_tilde = R_istar * a_i_tilde     # eqn [8]

        es_full = es_full - a_ri_tilde # eqn [9]
        ft_full = es_full^2 * var_X * (N-2) / (var_y - es_full^2 * var_X) # eqn [10]

es_deconvolved = pad_by_MAF(es_deconvolved, M, NAN) # pad es_deconvolved with NAN for locus with MAF < 0.05
## at this point the length of es_deconvolved is M (not Mmaf)

ft_deconvolved = es_deconvolved^2 * var_X * (N-2) / (var_y - es_deconvolved^2 * var_X)
```

The deconvolved SNP markers were then pruned, which was done by slicing $\boldsymbol{ft_{deconvolved}}$ were then sliced into segments of contiguous SNPs of equal length. Any arbitrary number of SNPs per segments could be chosen for this method, although the choice would affect the performance of the ranking process; a segment with a large number of SNPs reduces the precision of location that would be assigned as QTL from the proposed models, while a segment with a small number of SNPs increases the method's vulnerability toward heterogeneous linkage disequilibrium structures. For this study, 10 SNPs per segments was chosen for this study. Within each slice, the top test statistics within the slice, excluding the "NAN," were recorded into a vector with size $1 \times \lceil M/10 \rceil$, denoted as $\boldsymbol{ft_{sliced}}$.

Finally, the rank of the significance of association of a genomic region, defined as $\boldsymbol{rank}$, will be defined as a vector of indices that would sort $\boldsymbol{ft_{sliced}}$, with 1 being the region of least significance with the phenotype, and maximum value being region of most significance. Slices with all "NAN" were assigned with an index of 0 in $\boldsymbol{rank}$.

A simplified example of the ranking process was illustrated in Figure 5.3.

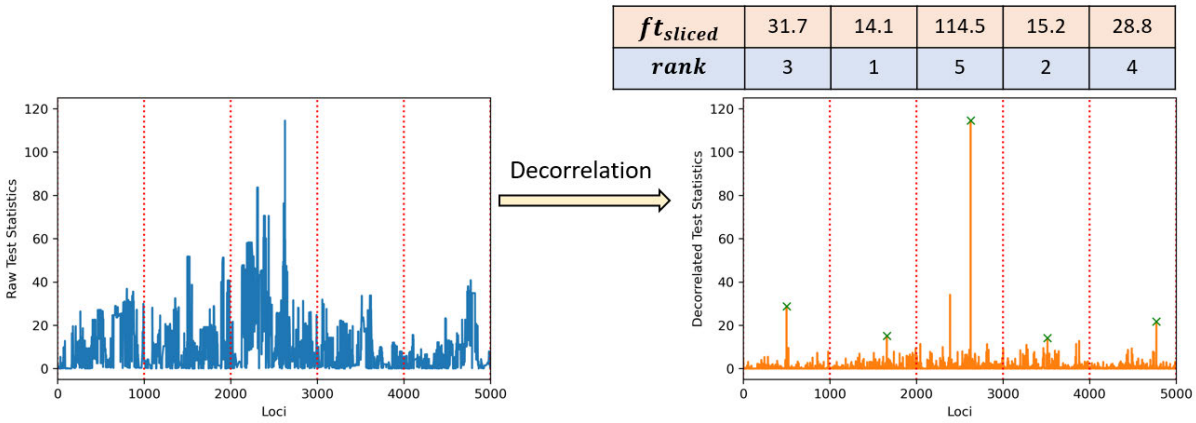

Figure 5.3: A simplified example of the ranking of significance of genomic region. The raw test statistics in the left panel was first deconvolved and have the effects of correlation and LD structures removed. The resulting test statistics (in the right panel) were subsequently sliced (red dotted lines demarcate the slicing points) and having the test statistics of the top loci (green crosses) recorded in $\boldsymbol{ft_{sliced}}$. Each element in $\boldsymbol{ft_{sliced}}$ were then assigned a rank in term of their significance, which were kept in $\boldsymbol{rank}$. In this example, 1000 SNPs were chosen per slice for clarity; 10 SNPs per slice were used in the actual algorithm.

5.5.1.2. Obtaining $\mathbb{D}_{FT}^{2}$ from Observed Phenotypes ($\mathbb{D}_{FT_{obs}}^{2}$)

The aim for this step is to obtain a sequence of ECDFs of test statistics from the GWAS between input genotype and phenotype ($\mathbb{D}_{FT_{obs}}^{2}$). These distributions serve as reference distributions which the algorithm would attempt to fit, with the fitting model parameterized in term of $[\mathbb{k}, \mathbb{a}, \mathbb{b}]$. An illustrative figure for this step was provided in Figure 5.4.

This step starts by calculating the distribution of GWAS test statistics $\mathbb{D}_{FT_{obs}}^{1}$ using the association between genotype array and the observed phenotypes, which was conducted through single SNP regression. The rationale behind single SNP regression is its simplicity, speed of calculation and the capability of being parallelized. As in the previous step, markers with MAF less than 0.05 were excluded from this GWAS.

Changing the genetic architecture would only produce minute changes in the $\mathbb{D}_{FT}^{1}$, with such changes strongly concentrated at the tail region of the distribution (details provided in

Appendix B). Despite this, it is also noted that $\mathbb{D}^1_{FT_{obs}}$ is a noisy distribution; even with the same underlying parameters (i.e., genetic architecture, allele frequency distribution or correlation structures), the $\mathbb{D}^1_{FT_{obs}}$ can vary substantially between each replication. This can cause problems in the detection of signals from the changing genetic architecture, where the small amount of data available can be easily overwhelmed by noise. Thus, rather than relying on one $\mathbb{D}^1_{FT_{obs}}$, a sequence of $\mathbb{D}^1_{FT_{obs}}$ will be generated. This is achieved by resampling a number of individuals (denoted as N_{rsamp}) from $\boldsymbol{X_{full}}$ and $\boldsymbol{y_{full}}$ without replacement, with number of resamples denoted as n_{obs}. The resampled genotype array is denoted as $\boldsymbol{X_{o,r,m}}$, where subscript o, r and m denote the index for resamples, index of resampled individuals and index of SNP markers respectively, and $\boldsymbol{X}$ would have the size $n_{obs} \times N_{rsamp} \times M$, and the resampled phenotype is denoted as $\boldsymbol{Y_r}$, and would have size $n_{obs} \times N_{rsamp}$.

The marker test statistics from each of the resamples of genotype and phenotypes were calculated using equations [5] and [6], but with the $\boldsymbol{X_{o,*,m}}$ and $\boldsymbol{y_o}$ being used in place of $\boldsymbol{X_{full_i}}$ and $\boldsymbol{y_{full}}$, where $\boldsymbol{X_{o,*,m}}$ is a vector of length N_{rsamp} containing the resampled genotypes and $\boldsymbol{y_o}$ is a vector of length N_{rsamp} extracted from $\boldsymbol{Y_r}$ that containing the resampled phenotype. The resulting test statistics were recorded in a 2-dimensional array, denoted as $\boldsymbol{FT_{obs}}$, of size $n_{obs} \times M_{maf}$.

The $\boldsymbol{FT_{obs}}$ array is used to build a scaled complement of the empirical distribution function ($MECDF$). Given a vector $\boldsymbol{x}$ the $MECDF$ is defined as follows (Singer and Andrade, 2010):

$$MECDF(x) = \sum_{i=1}^{M_{maf}} 1(x_i \geq x) \qquad [11]$$

Where $1(x)$ is the indicator function defined as follows:

$$1(x) = \begin{cases} 1; if\ x_i\ \in \boldsymbol{x} \\ 0; if\ x_i\ \notin \boldsymbol{x} \end{cases} \qquad [12]$$

Where $\boldsymbol{x}$ is a vector of random variables which may or may not have x_i as its element. In essence the $MECDF$ at x is the number of elements in vector $\boldsymbol{x}$ that have their values equal to or larger than x.

Using the equation [11], the $MECDF$ was calculated along n_{obs} margin of $\boldsymbol{FT_{obs}}$, and this resulted in a sequence of $\mathbb{D}^1_{FT_{obs}}$ with length n_{obs}, which collectively become $\mathbb{D}^2_{FT_{obs}}$. The $\boldsymbol{FT_{obs}}$ and $\mathbb{D}^2_{FT_{obs}}$, as well as $\boldsymbol{X_{o,r,m}}$ were kept for the subsequent steps.

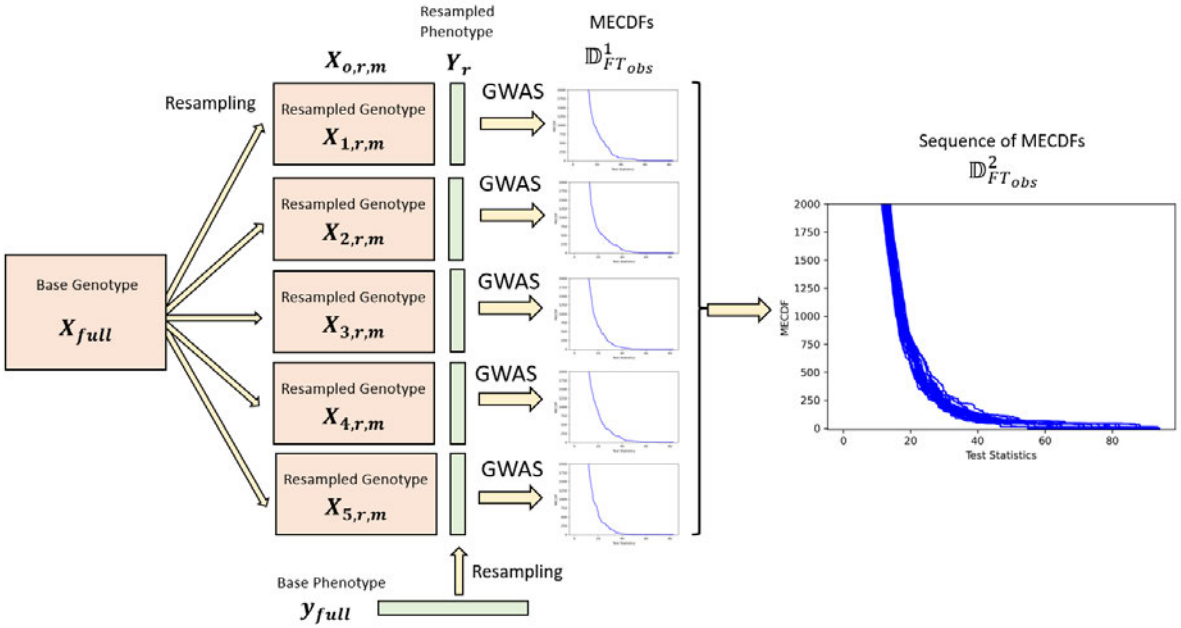

Figure 5.4: Illustration on the generation of "observed distributions" $\mathbb{D}^2_{FT_{obs}}$. The base (input) genotypes and phenotypes were first resampled, with N_{rsamp} individuals chosen per resamples. GWASes were then conducted for each pair of genotype-phenotype resamples, from which the scaled empirical cumulative distribution function (MECDF, which represents the $\mathbb{D}^1_{FT_{obs}}$ for the algorithm) were generated. The $\mathbb{D}^1_{FT_{obs}}$s were then collated into the sequence of observed distributions $\mathbb{D}^2_{FT_{obs}}$.

5.5.1.3. Sampling for Combinations of 𝕜s and 𝕒s Tested

For the calculation of ECDF for the proposed model, a number of QTL 𝕜 and shape parameter for QTL effect size distribution 𝕒s were sampled. In total n_{kix} number of 𝕜s and n_{aix} number of 𝕒s were sampled. From these sampled values, a grid of size $n_{kix} \times n_{aix}$ that contains all possible combinations of $[\mathbb{k}, \mathbb{a}]$ were generated, which will be evaluated by the method.

As the parameter 𝕜 can range between 0 and M, this introduces a large parameter space that needs to be tested, which could impede the feasibility of the method. This however can be resolved by choosing some of the values of 𝕜 that would be tested by the method. For this study, a "geom-linear sequence" of length n_{kix} will be used. This sequence starts by initially building a geometric progression that ranges from a starting value up to M, and this is

followed by linear interpolation between each consecutive pair within the geometric sequence. A detailed example for the generation of this sequence is provided in Appendix D.

For a shape parameter $\mathbb{a}$, a linear sequence of length n_{aix} with values in the range $0 < \mathbb{a} \leq 1$ will be utilized. This is based on the observation that these are the $\mathbb{a}$s that produce the correct shape for the gamma distribution. A more mathematically in-depth explanation is provided in Appendix D.

5.5.1.4. Generation of Simulated QTL Effect Sizes Random Variates

For each parameter combinations $[\mathbb{k}, \mathbb{a}]$, a vector of random variates of length $\mathbb{k}$ is generated (denoted as $\boldsymbol{q_{sim}}$), with the random variates following a gamma distribution $\Gamma(\mathbb{a}, 1)$. These random variates represent the QTL effect sizes from the proposed parameter combinations. This vector will be padded with $M - \mathbb{k}$ zeros, which represents the effect sizes for null markers. This results in a $\boldsymbol{q_{sim}}$ vector of length M that contains the effect sizes for all the markers.

To handle the effects of heterogeneous linkage disequilibrium structures, the $\boldsymbol{q_{sim}}$ was rearranged using the vector $\boldsymbol{rank}$. The vector is first sliced into segments of equal length, with the same number of SNP per slice as in the calculation of vector $\boldsymbol{rank}$ (i.e., 10 SNPs per segment for this study). For each slice of $\boldsymbol{q_{sim}}$, the effect sizes were summed and, using these sums, these slices were rearranged based on the indices from the vector $\boldsymbol{rank}$.

The rearrangement was conducted as follows: the slice of $\boldsymbol{q_{sim}}$ with the largest sum of effect sizes is assigned to the region that ranked at the maximum value (i.e., region with the strongest significance of association, or region with maximum $\boldsymbol{rank}$ value). The slice with second largest sum is allocated to the second most significant region, and slice with third largest sum to the third most significant, and so on. This step is to construct a new $\boldsymbol{q_{sim}}$ that best describes the Manhattan plot from the GWAS with observed phenotype (i.e., aligning the slices of $\boldsymbol{q_{sim}}$ in accordance with the peaks from the Manhattan plot while ranking the slices in accordance with the magnitude of the peaks). A simplified example of generation and rearrangement were provided in Figure 5.5.

To alleviate potential issues caused by noises in the distributions, a sequence of $\boldsymbol{q_{sim}}$ were generated and rearranged, with the number of $\boldsymbol{q_{sim}}$ generated denoted as n_{sim}. These $\boldsymbol{q_{sim}}$s are then compiled into an array of size $n_{sim} \times M$, which was denoted as $\boldsymbol{Q_{sim}}$. To ensure the

compatibility in size between the resampled genotype $\boldsymbol{X_{o,r,m}}$ and $\boldsymbol{Q_{sim}}$, the n_{sim} was set equal to n_{obs}.

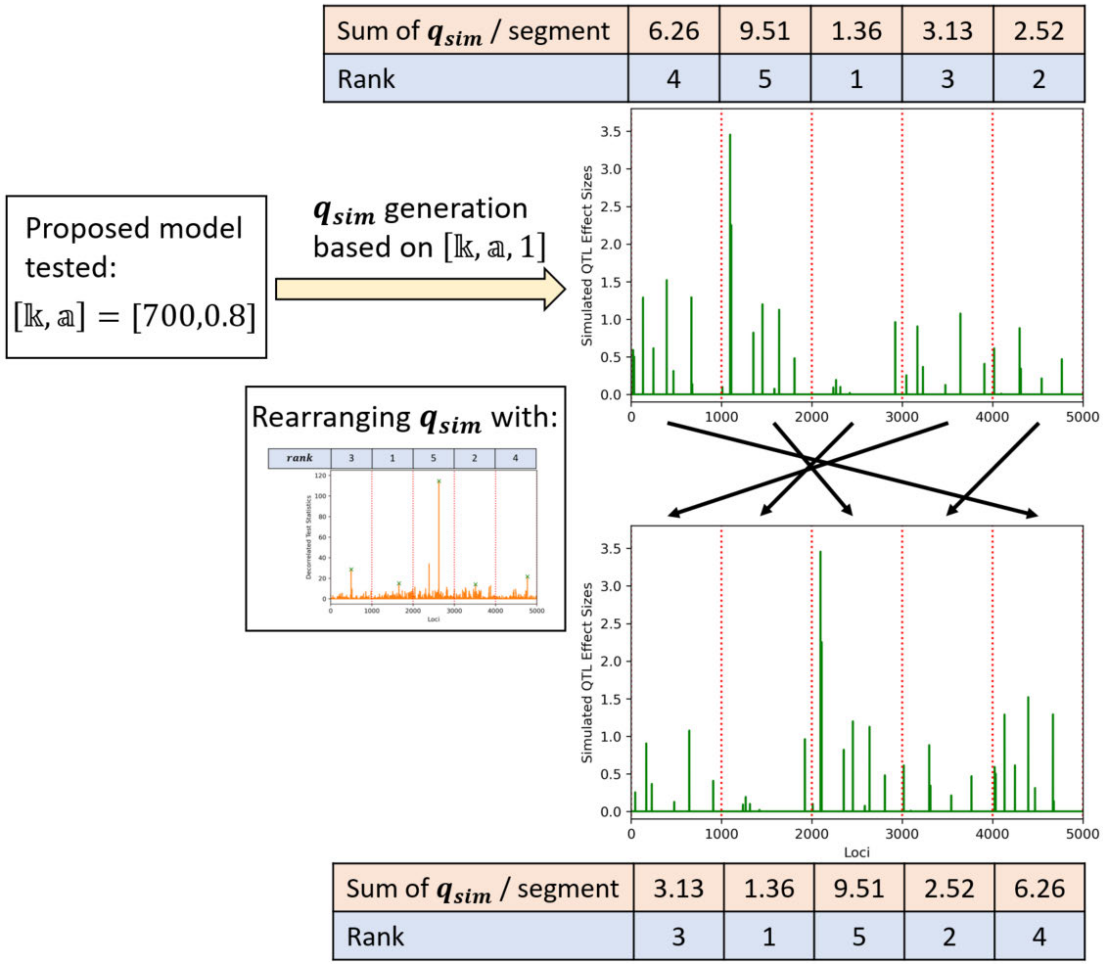

Figure 5.5: A simplified example of generation and reranking of simulated QTL effect sizes. Using a proposed values of $[\mathbb{k}, \mathbb{a}]$ (as example in this case, $\mathbb{k} = 700$ and $\mathbb{a} = 0.8$), a set of QTL effect sizes $\boldsymbol{q_{sim}}$ were simulated. At this point, the scale parameter $\mathbb{b}$ was kept at 1. Using the same slicing schemes as in Figure 5.3, the $\boldsymbol{q_{sim}}$ was sliced and the effect sizes were summed, and the total effect sizes per slices were ranked. Using the $\boldsymbol{rank}$ calculated from Figure 5.3, the slices of $\boldsymbol{q_{sim}}$ were rearranged such that resulting ranks correspond to that of $\boldsymbol{rank}$. This process was then repeated n_{sim} number of times, from which n_{sim} number of rearranged $\boldsymbol{q_{sim}}$ were produced before being compiled into $\boldsymbol{Q_{sim}}$. In this example 1000 SNPs were chosen per slice for clarity, 10 SNPs per slices was used in the actual algorithm.

5.5.1.5. The Calculation of Simulated Phenotypes

The simulated phenotype $\boldsymbol{Y_{sim}}$ was calculated using the resampled genotype $\boldsymbol{X_{o,r,m}}$, the effect sizes of the simulated QTL $\boldsymbol{Q_{sim}}$, as well as the full observed phenotype $\boldsymbol{y_{full}}$ and additive

genetic variance $v_{A_{obs}}$. The simulated phenotype $\boldsymbol{Y_{sim}}$ is a $n_{sim} \times N_{rsamp}$ 2-dimensional array. A flowchart for this step was provided in Figure 5.6.

The additive genetic component of $\boldsymbol{Y_{sim}}$, denoted as $\boldsymbol{Y_{simA}}_{[\mathbb{b}=1]}$, is calculated through the following multidimensional array multiplication:

$$\boldsymbol{Y_{simA}}_{[\mathbb{b}=1]_{s,r}} = \sum_{i=1}^{M} \boldsymbol{X}_{s,r,i} * \boldsymbol{Q_{sim}}_{s,i} \quad [13]$$

Where the subscript s and r are the index of n_{sim} and N_{rsamp} respectively. The additive genetic variance of the simulated phenotype (denoted as $\boldsymbol{v_{A}}_{[\mathbb{k},\mathbb{a},\mathbb{b}=1]}$) was calculated as follows:

$$\boldsymbol{v_{A}}_{[\mathbb{k},\mathbb{a},\mathbb{b}=1]_s} = var\left(\boldsymbol{y_{simA}}_{[\mathbb{b}=1]_{s,*}}\right) \quad [14]$$

The resulting $\boldsymbol{v_{A}}_{[\mathbb{k},\mathbb{a},\mathbb{b}=1]}$ is a vector of length n_{sim}.

It's worth noting that up to this point the simulated QTL effect sizes in $\boldsymbol{Q_{sim}}$ still have their scale parameter set at $\mathbb{b} = 1$, which might not be the case for the $\mathbb{d}_{QTL}$ that need to be estimated. It has been noted however that the sole effect of scale parameter $\mathbb{b}$ is that it scales the random variates of the $\mathbb{d}_{QTL}$ by a fixed amount (Mun, 2012):

$$\Gamma(\mathbb{a}, \mathbb{b}) = \mathbb{b} * \Gamma(\mathbb{a}, 1) \quad [15]$$

Using this scaling property, one can calculate the expected additive genetic variance of the phenotype if the $\mathbb{d}_{QTL}$ follows a $\mathbb{b}$ other than 1 by simply multiplying it with $\mathbb{b}^2$. A vector of length n_{sim} containing estimated $\mathbb{b}$s (denoted as $\hat{\mathbf{b}}$) could thus be calculated, with the sth entry of $\hat{\mathbf{b}}$ (denoted as $\widehat{\mathbb{b}_s}$) defined as follows:

$$\widehat{\mathbb{b}_s} = \sqrt{\frac{v_{A_{obs}}}{v_{A_{[\mathbb{k},\mathbb{a},\mathbb{b}=1]_S}}}} \quad [16]$$

Where $v_{A_{obs}}$ is the observed additive genetic variance. This operation has the implication of reducing the number of parameters that need to be estimated, hence simplifying the method. From this point onward, only $\mathbb{k}$ and $\mathbb{a}$ remained that need to be estimated.

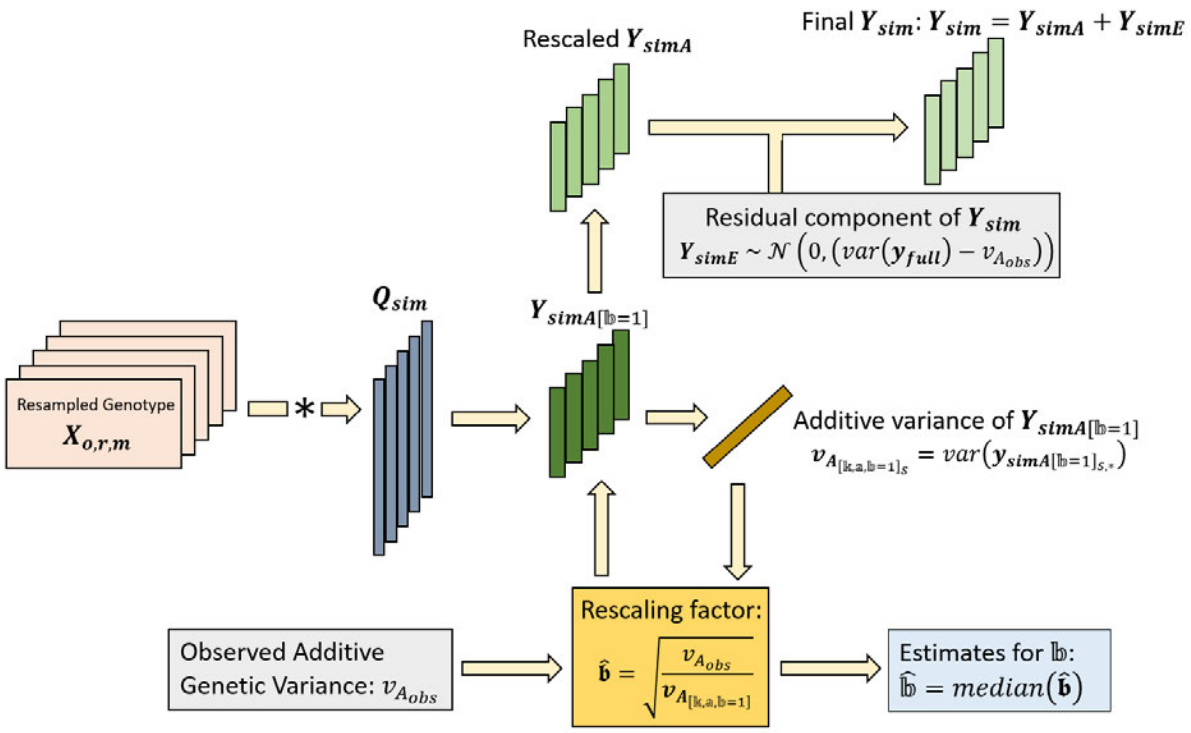

Figure 5.6: Flowchart for the calculation of simulated phenotypes $\boldsymbol{Y_{sim}}$. Starting from the resampled genotypes $\boldsymbol{X_{o,r,m}}$ and $\boldsymbol{Q_{sim}}$, the unscaled simulated additive phenotypes $\boldsymbol{Y_{simA[\mathbb{b}=1]}}$ were calculated. The scaling factor $\hat{\mathbf{b}}$ was calculated using the variance of $\boldsymbol{Y_{simA[\mathbb{b}=1]}}$ and observed additive genetic variance $v_{A_{obs}}$, which then be used to rescale $\boldsymbol{Y_{simA[\mathbb{b}=1]}}$ into $\boldsymbol{Y_{simA}}$. The median of $\hat{\mathbf{b}}$ serves as the estimate for $\mathbb{b}$. The residual component ($\boldsymbol{Y_{simE}}$, generated using normal distribution with zero mean and observed residual variance) were finally added into the $\boldsymbol{Y_{simA}}$, producing the final $\boldsymbol{Y_{sim}}$.

The vector $\hat{\mathbf{b}}$ would then be used to scale the vector of additive genetic component of the simulated phenotype, with the scaled vector denoted as $\boldsymbol{Y_{simA}}$. The sth row of $\boldsymbol{Y_{simA}}$ (denoted as $\boldsymbol{Y_{simA_s}}$) was scaled as follows:

$$\boldsymbol{Y_{simA_s}} = \boldsymbol{Y_{simA[\mathbb{b}=1]_s}} * \widehat{\mathbb{b}_s} \quad [17]$$

Where $\boldsymbol{Y_{simA[\mathbb{b}=1]_s}}$ is the s^{th} row of $\boldsymbol{Y_{simA[\mathbb{b}=1]}}$.

The residual component for the simulated phenotypes (denoted as $\boldsymbol{Y_{simE}}$ with size $n_{sim} \times N_{rsamp}$) is generated using the following normal distribution:

$$\boldsymbol{Y_{simE}} \sim \mathcal{N}\left(0, \left(var(\boldsymbol{y_{full}}) - v_{A_{obs}}\right)\right) \quad [18]$$

And finally, the $\boldsymbol{y_{sim}}$ is calculated by summing $\boldsymbol{Y_{simA}}$ and $\boldsymbol{Y_{simE}}$:

$$\boldsymbol{Y_{sim}} = \boldsymbol{Y_{simA}} + \boldsymbol{Y_{simE}} \quad [19]$$

The simulated phenotype $\boldsymbol{Y_{sim}}$ was utilized in the calculation of $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{sim}}$. For the vector of the estimated scale parameters $\hat{\mathbf{b}}$, the median of the vector (denoted as $\widehat{\mathbb{b}}$) was calculated and kept for section 5.5.1.8.

5.5.1.6. Obtaining $\mathbb{D}^2_{FT}$ from Simulated Phenotype ($\mathbb{D}^2_{FT_{sim}}$)

As in the calculation of $\mathbb{D}^1_{FT_{obs}}$ and $\mathbb{D}^2_{FT_{obs}}$ from the observed phenotype, the distributions from the proposed model, $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{sim}}$, were calculated using single SNP regression of the simulated phenotype $\boldsymbol{Y_{sim}}$ on the resampled genotype $\boldsymbol{X_{o,r,m}}$ and from which an array of test statistics $\boldsymbol{FT_{sim}}$ with size $n_{sim} \times M_{maf}$ was generated. The steps taken for the calculation of $\boldsymbol{FT_{sim}}$, $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{sim}}$ is identical to those utilized for the calculation of $\boldsymbol{FT_{obs}}$, $\mathbb{D}^1_{FT_{obs}}$ and $\mathbb{D}^2_{FT_{obs}}$ in section 5.5.1.2 to ensure the consistency and validity during the comparison of the distributions. The resulting $\mathbb{D}^2_{FT_{sim}}$ is a sequence of $\mathbb{D}^1_{FT}$ distributions with length n_{sim}, and was used in equality testing with $\mathbb{D}^1_{FT_{obs}}$. An illustrative figure for this step was provided in Figure 5.7.

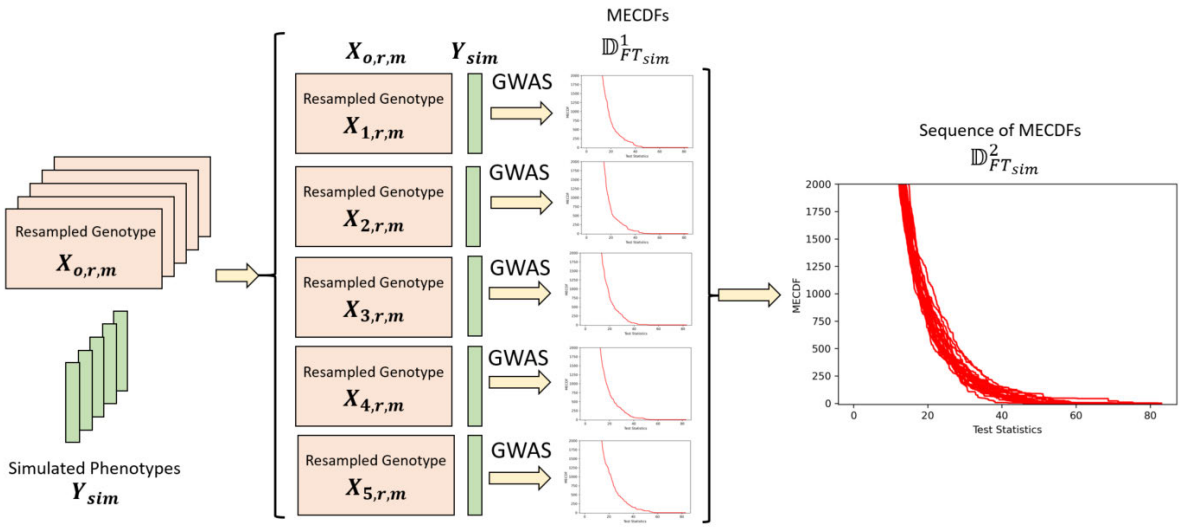

Figure 5.7: The generation of a sequence of "simulated distribution" $\mathbb{D}^2_{FT_{sim}}$. For each resamples of genotypes $\boldsymbol{X_{o,r,m}}$ and their corresponding simulated phenotypes $\boldsymbol{Y_{sim}}$, GWASes were conducted between each genotype-phenotype pairs, from which the simulated distribution $\mathbb{D}^1_{FT_{sim}}$ was generated. The $\mathbb{D}^1_{FT_{sim}}$ from all the resamples were collated into simulated distribution sequence $\mathbb{D}^2_{FT_{sim}}$.

5.5.1.7. Testing the Equality between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$

In this study, the equality between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ is defined as the goodness of fit between the each of the $\mathbb{D}^1_{FT}$ in $\mathbb{D}^2_{FT_{sim}}$ and each of $\mathbb{D}^1_{FT}$ in $\mathbb{D}^2_{FT_{obs}}$. This involves the calculation of amount of discrepancy between each of the $\mathbb{D}^1_{FT}$s in $\mathbb{D}^2_{FT_{sim}}$and those in $\mathbb{D}^2_{FT_{obs}}$. As there are n_{sim} number of $\mathbb{D}^1_{FT}$ in $\mathbb{D}^2_{FT_{sim}}$ and n_{obs} number of $\mathbb{D}^1_{FT}$ in $\mathbb{D}^2_{FT_{obs}}$, the resulting $\mathbb{D}^2$ test statistic $t_{\mathbb{D}^2}$ that describes the equality between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ is a $n_{sim} \times n_{obs}$ 2-dimensional array that contains the pairwise goodness of fit between each of the $\mathbb{D}^1_{FT}$s within $\mathbb{D}^2_{FT_{sim}}$ and that of $\mathbb{D}^2_{FT_{obs}}$. An example for the calculation of $\mathbb{D}^2$ Kolmogorov-Smirnov test statistic $t_{\mathbb{D}^2{}_{KS}}$ was provided in Figure 5.8.

Previous publications such as Cirrone et al. (2004) commented the lack of power of some of the statistical tests such as Kolmogorov-Smirnov test in detecting discrepancies at the tail region of $\mathbb{D}^1_{FT}$, where the effect of changing genetic architecture is the most observable. This is further burdened by the reduced amount of data at the tail region, which reduces the reliability of any statistical test. Another complication on testing the equality of distributions arises from the fact that $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ are mixtures of F-distribution and non-central F-distribution. By Fisher-Darmois-Pitman-Koopman theorem which, loosely stated, given a set of independent and identically distributed random variables, this set of random variable would have a sufficient summary statistics that can fully capture all the information pertaining to said set if and only if the variables follow an exponential family distribution, provided the support of the distribution do not changes with the parameters (Barankin and Maitra, 1963; Koopman, 1936). As F-distributions and non-central F-distributions are not exponential family distribution and yet with fixed supports (i.e. the mixture distributions range from zero to positive infinity, regardless the parameters), in addition with the lack of independence between the marker test statistics, all these break the assumption needed for the Fisher-Darmois-Pitman-Koopman theorem. This means for any summary statistics used on $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ there would be a loss of information on these distributions, as these summary statistics failed to capture all the information from these distribution. This in turn led to a loss of power in detecting the discrepancies between the distribution and increases the signals' vulnerability to be drown by noises in the test of equality between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$.

To improve the power and reliability of detection, and to capture as much information pertaining to these distributions as possible, a battery of 703 nonparametric statistical tests were employed to test the goodness of fit between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$. Some of the test statistics used include truncated Kolmogorov-Smirnov Test, Wasserstein's statistics, DTS statistics and quantile-based statistics as well as their generalizations. A description for the battery of statistics is provided in Appendix C.

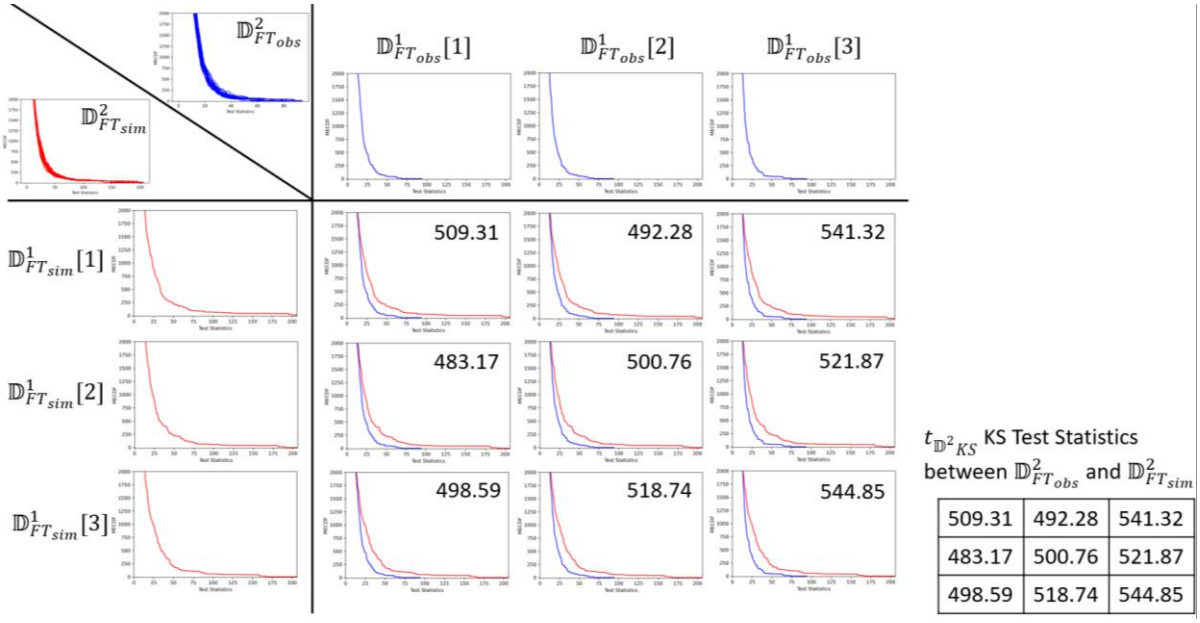

Figure 5.8: A simplified example of calculation of pairwise goodness of fit Kolmogorov-Smirnov (KS) test statistics between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ (denoted as $t_{\mathbb{D}^2 KS}$). For this simplified example, $n_{sim} = n_{obs} = 3$ distributions were tested. To conduct the KS test between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$, the KS test was conducted between each of the $\mathbb{D}^1_{FT_{obs}}$ in $\mathbb{D}^2_{FT_{obs}}$ (blue distributions column-wise) and each of the $\mathbb{D}^1_{FT_{sim}}$ in $\mathbb{D}^2_{FT_{sim}}$ (red distributions row-wise), from which their test statistics were recorded (numbers at the top right corner for each subplots). These number were then collated into a matrix of size $n_{sim} \times n_{obs}$, which become the test statistics for KS test between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$, $t_{\mathbb{D}^2 KS}$. This process was then repeated for all the statistics within the battery.

From this battery of statistics, a 3-dimensional array of size $n_{sim} \times n_{tst} \times n_{obs}$ (denoted as $\boldsymbol{T}_{[\mathbb{k},\mathbb{a}]_{s,t,o}}$) containing the test statistics from all 703 tests was compiled, with n_{tst} denoting the number of test statistics used in this experiment ($n_{tst} = 703$ in this study). This array was then used to compare the goodness of fit of the distribution from all the proposed models and the observed phenotypes. These test statistics can vary significantly in their scale, thus would need to be normalized in the filtering step in section 5.5.1.9.1.1.

5.5.1.8. Brute Force Searching the Problem

The procedures from the Step 5.5.1.4 down to Step 5.5.1.7 are repeated for each of the $[\mathbb{k}, \mathbb{a}]$ using a brute-force search method. This search method is chosen for its robustness against noisy statistics and its guarantee in finding a solution if one exists.

For each $[\mathbb{k}, \mathbb{a}]$ evaluated, a block of test statistics $\boldsymbol{T}_{[\mathbb{k},\mathbb{a}]_{s,t,o}}$ was calculated, and in total $n_{kix} \times n_{aix}$ blocks of $\boldsymbol{T}_{[\mathbb{k},\mathbb{a}]_{s,t,o}}$ were generated. These blocks of test statistics were compiled into a 5-dimensional array of size $n_{obs} \times n_{kix} \times n_{aix} \times n_{tst} \times n_{sim}$, where n_{kix} and n_{aix} are the numbers of $\mathbb{k}$s and $\mathbb{a}$s tested in the method. This array was denoted as $\boldsymbol{S}_{\boldsymbol{o},\boldsymbol{k},\boldsymbol{a},\boldsymbol{t},\boldsymbol{s}}$, with the subscript $\boldsymbol{o}$ being the index of n_{obs}, $\boldsymbol{k}$ being the index of n_{kix}, $\boldsymbol{a}$ being the index of n_{aix}, $\boldsymbol{t}$ being the index of n_{tst} and $\boldsymbol{s}$ being the index of n_{sim}. The $\boldsymbol{S}_{\boldsymbol{o},\boldsymbol{k},\boldsymbol{a},\boldsymbol{t},\boldsymbol{s}}$ array was then used to determine the goodness of fit of the distributions generated from a $[\mathbb{k}, \mathbb{a}]$, with the aim of finding a $[\mathbb{k}, \mathbb{a}]$ that best fit the observed distribution $\mathbb{D}^2_{FT_{obs}}$.

To feature the patterns observable through the $n_{kix} \times n_{aix}$ slices of $\boldsymbol{S}_{\boldsymbol{o},\boldsymbol{k},\boldsymbol{a},\boldsymbol{t},\boldsymbol{s}}$, as well as the results from subsequent processing, a 2-dimensional raster plot termed "K-a" plot were used to visualise the goodness of fit. This type of plot has a vertical axis denoting the value of $\mathbb{a}$s (shape parameter for QTL effect size distribution) and the horizontal axis the value of $\mathbb{k}$s (number of QTL), and the colour of each of the pixels in this plot represent the goodness of fits between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ of each of the models. An example of this plot is featured in Figure 5.9.

Besides the test statistics, the estimates of the scale parameter $\widehat{\mathbb{b}}$ were compiled into a 2-dimensional array of size $n_{kix} \times n_{aix}$, which was denoted as $\widehat{\mathfrak{B}}$.

5.5.1.9. Filtering the Statistics

While brute force search is one of the most robust methods available, the results from any given statistic could still be unreliable. This is due to the weak signals from the changing genetic architectures (Figure 5.10). To further improve the reliability of the method, the test statistics need to be filtered.

5.5.1.9.1. Types of Filters

Two types of filters were employed for this method: Quantile filter and Median-mode filter.

5.5.1.9.1.1. Quantile Filter

In a similar vein as in the detection of significant markers in a GWAS experiment, the test statistics within $\boldsymbol{S_{o,k,a,t,s}}$ was filtered using a quantile threshold. The aims of this process are to amplify the signals from the $[\mathbb{k}, \mathbb{a}]$ array that minimized the test statistics by nullifying those that failed the minimization, normalizing the scale of the test statistics, and to improve the reliability of the statistics caused by the small amount of data at the tail of the distribution.

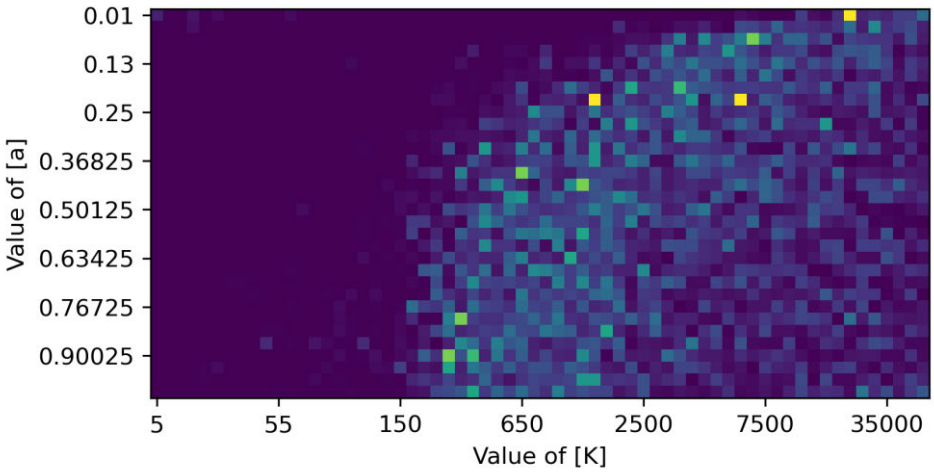

Figure 5.9: An example of the "K-a" plot. The colour of pixels featured in this plot signifies the magnitude of test statistics that test the goodness of fit between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ for each of the tested model $[\mathbb{k}, \mathbb{a}]$, with brighter pixels indicated lower test statistics (i.e., better goodness of fit). The horizontal axis denotes the values of $\mathbb{k}$ (number of QTL) and the vertical axis the value of $\mathbb{a}$ (shape parameter for QTL effect size distribution).

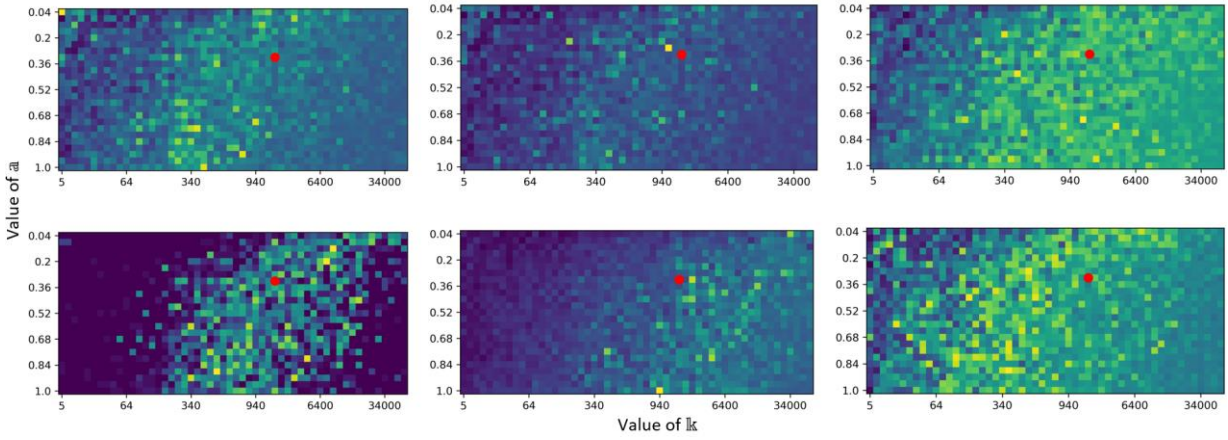

Figure 5.10: The differing performances of the test statistics. Each of the pixels in these 2-dimensional raster plots represent the goodness of fits between the distribution from observed phenotypes $\mathbb{D}^2_{FT_{obs}}$ and distributions from simulated phenotypes $\mathbb{D}^2_{FT_{sim}}$ for each of the parameter combination $[\mathbb{k}, \mathbb{a}]$, with lighter pixel indicates a higher goodness of fit (i.e., lower test statistics). The red dot on the plots indicated to the true parameter combinations.

This quantile filtering would operate along the sheets of $n_{kix} \times n_{aix}$ in $\boldsymbol{S_{o,k,a,t,s}}$, which is denoted as $\boldsymbol{S_{k,a}}$. Using a quantile threshold $q_{cut} = 0.05$, the critical values of the test statistics (denoted as s_{crit}) are obtained by defining as the bottom 5% of all test statistics recorded within the sheet $\boldsymbol{S_{k,a}}$:

$$\Pr\left(\boldsymbol{S_{k,a}} \leq s_{crit}\right) = 0.05 \qquad [20]$$

The critical value s_{crit} was used as the filter for the test statistics. From this critical value and $\boldsymbol{S_{k,a}}$, a 2-dimensional array of the size $n_{kix} \times n_{aix}$ was calculated, containing the scores of acceptance-rejection of the test statistics based on the critical value. Test statistics that have a value smaller than s_{crit} were marked with "1" and they were marked as "0" otherwise. These 2-dimensional arrays were then recompiled into another 5-dimensional array (denoted as $\boldsymbol{V_{o,k,a,t,s}}$) containing all the "votes" (i.e., "1" s) from each of the $\boldsymbol{S_{k,a}}$. An example of this filtering process is provided in Figure 5.11.

To amplify the signals, the $\boldsymbol{V_{o,k,a,t,s}}$ were then summed along the n_{sim} and n_{obs} axes, producing a 3-dimensional array of size $n_{kix} \times n_{aix} \times n_{tst}$ denoted as ($\boldsymbol{V_{k,a,t}}$):

$$\boldsymbol{V_{k,a,t}} = \sum_{s=1}^{n_{sim}} \sum_{o=1}^{n_{obs}} \boldsymbol{V_{o,k,a,t,s}} \qquad [21]$$

The $\boldsymbol{V_{k,a,t}}$ array contains a tally of votes for each of the model [𝕜, 𝕒] from each of the test statistics across the replicates. An example of the $\boldsymbol{V_{k,a,t}}$ is provided in the K-a plots in 5.12.

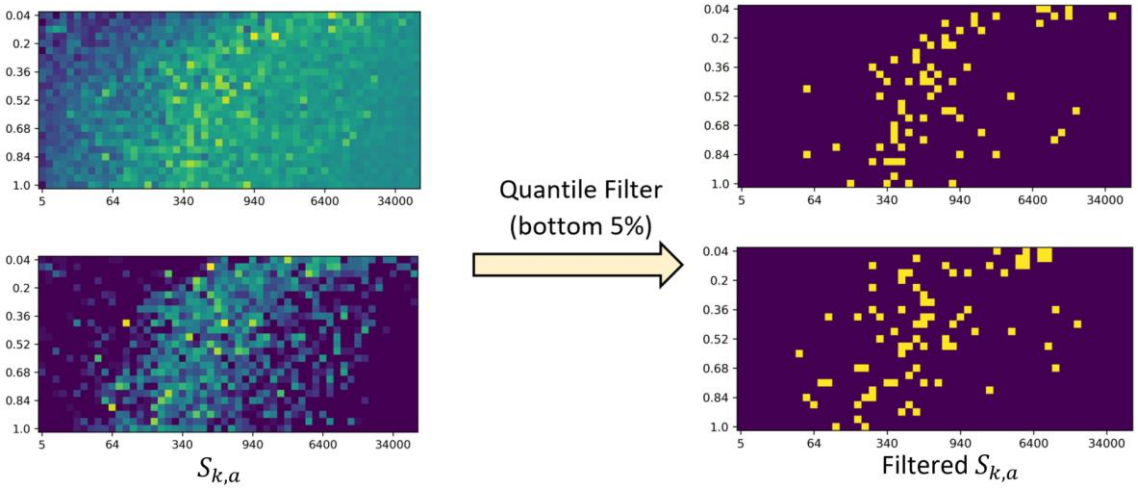

Figure 5.11: An example of quantile filtering in operation. The plots in the left contained the raw test statistics sliced from the $\boldsymbol{V_{o,k,a,t,s}}$, while the plots on the right contained the filtered array, with yellow pixels denoted as "1" (i.e., test statistics at the bottom 5%) and dark blue as "0".

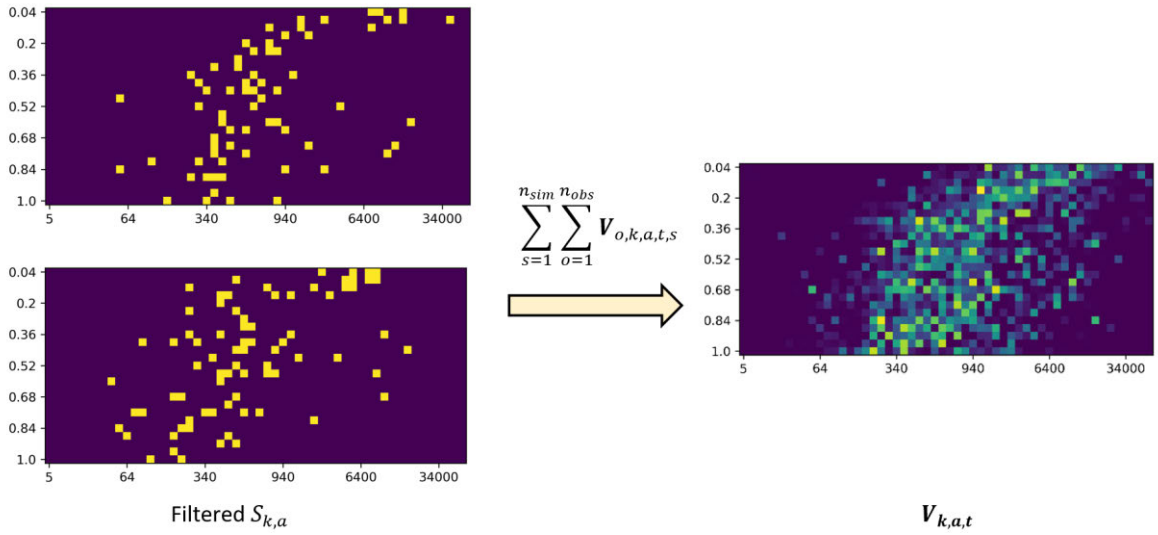

Figure 5.12: The building of the 3-dimensional array $\boldsymbol{V_{k,a,t}}$ through the summation of $\boldsymbol{V_{o,k,a,t,s}}$. Note that only a slice of $\boldsymbol{V_{k,a,t}}$ was featured in this figure.

The steps within the quantile filtering can also be expressed in the form of the following pseudocode:

```
## Input for this step : S_okats (5-dimensional array containing all the test statistics from [k, a]
## k is number of QTL, a is shape parameter for the QTL effect sizes

n_obs, n_kix, n_aix, n_tst, n_sim = S_okats.shape

## n_obs: number of observed distribution from input genotype and phenotype
#### (i.e. number of D1_FT_obs in D2-FT_obs)
## n_kix: number of levels of [k] tested
## n_aix: number of levels of [a] tested
## n_tst: number of statistical test on equality of distribution conducted
## n_sim: number of simulated distribution from resampled genotypes and simulated phenotype
#### (i.e. number of D1_FT_sim in D2_FT_sim)
#### simulated phenotypes calculated from simulated QTL generated from [k,a] and scaled by [b]

V_okats = matrix(0, S_okats.shape) # initialize a 5-d array to contain all the votes, starting values
set at 0

for ox in range(n_obs):
        for tx in range(n_tst):
                for sx in range(n_sim):
                        S_ka = S_okats[ox,:,:,tx,sx]

                        # calculating the bottom 5% quantile within S_ka
                        S_ka_quantile_05 = quantile(S_ka, 0.05)

                        ## for each of the S_ka, a vote is casted to the [k,a] that has its test
                        ## statistics below the 5% quantile (i.e. [k,a] with minimal test statistics)

                        which_ka_has_minimal_teststats = where(S_ka <= S_ka_quantile_05)
                        V_okats[ox,:,:,tx,sx][which_S_ka_has_minimal_teststats] = 1

## sum V_okats across n_obs and n_sim axis -> V_kat
V_kat = sum(sum(V_okats, axis=0), axis=4)
```

5.5.1.9.1.2. Median-mode Consensus Filter

While quantile filtering could amplify the signals, on some occasion the statistics can still be noisy, and this can be attributed to the dispersion of the distribution, which might cause some of the statistics to converge toward an outlying result. This is especially problematic for a polygenic trait, for which the signals from the changing [𝕜, 𝕒]s are sufficiently weak that noises from the dispersion can easily overwhelm the signal, producing a characteristic "lower right quadrant solutions" where an entire lower right region of the plot is marked as positive with poor differentiation between the band of solutions and the lower right quadrant (Figure 5.13). For this reason, a "median-mode consensus filter" was employed. This filtering method is designed in attempt to find a "consensus" among the test statistics.

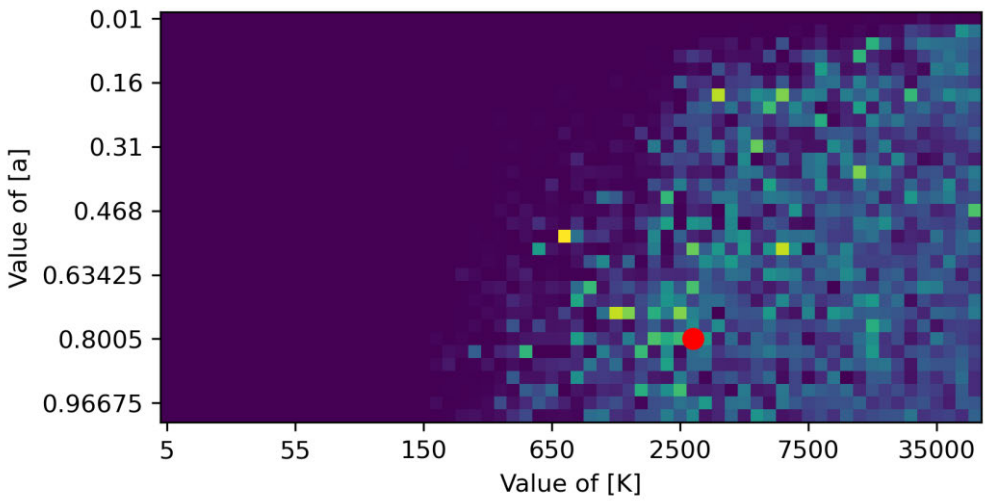

Figure 5.13: An example of the "lower right quadrant solutions" where the entire region in the lower right was marked as positive, causing poor differentiation in the band of solution and the lower right quadrant.

The median-mode consensus aimed to filter the test statistics based on their empirical distribution and attempt to achieve a consensus among the statistics by accepting those that located close to the mode of their empirical distribution. Through this filter, outlying statistics, which are the major contributor of noise in the solutions, could be weeded out.

5.5.1.9.1.2.1. The Median Phase: Calculation of Median 𝕜 Given a Test Statistics and 𝕒

This filter starts by slicing the vote tally array $\boldsymbol{V_{k,a,t}}$ array along the second (i.e. $\boldsymbol{a}$) axis, produces a 2-dimensional array of size $n_{kix} \times n_{tst}$ denoted as $\boldsymbol{V_{k,t}}$. The filter was then applied for each individual $\boldsymbol{V_{k,t}}$ slices.

For the initiation of the median phase, the $\boldsymbol{V_{k,t}}$ was further sliced along the column axis (i.e. $\boldsymbol{t}$ axis). This produces a vector of length n_{kix} denoted as $\boldsymbol{v_{k,a_x,t_x}}$ containing the tally of votes

for each of the 𝕜s for the shape parameter 𝕒 at index a_x, tested using the test statistics indexed at t_x. From this vector, the empirical CDF (denoted as F_{v,a_x,t_x}) was constructed using the following equation:

$$F_{v,a_x,t_x}(k_x) = \frac{\sum_{i=1}^{n_{kix}} (\boldsymbol{v_{k\leq i\ ,a_x,t_x}})}{\sum_{i=1}^{n_{kix}} \boldsymbol{v_{k,a_x,t_x}}} \quad [22]$$

Where the k_x is the index of 𝕜s tested in the method. The "arg-median" of this empirical CDF, defined as the argument for a distribution function that produces the median of the distribution, is obtained by finding k_x that fulfil this equation:

$$F_{v,a_x,t_x}(k_x) = 0.5 \quad [23]$$

This calculation was conducted across all n_{tst} columns of $\boldsymbol{V_{k,t}}$, from which their arg-medians k_x were obtained. These k_xs were collected in a vector of length n_{tst} denoted as $\boldsymbol{k_{med_t}}$. An example of $\boldsymbol{v_{k,a_x,t_x}}$ and F_{v,a_x,t_x} is provided in Figure 5.14.

The use of median here served two purposes: to estimate the location of the peak of the empirical distribution and, in a way, "tag" the distribution. The mean does not serve as a reliable indicator of the peak of the distribution as it is easily influenced by outliers, whereas the mode is easily influenced by the noise in the distribution. If the distributions are similar in shape and central tendency, then the median tend to be close together. Using this property, we can "tag" and classify the distribution based on its medians, thus allowing the identification of the "consensus" among the test statistics.

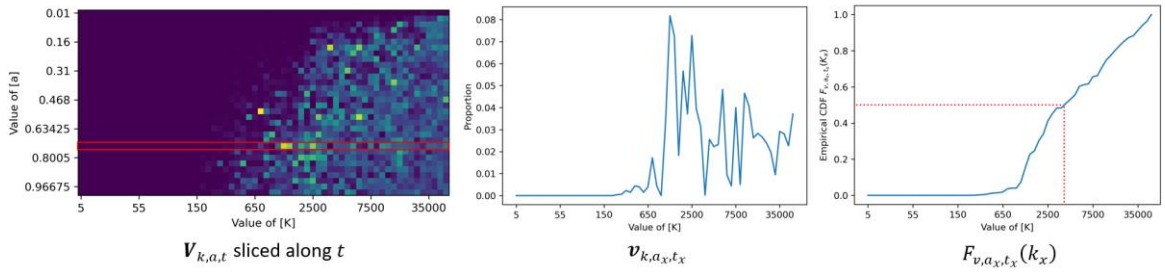

Figure 5.14: The calculation of median of the $\boldsymbol{v_{k,a_x,t_x}}$ vector. In Figure (a) the $\boldsymbol{V_{k,a,t}}$ array is first sliced along n_{aix} and n_{tst} (denoted as the red box), which the resulting vector is $\boldsymbol{v_{k,a_x,t_x}}$, as featured in Figure (b). The empirical CDF of (b) is then calculated, and the resulting distribution F_{v,a_x,t_x} are featured in Figure (c). The red lines in (c) defines the median of the empirical CDF, and the intersection point of the red line with the x-axis defines the "arg-median" (k_x) of the distribution.

5.5.1.9.1.2.2. The Mode Phase: Finding the Consensus (Modal) 𝕜 across all Test Statistics Given an 𝕒

In this phase, the mode of the vector arg-medians of 𝕜s, $\boldsymbol{k_{med_t}}$, was determined, with the modal 𝕜 value be denoted as $\mathbb{k}_{mode}$. The $\mathbb{k}_{mode}$ was assigned to be the consensus 𝕜 value among all the test statistics for the value 𝕒, and any test statistics that suggested a 𝕜 that is proximal with $\mathbb{k}_{mode}$ would be selected.

The arg-medians of 𝕜s are not discrete values however, thus the traditional notion of mode is not applicable. Thus, the $\mathbb{k}_{mode}$ were determined using a method based on averaged shifted histograms proposed by Scott (1985). This involves plotting the histograms of $\boldsymbol{k_{med_t}}$ under varying bin sizes, and from each of the histograms the 𝕜s from a few of the top peaks were selected (Figure 5.15). These 𝕜s were collected into a pool of peak 𝕜s denoted as $\{\mathbb{k}\}_{peak}$, and the $\mathbb{k}_{mode}$ was defined as the median of $\{\mathbb{k}\}_{peak}$.

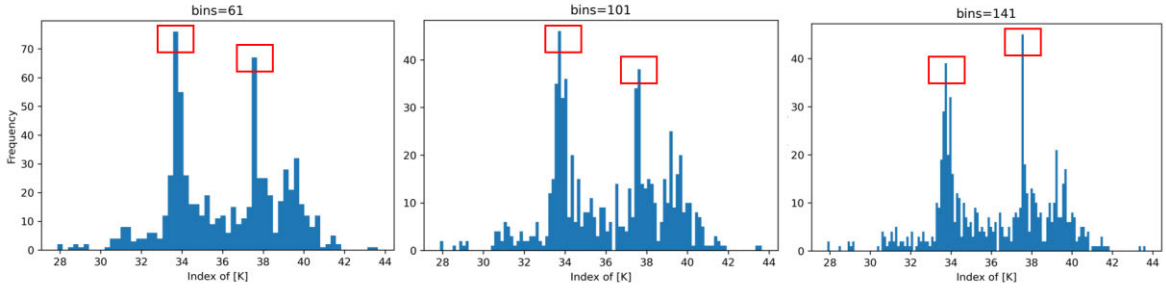

Figure 5.15: Examples of a series of histograms built from the same $\boldsymbol{k_{med_t}}$ under varying number of bins. The 𝕜s from the top few peaks from each of the histogram, demarcated with the red boxes, are then being collected into a pool denoted as $\{\mathbb{k}\}_{peak}$.

The proximity of 𝕜 suggested by a statistical test t (denoted as $\mathbb{k}(t)$) with $\mathbb{k}_{mode}$ is first calculated as the absolute deviation of 𝕜 from $\mathbb{k}_{mode}$:

$$k_{dev}(t) = |\mathbb{k}(t) - \mathbb{k}_{mode}| \qquad [24]$$

The deviation $k_{dev}(t)$ is stored in a vector of length n_{tst} denoted as $\boldsymbol{k_{dev}}$. Quantile filtering was then applied onto $\boldsymbol{k_{dev}}$, with the indices of test statistics that has its k_{dev} located at the bottom 5% of $\boldsymbol{k_{dev}}$ (i.e., most proximal from $\mathbb{k}_{mode}$) being selected. The indices of the chosen test statistics were used to trim $\boldsymbol{V_{k,t}}$ array.

5.5.1.9.1.2.3. The Filtering Phase: Filtering the Test Statistics For an 𝕒

Using the indices of the chosen test statistics, the $\boldsymbol{V_{k,t}}$ was trimmed along the column (i.e., test statistic n_{tst}). This is done by removing columns with outlying test statistics while retaining columns with indices of the chosen test statistics. The resulting array is a trimmed vote tally array, a 2-dimensional array of size $n_{kix} \times n_{tst_{filt}}$, where $n_{tst_{filt}}$ is the number of chosen test statistics, denoted as $\boldsymbol{V_{k,t_f}}$.

This process from Section 5.5.1.9.1.2.1 up to 5.5.1.9.1.2.3 was repeated for all n_{aix} slices of $\boldsymbol{V_{k,t}}$. The resulting $\boldsymbol{V_{k,t_f}}$s from all slices were recompiled into a 3-dimensional array of size $n_{kix} \times n_{aix} \times n_{tst_{filt}}$ denoted as $\boldsymbol{V_{k,a,t_f}}$, and this is the end result of the median-mode consensus filter.

The methodology for the median-mode consensus filter can also be expressed in the following pseudocode:

```
## Input for median-mode consensus filter: V_kat
n_kix, n_aix, n_tst = V_kat.shape
## n_kix: number of levels of [k] tested
## n_aix: number of levels of [a] tested
## n_tst: number of statistical test on equality of distribution conducted

temp_V_ktf_storage = []

for ax in range(n_aix):
        ## MEDIAN PHASE: calculation of median [k] given a test statistics and [a]
        V_kt = V_kat[:,ax,:] ## shape of V_kt: n_kix * n_tst

        # calculating the median [k] value across all test statistics given an [a] value
        k_med_t = zeros(n_tst)
        for tx in range(n_tst):
                v_k_ax_tx = V_kt[:,tx]

                ## calculation of empirical CDF for v_k_ax_tx across [k]
                F_v_ax_tx = cumsum(v_k_ax_tx) / sum(v_k_ax_tx) # eqn [22]
                list_of_all_possible_kx = 1:n_kix
                median_kx = list_of_all_possible_kx[where(F_v_ax_tx == 0.5)] # eqn[23]
                ## median_kx can be non-integer; this expression is used here for clarity
                k_med_t[tx] = median_kx

        # MODE PHASE: Finding the consensus (i.e. modal) [k] across all test statistics
        k_peak = ASH_mode(k_med_t) # mode calculated using averaged shifted histogram (ASH) by Scott
        ## (1985)
        k_mode = median(k_peak)

        ## calculating the deviation of [k] suggested by each test statistics with that of k_mode
        k_dev = abs(k_med_t - k_mode) # eqn [24]
        k_dev_quantile05 = quantile(k_dev, 0.05) # fine the 5% quantile within k_dev_t
        k_dev_which_has_minimal_deviation = where(k_dev <= k_dev_quantile05)

        ## FILTERING PHASE: Filtering the test statistics given an [a]
        V_ktf = V_kt[:,k_dev_which_has_minimal_deviation]
        temp_V_ktf_storage = temp_V_ktf_storage + [V_ktf] # store the filtered V_kt as temporary list

V_katf = as.array(temp_V_ktf_storage) # recompile the temporary list into 3-d array V_katf
```

5.5.1.10. Extracting the Solutions for Estimated Parameters of Genetic Architecture

For the preparation of extraction of solutions of estimated parameters of genetic architecture, the $\boldsymbol{V_{k,a,t_f}}$ was summed across the third axis (i.e. $\boldsymbol{t_f}$ axis), producing a 2-dimensional array of size $n_{kix} \times n_{aix}$, which was denoted as $\boldsymbol{V_{k,a}}$.

$$\boldsymbol{V_{k,a}} = \sum_{i=1}^{n_{tst_{filt}}} \boldsymbol{V_{k,a,t_{f_i}}} \quad [25]$$

The $\boldsymbol{V_{k,a}}$ array was used in the extraction of solution for the estimation of the parameters associated with the genetic architecture.

Using the summed vote tally array $\boldsymbol{V_{k,a}}$, the most likely solutions of $\mathbb{k}$s for each of the $\mathbb{a}$ tested was extracted via a consensus approaches. Theoretically the most likely solutions can be defined as follows: Given a value for parameter $\mathbb{a}$, which $\mathbb{k}$s have successfully minimized the most test statistics (i.e., column-wise mode of the $\boldsymbol{V_{k,a}}$ array). In practice however, directly applying the maximum value on the $\boldsymbol{V_{k,a}}$ array has its own issue, as the dispersion of $\mathbb{D}_{FT}^2$ reduces the stability of vote counts in $\boldsymbol{V_{k,a}}$ and a solution's reliability (Figure 5.16). For this reason, further processing on the $\boldsymbol{V_{k,a}}$ is still required.

For this step, smoothing algorithms were employed. A two-dimensional cubic spline was utilized on the $\boldsymbol{V_{k,a}}$. From this smoothed array, the arguments of the maximum (i.e., the modal $\mathbb{k}$ values) for each column of $\boldsymbol{V_{k,a}}$ were recorded into a vector of length n_{aix} denoted as $\widehat{\boldsymbol{\mathfrak{K}}}_{\boldsymbol{raw}}$. The index of modal $\mathbb{k}$, denoted as $k_{x_{mode}}$, were also recorded for further indexing purposes. An example of the implementation of the two-dimensional spline is illustrated in Figure 5.17.

To further smoothen $\widehat{\boldsymbol{\mathfrak{K}}}_{\boldsymbol{raw}}$, a Savitsky-Golay filter was utilized (Savitzky and Golay, 1964). There are several advantages of the Savitsky-Golay filter compared to cubic splines; the Savitsky-Golay filter preserves many of the essential properties of a distribution such as the moments, width and height of the curve, area under the curve, central tendencies, derivatives and symmetries of the curve while maintaining a least squared fitting, which made it preferable over cubic splines for this purpose (Schafer, 2011; Ziegler, 1981). For this reason, the Savitzky-Golay filter was used to smooth $\widehat{\boldsymbol{\mathfrak{K}}}_{\boldsymbol{raw}}$, and the resulting vector (denoted as

$\widehat{\mathfrak{K}}_{smoothed}$) would become the solution of the estimated number of QTL $\widehat{\mathbb{k}}$ for a value of $\mathbb{a}$. An example of application of cubic spline and Savitzky-Golay filter onto the $\boldsymbol{V}_{k,a}$ array is provided in Figure 5.18.

For each of the $\widehat{\mathbb{k}}$s in $\widehat{\mathfrak{K}}_{smoothed}$, they were paired with its corresponding shape parameter for the QTL effect size distribution $\mathbb{a}$, producing a parameter pair $[\widehat{\mathbb{k}}, \mathbb{a}]$. Using the index of $\mathbb{a}$, a_x, and the index for the corresponding modal $\mathbb{k}$, $k_{x_{mode}}$, the estimated scale parameter $\widehat{\mathbb{b}}$ was indexed from $\widehat{\mathfrak{B}}$, from row $k_{x_{mode}}$ and column a_x. This $\widehat{\mathbb{b}}$ was paired with $[\widehat{\mathbb{k}}, \mathbb{a}]$, and this produced a triplet of estimated parameter values $[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}]$, comprises of estimated number of QTL, shape parameter and scale parameter of the QTL effect size distribution. As there are n_{aix} number of $\mathbb{a}$s tested, there would be same number of triplets, which were compiled into a 2-dimensional array of size $n_{aix} \times 3$ denoted as $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$, with each of the columns containing the estimated number of QTL, shape parameter and scale parameter of the QTL effect size distribution respectively. This is then the final solution for the estimation of the genetic architecture parameters.

This proposed method does not attempt to simplify the $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$ any further. This is due to the non-uniqueness in the solution; there are only two equations available (i.e., the test statistic distribution $\mathbb{d}_{FT}^{1}$ and the additive genetic variance $v_{A_{obs}}$) but with three unknowns that need to be estimated (i.e. $\mathbb{k}$, $\mathbb{a}$ and $\mathbb{b}$ for number of QTL, and shape and scale parameter for QTL effect size distribution respectively). This causes the phenomenon that for each $\mathbb{a}$ provided, there would be a corresponding $\mathbb{b}$ and $\mathbb{k}$ that can be assigned that would produce a set of indistinguishable results (details are provided in Appendix B). Given that any one of the triplets in the $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$ could be the true underlying set of parameters for the genetic architecture but with no additional constraints, it might not be appropriate to further restrict the solutions. Thus, the solution array $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$ was the final output of this algorithm.

An example pseudocode for this step can be expressed as follows:

```
## Extracting the solution from estimated parameters of genetic architecture
## input for this step: V_katf, B_hat, list_of_k, list_of_a
## V_katf : 3-d array that contained filtered number of votes each [k,a] has received
### the more votes the [k,a] had received, the more test statistics that [k,a] has successfully
minimized, this implied a better fit distribution, and more likely being the solution
## B_hat : a 2-d array that contains estimated scale parameter (b_hat) across all [k,a] values
#### b_hat estimated in step 5.5.1.5 and collected across brute-force search in 5.5.1.8
## list_of_k : list of values of [k] tested by the method
## list_of_a : list of values of [a] tested by the method

n_kix, n_aix, n_tstf = V_katf.shape

V_ka = sum(V_katf, axis=2) # shape of V_ka : n_kix * n_aix
V_ka = 2d_spline(V_ka, 11) ## 2-spline with smoothing parameter set at 11; any value could be used.

k_x_mode = numeric(length = n_aix)
K_fraktur_raw = numeric(length = n_aix)
for ax in range(n_aix):
        k_x_mode[ax] = where(V_ka[:,ax] == max(V_ka[:,ax])) # index of the modal k value
        K_fraktur_raw[ax] = list_of_k[k_x_mode[ax]]

K_fraktur_smoothed = savitzky_golay(K_fraktur_raw, 11) ## Savitzky-Golay filter
## with degree of polynomial set at 11; any value could be used for smoothing parameter

## extract the [b] from B_hat
b_fraktur = numeric(length = n_aix)
for ax in range(n_aix):
        b_fraktur[ax] = B_hat[ax,k_x_mode[ax]]

## final solution of estimated parameters
kab_sln = matrix(0, nrow=length(K_fraktur_smoothed), ncol=3)
kab_sln[:,0] = K_fraktur_smoothed
kab_sln[:,1] = list_of_a
kab_sln[:,2] = b_fraktur
```

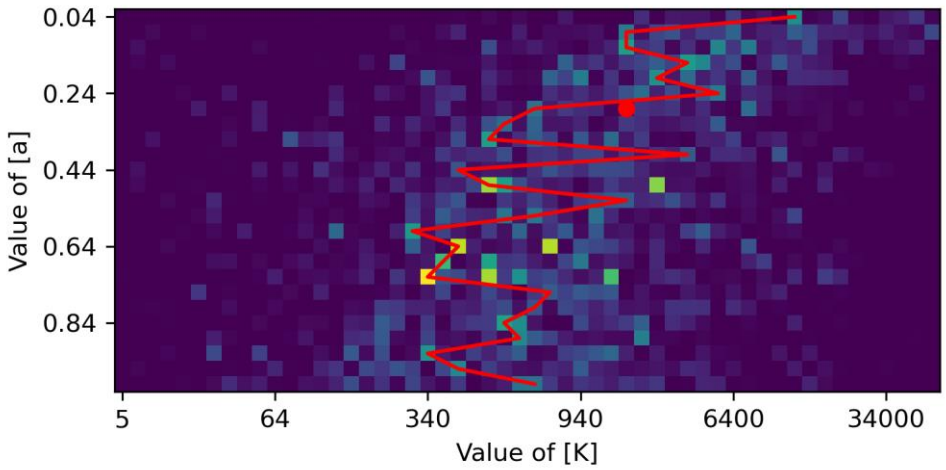

Figure 5.16: The K-a plot showing the filtered $\boldsymbol{V_{k,a}}$ array from Figure 5.15, with the overlying red line connecting the $\mathbb{k}$s that have successfully minimize the most statistics for each of the $\mathbb{a}$s. The red dot signifies the true genetic architecture parameters $Q(2000, 0.3, 1)$ in this example.

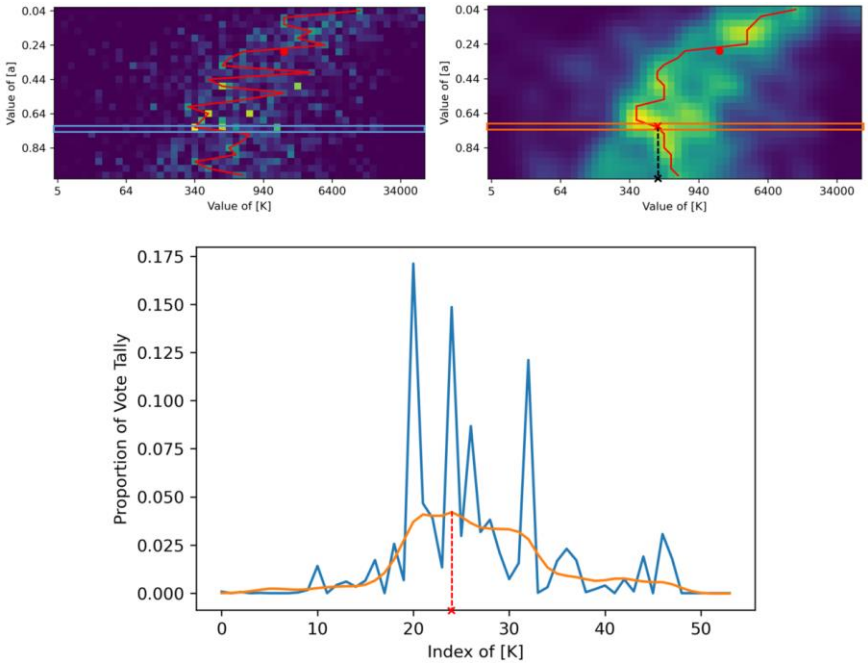

Figure 5.17: The implementation of two-dimensional spline on the $\boldsymbol{V_{k,a}}$ array. Featured in figure (a) is the raw $\boldsymbol{V_{k,a}}$ array that was featured in Figure 5.16, and in figure (b) is the $\boldsymbol{V_{k,a}}$ array that was smoothened. Figure (c) illustrated the histogram obtained by slicing the $\boldsymbol{V_{k,a}}$ array along the n_{aix} axis (in this example, $n_{aix} = 17$, which correspond to $\mathbb{a} = 0.68$), with the blue line obtained by slicing the blue box in the raw $\boldsymbol{V_{k,a}}$ array in (a), and the orange line obtained by slicing the orange box in the smoothened $\boldsymbol{V_{k,a}}$ array in (b). The red line in (b) denotes the $\widehat{\mathfrak{K}}_{\boldsymbol{raw}}$, and the index of modal $\mathbb{k}$, denoted as $k_{x_{mode}}$, is defined the red cross in (c), and the solution of $\mathbb{k}$, $\mathbb{k}_S$, was defined using the black cross in the "Value of [K]" axis in (b). The red dots in (a) and (b) signifies the true genetic parameter architecture $Q(2000, 0.3, 1)$.

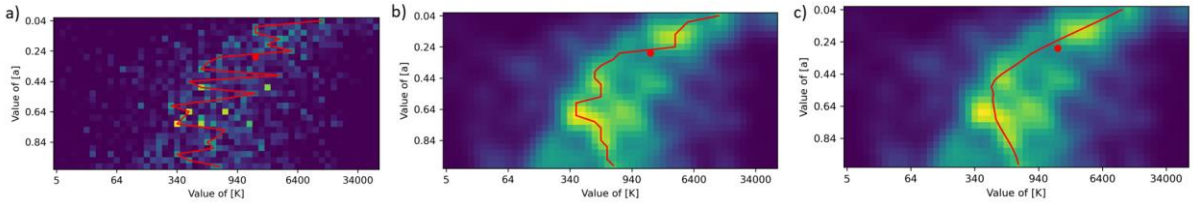

Figure 5.18: The applications of the smoothing methods on the $\boldsymbol{V_{k,a}}$ array, with Figure (a) showing the raw $\boldsymbol{V_{k,a}}$ and the associated estimated solutions. Figure (b) shows the $\boldsymbol{V_{k,a}}$ array that was smoothened by the two-dimensional cubic splines, with the red line containing $\widehat{\mathfrak{K}}_{\boldsymbol{raw}}$. Figure (c) shows the smoothened $\boldsymbol{V_{k,a}}$ array along with the red line containing the $\widehat{\mathbb{k}}$ solutions smoothened by Savitzky-Golay filter $\widehat{\mathfrak{K}}_{\boldsymbol{smoothed}}$.

5.6. Simulation Study for Testing of the Algorithm

5.6.1. Layout of the Experiment

The method was tested through simulation using Python (version 3.9.7, released 30 August 2021) and R (version 3.6.1, released 5 July 2019). The experiment was conducted using genotypic array encoded in the format of {0,1,2}, phenotypes and narrow sense heritability, which the last was assumed to have been estimated using methods outside this chapter. This experiment was conducted on a PC with the following specification: 8-core Intel i7-8665U at 1.90 GHz with 16 GB RAM, with all 8 cores being used.

For this experiment, genotype arrays of sample size of $N = 3000$ and $M = 50{,}000$ markers were utilized. Two genotype arrays were used for this test. The first genotype array tested (denoted as $\boldsymbol{X_1}$) is simulated with homogenous linkage disequilibrium structures, with the allele frequency distribution following a Beta distribution. Correlations between markers were generated by copying part of the genotype from a marker to the adjacent markers, with amount of copying was determined through the level of correlation (denoted as R^2_{tested}). For this study the R^2_{tested} was set at 0.9. The second genotype array (denoted as $\boldsymbol{X_2}$) is generated through coalescence using the R package "AlphaSimR" (Gaynor et al., 2021). For this simulation, the command "RunMac2" was used, with effective population size set at 100, and the mutation rate at 2.5×10^{-8} per base pair per generation. The small effective population size produces heterogeneous linkage disequilibrium structures for this genotype array (Gondro, 2015), and was used to test the vulnerability of this method toward such heterogeneity.

At the same time, a number of markers were nominated as the QTL of the phenotype. For each of these markers a QTL effect size was associated, with the effect sizes follow a gamma distribution with a shape and scale parameters. The genetic architecture parameters tested $Q(\mathbb{k}, \mathbb{a}, \mathbb{b})$ are provided in Table 5.1. These parameter values were designated as the "true parameter values," and they were the target of estimation for this method. A vector of effect sizes of all markers in $\boldsymbol{X_{full}}$, denoted as $\boldsymbol{q_{true}}$, was also generated.

Using the genotype arrays $\boldsymbol{X_{full}}$ and the marker effect sizes $\boldsymbol{q_{true}}$, a vector of phenotypes ($\boldsymbol{y_{full}}$) was generated. The phenotypes are assumed to follow a purely additive model, and is calculated as follows:

$$\boldsymbol{y_{full}} = \boldsymbol{X_{full}q_{true}} + \boldsymbol{y_{full_E}} \quad [26]$$

where the residual component $\boldsymbol{y_{full_E}}$ is generated using the following normal distribution:

$$\boldsymbol{y_{full_E}} \sim \mathcal{N}\left(0, \frac{var\left(\boldsymbol{X_{full}q_{true}}\right) * (1 - h^2)}{h^2}\right) \quad [27]$$

For this experiment, the h^2 was set at 0.3 for all the parameters tested. The $\boldsymbol{y_{full}}$ would become the "observed phenotypes" mentioned in the previous sections.

The $\boldsymbol{X_{full}}$, $\boldsymbol{y_{full}}$ and h^2 were utilized in this method. The genotype array was resampled 27 times (i.e. $n_{sim} = n_{obs} = 27$) for this study, and the number of individuals resampled was set at $N_{rsamp} = 2000$. For the number of QTL $\mathbb{k}$ tested, the following geom-linear progression was used:

$$\mathbb{k} \in \{5,10,16, \dots 94,100,160, \dots 940,1000,1600, \dots 9400,10000,16000, \dots 40000,46000,50000\} \quad [29]$$

In total 54 values of $\mathbb{k}$ were tested. For the shape parameter of QTL effect size distribution $\mathbb{a}$ tested, linear progression was used, starting from $\mathbb{a} = 0.04$, and the common differences for $\mathbb{a}$ set at 0.04. In total 25 $\mathbb{a}$s were tested. This represents a total of 54*25 = 1350 combinations of $[\mathbb{k}, \mathbb{a}]$ being tested. From this method an array of estimated values of the genetic architecture parameter, $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$, was generated.

5.6.2. Genetic Architecture Parameter Tested

For each of the genotype arrays, the method was tested under varying genetic architecture parameters, with the default and alternative values of the number of QTL ($\mathbb{k}$), shape ($\mathbb{a}$) and scale parameter ($\mathbb{b}$) of $\mathbb{d}_{QTL}$ provided in Table 5.1. In total, four combinations of genetic architecture parameters were used.

For each genetic architecture and genotype tested, three replications of the experiment were done. In total 24 (4*2*3 = 24) tests were conducted in this study.

Table 5.1: Genetic architecture parameters tested in this experiment, with the first value in $Q(\mathbb{k}, \mathbb{a}, \mathbb{b})$ denotes the number of QTL, the second and third values denote the shape and scale parameter of the true QTL effect size distribution.

Genetic Architectures	Parameter Value Tested
Defaults	$Q(300, 0.3, 1)$
Alternatives	$Q(2000, 0.3, 1)$
	$Q(300, 0.8, 1)$
	$Q(300, 0.3, 3)$

5.6.3. Testing the Performance of the Algorithm

In this study, the performance of the method is defined as the method's capability of estimating the true $\mathbb{d}_{QTL}$ by producing a solution of $[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}]_{\boldsymbol{sln}}$ that their distributions accurately reflect said true $\mathbb{d}_{QTL}$. The closer the estimated distributions (denoted as $\widehat{\mathbb{d}}_{QTL}$) are to the true $\mathbb{d}_{QTL}$, the higher the performance of the method.

Three measures were employed to test the performance of the method. The first measure is a modified version of Wasserstein's statistic, which is defined as the area between the $1 - CDF$ of the $\widehat{\mathbb{d}}_{QTL}$ and $\mathbb{d}_{QTL}$ (denoted as $\widehat{\mathbb{D}}_{QTL}$ and $\mathbb{D}_{QTL}$ respectively).

For this experiment, the area under the curves of $\widehat{\mathbb{D}}_{QTL}$ and $\mathbb{D}_{QTL}$ (denoted as $A_{\widehat{\mathbb{D}}}$ and $A_{\mathbb{D}}$) were defined as follows:

$$A_{\widehat{\mathbb{D}}_i} = 50000 * \int_{-\infty}^{\infty} \widehat{\mathbb{D}}_{QTL_i}\, d(QTL\ size) \quad [28]$$

$$A_{\mathbb{D}} = 50000 * \int_{-\infty}^{\infty} \mathbb{D}_{QTL}\, d(QTL\ size) \quad [29]$$

And the performance of the method in term of the modified Wasserstein's statistic is defined as follows:

$$\boldsymbol{p}_{\boldsymbol{WAS}_i} = 1 + \frac{\left[50000 * \int_{-\infty}^{\infty} \widehat{\mathbb{D}}_{QTL_i} - \mathbb{D}_{QTL}\, d(QTL\ size)\right]}{\left[50000 * \int_{-\infty}^{\infty} \mathbb{D}_{QTL}\ d(QTL\ size)\right]} \quad [30]$$

where the square brackets $[x]$ mean "round the numbers to the nearest integers."

Rather than the absolutize differences between $\widehat{\mathbb{D}}_{QTL}$ and $\mathbb{D}_{QTL}$ as in the usual Wasserstein's statistic, the raw differences were taken for this statistic. This is to test the level of overestimation or underestimation of the number of QTL by the method: if the $\widehat{\mathbb{D}}_{QTL}$ matches perfectly with $\mathbb{D}_{QTL}$, the $\boldsymbol{p_{WAS}}$ would be 1. If $\boldsymbol{p_{WAS}}$ is more than 1, the method has overestimated the number of QTL, and vice versa. This measure can range from 0, when $\widehat{\mathbb{D}}_{QTL}$ is a constant value of 0, up to infinity. The proportionality of this statistics is to standardize the areas under the curves under varying parameter values.

It is also noted that if the area under the curves of $\widehat{\mathbb{D}}_{QTL}$ and $\mathbb{D}_{QTL}$ are the same, the measures would be 1, regardless of how severe the actual discrepancies between the two distributions. This is the reason for employing the second and third measurements for the performance of this method, which are the number of QTL with certain effect sizes.

Let the effect size tested be denoted as a_{cut}, and the $\mathbb{D}_{QTL}(a_{cut})$ and $\widehat{\mathbb{D}}_{QTL}(a_{cut})$ be defined as the true and estimated number of QTL with effect size of a_{cut}. The performance of the method in term of number of QTL was defined as follows:

$$\boldsymbol{p_{QTL=a_{cut}}} = 1 + \frac{\widehat{\mathbb{D}}_{QTL_i}(a_{cut}) - \mathbb{D}_{QTL}(a_{cut})}{\mathbb{D}_{QTL}(a_{cut})} \qquad [31]$$

As in $\boldsymbol{p_{WAS}}$, the $\boldsymbol{p_{QTL=a_{cut}}}$ could ranges from 0, when $\widehat{\mathbb{D}}_{QTL_i}(a_{cut}) = 0$, up to infinity, when $\mathbb{D}_{QTL}(a_{cut}) = 0$. If the statistic is greater than 1, the method has overestimated the number of QTL with effect size a_{cut}, and vice versa. In this study, the second measurement would have $a_{cut} = 0.1\ \sigma_e$ and the third measurement with $a_{cut} = 1.0\ \sigma_e$.

As there are n_{aix} number of solution triplets in $\left[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}\right]_{\boldsymbol{sln}}$, the $\boldsymbol{p_{WAS}}$ and $\boldsymbol{p_{QTL=a_{cut}}}$ would also be a vector of length n_{aix}. For this reason, the performance scores were represented by their median. From each of the replicates, the median of $\boldsymbol{p_{WAS}}$ as well as $\boldsymbol{p_{QTL=a_{cut}}}$ with $a_{cut} = 0.1\ \sigma_e$ and $a_{cut} = 1.0\ \sigma_e$. The overall performance for each measurement was defined as the mean of the medians across all replicates.

5.7. Results

5.7.1. The Performance of the Algorithm

The area under the curve for the $1 - CDF$ of the true QTL effect size distribution, denoted as $A_{\mathbb{D}}$, and the median of the area under the curve for those of estimated QTL effect size

distribution, denoted as $A_{\mathbb{D}}$, for genotype array $\boldsymbol{X_1}$, as well as the true and estimated number of QTL with effect size of 0.1 σ_e and 1.0 σ_e, are provided in Table 5.2. The performance of the method to estimate the parameters of the genetic architecture, measured in terms of areas between curves and number of QTLs, for genotype array 1, is provided in Table 5.3. For genotype array $\boldsymbol{X_2}$, the area under the curves for the $1 - CDF$ of the true and estimated QTL effect size distribution, as well as the true and estimated number of QTL with effect sizes of 0.1 σ_e and 1.0 σ_e, are provided in Table 5.4, and the performance of the method in terms of these measures is provided in Table 5.5.

In general, the method successfully provided an estimate for the number of QTL and the distribution of its effect sizes. This is evident from the proximity of estimated QTL effect size distributions with those of true QTL effect size distributions in all genetic architecture tested for both genotype arrays (Figure 5.19 – 20).

Under default parameters the number of QTL with effect size 0.1 σ_e estimated by the method is 148.8% of the true number of QTL for genotype array $\boldsymbol{X_1}$, and for genotype array $\boldsymbol{X_2}$ it is 133.7% of the true number of QTL. Whereas for QTL with effect size 1.0 σ_e, the number of QTL estimated for genotype array $\boldsymbol{X_1}$ was 132.4% of the true number of QTL, and for genotype array $\boldsymbol{X_2}$ it was 106.8% of the true number of QTL. Across all genetic architecture parameters and genotype arrays tested, the estimated number of QTL with effect size 0.1 σ_e ranges from 69.9% to 167.0%, with an average of 109.8% of the true number of QTL, and for estimated number of QTL with effect size 1.0 σ_e ranges from 101.6% to 175.8%, with an average of 123.6% of the true number of QTL (Table 5.2 – 5).

5.7.2. Overviews on the Trends of the Outputs

While the method is able to provide an estimate of QTL effect size distribution, the non-uniqueness in the solution of the estimated parameter values has introduced ambiguity in the proposed models, which manifested as a thick band of estimated distributions (in red) rather than one singular distribution (Figures 5.12-5.13). The effects of non-uniqueness in the solution can also where the solution manifested itself as a band of solutions (in yellow) rather than one singular spot on the plot, with the width of the band as the distribution around the estimated parameter values, analogous to a confidence interval (Figure 5.21). This band of solutions contains combinations of $[\mathbb{k}, \mathbb{a}]$s that produces similar QTL effect size distributions.

Table 5.2: The medians of measures for various genetic architecture parameter tested with Genotype Array $\boldsymbol{X_1}$. The measures were defined in terms of area under the curves of true and estimated QTL effect size distribution ($A_{\mathbb{D}}$ and $A_{\widehat{\mathbb{D}}}$ respectively), as well as the true and estimated number of QTL ($\mathbb{D}_{QTL}$ and $\widehat{\mathbb{D}}_{QTL}$ respectively) with effect size of 0.1 σ_e and 1.0 σ_e.

Genetic Architecture Parameters		Representative Medians of Measures					
		Area Under the Curves		Number of QTL with 0.1 σ_e		Number of QTL with 1.0 σ_e	
		$A_{\mathbb{D}}$	$A_{\widehat{\mathbb{D}}}$	$\mathbb{D}_{QTL}$	$\widehat{\mathbb{D}}_{QTL}$	$\mathbb{D}_{QTL}$	$\widehat{\mathbb{D}}_{QTL}$
Default	$Q(300, 0.3, 1)$	98.32	134.77	142.67	211.67	27.56	35.77
Alternative	$Q(2000, 0.3, 1)$	601.69	662.03	898.78	626.23	159.33	220.23
	$Q(300, 0.8, 1)$	241.35	335.95	252.00	406.52	86.33	108.64
	$Q(300, 0.3, 3)$	253.63	398.29	181.89	294.26	70.67	123.61

Table 5.3: The performance of the method in the estimation of genetic parameter architectures in Genotype Array $\boldsymbol{X_1}$, evaluated in term of Wasserstein's statistics and number of QTL with effect size of 0.1 σ_e and 1.0 σ_e.

Genetic Architecture Parameters		Performance		
		Wasserstein's Statistics	Number of QTL with 0.1 σ_e	Number of QTL with 1.0 σ_e
Default	$Q(300, 0.3, 1)$	1.374	1.488	1.324
Alternative	$Q(2000, 0.3, 1)$	1.101	0.699	1.380
	$Q(300, 0.8, 1)$	1.400	1.613	1.264
	$Q(300, 0.3, 3)$	1.571	1.609	1.758

Table 5.4: The medians of measures for various genetic architecture parameter tested with Genotype Array $\boldsymbol{X_2}$. The measures were defined in terms of area under the curves of true and estimated QTL effect size distribution ($A_{\mathbb{D}}$ and $A_{\widehat{\mathbb{D}}}$ respectively), as well as the true and estimated number of QTL ($\mathbb{D}_{QTL}$ and $\widehat{\mathbb{D}}_{QTL}$ respectively) with effect size of 0.1 σ_e and 1.0 σ_e.

Genetic Architecture Parameters		Representative Medians of Measures					
		Area Under the Curves		Number of QTL with $0.1\sigma_e$		Number of QTL with $1.0\sigma_e$	
		$A_{\mathbb{D}}$	$A_{\widehat{\mathbb{D}}}$	$\mathbb{D}_{QTL}$	$\widehat{\mathbb{D}}_{QTL}$	$\mathbb{D}_{QTL}$	$\widehat{\mathbb{D}}_{QTL}$
Default	$Q(300, 0.3, 1)$	100.19	127.62	142.78	193.42	30.33	32.46
Alternative	$Q(2000, 0.3, 1)$	569.79	592.37	866.67	869.39	156.78	169.89
	$Q(300, 0.8, 1)$	242.68	300.10	252.67	429.99	88.78	89.01
	$Q(300, 0.3, 3)$	274.29	293.90	173.89	164.78	77.67	82.32

Table 5.5: The performance of the method in the estimation of genetic parameter architectures in Genotype Array $\boldsymbol{X_2}$, evaluated in terms of Wasserstein's statistics and number of QTL with effect size of 0.1 σ_e and 1.0 σ_e.

Genetic Architecture Parameters		Performance		
		Wasserstein Statistics	Number of QTL with 0.1 σ_e	Number of QTL with 1.0 σ_e
Default	$Q(300, 0.3, 1)$	1.293	1.337	1.068
Alternative	$Q(2000, 0.3, 1)$	1.052	1.018	1.089
	$Q(300, 0.8, 1)$	1.253	1.670	1.016
	$Q(300, 0.3, 3)$	1.076	0.932	1.090

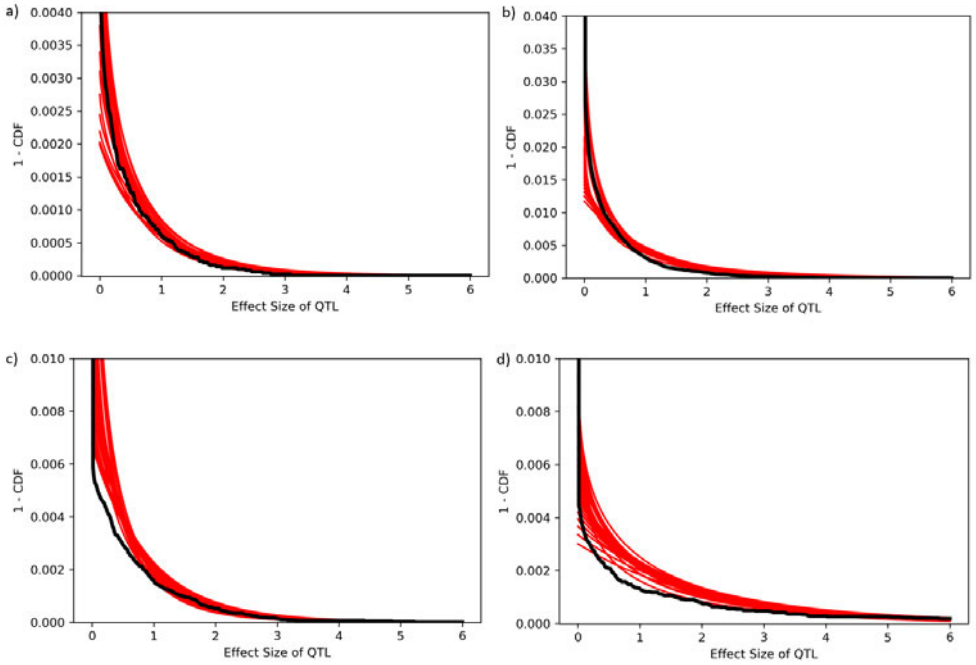

Figure 5.19: The comparison plots between the true QTL effect size distribution (black line) and the estimated QTL effect size distribution (red lines) for each of the parameter combination tested in Genotype Array $\boldsymbol{X_1}$. The genetic architecture parameters tested is as follows: (a) $Q(300, 0.3, 1)$; (b) $Q(2000, 0.3, 1)$; (c) $Q(300, 0.8, 1)$ and (d) $Q(300, 0.3, 3)$.

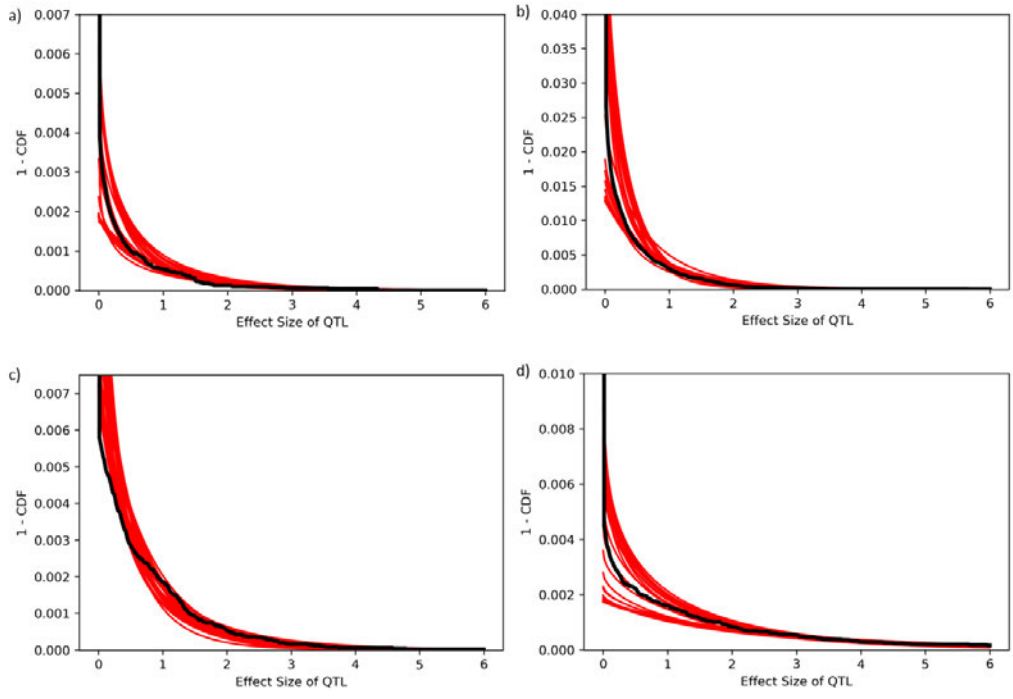

Figure 5.20: The comparison plots between the true QTL effect size distribution (black line) and the estimated QTL effect size distribution (red lines) for each of the parameter combination tested in Genotype Array $\boldsymbol{X_2}$. The genetic architecture parameters tested is as follows: (a) $Q(300, 0.3, 1)$; (b) $Q(2000, 0.3, 1)$; (c) $Q(300, 0.8, 1)$ and (d) $Q(300, 0.3, 3)$.

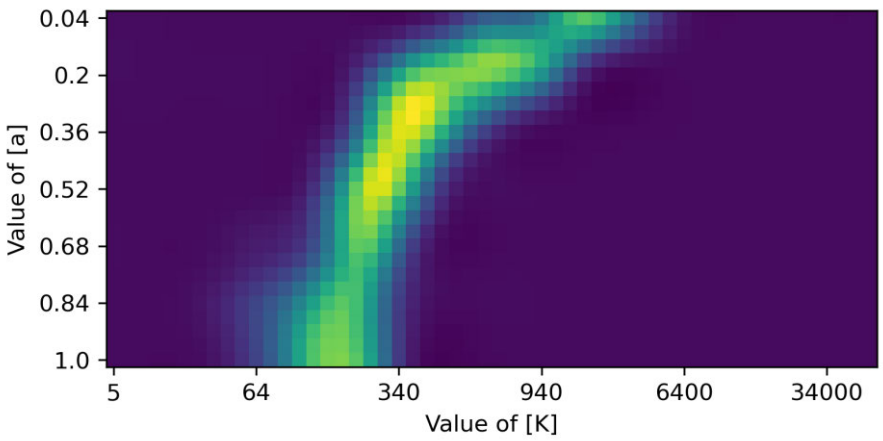

Figure 5.21: An example of the K-a plot, showing a yellow band of solutions that cut across all values of 𝕒s. This plot is generated using genetic architecture parameters $Q(300, 0.3, 1)$ on Genotype Array $\boldsymbol{X_1}$.

The value of shape parameter 𝕒 also influences the estimated number of QTL $\widehat{\mathbb{k}}$; when the 𝕒 is small, the estimated value of $\widehat{\mathbb{k}}$ would also be large, and vice versa. This relationship of $\widehat{\mathbb{k}}$ and 𝕒 can also be observed in the K-a plots, where the solutions produced by the method have large $\widehat{\mathbb{k}}$ when the 𝕒 is small, and as 𝕒 increases the corresponding $\widehat{\mathbb{k}}$ decreases, reaching a minimum with $\mathbb{a} = 1$ (Figure 5.22).

Another notable observation from the output of this method is that the global maxima across all [𝕜, 𝕒]s in the K-a plot do not necessarily correspond to the true combinations of the underlying genetic architecture parameters. This is true even if the global maxima are located in the band of solution, which in this case only represent one of the infinite possible solutions. This can be observed as a mismatch between the global maximum and the true QTL effect size parameters in the K-a plot (Figure 5.23).

Heterogeneity in linkage disequilibrium structures also has significant effects on the estimated distribution from the method. A heterogeneous linkage disequilibrium structure was associated with an increased dispersion of the distribution of the vote tally from vector $\boldsymbol{v}_{k,a_x,t_x}$ (Figure 5.24). This in turn causes an increased error of estimation of 𝕜s for a value of 𝕒, which can be seen by the increased width and reduced regularity of the solution band in the K-a plots (Figure 5.25). Despite this, even with heterogeneity in the linkage disequilibrium structures, the proximity of the estimated QTL effect size distribution from those from true QTL effect size distribution for both genotypic arrays in Figure 5.19 – 20 suggested this method is able to handle the effects from heterogeneity of linkage disequilibrium structures.

Generally, the error of estimation is greater for smaller QTL effect sizes. This error can manifest itself in several ways, such as differing estimated number of QTL with small effect sizes compared to true number of QTL) (Figure 5.26(a)), and the increased dispersion of estimated number of QTL across the solutions (Figure 5.26(b)). The dispersion effect is especially severe for large $\mathbb{b}$ in Genotype Array $\boldsymbol{X_2}$, where the distributions from the solution spread out near the left side of the plot (Figure 5.26(b-c)). These errors of estimation became less apparent for QTL with larger effect sizes, and this observation is more obvious for Genotype Array $\boldsymbol{X_2}$, where the estimations are less than 10% away from the true number of QTL.

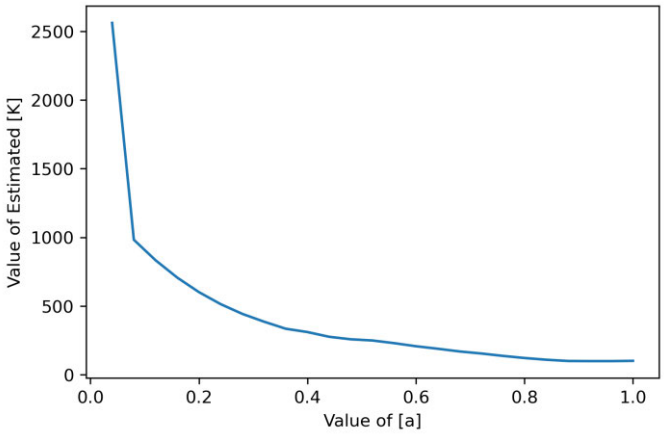

Figure 5.22: Plot of estimated value of $\widehat{\mathbb{k}}$ over varying values of $\mathbb{a}$. The true genetic architecture for this plot is $Q(300, 0.3, 1)$ and is conducted on Genotype Array $\boldsymbol{X_1}$.

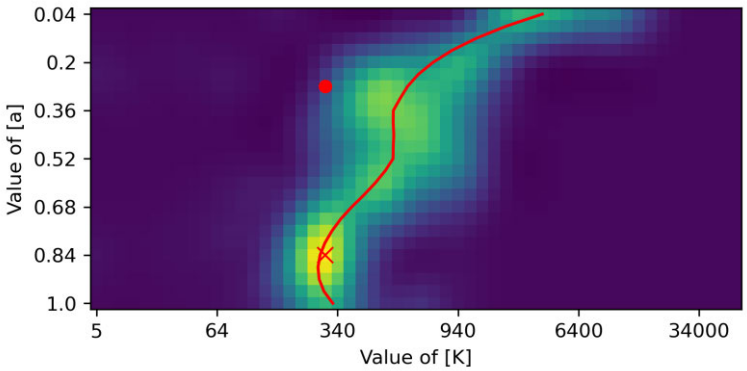

Figure 5.23: A K-a plot showing the goodness of fit between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ for each of the proposed $[\mathbb{k}, \mathbb{a}]$ models, showing a band of solutions that successfully maximized the goodness of fit. The lighter the pixel the greater the goodness of fit is. The red line denotes the solution $[\mathbb{k}, \mathbb{a}]$ models estimated by the method. The red dot denotes the true parameter combinations for the underlying genetic architecture (i.e. $Q(300, 0.3, 3)$ in this plot), while the red cross denotes the global maximum of the K-a plot. This plot is generated from Genotype Array $\boldsymbol{X_1}$.

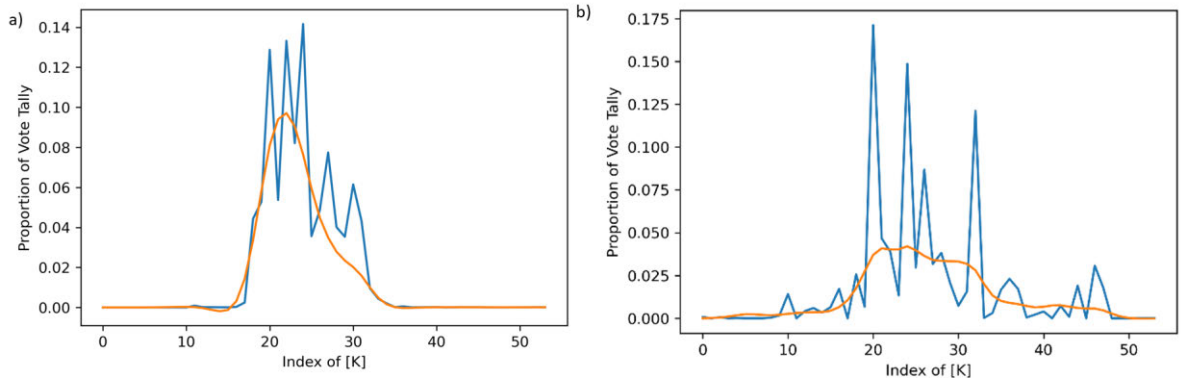

Figure 5.24: The distribution of the vote tally of $\boldsymbol{v}_{k,a_x,t_x}$ across the index of $\mathbb{k}$s from (a) Genotype Array $\boldsymbol{X_1}$ with homogenous linkage disequilibrium structures, and (b) Genotype Array $\boldsymbol{X_2}$ with heterogeneous linkage disequilibrium structures. Raw distribution is show in blue, and smoothed distributions in orange.

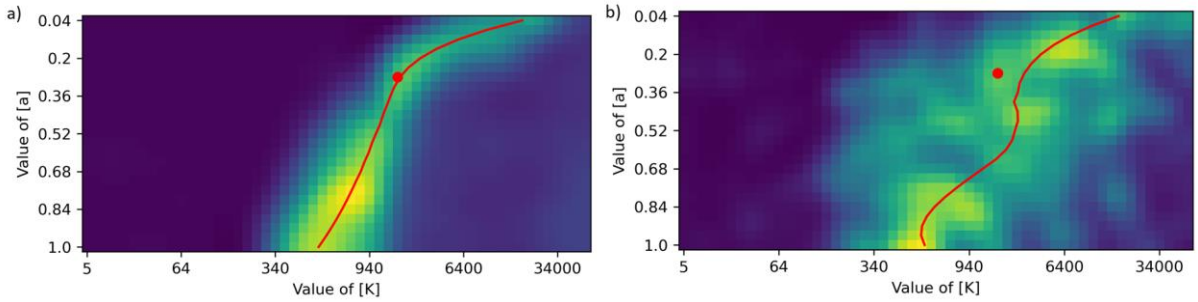

Figure 5.25: The effects of heterogeneity in linkage disequilibrium structures on the estimated QTL effect size distribution from the method. The K-a plot in (a) is generated with Genotype Array $\boldsymbol{X_1}$ with homogenous linkage disequilibrium structures, while those in (b-d) are generated from Genotype Array $\boldsymbol{X_2}$ with heterogeneous structures. Both figures are generated from the genetic architecture parameter of $Q(2000, 0.3, 1)$.

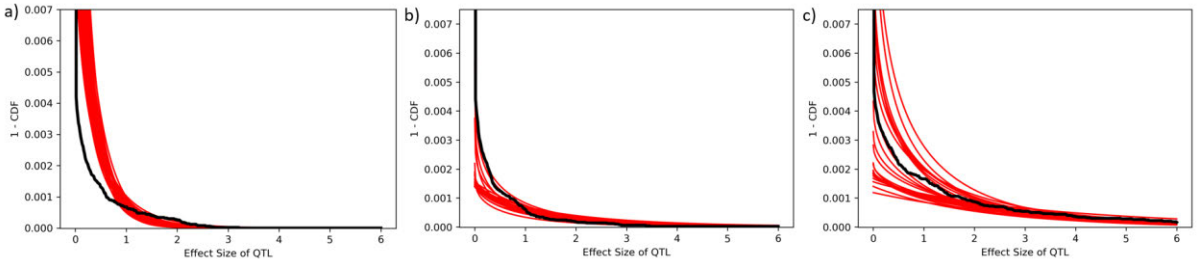

Figure 5.26: Comparison plots showing the error of estimation near regions of small effect sizes between the true QTL effect size distribution (black lines) and estimated QTL effect size distributions (red lines). Figure (a) shows the biased estimated number of QTL of small effect size and the correspondingly decreased estimated number of QTL with large effects. Figure (b) and (c) shows the increased dispersion of estimated number of QTL with small effect sizes. Both Figure (a) and (b) are generated from genetic architecture parameters $Q(300, 0.3, 1)$ while (c) is generated from $Q(300, 0.3, 3)$. Genotype Array $\boldsymbol{X_2}$ was utilized for all of these graphs.

5.8. Discussion

In this study a method that estimates the number of QTL associated with a trait, as well as the parameters associated with the distribution of the QTL effect sizes, has been proposed. Despite the non-uniqueness in the solution proposed, in general this method has successfully estimated the number of QTL and its distribution. Rather than attempting to estimate the genetic architecture parameters such as number of QTL using GWAS results that could be affected by various confounding factors such as allele frequencies and their distribution and correlations between markers, this method takes into account their effects by modelling the expected distribution of the test statistics of the GWAS with the effects of these confounding factors included. This reduces the severity of their impacts on the estimation of genetic architecture parameters. As an example, linkage disequilibrium structures affect the false positive rates of a GWAS (Kaler and Purcell, 2019), which in turn could affect the estimation of genetic architecture parameters. Despite this, this method has successfully estimated genetic architecture parameters using two genotypic arrays with differing linkage disequilibrium structures, which suggested its robustness against the effects of the linkage disequilibrium structures.

There are numerous prospective uses for this method. The main utility of this method is to estimate the number and distribution of the QTL effect size. While methods of similar nature have been published in previous work (Cheng et al., 2020; Park et al., 2010; Zhang et al., 2018), as well as several Bayesian-based methods (as example, Habier et al., 2011; Meuwissen et al., 2001; Moser et al., 2015), this method is different such that it attempts to achieve such aim while considering the effects from varying confounding factors such as correlation between markers, which increases the dispersion of the distribution of the GWAS test statistics, and heterogeneity in linkage disequilibrium, which could affect the number of QTL estimated. Unlike Park et al. (2010), this method also does not require previously published GWAS results, which can be used in studying the genetic architectures of a new trait. This method also not requiring a user-defined cut-off point between null and non-null markers as in Cheng et al. (2020), thus would not be affected by the optimality of the cut-off point. While simulated data were used to test the performance of this method, it can also be used in real-life scenarios since the method requires only genotype, phenotype and narrow sense heritability for the estimation. Further study can be done to test this method with real data.

With these advantages, this method would be suited to populations with small effective population sizes, such as that in a livestock production system (Gondro, 2015; Toro Ospina et al., 2019). Small effective population sizes increase the linkage disequilibrium between the markers (Sved, 1971; Waples et al., 2016), which in turn affects the estimation of genetic architecture parameters by altering the distribution of the estimated marker effect sizes and test statistics from a GWAS (detailed in Appendix B). Besides this, selection processes in a livestock production system induce the formation of haplotype blocks due to heterogeneous rate of recombination across the genome, which produces heterogeneity in the linkage disequilibrium structures (Ardlie et al., 2002). Such structures also affect the distributions of estimated effect sizes and test statistics, which in principle would also affects the estimation of the genetic architecture parameters. By taking the effects of linkage disequilibrium into account. Indeed, Lloyd-Jones et al. (2019) commented on the negative effects of linkage disequilibrium on the convergence of their model. By taking the effects of these linkage disequilibrium into account, it is anticipated this method could provide a more accurate picture of the genetic architecture of a trait.

The proposed method can also be used to estimate the effect sizes of the markers, as in a GWAS experiment. This could be done by feeding the $[\widehat{\mathbb{k}}, \mathbb{a}, \widehat{\mathbb{b}}]_{\boldsymbol{sln}}$ back into the Step 5.5.1.4 in this method and extracting the resulting $\boldsymbol{q_{sim}}$. Unlike GWAS however, rather than returning an estimated effect sizes that could be confounded by other factors, it returned an estimate of marker effect sizes while considering those confounding factors (i.e., what is the expected effect sizes if such confounding factor exists in the dataset). The output from this method can be directly used to estimate the additive effect size of a QTL, a genomic region, or an animal. Using a modest sample size of a few thousand (e.g., 3000 in this study), this method can also be used to model the distribution of the QTL with effect sizes generally undetectable across the genomic sequence using a conventional GWAS. By doing so, this method will provide an insight into the statistical behaviour of a GWAS experiment, which could also serve as a stepping stone for solving the missing heritability problem (Hall et al., 2016; Maher, 2008; Manolio et al., 2009).

Another feature for this method is the flexibility of distribution assumed by the method. Unlike previous publications such as Cheng et al. (2020); Moser et al. (2015) and Zhang et al. (2018), this method did not force one specific distribution for the QTL effect sizes. While the method in this study utilized a gamma distribution, any one or two parameter distributions

that its range of parameters can be discretised into a finite number of steps can be used for this method. One such distribution is Weibull distribution, defined as follows (Mun, 2012):

$$f(x; a, b) = \frac{a}{b}\left(\frac{x}{b}\right)^{a-1} e^{-\left(\frac{x}{b}\right)^{a}} \quad [34]$$

where a and b are the shape and rate parameters for the Weibull distribution respectively. Similar to gamma distribution, a Weibull distribution has the "large number of small QTL, small number of large QTL" shape if $0 < a \leq 1$, which means the a in this distribution can be discretised into a finite number of steps. The b can be treated in a similar fashion as in $\mathbb{b}$ in this study. Thus, a Weibull distribution can also be used in this modelling.

Another distribution that could be considered is the q-exponential distribution, which its normalized form is defined as follow (Picoli et al., 2009):

$$f(x; q, b) = \frac{2-q}{b}\left(1 - \frac{(1-q)x}{b}\right)^{\frac{1}{1-q}} \quad [35]$$

If $q = 1$ then the distribution reduces to standard exponential distribution (Picoli et al., 2009). This particular distribution is only defined if $0 < q \leq 2$, thus can be discretised into finite number of qs, and the b in the same way as in the gamma or Weibull distributions counterparts. This means q-exponential distribution could also be used for this method.

The advantage for allowing a flexible choice of distribution lies in its improved performance in detecting varying genetic architecture for a trait (Zeng and Zhou, 2017). Previously published methods such as Moser et al. (2015) and Habier et al. (2011) assumed the QTL effect sizes to be distributed according to a normal or a mixture of normal distributions, which might fail to properly capture the tail of the effect size distribution due to the fixed kurtosis of these distribution, especially if large causal variants are present (Mun, 2012; Zeng and Zhou, 2017). While the true underlying distribution effect sizes of a trait is usually unknown (Moser et al., 2015), the flexibility of this method means we can test the genetic architecture of a trait using multiple types of distribution, which could potentially lead to a clearer picture on the true underlying distribution of a trait. This flexibility in choice of distribution would also mean this method can be used in a wide range of traits, be it an oligogenic trait, such as horned/polled phenotypes in goats (Guo et al., 2021), as well as in polygenic traits, such as milk yield in cattle (Nayeri et al., 2016), thus allowing the dissection of their genetic architectures.

From a pure mathematical point of view, the proposed method has also introduced several new classes of nonparametric statistical tests that are powerful in detection of discrepancies near the tail of the distribution and yet insensitive toward discrepancies at the head of the distribution, which is useful in testing discrepancy at the tail of the distribution. This could be important in modelling extreme events in ecology (Batt et al., 2017), disaster management (Alvarado et al., 1998; Tippett et al., 2016), engineering (Fortin and Clusel, 2015; Orsini et al., 2019) and finances (Chavez-Demoulin and Embrechts, 2004). This study has also introduced several techniques available to amplify a weak signal that could not be easily detected, especially from a noisy dataset. This study has also introduced the concept of multi-level distributions and its application in statistical testing, which could be useful in testing the replicability of an experiment, and the expected distribution from data from an experiment. In term of quantitative genetics and genomics, this study could provide insights into the distribution of QTL effects based on a GWAS under varying confounding factors and genetic architecture, which can be useful in optimizing the power and false positive rate of a GWAS.

There are several aspects in this method that could be further improved. One such aspect is the speed of the method; the brute force search of this method means all the parameter combinations have to be evaluated, which reduces the speed of running the method (Bergstra and Bengio, 2012). This could be problematic for high density genetic markers or Whole Genome Sequence (WGS). For this reason, faster methods that could cover the parameter spaces could be considered, such as successive halving method (SHA) that weeds out parameter combinations that produces poor results (Jamieson and Talwalkar, 2015).

The unbalanced number of equations and variables has introduced non-uniqueness into the solutions, which causes global maxima in the $\boldsymbol{V}_{\boldsymbol{k},\boldsymbol{a}}$ array that no longer correspond to the true underlying parameters $[\mathbb{k}, \mathbb{a}]$. There are several potential methods to resolve this issue, all of which involve the removal of the excessive degree of freedom. The most straightforward but not necessarily simplest method is to find the third independent equation that can further constrain the solutions. The third independent equation should have its output variable alongside the genetic architecture parameters. A shortcoming for this method is the fact that the independent equation might not exist. The distribution for the estimated effect sizes $\mathbb{d}_{ES}^{1}$ cannot be used as the third equation as $\mathbb{d}_{ES}^{1}$ is not independent from $\mathbb{d}_{FT}^{1}$; one can derive the $\mathbb{d}_{FT}^{1}$ using the $\mathbb{d}_{ES}^{1}$, rendering it non-independent (as suggested in equation [6]). Another method is to constrain the value of one of the parameters, such as the scale parameter $\mathbb{b}$. This however came with the cost of compromising the flexibility of the method, and this could

compromise the accuracy of the estimated genetic architectures. An alternative method is to model the genetic architecture with a distribution that require less parameters, but such distributions tend to have fewer flexible shapes. Methods to resolve this issue could warrant further study.

It is also noted that as the value of $\mathbb{k}$s increases, the effects of changing a fixed amount of values diminishes, and this resulted in a weakening of signals as the $\mathbb{k}$ increases. This increases the chance of having signals from the changing $\mathbb{k}$s swamped by the noises, thus suggesting a potential weakness in this method in estimating the QTL effect size distribution for a strongly polygenic trait. There are several methods of mitigating such issue, with one obvious way to increase the resampled sample size N_{rsamp}, which might improve the signal of changing genetic architecture while reducing the amount of noises but would come at the cost of reducing the speed and feasibility of the method.

As this method utilized GWAS as part of its operations, it would also suffer from the same inherent issues that plague a GWAS. For example, extreme allele frequencies are known to reduces the test statistics of a GWAS (Spencer et al., 2009). This means that if the QTL is located near a locus with extreme allele frequencies, this method might not be able to detect it. Population stratification could produce false positives in a GWAS experiment (Panagiotou and Ioannidis, 2012), and this method would also suffer from these issues. Further studies could also be conducted to improve the method over these shortcomings.

In conclusion, a method that estimates the genetic architecture parameters such as number of QTL and shape of QTL effect size distribution with consideration on the effects of confounding factors of heterogeneity in linkage disequilibrium structure, allele frequencies and their distribution and correlation between markers has been proposed in this study. Using a modest sample size, this method successfully detected the number of QTL and the distribution of the effect sizes that is associated with a trait. In the process this method has also introduced new techniques and classes of statistics that is powerful at the tail of the distribution. Despite this, further studies could be done to improve the speed of the method, as well as improving the power of this method to estimate the distribution of QTL effect sizes in a strongly polygenic trait. Finally, further studies could also be conducted on the statistics to improve their power while reducing their false positives, thereby improving the robustness of the method.

Chapter 6. An Optimal Contribution Selection Algorithm that Utilizes Non-additive Genetic Effects

Zhi Loh, Julius H. J. van der Werf, Sam Clark

6.1. Abstract

The aim of this study is to propose an Optimal Contribution Selection (OCS) algorithm that utilizes both additive and dominance genetic variance while constraining inbreeding. Using a genetic algorithm, the contribution of sires and dams to the next generation, as well as their mate allocation, was optimised. EBVs and expected progeny heterozygosity were used as scores of optimizations for additive and dominance genetic components, respectively. Under the same constrained inbreeding rate of 1%, narrow sense heritability of 0.3, dominance to additive genetic variance ratio of 15% and 500 sires and dams, the OCS algorithm increased the genetic gain compared to truncation selection, with the additive genetic component up by 87.0% from +1.34 to +2.59 while the increase in genetic merit due to the dominance genetic component improved from -0.36 in truncation selection to +4.98 in the first generation, with the corresponding total genetic merit increasing from +0.98 to +7.57. Therefore, the genetic lift in the first generation is approximately equal to two generations of additive genetic gain when dominance variance was 15% of the additive genetic variance. By including dominance genetic component optimization in an OCS, the total genetic merit improved from +2.55 to +7.57 in the first generation. The optimization of the dominance genetic component result in additional genetic merit only in the first generation, with additional merits from this component not increasing after the first generation despite continued optimization. In conclusion, the inclusion of dominance genetic component in an OCS significantly improved the genetic merit of the offspring. While simulated data was used in this study, it is anticipated this method can be used with real data in a selective breeding program in a livestock production system.

6.2. Introduction

While selective breeding has played a major role in livestock production systems, the selection process has also been associated with an increased level of inbreeding, which can lead to inbreeding depression and a loss of genetic variation (Falconer, 1989). Inbreeding depression could manifest in a multitude of manners such as genetic defects and reduced fertility in breeding animals, which could incur significant economic loss in animal production, compromises the welfare of the animals, and in extreme cases causes extinction of the entire breed, population or species (Ryder and Wedemeyer, 1982; Schlie, 1967; Sevinga et al., 2004). For this reason, the inbreeding level should be managed when conducting a breeding program.

The need to constrain the level of inbreeding while maximizing the genetic gain from a selective breeding program has given rise to the optimal contribution selection (OCS) theory. First introduced by Wray and Goddard (1994) and further developed by Meuwissen (1997), this method seeks to optimize the contributions from each selection candidate to be propagated into the next generation by attempting to maximize the genetic gain to the next generation while constraining the increase in the inbreeding level so that the selection response could be maintained in the long term. The OCS method uses the estimated breeding values of the selection candidates and a matrix with relationships between them and various algorithms can be used to find an optimal solution (Clark et al., 2013). Compared with truncation selection and with the constraint of inbreeding rate $\Delta F = 0.01$, previous studies have suggested an improvement of genetic gain ranging from 16% to 81% per generation (Clark et al., 2013; Meuwissen, 1997; Nielsen et al., 2011).

While OCS allows the maximization of genetic gain given a constraint in the changes in inbreeding level in a selective breeding program, it uses estimates of breeding values, i.e., additive genetic effects, and the method would usually not optimize the non-additive genetic components of a trait, such as its dominance effect. While not heritable, these non-additive genetic components could significantly lift the mean genetic merit in a population. Dominance effects are typically exploited in crossbreeding where heterosis is observed when mating individuals of two different lines or populations, e.g., such as in crossbred maize (Brieger, 1950) and poultry (Goto and Nordskog, 1959). Attempts to utilize them in an OCS method remain lacking, however.

Until recently, utilizing dominance genetic effects in mating designs has been difficult as its effects within the population are difficult to estimate accurately. The dominance effects are mating-specific and are not easily replicable unless a large litter size is available (de Boer and Hoeschele, 1993). With the advent of genomic information however, a more direct prediction of level of genome-wide heterozygosity in the offspring has become feasible, which allows the optimization of dominance effects especially in polygenic traits. Using predicted heterozygosity could be a practical method to utilize dominance effects in OCS and optimized individual mating.

With this in mind, the aim of this study is to develop a framework for the optimal contribution selection that considers the dominance genetic component. This framework would maximize both additive and dominance genetic effects on the phenotypes in the next generation while constraining the increase in inbreeding level. The usability of genomic-based data on additive and dominance component was also tested in this study. It is anticipated that this algorithm could successfully maximize the additive and dominance genetic component, although the performance might depend on the genetic architecture, population size and quality of information, which was tested through simulation under varying parameters across multiple generations.

6.3. Definitions and Model Assumed by the Algorithm

In this study, given N individuals and M genetic markers, the phenotypes are based on the model by Duenk (2020):

$$\boldsymbol{y} = \boldsymbol{X_a}\boldsymbol{\alpha} + \boldsymbol{X_d}\boldsymbol{\delta} + \boldsymbol{e} \quad [1]$$

Where $\boldsymbol{y}$ is a vector of length N containing the phenotype of N individuals; $\boldsymbol{X_a}$ being a matrix of size $N \times M$ containing the scores of the M additive loci in the form of $\{0, 1, 2\}$, where the values denote the number of alleles with nonzero effect sizes possessed by an individual; $\boldsymbol{\alpha}$ being a vector of length M containing the additive effect sizes of each of the markers, $\boldsymbol{X_d}$ being a matrix of size $N \times M$ containing the scores for the dominance loci in the form of $\{0, 1, 0\}$, where the values denote whether the marker is in a heterozygous state; $\boldsymbol{\delta}$ being a vector of length M containing the dominance effect sizes of each of the markers, and $\boldsymbol{e}$ being a vector of length N containing the residual component of the phenotypes.

The additive genetic component of the phenotypes is defined as the genetic component that contributed additively to the phenotypes, be it additive from number of copies of non-null

alleles within a locus, or a multitude of loci that contributed additively to a trait (Falconer, 1989). It is represented by the $\boldsymbol{X_a\alpha}$ component in equation [1].

The dominance genetic component of the phenotypes is defined as the sum over loci of the dominance deviation at each locus (denoted as δ), which is defined as the deviation of the phenotypic values of the heterozygotes (f_1) from the expected mid-homozygote values (f_0 and f_2) (Falconer, 1989):

$$\delta = f_1 - \frac{f_2 - f_0}{2} \quad [2]$$

6.4. The Basics of Optimal Contribution Selection (OCS) and Genetic Algorithm

The aim of the optimal contribution selection (OCS) is to maximize the genetic gain in the next generation while constraining the increase in the inbreeding level. If the dominance genetic component is not being considered, the algorithm achieved said aim by finding the balance between the increase in the additive genetic gain and the weighted increase in the inbreeding level (Meuwissen, 1997; Clark et al., 2013). For the algorithm proposed in this study, it is assumed that only sires are subjected to selection, and all the dams would be utilized in a breeding program.

6.4.1. The Basics of Additive-only OCS

Given a number of candidate sires N_m, the objective function for an additive-only OCS (denoted as f_{OCS}) is defined as follows:

$$f_{OCS}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{b} - \lambda\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x} \quad [3]$$

Where $\boldsymbol{x}$ is a column vector of length N_m containing the proportion of contribution of each of the sires toward the next generation; $\boldsymbol{x}^T$ be the transpose of $\boldsymbol{x}$; $\boldsymbol{b}$ being a column vector of length N_m containing the true or estimated additive breeding values of each of the sires. $\boldsymbol{G}$ is a matrix of size $N_m \times N_m$ containing the pedigree or genomic based relationship matrix between each of the sires in the current generation and λ is a scalar value that serves as a penalty for an increment in the inbreeding. In effect the algorithm calculates the expected genetic gain (denoted by the $\boldsymbol{x}^T\boldsymbol{b}$ component in equation [3]) and increase in inbreeding level scaled by λ (denoted by $\lambda\boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x}$ component in equation [3]) from a combination of sires to be

used in the propagation of the next generation (Meuwissen, 1997). While the ideal source of information for $\boldsymbol{b}$ is the true breeding values or QTL effect sizes, in practice such information is not attainable, thus the estimates $\widehat{\boldsymbol{b}}$ would be used.

The maximizing of genetic gains under constrained increase in inbreeding level can be achieved through genetic algorithm (Srinivas and Patnaik, 1994). Inspired by genetic processes observed in selection, genetic algorithms are iterative algorithms aimed at finding an optimal solution among the pool of candidate solutions by maximizing the objective function associated with the solution. Unlike other optimization algorithms such as gradient descent, this method is capable of finding optimal solutions in a large multimodal landscape (Srinivas and Patnaik, 1994; Taherdangkoo et al., 2012). Using [3] as example, the genetic algorithm attempts to find $\boldsymbol{x}$ that would maximize $f_{OCS}(\boldsymbol{x})$ (Clark et al., 2013).

The simplest genetic algorithm starts by randomly initializing a population of candidate solutions. Using example in [3], each member of the population is a vector of proportion of contribution from each sires $\boldsymbol{x}$. The objective function $f_{OCS}(\boldsymbol{x})$ for each of the $\boldsymbol{x}$s in the population is evaluated, and from the pool of $\boldsymbol{x}$s, those with highest values of $f_{OCS}(\boldsymbol{x})$s were selected to be propagated into the next generation. For the next generation population, the $\boldsymbol{x}$s were subjected to various genetic operators such as mutation, where parts of the solution were substituted with new values, crossover where parts of the solutions were exchanged, and inversion where the sequence is inverted. The resulting population would have new combinations of sires' contributions. This process was iterated until convergence, when the operators no longer producing a more optimal solution (Srinivas and Patnaik, 1994; Taherdangkoo et al., 2012).

6.4.2. Modifications Needed for Additive-Dominance OCS

For OCS involving dominance genetic components, given the number of sires N_m and number of dams N_f, a generalized model for the OCS could be defined as follows:

$$f_{OCS}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{b} + \lambda_d \sum_{i=1}^{N_f} \boldsymbol{d_i} - \lambda_b \boldsymbol{x}^T\boldsymbol{G}\boldsymbol{x} \qquad [4]$$

Where the $\boldsymbol{d_i}$ is the ith element of the vector $\boldsymbol{d}$, which defines the dominance scores for each of the sires when aired with the ith dam. The dominance scores can be calculated from true or estimated dominance effects depending on the sources of information, and λ_d and λ_b are

weights for the dominance term and a penalty on inbreeding, respectively. The calculation of $\boldsymbol{d}$ will be detailed in Section 6.5.6.

While theoretically optimizing the additive and dominance components while constraining the increase in the level of inbreeding can be achieved by optimizing [4] using genetic algorithm, in practice optimizing the $f_{OCS}(\boldsymbol{x})$ in equation [4] directly would not yield the optimal solution. One such reasons is the competition between the additive and dominance genetic components, which causes the optimization to favour the component with larger variance at the expense of the component with smaller variance. Another more practical aspect in the optimization of all three component simultaneously is the computational intensity and time required for the convergence of solution. With the increased number of components that need to be optimized, this increases the sample space that need to be tested, which increases the chance of convergence toward a local optimum. Additionally, unlike additive and inbreeding calculation, in which the exact ordering of the sires does not matter, the dominance genetic components depend on the exact permutation of the sires, which significantly increases the solution space. The search space and the number of solutions is detailed provided in Appendix E. To worsen the situation, the exact sire permutation is not expressible through vector $\boldsymbol{x}$, thus precluding its use in the optimization of dominance genetic component.

Due to these challenges, modification of the original algorithm might be required. One applicable modification is by separating the optimization of the additive from the dominance genetic component. This setup would have optimization on the additive genetic component and inbreeding level in the first phase, and the results from this phase would become the input for the second phase where the dominance genetic component was optimized. In this setup, the objective function for the first phase (denoted as $f_{OCS_{AI}}$) was as defined in equation [3], and the second phase (denoted as f_{OCS_D}) was as follows:

$$f_{OCS_D}(\boldsymbol{s}) = \sum_{i=1}^{N_f} \boldsymbol{d_i} \qquad [5]$$

Where $\boldsymbol{s}$ is a vector of length N_f termed the "sire index vector", and is defined such that the i-th element of this vector contains which sire that would be mated with dam i. This formatting is needed due to the fact that the dominance component varies with sire-dam mating configurations. For this reason, the exact configurations of sire-dam mating need to be

specified, hence the use of vector $\boldsymbol{s}$. The aforementioned vector $\boldsymbol{x}$ would in turn be referred as "sire proportion vector." A sire index vector can be translated into its corresponding sire proportion vector using the following method:

$$\boldsymbol{x_i} = \frac{\#\{\boldsymbol{s}: \boldsymbol{s} = i\}}{N_f} \quad [6]$$

where $\boldsymbol{x_i}$ is the i-th elements of vector $\boldsymbol{x}$, and $\#\{\boldsymbol{s}: \boldsymbol{s} = i\}$ is defined as "the number of occurrence of value i in vector $\boldsymbol{s}$. Equation [6] can also be expressed in the following pseudocode (as the function "`s_to_x`"):

```
## converting sire index vector to sire proportion vector
## INPUT: s, nmal
### s: sire index vector, containing the indices of which sires to be mated with a dam
### nmal: number of males

def s_to_x(s, nmal):
	nfem = length(s)
	x = numeric(length = nmal)
	for nm in range(nmal):
		x[nm] = length(which(s == nm)) / nmal
		## assuming the sires are indexed 1 to nmal
	return x
```

This separation of optimization phases bypasses the competition between the additive and dominance genetic components. To ensure a balance between the additive genetic and the dominance effects, the $f_{OCS_{AI}}$ and f_{OCS_D} were combined in the third phase, yielding the final objective function as defined in equation [4].

It should be noted that such cascading genetic algorithm design is not possible for all problems, as in some problems the optimization at the second phase would disrupt the already optimal solutions in the first phase, losing the progress from the optimization in the first phase. This setup is possible for this problem however as some genetic operators such as vertical recombination and horizontal inversion only affect f_{OCS_D}; they have no effects on $f_{OCS_{AI}}$. This is due to the fact that f_{OCS_D} depends on both proportion and permutation of sires, whereas $f_{OCS_{AI}}$ only depends on the proportion of sires. Thus, any genetic operators that only affect the sire permutation would only affect f_{OCS_D}, with no effects on $f_{OCS_{AI}}$. Therefore, by applying these genetic operators at the second phase of the algorithm, the f_{OCS_D} can be optimized while preserving the $f_{OCS_{AI}}$ from the first phase, allowing the use of cascading genetic algorithm for this problem.

One caveat for cascading genetic algorithm is the constriction of diversity in the solution pool from the first phase optimization, which could be mitigated through parallelization of the first phase optimization. This involves running the first phase multitude of times, rather than once, and the pool of solutions collected from the several first phase optimizations would become the input for the second phase optimization. By increasing the number of reruns for the first phase of optimization, the diversity in the input for the second phase would be increased, thus improving the chance of finding the global optimum for this phase while reducing the chance of missing out sires that benefit the second phase optimization (Baluja and Caruana, 1995). For this reason, parallelization has been used in this algorithm.

Due to its stochasticity of the genetic operators, the genetic algorithm also tends to disrupt an already optimal solution during the process of optimization, which causes the algorithm to miss out the global optimum. For this reason, an elitist genetic algorithm strategy can be employed. Rather than having the entirety of the population replaced by the mutant and recombinants, an elitist genetic algorithm ensures the top solutions be propagated into the next generation unaltered. This way, if the top solutions are indeed the global optimum, their survival into the subsequent generations was guaranteed as any mutants and recombinants would not have a fitness score greater than them, thus preserving the potential global optimum (Baluja and Caruana, 1995). For this reason, elitism was employed for this algorithm.

6.4.3. Genetic Operators and Their Hyperparameters

For this genetic algorithm in this OCS, five genetic operators were employed: (1) mutation; (2) horizontal recombination; (3) vertical recombination; (4) horizontal inversion and; (5) vertical inversion. These genetic operators were employed on a population of sire index vectors, which would be compiled into a sire index array (denoted as $\boldsymbol{S}$). This array is of size $N_s \times N_f$, where N_s is the number of sire index vectors passed into the genetic algorithm.

In mutation some of the sires in $\boldsymbol{s}$s are replaced with new sires. In horizontal recombination, part of the rows in $\boldsymbol{S}$ are exchanged, whereas in vertical recombination, part of the columns in $\boldsymbol{S}$ are exchanged. For horizontal inversion, part of the sequence within a row of $\boldsymbol{S}$ is inverted, and for vertical recombination, part of the sequence within a column of $\boldsymbol{S}$ is inverted. The hyperparameters associated with these genetic operators is provided in Table 6.1.

Table 6.1: Hyperparameters used for the genetic operators and number of solutions

Genetic Operators		Phase 1 Values	Phase 2 Values
Mutation			
	Mutation Rate (p_m)	0.05	-
Horizontal			
Recombination	Recombination Rate (p_{hcr})	0.3	-
	Average Recombination Block Size (p_{hcr})	0.1	-
Vertical			
Recombination	Recombination Rate (p_{vcr})	0.3	0.4
	Average Recombination Block Size (p_{vcb})	0.1	0.1
Horizontal			
Inversion	Inversion Rate (p_{hir})	0.3	0.4
	Average Inversion Block Size (p_{hib})	0.1	0.1
Vertical			
Inversion	Inversion Rate (p_{vir})	0.3	-
	Average Inversion Block Size (p_{vib})	0.1	-
Number of Solutions			
Number of	Number of Solution Tested per Iteration (N_s)	1500	3000
Solutions	Number of Solution Selected per Iteration (N_{top})	2	2
	Number of Parallelization Threads (N_{par})	8	8

For the first phase optimization, all five genetic operators were utilized, whereas for the second phase only vertical recombination and horizontal inversion were utilized. The hyperparameters associated with these operators are provided in Table 6.1. For each of these hyperparameter values, an adaptive scaling factor was employed, which is calculated based on the population performance of the solutions. This scaling factor was multiplied into the hyperparameter values within each iteration of optimization. This adaptive scaling factor is based on Srinivas and Patnaik (1994), and is defined as follows:

$$scaling\ factor = \frac{f_{max} - f_{mean}}{f_{max} - f_{min}} \quad [6]$$

Where f_{max} is the maximum f_{OCS} score within the population; f_{mean} is the average value of all f_{OCS} scores within the population, and f_{min} being the minimum f_{OCS} score within the

population. This scaling factor is to preserve the already highly optimal solution from being disrupted by the operators as the population performance converges (Srinivas and Patnaik, 1994). The more optimized the population is, the smaller the range of $f_{max} - f_{mean}$, and a smaller scaling factor being imposed onto the hyperparameters, thus reducing the likelihood of disrupting an optimal solution.

Besides the genetic operators, the number of solutions tested per iteration, as well as the number of top solutions chosen per iterations, is provided in Table 6.1.

6.5. Layout of the Algorithm

For the purpose of this study, the algorithm requires several inputs: sire genotype array of size $N_m \times M$ in {0, 1, 2} format (denoted as $\boldsymbol{Z_m}$), where M is the number of markers; dam genotype array of the same format of size $N_f \times M$ (denoted as $\boldsymbol{Z_f}$); as well as information on the additive genetic components, which could be genomic (as an example, solutions for marker effects or test statistics from GWAS) or animal-based (as example, estimated breeding values (EBVs) of the animals). For optimization of the dominance genetic component, only genomic information can be used as it depends on the parental genotypic configuration (de Boer and Hoeschele, 1993). Therefor the genotype arrays are needed for the optimization of the dominance genetic component.

A targeted average coancestry among the selection candidates ΔI_t would need to be specified. This is the acceptable increment of inbreeding level that the algorithm needs to constrain at. In our study, the selection was conducted on sires only, and all the dams were used in the selective breeding.

6.5.1. The Building of Sire's Relationship Matrix (NRM or GRM)

This algorithm starts by building the relationship matrix among the sires, which was used in the calculation of expected ΔI. This algorithm could utilize either a Numerator Relationship Matrix (NRM) based on pedigree, or a Genomic Relationship Matrix (GRM) based on genomic data. The latter will be used in this study.

The sire GRM (denoted as $\boldsymbol{G_{GRM}}$) was built using the first method as proposed by VanRaden (2008). While the GRM is advantageous compared to the NRM in that it could reveal information on the relationship of an individual that has appeared to have no pedigree

relationship with other individuals within the population, thus improving the accuracy of predicting the coefficient of consanguinity with said individual (Gondro, 2015), it has several less desirable properties. One notable property is that, with the assumption of sires being non-inbred and without consanguinity with each other, the column-wise expected values of GRM is zero, which produces discrepancies if GRM is utilized in place of the NRM (a simplified example along with mathematical explanation is provided in Appendix F).

To alleviate these problems, some adjustments could be made onto the GRM. One applicable adjustment was the addition of a constant to the off-diagonal value of the GRM. As observed in the column-wise averages of an NRM, the constant should shift the column-wise average toward $\frac{1}{N_m}$ for unrelated, non-inbred sires. With some algebraic manipulation, provide the number of sires is large ($N_m \geq 50$), the constant that should be added to the off-diagonals of GRM could be approximated to be $\frac{1}{N_m - 1}$ (a mathematical derivation of this constant, as well as a more precise but elaborated expression, is provided in Appendix F). With this constant, the adjusted GRM (denoted as $\boldsymbol{G}^*_{\boldsymbol{GRM}}$) that shall be employed for the calculation of ΔI can be defined as follows:

$$\boldsymbol{G}^*_{\boldsymbol{GRM}} = \boldsymbol{G}_{\boldsymbol{GRM}} + \frac{1}{N_m - 1}\left(\mathbf{1}_{N_m} - \boldsymbol{I}_{N_m}\right) \quad [7]$$

where $\mathbf{1}_{N_m}$ is a matrix of ones of size $N_m \times N_m$, and $\boldsymbol{I}_m$ is an identity matrix of size $N_m \times N_m$. The $\boldsymbol{G}^*_{\boldsymbol{GRM}}$ would provide additional weight to the ΔI based on the number of sires available for selection. For these reasons, $\boldsymbol{G}^*_{\boldsymbol{GRM}}$ was used in place of $\boldsymbol{G}_{\boldsymbol{GRM}}$ in the calculation of ΔI for this algorithm.

6.5.2. Calculation of Additive Genetic Values

The additive genetic values can be predicted using various sources of information such as EBVs of each individual, or from genomic information. The methods and calculations used for these predictions are detailed below.

6.5.2.1. Estimated Breeding Values (EBVs)

One commonly described method of obtaining additive genetic values for an OCS is the sires' EBVs (Kinghorn, 2000). For this study the EBVs (denoted as $\widehat{\boldsymbol{b}}_{\boldsymbol{m}}$) was calculated using the methods detailed in Gondro (2015) using the aforementioned sires' GRM (denoted as $\boldsymbol{G}_{\boldsymbol{GRM}_{\boldsymbol{m}}}$):

$$\widehat{\boldsymbol{b}}_{\boldsymbol{m}} = \left[\boldsymbol{I}_{N_m} + \left(\frac{1-h^2}{h^2}\right) * \boldsymbol{G}_{\boldsymbol{GRM}_{\boldsymbol{m}}}^{-\mathbf{1}}\right]^{-\mathbf{1}} * \boldsymbol{y}_{\boldsymbol{m}} \quad [8]$$

where $\boldsymbol{y}_{\boldsymbol{m}}$ is a column vector of length N_m containing the mean-centred sires' phenotypes. Besides GRMs, NRMs can also be used in equation [8] for prediction of EBVs.

6.5.2.2. Marker-based Information

Besides the EBVs, marker-based information from a Genome-Wide Association Study (GWAS) such as the marker test statistics can also be used as the additive genetic component of the OCS. For this example, the t-test statistics of the markers would be used.

Given a column vector of t-test statistics for each of the M markers $\boldsymbol{\Theta} = [t_1, t_2, t_3 \ldots t_M]^T$, the test statistics based additive score of the sires (denoted as $\hat{\boldsymbol{t}}_{\boldsymbol{m}}$) was calculated as follows:

$$\hat{\boldsymbol{t}}_{\boldsymbol{m}} = \boldsymbol{Z}_{\boldsymbol{m}}\boldsymbol{\Theta} \quad [9]$$

The $\hat{\boldsymbol{t}}_{\boldsymbol{m}}$ can then be used in place of $\widehat{\boldsymbol{b}}_{\boldsymbol{m}}$, and the remaining calculation remained unchanged.

6.5.3. Calculation of Dominance Genetic Values

The dominance score between a sire and a dam is defined as the sum of dominance effects over all loci. As in the additive genetic component, predicting the dominance scores in offspring can also be done using various sources of information. While knowledge of true dominance effect sizes at the QTL would be ideal, estimation of these effects is difficult unless there is data on large full sib families. With the availability of genomic data, genome-wide heterozygosity can now be used as a more practical way to predict dominance effects in future phenotypes and in this study, we will test this alternative approach. We will compare this approach with using known dominance effects to measure its effectiveness.

Unlike the additive genetic values, which can be represented using vectors of parental values, the dominance genetic values depend on the sire-dam pairings. For this reason, a dominance score array of size $N_m \times N_f$ is needed.

6.5.3.1. Dominance Effect Sizes

The dominance effect sizes could be used to construct the score array for the dominance genetic component. Using the dominance effect size of a marker of locus k, denoted as δ_k, the dominance score for the offspring of each combination of sires of index i and dams of index j (denoted as $d_{i,j}$) could thus be calculated as follows:

$$d_{i,j} = \frac{1}{M}\sum_{k=1}^{M} u_{i_k j_k}\delta_k \quad [10]$$

Where $u_{i_k j_k}$ is a scalar value in a 3×3 incidence matrix $\boldsymbol{U}$ describing the likelihood of producing heterozygous offspring for each genotype combinations at locus k between sire i and dam j. The matrix $\boldsymbol{U}$ is defined as follows:

$$\boldsymbol{U} = \begin{bmatrix} 0 & 0.5 & 1 \\ 0.5 & 0.5 & 0.5 \\ 1 & 0.5 & 0 \end{bmatrix} \quad [11]$$

and is indexed at first, second and third rows for sire with genotypic state of {0,1,2} respectively, and first, second and third column for dam with genotypic state of {0,1,2} respectively. As an example, if a sire has a genotypic state of 2 and the dam has a 0 state, the $u_{i_k j_k}$ would be the third row, first column of $\boldsymbol{U}$, and $u_{i_k j_k} = 1$.

The $d_{i,j}$ can then be stored in dominance score array of size $N_m \times N_f$ denoted as $\boldsymbol{D}$, which was defined as follows:

$$\boldsymbol{D} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,N_f} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,N_f} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N_m,1} & d_{N_m,2} & \cdots & d_{N_m,N_f} \end{bmatrix} \quad [12]$$

6.5.3.2. Heterozygosity

Due to the difficulty of estimation of effect sizes, obtaining the idealized input of true dominance effect sizes is not a trivial problem. For this reason, an estimate for the dominance genetic component would be desirable, with one such estimates being the expected heterozygosity of the offspring.

Given a sire of index i and dam of index j, the heterozygosity score (denoted as $h_{i,j}$) was defined as follows:

$$h_{i,j} = \sum_{k=1}^{M} u_{i_k j_k} \quad [13]$$

The $h_{i,j}$ can then be stored in a heterozygosity score array of size $N_m \times N_f$ (denoted as $\boldsymbol{H}$), which is defined as in $\boldsymbol{D}$ from equation [12], but with $h_{i,j}$ in place of $d_{i,j}$.

One major advantage of using $\boldsymbol{H}$ compared to $\boldsymbol{D}$ is that $\boldsymbol{H}$ can be used without knowing the dominance effect size of the QTL. This is advantageous in cases when accurate estimates for dominance effect sizes or dominance genetic variance are unavailable. For this study it is assumed that dominance variance is unavailable, and the heterozygosity is the only practical proxy for the optimization of dominance component.

One downside for using heterozygosity is that the realized benefit of optimization depends on dominance effects δ_k; if there exist some negative δ_k, maximizing the heterozygosity would not yield a maximal dominance effect. This would be the case for underdominance where heterozygotes have lower phenotypic value than both homozygotes (Magori and Gould, 2006; Nehrenberg et al., 2010; Newberry et al., 2016). Thus, if underdominance is present, maximizing heterozygosity would not maximize the dominance effects. For this phenotypic model, it is assumed that underdominance is uncommon and has negligible impact on the optimization. Further studies on improving the estimates of the marker-wise dominance effect sizes is desirable.

6.5.4. Initialization of Solution Pool

For the initialization of the solution, a number of N_f sires were randomly chosen with replacement, which their indices were kept as sire index vector $\boldsymbol{s}$. To initialize a pool of solutions, an N_s number of vectors $\boldsymbol{s}$s was initialized, which was compiled into sire index array $\boldsymbol{S}$. This solution pool was used in the genetic algorithm.

6.5.5. Phase 1: Optimization of Additive and Inbreeding Scores

The aim of this phase is to maximize the additive genetic of the solutions while constraining the changes of inbreeding coefficient within the ΔI_t. This involves the calculation of $f_{OCS_{AI}}$ of each of the rows of sire index array $\boldsymbol{S}$. A pseudocode for this phase of genetic algorithm is provided in Appendix G.

This phase starts by converting the sire index array $\boldsymbol{S}$ into sire proportion array $\boldsymbol{X}$. This is done by converting each rows of $\boldsymbol{S}$ into its corresponding sire proportion vectors using [6], before compiling the N_s sire proportion vectors into the sire proportion array that would have a size of $N_s \times N_m$.

If the sires' EBVs are used in the calculation, the additive score of each row of $\boldsymbol{X}$ (denoted as $\boldsymbol{a_s}$) is a vector of length N_s defined as follows:

$$\boldsymbol{a_s} = \boldsymbol{X}\widehat{\boldsymbol{b}}_{\boldsymbol{m}} \quad [15]$$

Whereas if test statistics based additive scores had been used, the $\hat{\boldsymbol{t}}_{\boldsymbol{m}}$ was used in place of $\widehat{\boldsymbol{b}}_{\boldsymbol{m}}$ for equation [15], and the subsequent calculations would remain unchanged.

Using the matrix $\boldsymbol{X}$, the expected changes in average coancestry of the selection candidates for each row of matrix $\boldsymbol{X}$ (denoted as $\Delta\boldsymbol{i_s}$) is a vector of length N_s defined as follows:

$$\Delta\boldsymbol{i_s} = diag(\boldsymbol{X}\boldsymbol{G}^*_{\boldsymbol{GRM}}\boldsymbol{X}^{\boldsymbol{T}}) \quad [16]$$

Finally, the vector of size N_s containing the $f_{OCS_{AI}}(\boldsymbol{x})$s of each of the rows of $\boldsymbol{X}$ (denoted as $\boldsymbol{f_{OCS_{AI}}}$) is defined as the weighted sum of $\boldsymbol{a_s}$ and $\Delta\boldsymbol{i_s}$:

$$\boldsymbol{f_{OCS_{AI}}} = \boldsymbol{a_s} + \lambda\Delta\boldsymbol{i_s} \quad [17]$$

With the λ being the penalty factor the inbreeding coefficient, which was used to penalize solutions with overly high inbreeding coefficient. The initial λ was set at zero, and can range from negative infinity (although realistically a large but finite negative number), where the model would attempt to constrain the increment in inbreeding level while disregarding its impact on the additive genetic gains, up to zero, where the additive genetic gains was maximized while ignoring its impacts on the increment of level of inbreeding. To consider a potentially wide range of values of $\boldsymbol{a_s}$ and $\Delta\boldsymbol{i_s}$, an adaptive approach was utilized to calculate λ; for each iteration, the value of λ was updated based on the inbreeding coefficients of the solutions.

Let λ_i be the λ value at iteration i, with ΔI_{avg} being the average of vector $\Delta\boldsymbol{i_s}$ (i.e. $\Delta I_{avg} = average\Big(diag(\boldsymbol{X}\boldsymbol{G}^*_{\boldsymbol{GRM}}\boldsymbol{X}^{\boldsymbol{T}})\Big)$), and ΔI_t being the user-specified targeted average coancestry of the selection candidates. For each iteration, the updated value of λ (denoted as λ_{i+1}) was calculated as follows:

$$\lambda_{i+1} = \lambda_i + z_\lambda * \left(\Delta I_{avg} - \Delta I_t\right) \quad [18]$$

Where z_λ is a scaling factor that translates the discrepancy between average inbreeding coefficient from the targeted inbreeding coefficient to the penalty weight λ. For this study, the z_λ is set at 100.0 per 1 unit of $\Delta I_{avg} - \Delta I_t$ discrepancy. Using this updating system, if the

average ΔI in the solution pool is overly high (i.e. $\Delta I_{avg} > \Delta I_t$) additional weightage was added onto λ, increasing the penalty of inbreeding. Conversely if the average ΔI is overly low, the λ was reduced in attempt to exploit additional genetic gains by allowing solutions with greater inbreeding coefficient. This produces a self-stabilizing system; with sufficiently large number of iterations, the ΔI_{avg} would eventually converge toward ΔI_t, achieving the targeted average coancestry (i.e. $\Delta I_{avg} = \Delta I_t$).

From the vector $\boldsymbol{f_{OCS_{AI}}}$, indices for the maximal values in this vector was recorded. These indices were used to subset the sire index array $\boldsymbol{S}$, and the resulting matrix of size $N_{top} \times N_f$, denoted as $\boldsymbol{S}_{top}$, would become the seed for propagation into the next iteration. The rows of $\boldsymbol{S}_{top}$ were sampled $N_s - N_{top}$ times, and the resulting sampled matrix, denoted as $\boldsymbol{S}_{off}$, was subjected to the Phase 1 genetic operators. The altered $\boldsymbol{S}_{off}$ was then combined with unaltered $\boldsymbol{S}_{top}$ to produce a new sire index array, which was used in the next iteration. The process in Phase 1 was repeated until convergence of $f_{OCS_{AI}}$.

In this algorithm, the point of convergence of $f_{OCS_{AI}}$ in the solution array is defined as the point where the $\boldsymbol{a_s}$ no longer significantly improves with each iteration, and $\Delta \boldsymbol{i_s}$ has largely stabilized at ΔI_t (or, if such ΔI_t is impossible to achieve, no longer fluctuates with each iteration). This is determined using the slope of the tangent of the additive scores and inbreeding coefficient from the last 50 iterations, and is considered to have converged if $f_{OCS_{AI}}(\boldsymbol{x})$ reaches a slope of less than 5×10^{-3} across the last 50 iterations. The end result of this convergence was the $\boldsymbol{S}_{top}$ from the last iteration of the genetic algorithm. For each of the convergence event, the λ from the last iteration (denoted as λ_{fin}) was also recorded.

To preserve the diversity in the solution array for Phase 2 of the algorithm, the Phase 1 genetic algorithm was parallelized N_{par} number of times, with the $\boldsymbol{S}_{top}$ from each repeat being collated into a new 2 dim-array of size $(N_{par} * N_{top}) \times N_f$ which was denoted as $\boldsymbol{S}_{allP1}$. The average of λ_{fin}s from all the repeats (denoted as λ_{avg}) was also calculated, to be used in Phase 3 of the algorithm. The $\boldsymbol{S}_{allP1}$ would become the input for the Phase 2 of the algorithm.

6.5.6. Phase 2: Optimization of Dominance Scores

The aim of this phase is to optimize the dominance genetic component from the solution pool $\boldsymbol{S}_{allP1}$. This involves the calculation of f_{OCS_D} of each of the rows in $\boldsymbol{S}_{allP1}$. A pseudocode for this phase of genetic algorithm is provided in Appendix G.

This phase starts by generating an offspring solution array sampled from $\boldsymbol{S}_{allP1}$. In total $N_s - \left(N_{par} * N_{top}\right)$ new solutions were sampled from $\boldsymbol{S}_{allP1}$. This offspring solution array was subjected to Phase 2 genetic operators, with the hyperparameters for these genetic operations being provided in Table 6.1. This altered offspring array was combined with $\boldsymbol{S}_{allP1}$, producing a new solution array $\boldsymbol{S}$. This array would become the starting solutions for Phase 2 of the algorithm.

Let $\boldsymbol{s}$ be a vector representing a row in $\boldsymbol{S}$ that contains the indices of sires chosen to be mated with the dam, with i^{th} index in vector $\boldsymbol{s}$ denoted as s_i. Using the dominance genetic effect score array (in this example, array $\boldsymbol{D}$), the dominance score for vector $\boldsymbol{s}$ (denoted as d_s) was calculated as follows:

$$d_s = \sum_{i=1}^{N_f} \boldsymbol{D}_{s_i,i} \qquad [19]$$

With the notation $\boldsymbol{D}_{x,y}$ being the xth row and yth column of array $\boldsymbol{D}$. In this study, the d_s serves as the objective function term f_{OCS_D} as defined in equation [5]. This operation was repeated for all the $\boldsymbol{s}$ in $\boldsymbol{S}$, from which a vector of length N_s denoted as $\boldsymbol{d_s}$ containing the dominance scores of each of the rows in $\boldsymbol{S}$. If the dominance score array is built using heterozygosity, array $\boldsymbol{H}$ can be used in place of $\boldsymbol{D}$ in equation [19], and subsequent calculations remained unchanged.

From the vector $\boldsymbol{d_s}$, the indices of maximal values in this vector were chosen, from which the rows of $\boldsymbol{S}$ were sliced, and from which the $\boldsymbol{S}_{top}$ of size $N_{top} \times N_f$ was generated. The $\boldsymbol{S}_{top}$ was sampled $N_s - N_{top}$ times, and the resulting matrix, denoted as $\boldsymbol{S}_{off}$, was subjected to the vertical recombination and horizontal inversion. As in Phase 1, the rate of recombination and inversion was also scaled with the scaling factor as defined in equation [6]. The altered $\boldsymbol{S}_{off}$ was then recompiled with $\boldsymbol{S}_{top}$, and the resulting $\boldsymbol{S}$ array would become the input for the next iteration of the genetic algorithm.

The point of convergence in Phase 2 was defined as the point where the maximal values of $\boldsymbol{d_s}$ no longer substantially improve with each iteration. This refers to the point where the slope of the tangent of the curve of $\boldsymbol{d_s}$ across the last 200 iterations is less than 1×10^{-5}. At the point of convergence, the most optimal solution, denoted as $\boldsymbol{s}_{optP2}$, was generated.

As the dominance genetic component has a significantly larger sample space than both additive and inbreeding coefficient, there is an increased probability of convergence toward local optima. For this reason, this phase was repeated N_{par} times, using the same $\boldsymbol{S}_{allP1}$ as the input. The $\boldsymbol{s}_{optP2}$ from each repeat was kept as $\boldsymbol{S}_{allP2}$, which is a 2-dimensional array of size $N_{par} \times N_f$. The $\boldsymbol{S}_{allP2}$ was used for the third phase of the algorithm.

6.5.7. Phase 3: Combining the Dominance Scores to Additive and Inbreeding Scores

The aim of this phase is to identify the optimal solution from $\boldsymbol{S}_{allP2}$ by combining the dominance scores from the second phase into the additive and inbreeding scores from the first phase. Given the additive scores, dominance scores and inbreeding coefficients of each row in $\boldsymbol{S}_{allP2}$, denoted as $\boldsymbol{a_s}$, $\boldsymbol{d_s}$ and $\Delta\boldsymbol{i_s}$ respectively, the objective function for this phase (denoted as $f_{OCS_{ADI}}$) was defined as follows:

$$f_{OCS_{ADI}} = \boldsymbol{a_s} + \lambda_d \boldsymbol{d_s} - \lambda_{avg}(\Delta \boldsymbol{i_s} - \Delta I_t) \quad [20]$$

where λ_{avg} is the average of all penalties of inbreeding coefficients at the point of convergence across the repeats in Phase 1, and λ_d being the weight for the dominance genetic effects, which for this study was set to $\lambda_d = 1$.

This operation could then be applied to all $\boldsymbol{s}$ in $\boldsymbol{S}_{allP2}$, generating a vector of length N_{par} containing the $f_{OCS_{ADI}}$ of all the $\boldsymbol{s}$s, which can be denoted as $\boldsymbol{f_{OCS_{ADI}}}$. The optimal solution for this phase, denoted as $\boldsymbol{s_{opt}}$, is defined as the $\boldsymbol{s}$ that produces the maximal $f_{OCS_{ADI}}$ value in $\boldsymbol{f_{OCS_{ADI}}}$. The $\boldsymbol{s_{opt}}$ would also be regarded as output for the OCS algorithm and could then be applied to the selective breeding program.

6.5.8. The Difficulty of Optimization Across Multiple Generation

The aforementioned algorithm optimizes the breeding pairs across one generation. For optimization across multiple generations, the most straightforward method is to optimize the breeding pair for the current generation and feed the algorithm with the predicted offspring genotypes. The feasibility of direct optimization of breeding pairs across multiple generations is limited as it requires the knowledge of the exact genotype of the population that need to be selected, which due to the randomness from Mendelian sampling, would not be available until the production of offspring.

This is further complicated by the recombination process, which alter the allelic composition of a haplotype. While recombination would not alter the expected values for the additive, dominance and inbreeding scores, as there are no net changes in the expected allelic composition and heterozygosity between the four possible combinations of sire-dam haplotypes. Recombination does alter these quantities within each of the haplotypic combinations, and these quantities would be relevant to the allelic composition and heterozygosity as the parent transmitted the haplotypes to the offspring. For these reasons, without additional knowledge on the exact genotypes of the offspring, it might not be feasible to predict the exact permutation of sires required for optimal mating in the subsequent generation.

6.6. Testing the Algorithm

6.6.1. Layout of the Experiment

The OCS algorithm was tested using simulation, which was conducted using Python (version 3.9.7, released 30 August 2021). This OCS was tested against other methods of genomic selection. The merits of this OCS were also tested under varying sample sizes, genetic architecture as well as sources of information. The inputs for this experiment were the sire and dam genotype arrays, sire and dam phenotype vectors, narrow sense heritability (assumed to have been known) and a targeted coancestry level between selection candidates. The experiment was conducted on a PC with the following specification: 8-core Intel i7-8665U at 1.90 GHz with 16 GB RAM, with all 8 cores being used in the OCS.

This simulation starts by generating the genotype arrays for the base population. For this simulation, a pair of sire and dam genotype arrays, $\boldsymbol{Z_m}$ and $\boldsymbol{Z_f}$ respectively, were generated. Equal numbers of sires and dams were simulated for the base population. 20k markers were generated for both genotype arrays, with the number of sires and dams being provided in Table 6.2. Correlation between markers was simulated however, by copying part of the genotypes from a marker into the adjacent markers, with the pairwise marker correlation being set at 0.9. The allele frequencies for each marker were assumed to follow a symmetric Beta distribution with shape parameter set at 0.5.

Some of the markers were nominated as a QTL, and they were allocated an effect size. The additive effect sizes were randomly generated using a gamma distribution, with the distribution parameter tested provided in Table 6.2. In total 500 markers were designated to be additive QTL. These effect sizes were padded with zeros for null markers, and the vector for additive effect sizes of all markers, $\boldsymbol{\alpha}$, was used to calculate the sires' true breeding values (TBVs) as follows:

$$\boldsymbol{b} = \boldsymbol{X_m}\boldsymbol{\alpha} \qquad [21]$$

Besides additive effect sizes, these QTL were allocated a dominance effect sizes (denoted as δ_k) that would contribute toward the phenotype of individuals heterozygous for that particular locus genotype. For the dominance genetic component, the dominance effect size was simulated using a half-normal distribution, with the distribution tested provided in Table 6.2. In total 500 dominance QTL were simulated, with the collection of dominance effect sizes denoted as $\boldsymbol{\delta}$. The δ_k was used to calculate the matrix $\boldsymbol{D}$ as defined in equation [12]. The heterozygosity-based score array $\boldsymbol{H}$ for the dominance genetic component can also be calculated using the sire and dam genotype arrays as defined in equation [12] and [13].

Using the same $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$, the phenotype of the sires and dams are also calculated in accordance with equation [1]. For the simulation of the phenotype, the narrow sense heritability for the additive genetic component was set at 0.3. The phenotype was in turn used to estimate the EBVs of the sires using equation [8], with the resulting EBVs denoted as $\widehat{\boldsymbol{b}}$. For the test statistics $\boldsymbol{\Theta}$, a single SNP regression of the phenotype on individual marker genotype was conducted. All sires and dams are utilized in the single SNP regression. The $\boldsymbol{\Theta}$ was then used to build a $\hat{\boldsymbol{t}}_{\boldsymbol{m}}$ as defined in equation [9]. Markers with minor allele frequencies of less than 0.05 are filtered out. No threshold was applied for the test statistics.

The genotype arrays and additive and dominance scores was used in the OCS algorithm, with the targeted coancestry level between selection candidates set at $\Delta I_t = 0.01$. This algorithm produced the vector $\boldsymbol{s_{opt}}$ that contains the permutation of mating between sire and dams that would optimize the additive and dominance genetic component while constraining the increase in inbreeding level. The $\boldsymbol{s_{opt}}$ was then used in the selection of sires for the next generation. The selection intensity of the sires was determined by the constraint set by ΔI_t. No selection was conducted on the dam side.

The OCS was used in a 4-generation genomic selection program. The generations were assumed to be non-overlapping. The OCS was applied on per generational basis, with the $\boldsymbol{s_{opt}}$ from each generation being used to generate the offspring genotype arrays. Each dam produces two offspring per generation interval, with an equal number of males and females, and the total number of animals remained unchanged between generations. From each generation, the additive scores (a_{opt}), dominance scores (d_{opt}) and the increase in inbreeding coefficients (Δi_{opt}) of the $\boldsymbol{s_{opt}}$ was recorded. Besides these, the total genetic merits (t_{opt}), defined as the sums of a_{opt} and d_{opt}, was calculated as well. To compare the performance of the selective breeding, the a, d and t of the base population (denoted as a_{base}, d_{base} and t_{base} respectively) were recorded.

6.6.2. Parameters Tested in this Experiment

The parameters, as well as their associated default and alternative values, tested in this experiment are provided in Table 6.2.

Table 6.2: Parameter tested in this experiment

Parameter Tested	Default Values	Alternative Values
Number of Sires and Dams	500	1000
Distribution of Additive QTL Effect Sizes	$Gamma(0.3, 1.0)$	$Gamma(0.9, 1.0)$
Distribution of Dominance Coefficient	$Half - Normal(0.3)$	$Half - Normal(0.8)$

The default parameter values were chosen such that the distribution of additive QTL effect sizes and dominance coefficients the dominance genetic variance is approximately 15% the additive genetic variance. This value was chosen based on values reported by several

previous publications (Garcia-Baccino et al., 2020; Vitezica et al., 2016). When the effects of one of the parameters were tested, the values for all other parameters were kept at their default values.

6.6.3. Testing the Performance of the OCS

To test the performance of the proposed OCS, several methods of genomic selection were conducted. Methods tested in this study, as well as the model of selection optimization, are provided in Table 6.3.

For selections that includes true additive and dominance scores (i.e. $\boldsymbol{b}$ for SWUA, SWAI, SWAD, SWD1 and SWH1, and $\boldsymbol{D}$ for SWAD, SWED, SWTD, and SWD1), the aim for these scenarios was to test the theoretical improvement achievable with perfectly accurate sources of information, whereas the estimates (i.e. $\widehat{\boldsymbol{b}}$ and $\hat{\boldsymbol{t}}$ for additive and $\boldsymbol{H}$ for dominance genetic components) were used to test the performance of this OCS under realistic sources of information that might not be perfectly accurate. The usability of heterozygosity as a proxy of optimizing dominance effects were also tested by comparing the performance between SWEI (i.e. EBVs without dominance component), SWED (i.e. EBVs with true dominance effect sizes) and SWEH (i.e. EBVs with heterozygosity). Any improvement in the total genetic merit in SWEH and SWED compared to SWEI were deemed to be the additional gain obtainable from the inclusion of heterozygosity and dominance component in the OCS respectively.

The SWD1 and SWH1 scenarios are used to test the effects of early termination of optimization of the dominance genetic component in the breeding program on its scores in the subsequent generations. For those optimized genomic selection scenarios that excluded the optimization of dominance genetic component (i.e., SWAI, SWEI, SWD1 and SWH1), $\boldsymbol{S}_{allP1}$ was used in place of $\boldsymbol{S}_{allP2}$ (i.e., bypassing Phase 2 of the algorithm), and the term d_s was omitted in equation [20] in Phase 3 of the algorithm. For the un-optimized genomic selection, only the additive genetic component was used in the selection process. For optimizations that utilized estimates, the corresponding true additive and dominance scores ($\boldsymbol{b}$ and $\boldsymbol{D}$ respectively) were also recorded. This allows valid comparisons on the true additive and dominance genetic gains between different optimization methods.

Table 6.3: Method of genomic selection and model of selection optimization tested in this study. For the models of selection optimization, $\boldsymbol{b}$ represents optimization of additive using sire true QTL effect sizes; $\widehat{\boldsymbol{b}}$ represents that of sire estimated additive breeding values; $\hat{\boldsymbol{t}}$ represents that of sire marker test statistics from GWAS; $\boldsymbol{D}$ represents optimization of non-additive using true dominance effect sizes, $\boldsymbol{H}$ represents that of heterozygosity and $\boldsymbol{I}$ represents selection with constraint on inbreeding coefficient.

Genomic selection method	OCS Utilized	Model of selection optimization
NSEL	No	No selection
SWUA	No	$\boldsymbol{b}$
SWUE	No	$\widehat{\boldsymbol{b}}$
SWAI	Yes	$\boldsymbol{b} + \boldsymbol{I}$
SWEI	Yes	$\widehat{\boldsymbol{b}} + \boldsymbol{I}$
SWAD	Yes	$\boldsymbol{b} + \boldsymbol{D} + \boldsymbol{I}$
SWAH	Yes	$\boldsymbol{b} + \boldsymbol{H} + \boldsymbol{I}$
SWED	Yes	$\widehat{\boldsymbol{b}} + \boldsymbol{D} + \boldsymbol{I}$
SWEH	Yes	$\widehat{\boldsymbol{b}} + \boldsymbol{H} + \boldsymbol{I}$
SWTD	Yes	$\hat{\boldsymbol{t}} + \boldsymbol{D} + \boldsymbol{I}$
SWTH	Yes	$\hat{\boldsymbol{t}} + \boldsymbol{H} + \boldsymbol{I}$
SWD1	Yes	$\boldsymbol{b} + \boldsymbol{D} + \boldsymbol{I}$ for generation 1; $\boldsymbol{b} + \boldsymbol{I}$ for later generations
SWH1	Yes	$\boldsymbol{b} + \boldsymbol{H} + \boldsymbol{I}$ for generation 1; $\boldsymbol{b} + \boldsymbol{I}$ for later generations

For the SWUA and SWUE tests, additional work was done to calculate the expected number of selected top sires required to generate the targeted amount of increment in the inbreeding coefficient. It is defined as follows: let k be the number of top sires being chosen from a total of N_m sires. These sires were used to generate a sire index array $\boldsymbol{S_{SWU}}$, which through equation [14] translated into its corresponding sire proportion array $\boldsymbol{X}$ ($\boldsymbol{X_{SWU}}$). The number of selected top sires $k_{\Delta I_t}$ was defined as the value k such that when used to generate the $\boldsymbol{X_{SWU}}$ the $\boldsymbol{X_{SWU}}$ fulfil the following equation:

$$average\left(diag\left(\boldsymbol{X_{SWU}}\boldsymbol{G^*_{GRM}}\boldsymbol{X_{SWU}}^{\boldsymbol{T}}\right)\right) = \Delta I_t \qquad [22]$$

Using the top $k_{\Delta I_t}$ number of sires, the $\boldsymbol{S_{SWU}}$ array and $\boldsymbol{X_{SWU}}$ arrays were generated, and their additive, dominance and rate of inbreeding coefficient were calculated as in other types of genomic selection tested in this experiment.

For the additive and dominance genetic component, as well as the total genetic merit for each of the methods of genomic prediction tested, they were compared with the base values, which is defined as the expected values had the mating being done randomly and no selection is in place (i.e., the NSEL). For the inbreeding coefficient, the absolute values of changes between generations were reported.

Besides the optimization performance of the OCS, the runtime performance of the OCS was also recorded. This includes the time required for a run of genetic algorithm to converge in each phases, the total runtime of the OCS across all reruns for each generations, and the number of iterations before the convergences.

6.7. Results

6.7.1. Overall Results of Optimization

Given a constraint on the rate of inbreeding, genomic selection that utilized OCS (i.e., SWAI, SWEI, SWAD, SWAH, SWED, SWEH, SWTD, SWTH, SWD1, SWH1) has a larger increase in the additive genetic component in comparison with un-optimized genomic selection (i.e., SWUA, SWUE). Among the genomic selection methods that utilized OCS however, those that optimized the dominance genetic component (i.e., SWAD, SWAH, SWED and SWEH, as well as SWD1 and SWH1 in the first generation) achieved a significantly higher dominance genetic response compared to those that did not do such optimization (i.e., SWAI and SWEI). Under default conditions, OCS methods that have the true dominance effects included have an average gain of the dominance genetic component of +8.41, compared to +0.07 for those an optimization that did not include dominance (Figure 6.1(a)).

The increase in genetic merit due to dominance genetic effects translated to an increased total genetic merit of +12.66 in the first generation for OCS that exploited dominance effects, compared to +2.75 for OCS without considering dominance effects, a difference of +9.91. A similar difference in total genetic merit was observed in the final generation, with total genetic merit of +19.16 for those that include dominance genetic component, compared to +9.01 for those that did not, a difference of +10.15 (Figure 6.1(b)), in other words, the additional gain due to exploiting dominance effects was a genetic lift that was mainly due to selection in the first generation.

The OCS remained successful even if only estimated data were utilized (i.e. $\widehat{\boldsymbol{b}}$ or $\hat{\boldsymbol{t}}$ for additive optimization and $\boldsymbol{H}$ for dominance optimization). With the use of EBV $\widehat{\boldsymbol{b}}$, compared with un-optimized truncation selection SWUE, the OCS increased the additive genetic component from +1.31 to +2.59 in SWEH. For SWEH, where the non-additive optimization was included, the non-additive genetic component reached +4.98 in the first generation, compared to +0.26 for SWEI if the non-additive genetic component was omitted (Figure 6.2(a)). In terms of total genetic merit, the SWEH method achieved a value of +7.57, compared to +2.55 for SWEI (Figure 6.2(b)). Whereas in the SWTH strategy with the use of marker test statistics to form $\hat{\boldsymbol{t}}$, the additive, non-additive and total genetic merit were increased to +3.03, +7.47 and +10.50, respectively.

While the OCS successfully improved the offspring genetic merits, the accuracy of the additive scores used in the optimization has significant effects on the additive genetic gains of the offspring. As example, for SWAD, which use the true QTL effect size $\boldsymbol{b}$, the OCS achieved an additive genetic gain of +5.10 in the first generation, whereas for SWED, where EBVs $\widehat{\boldsymbol{b}}$ was used, the optimization only achieved an average additive genetic gain of +2.48 in the first generation, and for SWTD, where test statistics $\hat{\boldsymbol{t}}$ was used, an additive genetic gain of +2.78 was achieved. These correspond to 48.6% and 54.5% of the maximum genetic gain based on using true additive genetic effects in the method (Figure 6.3). These figures should represent the accuracy of genomic selection.

Similar observations have also been made on the effects of accuracy of dominance scores used in the optimization. One such example was SWEH, which uses heterozygosity score array $\boldsymbol{H}$, the algorithm optimized the dominance genetic component up to +4.98 in the first generation, whereas for SWED, which uses true dominance effect size score array $\boldsymbol{D}$, an optimization of +6.34 in the first generation was achieved. This corresponds to a 78.5% efficiency in using the dominance genetic component gained from optimization when approximating dominance gain by using predicted genome-wide heterozygosity in progeny (Figure 6.1(a)). The improvement in dominance component is sufficiently substantive when using heterozygosity as a score of non-additive genetic component optimization, and in most cases, this is likely higher than using estimated dominance effects.

The decline in the dominance genetic component optimization also translated into lower total genetic merits in the optimized solutions; for heterozygosity score array, the total genetic merit of optimized solution has an average of +11.96 in the first generation, compared to

+13.35 if true dominance score array was used. This translated to a decline of 10.4% in total genetic merit in the use of heterozygosity instead of true dominance effect sizes (Figure 6.1(b)).

In terms of runtime performance, under the default set of parameters, the genetic algorithm in the first phase converges after an average of 281.28 iterations at an average of 43.1 seconds per run. Given that there were 8 runs in Phase 1 per generation, an average of 328.35 seconds was needed to complete this phase. Whereas for Phase 2, the genetic algorithm converges after an average of 2325.94 iterations over 278.21 seconds per run. Across all 8 runs of Phase 2, the total average runtime for this phase is 2376.53 seconds. Similar observations in term of runtime performances were made for all parameter values tested, with the exception of the case with 1000 sires and dams. For this parameter values, the first phase genetic algorithm converges after an average of 338.44 iterations at an average of 223.17 seconds per run. The average runtime for Phase 1 across all 8 reruns was 1651.42 seconds. Whereas for Phase 2, the genetic algorithm converges at average of 4777.86 iterations, which requires 723.76 seconds. Across all 8 reruns, the total runtime for this phase is 4894.16 seconds.

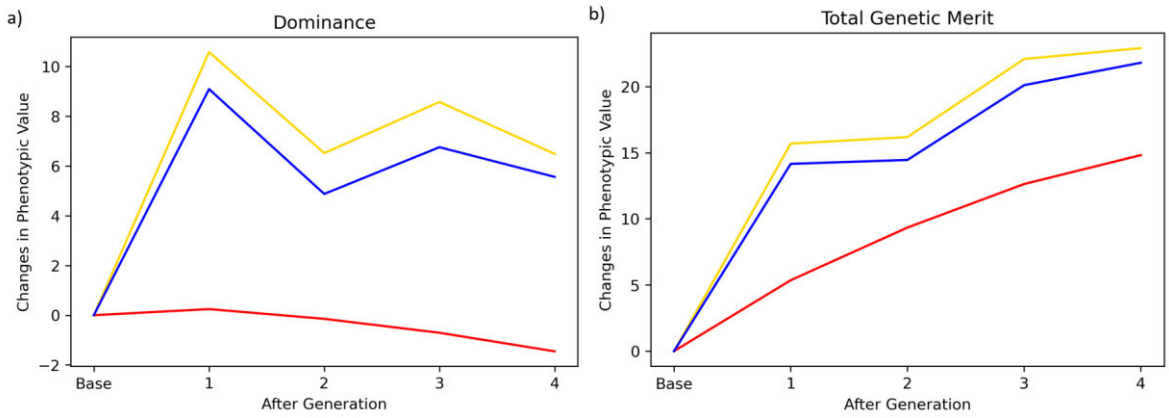

Figure 6.1: Response to selection on the (a) dominance genetic component and (b) total genetic merit across generations. Red line represents SWAI, the OCS that omits dominance component ($\boldsymbol{b} + \boldsymbol{I}$), yellow line represents SWAD, the OCS that utilized true dominance effects ($\boldsymbol{b} + \boldsymbol{D} + \boldsymbol{I}$) and blue represent SWAH, the OCS that utilized heterozygosity ($\boldsymbol{b} + \boldsymbol{H} + \boldsymbol{I}$). A full description of the methods and models are provided in Table 6.3. For all these methods, a default set of parameters were utilized.

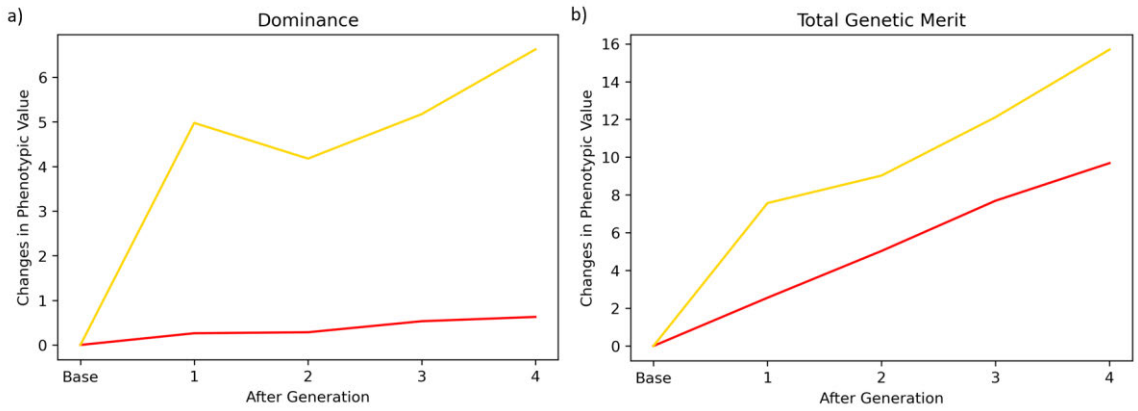

Figure 6.2: The effects of inclusion of the dominance component in mating optimization on (a) the dominance genetic component and (b) total genetic merit across generations in situations where only estimated data are utilized. The red line represents SWEI genomic selection method that ignores the non-additive component ($\widehat{\boldsymbol{b}} + \boldsymbol{I}$) and the yellow line represents the SWEH method that utilizes the heterozygosity score array ($\widehat{\boldsymbol{b}} + \boldsymbol{H} + \boldsymbol{I}$). A full description of the methods and models is provided in Table 6.3. The default set of parameters was used for all methods utilized.

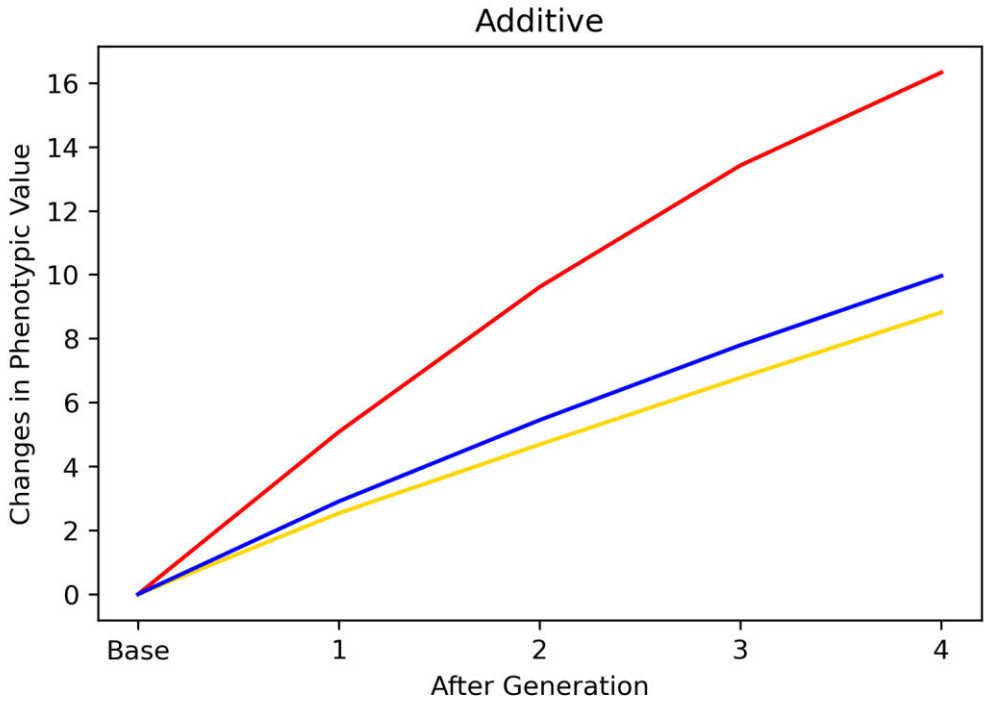

Figure 6.3: The effects of type of additive information on the optimization of the additive genetic component, with the red line being true QTL effect size $\boldsymbol{b}$ being used in the optimization, yellow line being EBV $\widehat{\boldsymbol{b}}$ and blue line being GWAS test statistics $\hat{\boldsymbol{t}}$. For all these methods, default set of parameters were utilized.

6.7.2. Effects of Cessation of Optimization of Dominance Genetic Component

Unlike the additive genetic component, the dominance genetic component is not cumulative from generation to generation. Thus, the cessation of optimization of this component would result in a decay of the dominance genetic component for the subsequent generations. This effect could be observed during the comparison of the total genetic merit between SWAI, where only the additive genetic component was optimized, and SWAH, where both additive and dominance genetic component were optimized for all generations, and SWH1, where the dominance component is optimized for the first generation only.

In the first generation, the total genetic merit of SWH1 reaches +14.17 from the optimization of the dominance genetic component, which is comparable to +14.15 for SWAH, and is significantly higher than +5.36 observed in truncation selection method SWAI. From the second generation onward however, optimization of dominance genetic component ceases in SWH1, thus the total genetic merit decays to +9.20, which is comparable to +9.33 for SWAI, and significantly lower than +14.45 in SWAH (Figure 6.4).

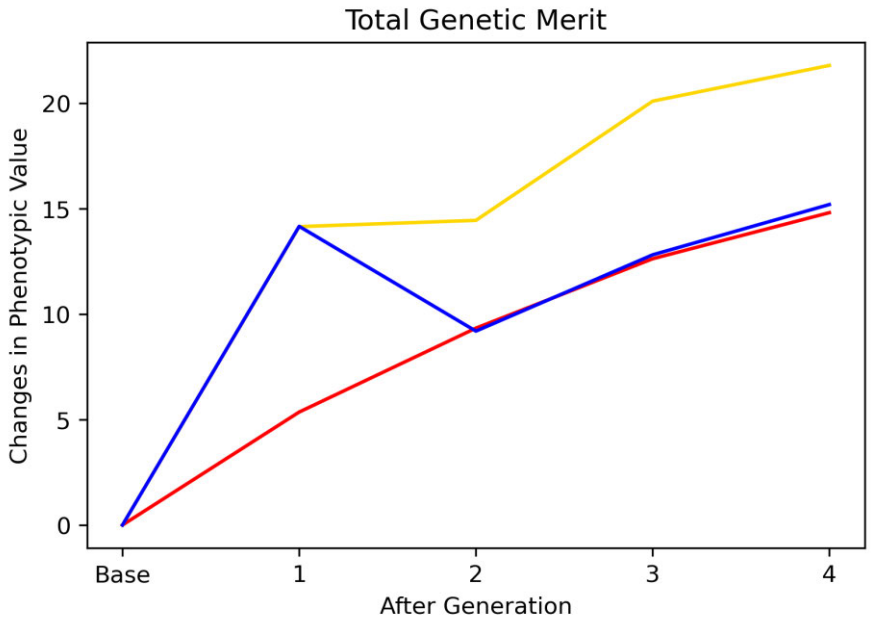

Figure 6.4: The effects of cessation of optimization of dominance genetic component on the total genetic merit from the OCS. Red line represents truncation selection method SWAI ($\boldsymbol{b} + \boldsymbol{I}$), yellow line represents SWAH method ($\boldsymbol{b} + \boldsymbol{H} + \boldsymbol{I}$) and blue line represent SWH1 ($\boldsymbol{b} + \boldsymbol{H} + \boldsymbol{I}$) in first generation, $\boldsymbol{b} + \boldsymbol{I}$ in subsequent generations). Full description of the methods and models were provided in Table 6.3. For all these methods, default set of parameters were utilized.

6.7.3. Effects of Parameters

6.7.3.1. Effects of Number of Sires and Dam

Besides the methods of genomic selection, the number of sires and dams have significant effects on the OCS. For simulations that utilized the true additive genetic component $\boldsymbol{b}$, by increasing the number of sires and dams from 500 to 1000, under the same constraint of inbreeding coefficient increment while allowing sire selection intensity to vary, the additive genetic component increases from an average of +5.11 to +6.27 in the first generation, an increment of 22.7%, and the total genetic merit from +14.89 to +15.66, an increment of 5.2%. When selecting on the EBV $\widehat{\boldsymbol{b}}$, the corresponding additive genetic component increases from +2.53 to +3.76, an increment of 48.6%, and total genetic merit increased from +8.20 to +10.26, an increment of 25.1%, with the same increment in number of sires and dams. One possible reason of this observation is that by increasing the number of sires and dams, a more stringent selection intensity could be afforded under the same increment of inbreeding coefficient, thus with increased additive genetic gain. Another possible reason is an increased accuracy of the EBVs estimated using a larger reference population.

For the OCS that utilized $\widehat{\boldsymbol{b}}$ for additive and $\boldsymbol{H}$ for dominance genetic component optimization (i.e., SWEH), increasing the number of sires and dams increased gain from the additive genetic component from +2.59 to +3.67 in the first generation, and slightly increased the dominance genetic component from +4.98 to +5.64, with the total genetic merit increasing from +7.57 to +9.31 (Figure 6.5).

6.7.3.2. Effects of Additive Genetic Variance

Genetic architecture parameters such as additive genetic variance also has significant effects on the OCS. By increasing the shape parameter of the additive QTL effect size distribution, which increases the additive genetic variance by approximately fourfold, the increment of additive genetic score in the first generation has increased from an average of +5.08 to +13.34 for both SWAD and SWAH, which correspond to an increment of 162.6%. This increment in the additive genetic score has translated into an increase of total genetic merit from +14.92 to +22.89, which corresponds to an increase of 53.4%. Additive QTL effect size distribution do not have significant effects on the dominance scores.

For the OCS that utilized $\widehat{\boldsymbol{b}}$ for additive and $\boldsymbol{H}$ for dominance genetic component optimization (i.e., SWEH), increasing the additive genetic variance increases the additive

genetic component from +2.59 to +9.20 in the first generation, and the non-additive genetic component from +4.98 to +6.84. This produces an increase in the total genetic merit from +7.57 to +16.04, an increment of 112% (Figure 6.6).

6.7.3.3. Effects of Dominance QTL Genetic Architecture Parameters

The genetic architecture for the dominance genetic component has significant effects on the OCS. By increasing the dominance QTL effect size distribution variance from $Half - Normal(0.3)$ to $Half - Normal(0.8)$ increases the dominance genetic scores from an average of +9.11 to +24.36 if for true dominance effect size score array $\boldsymbol{D}$ was used, and an average from +7.71 to +20.14 for heterozygosity score array $\boldsymbol{H}$. This could be attributed to an increased variance in the dominance genetic component.

It is also noted that the choice of the additive genetic scores altered the effects of genetic architecture for the dominance genetic component on the OCS. If the true additive genetic component $\boldsymbol{b}$ was used in the optimization process, the same changes in dominance QTL effect size distribution slightly increases the additive genetic component increases from an average of +5.12 to +5.80, which corresponds to an increment of 13.3%. This is not the case for EBVs $\widehat{\boldsymbol{b}}$ however; the same changes in the dominance QTL effect size distribution, the additive genetic component decreases from +2.54 to +1.42, a decline of 44.1%. Despite this, in comparison with truncation selection with $\widehat{\boldsymbol{b}}$ SWUE, the OCS still has significantly higher additive genetic component; the implementation of OCS in SWEH has successfully increases the additive genetic component from +1.31 in SWUE to +2.59 in SWEH for $Half - Normal(0.3)$, and from +0.50 in SWUE to +1.44 in SWEH for $Half - Normal(0.8)$.

For OCS that utilized $\widehat{\boldsymbol{b}}$ and $\boldsymbol{H}$ for the additive and dominance genetic component optimization (i.e., SWEH), increasing the variance of the dominance QTL effect size distribution decreases the additive genetic component from +2.59 to +1.44, a decline of 44.4%, but increases the dominance genetic component from +4.98 to +9.20. This produces an increase in the total genetic merit from +7.57 to +10.64, an increment of 40.6% (Figure 6.7).

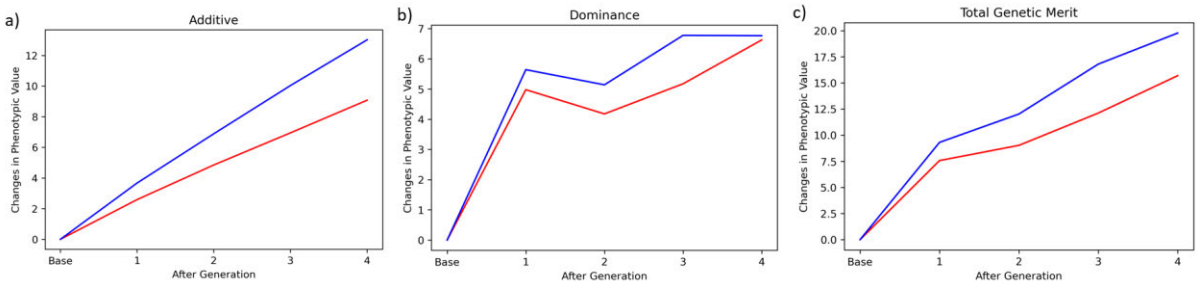

Figure 6.5: The effects of number of sires and dams on the optimization of (a) additive genetic component, (b) dominance genetic component and (c) total genetic merit. Red line represents optimization with 500 sires and dams, and blue line represents optimization with 1000 sires and dams. For these plots, SWEH method ($\widehat{\boldsymbol{b}} + \boldsymbol{H} + \boldsymbol{I}$) was used.

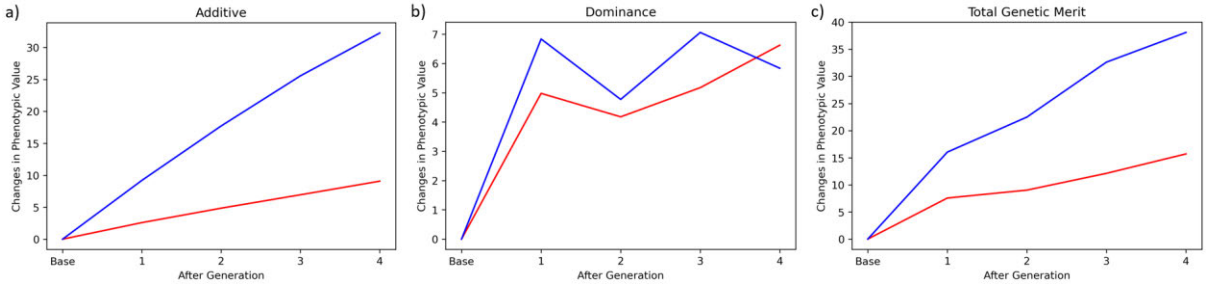

Figure 6.6: The effects of additive genetic variance on the optimization of (a) additive genetic component, (b) dominance genetic component and (c) total genetic merit. Red line represents optimization on a trait with many small QTL (i.e. $Gamma(0.3,1.0)$), and blue lines on trait with few large QTL (i.e. $Gamma(0.9,1.0)$). For these plots, SWEH method ($\widehat{\boldsymbol{b}} + \boldsymbol{H} + \boldsymbol{I}$) was used.

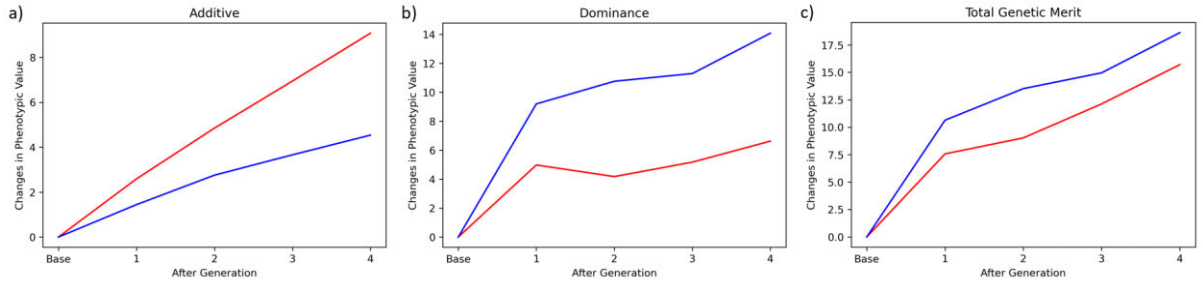

Figure 6.7: The effects of variance of the dominance QTL effect size distribution on the optimization of (a) additive genetic component, (b) dominance genetic component and (c) total genetic merit. Red line represents optimization on a trait with its additive QTL effect size distribution follows $Half - Normal(0.3)$, and blue lines on distribution that follows $Half - Normal(0.8)$. For these plots, OCS method with heterozygosity and EBVs SWEH ($\widehat{\boldsymbol{b}} + \boldsymbol{H} + \boldsymbol{I}$) was used.

6.8. Discussion

In this experiment the new optimal contribution selection algorithm that takes into account the optimization of non-additive genetic effects has been proposed and tested under varying genetic architectures, sample size and inbreeding level of the base population. By optimizing the contribution of sires to the next generation, the algorithm has successfully optimized the additive genetic component of the offspring within a constraint on inbreeding rate in a multi-generational selective breeding program, achieving an additive genetic component superior to those expected from truncation selection. Using the true additive genetic value, the OCS increase the additive genetic component from an average of +1.91 to +5.11 in the first generation, and from an average of +7.29 to +16.25 for the final generation. This corresponds to an increment of 167.54% in the first generation and 122.9% in the final generation.

Such an increment could be considered overly optimistic however, as obtaining such accurate additive genetic values is not feasible. In practice only the EBVs or the estimates of marker effects from GWAS are available. Despite this, even with more realistic accuracies, the OCS was still able to substantially increase the additive genetic gains from the un-optimized truncation selection; using the EBVs, the OCS increased the additive genetic gain from +1.31 to +2.54 in the first generation, and for the final generation, an average increment from +5.47 to +8.82. This corresponds to an increment of 87.0% for the first generation, and 62.7% for the final generation. While this is significantly higher than the 21% to 27% as reported by Meuwissen (1997), it is comparable to the 81% reported by Nielsen et al. (2011), suggesting its success.

Besides the additive genetic component, through this OCS we also managed to optimize the mating pairs while exploiting the non-additive genetic variance, simulated as the dominance effects in this study. This is up to authors' knowledge the first study that provided an explicit formulation on the ways of incorporating the non-additive genetic component into an OCS algorithm. This algorithm would be important for breeders that wished to exploit the heterosis effects in the animal production system, as well as in the situations when the maximization of the non-additive traits is required. The algorithm has also successfully exploited varying types of additive information (EBVs and genomic information) and non-additive information (true dominance effect sizes and heterozygosity). It is also noted that the increase in dominance component is one-off however, with additional merits not increased after the first generation despite continued optimization.

While it is arguable that the inclusion of dominance component would almost certainly led to a significantly higher dominance genetic response and thus total genetic merits, this depends on having an accurate estimate of dominance genetic component in the parental generation. There were attempts on estimating the dominance genetic component of the population which based on mixed model approaches (as example, Vitezica et al. (2013) and Aliloo et al. (2017)). These methods require an estimated dominance genetic variance however (Lynch and Walsh, 1998), which might not be available for all commercially important traits. The successful utilization of heterozygosity as a proxy of optimization bypassed this constraint, thus allowing a practical method of the optimization of offspring dominance genetic components.

The performance of this method depends on a positive correlation between the score used in optimization and the true genetic values. As example, while heterozygosity serves as a suitable estimate for this study, this is with the assumption that loci with a negative value are uncommon and their effects deemed negligible, which might not always be the case. Thus, the suitability of such estimates needs to be established before feeding it into the algorithm. While it is anticipated that an increased dominance performance could be obtained if a more precise quantification of dominance effect of a loci could be included in the algorithm, attempts to elucidate the genetic architecture for non-additive genetic component of a trait is scarce due to difficulty in estimating the dominance effect sizes, especially without information on the dominance variance, and there is lack of reliable method of estimating variance contributed by epistatic effects (Lynch and Walsh, 1998; Vitezica et al., 2018). Future studies could be dedicated in search for a method that could estimate these non-additive effect sizes, potentially without the use of these variance estimates.

There are also several aspects of the algorithm worth improving. One such aspect was the improvement of hyperparameters chosen for the genetic algorithm in this study. Previous publications on the use of genetic algorithms suggested that the choice of hyperparameters could affect the balance between exploration, where a new region of solution space is searched, and exploitation, where the optima is determined within a region of solution space, which could affect the outcome of the convergence (Srinivas and Patnaik, 1994; Taherdangkoo et al., 2012). Further study could also be dedicated in search for a more optimal hyperparameters, such as those suggested by Heider and Drabe (1997). Alternatively, as it is likely that a poor solution has its configuration far away from those of optimal solution, another method that could be tested would include the adjustments of the mutation,

recombination and inversion rate based on the performance of the solutions (Srinivas and Patnaik, 1994). The possible shortcoming from the two phased genetic algorithm could potentially be alleviated further by “seeding” the solution pool in the first phase optimization with sire-dam pairs with high dominance scores, which increases the proportion of encountering sire-dam pairs with high dominance scores in the second phase, thus improving the chance of optimized pairing in this phase. These alternative optimization methods could warrant further studies.

In conclusion, a new framework for optimal contribution selection that take into account the effects of non-additive genetic component have been proposed in this study. Besides improving the additive genetic component under a constraint on the increment of the inbreeding level, this algorithm has also successfully improved the non-additive genetic component compared to un-optimized genomic selection. The algorithm was successfully tested under varying population and genetic architecture parameters as well as different degrees of accuracy of estimating marker effects and EBV. Further studies could be dedicated to improving the algorithm especially in terms of more optimal hyperparameter values, as well as elucidating the additive and non-additive genetic architecture of a trait.

Chapter 7. General Discussion and Conclusion

The aim of the project is to design a framework for the optimization of selection of breeding pairs in a breeding program using artificial intelligence, with emphasis being placed on the optimization of the additive and non-additive genetic components, using genomic information derived from a GWAS experiment. Due to the massive sample space of all possible breeding pairs, evaluating the performance from each breeding pairs would not be feasible, thus a genetic algorithm, a form of artificial intelligence, has been employed.

While it might be possible to use genomic information such as t-test statistics in the optimization of breeding pair, its performance depends on the quality of the GWAS, which was liable to all the shortcomings a GWAS would suffer. This could include any factors that could affect the power and false positive rate of the GWAS. Some of these factors include allele frequencies and their distribution, narrow sense heritability, genetic architecture of the trait, linkage disequilibrium structures, correlation between markers, sample size and threshold of the experiment. While the effects for some of these factors, such as sample size and narrow sense heritability, have their effects well-characterized in previous publications (Spencer et al., 2009; Visscher et al., 2017), others, especially those related to the genetic architecture, are much less frequently studied. This has become the reason why this project starts by investigating the impact from these factors. This involves a comprehensive study on the impact of these factors on the power and false positive rate of the GWAS. Results from this investigation would serve as the foundational work for techniques to improve the quality of the genomic data.

This study suggests significant impact from various factors on the power and false positive rate of the GWAS. Small sample sizes, extreme allele frequencies and polygenic genetic architectures significantly reduce the power of GWAS in detecting a QTL. Heterogeneity in linkage disequilibrium structures, correlations between markers and small sample sizes increases the false positive rate of the GWAS, especially with a large number of markers. These observations were important in establishing the suitability of using GWAS-based results on the optimization process. The vulnerability of GWAS toward small sample size could impede its use in the optimization, which in such case animal-based data might be more suitable for the optimization.

Besides impacting the power and false positive rate of the GWAS, these confounding factors could also affect the distribution of estimated effect sizes and test statistics from a GWAS experiment, in which a detailed description has been provided in Appendix B. With decreased sample sizes, the variance of the estimated effect sizes of the GWAS increases significantly, while the kurtosis of the distribution of the test statistics decreases. Small sample sizes and allele frequency distribution with higher proportion of extreme allele frequencies also reduces the proportion of markers at the tail of the distribution being actual QTL.

This study has also suggested the sub-optimality of the commonly used multiple testing correction method such as the Bonferroni method and Benjamini-Hochberg False Discovery Rate (FDR), especially under changing genetic architectures and experimental designs. This is due to the fact that the calculation of both thresholds does not take into account the behaviour of the p-values or test statistics of the markers under such changing parameters. This causes the Bonferroni method to be insensitive toward the effects of these parameters, and for Benjamini-Hochberg FDR the way the threshold changes do not always line up with what is needed for an improved optimality. One such example is a polygenic trait, which is known to reduce the significance of the markers (Gondro, 2015). The Bonferroni method do not vary with the polygenicity of the trait, thus causing less QTL to have sufficient significance to cross the threshold and be detected. Whereas for Benjamini-Hochberg FDR, instead of decreasing the stringency of the threshold, the threshold stringency increases, thus reduces the optimality of the threshold.

Another aspect worth considering for genomic information obtained from this investigation is the type of information used in the optimization. There are several types of information derivable from a GWAS experiment, two of which being the estimated effect sizes and t-test statistics of the markers. The latter was chosen for the optimization algorithm due to its reduced vulnerability toward extreme allele frequencies. As the estimated effect size has its variance inversely proportional to the $p(1-p)$, where p is the allele frequency (Wang and Xu, 2019), an extreme allele frequency increases the amount of noises in the estimated effect sizes, reducing its suitability for optimization. Whereas for test statistics, it is directly proportional to the $p(1-p)$ (Spencer et al., 2009), which scale down the noises in the test statistics. This increases the signal-to-noise ratio of the top markers (i.e., the proportion of top markers being actual QTL), and indicated an increased likelihood of markers with large test statistics being an actual QTL. This observation has also been validated through additional

simulations (further details provided in Appendix B), where the tail of estimated effect size distribution has lower percentages of actual QTL compared to the tail of test statistics. This also indicate that if marker selection is to be conducted, test statistics of a GWAS instead of estimated effect sizes should be used.

Due to the stringency of the thresholds, the number of detected QTL obtained from a GWAS in this study is significantly less than the actual number of nontrivial QTL included in the simulation. This observation highlighted the difficulty of detecting all or even most of the QTL associated with a trait, which obscures the true underlying number of QTL associated with a trait. This obscurity is exacerbated by the effects of linkage disequilibrium and correlation between markers, which blends the effect sizes across multiple QTL (details provided in Appendix B), and the errors in effect size estimation. To take this issue further, it is unlikely the true number of QTL can ever be definitively determined. This is because the "number of QTL" depends on the definition used to define a QTL in the context of a GWAS; in theory, a QTL can be defined as a locus that explained a nonzero genetic variance. In practice however, this definition also includes a locus that contributed a negligible amount of genetic variance that are virtually undetectable by a GWAS, which raises the question of should we expect a GWAS to be able to detect them. This is especially true if infinitesimal model, where all the loci would contribute a minute amount of genetic variance, is assumed. In this case, the number of QTL is substantially larger than what obtained if finite QTL model is assumed. The estimated number of QTL also depends on the distribution of effect size assumed by the methods used to estimate it. This phenomenon has been noted by Park et al. (2010) who mentioned the use of Weibull distribution resulted in a larger number of QTL estimated compared to that from an exponential distribution, with a large portion of the increment came from QTL with small effect sizes. For all these reasons, it is unreasonable to expect a concrete number of QTL obtainable through a GWAS, especially for QTL with small effect sizes.

These findings were then being used to develop methods and techniques that can be used to improve the quality and suitability of genomic data in the breeding pair optimization algorithm. This includes the proposal of a method to calculate an optimal threshold that could balance the power and false positive rate of the GWAS for Chapter 4, which have previously been reported to be highly dependent on various factors such as genotypic correlation structures (Hoggart et al., 2008; Panagiotou and Ioannidis, 2012). Using simulated genotypes and phenotypes with the assumption of known genetic architectures (i.e. number of QTL and

distribution of effect sizes), the performance of the optimal threshold calculated from this method was tested under varying parameter values. For the performance in binary classification for gene discovery, when measured using Matthews Correlation Coefficient (MCC) scores, the use of the optimal threshold has led to an improvement of up to 54.9% compared to the Bonferroni method, and 11.0% compared to the Benjamini-Hochberg FDR method.

The ROC-based thresholds have significantly better binary classification performance than the commonly used Bonferroni correction and the BH-FDR for all parameters tested. This would aid the gene discovery process in a GWAS, as this threshold increases a GWAS's ability in detecting markers with moderate significance (i.e. markers with peaks that clearly distinguished from the less significant noises but with insufficient significance to cross the threshold), which could aid the phenotypic predictive capability of the detected markers in a polygenic trait, given the observation that one leading cause for the low predictive capability of GWAS output is the failure to detect QTL with small effect sizes (Hall et al., 2016; Kooperberg et al., 2010; Kraft and Hunter, 2009). Furthermore, the significantly improved performance of ROC-based thresholds in binary classification with an increased number of markers also suggests their potential utility in cases where high density markers and Whole Genome Sequence (WGS) is used. In these cases the Bonferroni correction could become overly stringent and thus decimating the power of GWAS..

Despite its apparently excessive leniency, the use of an optimal threshold as a truncation point for marker selection in genomic prediction has led to an improvement in accuracy up to 16.8% compared to the Bonferroni method and 7.0% compared to Benjamini-Hochberg False Discovery Rate (FDR) method. This suggested if markers are selected for prioritization in genomic prediction, a more lenient threshold, such as that proposed by the ROC, would be more suitable. The significantly improved genomic prediction accuracy of the ROC-based thresholds in a strongly polygenic trait compared to Benjamini-Hochberg FDR also suggested its suitability to be used in marker prioritization of such traits, such as milk yield in cattle (Laodim et al., 2019) and mature body size in sheep (Posbergh and Huson, 2021).

The improved performances of ROC-based thresholds in binary classification and marker selection also indicated their potential in an OCS. While threshold is not imperative for the OCS used in this study, this is the case if and only if all the QTL has a positive effect, which in practice do not occur. For the maximization of offspring performance, the number of

positive alleles shall be maximized, and deleterious alleles be excluded, which involves the assignment of positive scores for the former and negative scores for the latter. If the deleterious alleles were assigned with positive scores, maximizing the scores would not yield optimal offspring. This is especially problematic for alleles with small deleterious effects, where the errors of estimation have increased the chance of pushing their scores into the positives. This issue could be mitigated with the use of a threshold, which ensure only the markers with strong positive effects receive a positive score. It is for this reason a method for determining the optimal threshold was included in this study. Despite this, a formal study on the utility of these thresholds on the OCS is desirable, which would be an avenue for further study.

Investigation from Chapter 3 suggested significant effects of genetic architecture on the optimality of a threshold. Furthermore, the calculation of optimal thresholds proposed in Chapter 4 require information on the genetic architectures. For these reasons, we did additional work for Chapter 5 in an attempt to estimate the genetic architecture parameters, such as number of QTL, as well as shape and scale parameters for the QTL effect size distribution, while simultaneously take into account the impact from various confounding factors such as allele frequency distribution, linkage disequilibrium structures and correlation between markers. Using simulated genotype, phenotype and narrow sense heritability, the estimated number of QTL with effect size 0.1 σ_e ranges from 69.9% to 167.0% (average 109.8%) of the true number of QTL, and for effect size 1.0 σ_e from 101.6% to 175.8% (average 123.6%). This method can also be used to estimate the QTL effect sizes in consideration with various confounding factors. This work is important for a more accurate prediction of marker effect sizes, which could also be fed into the ROC-threshold calculation in Chapter 4 and subsequently marker prioritization in the OCS for Chapter 6. While simulated dataset were used in this study, it is anticipated this method could be used with real data, and this could be an avenue for further testing.

One distinguishing features of our method of estimating genetic architecture parameters is the flexibility in the choice of QTL effect size distribution. Unlike previous methods such as Cheng et al. (2020), Moser et al. (2015) and Zhang et al. (2018, 2021) who use a normal distribution or a mixture of normal distributions, our method allows users to choose a wide range of distributions such as gamma distribution, Weibull distribution and q-exponential distributions, which have increased flexibility in the shape of distribution, thus better capture the shape of the QTL effect size distribution, especially if the distribution is strongly

leptokurtic. The changing number of QTL with varying shape parameters of the distribution was supported by Park et al (2010), who stated the number of QTL depends on distributions assumed by the estimation method, and our method further extended this statement by enumerating the range of expected number of QTL under varying shape parameters. This capability of our method could be beneficial in resolving the genetic architecture for a wide range of traits, be it a polygenic trait such as milk yield in cattle (Nayeri et al., 2016) or an oligogenic trait such as polledness in cattle (Scheper et al., 2021) and goats (Guo et al., 2021).

The success of this method in estimating the number of QTL with small effect size (i.e. 0.1 σ_e) as well as their effect size distribution shows prospect for the method's utility in a high density marker system, such as with WGS data. It is anticipated that the extreme stringency in threshold in a traditional GWAS with high density markers could severely reduce a GWAS's power in detecting QTL with small effect sizes (Tam et al., 2019). As our method does not rely on having markers reach the threshold of significance, but rather it utilizes the statistical properties of the distributions of marker test statistics obtained from a GWAS, it could help uncover genomic regions associated with a trait that were previously undetectable due to their small effect size. The increased linkage disequilibrium in high density markers and WGS data could further aid the performance of this method. Unlike some of the previously published methods, our proposed method does not rely on arbitrarily-set thresholds, such as linkage disequilibrium threshold required by Zhang et al. (2018) or trivial effect size threshold as required by Park et al. (2010) and Cheng et al. (2020). Such arbitrary threshold could impact the performance for these methods under changing marker density. For these reasons, in conjunction with WGS, this method has the potential of help resolving the genetic architectures of a complex trait.

In the final chapter (Chapter 6), results obtained from the investigations, such as the preference of test statistics over estimated effect sizes for optimization of additive genetic components, were incorporated into a breeding pair optimization framework, with experimentation on the use of additive and non-additive genomic data in the optimization. For this study, simulated genotype, phenotype and narrow sense heritability were used to test the performance of this optimization method. An avenue for further study would be the testing of this method using real dataset.

Results from this chapter suggested a successful improvement of additive and non-additive genetic component of the offspring; under the same constraint of increment in level of inbreeding, this method has successfully improved the additive genetic component by up to 87.0% in the first generation compared to truncation selection method. With the dominance genetic variance of 15% of the additive genetic variance, the genetic lift in the first generation is approximately equal to two generations of additive genetic gains. Similar improvements have also been observed with the use of estimated breeding values and GWAS test statistics as additive genetic data and heterozygosity as non-additive genetic data. This framework has also been successfully tested under varying population and genetic architecture parameters such as number of sires and dams, and additive and non-additive genetic variances. From these results it can be said that the main aim of the project has been achieved, a feat made possible with the use of genetic algorithm and the incorporation of findings from the investigations on factors that could affect the usability of genomic data and techniques to improve their usability.

There are several future prospects for the OCS, most notably the improvement of livestock production traits that have significant portion of non-additive genetic variance while constraining the increase in level of inbreeding. Previous published methods on exploiting non-additive genetic components in a selective breeding program have done so either through crossbreeding (e.g. Shepherd and Kinghorn, 1998) or through methods that do not constrain the level of inbreeding to a predefined level (e.g. González-Diéguez et al., 2019). Method proposed by González-Diéguez et al. (2019) also requires estimates of dominance effect sizes, which is not available without information on the dominance genetic variance. The use of heterozygosity bypassed this requirement, allowing the use of this OCS in these traits in these populations and approximating the maximum benefit that could be obtained if dominance effects were known, thus anticipated to produce significant economic benefit to the livestock industry. As non-additive effect sizes from QTL is more susceptible to linkage disequilibrium decay compared to the additive effect sizes (Visscher et al., 2017), the use of whole genome sequence data could further improve the performance of the OCS in optimizing the non-additive genetic component of the offspring, which worth further testing.

While this algorithm has successfully improved the additive and non-additive genetic components in the offspring generation, there are numerous aspects that could be improved, which served as avenues for further studies. One such aspect is the genetic algorithm used for optimization. Numerous variants of genetic algorithms have been proposed in the past, each

with its own strengths and weaknesses that could influence the optimization. As for example, while the cascading genetic algorithm proposed in this study could alleviate the negative effects from competition between additive and non-additive component, it could come with a cost of reduced optimization of the latter due to reduced diversity in solution from the optimization of the former. The genetic algorithm proposed by Heider and Drabe (1997) could produce a more optimal hyperparameter values but came at a cost of increased runtime from the search of said optimal hyperparameters. Further work could be dedicated in search for a more optimal genetic algorithm for this framework.

Another aspect worth further studying is the determination of effect sizes of non-additive genetic components such as dominance component. While there have been attempts to estimate the non-additive effect sizes (Aliloo et al., 2017; Goudey et al., 2013; Niel et al., 2015; Vitezica et al., 2013), they may not be feasible to implement in a practical level due to the complexity of the model (Niel et al., 2015) the requirement for large sample sizes (Visscher et al., 2017), and its reliability remained largely untested. This is exacerbated by the ever increasing genotypic density and declining genotyping cost, such as that for whole-genome sequence data (Visscher et al., 2017). Thus, a simple, scalable and reliable algorithm that could estimate the non-additive effect size of a polygenic trait would be desirable, which could warrant further studies.

The algorithm for the estimation of genetic architecture parameters proposed in this study also solely focused on the additive genetic component. This is due to a lack of previously published algorithms that could reliably estimate the non-additive QTL effect sizes. If such algorithm is available, one could further extend the estimation of genetic architecture parameters onto those of the non-additive genetic component, which theoretically could provide insights on the genetic architecture of the non-additive genetic components, including the number of non-additive QTL. This further emphasized the desirability of further study for a reliable algorithm that could estimate the non-additive QTL effect sizes. Despite this, even if such algorithm does not exist or infeasible to be implemented, this study has suggested the ability of optimizing the offspring non-additive genetic component with the use of expected progeny heterozygosity.

In conclusion, a framework for the optimization of breeding pairs using artificial intelligence has been developed, with successes in utilizing additive and non-additive genetic components while constraining the increment in inbreeding coefficient. While there are aspects that could

be improved, in general this algorithm has successfully achieved its aim. This algorithm would be useful for livestock producers and breed or species conservationists that wishes to improve the genetic merit of their livestock herds, while exploiting the non-additive genetic component. It is anticipated that all the methods proposed within this framework (i.e. calculation of ROC-based threshold from Chapter 4, genetic architecture estimator from Chapter 5 and the OCS from chapter 6) could be developed further into full-fledge products (such as Python packages that can be applied to real life mating system alongside with appropriate inputs) that could be utilized in a commercial setting.

Appendix A. The Mathematical Derivation of the Test Statistics and p-values of GWAS

This section of the appendix is to provide a layout on the mathematical derivation of the estimated effect sizes and test statistics for a marker during a single SNP regression in a Genome-Wide Association Study (GWAS). In this project, the derivation would be utilized in the calculation of power and false positive rate of a GWAS experiment, as well as estimation of genetic architecture parameters.

A.1. The Mathematical Derivation

The aim of GWAS is to test the level of correlation between the genotype of a marker and the phenotype, often with the assumption of linearity of the marker effects. For this reason, GWAS is often conducted using linear regression model that attempts to fit the phenotype as the response variable and genotype as explanatory variable (Gondro, 2015). With this assumption, the phenotype is modelled as follows:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{a} + \boldsymbol{e} \qquad [1]$$

Where $\boldsymbol{y}$ being the phenotypic vector; $\boldsymbol{X}$ being a matrix containing the genotypic states of all animals and markers; $\boldsymbol{a}$ being a vector containing the additive effect sizes of all markers and $\boldsymbol{e}$ being a vector containing the residual component of the phenotype.

Mathematically the distribution of a bi-allelic genetic variant with frequency p follows a Bernoulli distribution where the only possible outcomes of the alleles are 0 and 1 (Mun, 2012). For this distribution, the variance contributed by an allele j in locus i can be calculated as follows:

$$Var(\boldsymbol{X}_{ij}) = p_i(1 - p_i) \qquad [2]$$

Where the p is the probability of observing "1" (i.e., allele frequency in the context of SNP markers). Given that genotype of each marker comprises of sums of two independently assorted allele pairs (i.e., two independent Bernoulli distribution), the variance of the locus i (henceforth denoted as $var(\boldsymbol{X}_i)$) can be denoted as follows:

$$var(\boldsymbol{X}_i) = p_i(1 - p_i) + p_i(1 - p_i)$$

$$= 2p_i(1 - p_i) \quad [3]$$

Given that $var(\boldsymbol{X}_i)$ is a sample variance instead of population variance, the $var(X)$ need to be adjusted with a factor of $\frac{N-1}{N}$, where N is the sample size of GWAS, thus yielding the formula:

$$var(\boldsymbol{X}_i) = \frac{2p_i(1 - p_i)(N - 1)}{N} \quad [4]$$

The contribution of a marker genotype to each individual's phenotype can be obtained by multiplying the genotype with the allele substitution effect (henceforth denoted as a), as described by Falconer (1989). The sample variance of the additive genetic component of the phenotypic variance (henceforth denoted as $var(\boldsymbol{G}_i)$) can then be obtained from the product of this multiplication:

$$\begin{aligned} var(\boldsymbol{G}_i) &= var(a_i\boldsymbol{X}_i) \\ &= {a_i}^2\, var(\boldsymbol{X}_i) \\ &= \frac{2p_i(1 - p_i){a_i}^2(N - 1)}{N} \end{aligned} \quad [5]$$

The phenotypic variance (denoted as $var(\boldsymbol{y})$) can then be calculated by the sum of the sample variance of the additive genetic component and the sample variance of the residual component:

$$var(\boldsymbol{y}) = var(\boldsymbol{G_i}) + Var(\boldsymbol{e}) \quad [6]$$

Given the genotype of a marker $\boldsymbol{X}_i$ and phenotype vector $\boldsymbol{y}$, assuming the residual component of the phenotype is independent with the genotype (i.e. $cov(\boldsymbol{X}_i, \boldsymbol{e}) = 0$), the genotype-phenotype covariance can be calculated as follows:

$$\begin{aligned} cov(\boldsymbol{X}_i, \boldsymbol{y}) &= cov(\boldsymbol{X}_i, a\boldsymbol{X}_i + \boldsymbol{e}) \\ &= cov(\boldsymbol{X}_i, a\boldsymbol{X}_i) + cov(\boldsymbol{X}_i, \boldsymbol{e}) \\ &= a * cov(\boldsymbol{X}_i, \boldsymbol{X}_i) + 0 \\ &= a * var(\boldsymbol{X}_i) \\ &= \frac{2p_i(1 - p_i)a_i(N - 1)}{N} \end{aligned} \quad [7]$$

The squared correlation coefficient (henceforth denoted as R^2) is calculated as follows: (Gondro, 2015):

$$R_i{}^2 = \frac{SSR}{SST} \quad [8]$$

Where SSR is defined as the explained sum of squares and SST the total sum of squares. In the context of GWAS, SSR is represented by $cov^2(\boldsymbol{X}_i, \boldsymbol{y})$ and SST represented by $var(\boldsymbol{X}_i) * var(\boldsymbol{y})$. Substituting $SSR = cov^2(\boldsymbol{X}_i, \boldsymbol{y})$ and $SST = var(\boldsymbol{X}_i) * var(\boldsymbol{y})$ into [8] yields the following:

$$R_i{}^2 = \frac{cov^2(\boldsymbol{X}_i, \boldsymbol{y})}{var(\boldsymbol{X}_i) * var(\boldsymbol{y})} \quad [9]$$

The $cov^2(\boldsymbol{X}_i, \boldsymbol{y})$ can in turn written as follows:

$$\begin{aligned} cov^2(\boldsymbol{X}_i, \boldsymbol{y}) &= \left(\frac{2p_i(1-p_i)a_i(N-1)}{N}\right)^2 \\ &= a_i{}^2 * \left(\frac{2p_i(1-p_i)(N-1)}{N}\right)^2 \\ &= a_i{}^2\, var^2(\boldsymbol{X}_i) \end{aligned} \quad [10]$$

Which can then be substituted into [9] as follows:

$$\begin{aligned} R_i{}^2 &= \frac{a_i{}^2\, var^2(\boldsymbol{X}_i)}{var(\boldsymbol{X}_i) * var(\boldsymbol{y})} \\ &= a_i{}^2 * \frac{var(\boldsymbol{X}_i)}{var(\boldsymbol{y})} \end{aligned} \quad [11]$$

The test statistic of a locus (henceforth denoted as T_i) to test whether if the slope a is significantly different from 0 (i.e., the $\boldsymbol{X}_i$ has no effect on $\boldsymbol{y}$) can be calculated as follows (Kremelberg, 2011):

$$T_i = \frac{R_i\sqrt{N-2}}{\sqrt{1-R_i{}^2}} \quad [12]$$

Substituting equation [4], [5], [6] and [11] into [12] yields the following:

$$\begin{aligned} T_i &= a_i * \sqrt{\frac{var(\boldsymbol{X}_i) * (N-2)}{var(\boldsymbol{y}) - a_i^2 var(\boldsymbol{X}_i)}} \\ &= a_i * \sqrt{\frac{2p_i(1-p_i)(N-2)}{Var(\boldsymbol{y}) - 2p_i(1-p_i)a_i{}^2}} \end{aligned} \quad [17]$$

If the marker is indeed null (i.e. $a_i = 0$) then the test statistics T_i would follow a standard Student's t-distribution. Given a test statistics T_i, one can calculate the p-values, which is defined as the probability of observing a test statistic T_i assuming the null hypothesis is correct. In a GWAS experiment, the null hypothesis would be the true QTL effect size is zero ($a_i = 0$), hence with zero slope of regression of phenotype on the genotype. Using T_i the negative logarithmically transformed p-value (henceforth denoted as $logpval$) was defined as follows:

$$logpval = -\log_{10}\left(2 * \int_{T_i}^{\infty} t(T_i;\ N-2)\ dx\right) \quad [18]$$

Where $t(x; v)$ is the probability density function (PDF) of Student's t-distribution, which is defined as follows (Mun, 2012):

$$t(x; v) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi} * \Gamma\left(\frac{v}{2}\right)} * \left(\frac{v}{v + x^2}\right)^{\frac{v+1}{2}} \quad [19]$$

The $logpval$ can then be used to test the significance of association of a marker with the phenotype. For this study, it can also be used to calculate the number of true and false positives, and ultimately the power and false positive rate of the GWAS.

Appendix B. The Distribution of Output from a GWAS Experiment

The aim of this appendix is to provide a layout on the distribution of estimated effect sizes and test statistics from a GWAS experiment. This section includes the general asymptotic properties of the distributions of estimated effect sizes and test statistics of a GWAS, as well as factors that would affect the distributions, with emphasis on the effects from the genetic architecture parameters. Observations detailed in this section were then incorporated into an algorithm that aimed to estimate the genetic architecture parameters of a trait while taking into account the effects of nuisance parameters such as sample sizes and correlations between markers. Results from this appendix were also be utilized in testing the reliability of estimated effect sizes and test statistics as scoring method for the optimization of breeding pairs.

B.1. The Asymptotic Distribution of Estimated Effect Sizes and Test Statistics, and Their Implications on the Algorithm Design

B.1.1. The Distribution of Estimated Effect Sizes ($\mathbb{d}^1_{ES}$)

The "asymptotic distribution" is defined as the limiting distribution from an infinite sequence of distribution, which can be approximated by normalized sums or averages of probability densities from sufficiently large number of distributions. For this paper, the asymptotic distribution for $\mathbb{d}^1$ is generated by averaging the sequence of $\mathbb{d}^1$s from multiple GWAS experiments. This is done to alleviate the complications caused by the dispersion of the distribution of random variables obtained from any single GWAS experiment.

In the idealized situation, the estimated effect size obtained from a GWAS experiment would correspond to those in true QTL effect size. Such ideal situation is impossible to achieve, and error of estimation on the QTL effect size is inevitable. With repeated measurements of the same marker however, the estimated effect size ($\hat{a}$) produced follows this normal distribution (Wang and Xu, 2019):

$$\hat{a} \sim \mathcal{N}\left(a, \frac{e^2}{2p(1-p)(N-2)}\right) \quad [1]$$

Where a is the QTL effect size, p the allele frequency, N being the sample size and e^2 being the residual variance. An example of this normal distribution is provided in Figure B.1.

The estimated effect sizes obtained from a marker in GWAS can be thought of as obtaining one sample from the normal distribution of $\hat{a}$ as defined by equation [1]. When conducted across all the markers, this pool of $\hat{a}$ obtained would form a new distribution of estimated effect sizes (henceforth denoted as $\mathbb{d}_{ES}^{1}$). Due to various factors however, the distribution of $\hat{a}$ from each locus would not be independent and identically distributed, even if all the loci have null effects, thus the often-cited central limit theorem does not apply in this distribution. Instead, the asymptotic distribution for $\mathbb{d}_{ES}^{1}$ follows a Student's t-distribution (Lukacs, 1942) (Figure B.2). A mixture distribution was obtained if QTL is present, and while the mixture distribution superficially appeared similar to a Student's t-distribution, there are some differences between the mixture distribution and the Student's t-distribution.

One main difference between the mixture distribution and the t-distribution is their moments, such as variance and kurtosis. In an all-null model where no QTL is associated with the phenotype, the probability of a marker that has its estimated effect sizes achieved extreme values was reduced, and the resulting distribution would have reduced kurtosis, resulting in a more normal-like distribution. Whereas for QTL model, where some markers are associated with the phenotype, the presence of QTL increases the proportion of markers that have their expected value far away from what is expected for a null marker (i.e. $a = 0$), and this pushes the estimated effect size toward the tail of $\mathbb{d}_{ES}^{1}$, increasing its kurtosis.

For the same reason, the effect of QTL is also most observable at the tail of $\mathbb{d}_{ES}^{1}$, suggested that this is the part of the distribution that shall be utilized in the estimation of QTL effect size distribution. Conversely for head of $\mathbb{d}_{ES}^{1}$ (i.e., estimated effect sizes that proximal to zero), the noises from the null markers can easily drown out the signals from the non-null markers. The decreased signal-to-noise ratio in this region of the distribution could reduce the power of detecting non-null markers and hampering any attempt to estimate the underlying QTL effect size distribution. Despite this, the overall changes in the distribution of estimated effect sizes are miniscule, and strongly concentrated at the tail of the distribution (Figure B.3).

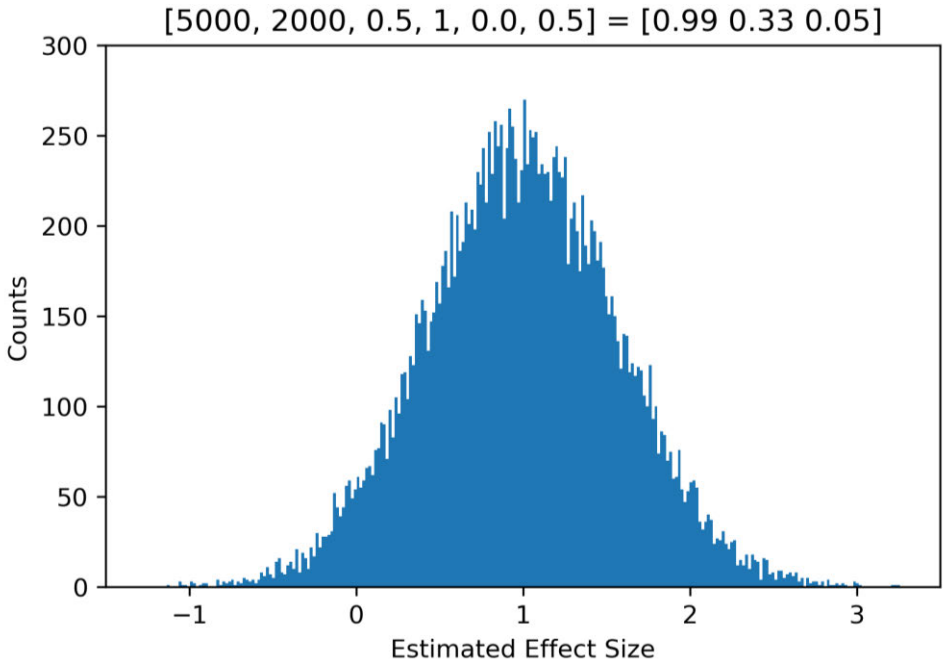

Figure B.1: Histogram showing the distribution of estimated effect size of a QTL with an effect size of 1.0 σ_e. This histogram was generated from 20000 replicates of GWAS with sample size of 5000, allele frequency of 0.5 and with phenotypic variance set at 750.

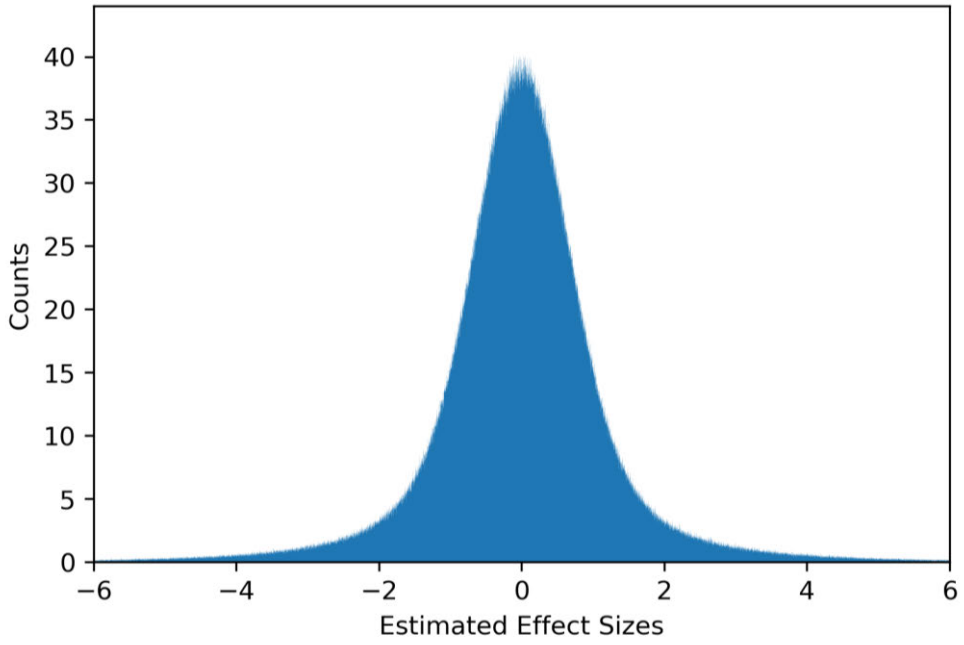

Figure B.2: Histogram of estimated effect size obtained from 100 replicates of GWAS experiment with 50k independent markers, showing the Student's t-distribution. In this plot, a sample size of 5000 was employed, with all markers being null. The distribution is generated by averaging the histograms of estimated effect sizes across all replicates of GWAS.

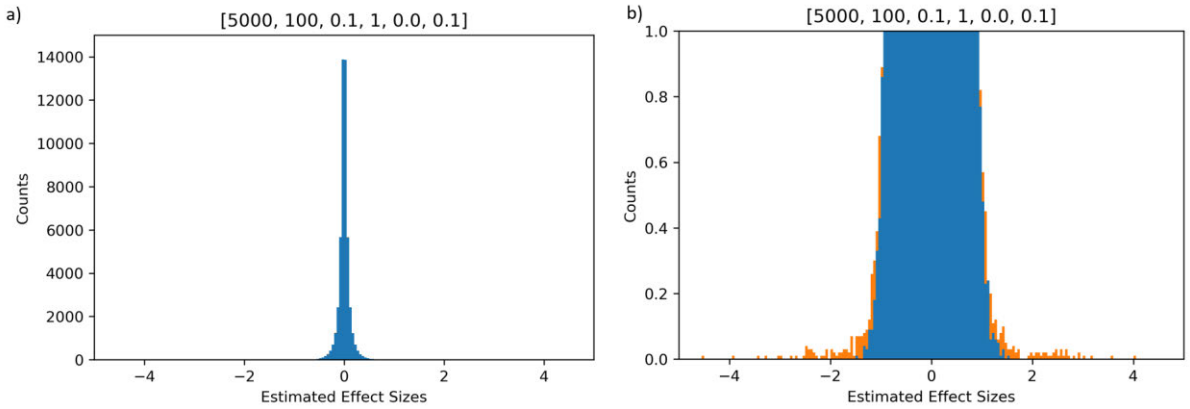

Figure B.3: Histogram showing the distribution of estimated effect sizes of GWAS experiment for an all-null markers (blue) and with 100 QTL that follows a gamma distribution with shape parameter of 0.1 and scale parameter 1 (orange). Figure (a) shows the overall distribution of the estimated effect sizes, while (b) showing the same plot focused on the bottom 1 counts. This GWAS was conducted using 50k independent markers on sample size of 5000, averaged over 100 replicates. The allele frequency distribution tested follows a Beta distribution $Beta(0.1, 0.1)$. The excess kurtosis for the all-null model is 3.95, compared to 9.83 for QTL model.

B.1.2. Distribution of Test Statistics ($\mathbb{d}_{FT}^{1}$)

Estimated effect sizes are not the only data obtainable from a GWAS experiment; a GWAS experiment can also return the test statistics and the associated p-values for each marker, which are used to test the probability of observing a level of genotypic-phenotypic correlation, assuming the null hypothesis being correct (i.e. $a = 0$) (Gondro, 2015). The test statistics of a marker F_i is calculated as follows (Gondro, 2015; Kremelberg, 2011):

$$F_i = \frac{2p_i(1-p_i)(N-2)\hat{a}_i^{\,2}}{Var(\boldsymbol{y}) - \ 2p_i(1-p_i)\hat{a}_i^{\,2}} \qquad [2]$$

Where $Var(\boldsymbol{y})$ is the phenotypic variance. As in the estimated effect sizes, one can also build the distribution for the test statistics (henceforth denoted as $\mathbb{d}_{FT}^{1}$). This distribution is also influenced by the presence of QTL; if all the markers are indeed null, the $\mathbb{d}_{FT}^{1}$ follows a standard F-distribution, and a mixture distribution in case of the presence of QTL (Figure B.4).

As reflected in the distribution of estimated effect sizes, the overall changes in the distribution of test statistics are miniscule, and strongly concentrated at the tail, which is also associated with an increased kurtosis (Figure B.5). This suggested the possibility of utilizing the test statistics of a GWAS experiment in detecting signals associated with the changing underlying QTL effect sizes.

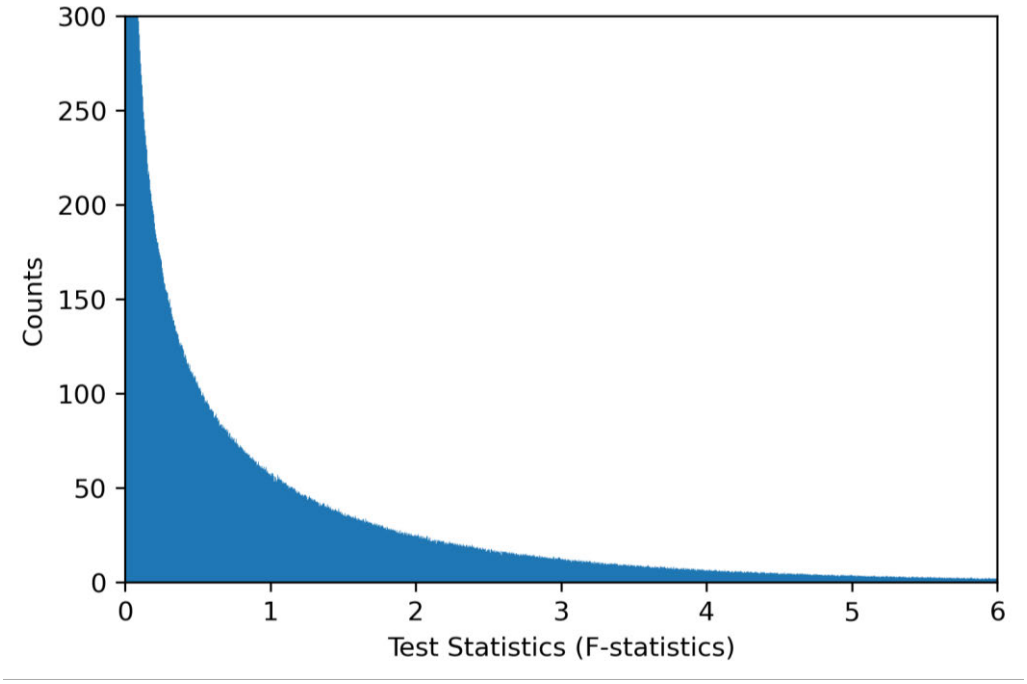

Figure B.4: Histogram of the squared test statistics obtained from 100 replicates of GWAS experiment with 50k independent markers, showing the F-distribution. In this plot, a sample size of 5000 was employed, and all markers are null, and was generated by averaging the histogram of test statistics from all replicates of GWAS.

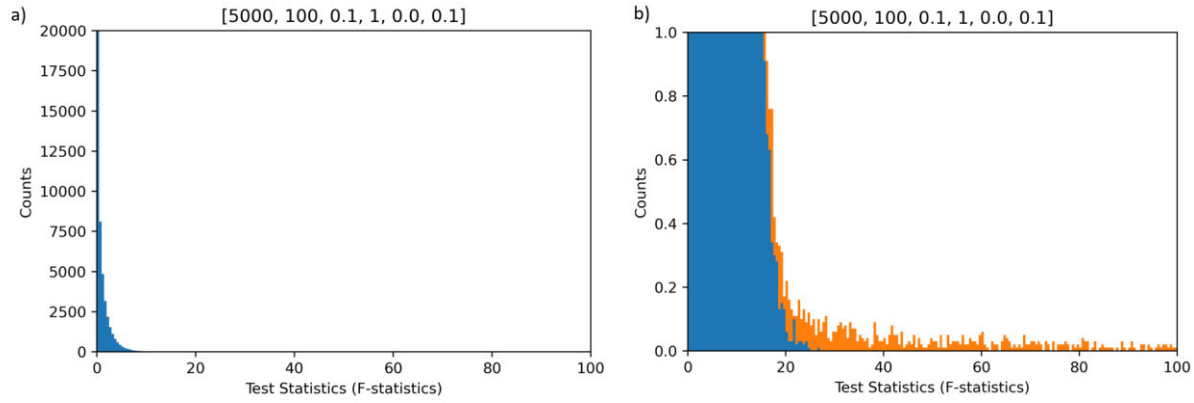

Figure B.5: Histogram showing the distribution of test statistics of GWAS experiment for all-null markers (blue) and with 100 QTL that follows a gamma distribution with shape parameter of 0.1 and scale parameter 1 (orange). Figure (a) shows the overall distribution of the test statistics, while (b) showing the same plot focused on the bottom 1 count. This GWAS was conducted using 50k independent markers on sample size of 5000, averaged over 100 replicates. The allele frequency distribution tested follows a Beta distribution $Beta(0.1, 0.1)$. The excess kurtosis for the distribution of test statistics for all null markers is 11.96, compared to 28896.64 for 100 QTL model.

B.1.3. The Signal-to-Noise Ratio at the Tail of $\mathbb{d}^1_{ES}$ and $\mathbb{d}^1_{FT}$

While the presence of QTL can affect the shape of the tail of distribution of estimated effect sizes $\mathbb{d}^1_{ES}$ and test statistics $\mathbb{d}^1_{FT}$, not all the loci that are located at the tail of $\mathbb{d}^1_{ES}$ and $\mathbb{d}^1_{FT}$ came from QTL. In fact, the proportion of loci at the tail of $\mathbb{d}^1_{ES}$ that originates from QTL is significantly lower than the proportion of loci at the tail of $\mathbb{d}^1_{FT}$ that originates from QTL (Figure B.6).

The reduced proportion of loci at the tail of $\mathbb{d}^1_{ES}$ that originates from QTL is caused by the varying allele frequencies; while the expected value for the estimated effects size is independent of the allele frequency p, its variance is inversely proportional to $p(1-p)$, which attained maximal value when $p = 0.5$ and decreases with extreme allele frequencies. Therefore, extreme allele frequencies increase the variance of estimated effect size. Thus, given a large estimated effect size from a marker (thus at the tail of $\mathbb{d}^1_{ES}$) it could be caused by either the marker is associated with a large effect size, or by extreme allele frequency. It is the latter that contributed to the reduced proportion of loci at the tail of the distribution that originates from QTL. This is not the case for test statistics however, which is directly proportional to $p(1-p)$ (Spencer et al., 2009). The scaling down effects of $p(1-p)$ in the test statistics from extreme allele frequencies reduces the noise from error of estimation near the tail of $\mathbb{d}^1_{FT}$, thus increasing the proportion of loci at the tail of $\mathbb{d}^1_{FT}$ originates from QTL.

This comparison could also be illustrated through additional simulations, of which the results are being provided in Figure B.6. In average only 6.38% of the top 50 markers in term of estimated effect sizes from each GWAS came from non-null markers, compared to 89.48% of the top 50 markers in term of test statistics. The increased signal-to-noise ratio in $\mathbb{d}^1_{FT}$ hinted the desirability of using the $\mathbb{d}^1_{FT}$ rather than the $\mathbb{d}^1_{ES}$ for the detection of QTL, and this could potentially aid the detection of signals from changing underlying QTL effect size distribution.

B.2. Effects of Genetic Architecture on $\mathbb{d}^1_{ES}$ and $\mathbb{d}^1_{FT}$

Besides the presence of QTL, the $\mathbb{d}^1_{ES}$ and $\mathbb{d}^1_{FT}$ are also affected by the genetic architecture of a trait, which can be defined in terms of number of QTL associated with a trait (denoted as $\mathbb{k}$), as well as the distribution of the underlying QTL effect sizes (denoted as $\mathbb{d}_{QTL}$). Understanding the effects of these parameters on $\mathbb{d}^1_{ES}$ and $\mathbb{d}^1_{FT}$ would be crucial for the designing of this algorithm.

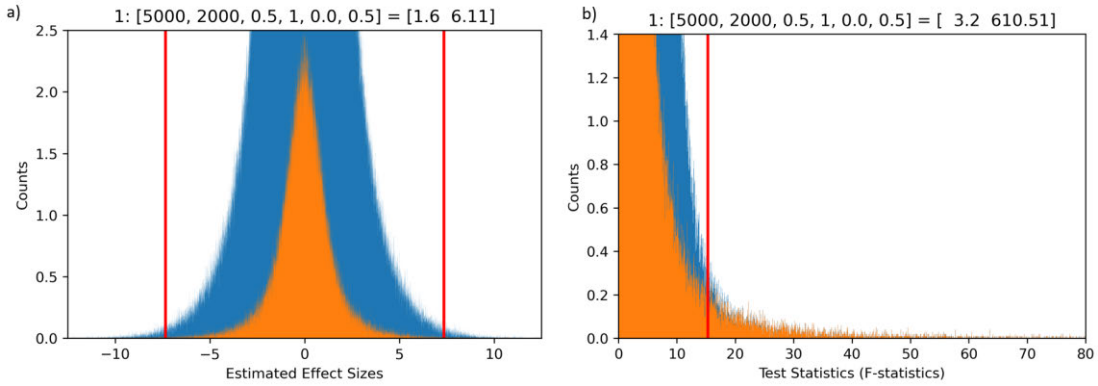

Figure B.6: The histogram of (a) estimated effect sizes and (b) test statistics obtained from 100 replicates of GWAS experiment with sample size of 5000 and 50k independent markers. The blue bars are the null markers, and the orange bar from QTL. 2000 QTL were simulated in this histogram, with their effect sizes following a gamma distribution $\Gamma(0.5, 1)$. The allele frequency distribution follows a beta distribution $Beta(0.5, 0.5)$. From both figures, the top 0.1% of all markers (i.e., 50) from each replicate of GWAS had been extracted and tested, with the red lines indicating the cut-off points.

For this study, the $\mathbb{d}_{QTL}$ is assumed to follow a gamma distribution, with the shape and scale parameter denoted as $\mathbb{a}$ and $\mathbb{b}$ respectively:

$$\mathbb{d}_{QTL} \sim \Gamma(\mathbb{a}, \mathbb{b}) \quad [3]$$

The shape parameter $\mathbb{a}$ dictates the shape of $\mathbb{d}_{QTL}$, with smaller $\mathbb{a}$ produces a more leptokurtic distribution (i.e., distribution with larger kurtosis) and increases the proportion of QTL with small effect sizes over those with large effect sizes (Figure B.7(a)). The scale parameter $\mathbb{b}$ scales the random variates of the $\mathbb{d}_{QTL}$ as follow (Figure B.7(b)) (Mun, 2012):

$$\Gamma(\mathbb{a}, \mathbb{b}) = \mathbb{b} * \Gamma(\mathbb{a}, 1) \quad [4]$$

Under these notations, the parameters for the genetic architecture of the trait were denoted as follows:

$$genetic\ architecture \sim Q(\mathbb{k}, \mathbb{a}, \mathbb{b}) \quad [5]$$

The rationale of using a gamma distribution to model the distribution of QTL effect sizes in this study is its flexibility in its shape. Previous publications such as Cheng et al. (2020); Hall et al. (2016) and Zhang et al. (2018) have assumed normal or exponential distribution of the QTL effect sizes, which have a fixed kurtosis of 3 and 9 respectively. The fixed kurtosis restricts the flexibility of the model, which might cause failure in capturing the full aspect of the effect size distribution. Indeed, this is a concern voiced by Zeng and Zhou (2017) where

the authors stated the assumption of normal distribution for Linear Mixed Model (LMM) reduces the estimation performance. This could be mitigated by having a more flexible model for the distribution such as gamma distribution, which can have its kurtosis varied based on the shape parameter (i.e. $excess\ kurtosis = 6/\mathbb{a}$) (Mun, 2012). The flexibility of a gamma distribution could better capture the shape of the tail of $\mathbb{d}_{QTL}$, especially if $\mathbb{d}_{QTL}$ is a strongly leptokurtic distribution.

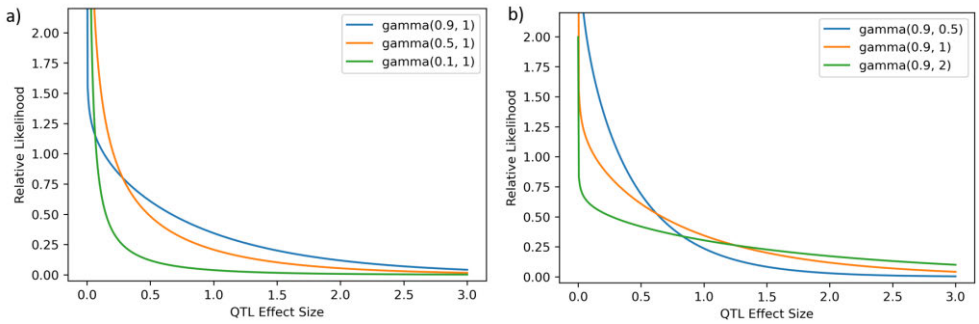

Figure B.7: The effects of varying values of (a) shape parameter $\mathbb{a}$ and (b) scale parameter $\mathbb{b}$ on the relative frequency of QTL effect sizes.

B.2.1. Effects of Number of QTL $\mathbb{k}$

The number of QTL associated with a trait has a significant impact on the $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$. Given a fixed narrow sense heritability h^2, shape and scale parameter as $\mathbb{a}$ and $\mathbb{b}$, a small $\mathbb{k}$ (i.e., oligogenic trait) reduces the variance and increases the kurtosis of $\mathbb{d}_{ES}^{1}$ (Figure B.8). In Figure B.8(a) the variance and excess kurtosis of $\mathbb{d}_{ES}^{1}$ are 0.16 and 7.03 respectively, compared to 1.61 and 6.14 in Figure B.8(b). The reduced proportion of additive genetic variance explained by a QTL in a polygenic trait (i.e. trait with large $\mathbb{k}$) also reduces the proportion of markers with top estimated effect sizes come from QTL; in Figure B.8(a), where only 200 QTL are associated with the trait, 14.40% of top 50 markers from each GWAS are QTL, compared to 6.54% in Figure B.8(b) with 2000 QTL.

Whereas for $\mathbb{d}_{FT}^{1}$, due to an increased proportion of variance being explained by a QTL in an oligogenic trait (i.e., trait with small $\mathbb{k}$), the test statistics deviate further from what is expected in null markers (i.e. $F_i = 0$), and this increases the kurtosis of $\mathbb{d}_{FT}^{1}$. The opposite is true for a polygenic trait. An example is provided in Figure B.9(a-b); in (a) where $\mathbb{k} = 200$ the $\mathbb{d}_{FT}^{1}$ has an excess kurtosis of 10881.14, compared to $\mathbb{k} = 2000$ in (b) which has an excess kurtosis of 737.44. Its effects are also observable from the Manhattan plot; with small $\mathbb{k}$, the Manhattan plots have few but strong and well-distinguished peaks and, if correlation

between markers is present, null markers flanking the peak (Figure B.9(c)); whereas for large $\mathbb{k}$, the Manhattan plot has numerous, but less well-defined peaks that could be difficult to be distinguished from noises (Figure B.9(d)).

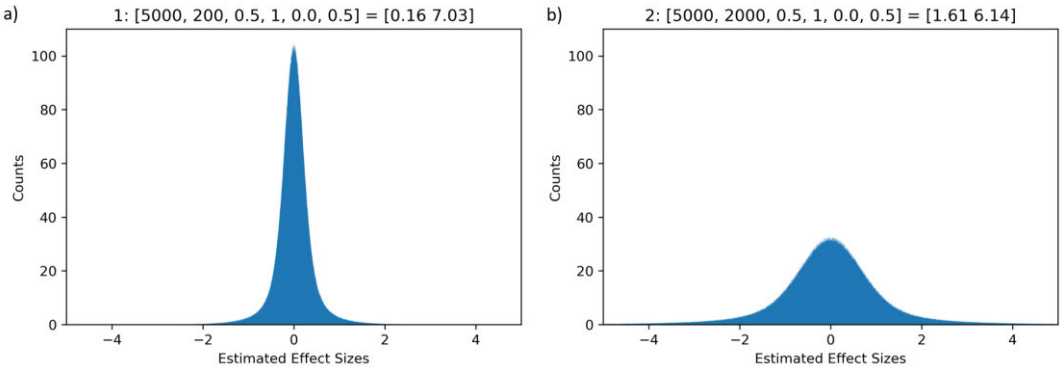

Figure B.8: Histogram showing the effects (in units of residual standard deviation) of number of QTL $\mathbb{k}$ on $\mathbb{d}_{ES}^{1}$, averaged from 100 GWAS experiments, with sample size of 5000 over 50k independent markers. In (a) 200 QTL were simulated, and (b) 2000 QTL were simulated. In all replicates, the distribution of QTL effect size is set with gamma distribution $\Gamma(0.5, 1)$, and allele frequency distribution followed a beta distribution $Beta(0.5, 0.5)$.

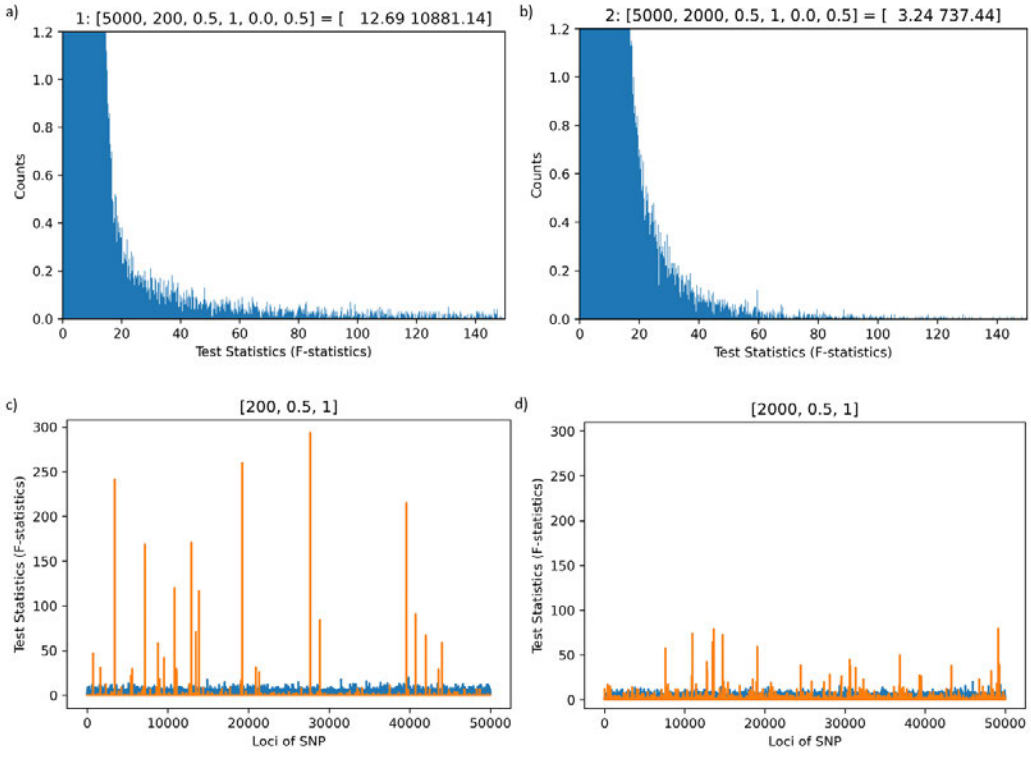

Figure B.9: The effects of number of QTL $\mathbb{k}$ on (a-b) $\mathbb{d}_{FT}^{1}$ and (c-d) the Manhattan plots of the GWAS. In (a-b) 100 GWASes with 50k independent markers and sample size of 5000 were simulated and have their histograms averaged. In (a) and (c) 200 QTL were simulated, and (b) and (d) 2000 QTL were simulated. In all simulations, the distribution of QTL effect size is set with gamma distribution $\Gamma(0.5, 1)$, and allele frequency distribution followed a beta distribution $Beta(0.5, 0.5)$.

B.2.2. Effects of Shape Parameter for the QTL Effect Size Distribution $\mathbb{a}$

Besides $\mathbb{k}$, the shape parameter $\mathbb{a}$ also has significant effects on the $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$. Provided all other parameters being kept constant, a more leptokurtic $\mathbb{d}_{QTL}$ (i.e., smaller $\mathbb{a}$) reduces the variance while increasing the kurtosis of $\mathbb{d}_{ES}^{1}$ (Figure B.10). In Figure B.10(a) the variance and excess kurtosis of the $\mathbb{d}_{ES}^{1}$ are 0.24 and 6.52 respectively, compared to 3.66 and 6.10 in $\mathbb{d}_{ES}^{1}$ featured in Figure B.10(b). This can be attributed to a reduced proportion of QTL with large effect size with smaller $\mathbb{a}$, an effect similar to those observed in small $\mathbb{k}$. The same effect also increases the proportion of top markers being QTL; in Figure B.10(a), where $\mathbb{a} = 0.1$, 13.16% of all top 50 markers originated from QTL, compared to 5.3% in Figure B.10(b) when $\mathbb{a} = 0.9$.

Whereas for $\mathbb{d}_{FT}^{1}$, a more leptokurtic $\mathbb{d}_{QTL}$ increases the kurtosis of $\mathbb{d}_{FT}^{1}$, which is observable as increased number of markers with large test statistics, and fatter tail of the test statistics distribution (Figure B.11(a-b)). In (a) where $\mathbb{a} = 0.1$ the $\mathbb{d}_{FT}^{1}$ has an excess kurtosis of 7037.54, compared to 177.34 in (b) where $\mathbb{a} = 0.9$. The leptokurtic QTL effect size distribution also resulted in stronger and well-defined peaks, similar to those observed in a trait with small $\mathbb{k}$ (Figure B.11(c-d)).

The similarity of effects of a genetic architecture with small $\mathbb{a}$ with those with small $\mathbb{k}$ highlighted one potential issue during the estimation the $\mathbb{d}_{QTL}$: since the distributions from a trait with small $\mathbb{a}$ are not distinguishable from those with small $\mathbb{k}$, any attempts that utilized the distributions might not be able to return a unique solution. This concern has indeed been validated through additional simulation, for which the results of simulation are provided in Figure B.12, where the distribution from a trait with small $\mathbb{k}$ but large $\mathbb{a}$ is practically indistinguishable from the distribution of a trait with small $\mathbb{a}$. Thus, if $\mathbb{d}_{FT}^{1}$ were used to estimate the $\mathbb{d}_{QTL}$, and no additional constraints are available, the solutions were not unique, and instead an infinite solution that described the relationship between $\mathbb{k}$ and $\mathbb{a}$ was obtained. Similar observations have also been made for $\mathbb{d}_{ES}^{1}$.

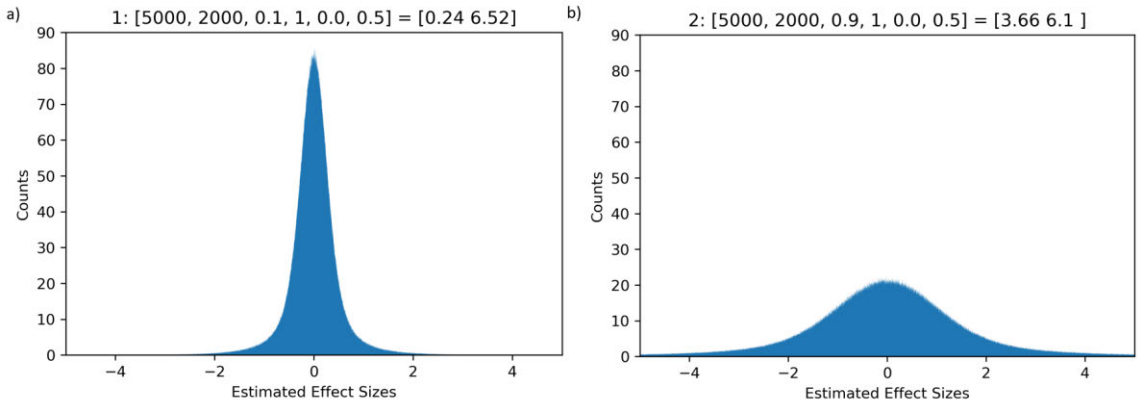

Figure B.10: Histogram showing the effects of shape parameter $\mathbb{a}$ on $\mathbb{d}_{ES}^{1}$, with the $\mathbb{d}_{QTL}$ in (a) followed the gamma distribution $\Gamma(0.1, 1)$ and in (b) $\Gamma(0.9, 1)$. The histograms are generated by averaging the histograms from 100 GWASes with 50k independent markers and sample size of 5000. In both scenarios, 2000 QTL has been simulated, and the allele frequency distribution followed a beta distribution $Beta(0.5, 0.5)$.

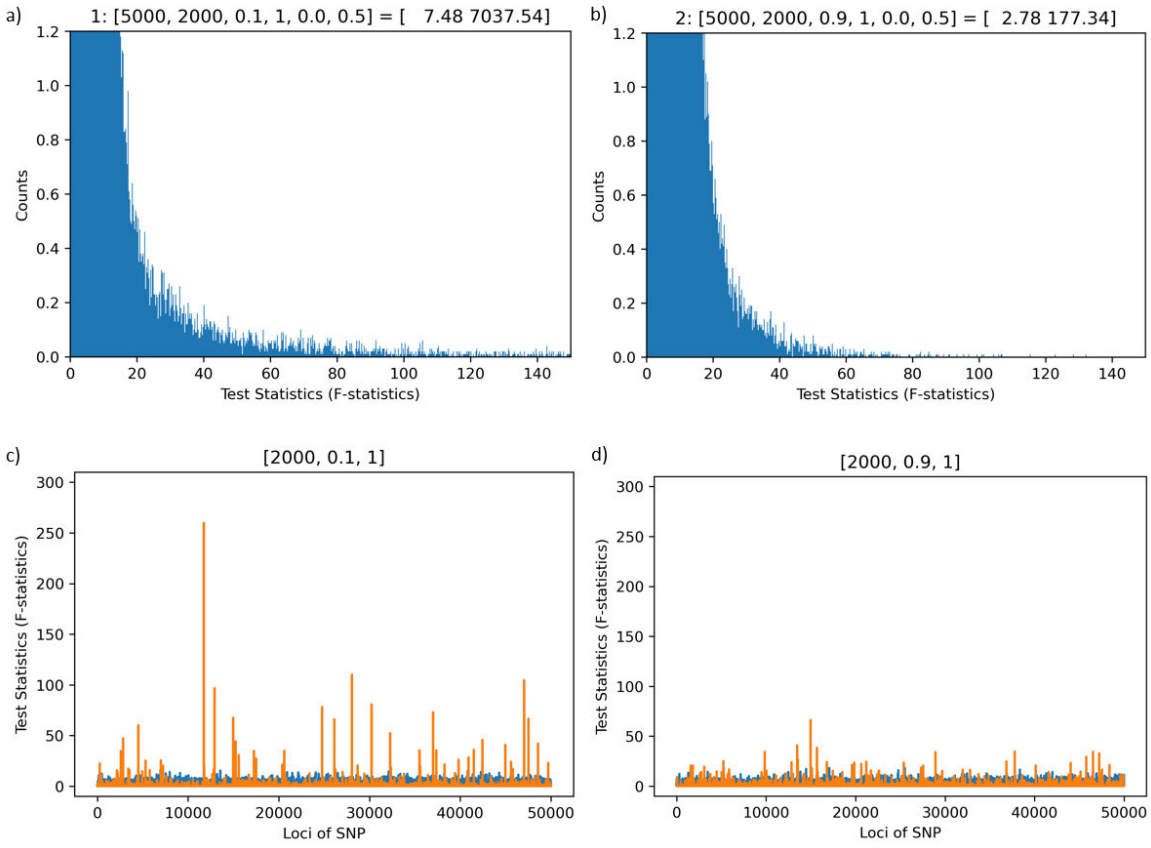

Figure B.11: Histograms showing the effects of shape parameter $\mathbb{a}$ of the $\mathbb{d}_{QTL}$ on (a-b) $\mathbb{d}_{FT}^{1}$ and (c-d) Manhattan plots. In Figure (a) and (c), the $\mathbb{d}_{QTL}$ had the shape parameter $\mathbb{a} = 0.1$, and (b) and (d) has shape parameter $\mathbb{a} = 0.9$. The histograms were generated by averaging 100 GWAS with sample size 5000 and 50k independent markers, and number of QTL $\mathbb{k} = 2000$ and scale parameter $\mathbb{b} = 1$. The allele frequency distribution followed a beta distribution $Beta(0.5, 0.5)$.

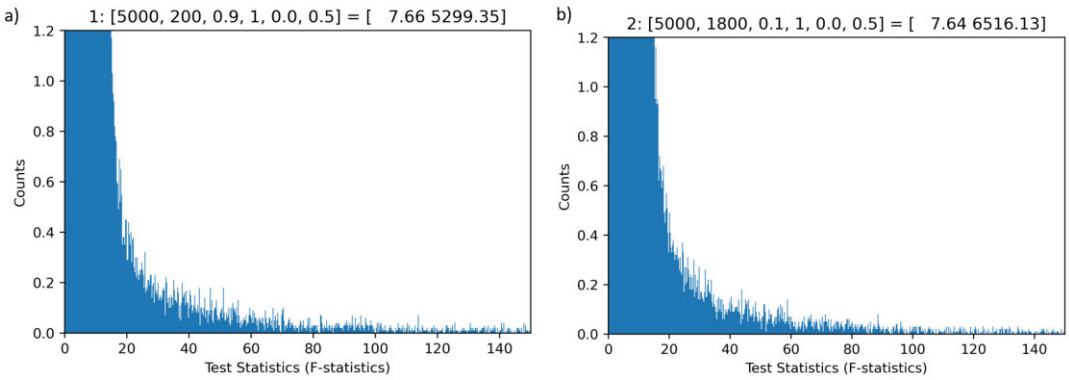

Figure B.12: Histograms showing the $\mathbb{d}_{FT}^{1}$ of (a) genetic architecture $Q(200, 0.9,1)$ and (b) genetic architecture $Q(1800, 0.1, 1)$. The figures were generated by averaging histograms from 100 GWAS with sample size of 5000 and 50k independent markers, and the allele frequency distribution followed a beta distribution $Beta(0.5, 0.5)$.

B.2.3. Effects of Scale Parameters for the QTL Effect Size Distribution $\mathbb{b}$

Unlike the parameter $\mathbb{k}$ and $\mathbb{a}$, which have effects on both $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$, the scale parameter for the genetic architecture $\mathbb{b}$ only affects $\mathbb{d}_{ES}^{1}$ (Figure B.13(a-b)). Provided all other parameters remained unchanged, $\mathbb{b}$ scales the variance of $\mathbb{d}_{ES}^{1}$ in the proportion of $\mathbb{b}^2$. From the example provided in Figure B.13(a) with the scale parameter $\mathbb{b} = 1$, the variance of $\mathbb{d}_{ES}^{1}$ is 1.58. Compared to 14.45 in Figure B.13(b) when $\mathbb{b} = 3$, this represents a scaling of variance by approximately 9 (14.45/1.58 = 9.15).

This observation suggested that, if the observed additive genetic variance $v_{\mathrm{A_{obs}}}$ can be calculated, and the value of $\mathbb{k}$ and $\mathbb{a}$ can be estimated, parameter $\mathbb{b}$ can be defined as follows:

$$\mathbb{b} = \sqrt{\frac{v_{\mathrm{A_{obs}}}}{v_{\mathrm{A}_{[\mathbb{k},\mathbb{a},\mathbb{b}=1]}}}} \qquad [6]$$

Where $v_{\mathrm{A}_{[\mathbb{k},\mathbb{a},\mathbb{b}=1]}}$ is the expected additive genetic variance if the $\mathbb{k}$ and $\mathbb{a}$ are as determined from calculation, but with $\mathbb{b}$ set as 1. It also implies that if the additive genetic variance can be determined, it reduces the number of parameters of the genetic architecture that need to be estimated by dropping the parameter $\mathbb{b}$, and only $\mathbb{k}$ and $\mathbb{a}$ need to be determined. Changing the $\mathbb{b}$ have no effects on the proportion of variance explained by each of the QTL, thus with no effects on the $\mathbb{d}_{FT}^{1}$ (Figure B.13(c-d)). This suggests that $\mathbb{d}_{FT}^{1}$ cannot be used to estimate parameter $\mathbb{b}$, and only $\mathbb{d}_{ES}^{1}$ and $v_{\mathrm{A_{obs}}}$ can be used for this estimation.

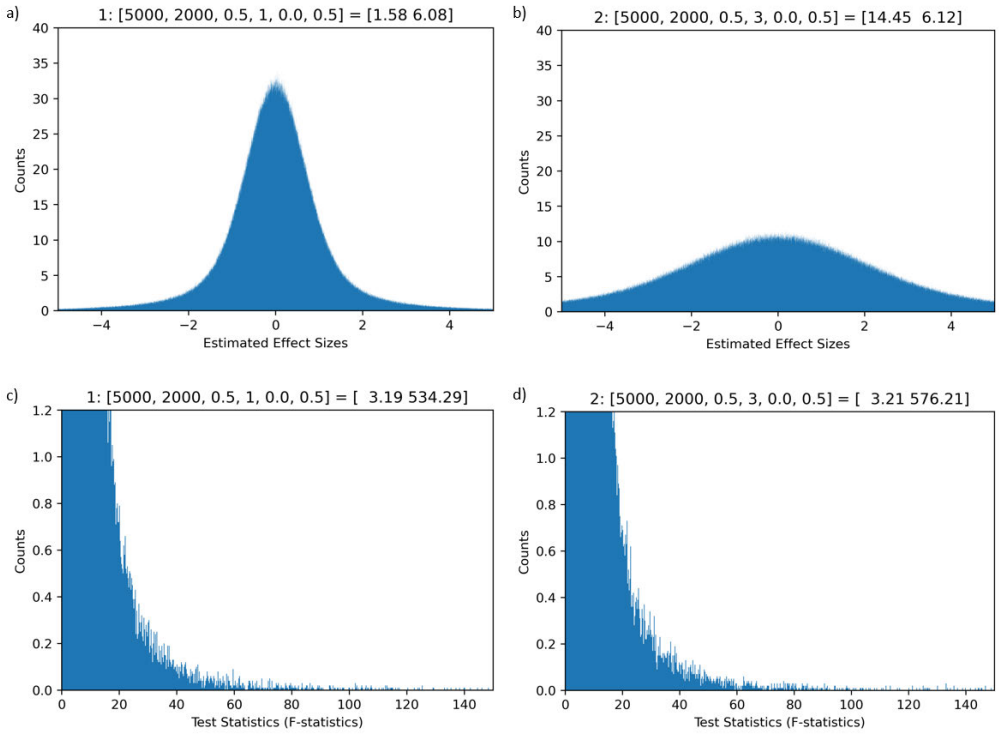

Figure B.13: Histograms showing the effects of scale parameters $\mathbb{b}$ on (a-b) $\mathbb{d}_{ES}^{1}$ and (c-d) $\mathbb{d}_{FT}^{1}$. Figure (a) and (c) has scale parameter $\mathbb{b} = 1$, while (b) and (d) has scale parameter $\mathbb{b} = 3$. The figures were generated by averaging histograms from 100 GWAS with sample sizes of 5000 and with 50k independent markers, with number of QTL set at 2000 and shape parameter of 0.1, and the allele frequency distribution followed a beta distribution $Beta(0.5, 0.5)$.

B.3. Confounding Factors that Affect $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$

Genetic architecture is not the only factor that affects the asymptotic distribution of $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$. They are confounded by numerous other factors, which could affect the estimation of the architecture parameters. This necessitates the investigation on the effects of these factors so that their effects could be taken into account during the estimation.

B.3.1. Allele Frequency Distribution

One of the factors that affects the $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$ is the distribution of the allele frequencies. Extreme allele frequencies reduce the GWAS's ability to accurately estimate the effect size of the marker and increases the error of effect size estimation. The errors in estimation could then be extended toward the overall distribution of the allele frequencies, which can be modelled with a symmetric Beta distribution (i.e. $Beta(x, x)$) where x is the shape parameter (Daetwyler et al., 2013). A genotype array with more loci with extreme allele frequencies produce a Beta distribution with smaller x (Figure B.14).

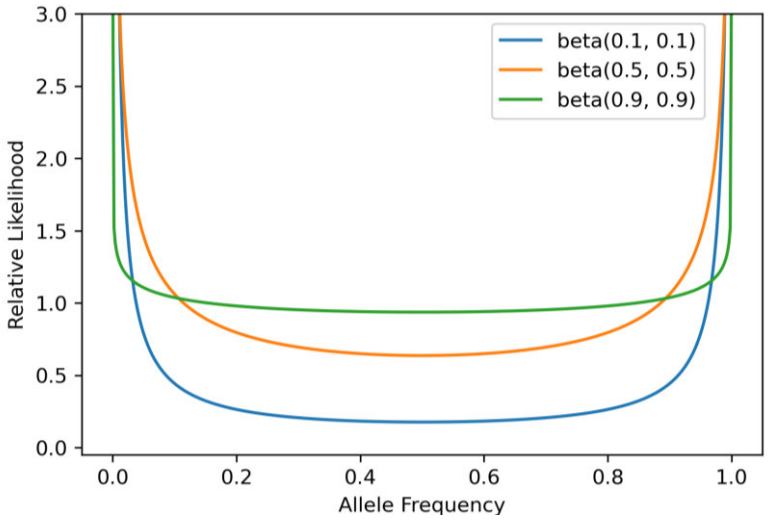

Figure B.14: The relative likelihood of allele frequency distribution under varying shape parameter for the symmetric Beta distribution.

A genotype array with smaller x has an increased proportion of loci with extreme allele frequencies, and this increases the variance of $\hat{a}$ and $\mathbb{d}_{ES}^{1}$. This increase in variance is illustrated in Figure B.15(a-b), where in (a) with allele frequency distribution $Beta(0.1, 0.1)$ the variance of $\mathbb{d}_{ES}^{1}$ is 2.01, compared to 1.18 in (b) where the allele frequency distribution is $Beta(0.9, 0.9)$.

This error of estimation not only obfuscate the $\mathbb{d}_{QTL}$, but it also drowns out the signals from small QTL near the head of the distribution, making them undetectable. With changing allele frequency distribution, such errors could be sufficiently strong that it affects the tail of $\mathbb{d}_{ES}^{1}$, which affects the proportion of top markers being QTL. In the example featured in Figure B.15(c), with the genotype array with its allele frequencies following the Beta distribution $Beta(0.1, 0.1)$, 7.90% of the top 50 markers in terms of estimated effect sizes from each GWAS came from QTL, compared to Figure B.15(d) where only 6.52% when the allele frequency following the Beta distribution $Beta(0.9, 0.9)$. This observation hinted that the shape of the tail of $\mathbb{d}_{ES}^{1}$ is confounded by the allele frequency distribution, making it an unreliable indicator on the underlying genetic architecture of the trait.

Due to the scaling down effects of $p(1-p)$ from extreme allele frequencies, markers with extreme allele frequencies do not have significant effects on the tail of $\mathbb{d}_{FT}^{1}$, and thus less vulnerable to the changing the allele frequency distribution (Figure B.16(a-b)). The proportion of top 50 markers in terms of test statistics also do not change significantly with varying allele frequency distribution; from the example provided in Figure B.16(c), where the

allele frequency distribution follows the Beta distribution $Beta(0.1, 0.1)$, 89.3% of the top 50 markers came from QTL, which is comparable with the 89.0% from Figure B.16(d) where the allele frequency distribution follows the Beta distribution $Beta(0.9, 0.9)$. The invulnerability of the tail of $\mathbb{d}_{FT}^{1}$ toward changing allele frequency distribution also hinted the desirability of using this distribution in the estimation of $\mathbb{d}_{QTL}$.

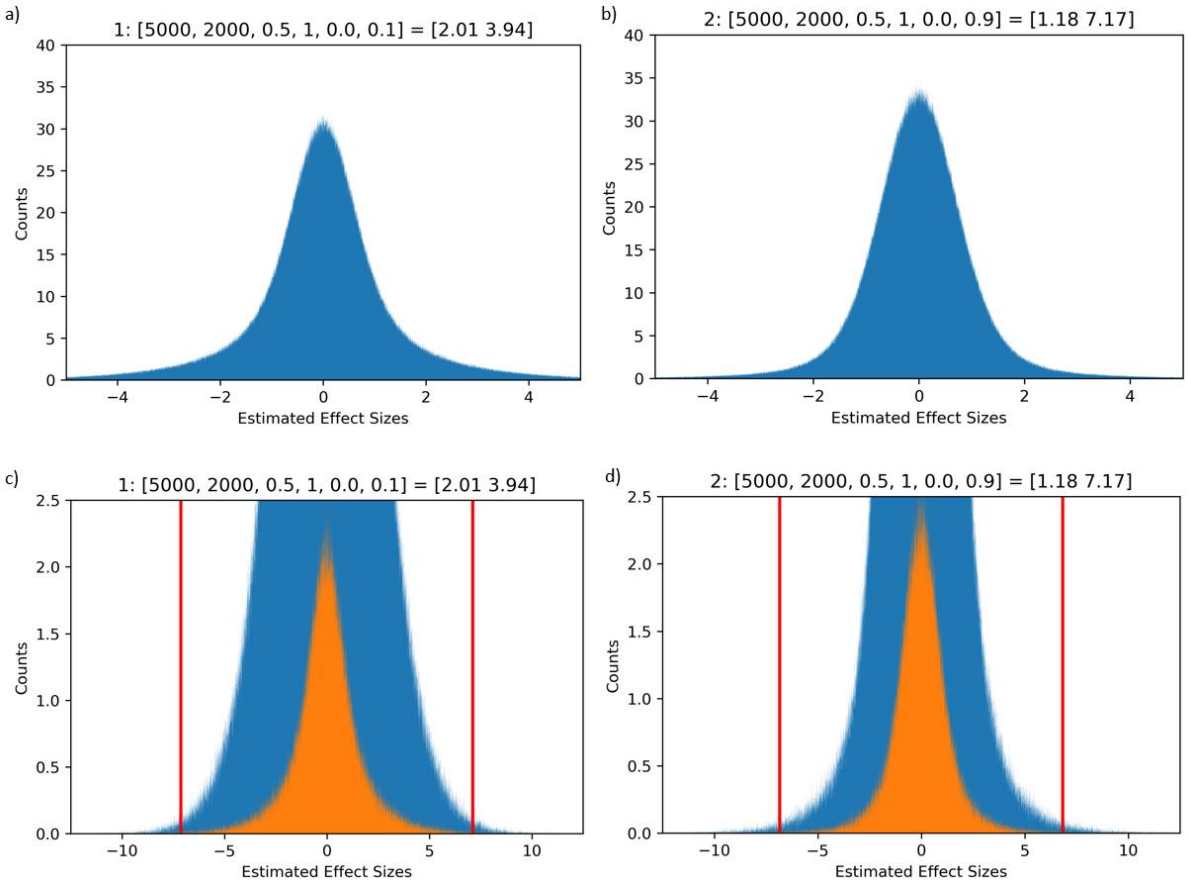

Figure B.15: Histograms showing the effects of allele frequency distributions on the (a-b) overall shape of $\mathbb{d}_{ES}^{1}$, and (c-d) the distribution of estimated effect sizes of the null markers (blue bars) and non-null markers (orange bars), with the red lines indicating the top 0.1% of all markers in term of estimated effect sizes. In Figure (a) and (c) the allele frequency distributions follow the Beta distribution $Beta(0.1, 0.1)$ and (b) and (d) follow the $Beta(0.9, 0.9)$. The figures were generated by averaging histograms from 100 GWAS with sample size of 5000 and with 50k independent markers. The genetic architecture parameters in all figures were set at $Q(2000, 0.5, 1)$.

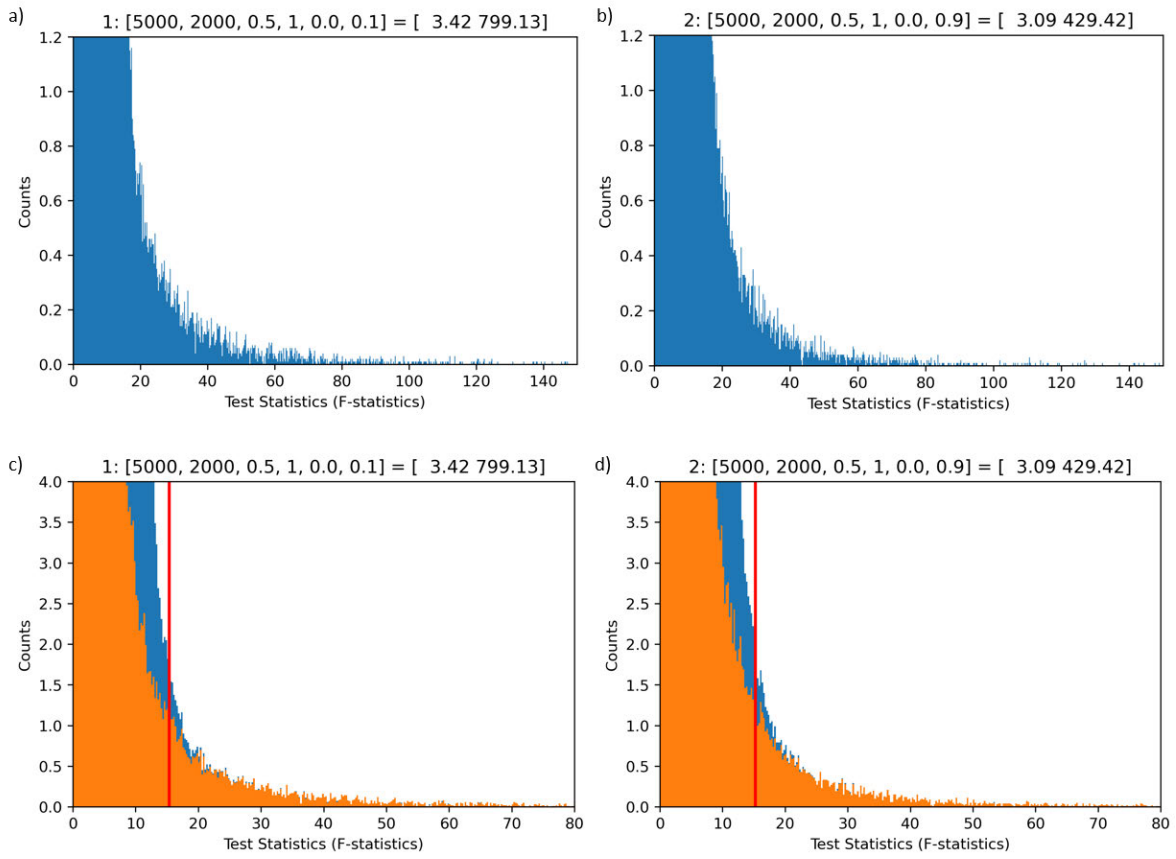

Figure B.16: The effects of allele frequency distributions on (a-b) overall shape of $\mathbb{d}_{FT}^{1}$ and (c-d) the distribution of test statistics of the null markers (blue bars) and non-null markers (orange bars), with the red lines indicating the top 0.1% of all markers in terms of test statistics. In Figure (a) and (c) the allele frequency distributions follow the Beta distribution $Beta(0.1, 0.1)$ and (b) and (d) follow the $Beta(0.9, 0.9)$. The figures were generated by averaging histograms from 100 GWAS with sample size of 5000 and with 50k independent markers. The genetic architecture parameters in all figures were at $Q(2000, 0.5, 1)$.

B.3.2. Sample Size

Besides the allele frequencies, sample size also affects the $\mathbb{d}_{ES}^{1}$. Due to an increased variance in the distribution of $\hat{a}$ from equation [1], the variance of the $\mathbb{d}_{ES}^{1}$ increased significantly with decreased sample size. This phenomenon is illustrated in Figure B.17(a-b); in (a) where the sample size utilized is 5000, the variance of $\mathbb{d}_{ES}^{1}$ is 1.62, compared to (b) where the sample size is 2000, the corresponding variance is 4.04. The increased estimation error has further drowned out signals from QTL and obfuscate the underlying $\mathbb{d}_{QTL}$. Indeed, from the top 50 markers as presented in Figure B.17(c), 6.72% of the markers originate from true QTL if the sample size is 5000, compared to 4.52% for sample size of 2000 in Figure B.17(d). This suggested the susceptibility of $\mathbb{d}_{ES}^{1}$ toward small sample size, which further confound with the effects of allele frequency distribution.

Decreasing the sample size also reduces the magnitude of the test statistics, thus scaling down the variance of $\mathbb{d}_{FT}^{1}$. The decreased magnitude of test statistics also reduces the proportion of test statistics with extreme values, thus decreasing the kurtosis of $\mathbb{d}_{FT}^{1}$. This phenomenon is illustrated in Figure B.18(a-b); in (a) where the sample size utilized is 5000, the excess kurtosis of $\mathbb{d}_{FT}^{1}$ is 545.66, compared to (b) where the sample size is 2000, the corresponding excess kurtosis is 51.92. A reduced sample size also decreases the proportion of top markers being true QTL; in the example featured in Figure B.18(c), 88.2% of the top 50 markers originate from true QTL when sample size is 5000, compared to 44.0% for sample size of 2000 in Figure B.18(d). This observation supported the previous reports on increased power in GWAS due to an increased sample size (Spencer et al., 2009), and also hinted that while the tail of the $\mathbb{d}_{FT}^{1}$ is susceptible to small sample size, placing a lower limit on the sample size for a feasible estimation of $\mathbb{d}_{QTL}$.

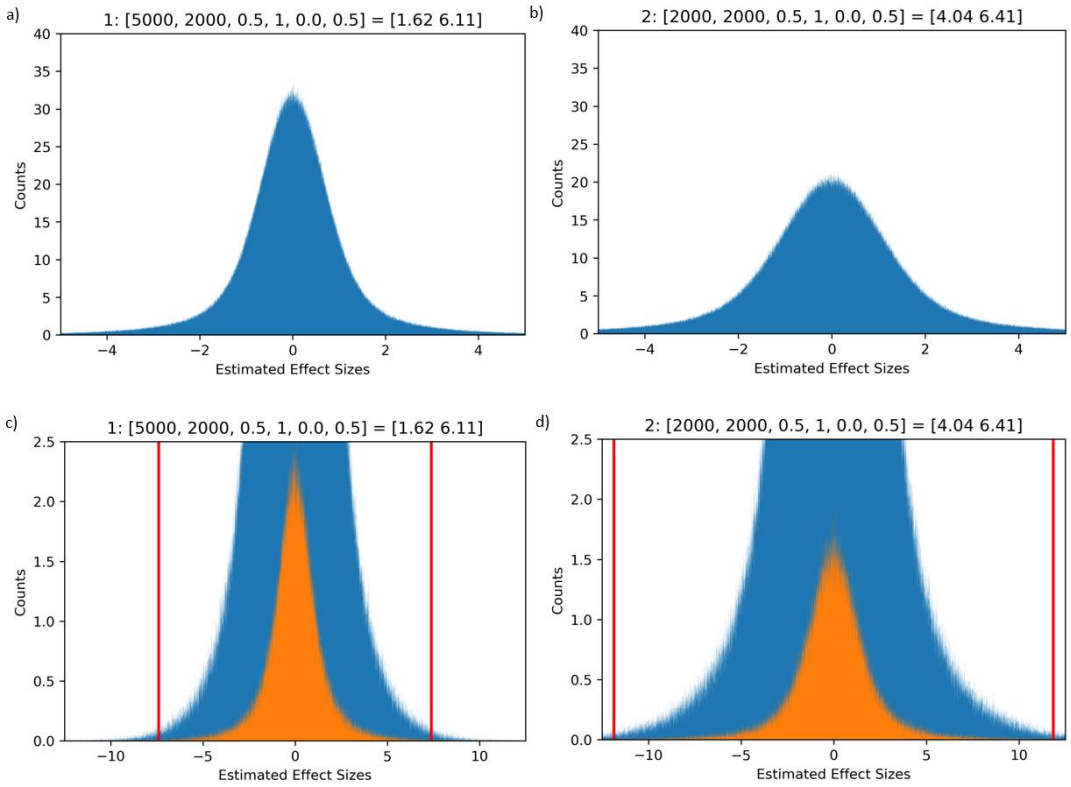

Figure B.17: Histograms showing thee effects of sample size used in the GWAS on $\mathbb{d}_{ES}^{1}$, with (a-b) showing the overall shape of the distributions, and (c-d) the distribution of estimated effect sizes of the null markers (blue) and non-null markers (orange), with red lines indicating the top 0.1% of all markers in term of estimated effect sizes. Figure (a) and (c) was simulated with sample size of 5000, and (b) and (d) with sample size of 2000. The figures were generated by averaging histograms from 100 GWAS with 50k independent markers that have their allele frequency distribution of $Beta(0.5, 0.5)$, and the genetic architecture parameters in all figures were set at $Q(2000, 0.5, 1)$.

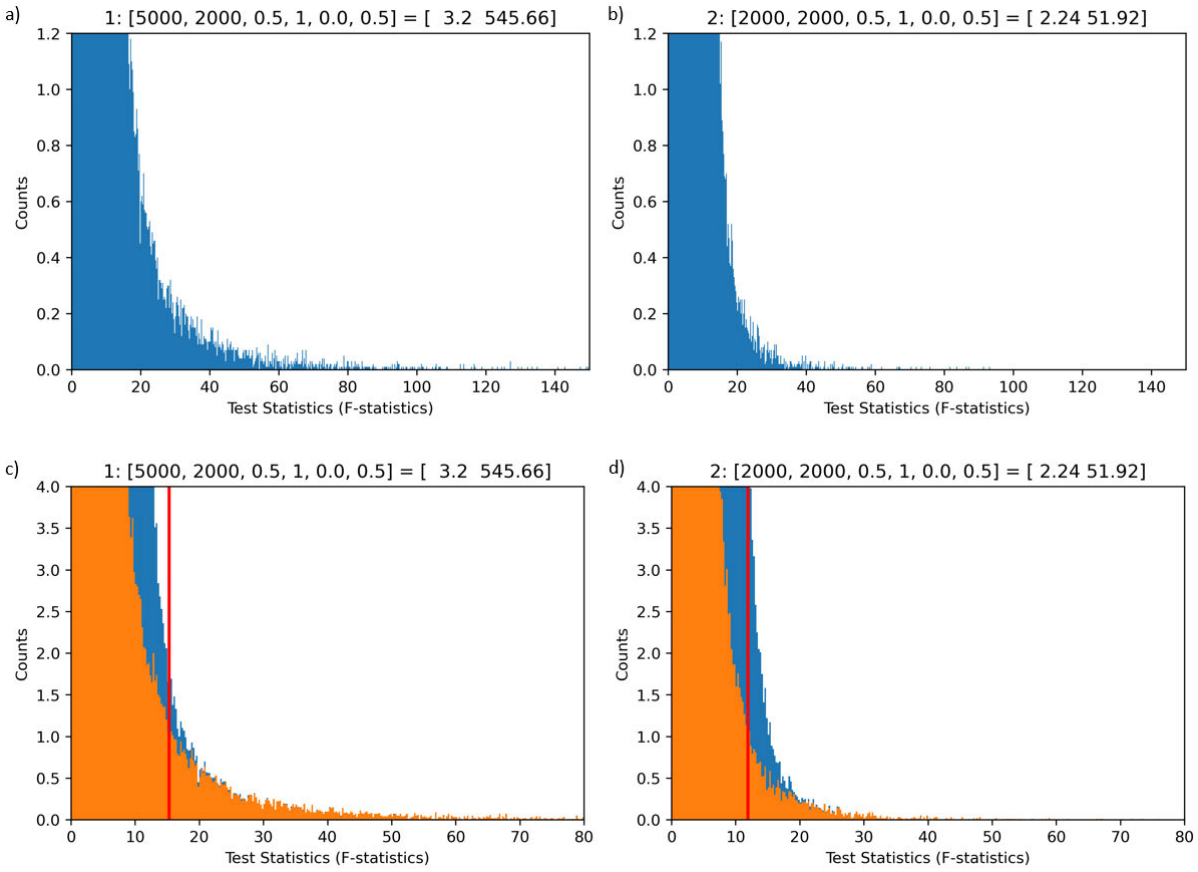

Figure B.18: Histogram showing the effects of sample size of (a-b) the overall shape of $\mathbb{d}_{FT}^{1}$ and (c-d) the proportion of null markers (blue) and non-null markers (orange),with red lines indicating the top 0.1% of all markers in term of test statistics. Figure (a) and (c) was simulated with sample size of 5000, and (b) and (d) with sample size of 2000. The figures were generated by averaging histograms from 100 GWAS with 50k independent markers that have their allele frequency distribution of $Beta(0.5, 0.5)$, and the genetic architecture parameters in all figures were set at $Q(2000, 0.5, 1)$.

B.3.3. Correlation Between Markers

Correlation between markers also has significant impact on these distributions. As correlation introduced a degree of similarity in the genotype between two markers, if one of the markers is correlated with a QTL, the other marker would also exhibit some of the effect size from the QTL. This creates a characteristic "bleeding" effect on the Manhattan plot where a peak of QTL is flanked by several null markers (an example provided in Figure 3.7). The effect size experienced by a null marker that correlates with a QTL can be calculated as follows:

$$\hat{a}_i = R_{LD_{i,j}} * a_{QTL} + \epsilon \qquad [7]$$

Where $R_{LD_{i,j}}$ is the correlation coefficient between loci i and j, and ϵ is the error of estimation of loci i. The $R_{LD_{i,j}}$ can be calculated as follows:

$$R_{LD_{i,j}} = \frac{cov(X_i, X_j)}{var(X_i) * var(X_j)} \quad [8]$$

When several QTL correlates with one another, their estimated effect size interacted additively as follows:

$$\hat{a}_i = a_i + \sum_{j=1}^{nQTL} R_{LD_{i,j}} * a_j + \epsilon \quad [9]$$

The impact of correlation on the estimation of QTL effect sizes are extensive. Due to the additional terms, the expected values for the estimated effect sizes no longer correspond to those of the original QTL effect size, but instead approach $\hat{a}_i$ as defined in equation [9]. This suggests that if correlation between markers is present, the distribution of estimated effect size can no longer be directly used as the underlying QTL effect sizes distribution, even if error in estimation can be avoided.

The additional terms also increase the variance of the estimated effect sizes (Figure B.19(a-b)); in (a), where the markers are independent, the variance of the estimated effect size is 0.33, whereas in (b) where the pairwise marker correlation $R_{LD_{i,j}} = 0.98$, the variance increases to 1.85. The presence of correlation also altered the kurtosis of the distribution; in (a) the excess kurtosis is 0.07, which is not significantly different compared to a normal distribution, whereas in (b) the excess kurtosis is 2.48.

Due to the shifted expected value and increased variance of the estimated effect sizes of the markers, the resulting distribution of estimated effect size $\mathbb{d}_{ES}^1$ would have an increased variance and decreased kurtosis (Figure B.20(a-b)); in (a), where no correlation between markers is present, the variance and excess kurtosis of the $\mathbb{d}_{ES}^1$ are 1.57 and 6.11, respectively, whereas in (b) where the pairwise marker correlation $R_{LD_{i,j}} = 0.98$, the corresponding variance and excess kurtosis are 3.03 and 2.44 respectively. Similar changes can also be observed in the $\mathbb{d}_{FT}^1$, where in Figure B.20(a) the variance and excess kurtosis of $\mathbb{d}_{FT}^1$ are 3.15 and 488.54, compared to Figure B.20(b) with pairwise marker correlation $R_{LD_{i,j}} = 0.98$ the corresponding variance and excess kurtosis are 54.68 and 70.59 respectively.

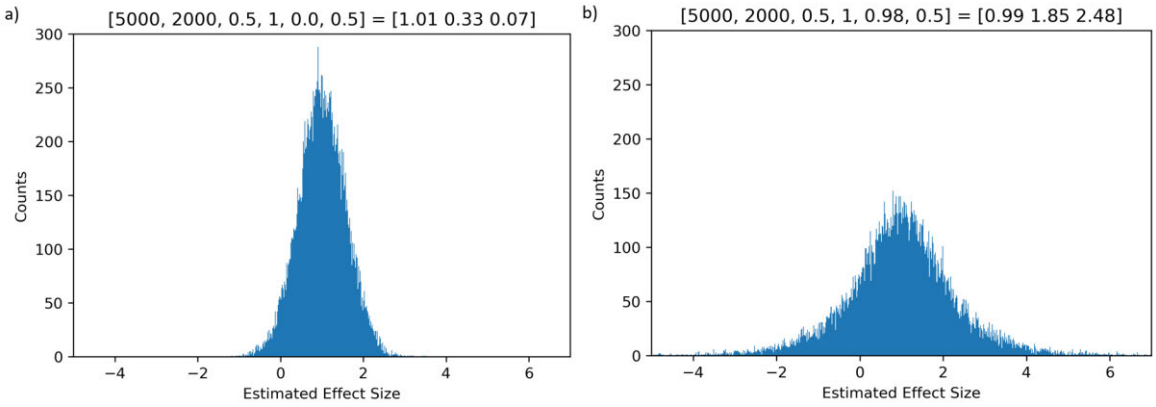

Figure B.19: Histogram showing the distribution of estimated effect size of a QTL under (a) independent markers and (b) average pairwise correlation of 0.98, with genetic architecture parameter $Q(2000, 0.5, 1)$. The distributions of estimated effect sizes featured in both figures have a true effect size is 1.0 σ_e, and the GWAS is generated with allele frequency of 0.5, sample size of 5000 and phenotypic variance set at 750.

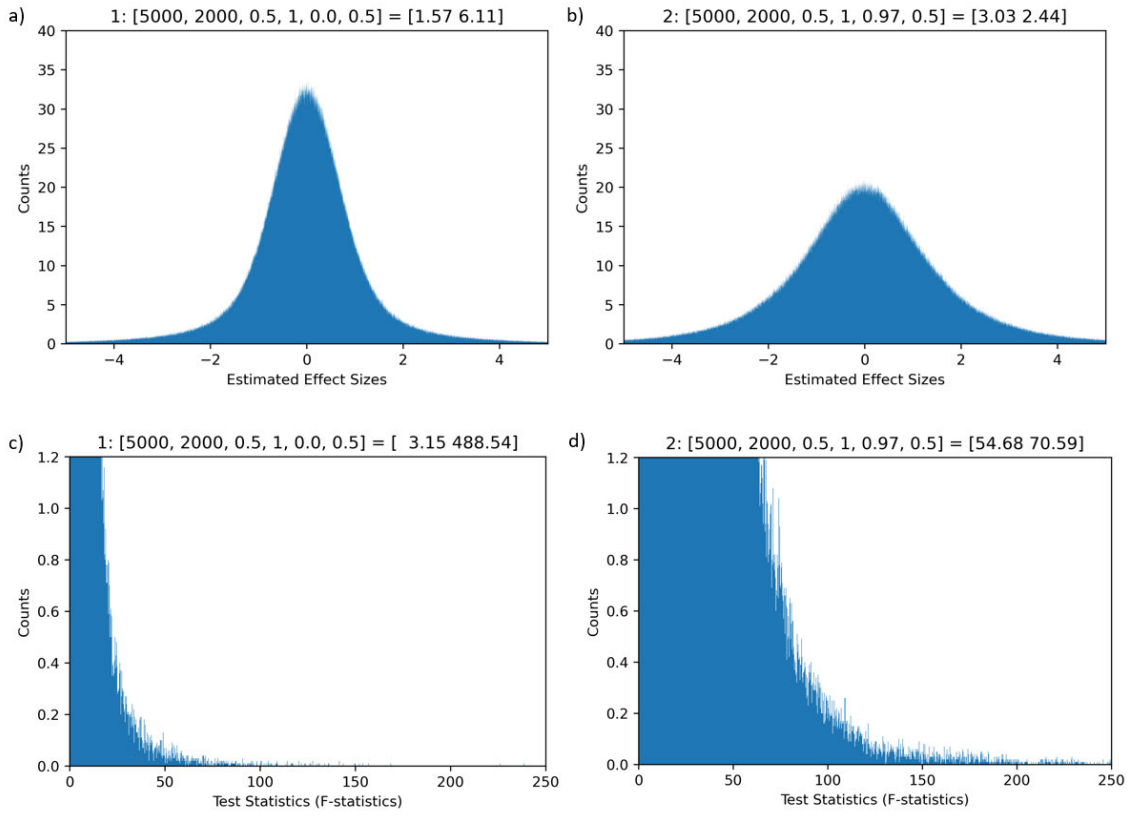

Figure B.20: The effects of correlation between markers in (a-b) the $\mathbb{d}_{ES}^{1}$ and (c-d) $\mathbb{d}_{FT}^{1}$, with (a) and (c) being the distribution if the markers are independent, and (b) and (d) if the pairwise marker correlation is set at 0.97. In all figures, the sample size of GWAS was set at 5000, and number of markers is 50k. The number of QTL in all figures was set at 2000, with $\mathbb{d}_{QTL} \sim (0.5, 1)$. The allele frequency distribution follows a symmetric Beta distribution $Beta(0.5, 0.5)$. 100 replicates were generated, and the figures were generated by averaging the histograms from each of the replicates.

Besides altering the shape of $\mathbb{d}_{FT}^{1}$ and $\mathbb{d}_{ES}^{1}$, correlation between markers also limits a GWAS's ability to uniquely identify the number of QTL under a peak. It means given a peak observed in a Manhattan plot, it could be caused by few QTL with large effect sizes, or by numerous QTL with small effect sizes, and it might not be observable through estimated effect size alone, even with high density markers such as whole-genome sequence data (Figure B.21). While algorithms that "de-correlate" the estimate effect sizes are available (as an example, pruning of markers), the number of QTL is no longer recoverable as there is no way of identifying which markers contributed to the peaks.

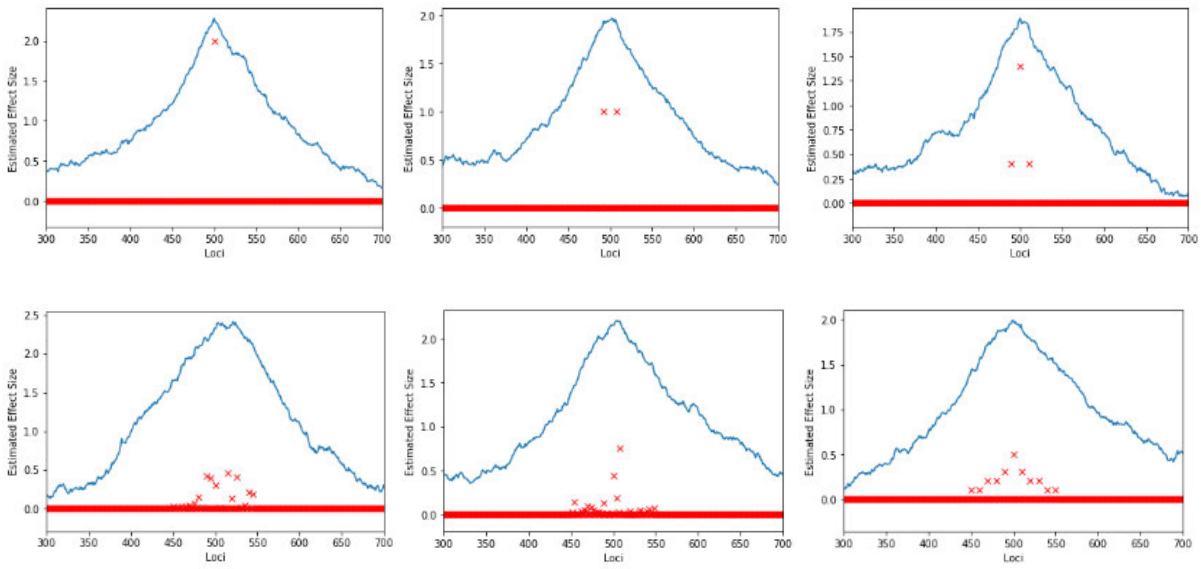

Figure B.21: The numerous possibilities of the underlying QTL effect size distribution (red crosses) given a peak being observed in the estimated effect sizes of a GWAS experiment (blue line). In all these plots, the pairwise marker correlation is set at 0.98.

This observation also suggested an ambiguity in identifying the mechanism of peak formations in the Manhattan plots of a GWAS. Given a peak observed in a Manhattan plot, there are two extremes of mechanism on how the peak is formed; the first mechanism is the "effect size peak" where the peaks are created by one or a few QTL with large effect sizes. The second mechanism is "QTL density peak" where the peak is caused by many QTL with small effect size being correlated to one another, meanwhile being exceptionally dense under the peak. For any given peak, it is likely that the mechanism lies somewhere between the extremes, although in general, a Manhattan plot with more "effect size peaks" tend to have smaller $\mathbb{k}$ and larger $\mathbb{a}$, and the opposite is true for those with more "QTL density peaks". In the context of architecture parameter, the proportion of QTL associated with the type of

peaks can be captured by parameter $\mathbb{a}$; a smaller $\mathbb{a}$ indicates a larger proportion of the QTL will be associated with the QTL density peak, and vice versa.

B.3.4. Other Confounding Factors

Other confounding factors that could affect the estimation of $\mathbb{d}_{QTL}$ include heterogeneity in linkage disequilibrium structures in the genome which could produces false positives in a GWAS experiment (Kaler and Purcell, 2019). Population stratification has also been reported as a source of false positives in GWAS through non-random segregation of genotypes based on the subpopulation structures (Panagiotou and Ioannidis, 2012). Finally, as previously reported in several publications such as Beavis (1994) and Hall et al. (2016), GWAS have a tendency to overestimate the QTL effect sizes, especially for the QTL with small effect sizes, which could also impact the estimation of $\mathbb{d}_{QTL}$ if $\mathbb{d}_{ES}^{1}$ is used. Further study could be dedicated to investigating the effects of these additional confounding factors on $\mathbb{d}_{ES}^{1}$ and $\mathbb{d}_{FT}^{1}$.

B.3.5. Conclusion

In conclusion, there are numerous confounding factors that could affect the estimation of $\mathbb{d}_{QTL}$. Overall, it appears that $\mathbb{d}_{FT}^{1}$ is less vulnerable toward these confounding factors compared to $\mathbb{d}_{ES}^{1}$, with the former being less vulnerable to allele frequency distribution that favoured extreme frequencies. This improves the signal-to-noise ratio in the tail of $\mathbb{d}_{FT}^{1}$ compared to $\mathbb{d}_{ES}^{1}$, suggested that the former is preferable for the estimation of the genetic architecture parameters such as number of QTL. Despite this, the vulnerability of both $\mathbb{d}_{FT}^{1}$ and $\mathbb{d}_{ES}^{1}$ toward small sample size placed a lower limit on the sample size where the estimation of $\mathbb{d}_{QTL}$ remained feasible (i.e., signals from QTL not being drowned out by noises from null markers). Even with smaller sample sizes however, the signal-to-noise ratio remained higher in the tail of $\mathbb{d}_{FT}^{1}$ compared to $\mathbb{d}_{ES}^{1}$, thus suggested the former being preferable for such estimation.

The higher signal-to-noise ratio in the tail of $\mathbb{d}_{FT}^{1}$ also suggested an increased proportion of top markers in term of test statistics being actual QTL compared to $\mathbb{d}_{ES}^{1}$, as well as reduced amount of noises from the null markers. This suggested that if the optimization of breeding pair is to be done using genomic information, test statistics would be the more reliable method of scoring for the additive genetic component of the offspring.

Appendix C. Test Statistics for Equality between Distributions of GWAS

The aim of this appendix section is to provide a layout of statistical tests that were utilized in the estimation of genetic architecture parameters. These tests are chosen or tailored to detect the effects of changing genetic architectures on the distributions of the test statistics from a GWAS experiment, with the null hypothesis being the expected distribution from a proposed set of genetic architecture parameters (denoted under the notation $[\mathbb{k}, \mathbb{a}, \mathbb{b}]$) being equal to the observed distribution from a GWAS experiment.

C.1. Properties Required for the Test Statistics in the Estimation of Genetic Architecture Parameters

While the genetic architectures such as number of QTL and shape of underlying QTL effect size distribution (denoted as $\mathbb{d}_{QTL}$) have detectable influences on the distribution of estimated effect sizes ($\mathbb{D}_{ES}^{1}$) and test statistics ($\mathbb{D}_{FT}^{1}$) from a GWAS experiment, there are several properties of the distributions that need to be taken into account when choosing a test for their equality.

One important aspect that needs to be taken into account being the difficulty of having the distributions to be defined mathematically. As the distribution is built from data of null and non-null marker, the resulting distributions are mixture distributions represented by a finite amount of data, thus are not smooth and not differentiable. This would call for statistics that do not make assumptions on the underlying distributions, such as nonparametric statistics (Cirrone et al., 2004; Fagerland, 2012). These nonparametric statistics work by detecting discrepancies in a multitude of aspects of a distribution, such as their locations and dispersion, which can then be used to test the equality of the distributions (Fagerland, 2012; Hart, 2001).

The signals from changing $\mathbb{d}_{QTL}$ on $\mathbb{D}_{FT}^{1}$ are miniscule but strongly concentrated at the tail region. This observation suggested that any statistics that are to be used in this algorithm need to be powerful at detecting the discrepancy between two distributions at the tail region. This requirement has, however, disqualified large number of statistics that potentially could be

utilized, as many of the statistics such as Kolmogorov-Smirnov test and Cramer von Mises test are sensitive at the median of the distribution and weak at the tail region (Lanzante, 2021; Mason and Schuenemeyer, 1983). This is the region of distribution that can be easily dominated by noise, which could easily drown out the signal from changing $\mathbb{d}_{QTL}$.

This could be partially alleviated through several methods. One of the simplest methods is to truncate $\mathbb{D}_{FT}^{1}$ at the tail region. By focusing the statistics on the tail region, any discrepancies at this region can be reliably detected, rather than being drowned out by noises near the head of $\mathbb{D}_{FT}^{1}$. An additional benefit of truncation of $\mathbb{D}_{FT}^{1}$ is the improvement of computational speed; as truncation removes a large portion of data points, the amount of calculation required decreases significantly, thus with added benefit of increasing the feasibility of the algorithm. Despite this, care needs to be taken to avoid excessive truncation as this could lead to overly small number of data points that reduce the reliability of the statistics.

An issue that should be noted is these modifications break the original validity of the tests, and thus no p-values that can be calculated, and no acceptance or a rejection of a model can be made. Despite this, minimization of the test statistics between a sequence of expected distributions of test statistics (denoted as $\mathbb{D}_{FT_{sim}}^{2}$) and a sequence of observed distribution of test statistics (denoted as $\mathbb{D}_{FT_{obs}}^{2}$) is still possible, and this would occur if the proposed genetic architecture parameters for $\mathbb{D}_{FT_{sim}}^{2}$ matches with true genetic architecture parameters (denoted under the notation $Q(\mathbb{k}, \mathbb{a}, \mathbb{b})$) from $\mathbb{D}_{FT_{obs}}^{2}$.

Utilizing the tail of $\mathbb{D}_{FT}^{1}$ had also entailed its own difficulty as well; with less data available at the tail of the distribution, it is more vulnerable toward the dispersion of $\mathbb{D}_{FT}^{1}$. This reduces the robustness of any statistical tests, and the presence-absence of a data point has greater effects if the data point is located at the tail region compared to those located at the head region. As an example, if equality in kurtosis is to be used to test the equality of two distribution, adding data points at the tail of the distribution has stronger influences on the test compared to adding data points to the head of the distribution (Figure C.1). If the added data points came from false positives, this could muddle the test statistics.

This issue could be partially alleviated in several ways, with the resampling of genotype and phenotype data being one such way. Besides this resampling, the use of a multitude of statistics that capture different aspects of the distributions, such as the shape of the distribution and maximal points in the distributions, could also be conducted. That way, if

one of the statistics failed (for example, the statistics are undefined, or returned anomalous results) other statistics can be used to supplement or back up the failures, thus improving the robustness of the algorithm. The use of multiple types of statistics has the added bonus of improving the accuracy of statistics that would otherwise be biased, and allowing an alternative formulation of "goodness of fit" between two distribution by defining it as the number of statistics that the model has successfully minimized (as an example, a model that minimizes seven out of ten statistics is a better fit than a model that minimizes two). It is also noted that the use of large number of statistics increases the chance of false positives however, thus the use of p-values should be avoided in this estimation, and the sole focus shall be the minimization of the test statistics.

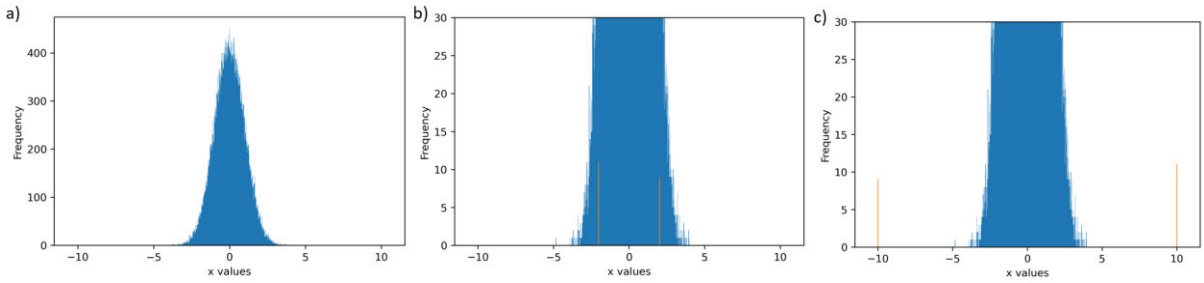

Figure C.1: The effects of additional data points at different part of the distribution on the kurtosis of the distribution. Figure (a) illustrated a histogram from 10,000 normal random variates, with excess kurtosis of -0.016. Figure (b) illustrated the histograms from the same set of random variates but with 20 additional points (denoted as orange bars), sampled randomly between -3.0 and 3.0, were included. The excess kurtosis for this distribution is -0.018. In Figure (c) the histogram also has the same set of random variates but with 20 additional points, also denoted as orange bars, sample randomly between -10 and 10, been included. The excess kurtosis for this set of random variates is 3.421.

The use of a sequence of $\mathbb{D}_{FT}^{1}$s (in the form of $\mathbb{D}_{FT}^{2}$) rather than one singular $\mathbb{D}_{FT}^{1}$ also has the added benefit of introducing more types of statistics that could be tested. This could be attributed to the emergent properties from a sequence of $\mathbb{D}_{FT}^{1}$ compared to a sequence of random variables. One such example is the convergence of maximum values of the random variables. It is known that the maximum values of the random variables are heavily affected by the outliers, and any statistics that utilized such values would have low robustness and poor reliability (Nazir, 2014). Conversely if the maximum values are collected among the random variables from each of the $\mathbb{D}_{FT}^{1}$ within a $\mathbb{D}_{FT}^{2}$, these maximum values converge toward a new distribution, as described by Fisher-Tippett-Gnedenko theorem (Charras-Garrido and Lezaud, 2013; Fisher and Tippett, 1928), which could then be used to test the

equality of $\mathbb{D}^2_{FT}$. This allows the development of new statistics that could be done to test the equality of $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$.

To further increase the number of tests available, this algorithm would repeat the tests but with varying the options for the tests. An example of the varying options is changing the tail cut-off point at, for example, top 1%, 0.5%, 0.2% and 0.1%, and then run the statistics at these varying cut-off points. This would also alleviate issues that might arise due to ambiguity of the optimal cut-off point for the statistics; while the signal-to-noise ratio increases further into the tail of the distribution, a cut-off point too far into the tail reduces the reliability of the test. Using a multitude of cut-off points alleviate this ambiguity.

C.2. Tests Utilized in Genetic Architecture Parameters Estimation

C.2.1. Maximal Distance Statistics

This class of statistics tests the equality of two distributions by testing the one-dimensional maximal distance between them. The rationale for this class of statistics is that if the two distributions came from the same underlying distribution, the maximal distance between them is asymptotically minimized (Cirrone et al., 2004; Simard and L'Ecuyer, 2011).

C.2.1.1. Kolmogorov-Smirnov Test

Perhaps the most well-known test statistic for maximal distance is the Kolmogorov-Smirnov test, which the calculation of test statistics has been defined in equation [1] in Chapter 5 (Cirrone et al., 2004; Lanzante, 2021; Stephen, 1970). It can be defined as the maximum difference between two empirical distribution functions (ECDF) along the y-axis (Figure C.2). One commonly raised shortcoming for the test is its lack of power in detecting discrepancies at the tail of the distribution (Cirrone et al., 2004; Lanzante, 2021). For this, adjustment needs to be made to improve the Kolmogorov-Smirnov test at the tail of $\mathbb{D}^1_{FT}$.

One of the ways to improve the power of Kolmogorov-Smirnov test at the tail of $\mathbb{D}^1_{FT}$ is to truncate the distribution at the tail region, which is defined using a y-axis cut-off point y_c. Using y_c, the x-axis cut-off point (denoted as x_c) is defined as the minimum between the arguments on $\mathbb{D}^1_{FT}$ where $\mathbb{D}^1_{FT} = y_c$:

$$x_c = \min\left(\left(\mathbb{D}^1_{FT_{sim}} : \mathbb{D}^1_{FT_{sim}} = y_c\right), \left(\mathbb{D}^1_{FT_{obs}} : \mathbb{D}^1_{FT_{obs}} = y_c\right)\right) \quad [1]$$

Using x_c, the test statistics for this truncated Kolmogorov-Smirnov test (denoted as $t_{KS_{trc}}$) is defined as the maximal distance between two scaled empirical CDFs (denoted as $MECDF$s) at xs larger than x_c:

$$t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right) = \sup\left|\mathbb{D}^1_{FT_{sim}}(x \geq x_c) - \mathbb{D}^1_{FT_{obs}}(x \geq x_c)\right| \quad [2]$$

An example of this truncated Kolmogorov-Smirnov test is provided in Figure C.2(b).

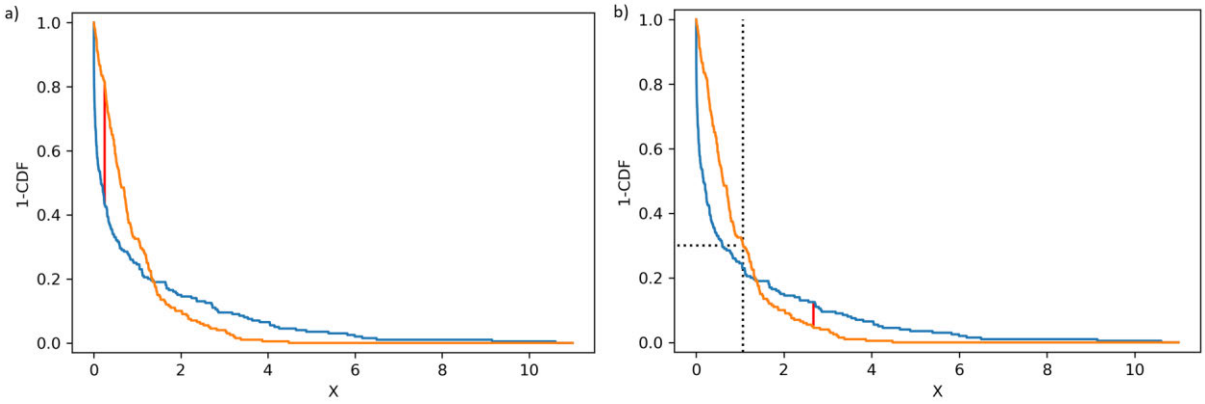

Figure C.2: Examples of (a) Kolmogorov-Smirnov test and (b) truncated Kolmogorov-Smirnov test. In both graphs the test statistics are defined by the length of the red line. In (b) the y-axis truncation point y_c is set at 0.3.

In this study, given a cut-off point y_c, the truncated Kolmogorov-Smirnov test between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ is defined as conducting this test between each of the $\mathbb{D}^1_{FT}$ with $\mathbb{D}^2_{FT_{sim}}$ each of the $\mathbb{D}^1_{FT}$ in $\mathbb{D}^2_{FT_{obs}}$, and the resulting $\mathbb{D}^2_{FT}$ test statistics $t_{\mathbb{D}^2_{KS_{trc}}}$ is a $n_{sim} \times n_{obs}$ 2-dimensional array of $t_{KS_{trc}}$ structured as follows:

$$t_{\mathbb{D}^2_{KS_{trc}}} = \begin{bmatrix} t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_1}}, \mathbb{D}^1_{FT_{obs_1}}\right) & t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_1}}, \mathbb{D}^1_{FT_{obs_2}}\right) & \cdots & t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_1}}, \mathbb{D}^1_{FT_{obs_o}}\right) \\ t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_2}}, \mathbb{D}^1_{FT_{obs_1}}\right) & t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_2}}, \mathbb{D}^1_{FT_{obs_2}}\right) & & t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_2}}, \mathbb{D}^1_{FT_{obs_o}}\right) \\ \vdots & & \ddots & \vdots \\ t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_s}}, \mathbb{D}^1_{FT_{obs_1}}\right) & t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_s}}, \mathbb{D}^1_{FT_{obs_2}}\right) & \cdots & t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim_s}}, \mathbb{D}^1_{FT_{obs_o}}\right) \end{bmatrix} \quad [3]$$

With the subscript $(1,2,\ldots,s)$ for $\mathbb{D}^1_{FT_{sim}}$ and $(1,2,\ldots,o)$ for $\mathbb{D}^1_{FT_{obs}}$ denoting the sequence index of $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ within $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ respectively, and n_{sim} and n_{obs} are the number of $\mathbb{D}^1_{FT}$ within $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ respectively.

In this study a multitude of y_c were employed. The following y_c was utilized for this statistic: $y_c = 0.01, 0.008, 0.006, 0.005, 0.004, 0.003, 0.002, 0.0015, 0.001, 0.0005, 0.0004, 0.0003,$

0.0002, 0.0001 and 0. The $t_{\mathbb{D}^2_{KS_{trc}}}$ from each y_c is stacked into a 3-dimensional array with size $n_{sim} \times l_{y_c} \times n_{obs}$, with l_{y_c} being the number of y_cs being tested in this study (i.e. $l_{y_c} = 15$).

C.2.1.2. Kuiper's Test

A close relative to the Kolmogorov-Smirnov test is the Kuiper's test. It was introduced by Kuiper (1960) as a modification of Kolmogorov-Smirnov test for testing the averages of azimuths a group of migratory birds would take during the migration process. The test statistic (t_{KU}) is defined as the sum of the absolute values of supremum distance and infimum distance between two ECDFs (Kuiper, 1960):

$$t_{KU} = \left|\sup\left(\mathbb{D}^1_{FT_{sim}} - \mathbb{D}^1_{FT_{obs}}\right)\right| + \left|\inf\left(\mathbb{D}^1_{FT_{sim}} - \mathbb{D}^1_{FT_{obs}}\right)\right| \quad [4]$$

An example of the implementation of Kuiper's test is provided in Figure C.3(a).

Cirrone (2004) and Lanzante (2021) suggested this statistic is equally powerful in detecting discrepancies at the median and the tail of the distribution. The signal-to-noise ratio near the median of $\mathbb{D}^1_{FT}$ is low however, thus discrepancy in that region could be caused by noise rather than signal, muddling the test statistics. For this reason, truncation of the statistics was also employed in this test statistic, using the same y_c and x_c as in the calculation of truncated Kolmogorov-Smirnov test. For this the test statistics of the truncated Kuiper's test ($t_{KU_{trc}}$) it is calculated in a similar way as in truncated Kolmogorov-Smirnov test:

$$\begin{aligned} t_{KU_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right) = & \left|\sup\left(\mathbb{D}^1_{FT_{sim}}(x \geq x_c) - \mathbb{D}^1_{FT_{obs}}(x \geq x_c)\right)\right| \\ & + \left|\inf\left(\mathbb{D}^1_{FT_{sim}}(x \geq x_c) - \mathbb{D}^1_{FT_{obs}}(x \geq x_c)\right)\right| \end{aligned} \quad [5]$$

The truncated Kuiper's test is illustrated in Figure C.3(b).

Given a cut-off point y_c, the test statistics for the truncated Kuiper's test between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ (denoted as $t_{\mathbb{D}^2_{KU_{trc}}}$) is a $n_{sim} \times n_{obs}$ 2-dimensional array structured in a similar fashion as in $t_{\mathbb{D}^2_{KS_{trc}}}$ in equation [3], with $t_{KU_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right)$ from equation [5] being used in place of $t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right)$.The same set of y_cs as in the truncated Kolmogorov-Smirnov test was used, and the test statistics is kept as a 3-dimensional array with size $n_{sim} \times l_{y_c} \times n_{obs}$.

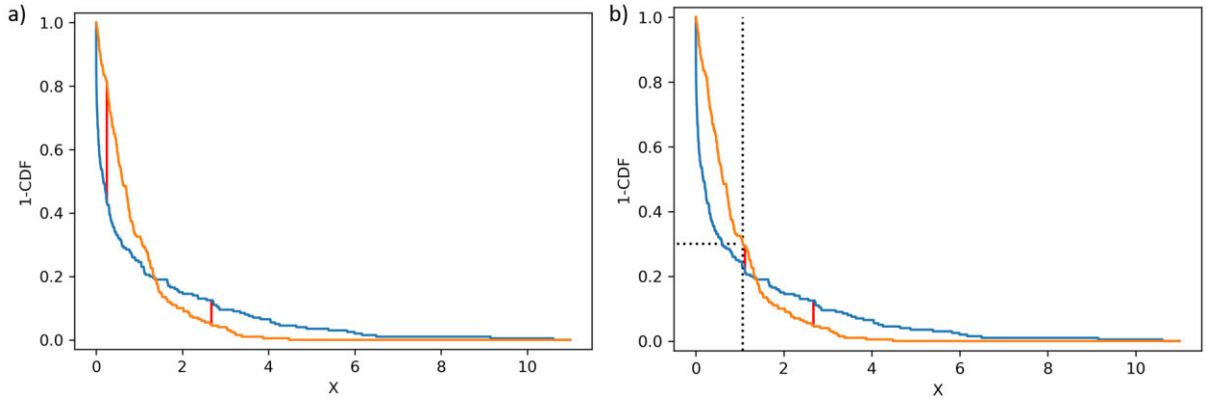

Figure C.3: Example of (a) Kuiper's test and (b) truncated Kuiper's test. In both graphs the test statistics are defined as the sums of length of the red lines in each graph. In (b) the y-axis truncation point y_c is set at 0.3.

C.2.1.3. Maximal x-axis Distance Test

While the Kolmogorov-Smirnov test and Kuiper's test relies on the maximal distance between two distributions along the y-axis, it is also possible to test the equality of two distribution using the distance along the x-axis. One such statistic, which was termed "maximal x-axis distance test" in this study, utilized the latter. This is a newly developed statistics designed to detect differences in magnitude of outliers between two distributions.

The mechanism of this statistics relies on the aforementioned convergence of the maximum values of the random variables in $\mathbb{D}_{FT}^{2}$. The rationale behind this statistic is that if the two empirical distributions came from the same underlying distribution, provided that none of the random variates is undefined and they have the same data size, the maximum values from the empirical distribution converge asymptotically, forming a new distribution (Figure C.4(a), Figure C.4(c)). Whereas if the two empirical distributions do not come from the same underlying distribution, with resampling, they would not asymptotically converge toward the same distribution (Figure C.4(b), Figure C.4(d)).

In this study, the maximal x-axis distance test is defined as the differences of maximum values between two distributions:

$$t_{XML}(\mathbb{D}_{FT_{sim}}^{1}, \mathbb{D}_{FT_{obs}}^{1}) = \left|\max\left(\mathbb{D}_{FT_{sim}}^{1}\right) - \max\left(\mathbb{D}_{FT_{obs}}^{1}\right)\right| \quad [6]$$

The $\mathbb{D}^{2}$ test statistics of maximum x-axis distance test $t_{\mathbb{D}_{XML}^{2}}$ is a 2-dimensional array of size $n_{sim} \times n_{obs}$ containing the t_{XML} between each of the $\mathbb{D}_{FT_{sim}}^{1}$ and each of the $\mathbb{D}_{FT_{obs}}^{1}$, and is structured in a similar fashion as in $t_{\mathbb{D}_{KS_{trc}}^{2}}$, but with $t_{XML}(\mathbb{D}_{FT_{sim}}^{1}, \mathbb{D}_{FT_{obs}}^{1})$ in place of $t_{KS_{trc}}\left(\mathbb{D}_{FT_{sim}}^{1}, \mathbb{D}_{FT_{obs}}^{1}\right)$.

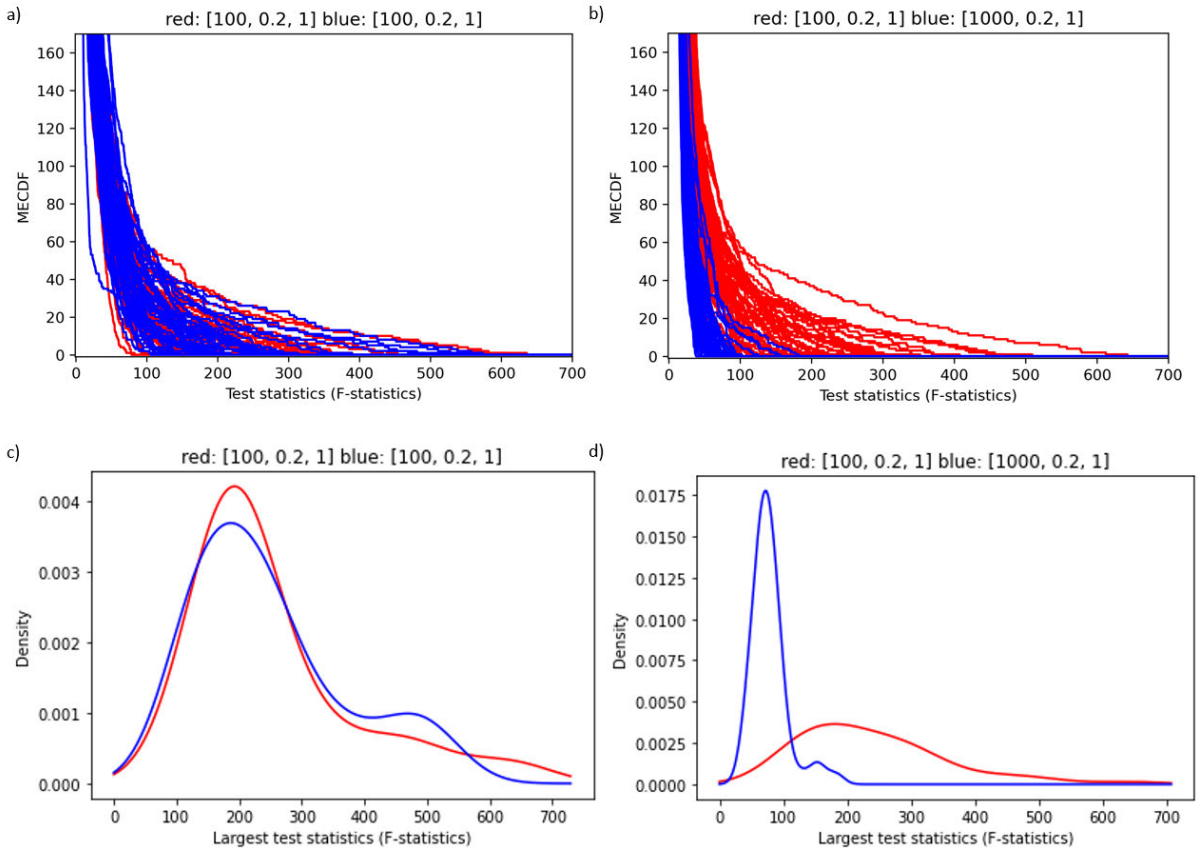

Figure C.4: The mechanism of maximal x-axis distance test. Figure (a) and (b) shows the scaled ECDF (denoted as $MECDF$s) of the test statistics, whereas (c) and (d) shows the distributions of the maximum test statistics (i.e., maximum x-axis values) from each of the $MECDF$s. In (a) and (c) the red and blue lines are generated from the same genetic architecture parameters $Q(100, 0.2, 1)$, whereas for (b) and (d) the red lines are generated from $Q(100, 0.2, 1)$ whereas the blue lines are generated from $Q(1000, 0.2, 1)$. The maximal x-axis distance uses the differences in the distribution in (c) and (d) to test the equality of the $MECDF$s.

C.2.2. Area-based Statistics

Besides maximal distances, discrepancies between two distributions can also be measured through the area between the distributions. The rationale behind this class of statistics is that if the distribution comes from the same underlying distribution, the area between the distributions would be minimized (Dobrushin, 1970; Dowd, 2020). Generally, area-based statistics are more powerful than the maximal distance statistics as the comparison utilized the entirety range of x-axis rather than one single point (Cirrone, 2004, Dowd, 2020).

C.2.2.1. Wasserstein's Statistics

Perhaps the most straightforward and important area-based statistics is the Wasserstein's statistics (Cabrelli and Molter, 1995). It was first formulated by Kantorovich (1939) and

Kantorovich (1958) as a process to achieve optimized distribution of workforces using the least amount of work. It was unknowingly utilized by Vasershtein (1969) as part of the probability modelling of a Markov process, before popularized by Dobrushin (1970) as a way of measuring the amount of discrepancies between two distributions.

The test statistic for Wasserstein's statistics (t_{WS}) is defined as the area between curves of $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ across all test statistics FT (Cabrelli and Molter, 1995; Dobrushin, 1970):

$$t_{WS} = \int_{-\infty}^{\infty} \left| \mathbb{D}^1_{FT_{sim}}(FT) - \mathbb{D}^1_{FT_{obs}}(FT) \right| \, dFT \qquad [7]$$

Where $\mathbb{D}^1_{FT_{sim}}(FT)$ and $\mathbb{D}^1_{FT_{obs}}(FT)$ are defined as the values of $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ evaluated at x-axis point $x = FT$ respectively. An example of this statistic is provided in Figure C.5(a).

The Wasserstein's statistics is generally more powerful than Kolmogorov-Smirnov test and Kuiper's test (Cirrone, 2004). Despite this, similar to Kolmogorov-Smirnov test, this test is powerful at the median and weak at the tail of the $\mathbb{D}^1_{FT}$. This can be attributed to the unequal variances of the test statistic across the y-axis of the ECDF, with maximal variance at $ECDF = 0.5$ (Dowd, 2020). For this reason, truncation would be applied, with the truncated Wasserstein's statistics defined as follows:

$$t_{WS_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right) = \int_{FT_c}^{\infty} \left| \mathbb{D}^1_{FT_{sim}}(FT) - \mathbb{D}^1_{FT_{obs}}(FT) \right| \, dFT \qquad [8]$$

Where FT_c is the test statistic that served as the cut-off point of the truncated Wasserstein's statistics. The calculation of FT_c is equivalent to the calculation of x_c for the truncated Kolmogorov-Smirnov statistic. The truncated Wasserstein's statistic is illustrated in Figure C.5(b).

The test statistic for the truncated Wasserstein's statistic between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ is defined (denoted as $t_{\mathbb{D}^2_{WS_{trc}}}$) as a 2-dimensional array of size $n_{sim} \times n_{obs}$ containing $t_{WS_{trc}}$ between each $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$, and is structured in a similar way as in $t_{\mathbb{D}^2}$ for the truncated Kolmogorov-Smirnov statistic. With repetition of $t_{\mathbb{D}^2_{WS_{trc}}}$ across varying y-axis cut-off points y_c, the resulting test statistic is a 3-dimensional array of size $n_{sim} \times l_{y_c} \times n_{obs}$. The y-axis cut-off points tested in truncated Wasserstein's statistic is the same as those defined in truncated Kolmogorov-Smirnov test.

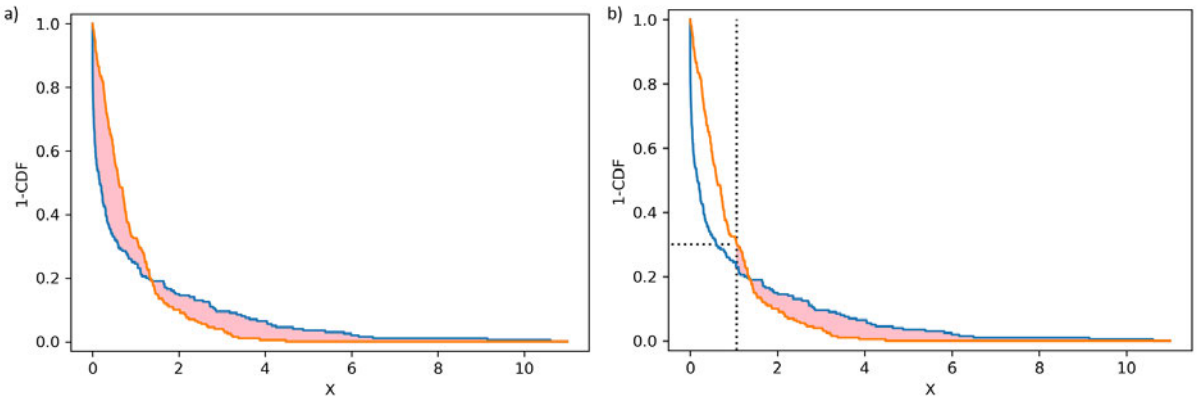

Figure C.5: Example of (a) Wasserstein's statistics and (b) truncated Wasserstein's statistics, evaluated across a range of x-axis. The test statistics are defined by the area of the pink regions in each of the graphs. In (b) the y-axis truncation point y_c is set at 0.3.

C.2.2.2. DTS Statistics and its Generalization

The weakness of Wasserstein's statistic toward discrepancies at the tail of the distributions has led to development of new statistics, and one such statistic is the DTS statistic, a modified version of Wasserstein's statistic introduced by Dowd (2020).

The inspiration for this statistic is the Anderson-Darling test, which is equivalent to the Cramer-von Mises test that has been weighted by the variance factor $y(1-y)$ in order to improve the latter's power toward discrepancies at the tail of the distribution (Anderson and Darling, 1952). For this, the DTS test statistics (t_{DTS}) between $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ is defined as follows (Dowd, 2020):

$$t_{DTS} = \int_{-\infty}^{\infty} \frac{\left|\mathbb{D}^1_{FT_{sim}}(FT) - \mathbb{D}^1_{FT_{obs}}(FT)\right|}{\hat{Y}(FT)}\, dFT \qquad [9]$$

Where $\hat{Y}(FT)$ is the weightage function at point $x = FT$. In the original publication by Dowd (2020) the $\hat{Y}(FT)$ is defined as follow (denoted as $\hat{Y}(FT)_m$):

$$\hat{Y}(FT)_m = \left(\frac{\mathbb{D}^1_{FT_{sim}}(FT) + \mathbb{D}^1_{FT_{obs}}(FT)}{2}\right) * \left(1 - \frac{\mathbb{D}^1_{FT_{sim}}(FT) + \mathbb{D}^1_{FT_{obs}}(FT)}{2}\right) \qquad [10]$$

An example of utilization of DTS statistics with $\hat{Y}(FT)_m$ is provided in Figure C.6(a).

There are several possible interpretations of $\hat{Y}(FT)$. Indeed, from the original publication for the Anderson-Darling test (Anderson and Darling, 1952; Anderson and Darling, 1954) the authors did not explicitly define the $y(1-y)$ as the mandatory weighting factor, and instead

allow the users to choose any suitable weighting factor as long as the factor is nonnegative. For this, besides the original definition of $\hat{Y}(FT)$, an alternative version for the weighting factor was formulated, which is defined as follows (denoted as $\hat{Y}(FT)_A$):

$$\hat{Y}(FT)_A = \frac{\left(\mathbb{D}^1_{FT_{sim}}(FT) * \left(1 - \mathbb{D}^1_{FT_{sim}}(FT)\right)\right) + \left(\mathbb{D}^1_{FT_{obs}}(FT) * \left(1 - \mathbb{D}^1_{FT_{obs}}(FT)\right)\right)}{2} \quad [11]$$

There are slight differences between $\hat{Y}(FT)_m$ and $\hat{Y}(FT)_A$; the $\hat{Y}(FT)_m$ can be seen as the expected variance of the test statistic evaluated at the midpoint between $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ at point $x = FT$, whereas for $\hat{Y}(FT)_A$, it can be seen as the midpoint between two expected test statistic variances evaluated at $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ at point $x = FT$.

While DTS statistics have more power to detect discrepancies at the tail of the distributions, it is still sensitive toward discrepancies near the median of the distribution, which is undesirable in this situation. Thus, truncation would also be applied onto DTS statistics. For the truncated DTS test statistics that utilized $\hat{Y}(FT)_m$ is denoted as $t_{DTSm_{trc}}$ and is defined as follows:

$$t_{DTSm_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right) = \int_{FT_c}^{\infty} \frac{\left|\mathbb{D}^1_{FT_{sim}}(FT) - \mathbb{D}^1_{FT_{obs}}(FT)\right|}{\hat{Y}(FT)_m}\, dFT \quad [12]$$

and the truncated DTS test statistics that utilized $\hat{Y}(FT)_A$ is denoted as $t_{DTSA_{trc}}$ and is defined as follows:

$$t_{DTSA_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right) = \int_{FT_c}^{\infty} \frac{\left|\mathbb{D}^1_{FT_{sim}}(FT) - \mathbb{D}^1_{FT_{obs}}(FT)\right|}{\hat{Y}(FT)_A}\, dFT \quad [13]$$

An example of truncated DTS statistics with $\hat{Y}(FT)_m$ is provided in Figure C.6(b).

The truncated DTS test statistic between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ is defined as a pair of 2-dimensional arrays of size $n_{sim} \times n_{obs}$. The first array, denoted as $t_{\mathbb{D}^2_{DTSm_{trc}}}$, contains $t_{DTSm_{trc}}$ between each of the $\mathbb{D}^1_{FT_{sim}}$ and each of the $\mathbb{D}^1_{FT_{obs}}$, and is structured in a similar way as in the $t_{\mathbb{D}^2_{KS_{trc}}}$, but with $t_{DTSm_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right)$ in place of $t_{KS_{trc}}\left(\mathbb{D}^1_{FT_{sim}}, \mathbb{D}^1_{FT_{obs}}\right)$. The second array, denoted as $t_{\mathbb{D}^2_{DTSA_{trc}}}$, contains $t_{DTSA_{trc}}$ between the $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$. This calculation of $t_{\mathbb{D}^2_{DTSm_{trc}}}$ and $t_{\mathbb{D}^2_{DTSA_{trc}}}$ was repeated across varying y-axis cut-off points y_c, with the cut-off points being the same as those utilized in truncated Kolmogorov-Smirnov

tests. The eventual test statistics is kept as a 3-dimensional array of size $n_{sim} \times (2 * l_{y_c}) \times n_{obs}$ formed by compiling the $t_{\mathbb{D}^2_{DTSm_{trc}}}$ and $t_{\mathbb{D}^2_{DTSA_{trc}}}$ across all y_c.

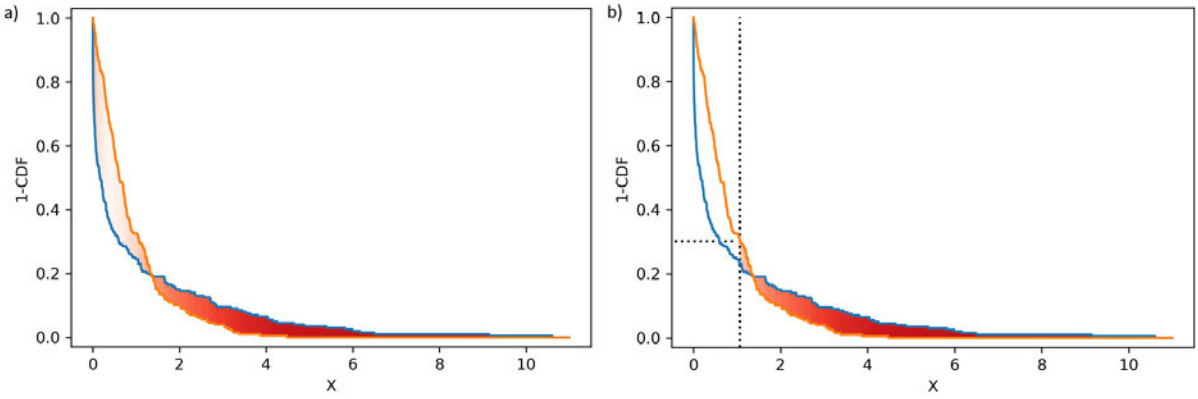

Figure C.6: Example of (a) DTS statistics and (b) truncated DTS statistics, evaluated across a range of x-axis. The test statistics is defined as the area of the shaded region, with the intensity of the shade denotes the relative weights for the statistics. The darker the shades the heavier the relative weights. In (b) the y-axis truncation point y_c is set at 0.3.

C.2.3. Quantile-based Statistics

While the x-axis cut-off points x_c are utilized in the previously mentioned statistics as the truncation points for the $\mathbb{D}^1_{FT}$s, the x_c by itself reveal some of the properties for the distributions. For example, if the $\mathbb{D}^1_{FT}$s originate from the same underlying distribution, then the x_cs converge toward a distribution. This is further aided by the internal consistency of the distributions, which allow a sequence of distributions to converge toward an asymptotic distribution (i.e., the distributions would distribute close to their asymptotic distribution, similar to how a set of normal random variates distribute close to their mean). Thus, several statistics of this class were utilized to detect the discrepancies between the $\mathbb{D}^1_{FT}$s.

C.2.3.1. Equivalence in Quantiles

This is a newly developed method to test the equality between two sequences of distributions. The rationale for this method is that if two $\mathbb{D}^2_{FT}$s come from the same underlying distribution, then the quantile from each of the $\mathbb{D}^1_{FT}$ within the two $\mathbb{D}^2_{FT}$ would form a similar distribution, and the opposite is true if the two $\mathbb{D}^2_{FT}$s do not come from the same distributions (Figure C.7).

Given a quantile that would be used as a cut-off point y_q, the corresponding x-axis cut-off point x_{yq} of $\mathbb{D}^1_{FT}$ is defined as follows:

$$y_q = \mathbb{D}^1_{FT}(x_{yq}) \quad [14]$$

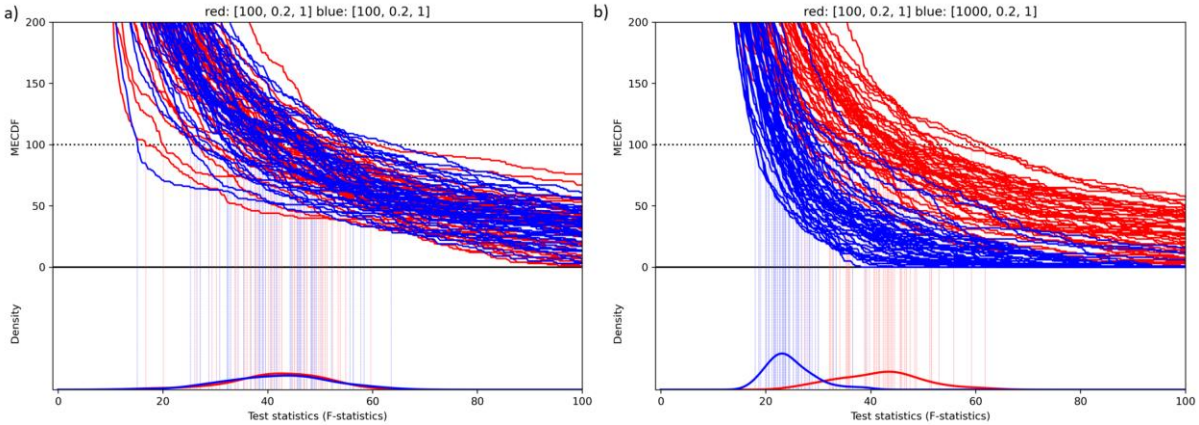

Figure C.7: The mechanism of "Equivalence in Quantiles" test. The top panels in both figures represent the $MECDF$s generated from (a) two of the same genetic architecture parameters $Q(100, 0.2, 1)$ and (b) two different genetic architecture parameters with red distributions being $Q(100, 0.2, 1)$ and blue $Q(1000, 0.2, 1)$. The total number of markers used is 50,000, and the quantile that would be used as cut-off point is set at 0.002. This correspond to $MECDF$ cut-off point at 50,000*0.002 = 100, which is delineated with a dotted black line. From each of the $MECDF$s the corresponding x-axis values were recorded, which is represented by the vertical lines down to the x-axis. These x-axis values form a distribution, which is featured in the bottom panels. The "Equivalence in Quantiles" test tests the similarity of distribution of the x-axis values from the blue and red $MECDF$s.

This calculation can then be applied onto all $\mathbb{D}^1_{FT}$ within $\mathbb{D}^2_{FT_{sim}}$, and this produces a vector of length n_{sim} (denoted as $\boldsymbol{x_{yq_{sim}}}$) containing x_{yq} for each of the $\mathbb{D}^1_{FT_{sim}}$. These x_{yq} are denoted as s_x. This operation can then be repeated for $\mathbb{D}^2_{FT_{obs}}$, with the resulting vector denoted as $\boldsymbol{x_{yq_{obs}}}$ and the x_{yq} denoted as o_x. The vectors are structured as follows:

$$\boldsymbol{x_{yq_{sim}}} = \left[s_{x_1}, s_{x_2}, s_{x_3}, \dots, s_{x_{n_{sim}}}\right] \quad [15]$$

$$\boldsymbol{x_{yq_{obs}}} = \left[o_{x_1}, o_{x_2}, o_{x_3}, \dots, o_{x_{n_{obs}}}\right] \quad [16]$$

Finally, the test statistic for the equivalence in quantile (denoted as $t_{\mathbb{D}^2_{EQV}}$) is a $n_{sim} \times n_{obs}$ array and is calculated as follows:

$$t_{\mathbb{D}^2_{EQV}} = \begin{bmatrix} \left|s_{x_1} - o_{x_1}\right| & \left|s_{x_1} - o_{x_2}\right| & \cdots & \left|s_{x_1} - o_{x_{n_{obs}}}\right| \\ \left|s_{x_2} - o_{x_1}\right| & \left|s_{x_2} - o_{x_2}\right| & \cdots & \left|s_{x_2} - o_{x_{n_{obs}}}\right| \\ \vdots & \vdots & \ddots & \vdots \\ \left|s_{x_{n_{sim}}} - o_{x_1}\right| & \left|s_{x_{n_{sim}}} - o_{x_2}\right| & \cdots & \left|s_{x_{n_{sim}}} - o_{x_{n_{obs}}}\right| \end{bmatrix} \quad [17]$$

The calculation of $t_{\mathbb{D}^2_{EQV}}$ is repeated across varying cut-off points, with y_q tested being $y_q =$ 0.02, 0.015, 0.01, 0.009, 0.008, 0.007, 0.006, 0.005, 0.004, 0.003, 0.002, 0.0015, 0.001, 0.0005, 0.0004, 0.0003, 0.0002, 0.0001 and 0.00002. The resulting test statistics are kept as a 3-dimensional array of size $n_{sim} \times l_{y_q} \times n_{obs}$ where l_{y_q} is the number of y_q being tested in this study (i.e. $l_{y_q} = 19$).

C.2.3.2. Distance from Median

Using the same $\boldsymbol{x_{yq_{sim}}}$ and $\boldsymbol{x_{yq_{obs}}}$, another test that can be done is to compare the location x_{yq} from one of the vectors onto the other vector. As a simplified example, given an ordered vector of length 7 $\boldsymbol{z_1} = [0.7, 1.0, 1.3, 1.5, 1.6, 2.0, 2.4]$, the location of the median in this vector would be 4. Given a value that is to be tested, e.g. $z = 1.9$, if z is to be inserted into $\boldsymbol{z_1}$ in such a way that ordered state in $\boldsymbol{z_1}$ is preserved, then z would be positioned between index 5 and 6 in $\boldsymbol{z_1}$. As it turns out, if $\boldsymbol{z_1}$ forms a distribution, and z came from the same distribution as $\boldsymbol{z_1}$, then the expected location of z would be close to the median of $\boldsymbol{z_1}$. This is the basis of "distance from median" test. For this example, the distance from median test statistics (denoted as t_{LM}) is defined as follows:

$$t_{LM}(z, \boldsymbol{z_1}) = \left|(Number\ of\ z_i\ in\ \boldsymbol{z_1} \geq z) - \frac{length\ of\ \boldsymbol{z_1} + 1}{2}\right| \quad [18]$$
$$= \left|5 - \frac{7+1}{2}\right|$$
$$= 1$$

The maximum value attainable by t_{LM} is $\frac{length\ of\ \boldsymbol{z_1}+1}{2}$, and observing such t_{LM} would indicate that z fall outside the range of $\boldsymbol{z_1}$.

Now hypothetically if there is a second vector to be tested $\boldsymbol{z_2}$, and this vector is identically distributed as in $\boldsymbol{z_1}$, then the "distance from median" test can be applied to each element in $\boldsymbol{z_2}$ onto $\boldsymbol{z_1}$, from which a sequence of t_{LM}s could be obtained. While the individual t_{LM}s appeared to disperse randomly with no discernible pattern, collectively the t_{LM}s would still be minimized (i.e., few if any of the t_{LM}s would stray far away from the location of the median, minimizing the statistics). Whereas if $\boldsymbol{z_2}$ came from different distribution as in $\boldsymbol{z_1}$, especially if their median is far away from that in $\boldsymbol{z_1}$, then a larger number of the t_{LM}s would attain its maximum value, thus failing to minimize the t_{LM}s collectively. An example of this phenomenon is provided in C.8.

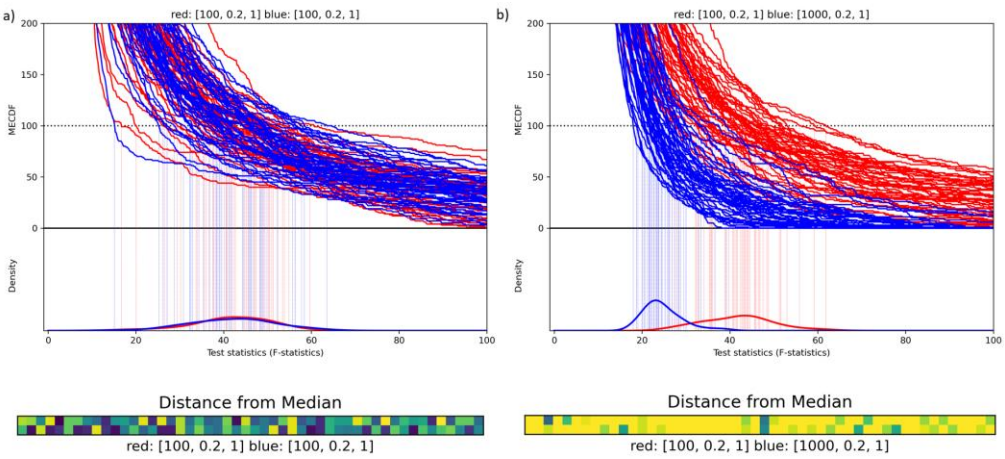

Figure C.8: The mechanism of "Distance from Median" test, using the same set of $MECDF$s with the same y-axis cut-off points of $y_q = 0.002$ and the same data and distribution of x-axis values as in Figure C.7. The red and blue $MECDF$s in (a) were generated from the same genetic architecture parameters $Q(100, 0.2, 1)$ while (b) were generated from different genetic architecture parameters, with red from $Q(100, 0.2, 1)$ and blue from $Q(1000, 0.2, 1)$. The raster plots at the bottom of (a) and (b) represent how far the location of each of the x-axis values from a set of $MECDF$s when compared with the median of those from the other sets of $MECDF$s. The first row of the raster plots is from each of the blue $MECDF$s compared to the medians of those in red $MECDF$s, and the second row if from each of the red $MECDF$s compared to median of the blue $MECDF$s. The distance is calculated as defined in equation [18], and the lighter the pixel is in the raster plot, the further the distance of an x-axis value to the median.

One shortcoming for this statistic is that it is only powerful toward differences in median of the $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$, and is weak against any other discrepancies in the distribution of $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$, such as differences variance and kurtosis in their distribution.

For this study, the distance from median test was conducted on the $\boldsymbol{x_{yq_{sim}}}$ and $\boldsymbol{x_{yq_{obs}}}$. If the comparison is made between two vectors there are two ways of calculating the test statistic: (1) compare each of s_x in $\boldsymbol{x_{yq_{sim}}}$ with vector $\boldsymbol{x_{yq_{obs}}}$, and (2) compare each of o_x in $\boldsymbol{x_{yq_{obs}}}$ with vector $\boldsymbol{x_{yq_{sim}}}$. The resulting test statistic from (1), denoted as $t_{\mathbb{D}^2_{LM(1)}}$ is a vector of length n_{sim} and is calculated as follows:

$$t_{\mathbb{D}^2_{LM(1)}} = \left[t_{LM}\left(s_{x_1}, \boldsymbol{x_{yq_{obs}}}\right) \quad t_{LM}\left(s_{x_2}, \boldsymbol{x_{yq_{obs}}}\right) \quad t_{LM}\left(s_{x_3}, \boldsymbol{x_{yq_{obs}}}\right) \quad \cdots \quad t_{LM}\left(s_{x_{n_{sim}}}, \boldsymbol{x_{yq_{obs}}}\right)\right] \quad [19]$$

And the resulting test statistic from (2), denoted as $t_{\mathbb{D}^2_{LM(2)}}$ is a vector of length n_{obs} and is calculated as follows:

$$t_{\mathbb{D}^2_{LM(2)}} = \left[t_{LM}\left(o_{x_1}, \boldsymbol{x_{yq_{sim}}}\right) \quad t_{LM}\left(o_{x_2}, \boldsymbol{x_{yq_{sim}}}\right) \quad t_{LM}\left(o_{x_3}, \boldsymbol{x_{yq_{sim}}}\right) \quad \cdots \quad t_{LM}\left(o_{x_{n_{obs}}}, \boldsymbol{x_{yq_{sim}}}\right)\right] \quad [20]$$

The final test statistic for the location from the median (denoted as $t_{\mathbb{D}^2_{LM}}$) is a $n_{sim} \times n_{obs}$ array and is calculated as follows:

$$t_{\mathbb{D}^2_{LM}} = \begin{bmatrix} t_{\mathbb{D}^2_{LM(1)}}(1) + t_{\mathbb{D}^2_{LM(2)}}(1) & t_{\mathbb{D}^2_{LM(1)}}(1) + t_{\mathbb{D}^2_{LM(2)}}(2) & \cdots & t_{\mathbb{D}^2_{LM(1)}}(1) + t_{\mathbb{D}^2_{LM(2)}}(n_{obs}) \\ t_{\mathbb{D}^2_{LM(1)}}(2) + t_{\mathbb{D}^2_{LM(2)}}(1) & t_{\mathbb{D}^2_{LM(1)}}(2) + t_{\mathbb{D}^2_{LM(2)}}(2) & \cdots & t_{\mathbb{D}^2_{LM(1)}}(2) + t_{\mathbb{D}^2_{LM(2)}}(n_{obs}) \\ \vdots & \vdots & \ddots & \vdots \\ t_{\mathbb{D}^2_{LM(1)}}(n_{sim}) + t_{\mathbb{D}^2_{LM(2)}}(1) & t_{\mathbb{D}^2_{LM(1)}}(n_{sim}) + t_{\mathbb{D}^2_{LM(2)}}(2) & \cdots & t_{\mathbb{D}^2_{LM(1)}}(n_{sim}) + t_{\mathbb{D}^2_{LM(2)}}(n_{obs}) \end{bmatrix} \quad [21]$$

Where $t_{\mathbb{D}^2_{LM(1)}}(i)$ and $t_{\mathbb{D}^2_{LM(2)}}(j)$ are the i^{th} and j^{th} entries for the $t_{\mathbb{D}^2_{LM(1)}}$ and $t_{\mathbb{D}^2_{LM(2)}}$, respectively. An example of the full $t_{\mathbb{D}^2_{LM}}$ is provided in Figure C.9.

The process of calculation $t_{\mathbb{D}^2_{LM}}$ was then repeated for all quantiles y_qs, with the quantile tested being the quantiles utilized in the "Equivalence of Quantile" test. The resulting test statistics was kept in a 3-dimensional array of size $n_{sim} \times l_{y_q} \times n_{obs}$ where l_{y_q} is the number of y_q being tested in this study (i.e. $l_{y_q} = 19$).

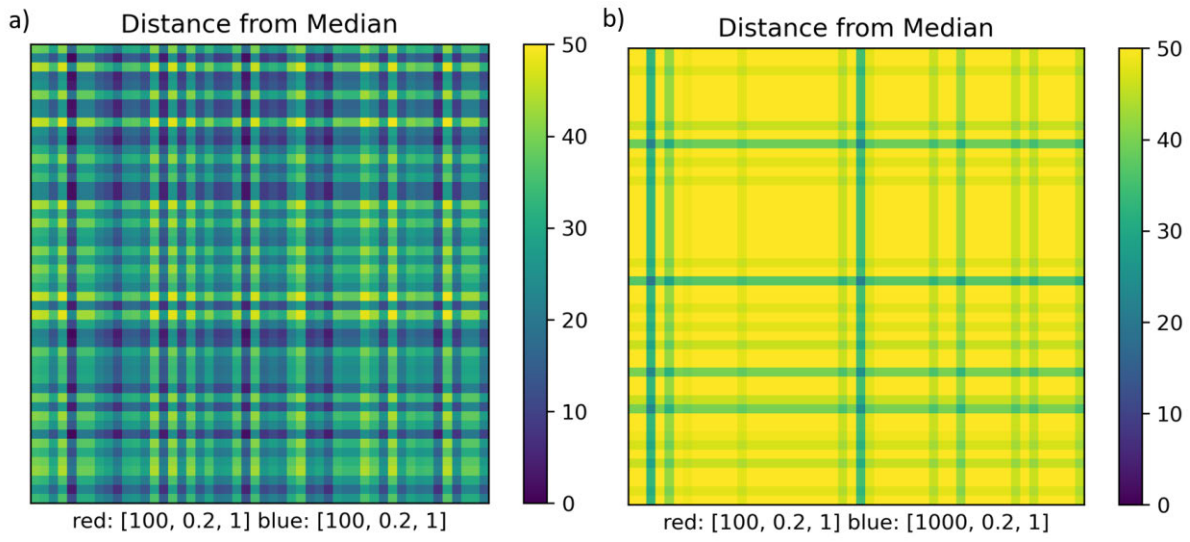

Figure C.9: Examples of $t_{\mathbb{D}^2_{LM}}$ calculated using equation [21] using the "Distance from Median" raster plots in Figure C.8.

C.2.4. Integral based statistics

As previously established, one way to improve the signals from the changing $\mathbb{d}_{QTL}$ is to transform the $\mathbb{D}^1_{FT}$ such that it would amplify any discrepancies at the tail region. Several transformation methods are available, and one such method is by integrating the $\mathbb{D}^1_{FT}$. This operation would become the basis of integral-based statistics.

C.2.4.1. The Mechanism of Error Amplification by Integration (EAI)

This class of statistics employs a newly developed technique known as "Error Amplification by Integration" (EAI). This technique relies on the accumulative property of integration. While this property of integration is usually used in the calculation of CDF of a distribution, a side effect from this process is the amplification of any discrepancies between two distributions. This could be illustrated using the following simplified example. Let y_1 be defined as the following constant function:

$$y_1 = 1.0 \quad [22]$$

Let y_2 be another function with similar definition as in y_1 but with value of 1.2 between $2 < x < 3$ instead of 1.0, thus defined as follows:

$$y_2 = \begin{cases} 1.2\ ;\ 2 < x < 3 \\ 1.0\ ;\ elsewhere \end{cases} \quad [23]$$

The plots for functions y_1 and y_2 are provided in Figure C.10(a).

The area between curves between the raw y_1 and y_2 evaluated in the range $0 < x < 5$ (denoted as $Area_1$) can be evaluated as follows:

$$\begin{aligned} Area_1 &= \int_0^5 |y_1 - y_2|\, dx \quad [24] \\ &= \int_0^2 |y_1 - y_2|\, dx + \int_2^3 |y_1 - y_2|\, dx + \int_3^5 |y_1 - y_2|\, dx \\ &= \int_0^2 |1.0 - 1.0|\, dx + \int_2^3 |1.0 - 1.2|\, dx + \int_3^5 |1.0 - 1.0|\, dx \\ &= \int_0^2 0\, dx + \int_2^3 |-0.2|\, dx + \int_3^5 0\, dx \\ &= 0 + (0.2x)_2^3 + 0 \\ &= (0.2 * 3) - (0.2 * 2) \\ &= 0.2 \end{aligned}$$

Let Y_1 and Y_2 be the integral of y_1 and y_2 respectively, which are defined as follows:

$$Y_1 = x \quad [25]$$

$$Y_2 = \begin{cases} x & x \leq 2 \\ 1.2x - 0.4 & 2 < x < 3 \\ x + 0.2 & x \geq 3 \end{cases} \quad [26]$$

The plots for functions Y_1 and Y_2 are provided in Figure C.10(b).

Using the same range of x, the area between curve between the Y_1 and Y_2 (denoted as $Area_2$) can then be evaluated as follows:

$$
\begin{aligned}
Area_2 &= \int_0^5 |Y_1 - Y_2|\, dx \\
&= \int_0^2 |Y_1 - Y_2|\, dx + \int_2^3 |Y_1 - Y_2|\, dx + \int_3^5 |Y_1 - Y_2|\, dx \\
&= \int_0^2 |x - x|\, dx + \int_2^3 |x - (1.2x - 0.4)|\, dx + \int_3^5 |x - (x + 0.2)|\, dx \\
&= \int_0^2 0\, dx + \int_2^3 |-0.2x + 0.4|\, dx + \int_3^5 |-0.2|\, dx \\
&= 0 + (|-0.1x^2 + 0.4x|)_2^3 + (|-0.2x|)_3^5 \\
&= 0 + 0.1 + 0.4 \\
&= 0.5
\end{aligned}
\qquad [27]
$$

Note the increment of the area between curves from 0.2 in $Area_1$ to 0.5 in $Area_2$. This is caused by the amplification effect from the integration of y_1 and y_2. As the integration ran from $x = 2$ to $x = 3$, the discrepancies between y_1 and y_2 was translated into additional y-axis distance (note that $Y_2 = 3.2$ when $x = 3$, compared to $Y_1 = 3$ for the same x). While the values of y_1 and y_2 coincides for $3 < x < 5$, this is no longer the case for Y_1 and Y_2. Instead, the discrepancies had been carried across the x-axis range, therefore translated into additional area between the curves. It can also be thought as "squashing" the original $Area_1$ into one of the dimensions for $Area_2$ while adding a new dimension onto $Area_2$, thus increasing the area between the curves. This phenomenon would become the basis of EAI.

C.2.4.2. The Utility of EAI

Using EAI, the two $\mathbb{D}_{FT}^1$s were first be integrated using a cumulative integral, with the integrated $\mathbb{D}_{FT}^1$ denoted as $\mathbb{D}^{\mathbb{I}}$:

$$\mathbb{D}^{\mathbb{I}}(FT) = \int_{-\infty}^{x} \mathbb{D}_{FT}^1\, dFT \qquad [28]$$

The integrated $\mathbb{D}_{FT_{sim}}^1$ and $\mathbb{D}_{FT_{obs}}^1$ were denoted as $\mathbb{D}_{sim}^{\mathbb{I}}$ and $\mathbb{D}_{obs}^{\mathbb{I}}$, respectively. This operation can then be applied to all the $\mathbb{D}_{FT}^1$s in $\mathbb{D}_{FT_{sim}}^2$ and $\mathbb{D}_{FT_{obs}}^2$, producing the following sequences of $\mathbb{D}^{\mathbb{I}}$s (denoted as $\mathbb{D}_{sim}^{\mathbb{I}2}$ and $\mathbb{D}_{obs}^{\mathbb{I}2}$ respectively):

$$\mathbb{D}_{sim}^{\mathbb{I}2} = \left[\mathbb{D}_{sim_1}^{\mathbb{I}}, \mathbb{D}_{sim_2}^{\mathbb{I}}, \mathbb{D}_{sim_3}^{\mathbb{I}}, \dots, \mathbb{D}_{sim_{n_{sim}}}^{\mathbb{I}}\right] \qquad [29]$$

$$\mathbb{D}_{obs}^{\mathbb{I}2} = \left[\mathbb{D}_{obs_1}^{\mathbb{I}}, \mathbb{D}_{obs_2}^{\mathbb{I}}, \mathbb{D}_{obs_3}^{\mathbb{I}}, \ldots, \mathbb{D}_{obs_{n_{obs}}}^{\mathbb{I}}\right] \quad [30]$$

An example of implementation of EAI in this study is presented in Figure C.11.

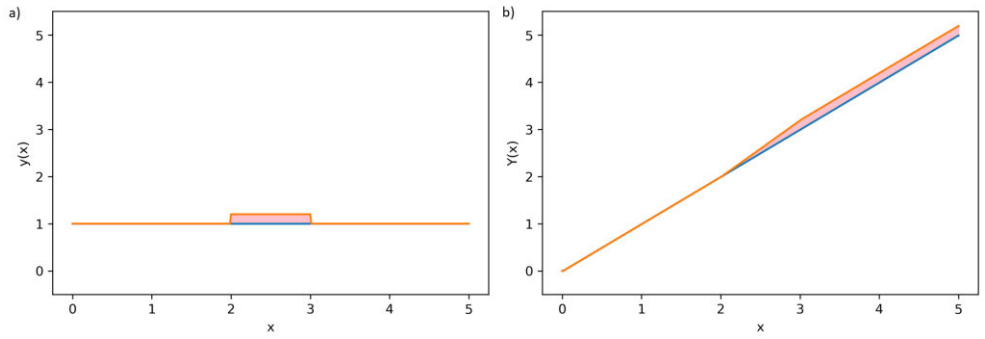

Figure C.10: Mechanism of Error Amplification by Integration (EAI) featured from the examples from equation [22] to [27]. Figure (a) featured the plots for y_1 (blue line) and y_2 (orange line), with the pink region denotes $Area_1$. Figure (b) featured the plots for the integral of y_1 and y_2, denoted as Y_1 (blue line) and Y_2 (orange line) respectively, with the pink region denotes $Area_2$.

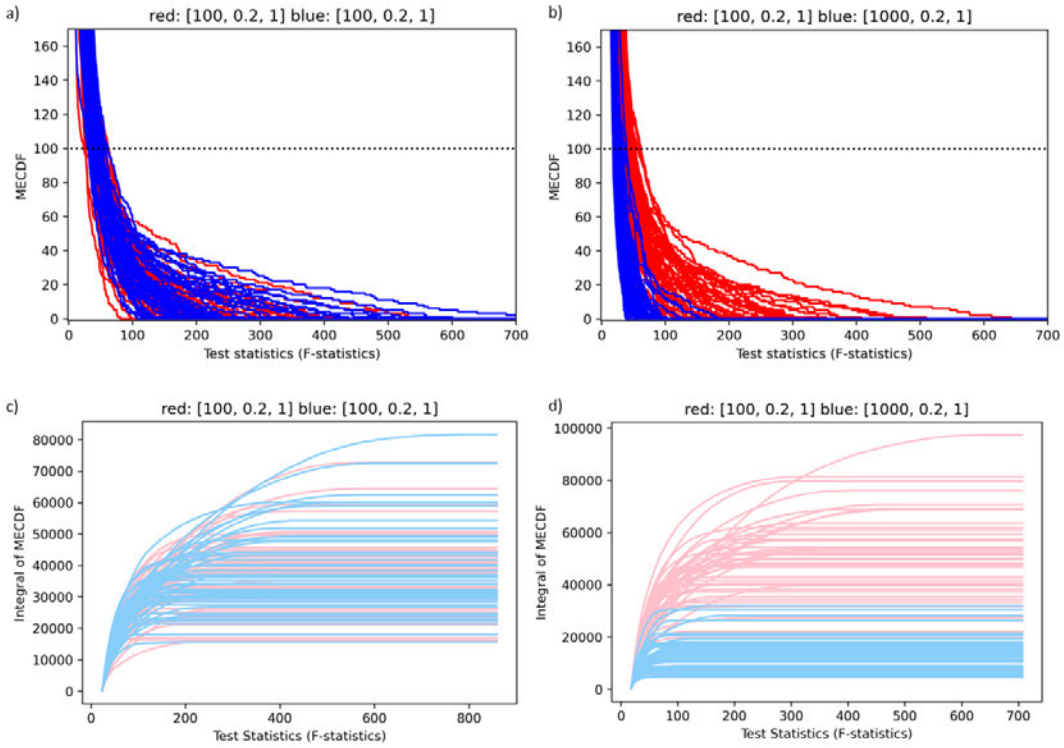

Figure C.11: The use of EAI in testing the equality of $\mathbb{D}_{FT}^{2}$s. Figure (a) and (b) illustrated the raw $MECDF$s while (c) and (d) illustrated the running integral of the $MECDF$s beyond the y-axis cut-off point (black dotted lines in (a) and (b)). The pink curves are the integrals of red $MECDF$s, and the light blue curves are the integrals of blue $MECDF$s in the figures. In (a) the red and blue $MECDF$s are generated from the same genetic architecture parameter $Q(100, 0.2, 1)$, whereas in (b) the red and blue $MECDF$s are generated from different genetic architecture parameters of $Q(100, 0.2, 1)$ and $Q(1000, 0.2, 1)$ respectively.

C.2.4.2.1. Amplified "Wasserstein"-like Statistics

The $\mathbb{D}^{\mathbb{I}2}_{sim}$ and $\mathbb{D}^{\mathbb{I}2}_{obs}$ can be utilized in testing the discrepancies between distributions in a fashion analogous to the previously described statistics. As an example, a "Wasserstein"-like statistics between $\mathbb{D}^{\mathbb{I}}_{sim}$ and $\mathbb{D}^{\mathbb{I}}_{obs}$ (denoted as $A_{\mathbb{D}^{\mathbb{I}}}$) can be defined as follows:

$$A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim}, \mathbb{D}^{\mathbb{I}}_{obs}\right) = \int_{D_{min}}^{D_{max}} \left|\mathbb{D}^{\mathbb{I}}_{sim} - \mathbb{D}^{\mathbb{I}}_{obs}\right| dFT \quad [31]$$

Where D_{max} and D_{min} are defined as follows:

$$D_{max} = \max\left(\mathbb{D}^{1}_{FT_{sim}}, \mathbb{D}^{1}_{FT_{obs}}\right) \quad [32]$$

$$D_{min} = \min\left(\mathbb{D}^{1}_{FT_{sim}}, \mathbb{D}^{1}_{FT_{obs}}\right) \quad [33]$$

The test statistic for the amplified Wasserstein statistics between $\mathbb{D}^{\mathbb{I}2}_{sim}$ and $\mathbb{D}^{\mathbb{I}2}_{obs}$ (denoted as $t_{\mathbb{D}^{2}_{AWS}}$) is a 2-dimensional array of size $n_{sim} \times n_{obs}$ calculated as follows:

$$t_{\mathbb{D}^{2}_{AWS}} = \begin{bmatrix} A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_1}, \mathbb{D}^{\mathbb{I}}_{obs_1}\right) & A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_1}, \mathbb{D}^{\mathbb{I}}_{obs_2}\right) & \cdots & A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_1}, \mathbb{D}^{\mathbb{I}}_{obs_{n_{obs}}}\right) \\ A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_2}, \mathbb{D}^{\mathbb{I}}_{obs_1}\right) & A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_2}, \mathbb{D}^{\mathbb{I}}_{obs_2}\right) & \cdots & A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_2}, \mathbb{D}^{\mathbb{I}}_{obs_{n_{obs}}}\right) \\ \vdots & \vdots & \ddots & \vdots \\ A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_{nsim}}, \mathbb{D}^{\mathbb{I}}_{obs_1}\right) & A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_{n_{sim}}}, \mathbb{D}^{\mathbb{I}}_{obs_2}\right) & \cdots & A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim_{n_{sim}}}, \mathbb{D}^{\mathbb{I}}_{obs_{n_{obs}}}\right) \end{bmatrix} \quad [34]$$

Truncation can be applied onto the EAI algorithm as well. Given a quantile cut-off point y_q, the x-axis cut-off point (x_{yq}) is first calculated using equation [14], and the vectors $\boldsymbol{x_{yq_{sim}}}$ and $\boldsymbol{x_{yq_{obs}}}$ are calculated as defined in equation [15] and [16]. The "universal x-axis cut-off point" (denoted as $x_{y_{min}}$) was defined as the minimum of all x_{yq}s across both vectors:

$$x_{y_{min}} = \min\left(\min\left(\boldsymbol{x_{yq_{sim}}}\right), \min\left(\boldsymbol{x_{yq_{obs}}}\right)\right) \quad [35]$$

The $x_{y_{min}}$ was used as the lower limit for the integral in equation [28]:

$$\mathbb{D}^{\mathbb{I}}(x) = \int_{x_{y_{min}}}^{x} \mathbb{D}^{1}_{FT}\, dFT \quad [36]$$

And this $\mathbb{D}^{\mathbb{I}}$ can then be used in the calculations from equation [29] up to $t_{\mathbb{D}^{2}_{AWS}}$ in [34].

C.2.4.2.2. Amplified "Kolmogorov-Smirnov"-like Statistics

Similarly, an amplified "Kolmogorov-Smirnov"-like y-axis maximal distance (denoted as $KS_{\mathbb{D}^{\mathbb{I}}}$) can also be calculated using $\mathbb{D}^{\mathbb{I}}_{sim}$ and $\mathbb{D}^{\mathbb{I}}_{obs}$ as follows:

$$KS_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim}, \mathbb{D}^{\mathbb{I}}_{obs}\right) = \sup\left|\mathbb{D}^{\mathbb{I}}_{sim} - \mathbb{D}^{\mathbb{I}}_{obs}\right| \quad [37]$$

This y-axis maximal distance between $\mathbb{D}^{\mathbb{I}2}_{sim}$ and $\mathbb{D}^{\mathbb{I}2}_{obs}$ was denoted as $t_{\mathbb{D}^2_{AKS}}$, and it is a 2-dimensional array of size $n_{sim} \times n_{obs}$ structured in a similar way as in $t_{\mathbb{D}^2_{AWS}}$ in equation [34], but with $KS_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim}, \mathbb{D}^{\mathbb{I}}_{obs}\right)$ used in place of $A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim}, \mathbb{D}^{\mathbb{I}}_{obs}\right)$.

C.2.4.2.3. Amplified "Distance from Median"-like Test

"Distance from Median"-like test can also be employed on $\mathbb{D}^{\mathbb{I}}_{sim}$ and $\mathbb{D}^{\mathbb{I}}_{obs}$. This is done by evaluating the maximal y-values of the $\mathbb{D}^{\mathbb{I}}$s in $\mathbb{D}^{\mathbb{I}2}_{sim}$ and $\mathbb{D}^{\mathbb{I}2}_{obs}$ (denoted as $s_{\mathbb{D}^{\mathbb{I}}}$ and $o_{\mathbb{D}^{\mathbb{I}}}$):

$$s_{\mathbb{D}^{\mathbb{I}}} = \mathbb{D}^{\mathbb{I}}_{sim}(D_{max}) \quad [38]$$

$$o_{\mathbb{D}^{\mathbb{I}}} = \mathbb{D}^{\mathbb{I}}_{obs}(D_{max}) \quad [39]$$

These maximal values were kept as a pair of vectors of length n_{sim} and n_{obs} denoted as $\boldsymbol{y_{q_{sim}}}$ and $\boldsymbol{y_{q_{obs}}}$ structured as follows:

$$\boldsymbol{y_{q_{sim}}} = \left[s_{\mathbb{D}^{\mathbb{I}}{}_1}, s_{\mathbb{D}^{\mathbb{I}}{}_2}, s_{\mathbb{D}^{\mathbb{I}}{}_3}, \dots, s_{\mathbb{D}^{\mathbb{I}}{}_{n_{sim}}}\right] \quad [40]$$

$$\boldsymbol{y_{q_{obs}}} = \left[o_{\mathbb{D}^{\mathbb{I}}{}_1}, o_{\mathbb{D}^{\mathbb{I}}{}_2}, o_{\mathbb{D}^{\mathbb{I}}{}_3}, \dots, o_{\mathbb{D}^{\mathbb{I}}{}_{n_{obs}}}\right] \quad [41]$$

The distance from the median test can then be applied on $\boldsymbol{y_{q_{sim}}}$ and $\boldsymbol{y_{q_{obs}}}$ as per equation [18], [19] and [20], with $\boldsymbol{y_{q_{sim}}}$ and $\boldsymbol{y_{q_{obs}}}$ been used in place of $\boldsymbol{x_{yq_{sim}}}$ and $\boldsymbol{x_{yq_{obs}}}$. The resulting vector pair of "Distance from Median" test statistic were denoted as $t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}$ and $t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}$, and were defined as follows:

$$t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}} = \left[t_{LM}\left(s_{\mathbb{D}^{\mathbb{I}}{}_1}, \boldsymbol{y_{q_{obs}}}\right) \quad t_{LM}\left(s_{\mathbb{D}^{\mathbb{I}}{}_2}, \boldsymbol{y_{q_{obs}}}\right) \quad t_{LM}\left(s_{\mathbb{D}^{\mathbb{I}}{}_3}, \boldsymbol{y_{q_{obs}}}\right) \quad \cdots \quad t_{LM}\left(s_{\mathbb{D}^{\mathbb{I}}{}_{n_{sim}}}, \boldsymbol{y_{q_{obs}}}\right)\right] \quad [42]$$

$$t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}} = \left[t_{LM}\left(o_{\mathbb{D}^{\mathbb{I}}{}_1}, \boldsymbol{y_{q_{sim}}}\right) \quad t_{LM}\left(o_{\mathbb{D}^{\mathbb{I}}{}_2}, \boldsymbol{y_{q_{sim}}}\right) \quad t_{LM}\left(o_{\mathbb{D}^{\mathbb{I}}{}_3}, \boldsymbol{y_{q_{sim}}}\right) \quad \cdots \quad t_{LM}\left(o_{\mathbb{D}^{\mathbb{I}}{}_{n_{obs}}}, \boldsymbol{y_{q_{sim}}}\right)\right] \quad [43]$$

Finally, the "Distance from median" statistics between $\mathbb{D}^{\mathbb{I}2}_{sim}$ and $\mathbb{D}^{\mathbb{I}2}_{obs}$ were defined as a 2-dimensional array of size $n_{sim} \times n_{obs}$ (denoted as $t_{\mathbb{D}^2_{ALM}}$) and is structured as follows:

$$t_{\mathbb{D}^2_{ALM}} = \begin{bmatrix} t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(1) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(1) & t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(1) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(2) & \cdots & t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(1) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(n_{obs}) \\ t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(2) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(1) & t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(2) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(2) & \cdots & t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(2) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(n_{obs}) \\ \vdots & \vdots & \ddots & \vdots \\ t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(n_{sim}) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(1) & t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(n_{sim}) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(2) & \cdots & t_{\mathbb{D}^{\mathbb{I}2}_{LM(1)}}(n_{sim}) + t_{\mathbb{D}^{\mathbb{I}2}_{LM(2)}}(n_{obs}) \end{bmatrix} \quad [44]$$

C.2.4.3. Iterated EAI

The algorithm can also be iterated, with each iteration increasing the magnitude of the discrepancies between $\mathbb{D}^1_{FT}$s. For iterated EAI, the iteratively integrated $\mathbb{D}^1_{FT}$ (denoted using the notation $\mathbb{D}^{\mathbb{I}^n}$, where n is the number of iterations) is first calculated as follows:

$$\mathbb{D}^{\mathbb{I}^n}(x) = \int_{-\infty}^{x} \left(\int_{-\infty}^{x} \dots \int_{-\infty}^{x} \left(\int_{-\infty}^{x} \mathbb{D}^1_{FT}\, dFT \right) dFT \dots dFT \right) dFT \quad [45]$$

With the number of integral signs corresponding to the number of iterations n. This iteratively integrated $\mathbb{D}^1_{FT}$ can then be used in various statistics. As an example, the area between curve between iteratively integrated $\mathbb{D}^1_{FT_{sim}}$ and $\mathbb{D}^1_{FT_{obs}}$ (denoted as $\mathbb{D}^{\mathbb{I}^n}_{sim}$ and $\mathbb{D}^{\mathbb{I}^n}_{obs}$ respectively) can be calculated as follows:

$$A_{\mathbb{D}^{\mathbb{I}^n}}\left(\mathbb{D}^{\mathbb{I}^n}_{sim}, \mathbb{D}^{\mathbb{I}^n}_{obs}\right) = \int_{D_{min}}^{D_{max}} \left| \mathbb{D}^{\mathbb{I}^n}_{sim} - \mathbb{D}^{\mathbb{I}^n}_{obs} \right| dFT \quad [46]$$

The $A_{\mathbb{D}^{\mathbb{I}^n}}\left(\mathbb{D}^{\mathbb{I}^n}_{sim}, \mathbb{D}^{\mathbb{I}^n}_{obs}\right)$ can then be substituted in place of $A_{\mathbb{D}^{\mathbb{I}}}\left(\mathbb{D}^{\mathbb{I}}_{sim}, \mathbb{D}^{\mathbb{I}}_{obs}\right)$ for the $\mathbb{D}^2$ test statistics $t_{\mathbb{D}^2_{AWS}}$ in equation [34].

For this study, amplified Wasserstein's statistics ($t_{\mathbb{D}^2_{AWS}}$), amplified Kolmogorov-Smirnov statistics ($t_{\mathbb{D}^2_{AKS}}$) and amplified Distance from Median ($t_{\mathbb{D}^2_{ALM}}$) between $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ were calculated. Iterated amplifications have also been implemented, with the number of iterations $n = 1$ (i.e., no iteration), 2, 3 and 4 being used. Truncation was applied during the calculation of this statistic, with the quantile cut-off points y_q set at $y_q = 0.01$, 0.008, 0.006, 0.005, 0.0045, 0.004, 0.0035, 0.003, 0.0025, 0.002, 0.0015, 0.001, 0.0005, 0.0004, 0.0003, 0.0002, 0.0001 and 0. Overall, the test statistic from this class of statistics were kept in a 3-dimensional array of size $n_{sim} \times \left(3 \times 4 \times l_{y_q}\right) \times n_{obs}$, where l_{y_q} is the number of quantiles tested (i.e. $l_{y_q} = 18$ in this case).

C.2.5. Moment based Statistics

C.2.5.1. The Basics of Moments

In statistics, the moments of a distribution can be defined as quantitative measures that describe the shape of a distribution (Ramsey et al., 2002). While some of the lower moments such as arithmetic mean (Legendre, 1805; Plackett, 1958) and variance (Bienaymé, 1867)

have been discussed by various authors, the generalized concept of moments was first formalized by Tchebichef (1874) and Tchebychef (1907) where the author described the relationship of the asymptotic value of the integral of a function with the mass distribution within a material. In particular, Tchebichef (1874) and Tchebychef (1907) are evaluating the integral of this form:

$$\mu = \int_{-\infty}^{\infty} x^m f(x)\, dx \qquad [47]$$

The asymptotic value μ would later develop into the concept of "moments" in statistics. With a finite set of random variables, the raw sample moments (denoted as μ_r) of a vector $\boldsymbol{x}$ can be defined as the averages of the power of the random variables (Ramsey et al., 2002):

$$\mu_r(\boldsymbol{x}, m) = \frac{1}{n} \sum_{i=1}^{n} x_i^m \qquad [48]$$

One notable examples of moment is the arithmetic mean of the random variable, which is defined as $\mu_r(\boldsymbol{x}, 1)$. This definition of moment is not translation-invariant however; adding a constant value to each of the x_i in $\boldsymbol{x}$ changes the μ_r (Ramsey et al., 2002; Zellinger et al., 2017). This means if the distribution is shifted along the x-axis the μ_r would change, which is undesirable if the shape and spread of the distribution is the properties of interest. For this reason, the random variables were centralized at the mean of the random variables, yielding the concept of "central moment" (μ_c), defined as the moment around the mean (Ramsey et al., 2002):

$$\mu_c(\boldsymbol{x}, m) = \frac{1}{n} \sum_{i=1}^{n} \big(x_i - \mu_r(\boldsymbol{x}, 1)\big)^m \qquad [49]$$

Where $\mu_r(\boldsymbol{x}, 1)$ is the arithmetic mean of $\boldsymbol{x}$. One notable example is variance, where under this notation is defined as $\mu_c(\boldsymbol{x}, 2)$ (Ramsey et al., 2002). Similarly, if the location and spread of the distribution are not the aspects of interest, a scale invariant measure was required, and for this the random variable can be further transformed by scaling the random variable with the measure of spread of the distribution. This would yield a "standardized moment" (denoted as μ_s), which is defined as follows:

$$\mu_s(\boldsymbol{x}, m) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_r(\boldsymbol{x}, 1)}{\sqrt{\mu_c(\boldsymbol{x}, 2)}} \right)^m \qquad [50]$$

Where $\mu_c(\boldsymbol{x}, 2)$ is the variance of the distribution. Some examples of standardized moments include skewness ($\mu_s(\boldsymbol{x}, 3)$) and kurtosis ($\mu_s(\boldsymbol{x}, 4)$) (Ramsey et al., 2002).

From Tchebichef's (1874) analogy and further development of the "problem of moments" (i.e., loosely speaking, given a sequence of moments, find the distribution $f(x)$ that produces said sequence of moments), the concept of moment eventually became tied to the shape of a distribution (Schmüdgen, 2020; Tian et al., 2017). This property also suggested the possibility of a class of statistics that could be utilized to test the equality of distributions.

C.2.5.2. Test Based on Differences in Sample Moments

Given that the distributions are internally consistent (i.e., the dispersion of distributions $V(\mathbb{D})$ is finite and quantifiable, analogous to how a set of normally distribution random variables have a finite variance), theoretically if the proposed model for the genetic architecture $[\mathbb{k}, \mathbb{a}, \mathbb{b}]$ matches the underlying genetic architecture $Q(\mathbb{k}, \mathbb{a}, \mathbb{b})$, then the shape of the $\mathbb{D}^1_{FT_{sim}}$ is similar to that obtained from the observed phenotypes $\mathbb{D}^1_{FT_{obs}}$. Therefore, in theory, the moments of the random variables that produce the former should also match up with those of the latter, minimizing the differences between the moments (Figure C.12). This becomes the basis of the "Differences in Sample Moments" test.

Given a power of moment m, a vector of test statistics from observed phenotype $\boldsymbol{ft_{obs}}$ and simulated phenotype $\boldsymbol{ft_{sim}}$, and a function for the calculation of moment $\mu(\boldsymbol{ft}, m)$, the difference in sample moments between $\boldsymbol{ft_{sim}}$ and $\boldsymbol{ft_{obs}}$ (denoted as t_{DM}) is defined as follows:

$$t_{DM}(\boldsymbol{ft_{sim}}, \boldsymbol{ft_{obs}}, m) = |\mu(\boldsymbol{ft_{sim}}, m) - \mu(\boldsymbol{ft_{obs}}, m)| \quad [51]$$

The function $\mu(\boldsymbol{ft}, m)$ used in equation [51] could be that of raw, central or standardized moments, with its method of calculation defined in equation [48], [49] and [50] respectively. Truncation can also apply onto this test, where only the data points larger than an x-axis cut-off point (x_{yq}) were utilized in the sample moment calculation.

The test statistics for the differences in sample moments between $\boldsymbol{FT_{sim}}$ and $\boldsymbol{FT_{obs}}$ for moment m (denoted as $t_{\mathbb{D}^2_{DM}}(m)$) is defined as a 2-dimensional array of size $n_{sim} \times n_{obs}$ as follows:

$$t_{\mathbb{D}^2_{DM}}(m) = \begin{bmatrix} t_{DM}(\boldsymbol{FT_{sim}}(1), \boldsymbol{FT_{obs}}(1), m) & t_{DM}(\boldsymbol{FT_{sim}}(1), \boldsymbol{FT_{obs}}(2), m) & \cdots & t_{DM}(\boldsymbol{FT_{sim}}(1), \boldsymbol{FT_{obs}}(n_{obs}), m) \\ t_{DM}(\boldsymbol{FT_{sim}}(2), \boldsymbol{FT_{obs}}(1), m) & t_{DM}(\boldsymbol{FT_{sim}}(2), \boldsymbol{FT_{obs}}(2), m) & \cdots & t_{DM}(\boldsymbol{FT_{sim}}(2), \boldsymbol{FT_{obs}}(n_{obs}), m) \\ \vdots & \vdots & \ddots & \vdots \\ t_{DM}(\boldsymbol{FT_{sim}}(n_{sim}), \boldsymbol{FT_{obs}}(1), m) & t_{DM}(\boldsymbol{FT_{sim}}(n_{sim}), \boldsymbol{FT_{obs}}(2), m) & \cdots & t_{DM}(\boldsymbol{FT_{sim}}(n_{sim}), \boldsymbol{FT_{obs}}(n_{obs}), m) \end{bmatrix} \quad [52]$$

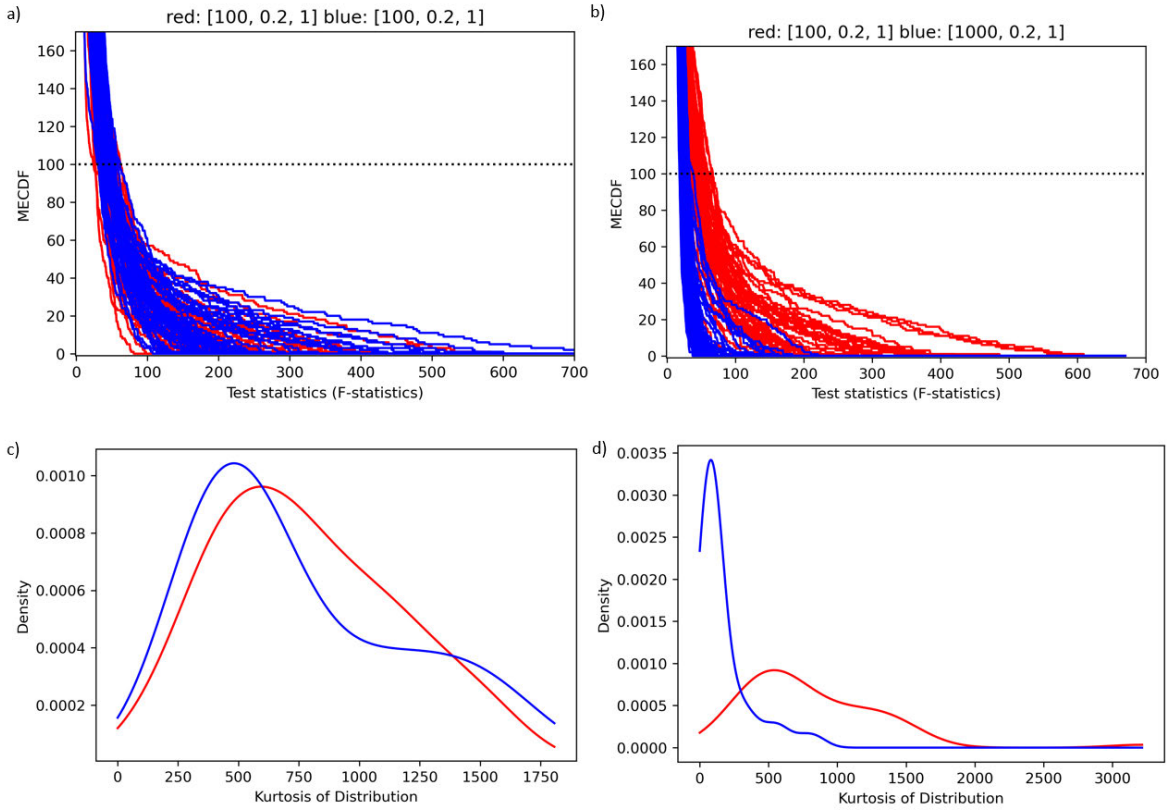

Figure C.12: The mechanism of "differences in moment" test. Figure (a) and (b) illustrated the $MECDF$s of the test statistics, whereas (c) and (d) illustrated the distribution of the kurtosis of the $MECDF$s. In (a) the red and blue $MECDF$s were generated from the same genetic architecture parameters of $Q(100, 0.2, 1)$, whereas in (b) the red and blue $MECDF$s were generated from different genetic architecture parameters, with the red being $Q(100, 0.2, 1)$ and blue being $Q(1000, 0.2, 1)$. The distribution of kurtosis of $MECDF$s in (c) are generated from those in (a), and those in (d) are generated from those in (b). The black dotted lines were used to truncate the random variables for the calculation of the kurtosis.

Where $\boldsymbol{FT_{sim}}(i)$ and $\boldsymbol{FT_{obs}}(j)$ are the i^{th} and j^{th} rows of $\boldsymbol{FT_{sim}}$ and $\boldsymbol{FT_{obs}}$, respectively.

For this study, the differences in raw moments, central moments and standardized moments were calculated. A multitude of powers of moments ms were tested in this study. For the raw moments, 10 ms were used: $m = 1, 2, 3, 4, 5, 6, 7, 8, 9$ and 10; while for central moments; 9 ms were used: $m = 2, 3, 4, 5, 6, 7, 8, 9$ and 10; and standardized moment, 8 ms were used: $m = 3, 4, 5, 6, 7, 8, 9$ and 10. Truncation had also been applied in this calculation, with quantile cut-off points y_q set at $y_q = 0.2, 0.17, 0.15, 0.13, 0.1, 0.07, 0.05, 0.03, 0.02, 0.01$ and

0.005. The resulting test statistics is a 3-dimensional array of size $n_{sim} \times \left((10 + 9 + 8) \times l_{y_q}\right) \times n_{obs}$ where $l_{y_q} = 10$ in this study.

C.2.5.3. Statistics based on Fractional Moments

Many of the most familiar moments such as mean, variance and kurtosis deal with the power m being natural numbers (Ramsey, 2002). One interesting observation that could be made on the equations for the calculation of the sample moments (i.e., equation [48], [49] and [50]) lies in the exponent m. Provided the bases of the exponentiation in these equations are nonnegative, there is no reason to restrict the m into positive integers. The equation is still properly defined for any value of m, and this includes non-integers. This observation introduces the concept of fractional moments, where the moment calculation no longer restricted to positive integers (Consortini and Rigal, 1998; Dremin, 1994). The equations for fractional sample moments are defined as in equation [48], [49] and [50].

Fractional moments shared many properties as in integer moments, such as similarity in magnitude of moments for similar distributions, which allows their use in the testing of equality of distribution (Figure C.13). Unlike integer moments however, the continuous and smooth nature of the fractional moments allowed more types of operations to be conducted. This opened up a new class of tests that could be used to test the equality in distributions.

With the fractional moment, one can define the equations for calculation of moments as a continuous function of m, which allows additional tests that can be used. For example, one can conduct "Wasserstein"-like statistics test where the area between the curves of the fractional moments (denoted as t_{FMW}) can be calculated as follows:

$$t_{FMW}(\boldsymbol{ft_{sim}}, \boldsymbol{ft_{obs}}, m) = \int_{m_{min}}^{m_{max}} |\mu(\boldsymbol{ft_{sim}}, m) - \mu(\boldsymbol{ft_{obs}}, m)| \, dm \qquad [53]$$

where m_{min} and m_{max} are the minimum and maximum moments tested. The reason for restricting the range of integration is due to the divergent nature of the integral (i.e., the area between curve increases without bound). An example of area between the curves for the fractional moment is provided in Figure C.14.

The test statistics for area between curve of fractional moment function (denoted as $t_{\mathbb{D}^2_{FMW_m}}$) is a 2-dimensional array of size $n_{sim} \times n_{obs}$ structured as follows:

$$t_{\mathbb{D}^2_{FMW_m}}$$

$$= \begin{bmatrix} t_{FMW}(\boldsymbol{FT_{sim}}(1), \boldsymbol{FT_{obs}}(1), m) & t_{FMW}(\boldsymbol{FT_{sim}}(1), \boldsymbol{FT_{obs}}(2), m) & \cdots & t_{FMW}(\boldsymbol{FT_{sim}}(1), \boldsymbol{FT_{obs}}(n_{obs}), m) \\ t_{FMW}(\boldsymbol{FT_{sim}}(2), \boldsymbol{FT_{obs}}(1), m) & t_{FMW}(\boldsymbol{FT_{sim}}(2), \boldsymbol{FT_{obs}}(2), m) & \cdots & t_{FMW}(\boldsymbol{FT_{sim}}(2), \boldsymbol{FT_{obs}}(n_{obs}), m) \\ \vdots & \vdots & \ddots & \vdots \\ t_{FMW}(\boldsymbol{FT_{sim}}(n_{sim}), \boldsymbol{FT_{obs}}(1), m) & t_{FMW}(\boldsymbol{FT_{sim}}(n_{sim}), \boldsymbol{FT_{obs}}(2), m) & \cdots & t_{FMW}(\boldsymbol{FT_{sim}}(n_{sim}), \boldsymbol{FT_{obs}}(n_{obs}), m) \end{bmatrix} \quad [54]$$

To ensure the monotonicity of the fractional moments, and to avoid the calculation of negative number power to a non-integer, only the raw moments been utilized in the fractional moments in this study. Further work could extend this statistic toward the central and standardized moments, and for negative random variables. For this study, the m_{min} is set at 1 and m_{max} set at 5. Truncation had also been utilized in this calculation, with quantile cut-off points y_q set at $y_q = 0.2, 0.17, 0.15, 0.13, 0.1, 0.07, 0.05, 0.03, 0.02, 0.01$ and 0.005. The resulting test statistics is a 3-dimensional array of size $n_{sim} \times l_{y_q} \times n_{obs}$ where $l_{y_q} = 10$ in this case.

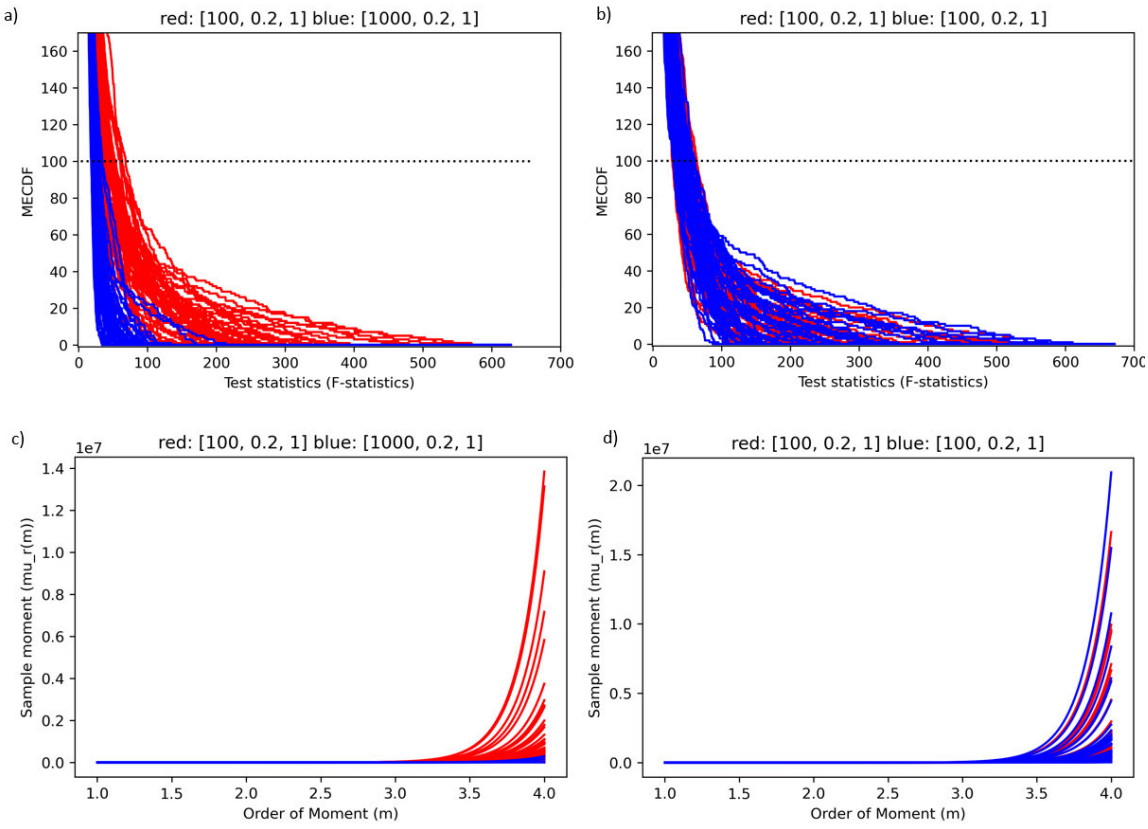

Figure C.13: An application of fractional moments in testing of equality of distributions. Figure (a) and (b) represent the $MECDF$s of the test statistics, whereas (c) and (d) illustrated the distribution of the kurtosis of the $MECDF$s, with the genetic architecture parameters utilized being defined in Figure C.12. The fractional moments of $MECDF$s in (a) are illustrated in (c), and those in (b) are illustrated in (d). The black dotted lines were used to truncate the random variables for the calculation of the fractional moments.

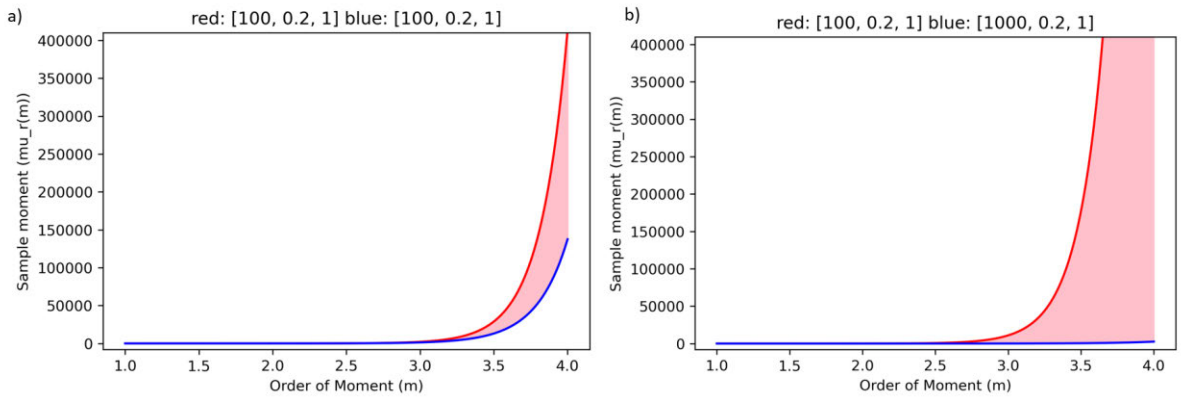

Figure C.14: The area between the curves of fractional moments, shaded in pink regions, being used to test the equality in distributions. In (a) the fractional moments of both red and blue curves were generated from the same genetic architecture parameters $Q(100, 0.2, 1)$, whereas for (b) the red curve was generated from $Q(100, 0.2, 1)$ and for blue from $Q(1000, 0.2, 1)$.

C.3. Stacking up the Statistics

From the several types of statistics proposed in the previous sections, there exists some statistics that can be "stacked". This involves the combinations of various elements from other statistics into a new statistic. From this process, even more types of statistics that test the equality of distributions can be built. This is further aided by the emergent properties of $\mathbb{D}_{FT}^{2}$ that do not exist in $\mathbb{D}_{FT}^{1}$, which allows combinations of elements from other statistics.

For example, Wasserstein's statistics involved the calculation of the area of difference between the curves, with the rationale that said area is minimized if two distributions came from the same underlying distribution. This however also hinted that the area under the curve between the two distributions should also be similar. From this observation, an alternative statistic that could be applied is to simply compare the area under the curves between the two distributions. For this, given a quantile cut-off point y_q and its corresponding x-axis point x_{yq}, the area under the curves (denoted as $a_{\mathbb{D}^1}$) is the integral of $\mathbb{D}_{FT}^{1}$ from x_{yq} onward:

$$a_{\mathbb{D}^1}(\mathbb{D}_{FT}^{1}) = \int_{x_{yq}}^{\infty} \mathbb{D}_{FT}^{1}\ dFT \qquad [55]$$

For this study, the vectors of $a_{\mathbb{D}^1}$ for $\mathbb{D}_{FT_{sim}}^{2}$ and $\mathbb{D}_{FT_{obs}}^{2}$ were denoted as $\boldsymbol{a}_{\mathbb{D}_{sim}^{2}}$ and $\boldsymbol{a}_{\mathbb{D}_{obs}^{2}}$ respectively, structured as follows:

$$\boldsymbol{a}_{\mathbb{D}_{sim}^{2}} = \left[a_{\mathbb{D}^1}\left(\mathbb{D}_{FT_{sim_1}}^{1}\right), a_{\mathbb{D}^1}\left(\mathbb{D}_{FT_{sim_2}}^{1}\right), a_{\mathbb{D}^1}\left(\mathbb{D}_{FT_{sim_3}}^{1}\right), \ldots, a_{\mathbb{D}^1}\left(\mathbb{D}_{FT_{sim_{n_{sim}}}}^{1}\right)\right] \qquad [56]$$

$$\boldsymbol{a}_{\mathbb{D}^2_{obs}} = \left[a_{\mathbb{D}^1}\left(\mathbb{D}^1_{FT_{obs_1}}\right), a_{\mathbb{D}^1}\left(\mathbb{D}^1_{FT_{obs_2}}\right), a_{\mathbb{D}^1}\left(\mathbb{D}^1_{FT_{obs_3}}\right), \ldots, a_{\mathbb{D}^1}\left(\mathbb{D}^1_{FT_{obs_{n_{obs}}}}\right)\right] \quad [57]$$

The differences in area under the curve between the vectors of $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$ (denoted as $t_{\mathbb{D}^2_{ADF}}$) can then be calculated and defined as a 2-dimensional array of size $n_{sim} \times n_{obs}$ defined as follows:

$$t_{\mathbb{D}^2_{ADF}} = \begin{bmatrix} \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(1) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(1)\right| & \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(1) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(2)\right| & \cdots & \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(1) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(n_{obs})\right| \\ \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(2) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(1)\right| & \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(2) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(2)\right| & \cdots & \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(2) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(n_{obs})\right| \\ \vdots & \vdots & \ddots & \vdots \\ \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(n_{sim}) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(1)\right| & \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(n_{sim}) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(2)\right| & \cdots & \left|\boldsymbol{a}_{\mathbb{D}^2_{sim}}(n_{sim}) - \boldsymbol{a}_{\mathbb{D}^2_{obs}}(n_{obs})\right| \end{bmatrix} \quad [58]$$

Similarly, the location from median test can also be used to compare the area under the curve of $\mathbb{D}^1_{FT}$s in $\mathbb{D}^2_{FT_{sim}}$ and $\mathbb{D}^2_{FT_{obs}}$, with the vectors $\boldsymbol{a}_{\mathbb{D}^2_{sim}}$ and $\boldsymbol{a}_{\mathbb{D}^2_{obs}}$ being used in place of $\boldsymbol{x_{yq_{sim}}}$ and $\boldsymbol{x_{yq_{obs}}}$:

$$t_{\mathbb{D}^2_{LM(1)}} = \left[t_{LM}\left(\boldsymbol{a}_{\mathbb{D}^2_{sim}}(1), \boldsymbol{a}_{\mathbb{D}^2_{obs}}\right) \quad t_{LM}\left(\boldsymbol{a}_{\mathbb{D}^2_{sim}}(2), \boldsymbol{a}_{\mathbb{D}^2_{obs}}\right) \quad \cdots \quad t_{LM}\left(\boldsymbol{a}_{\mathbb{D}^2_{sim}}(n_{sim}), \boldsymbol{a}_{\mathbb{D}^2_{obs}}\right)\right] \quad [59]$$

$$t_{\mathbb{D}^2_{LM(2)}} = \left[t_{LM}\left(\boldsymbol{a}_{\mathbb{D}^2_{obs}}(1), \boldsymbol{a}_{\mathbb{D}^2_{sim}}\right) \quad t_{LM}\left(\boldsymbol{a}_{\mathbb{D}^2_{obs}}(2), \boldsymbol{a}_{\mathbb{D}^2_{sim}}\right) \quad \cdots \quad t_{LM}\left(\boldsymbol{a}_{\mathbb{D}^2_{obs}}(n_{obs}), \boldsymbol{a}_{\mathbb{D}^2_{sim}}\right)\right] \quad [60]$$

The final test statistic for the distance from median (denoted as $t_{\mathbb{D}^2_{LMA}}$) is a $n_{sim} \times n_{obs}$ array and is calculated as in equation [21].

This technique allows the multiplication of the number of statistics that could be done, making the construction of a battery of 703 statistics designed to test the equality of tail distribution between the distribution from observed phenotypes $\mathbb{D}^2_{FT_{obs}}$ and simulated phenotypes $\mathbb{D}^2_{FT_{sim}}$ possible.

Appendix D. The Selection of Proposed Genetic Architecture Parameters

The aim for this appendix section is to provide a layout of methodology of sampling the genetic architecture parameters that were tested for the estimation of genetic architecture parameters. This includes the number of QTL (denoted as $\mathbb{k}$) and the shape parameters for the distribution of the QTL effect sizes (denoted as $\mathbb{a}$). This section will also detail the methodology and a simplified example of generating a "Geom-linear" sequence for the testing of number of QTL. This methodology will be used to generate a grid of $[\mathbb{k}, \mathbb{a}]$ combinations that were brute force searched during the estimation of genetic architecture parameters.

D.1. The selection of $\mathbb{k}$ and the Rationale of a "Geom-linear" Sequence

For this study, the possible range of $\mathbb{k}$ can span from 0 to total number of markers M. In the idealized situation all possible values of $\mathbb{k}$ would be tested. Given the large number of M however, this introduces a large parameter space that needs to be tested, which could impede the feasibility of the algorithm. This however can be resolved by choosing some of the values of $\mathbb{k}$ that were tested by the algorithm.

One possible approaches is a series of equally spaced $\mathbb{k}$, which ensures consistent coverages of all possible values for this parameter. This series suffered from poor scalability however; as the number of operations increases linearly with M, any increment in M would quickly overwhelm the practicality of the algorithm. For example, if $M = 5{,}000$ and the spacing between $\mathbb{k}$ is 100, this means 50 $\mathbb{k}$s need to be tested, and if $M = 500{,}000$, with the same spacing there were 5000 $\mathbb{k}$s that need to be tested, severely reducing the feasibility of the algorithm. Furthermore, given a fixed amount of change in the value of parameter $\mathbb{k}$, the effects of such change is greater if $\mathbb{k}$ is small (Figure D.1). Therefore, from the perspective of investigating the effects of $\mathbb{k}$ on the output, choosing an equal spacing for $\mathbb{k}$ would result in a poor resolution for the small values (i.e., overly large changes in outputs for each $\mathbb{k}$), and an

unnecessarily high resolution for larger values. Thus, an equal spacing of 𝕜 is not appropriate for this purpose.

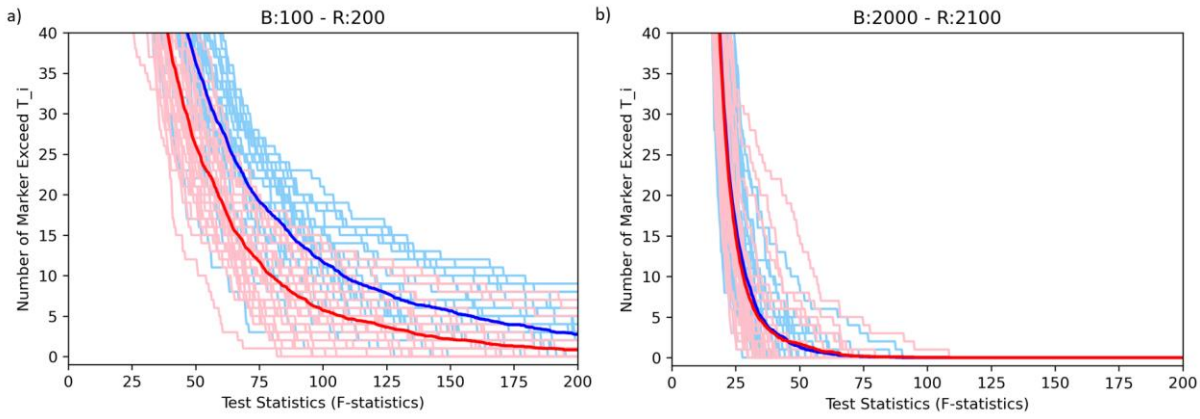

Figure D.1: The effects of changing a fixed amount of parameter 𝕜 on the $MECDF$ and their asymptotic distributions. The light blue $MECDF$s in (a) is generated from the genetic architecture of $Q(100, 0.5,1)$, whereas the $MECDF$s in pink is generated from $Q(200, 0.5,1)$, with a differences of 100 between the two 𝕜s. Whereas for (b) the light blue $MECDF$s are generated from the genetic architecture parameter of $Q(2000, 0.5,1)$ and the pink $MECDF$s are generated from $Q(2100, 0.5,1)$, the same amount of differences between the two 𝕜s. The red and blue distributions in both graphs are the asymptotic distributions of the pink and light blue $MECDF$s respectively.

Conversely, one can also choose geometric series for 𝕜 (i.e., two consecutive 𝕜s have equal ratio), which increases the density for small 𝕜 while decreasing the density for large 𝕜. This serie has a good scalability, quickly achieving large M with relatively small number of tested 𝕜s. Indeed, the expected number of 𝕜 that need to be tested increases logarithmically with larger M; using the previous example of 50 𝕜s for $M = 5{,}000$, the number of 𝕜s that need to be tested for $M = 500{,}000$ would be 78 (i.e. $\left\lceil 50 * log\left(\frac{500{,}000}{5{,}000}\right)\right\rceil = 78$).

This series has suffered from inconsistent coverage across the range of 𝕜 however; this series tends to be overly sparse for large 𝕜, overly dense for small 𝕜, while failing to maintain consistency of spacing. From the example of $M = 500{,}000$, 23 of them have values less than 50, while only 14 of them have values more than 50,000. This corresponds to an average distance of 2.17 markers per consecutive pairs of 𝕜s within the smallest 50 markers (50/30 = 2.17) and 32142.86 markers per consecutive pairs of 𝕜s for those larger than 50,000 ((500,000 – 50,000)/14 = 32142.86). The massive discrepancy of distance between 𝕜s could reduce the accuracy of the algorithm, especially if the trait is polygenic. Thus, a geometric series for 𝕜 was also unsuitable for this purpose.

One possible solution for this issue is to build a new progression that combines the benefits from both linear and geometric series. One such series, which was termed "geom-linear progression", would combine the consistent coverage of linear progression and scalability of geometric progression. Three numbers were required to build a geom-linear progression: the starting value C_0, the common ratio between geometric node Cr and number of linear nodes within each geometric node pairs, excluding the geometric nodes themselves, Cl. As an example, if $C_0 = 10$, $Cr = 5$ and $Cl = 4$, a geom-linear series can be built by first building a geometric progression:

$$\begin{aligned}(C_0, C_1, C_2, C_3, \dots) &= (C_0, C_0(Cr), C_0(Cr)^2, C_0(Cr)^3, \dots) \\ &= (10, 10(5), 10(5)^2, 10(5)^3, \dots) \\ &= (10, 50, 250, 1250, \dots) \end{aligned} \quad [1]$$

For each pair of C_n and C_{n+1}, the common differences of linear nodes for each geometric node pairs, denoted as Cd_n, are calculated. Using $C_0 = 10$ and $C_1 = 50$ pair as example, the Cd_1 is calculated as follows:

$$\begin{aligned}Cd_1 &= \frac{C_1 - C_0}{Cl + 1} \\ &= \frac{50 - 10}{4 + 1} \\ &= 8 \end{aligned} \quad [2]$$

The Cd_n was then be calculated from the linear progression between C_n and C_{n+1} up to Cl^{th} term, excluding the geometric node itself:

$$\left(C_{n_1}, C_{n_2}, C_{n_3}, \dots, C_{n_{Cl-1}}, C_{n_{Cl}}\right) = (C_n + Cd, C_n + 2Cd, \dots, C_n + (Cl - 1) * Cd, C_n + Cl * Cd) \quad [3]$$

And in the example above, with $C_0 = 10$, $C_1 = 50$ and $Cd_1 = 8$, the linear progression was as follows:

$$\begin{aligned}\left(C_{0_1}, C_{0_2}, C_{0_3}, C_{0_4}\right) &= \big(10 + 8, 10 + 2(8), 10 + 3(8), 10 + 4(8)\big) \\ &= (18, 26, 34, 42) \end{aligned} \quad [4]$$

The sequence in equation [4] can be inserted between C_0 and C_1 in equation [1]. This process was then repeated for all C_n and C_{n+1} pairs. The end result of the sequence would have this pattern:

$$\left(C_0, C_{0_1}, C_{0_2}, \dots, C_{0_{Cl}}, C_1, C_{1_1}, C_{1_2}, \dots, C_{1_{Cl}}, C_2, C_{2_1}, \dots, C_{2_{Cl}}, \dots\right) \quad [5]$$

From the example provided above, the $Cd_1 = 8$, $Cd_2 = 40$, $Cd_3 = 200$ and so on, and final sequence was as follows:

$$(10, 18, 26, 34, 42, 50, 90, 130, 170, 210, 250, 450, 650, 850, 1050, 1250 \dots) \quad [6]$$

The geom-linear series can also be seen as a finite linear approximation of an exponential curve, with the example featured above illustrated in Figure D.2. Using the example of 50 𝕜s for $M = 5{,}000$ with $C_0 = 1$ and $Cr = 10$, the resulting Cl was 15, and $Cd_n = 0.5625 * (10)^n$. If the same C_0, Cr, Cl and Cd_n are to be used on $M = 500{,}000$ the number of 𝕜s that need to be tested would be 82.

Compared to the 5000 𝕜s from linear progression, brute forcing the sequence of 82 𝕜s from geom-linear series have a better feasibility. Compared to geometric series, a "geom-linear" series also has a better consistency in coverage of 𝕜s. Compared to the 23 𝕜s with a value less than 50 in a geometric series, a geom-linear series has only 17 𝕜s. This translated into an average of 2.94 markers per consecutive pairs of 𝕜s. Whereas for 𝕜s larger than 50,000, compared to 14 𝕜s from a geometric series, geom-linear series also has 17 𝕜s, which translated to an average of 26470.59 markers per consecutive pairs of 𝕜s. Compared to the range of average number of markers from 2.17 to 32142.86 per consecutive pairs of 𝕜s, the geom-linear series produces a less extreme range of average number of markers per consecutive pairs of 𝕜s, therefore provides a better consistency of coverages. Therefore, the geom-linear progression could be used to sample the 𝕜s for the brute-force algorithm.

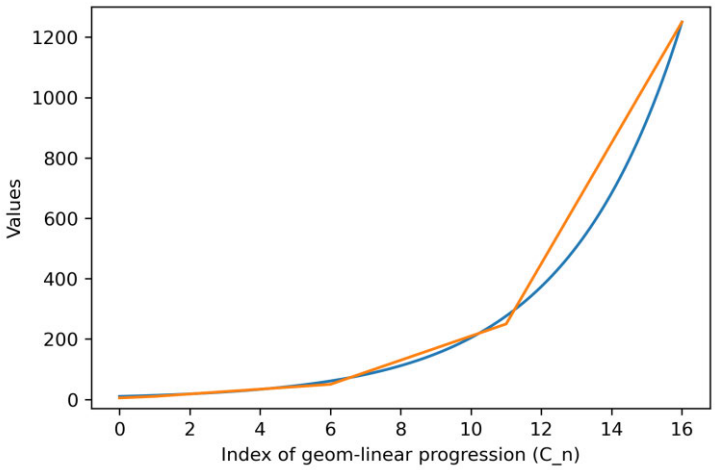

Figure D.2: An example of the "geom-linear" progression (orange line), in comparison with the regular geometric progression (blue line).

D.2. The Selection of $\mathbb{a}$

In a situation similar to parameter $\mathbb{k}$, testing each possible value of the shape parameter for QTL effect size distribution ($\mathbb{d}_{QTL}$), $\mathbb{a}$, is also not possible. This is due to the fact that $\mathbb{a}$ is a continuous variable that can take any positive real number (Mun, 2012), and thus impractical to brute-force. Therefore, the algorithm should only select a number of $\mathbb{a}$s that would be tested. Despite this, as $\mathbb{a}$ can range from 0 to infinity, that would still leave an infinite number of $\mathbb{a}$s that need to be tested.

The range of $\mathbb{a}$s that need to be tested can be restricted by inspecting the probability density function of the gamma distribution. As the gamma distribution is used to model the $\mathbb{d}_{QTL}$ that have large number of QTL with small effect sizes and small number of QTL with large effect sizes, this requirement has placed a restriction on the possible range of $\mathbb{a}$s that could fulfil such purpose.

Given the shape parameter $\mathbb{a}$ and scale parameter $\mathbb{b}$, the probability density function for a gamma distribution is defined as follows (Mun, 2012):

$$\Gamma(x, \mathbb{a}, \mathbb{b}) = \frac{\mathbb{b} * (\mathbb{b}x)^{\mathbb{a}-1} * e^{-\mathbb{b}x}}{\gamma(\mathbb{a})} \quad [7]$$

Where $\gamma(\mathbb{a})$ is the gamma function of $\mathbb{a}$, which is always positive for all positive values of $\mathbb{a}$.

The restriction for parameter $\mathbb{a}$ in the gamma distribution can be observed in its numerator, more precisely the exponent of the $(\mathbb{b}x)^{\mathbb{a}-1}$ part. The exponent $\mathbb{a} - 1$ hinted that if $\mathbb{a} < 1$, then the exponent would become negative, and this places the $\mathbb{b}x$ into the reciprocal (i.e. $\frac{1}{\mathbb{b}x}$). This reciprocal function means as x become smaller, $\Gamma(x, \mathbb{a}, \mathbb{b})$ would become larger, just like the model with large number of QTL with small effect sizes. This is further aided by the $e^{-\mathbb{b}x}$, where if $\mathbb{b}$ is positive, then it would reach its maximum of 1, when $x = 0$.

If $\mathbb{a} > 1$, the $\mathbb{b}x$ would remain in the numerator of $\Gamma(x, \mathbb{a}, \mathbb{b})$, thus for the range of x-axis with value less than its mode the $\Gamma(x, \mathbb{a}, \mathbb{b})$ decreases, reaching zero as $x = 0$. If the purpose of gamma distribution is to model the $\mathbb{d}_{QTL}$ with a large number of QTL with small effect size, then the condition of $\mathbb{a} > 1$ might not be appropriate for such modelling. If $\mathbb{a} = 1$, the gamma distribution simplifies into an exponential distribution, which have been used in previous attempts to estimate the $\mathbb{d}_{QTL}$ (Hall et al., 2016; Mun, 2012). Thus, $\mathbb{a} = 1$ is also a feasible model. Examples of $\mathbb{a} < 1$, $\mathbb{a} = 1$, $\mathbb{a} > 1$ and $\mathbb{a} \gg 1$ were presented in Figure D.3.

With these observations, one can restrict the brute-force search of $\mathbb{a}$ within the range $0 < \mathbb{a} \leq 1$, which greatly increases the feasibility of the search algorithm. In fact, one could use a linear progression as a way to discretise the parameter into a finite number of $\mathbb{a}$s that needs to be tested.

Using the $\mathbb{k}$s sampled from a geom-linear progression and the discretised $\mathbb{a}$s sampled from linear progression, one could build a grid of $[\mathbb{k}, \mathbb{a}]$ combinations that could be brute-force searched to find the combinations of parameters that best fit the observed distribution from a GWAS experiment.

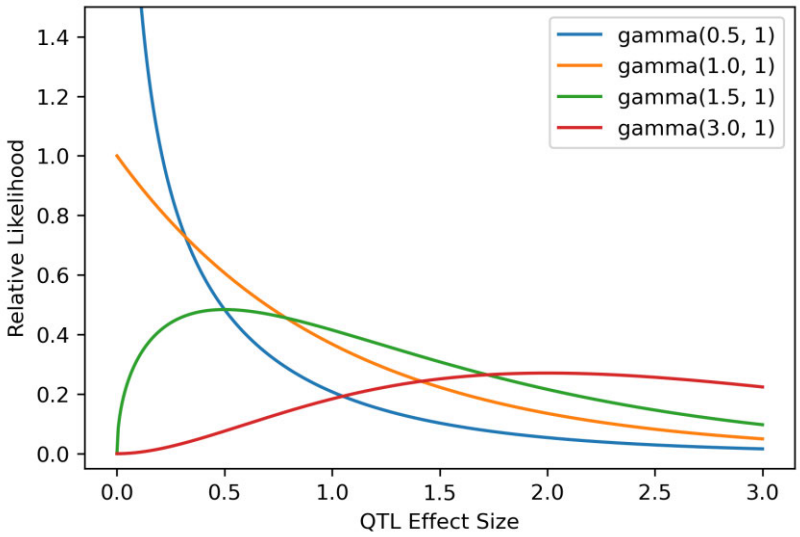

Figure D.3: Example of gamma distribution under varying shape parameter $\mathbb{a}$. Only the blue and the orange lines have the correct shape for the modelling of QTL effect size distributions.

Appendix E. Evaluation of Sizes of Sample Space for Additive, Non-additive and Inbreeding Coefficient

This appendix is to provide a description on the size of sample space of additive and non-additive effects and level of co-ancestry that needs to be evaluated by the genetic algorithm. The sample space for the non-additive genetic component is significantly larger than those of additive and inbreeding coefficients, and this appendix is to provide a mathematical proof for this assertion.

E.1. The Mathematical Proof

Given a number of sires N_m and number of dams N_f, the size of sample space for all possible values of additive and co-ancestry (denoted as $n_{\{A\}}$ and $n_{\{I\}}$ respectively) can be defined as the number of combinations of sires chosen with repetition, without considering the ordering of the sires. This can be calculated using the following binomial coefficient (Benjamin and Quinn, 2003):

$$n_{\{A\}} = n_{\{I\}} = \binom{N_m + N_f - 1}{N_f} = \frac{(N_m + N_f - 1)!}{(N_f)!\,(N_m - 1)!} \quad [1]$$

Where $\binom{n}{k}$ is the binomial coefficient of "n choose k" and $k!$ is denoted as the factorial of k. The size of sample space of all possible values of non-additive genetic value possible (denoted as $n_{\{D\}}$) was defined as number of combinations of sires chosen with repetition, but with consideration of ordering of the sires. Thus, the sample space for non-additive genetic values was defined as follows:

$$n_{\{D\}} = N_m^{N_f} \quad [2]$$

Let b be the binomial coefficient that represent both $n_{\{A\}}$ and $n_{\{I\}}$, which is defined as follows:

$$b = \binom{N_m + N_f - 1}{N_f} = \frac{(N_m + N_f - 1)!}{(N_f)!\,(N_m - 1)!} \quad [3]$$

The b in equation [3] can be expanded as follows:

$$\frac{(N_m + N_f - 1)!}{(N_f)!\,(N_m - 1)!} = \frac{(1)(2)(3)\dots(N_f - 1)(N_f)(N_f + 1)\dots(N_m + N_f - 3)(N_m + N_f - 2)(N_m + N_f - 1)}{\big((1)(2)(3)\dots(N_f - 2)(N_f - 1)(N_f)\big) * \big((1)(2)(3)\dots(N_m - 3)(N_m - 2)(N_m - 1)\big)}$$

$$= \frac{\cancel{(1)(2)(3)\dots(N_f - 1)(N_f)}(N_f + 1)\dots(N_f + N_m - 3)(N_f + N_m - 2)(N_f + N_m - 1)}{\cancel{\big((1)(2)(3)\dots(N_f - 2)(N_f - 1)(N_f)\big)} * \big((1)(2)(3)\dots(N_m - 3)(N_m - 2)(N_m - 1)\big)}$$

$$= \frac{(N_f + 1)(N_f + 2)(N_f + 3)\dots(N_f + N_m - 3)(N_f + N_m - 2)(N_f + N_m - 1)}{(1)(2)(3)\dots(N_m - 3)(N_m - 2)(N_m - 1)} \quad [4]$$

Let b^* be a new binomial coefficient defined as:

$$b^* = \binom{N_m + N_f}{N_f} = \frac{(N_m + N_f)!}{(N_f)!\,(N_m)!} \quad [5]$$

Which can be expanded as follows:

$$\frac{(N_m + N_f)!}{(N_f)!\,(N_m)!} = \frac{(1)(2)(3)\dots(N_f - 1)(N_f)(N_f + 1)\dots(N_m + N_f - 2)(N_m + N_f - 1)(N_m + N_f)}{\big((1)(2)(3)\dots(N_f - 2)(N_f - 1)(N_f)\big) * \big((1)(2)(3)\dots(N_m - 2)(N_m - 1)(N_m)\big)}$$

$$= \frac{\cancel{(1)(2)(3)\dots(N_f - 1)(N_f)}(N_f + 1)\dots(N_m + N_f - 2)(N_m + N_f - 1)(N_m + N_f)}{\cancel{\big((1)(2)(3)\dots(N_f - 2)(N_f - 1)(N_f)\big)} * \big((1)(2)(3)\dots(N_m - 2)(N_m - 1)(N_m)\big)}$$

$$= \frac{(N_f + 1)(N_f + 2)(N_f + 3)\dots(N_m + N_f - 2)(N_m + N_f - 1)(N_m + N_f)}{(1)(2)(3)\dots(N_m - 2)(N_m - 1)(N_m)} \quad [6]$$

Taking the ratio between b and b^* would yield the following value:

$$\frac{b}{b^*} = \frac{\binom{N_m + N_f - 1}{N_f}}{\binom{N_m + N_f}{N_f}}$$

$$= \frac{\cancel{(N_f + 1)(N_f + 2)(N_f + 3)\dots(N_f + N_m - 3)(N_f + N_m - 2)(N_f + N_m - 1)}}{\cancel{(1)(2)(3)\dots(N_m - 3)(N_m - 2)(N_m - 1)}}$$
$$\frac{\cancel{(N_f + 1)(N_f + 2)(N_f + 3)\dots(N_m + N_f - 2)(N_m + N_f - 1)}(N_m + N_f)}{\cancel{(1)(2)(3)\dots(N_m - 2)(N_m - 1)}(N_m)}$$

$$= \frac{1/1}{(N_m + N_f)/N_m}$$

$$= \frac{N_m}{N_m + N_f} \quad [7]$$

By rearranging [7], it could be shown that

$$b = \left(\frac{N_m}{N_m + N_f}\right) b^* \quad [8]$$

This identity is important to establish as it heavily simplifies the mathematics from here onward. Without the nuisance "-1" in the equation, and with sufficiently large N_m and N_f (i.e. $N_m > 2$ and $N_f > 2$), the b^* can be converted into a more tractable form using Stirling's approximation. This approximation is defined as follows (Pearson, 1924):

$$n! \cong \sqrt{2\pi n} * \left(\frac{n}{e}\right)^n \quad [9]$$

Where e is the Euler's number (i.e. $e \approx 2.71828$ ...). Substituting equation [9] into [8] yields the following:

$$\begin{aligned}
b &= \left(\frac{N_m}{N_m + N_f}\right) * b^* \\
&= \left(\frac{N_m}{N_m + N_f}\right) * \left(\frac{(N_m + N_f)!}{(N_f)!\,(N_m)!}\right) \\
&= \left(\frac{N_m}{N_m + N_f}\right) * \left(\frac{\left(\sqrt{2\pi(N_m + N_f)} * \left(\frac{N_m + N_f}{e}\right)^{N_m+N_f}\right)}{\left(\sqrt{2\pi N_f} * \left(\frac{N_f}{e}\right)^{N_f}\right) * \left(\sqrt{2\pi N_m} * \left(\frac{N_m}{e}\right)^{N_m}\right)}\right) \\
&= \left(\frac{N_m}{N_m + N_f}\right) * \left(\frac{1}{\sqrt{2\pi}}\right) * \left(\sqrt{\frac{N_m + N_f}{N_m N_f}}\right) * \left(\frac{e^{N_m+N_f}}{e^{N_m} * e^{N_f}}\right) * \left(\frac{(N_m + N_f)^{N_m+N_f}}{N_f^{N_f} * N_m^{N_m}}\right) \\
&= \left(\sqrt{\frac{N_m}{N_f(N_m + N_f)}}\right) * \left(\frac{1}{\sqrt{2\pi}}\right) * \left(\frac{e^{N_m+N_f}}{e^{N_m+N_f}}\right) * \left(\frac{(N_m + N_f)^{N_m} * (N_m + N_f)^{N_f}}{N_m^{N_m} * N_f^{N_f}}\right) \\
&= \frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m + N_f)}}\right) * \left(\left(\frac{N_m + N_f}{N_m}\right)^{N_m} * \left(\frac{N_m + N_f}{N_f}\right)^{N_f}\right) \quad [10]
\end{aligned}$$

To prove that $n_{\{D\}}$ grows faster than $n_{\{A\}}$ and $n_{\{I\}}$ with increasing large N_m and N_f, it needs to be shown that $N_m^{N_f}$ grows faster than [10]. Let d be a function that represent $n_{\{D\}}$:

$$d = N_m^{N_f} \quad [11]$$

With these definitions, function d can be said to grow faster than b if it fulfils the following condition:

$$\lim_{(N_m,N_f)\to(\infty,\infty)} \frac{b}{d} = \lim_{(N_m,N_f)\to(\infty,\infty)} \frac{\left(\frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m + N_f)}}\right) * \left(\left(\frac{N_m + N_f}{N_m}\right)^{N_m} * \left(\frac{N_m + N_f}{N_f}\right)^{N_f}\right)\right)}{N_m^{N_f}} = 0 \quad [13]$$

To evaluate this limit, some rearrangement and simplification of the function is required. This could be done as follows:

$$\frac{\left(\frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m+N_f)}}\right) * \left(\left(\frac{N_m+N_f}{N_m}\right)^{N_m} * \left(\frac{N_m+N_f}{N_f}\right)^{N_f}\right)\right)}{N_m^{N_f}}$$

$$= \frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m+N_f)}}\right) * \frac{\left(\frac{N_m+N_f}{N_m}\right)^{N_m}}{N_m^{N_f}} * \left(\frac{N_m+N_f}{N_f}\right)^{N_f}$$

$$= \frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m+N_f)}}\right) * \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m+N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m+N_f}{N_f}\right)^{N_f} \quad [14]$$

The evaluation of this limit also require the use of l'Hôpital's rule, which states the following (Lawlor, 2020):

$$\lim_{x\to c}\frac{f(x)}{g(x)} = \lim_{x\to c}\frac{f'(x)}{g'(x)} \quad [15]$$

The condition required for the application of l'Hôpital rule is that (1) the derivative of $f(x)$ and $g(x)$ (denoted as $f'(x)$ and $g'(x)$) exists at $x = c$, with (2) $\lim_{x\to c} f(x) = \lim_{x\to c} g(x) = 0$ and (3) $g'(c) \neq 0$. This function does not fulfil condition (2), but it could be transformed into a form that does fulfil such a requirement. Even so, l'Hôpital rule cannot be directly applied to this [13] as it is a multivariate function, whereas this rule could only be applied to a univariate function (Lawlor, 2020). For this reason, the limit in equation [13] needs to be split between the variables and inspect their behaviour of the limits. This is similar to taking the partial derivatives of [13] by N_m and N_f. The split limits were defined as follows:

$$\lim_{N_m\to\infty} \frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m+N_f)}}\right) * \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m+N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m+N_f}{N_f}\right)^{N_f} \quad [16]$$

$$\lim_{N_f\to\infty} \frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m+N_f)}}\right) * \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m+N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m+N_f}{N_f}\right)^{N_f} \quad [17]$$

The limit defined in function [16] and [17] comprises of products of several sub-functions, which, by using the distributive property of limit, could be inspected individually, most notably the following sub-functions:

$$s_1 = \sqrt{\frac{N_m}{N_f\left(N_m + N_f\right)}} \quad [18]$$

and

$$s_2 = \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m + N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m + N_f}{N_f}\right)^{N_f} \quad [19]$$

For limit [16], when $N_m \to \infty$, the sub-function s_1 in [18] can be evaluated as following:

$$\begin{aligned} \lim_{N_m \to \infty} s_1 &= \lim_{N_m \to \infty} \sqrt{\frac{N_m}{N_f\left(N_m + N_f\right)}} \\ &= \lim_{N_m \to \infty} \sqrt{\frac{N_m}{N_m N_f + N_f^2}} \\ &= \frac{1}{\sqrt{N_f}} \end{aligned} \quad [20]$$

For evaluating the limit of s_2 as $N_m \to \infty$, [19] can be further rearrange as follows:

$$\begin{aligned} s_2 &= \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m + N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m + N_f}{N_f}\right)^{N_f} \\ &= \left(\frac{N_m + N_f}{N_m}\right)^{N_m} * \left(\frac{N_m + N_f}{N_m N_f}\right)^{N_f} \\ &= \left(1 + \frac{N_f}{N_m}\right)^{N_m} * \left(\frac{1}{N_m} + \frac{1}{N_f}\right)^{N_f} \end{aligned} \quad [21]$$

The limit of s_2 could then be evaluated as follows:

$$\begin{aligned} \lim_{N_m \to \infty} s_2 &= \lim_{N_m \to \infty} \left(\left(1 + \frac{N_f}{N_m}\right)^{N_m} * \left(\frac{1}{N_m} + \frac{1}{N_f}\right)^{N_f}\right) \\ &= \lim_{N_m \to \infty} \left(1 + \frac{N_f}{N_m}\right)^{N_m} * \lim_{N_m \to \infty} \left(\frac{1}{N_m} + \frac{1}{N_f}\right)^{N_f} \end{aligned} \quad [22]$$

Using the following definition of Euler's number (Khattri and Witkowski, 2012):

$$e^z = \lim_{x \to \infty} \left(1 + \frac{z}{x}\right)^x \quad [23]$$

The limit in [22] can be further simplified as follows:

$$\lim_{N_m\to\infty}\left(1+\frac{N_f}{N_m}\right)^{N_m} * \lim_{N_m\to\infty}\left(\frac{1}{N_m}+\frac{1}{N_f}\right)^{N_f} = e^{N_f} * \left(\frac{1}{N_f}\right)^{N_f}$$

$$= \left(\frac{e}{N_f}\right)^{N_f} \quad [24]$$

By comparing [24] with the Stirling's approximation from [9], it can be shown that [24] is approximately equals to:

$$\left(\frac{e}{N_f}\right)^{N_f} \cong \frac{\sqrt{2\pi N_f}}{N_f!} \quad [25]$$

By combining [20] and [25] into [16], one can evaluate the limit in [16] as follows:

$$\lim_{N_m\to\infty}\frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f\left(N_m+N_f\right)}}\right) * \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m+N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m+N_f}{N_f}\right)^{N_f}$$

$$= \frac{1}{\sqrt{2\pi}} * \frac{1}{\sqrt{N_f}} * \frac{\sqrt{2\pi N_f}}{N_f!}$$

$$= \frac{1}{N_f!} \quad [26]$$

This implies that as the N_m increases toward infinity, the ratio between b and d reaches a constant value of $\frac{1}{N_f!}$. Given that $N_f!$ is always greater than 1 when $N_f > 1$, the ratio between b and d is always smaller than 1 for all positive integer values of $N_f > 1$ if the value of N_m is large. With the N_f in the denominator of the limit in [26], this also means that as N_f approaches infinity, the value of [26] would approach zero.

Similar to the limit defined in [16], the limit in [17] can also be evaluated by calculating the limit of the sub-functions s_1 and s_2 as N_f approaches infinity. As all the N_f terms are at the denominator of s_1, the limit of s_1 as N_f approaches infinity is as follows:

$$\lim_{N_f\to\infty}\sqrt{\frac{N_m}{N_f\left(N_m+N_f\right)}} = 0 \quad [27]$$

While for the limit of s_2 as N_f approaches infinity, it is defined as follows:

$$\lim_{N_f\to\infty} s_2 = \lim_{N_f\to\infty}\left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{\left(N_m+N_f\right)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m+N_f}{N_f}\right)^{N_f}$$

$$= \left(\frac{1}{N_m^{N_m}}\right) * \lim_{N_f \to \infty} \left(\frac{\left(N_m + N_f\right)^{N_m}}{N_m^{N_f}}\right) * \lim_{N_f \to \infty} \left(1 + \frac{N_m}{N_f}\right)^{N_f} \quad [28]$$

There are two additional limits in [28] that need to be evaluated:

$$s_{21} = \lim_{N_f \to \infty} \left(\frac{\left(N_m + N_f\right)^{N_m}}{N_m^{N_f}}\right) \quad [29]$$

and

$$s_{22} = \lim_{N_f \to \infty} \left(\left(1 + \frac{N_m}{N_f}\right)^{N_f}\right) \quad [30]$$

The limit from [30] can be easily evaluated by again comparing it with the equation from Euler's number, which yielded the value of:

$$s_{22} = \lim_{N_f \to \infty} \left(\left(1 + \frac{N_m}{N_f}\right)^{N_f}\right) = e^{N_m} \quad [31]$$

Evaluating the limit of s_{21} from [29] is significantly more difficult however and requires the use of l'Hôpital's rule. As mentioned, this rule requires both numerator and denominator to approach zero as N_f approaches infinity to be applied (Lawlor, 2020). It is obvious however that as N_f approaches infinity, so do the numerator and denominator, thus the rule could not be directly applied. Despite this, it is possible to transform the functions such that the rule become applicable, which can be achieved through iterated differentiation of the functions.

Let $f(N_f) = \left(N_m + N_f\right)^{N_m}$ and $g\left(N_f\right) = N_m^{N_f}$, with N_m being a positive integer, then

$$\frac{f\left(N_f\right)}{g\left(N_f\right)} = \frac{\left(N_m + N_f\right)^{N_m}}{N_m^{N_f}} \quad [32]$$

Differentiate the functions once, and the resulting ratio was as follows:

$$\frac{f'\left(N_f\right)}{g'\left(N_f\right)} = \frac{N_m\left(N_m + N_f\right)^{N_m - 1}}{N_m^{N_f} * \ln N_m} \quad [33]$$

Which can be differentiated again, yielding the following:

$$\frac{f''(N_f)}{g''(N_f)} = \frac{N_m(N_m-1)(N_m+N_f)^{N_m-2}}{N_m^{N_f} * (\ln N_m)^2} \quad [34]$$

This process could be iterated N_m times, with the last two iterations being as follows:

$$\frac{f^{(N_m-1)}(N_f)}{g^{(N_m-1)}(N_f)} = \frac{N_m(N_m-1)(N_m-2)\dots(3)(2)(N_m+N_f)}{N_m^{N_f} * (\ln N_m)^{N_m-1}} \quad [35]$$

and

$$\frac{f^{(N_m)}(N_f)}{g^{(N_m)}(N_f)} = \frac{N_m(N_m-1)(N_m-2)\dots(3)(2)(1)}{N_m^{N_f} * (\ln N_m)^{N_m}} = \frac{N_m!}{N_m^{N_f} * (\ln N_m)^{N_m}} \quad [36]$$

L'Hôpital's rule can finally be applied onto [36], yielding the following:

$$\lim_{N_f\to\infty} \frac{N_m!}{N_m^{N_f} * (\ln N_m)^{N_m}} = \frac{N_m!}{(\ln N_m)^{N_m}} * \lim_{N_f\to\infty} \frac{1}{N_m^{N_f}} = 0 \quad [37]$$

By substituting [31] and [37] into [28], the limit of s_2 as N_f approaches infinity can be evaluated as such:

$$\lim_{N_f\to\infty} s_2 = \frac{1}{N_m^{N_m}} * e^{N_m} * 0 = 0 \quad [38]$$

By combining [27] and [38], the limit in [17] as N_f approaches infinity can be evaluated as such:

$$\lim_{N_f\to\infty} \frac{1}{\sqrt{2\pi}} * \left(\sqrt{\frac{N_m}{N_f(N_m+N_f)}}\right) * \left(\frac{1}{N_m^{N_m}}\right) * \left(\frac{(N_m+N_f)^{N_m}}{N_m^{N_f}}\right) * \left(\frac{N_m+N_f}{N_f}\right)^{N_f} = \frac{1}{\sqrt{2\pi}} * 0 * 0 = 0 \quad [39]$$

This implies that as N_f increases toward infinity, the ratio between b and d converges toward zero for all values of N_m. Along with the proposition from [26], this suggests that as N_m and N_f grow toward infinity, the limit from [13] converges toward zero, thus showing that d, and henceforth $n_{\{D\}}$, grow faster than b from $n_{\{A\}}$ and $n_{\{I\}}$.

To fur further confirm the increased growth rate of $n_{\{D\}}$ compared to $n_{\{A\}}$ and $n_{\{I\}}$ especially when N_m and N_f are small, equation [14] has also been tested under varying values of N_m and N_f. Besides the trivial value of $N_m = 1$, where [14] returned the value of N_f, the maximal value attained by this equation is when $(N_m, N_f) = (2,2)$ and $(N_m, N_f) = (2,3)$, where equation [14] reaches 0.75 (Figure E.1). Given that at this maximal value the ratio between b and d is still less than 1, it means for any given value of $N_m > 1$, the d is larger than b, even for small N_m and N_f. With the ratio quickly approaching 0 for any larger value of N_m and N_f, this proves that with larger number of sires and dams, the growth rate of $n_{\{D\}}$ is greater than $n_{\{A\}}$ and $n_{\{I\}}$.

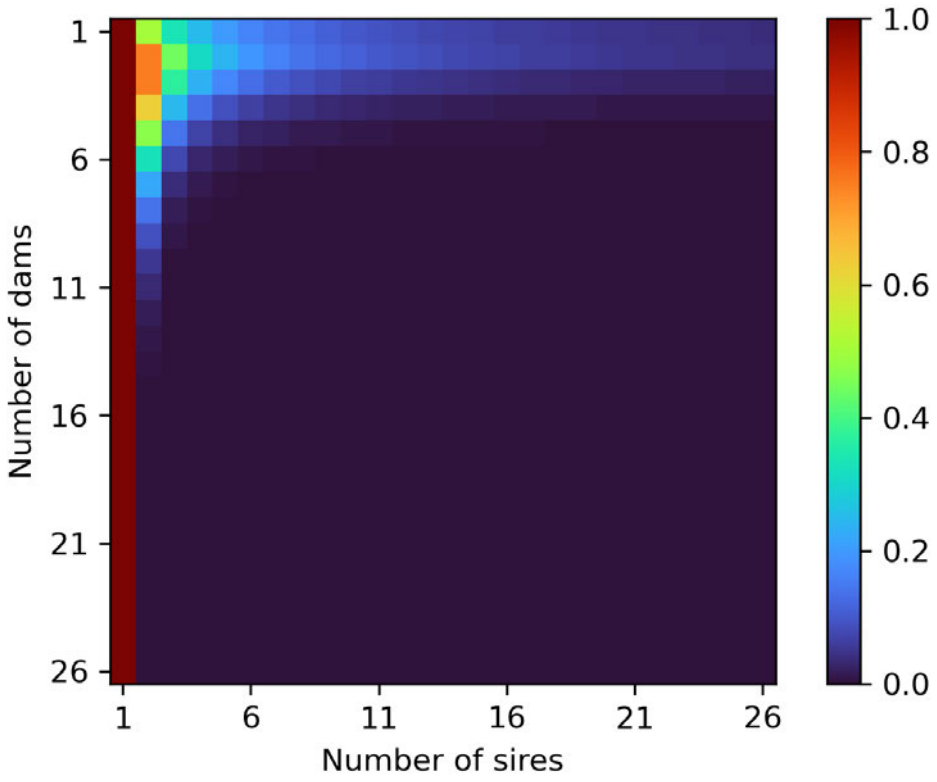

Figure E.1: The ratio between $n_{\{A\}}$ and $n_{\{D\}}$, defined by equation [14], evaluated across a number of sires and dams.

# Appendix F. The Discrepancies of ΔI Between NRM and GRM by VanRaden (2008) for Equally Contributed Sires

One of the most popular methods of calculating the Genomic Relationship Matrix (GRM) is one proposed by VanRaden (2008). This GRM has been utilized in optimal contribution selection (OCS) where it is used to calculate the increment in level of inbreeding (denoted as ΔI) in the offspring (Clark et al., 2013). Compared to a pedigree-based Numerator Relationships Matrix (NRM), there are some discrepancies in the ΔI calculated if the GRM is used. In particular, with the assumption of all sires contributed equally to the dams, with no inbreeding or consanguinity between the sires, the expected ΔI for NRM is $\frac{1}{2N_m}$ where the N_m being the number of sires selected, whereas for GRM, the expected ΔI is zero. This appendix section is to provide a mathematical explanation of such discrepancy, and the rationale for the adjustment of the GRM.

F.1. A Simplified Example

Consider a case with 5 unrelated, non-inbred sires being considered in a breeding program, which was denoted using a sire contribution vector $\boldsymbol{c}$. Assuming all the sires contributed equally to the breeding program (i.e. $N_m = 5$), the sire contribution vector was as follows:

$$\boldsymbol{c} = [0.2 \quad 0.2 \quad 0.2 \quad 0.2 \quad 0.2] \qquad [1]$$

If the sires are unrelated and non-inbred, the expected NRM between the sires (denoted as $\boldsymbol{G_{NRM}}$) is an identity matrix (Henderson, 1976):

$$\boldsymbol{G_{NRM}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \qquad [2]$$

Using the NRM and sire contribution vector, the ΔI can then be calculated as such (Clark et al., 2013):

$$\Delta I = \frac{1}{2}\boldsymbol{c}\boldsymbol{G}_{\boldsymbol{NRM}}\boldsymbol{c}^{T} \quad [3]$$

Which can be calculated as follows:

$$\Delta I = \frac{1}{2}\boldsymbol{c}\boldsymbol{G}_{\boldsymbol{NRM}}\boldsymbol{c}^{T} = \frac{1}{2} * [0.2 \quad 0.2 \quad 0.2 \quad 0.2 \quad 0.2] * \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} * \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix} = 0.1 \quad [4]$$

The value of $\Delta I = 0.1$ also corresponds to the previously reported expected ΔI in the offspring; if there is no inbreeding or relationship between sires, the expected increase in inbreeding level is $\frac{1}{2N_m}$ (Clark et al., 2013; Falconer, 1989; Meuwissen, 1997).

This is not the case for the GRM however; as one of the steps by VanRaden (2008) involves the removal of effects of allele frequency, this causes the GRM to have column-wise expected values of zero. Due to this, if the GRM is used in place of the NRM in equation [4], the resulting ΔI would become zero. The mathematical explanation for this observation is provided in Section F.2.

F.2. The Mathematical Explanation

Let $\boldsymbol{X}$ be a matrix of size $n \times k$ that represents a genotype array with n number of animals and k number of markers that have been coded in the form of $\{-1, 0, 1\}$ for genotype hh, Hh and HH. VanRaden (2008) call for the adjustment of each column of matrix $\boldsymbol{X}$ with their column-wise means (denoted using the vector $\bar{\boldsymbol{x}}$), which is defined as follows:

$$\bar{\boldsymbol{x}} = 2\boldsymbol{p} - 1 \quad [5]$$

Where $\boldsymbol{p}$ is the vector of allele frequencies. The adjustment is done by subtracting each column of $\boldsymbol{X}$ with the corresponding values of $\bar{\boldsymbol{x}}$ (denoted as $\bar{x}_j$) The adjusted matrix, denoted as $\boldsymbol{Z}$, is defined as follows:

$$\boldsymbol{Z} = \boldsymbol{X} - \bar{\boldsymbol{x}} \quad [6]$$

Where the ith row and jth column of matrix $\boldsymbol{Z}$ is calculated as follows:

$$\boldsymbol{Z}_{i,j} = \boldsymbol{X}_{i,j} - \bar{x}_j \quad [7]$$

The purpose for this adjustment is to remove the effects of allele frequency from each of the loci in $\boldsymbol{X}$ (VanRaden, 2008). This adjustment has also the effects of setting the column-wise means of $\boldsymbol{Z}$ into zero:

$$\sum_{i=1}^{n} \boldsymbol{Z}_{i,j} = 0 \qquad [8]$$

This observation is the key to the proof of the zero ΔI with the use of GRM.

As defined by VanRaden (2008), the GRM (denoted as $\boldsymbol{G_{GRM}}$) is defined as follows:

$$\boldsymbol{G_{GRM}} = \frac{\boldsymbol{Z}\boldsymbol{Z}^T}{2\sum_{j=1}^{k} p_j(1-p_j)} \qquad [9]$$

As the denominator of equation [9] (i.e. $2\sum_{j=1}^{k} p_j(1-p_j)$) is a nonnegative finite scalar factor, it is not relevant to the proof and thus can be ignored. In this case, the unscaled GRM, denoted as $\boldsymbol{M}$ be defined as such:

$$\boldsymbol{M} = \boldsymbol{Z}\boldsymbol{Z}^T = 2\sum_{j=1}^{k} p_j(1-p_j) * \boldsymbol{G_{GRM}} \qquad [10]$$

Without loss of generality, let a be the index of a column from matrix $\boldsymbol{M}$. This column of $\boldsymbol{M}$ (denoted as $\boldsymbol{M}_{*,a}$) is the product of $\boldsymbol{Z}$ and the ath column of $\boldsymbol{Z}^T$(denoted as $\boldsymbol{Z}^T_{*,a}$):

$$\boldsymbol{M}_{*,a} = \boldsymbol{Z}\boldsymbol{Z}^T_{*,a} \qquad [11]$$

This column vector $\boldsymbol{M}_{*,a}$ can be defined in term of $\boldsymbol{Z}$ as follows:

$$\boldsymbol{M}_{*,a} = \begin{bmatrix} \boldsymbol{Z}_{1,1}\boldsymbol{Z}^T_{1,a} + \boldsymbol{Z}_{1,2}\boldsymbol{Z}^T_{2,a} + \boldsymbol{Z}_{1,3}\boldsymbol{Z}^T_{3,a} + \cdots + \boldsymbol{Z}_{1,k}\boldsymbol{Z}^T_{k,a} \\ \boldsymbol{Z}_{2,1}\boldsymbol{Z}^T_{1,a} + \boldsymbol{Z}_{2,2}\boldsymbol{Z}^T_{2,a} + \boldsymbol{Z}_{2,3}\boldsymbol{Z}^T_{3,a} + \cdots + \boldsymbol{Z}_{2,k}\boldsymbol{Z}^T_{k,a} \\ \boldsymbol{Z}_{3,1}\boldsymbol{Z}^T_{1,a} + \boldsymbol{Z}_{3,2}\boldsymbol{Z}^T_{2,a} + \boldsymbol{Z}_{3,3}\boldsymbol{Z}^T_{3,a} + \cdots + \boldsymbol{Z}_{3,k}\boldsymbol{Z}^T_{k,a} \\ \vdots \\ \boldsymbol{Z}_{n,1}\boldsymbol{Z}^T_{n,a} + \boldsymbol{Z}_{n,2}\boldsymbol{Z}^T_{2,a} + \boldsymbol{Z}_{n,3}\boldsymbol{Z}^T_{3,a} + \cdots + \boldsymbol{Z}_{n,k}\boldsymbol{Z}^T_{k,a} \end{bmatrix} \qquad [12]$$

The zero column mean of $\boldsymbol{M}$ can be shown by multiplying a row vector of ones $\mathbf{1}_{1,n}$ and the unscaled GRM $\boldsymbol{M}$. It is from the observation that the multiplication of a row vector of ones onto a matrix has the same effect of taking the unweighted column-wise sums of the matrix. Therefore, the multiplication of row vector $\mathbf{1}_{1,n}$ into $\boldsymbol{M}$ has the following effect on $\boldsymbol{M}_{*,a}$:

$$\mathbf{1}_{1,n}\boldsymbol{M}_{*,a} = [1 \quad 1 \quad 1 \quad \cdots \quad 1] * \begin{bmatrix} \boldsymbol{Z}_{1,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{1,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{1,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{1,k}\boldsymbol{Z}_{k,a}^T \\ \boldsymbol{Z}_{2,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{2,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{2,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{2,k}\boldsymbol{Z}_{k,a}^T \\ \boldsymbol{Z}_{3,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{3,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{3,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{3,k}\boldsymbol{Z}_{k,a}^T \\ \vdots \\ \boldsymbol{Z}_{n,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{n,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{n,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{n,k}\boldsymbol{Z}_{k,a}^T \end{bmatrix} \quad [13]$$

Which can be expressed in term of the following summations:

$$\begin{aligned} \mathbf{1}_{1,n}\boldsymbol{M}_{*,a} = & \left[\boldsymbol{Z}_{1,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{1,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{1,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{1,k}\boldsymbol{Z}_{k,a}^T\right] + \\ & \left[\boldsymbol{Z}_{2,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{2,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{2,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{2,k}\boldsymbol{Z}_{k,a}^T\right] + \\ & \left[\boldsymbol{Z}_{3,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{3,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{3,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{3,k}\boldsymbol{Z}_{k,a}^T\right] + \\ & \cdots + \\ & \left[\boldsymbol{Z}_{n,1}\boldsymbol{Z}_{1,a}^T + \boldsymbol{Z}_{n,2}\boldsymbol{Z}_{2,a}^T + \boldsymbol{Z}_{n,3}\boldsymbol{Z}_{3,a}^T + \cdots + \boldsymbol{Z}_{n,k}\boldsymbol{Z}_{k,a}^T\right] \end{aligned} \quad [14]$$

By rearranging and factorizing the common terms in $\mathbf{1}_{1,n}\boldsymbol{M}_{*,a}$ (i.e. $\boldsymbol{Z}^T$s), the summation could be simplified as follows:

$$\begin{aligned} \mathbf{1}_{1,n}\boldsymbol{M}_{*,a} = & \boldsymbol{Z}_{1,a}^T\left(\boldsymbol{Z}_{1,1} + \boldsymbol{Z}_{2,1} + \boldsymbol{Z}_{3,1} + \cdots + \boldsymbol{Z}_{n,1}\right) + \\ & \boldsymbol{Z}_{2,a}^T\left(\boldsymbol{Z}_{1,2} + \boldsymbol{Z}_{2,2} + \boldsymbol{Z}_{3,2} + \cdots + \boldsymbol{Z}_{n,2}\right) + \\ & \boldsymbol{Z}_{3,a}^T\left(\boldsymbol{Z}_{1,3} + \boldsymbol{Z}_{2,3} + \boldsymbol{Z}_{3,3} + \cdots + \boldsymbol{Z}_{n,3}\right) + \\ & \cdots + \\ & \boldsymbol{Z}_{k,a}^T\left(\boldsymbol{Z}_{1,k} + \boldsymbol{Z}_{2,k} + \boldsymbol{Z}_{3,k} + \cdots + \boldsymbol{Z}_{n,k}\right) \end{aligned} \quad [15]$$

and by collecting the indices within the parentheses, as follows:

$$\mathbf{1}_{1,n}\boldsymbol{M}_{*,a} = \boldsymbol{Z}_{1,a}^T\sum_{i=1}^{n}\boldsymbol{Z}_{i,1} + \boldsymbol{Z}_{2,a}^T\sum_{i=1}^{n}\boldsymbol{Z}_{i,2} + \boldsymbol{Z}_{3,a}^T\sum_{i=1}^{n}\boldsymbol{Z}_{i,3} + \cdots + \boldsymbol{Z}_{k,a}^T\sum_{i=1}^{n}\boldsymbol{Z}_{i,k} \quad [16]$$

Note the equivalence in the summations $\sum_{i=1}^{n}\boldsymbol{Z}_{i,j}\ ; j \in \{1, \ldots, k\}$ in [16] with that in [8], which suggests that each of these summations would result in zeroes. Thus, by substituting [8] into [16], the $\mathbf{1}_{1,n}\boldsymbol{M}_{*,a}$ could be evaluated as such:

$$\begin{aligned} \mathbf{1}_{1,n}\boldsymbol{M}_{*,a} &= \boldsymbol{Z}_{1,a}^T(0) + \boldsymbol{Z}_{2,a}^T(0) + \boldsymbol{Z}_{3,a}^T(0) + \cdots + \boldsymbol{Z}_{k,a}^T(0) \\ &= 0 \end{aligned} \quad [17]$$

The $\mathbf{1}_{1,n}\boldsymbol{M}_{*,a} = 0$ suggests the column-wise sum of $\boldsymbol{M}_{*,a}$ sums to zero and, given the generality of index a, also mean any columns within $\boldsymbol{M}$ sums to zero and, by extension, the GRM as proposed by VanRaden (2008) sums to zero as well.

The zero column mean of $\boldsymbol{G_{GRM}}$ is the root cause for ΔI to become zero when it is used in place of $\boldsymbol{G_{NRM}}$ in equation [4]. This can be further proven as follows: let $\boldsymbol{c}$ be a sire contribution row vector of length N_m. If all the sires contributed equally to the breeding program, the sire contribution vector was defined as follows:

$$\boldsymbol{c} = \begin{bmatrix} \frac{1}{N_m} & \frac{1}{N_m} & \frac{1}{N_m} & \cdots & \frac{1}{N_m} \end{bmatrix} \quad [18]$$

Which can be rewritten as such:

$$\begin{aligned} \boldsymbol{c} &= \frac{1}{N_m} * [1 \quad 1 \quad 1 \quad \cdots \quad 1] \\ &= \frac{1}{N_m} * \mathbf{1}_{1,n} \end{aligned} \quad [19]$$

By substituting equation [19] and [9] into [3], the following expression for ΔI is obtained:

$$\Delta I = \frac{1}{2} * \left(\frac{1}{N_m} * \mathbf{1}_{1,n}\right) * \left(\frac{\boldsymbol{Z}\boldsymbol{Z}^T}{2\sum_{j=1}^{k} p_j(1-p_j)}\right) * \left(\frac{1}{N_m} * \mathbf{1}_{1,n}\right)^T \quad [20]$$

Which simplified into:

$$\Delta I = \frac{1}{4N_m \sum_{j=1}^{k} p_j(1-p_j)} * \mathbf{1}_{1,n}\boldsymbol{Z}\boldsymbol{Z}^T\mathbf{1}_{1,n}^T \quad [21]$$

As the denominator for equation [21] (i.e. $4N_m \sum_{j=1}^{k} p_j(1-p_j)$) is a nonnegative finite scalar factor, it is not relevant to the proof for a zero ΔI, thus can be ignored. In this case, the unscaled version of ΔI (denoted as ΔI_s) was as follows:

$$\Delta I_s = \mathbf{1}_{1,n}\boldsymbol{Z}\boldsymbol{Z}^T\mathbf{1}_{1,n}^T \quad [22]$$

Or in term of $\boldsymbol{M}$:

$$\Delta I_s = \mathbf{1}_{1,n}\boldsymbol{M}\mathbf{1}_{1,n}^T \quad [23]$$

As the column-wise sum of $\boldsymbol{M}$ had been shown to equate to zero, the product $\mathbf{1}_{1,n}\boldsymbol{M}$ in equation [23] yields a zero vector and, by extension, the full product $\mathbf{1}_{1,n}\boldsymbol{M}\mathbf{1}_{1,n}^T$. Therefore, the ΔI_s and, by extension, the ΔI from equation [20] would become zero, showing that the use of GRM as suggested by VanRaden (2008) in equally contributed sires would yield $\Delta I = 0$.

F.3. Adjusting the GRM for the OCS

While t the effects of discrepancies of ΔI between GRM and NRM could be easily ignored for conventional use of the OCS, the zero column-wise means of GRM has an undesirable consequence; as the column-wise means of the GRM is zero, the increase in inbreeding level become no longer relevant to the N_m if all the sires contributed equally to all the dams. This means if 5 sires are chosen, the calculated expected increase in inbreeding level would not differ from another scenario with, for example, 5000 sires (i.e., expected $\Delta I = 0$). This contradicts the theoretical results from previous publications. For this reason, adjustments on the calculation of GRM are proposed to make the expected ΔI more analogous with those expected from NRM. This could be done by comparing the numerical properties of GRM and NRM.

This calculation can be generalized to any columns of the GRM and NRM, but for convenience, let $\boldsymbol{G}_{\boldsymbol{NRM}_{*,1}}$ and $\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}}$ be the first column of the NRM $\boldsymbol{G}_{\boldsymbol{NRM}}$ and GRM $\boldsymbol{G}_{\boldsymbol{GRM}}$, respectively. If the sires are unrelated and non-inbred, the expected values of $\boldsymbol{G}_{\boldsymbol{NRM}_{*,1}}$ were as follows (Henderson, 1976):

$$\boldsymbol{G}_{\boldsymbol{NRM}_{*,1}} = [1 \quad 0 \quad 0 \quad \cdots \quad 0] \qquad [24]$$

Publication of the nature of $\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}}$ is less common but can be calculated as follows. Let $\boldsymbol{z} \in \{z_1, z_2, z_3 \ldots z_{N_m}\}$ be a set of covariance that built the $\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}}$, defined as follows:

$$\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}} = [z_1 \quad z_2 \quad z_3 \quad \cdots \quad z_{N_m}] \qquad [25]$$

Given the assumption of unrelated sires, from the perspective of the first sire (i.e., the first column and row of the GRM), it would share a similar level of relationship with all other individuals within the matrix. This implies that with the exception of z_1, all other covariance in $\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}}$ have the same value (i.e. $z_2 = z_3 = \cdots = z_{N_m}$). Thus equation [25] can be simplified into the following:

$$\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}} = [z_1 \quad z_2 \quad z_2 \quad \cdots \quad z_2] \qquad [26]$$

As the column-wise means of GRM has been shown to be zero (i.e. $\sum \boldsymbol{G}_{\boldsymbol{GRM}_{*,1}} = 0$), this also imply the following relationship between the zs:

$$z_1 + (N_m - 1) * z_2 = 0$$

$$z_2 = -\frac{z_1}{N_m - 1} \quad [27]$$

Substituting equation [27] into [26] yields the following:

$$\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}} = \begin{bmatrix} z_1 & -\frac{z_1}{N_m - 1} & -\frac{z_1}{N_m - 1} & \cdots & -\frac{z_1}{N_m - 1} \end{bmatrix} \quad [28]$$

From equation [28], the remaining value that needs to be determined was z_1, which is the covariance value of the first individual with itself. Additional simulations from 100,000 rounds of GRMs suggest that expected values for the z_1 can be approximated as follows:

$$z_1 \approx \frac{2N_m - 2}{2N_m - 1} \quad [29]$$

which can be substituted into [28], yielding the following:

$$\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}} = \begin{bmatrix} \frac{2N_m - 2}{2N_m - 1} & \frac{2 - 2N_m}{(2N_m - 1)(N_m - 1)} & \frac{2 - 2N_m}{(2N_m - 1)(N_m - 1)} & \cdots & \frac{2 - 2N_m}{(2N_m - 1)(N_m - 1)} \end{bmatrix} \quad [30]$$

The $\boldsymbol{G}_{\boldsymbol{GRM}_{*,1}}$ can then be compared with $\boldsymbol{G}_{\boldsymbol{NRM}_{*,1}}$, with their differences (denoted as $\boldsymbol{G}_{\boldsymbol{NRM-GRM}_{*,1}}$) be calculated as follows:

$$\begin{aligned} \boldsymbol{G}_{\boldsymbol{NRM-GRM}_{*,1}} &= \boldsymbol{G}_{\boldsymbol{NRM}_{*,1}} - \boldsymbol{G}_{\boldsymbol{GRM}_{*,1}} \\ &= \begin{bmatrix} 1 - \frac{2N_m - 2}{2N_m - 1} & \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} & \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} & \cdots & \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{2N_m - 1} & \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} & \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} & \cdots & \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} \end{bmatrix} \end{aligned} \quad [31]$$

which, provided that N_m and k are sufficiently large ($N_m \geq 50$ and $k \geq 10k$), can also be approximated using the following simplified expression:

$$\boldsymbol{G}_{\boldsymbol{NRM-GRM}_{*,1}} \approx \begin{bmatrix} 0 & \frac{1}{N_m - 1} & \frac{1}{N_m - 1} & \cdots & \frac{1}{N_m - 1} \end{bmatrix} \quad [31]$$

With an expected bias of $-\frac{1}{2N_m - 1}$ for the diagonal component, and $\frac{1}{(N_m - 1)(2N_m - 1)}$ for the off-diagonal component.

The $\boldsymbol{G}_{\boldsymbol{NRM-GRM}_{*,1}}$ in equation [31] suggests a constant shift of the off-diagonal component of the GRM by $\frac{1}{N_m - 1}$. If expressed in terms of the full GRM, the $\boldsymbol{G}_{\boldsymbol{NRM-GRM}_{*,1}}$ becomes the first column of the following adjustment matrix (denoted as $\boldsymbol{G}_{\boldsymbol{NRM-GRM}}$):

$$\boldsymbol{G_{NRM-GRM}} = \begin{bmatrix} 0 & \frac{1}{N_m-1} & \frac{1}{N_m-1} & \cdots & \frac{1}{N_m-1} \\ \frac{1}{N_m-1} & 0 & \frac{1}{N_m-1} & \cdots & \frac{1}{N_m-1} \\ \frac{1}{N_m-1} & \frac{1}{N_m-1} & 0 & \cdots & \frac{1}{N_m-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{N_m-1} & \frac{1}{N_m-1} & \frac{1}{N_m-1} & \cdots & 0 \end{bmatrix}$$

$$= \frac{1}{N_m-1} * \left(\mathbf{1}_{\boldsymbol{N_m}} - \boldsymbol{I}_{\boldsymbol{N_m}}\right) \quad [32]$$

The shifted GRM, denoted as $\boldsymbol{G^*_{GRM}}$, could thus be expressed as follows:

$$\boldsymbol{G^*_{GRM}} = \boldsymbol{G_{GRM}} + \boldsymbol{G_{NRM-GRM}}$$

$$= \boldsymbol{G_{GRM}} + \frac{1}{N_m-1} * \left(\mathbf{1}_{\boldsymbol{N_m}} - \boldsymbol{I}_{\boldsymbol{N_m}}\right) \quad [33]$$

Which can be used to take into account the effects of a changing number of equally contributing sires on ΔI if GRM has been used.

A less wieldy but more precise adjustment (suitable for $N_m < 50$) can also be defined as follows:

$\boldsymbol{G_{NRM-GRM}}$

$$= \begin{bmatrix} \frac{1}{2N_m-1} & \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \cdots & \frac{2N_m-2}{(2N_m-1)(N_m-1)} \\ \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \frac{1}{2N_m-1} & \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \cdots & \frac{2N_m-2}{(2N_m-1)(N_m-1)} \\ \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \frac{1}{2N_m-1} & \cdots & \frac{2N_m-2}{(2N_m-1)(N_m-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \frac{2N_m-2}{(2N_m-1)(N_m-1)} & \cdots & \frac{1}{2N_m-1} \end{bmatrix} \quad [34]$$

$$= \frac{1}{2N_m-1} * \boldsymbol{I}_{\boldsymbol{N_m}} + \frac{2N_m-2}{(2N_m-1)(N_m-1)} * \left(\mathbf{1}_{\boldsymbol{N_m}} - \boldsymbol{I}_{\boldsymbol{N_m}}\right)$$

$$= \frac{2N_m-2}{(2N_m-1)(N_m-1)} * \mathbf{1}_{\boldsymbol{N_m}} - \frac{1}{2N_m-1} * \boldsymbol{I}_{\boldsymbol{N_m}}$$

And the corresponding $\boldsymbol{G^*_{GRM}}$ was as follows:

$$\boldsymbol{G^*_{GRM}} = \boldsymbol{G_{GRM}} + \frac{2N_m-2}{(2N_m-1)(N_m-1)} * \mathbf{1}_{\boldsymbol{N_m}} - \frac{1}{2N_m-1} * \boldsymbol{I}_{\boldsymbol{N_m}} \quad [35]$$

Using the aforementioned simplified example with $N_m = 5$, the unadjusted $\boldsymbol{G_{GRM}}$ would have diagonal component with expected value of $\frac{2N_m-2}{2N_m-1} = \frac{8}{9}$ and off-diagonal component with expected value of $\frac{2-2N_m}{(2N_m-1)(N_m-1)} = -\frac{2}{9}$. The corresponding $\boldsymbol{G^*_{GRM}}$ would be as follows:

$$\begin{aligned}
\boldsymbol{G^*_{GRM}} &= \boldsymbol{G_{GRM}} + \frac{2N_m - 2}{(2N_m - 1)(N_m - 1)} * \mathbf{1}_{N_m} - \frac{1}{2N_m - 1} * \boldsymbol{I}_{N_m} \\
&= \boldsymbol{G_{GRM}} + \left(\frac{8}{9 * 4}\right) * \mathbf{1}_{N_m} - \left(\frac{1}{9}\right) * \boldsymbol{I}_{N_m} \\
&\approx \begin{bmatrix} \frac{8}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} \\ -\frac{2}{9} & \frac{8}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} \\ -\frac{2}{9} & -\frac{2}{9} & \frac{8}{9} & -\frac{2}{9} & -\frac{2}{9} \\ -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & \frac{8}{9} & -\frac{2}{9} \\ -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & \frac{8}{9} \end{bmatrix} + \begin{bmatrix} \frac{1}{9} & \frac{2}{9} & \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{1}{9} & \frac{2}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{1}{9} & \frac{2}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} & \frac{1}{9} & \frac{2}{9} \\ \frac{2}{9} & \frac{2}{9} & \frac{2}{9} & \frac{2}{9} & \frac{1}{9} \end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}
\end{aligned} \qquad [36]$$

Equation [36] suggests that the $\boldsymbol{G^*_{GRM}}$ is an identity matrix, the same matrix that defines an NRM for non-inbred and unrelated sires (Henderson, 1976). Thus, the $\boldsymbol{G^*_{GRM}}$ can be substituted in place of $\boldsymbol{G_{NRM}}$ equation [4] and produces the same results.

To illustrate the bias of the approximated $\boldsymbol{G^*_{GRM}}$ from equation [33] with a small number of sires, it would be defined as follows:

$$\begin{aligned}
\boldsymbol{G^*_{GRM}} &= \boldsymbol{G_{GRM}} + \frac{1}{N_m - 1} * \left(\mathbf{1}_{N_m} - \boldsymbol{I}_{N_m}\right) \\
&= \boldsymbol{G_{GRM}} + \frac{1}{4}\left(\mathbf{1}_{N_m} - \boldsymbol{I}_{N_m}\right) \\
&\approx \begin{bmatrix} \frac{8}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} \\ -\frac{2}{9} & \frac{8}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} \\ -\frac{2}{9} & -\frac{2}{9} & \frac{8}{9} & -\frac{2}{9} & -\frac{2}{9} \\ -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & \frac{8}{9} & -\frac{2}{9} \\ -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & -\frac{2}{9} & \frac{8}{9} \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{bmatrix}
\end{aligned}$$

$$= \begin{bmatrix} \frac{8}{9} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} \\ \frac{1}{36} & \frac{8}{9} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{36} & \frac{8}{9} & \frac{1}{36} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{8}{9} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{8}{9} \end{bmatrix} \quad [37]$$

With a bias of $-\frac{1}{9}$ for the diagonal and $\frac{1}{36}$ for the off-diagonals. Compared to $\boldsymbol{G}^*_{\boldsymbol{NRM}}$ of equation [36] however, these biases would not affect the outcome of ΔI; by substituting this $\boldsymbol{G}^*_{\boldsymbol{NRM}}$ into equation [4], the same result is obtained:

$$\Delta I = \frac{1}{2}\boldsymbol{c}\boldsymbol{G}^*_{\boldsymbol{NRM}}\boldsymbol{c}^T = \frac{1}{2} * [0.2 \quad 0.2 \quad 0.2 \quad 0.2 \quad 0.2] * \begin{bmatrix} \frac{8}{9} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} \\ \frac{1}{36} & \frac{8}{9} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{36} & \frac{8}{9} & \frac{1}{36} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{8}{9} & \frac{1}{36} \\ \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{1}{36} & \frac{8}{9} \end{bmatrix} * \begin{bmatrix} 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \\ 0.2 \end{bmatrix} = 0.1 \quad [38]$$

Thus, this approximated $\boldsymbol{G}^*_{\boldsymbol{NRM}}$ from equation [33] can still be used in the OCS to handle situation with equally contributed sires, while taking into account the effects of number of sires on the ΔI when GRM is utilized. Furthermore, given the fact that the biases would become less prominent with larger N_m, this suggested the suitability of using the approximated $\boldsymbol{G}^*_{\boldsymbol{NRM}}$ in the OCS.

Appendix G. Pseudocodes for the Phases of Genetic Algorithm in Chapter 6

G.1. Phase 1 Pseudocode for Chapter 6

The pseudocode for the Phase 1 of the cascading genetic algorithm, which aimed to optimize the additive and inbreeding component for the OCS, is as follows:

```
## INPUT: bhat, GRM, Z_fem (b for TBV control)
### bhat: sire EBVs # shape = (nmal, 1)
### GRM: sire's GRM # shape = (nmal, nmal)
### nmal: number of sires
### Z_fem: dam genotype array (included here just to obtain the number of dam)
### bhat and GRM are calculated using method proposed by VanRaden (2008)
### b: additive TBV for sires # shape = (nmal, 1)

nmal = nrow(bhat)
nfem = nrow(Z_fem)

phase1_Ns = 1500 # number of solutions for GA
phase1_Ntop = 2 # number of top solution chosen to propagate into the next iteration
phase1_Npar = 8 # number of parallelized threads

## inbreeding component targets
i_t = 0.01 # user-specified targeted level of consanguinity (Clark et al., 2013)
z_lambda = 100 # amount of lambda_i update per unit of departure from i_t

##### HYPERPARAMETERS
### GENETIC OPERATOR HYPERPARAMETERS FOR PHASE 1 (Table 6.1)
## mutation hyperparameters
p_m = 0.05 # mutation rate

## horizontal recombination hyperparameters
p_hcr = 0.3 # horizontal recombination rate
p_hcb = 0.1 # horizontal recombination block size

## vertical recombination hyperparameters
p_vcr = 0.3 # vertical recombination rate
p_vcb = 0.1 # vertical recombination block size

## horizontal inversion hyperparameters
p_hir = 0.3 # horizontal inversion rate
p_hib = 0.1 # horizontal inversion block size

## vertical inversion hyperparameters
p_vir = 0.3 # vertical inversion rate
p_vib = 0.1 # vertical inversion block size

## convergence hyperparameters
convergence_slope = 1e-3 ## slope of fOCS for which convergence is reached
convergence_lastiter = 50 ## convergence evaluated over the last ... iterations
maximum_iteration = 3000 ## maximum number of iterations

GRMstar = GRM + (matrix(1, shape=(nmal, nmal)) - I(nmal)) / (nmal - 1)
## I(x): a function that produces identity matrix of size x*x

## PHASE 1 GENETIC ALGORITHM: optimizing additive + inbreeding components
S_allP1 = matrix(NA, shape = (phase1_Npar, phase1_Ntop, nfem))
fOCS_AI_avg = numeric(length = maximum_iteration)
lamba_i_fin = numeric(length = phase1_Npar)
```

```
for npx in range(phase1_Npar): # parallelizing the GA

        # generating seed sire index array, an array containing indices of sires to be mated
        S = random(nmal, size=(phase1_Ns, nfem), replace=T)
        lambda_i = 0  # seed lambda i

        for itx in range(maximum_iteration):
                # converting sire index array S to sire proportion array X (using s_to_x)

                X = matrix(NA, shape=(phase1_Ns, nmal))
                for zs in range(nmal):
                        X[zs,:] = s_to_x(S[zs,:], nmal)

                a_s = X @ bhat # eqn [15]; @: matrix multiplication
                ## for TBV control: b is used in place of bhat
                i_s = diag(X @ GRMstar @ X.T) # eqn [16]; X.T: transpose of X

                fOCS_AI = a_s + lambda_i*i_s # eqn [17]
                fOCS_AI_avg[itx] = mean(fOCS_AI)

                ## scaling factor for genetic operators(eqn. [6], based on Srinivas and Patnaik (1994))
                f_min, f_mean, f_max = min(fOCS_AI), mean(fOCS_AI), max(fOCS_AI)
                scaling_factor = (f_max - f_mean) / (f_max - f_min)

                # updating lambda_i
                i_avg = mean(i_s)
                lambda_i = lambda_i + z_lambda*(i_avg - i_t)

                top_sln_idx = argsort(fOCS_AI)[-phase1_Ntop:] # extract top phase1_Ntop solutions
                S_top = S[top_sln_idx,:]

                ## testing convergence using slope of fOCS_AI
                if itx >= convergence_lastiter:
                        if slope(fOCS_AI_avg[-convergence_lastiter:]) <= convergence_slope:
                                break

                ## applying genetic operators, with hyperparameters scaled by scaling_factor
                S_off = S_top[random(phase1_Ntop, size=(phase1_Ns - phase1_Ntop)),:]
                # mutation
                S_off = mutation(S_off, p_m, scaling_factor)
                # horizontal recombination
                S_off = horizontal_recombination(S_off, p_hcr, p_hcb, scaling_factor)
                # vertical recombination
                S_off = vertical_recombination(S_off, p_vcr, p_vcb, scaling_factor)
                # horizontal inversion
                S_off = horizontal_inversion(S_off, p_hir, p_hib, scaling_factor)
                # vertical inversion
                S_off = vertical_inversion(S_off, p_vir, p_vib, scaling_factor)

                S = vstack((S_top, S_off)) # combining S_top and S_off -> new S

        lamba_i_fin[npx] = lambda_i
        S_allP1[npx,:,:] = S_top

S_allP1 = reshape(S_allP1, shape=(-1, nfem))
## reshape S_allP1 into a new shape: ((phase1_Ntop*phase1_Npar) * nfem)
```

G.2. Phase 2 Pseudocode for Chapter 6

The pseudocode for the Phase 2 of the cascading genetic algorithm, which aimed to optimize the dominance component for the offspring, is as follows:

```
##### PHASE 2 GENETIC ALGORITHM
###INPUT: S_allP1 , H_scorearray (D_scorearray for (true) dominance control)
## S_allP1: Solutions optimized in Phase 1
## H_scorearray: Heterozygosity score array
## D_scorearray: (True) Dominance Score Array
#### D_scorearray can also be calculated using estimated dominance effect sizes
#### (if such estimates are available)

## GENETIC OPERATORS HYPERPARAMETERS
# Vertical Recombination
p_vcr = 0.4 # vertical recombination rate
p_vcb = 0.1 # vertical recombination blocksize

# Horizontal Inversion
p_hir = 0.4 # horizontal inversion rate
p_hib = 0.1 # horizontal inversion blocksize

# Solution Hyperparameters
Phase2_Ns = 3000
Phase2_Ntop = 2
Phase2_Npar = 8

# Convergence Hyperparameters
convergence_slope = 1e-5
convergence_lastiter = 200
maximum_iteration = 30000

nmal, nfem = H_scorearray.shape ## D_scorearray if dominance effect sizes are used
nsln_from_p1 = nrow(S_allP1)

# GA for optimizing dominance components
S_allP2 = matrix(NA, shape=(Phase2_Npar, nfem))
for npx in range(Phase2_Npar): # parallelization of GA
        S_off = S_allP1[random(nsln_from_p1, size=(Phase2_Ns - nsln_from_p1)),:]
        ## genetic operators to remove duplicated solutions from the sampling process
        S_off = vertical_recombination(S_off, p_vcr, p_vcb) # vertical recombination
        S_off = horizontal_inversion(S_off, p_hir, p_hib) # horizontal inversion

        S = vstack((S_allP1, S_off)) # shape = (Phase2_Ns, nfem)
        fOCS_D = numeric(maximum_iteration)

        for itx in range(maximum_iteration):
                # calculating the dominance scores for each solutions fOCS_D (d_s_bold)
                d_s_bold = numeric(length = Phase2_Ns)
                for i in range(Phase2_Ns):
                        d_s = numeric(length = nfem)
                        for zf in range(nfem):
                                sire_from_S = S[i,zf]
                                d_s[zf] = H_scorearray[sire_from_S,zf]
                                ## D_scorearray if dominance effect sizes are used
                        d_s_bold[i] = sum(d_s)

                d_s_top_index = argsort(d_s_bold)[-Phase2_Ntop:] # extract top phase2_Ntop sols.
                S_top = S[d_s_top_index,:]
                fOCS_D[itx] = mean(d_s_bold)

                # testing convergence
                if itx >= convergence_lastiter:
                        if slope(fOCS_D[-convergence_lastiter:]) <= convergence_slope
                                break
```

```
            f_min, f_mean, f_max = min(d_s_bold), mean(d_s_bold), max(d_s_bold)
            scaling_factor = (f_max - f_mean) / (f_max - f_min)

            # generate new S array
            S_off = S_top[random(Phase2_Ntop, size=(Phase2_Ns - Phase2_Ntop)),:]
            # vertical recombination
            S_off = vertical_recombination(S_off, p_vcr, p_vcb, scaling_factor)
            # horizontal inversion
            S_off = horizontal_inversion(S_off, p_hir, p_hib, scaling_factor)

            S = vstack((S_top, S_off)) # stacking S_top and S_off -> new S

    # extract the most optimal solution
    s_optP2 = S[argmax(d_s_bold),:]
    S_allP2[npx,:] = s_optP2
```

References

Akanno, E. C., Chen, L., Abo-Ismail, M. K., Crowley, J. J., Wang, Z., Li, C., . . . Plastow, G. S. (2018). Genome-wide association scan for heterotic quantitative trait loci in multi-breed and crossbred beef cattle. *Genetics Selection Evolution, 50*(1), 48. doi:10.1186/s12711-018-0405-y

Akdemir, D., & Sánchez, J. I. (2016). Efficient Breeding by Genomic Mating. *Frontiers in Genetics, 7*, 210-210. doi:10.3389/fgene.2016.00210

Al-Mamun, H. A., A Clark, S., Kwan, P., & Gondro, C. (2015). Genome-wide linkage disequilibrium and genetic diversity in five populations of Australian domestic sheep. *Genetics Selection Evolution, 47*(1), 90. doi:10.1186/s12711-015-0169-6

Al-Mamun, H. A., Kwan, P., Clark, S. A., Ferdosi, M. H., Tellam, R., & Gondro, C. (2015b). Genome-wide association study of body weight in Australian Merino sheep reveals an orthologous region on OAR6 to human and bovine genomic regions affecting height and weight. *Genetics Selection Evolution, 47*(1), 66. doi:10.1186/s12711-015-0142-4

Aliloo, H., Pryce, J. E., González-Recio, O., Cocks, B. G., Goddard, M. E., & Hayes, B. J. (2017). Including nonadditive genetic effects in mating programs to maximize dairy farm profitability. *Journal of Dairy Science, 100*(2), 1203-1222. doi:https://doi.org/10.3168/jds.2016-11261

Alvarado, E., Sandberg, D. V., & Pickford, S. G. (1998). Modeling Large Forest Fires as Extreme Events. *Northwest Science, 72*.

An, B., Xu, L., Xia, J., Wang, X., Miao, J., Chang, T., . . . Gao, H. (2020). Multiple association analysis of loci and candidate genes that regulate body size at three growth stages in Simmental beef cattle. *BMC Genetics, 21*(1), 32. doi:10.1186/s12863-020-0837-6

Anderson, T. W., & Darling, D. A. (1952). Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics, 23*(2), 193-212, 120.

Anderson, T. W., & Darling, D. A. (1954). A Test of Goodness of Fit. *Journal of the American Statistical Association, 49*(268), 765-769. doi:10.2307/2281537

Ardlie, K. G., Kruglyak, L., & Seielstad, M. (2002). Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet, 3*(4), 299-309. doi:10.1038/nrg777

Baluja, S., & Caruana, R. (1995). *Removing the genetics from the standard genetic algorithm*. Paper presented at the Proceedings of the Twelfth International Conference on International Conference on Machine Learning, Tahoe City, California, USA.

Barankin, E. W., & Maitra, A. P. (1963). Generalization of the Fisher-Darmois-Koopman-Pitman Theorem on Sufficient Statistics. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002), 25*(3), 217-244.

Batt, R. D., Carpenter, S. R., & Ives, A. R. (2017). Extreme events in lake ecosystem time series. *Limnology and Oceanography Letters, 2*(3), 63-69. doi:https://doi.org/10.1002/lol2.10037

Beavis, W. D. (1994). *The power and deceit of QTL experiments: Lessons from comparative QTL studies*. Paper presented at the 49th Annual Corn and Sorghum Industry Research Conference, Washington DC.

Bedhane, M., van der Werf, J., Gondro, C., Duijvesteijn, N., Lim, D., Park, B., . . . Clark, S. (2019). Genome-Wide Association Study of Meat Quality Traits in Hanwoo Beef Cattle Using Imputed Whole-Genome Sequence Data. *Frontiers in Genetics, 10*(1235). doi:10.3389/fgene.2019.01235

Benjamin, A. T., & Quinn, J. J. (2003). Binomial Identities. In *Proofs that Really Count* (1 ed., Vol. 27, pp. 63-80): Mathematical Association of America.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics, 29*(4), 1165-1188.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Mach. Learn. Res., 13*(null), 281–305.

Bienaymé, I.-J. (1867). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés. *Journal de mathématiques pures et appliquées 2e série, 12*, 158-176.

Billiard, S., Castric, V., & Llaurens, V. (2021). The integrative biology of genetic dominance. *Biological Reviews, 96*(6), 2925-2942. doi:https://doi.org/10.1111/brv.12786

Bökönyi, S. (1974). *History of Domestic Mammals in Central and Eastern Europe*. Budapest: Akadémiai Kiadó.

Bossini-Castillo, L., Villanueva-Martin, G., Kerick, M., Acosta-Herrera, M., López-Isac, E., Simeón, C. P., . . . Martin, J. (2021). Genomic Risk Score impact on susceptibility to systemic sclerosis. *Annals of the Rheumatic Diseases, 80*(1), 118. doi:10.1136/annrheumdis-2020-218558

Boughorbel, S., Jarray, F., & El-Anbari, M. (2017). Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLOS ONE, 12*(6), e0177678. doi:10.1371/journal.pone.0177678

Braun, H. (1991). On solving travelling salesman problems by genetic algorithms. In H.-P. Schwefel & R. Männer (Eds.), *Parallel Problem Solving from Nature* (pp. 129-133). Berlin, Heidelberg: Springer Berlin Heidelberg.

Brieger, F. G. (1950). The Genetic Basis of Heterosis in Maize. *Genetics, 35*(4), 420-445. doi:10.1093/genetics/35.4.420

Broberg, P. (2005). A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics, 6*(1), 199. doi:10.1186/1471-2105-6-199

Brody, J. P., Williams, B. A., Wold, B. J., & Quake, S. R. (2002). Significance and statistical errors in the analysis of DNA microarray data. *Proceedings of the National Academy of Sciences, 99*(20), 12975-12978. doi:doi:10.1073/pnas.162468199

Brøndum, R. F., Su, G., Janss, L., Sahana, G., Guldbrandtsen, B., Boichard, D., & Lund, M. S. (2015). Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. *Journal of Dairy Science, 98*(6), 4107-4116. doi:https://doi.org/10.3168/jds.2014-9005

Brotherstone, S., & Goddard, M. (2005). Artificial selection and maintenance of genetic variance in the global dairy cow population. *Philos Trans R Soc Lond B Biol Sci, 360*(1459), 1479-1488. doi:10.1098/rstb.2005.1668

Bůžková, P. (2013). Linear Regression in Genetic Association Studies. *PLOS ONE, 8*(2), e56976. doi:10.1371/journal.pone.0056976

Caballero, A., Fernández, A., Villanueva, B., & Toro, M. A. (2022). A comparison of marker-based estimators of inbreeding and inbreeding depression. *Genetics Selection Evolution, 54*(1), 82. doi:10.1186/s12711-022-00772-0

Cabrelli, C. A., & Molter, U. M. (1995). The Kantorovich metric for probability measures on the circle. *Journal of Computational and Applied Mathematics, 57*(3), 345-361. doi:https://doi.org/10.1016/0377-0427(93)E0213-6

Cai, L., Wheeler, E., Kerrison, N. D., Luan, J. a., Deloukas, P., Franks, P. W., . . . Wareham, N. J. (2020). Genome-wide association analysis of type 2 diabetes in the EPIC-InterAct study. *Scientific Data, 7*(1), 393. doi:10.1038/s41597-020-00716-7

Cai, Z., Guldbrandtsen, B., Lund, M. S., & Sahana, G. (2019). Weighting sequence variants based on their annotation increases the power of genome-wide association studies in dairy cattle. *Genetics Selection Evolution, 51*(1), 20. doi:10.1186/s12711-019-0463-9

Chang, T., Xia, J., Xu, L., Wang, X., Zhu, B., Zhang, L., . . . Gao, H. (2018). A genome-wide association study suggests several novel candidate genes for carcass traits in Chinese Simmental beef cattle. *Animal Genetics, 49*(4), 312-316. doi:https://doi.org/10.1111/age.12667

Chapman, J. M., Cooper, J. D., Todd, J. A., & Clayton, D. G. (2003). Detecting Disease Associations due to Linkage Disequilibrium Using Haplotype Tags: A Class of Tests and the Determinants of Statistical Power. *Human Heredity, 56*(1/3), 18-31.

Charras-Garrido, M., & Lezaud, P. (2013). Extreme Value Analysis: an Introduction. *Journal de la société française de statistique, 154*(2), 66-97.

Chavez-Demoulin, V., & Embrechts, P. (2004). Smooth Extremal Models in Finance and Insurance. *The Journal of Risk and Insurance, 71*(2), 183-199.

Cheng, W., Ramachandran, S., & Crawford, L. (2020). Estimation of non-null SNP effect size distributions enables the detection of enriched genes underlying complex traits. *PLOS Genetics, 16*(6), e1008855. doi:10.1371/journal.pgen.1008855

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity, 87*(1), 52-58. doi:10.1046/j.1365-2540.2001.00901.x

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(1), 6. doi:10.1186/s12864-019-6413-7

Cirrone, G. A. P., Donadio, S., Guatelli, S., Mantero, A., Mascialino, B., Parlati, S., . . . Viarengo, P. (2004). A Goodness-of-Fit Statistical Toolkit. *IEEE Transactions on Nuclear Science, 51*, 2056-2063. doi:10.1109/tns.2004.836124

Clark, S. A., Kinghorn, B. P., Hickey, J. M., & van der Werf, J. H. J. (2013). The effect of genomic information on optimal contribution selection in livestock breeding programs. *Genetics Selection Evolution, 45*(1), 44. doi:10.1186/1297-9686-45-44

Consortini, A., & Rigal, F. (1998). Fractional moments and their usefulness in atmospheric laser scintillation. *Pure and Applied Optics: Journal of the European Optical Society Part A, 7*(5), 1013-1032. doi:10.1088/0963-9659/7/5/011

Cotsapas, C., & Mitrovic, M. (2018). Genome-wide association studies of multiple sclerosis. *Clin Transl Immunology, 7*(6), e1018. doi:10.1002/cti2.1018

Crow, J. F. (2010). On epistasis: why it is unimportant in polygenic directional selection. *Philos Trans R Soc Lond B Biol Sci, 365*(1544), 1241-1244. doi:10.1098/rstb.2009.0275

Crow, J. F., & Kimura, M. (1979). Efficiency of truncation selection. *Proc Natl Acad Sci U S A, 76*(1), 396-399. doi:10.1073/pnas.76.1.396

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R., de los Campos, G., & Hickey, J. M. (2013). Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. *Genetics, 193*(2), 347-365. doi:10.1534/genetics.112.147983

Daetwyler, H. D., Pong-Wong, R., Villanueva, B., & Woolliams, J. A. (2010). The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. *Genetics, 185*(3), 1021-1031. doi:10.1534/genetics.110.116855

Dakhlan, A., Moghaddar, N., Gondro, C., & Werf, J. H. J. v. d. (2017). Gene by birth type interaction in Merino lamb. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics Conference, 22*, 45 - 48.

de Almeida Filho, J. E., Guimarães, J. F. R., e Silva, F. F., de Resende, M. D. V., Muñoz, P., Kirst, M., & Resende, M. F. R. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity, 117*(1), 33-41. doi:10.1038/hdy.2016.23

de Boer, I. J. M., & Hoeschele, I. (1993). Genetic evaluation methods for populations with dominance and inbreeding. *Theoretical and Applied Genetics, 86*, 245 - 258.

de Smet, F., Moreau, Y., Engelen, K., Timmerman, D., Vergote, I., & De Moor, B. (2004). Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer, 91*(6), 1160-1165. doi:10.1038/sj.bjc.6602140

Dempfle, L. (1974). A note on increasing the limit of selection through selection within families. *Genetical Research, 24*(2), 127-135. doi:10.1017/S0016672300015160

Dempfle, L. (1990). Conservation, Creation, and Utilization of Genetic Variation. *Journal of Dairy Science, 73*(9), 2593-2600. doi:https://doi.org/10.3168/jds.S0022-0302(90)78946-1

Dobrushin, R. L. (1970). Prescribing a System of Random Variables by Conditional Distributions. *Theory of Probability and Its Applications, 15*, 458-486.

Dowd, C. (2020). A New ECDF Two-Sample Test Statistic. *arXiv: Methodology*.

Dremin, I. M. (1994). Fractional Moments of Distributions. *JETP Letters, 59*(9), 561 - 564.

Duenk, P., Bijma, P., Calus, M. P. L., Wientjes, Y. C. J., & van der Werf, J. H. J. (2020). The Impact of Non-additive Effects on the Genetic Correlation Between Populations. *G3: Genes|Genomes|Genetics, 10*(2), 783-795. doi:10.1534/g3.119.400663

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association, 56*(293), 52-64. doi:10.2307/2282330

Efron, B., Tibshirani, R., Storey, J. D., & Tusher, V. (2001). Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Association, 96*(456), 1151-1160.

Evans, B., & Walton, S. (2017). Aerodynamic optimisation of a hypersonic reentry vehicle based on solution of the Boltzmann–BGK equation and evolutionary optimisation. . *Applied Mathematical Modelling, 52*, 215 - 240. doi:doi:10.1016/j.apm.2017.07.024

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016). The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics, 24*(8), 1202-1205. doi:10.1038/ejhg.2015.269

Fagerland, M. W. (2012). t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC Medical Research Methodology, 12*(1), 78. doi:10.1186/1471-2288-12-78

Falconer, D. S. (1989). *Introduction to Quantitative Genetics* (3rd Edition ed.). Essex: Longman Scientific & Technical.

Fisher, R. A. (1949). *The Theory of Inbreeding*. London: Oliver and Boyd.

Fisher, R. A., & Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society, 24*(2), 180-190. doi:10.1017/S0305004100015681

Forstmeier, W., Wagenmakers, E.-J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – a practical guide. *Biological Reviews, 92*(4), 1941-1968. doi:10.1111/brv.12315

Fortin, J.-Y., & Clusel, M. (2015). Applications of extreme value statistics in physics. *Journal of Physics A: Mathematical and Theoretical, 48*(18), 183001. doi:10.1088/1751-8113/48/18/183001

Fraser, A. (1957). Simulation of Genetic Systems by Automatic Digital Computers I. Introduction. *Australian Journal of Biological Sciences, 10*(4), 484-491. doi:https://doi.org/10.1071/BI9570484

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., & Province, M. A. (2010). Avoiding the high Bonferroni penalty in genome-wide association studies. *Genetic epidemiology, 34*(1), 100-105. doi:10.1002/gepi.20430

Garcia-Baccino, C. A., Lourenco, D. A. L., Miller, S., Cantet, R. J. C., & Vitezica, Z. G. (2020). Estimating dominance genetic variances for growth traits in American Angus males using genomic models. *Journal of animal science, 98*(1), skz384. doi:10.1093/jas/skz384

Gaynor, R. C., Gorjanc, G., & Hickey, J. M. (2021). AlphaSimR: an R package for breeding program simulations. *G3 Genes|Genomes|Genetics, 11*(2), jkaa017. doi:10.1093/g3journal/jkaa017

Gerges, F., Zouein, G., & Azar, D. (2018). *Genetic Algorithms with Local Optima Handling to Solve Sudoku Puzzles*. Paper presented at the Proceedings of the 2018 International Conference on Computing and Artificial Intelligence, Chengdu, China. https://doi.org/10.1145/3194452.3194463

Ghasemi, M., Zamani, P., Vatankhah, M., & Abdoli, R. (2019). Genome-wide association study of birth weight in sheep. *Animal, 13*(9), 1797-1803. doi:https://doi.org/10.1017/S1751731118003610

Gibson, J., Morton, N. E., & Collins, A. (2006). Extended tracts of homozygosity in outbred human populations. *Human molecular genetics, 15*(5), 789-795. doi:10.1093/hmg/ddi493

Gill, J. J. B., & Harland, M. (1992). Maximal Maintenance of Genetic Variation in Small Population. In L. Alderson & I. Bodó (Eds.), *Genetic Conservation of Domestic Livestock* (Vol. 2, pp. 3-17). Walingford: CAB International.

Gondro, C. (2015). *Primer to analysis of genomic data using R*: Springer.

Gondro, C., & Kinghorn, B. P. (2009). *Application of Evolutionary Algorithms to Solve Complex Problems in Quantitative Genetics and Bioinformatics*. Guelph: University of Guelph.

González-Diéguez, D., Tusell, L., Carillier-Jacquin, C., Bouquet, A., & Vitezica, Z. G. (2019). SNP-based mate allocation strategies to maximize total genetic value in pigs. *Genetics Selection Evolution, 51*(1), 55. doi:10.1186/s12711-019-0498-y

Gorodkin, J. (2004). Comparing two K-category assignments by a K-category correlation coefficient. *Computational Biology and Chemistry, 28*(5), 367-374. doi:https://doi.org/10.1016/j.compbiolchem.2004.09.006

Goto, E., & Nordskog, A. W. (1959). Heterosis in Poultry: 4. Estimation of Combining Ability Variance from Diallel Crosses of Inbred Lines in the Fowl1. *Poultry Science, 38*(6), 1381-1388. doi:https://doi.org/10.3382/ps.0381381

Goudey, B., Rawlinson, D., Wang, Q., Shi, F., Ferra, H., Campbell, R. M., . . . Kowalczyk, A. (2013). GWIS--model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomics, 14 Suppl 3*(Suppl 3), S10-S10. doi:10.1186/1471-2164-14-S3-S10

Gourdine, J. L., Sørensen, A. C., & Rydhmer, L. (2012). There is room for selection in a small local pig breed when using optimum contribution selection: A simulation study1,2. *Journal of animal science, 90*(1), 76-84. doi:10.2527/jas.2011-3898

Grapes, L., Dekkers, J. C. M., Rothschild, M. F., & Fernando, R. L. (2004). Comparing Linkage Disequilibrium-Based Methods for Fine Mapping Quantitative Trait Loci. *Genetics, 166*(3), 1561-1570. doi:10.1534/genetics.166.3.1561

Griffiths, A. J. F., Wessler, S. R., Carroll, S. B., & Doebley, J. (2015). *Introduction to Genetic Analysis*. New York: Macmillan Learning.

Guo, J., Jiang, R., Mao, A., Liu, G. E., Zhan, S., Li, L., . . . Zhang, H. (2021). Genome-wide association study reveals 14 new SNPs and confirms two structural variants highly

associated with the horned/polled phenotype in goats. *BMC Genomics, 22*(1), 769. doi:10.1186/s12864-021-08089-w

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia medica, 26*(3), 297-307. doi:10.11613/BM.2016.034

Habier, D., Fernando, R. L., Kizilkaya, K., & Garrick, D. J. (2011). Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics, 12*(1), 186. doi:10.1186/1471-2105-12-186

Hall, D., Hallingbäck, H. R., & Wu, H. X. (2016). Estimation of number and size of QTL effects in forest tree traits. *Tree Genetics & Genomes, 12*(6), 110. doi:10.1007/s11295-016-1073-0

Hart, A. (2001). Mann-Whitney test is not just a test of medians: differences in spread can be important. *BMJ (Clinical research ed.), 323*(7309), 391-393. doi:10.1136/bmj.323.7309.391

Haupt, R. L., & Werner, D. H. (2007). *Genetic Algorithms in Electromagnetics*. Hoboken: John Wiley & Sons, Inc.

Hay, E. H., & Roberts, A. (2018). Genome-wide association study for carcass traits in a composite beef cattle breed. *Livestock Science, 213*, 35-43. doi:https://doi.org/10.1016/j.livsci.2018.04.018

Hayes, B. (2013). Overview of Statistical Methods for Genome-Wide Association Studies (GWAS). In C. Gondro, J. van der Werf, & B. Hayes (Eds.), *Genome-Wide Association Studies and Genomic Prediction* (pp. 149-169). Totowa, NJ: Humana Press.

Heider, H., & Drabe, T. (1997). A cascaded genetic algorithm for improving fuzzy-system design. *International Journal of Approximate Reasoning, 17*(4), 351-368. doi:https://doi.org/10.1016/S0888-613X(97)00003-0

Heller, R., & Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics, 8*(1), 481-498, 418.

Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics, 31*(2), 423-447. doi:10.2307/2529430

Henderson, C. R. (1976). A Simple Method for Computing the Inverse of a Numerator Relationship Matrix Used in Prediction of Breeding Values. *Biometrics, 32*(1), 69-83. doi:10.2307/2529339

Hill, W. G. (2010). Understanding and using quantitative genetic variation. *Philos Trans R Soc Lond B Biol Sci, 365*(1537), 73-85. doi:10.1098/rstb.2009.0203

Hoerl, A. E., & Kennard, R. W. (2000). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics, 42*(1), 80-86. doi:10.2307/1271436

Högbom, J. A. (1974). Aperture Synthesis with a Non-Regular Distribution of Interferometer Baselines. *Astronomy and Astrophysics Supplement Series, 15*, 417.

Hoggart, C. J., Clark, T. G., De Iorio, M., Whittaker, J. C., & Balding, D. J. (2008). Genome-wide significance for dense SNP and resequencing data. *Genetic epidemiology, 32*(2), 179-185. doi:10.1002/gepi.20292

Hong, E. P., & Park, J. W. (2012). Sample size and statistical power calculation in genetic association studies. *Genomics & informatics, 10*(2), 117-122. doi:10.5808/GI.2012.10.2.117

Hou, Z., & Ochoa, A. (2023). Genetic association models are robust to common population kinship estimation biases. *Genetics*. doi:10.1093/genetics/iyad030

Hu, Z., Wang, Z., & Xu, S. (2012). An Infinitesimal Model for Quantitative Trait Genomic Value Prediction. *PLOS ONE, 7*(7), e41336. doi:10.1371/journal.pone.0041336

Huang, Qin Q., Ritchie, S. C., Brozynska, M., & Inouye, M. (2018). Power, false discovery rate and Winner's Curse in eQTL studies. *Nucleic Acids Research, 46*(22), e133-e133. doi:10.1093/nar/gky780

Ibeagha-Awemu, E. M., Peters, S. O., Akwanji, K. A., Imumorin, I. G., & Zhao, X. (2016). High density genome wide genotyping-by-sequencing and association identifies common and low frequency SNPs, and novel candidate genes influencing cow milk traits. *Scientific Reports, 6*(1), 31109. doi:10.1038/srep31109

Ioannidis, J. P. A. (2007). Non-Replication and Inconsistency in the Genome-Wide Association Setting. *Human Heredity, 64*(4), 203-213.

Ionita-Laza, I., Cho, M. H., & Laird, N. M. (2013). Statistical Challenges in Sequence-Based Association Studies with Population- and Family-Based Designs. *Statistics in Biosciences, 5*(1), 54-70. doi:10.1007/s12561-012-9062-9

Isik, F., Li, B., & Frampton, J. (2003). Estimates of Additive, Dominance and Epistatic Genetic Variances from a Clonally Replicated Test of Loblolly Pine. *Forest Science, 49*(1), 77-88. doi:10.1093/forestscience/49.1.77

Jamieson, K., & Talwalkar, A. (2015). *Non-stochastic best arm identification and hyperparameter optimization.* Paper presented at the International Conference on Artificial Intelligence and Statistics (AISTATS).

Jansen, P. R., Watanabe, K., Stringer, S., Skene, N., Bryois, J., Hammerschlag, A. R., . . . Posthuma, D. (2019). Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet, 51*(3), 394-403. doi:10.1038/s41588-018-0333-3

Jasielczuk, I., Gurgul, A., Szmatoła, T., Semik-Gurgul, E., Pawlina-Tyszko, K., Stefaniuk-Szmukier, M., . . . Bugno-Poniewierska, M. (2020). Linkage disequilibrium, haplotype blocks and historical effective population size in Arabian horses and selected Polish native horse breeds. *Livestock Science, 239*, 104095. doi:https://doi.org/10.1016/j.livsci.2020.104095

Jiang, L., Li, X., & Zhang, J. (2009). Design of high performance multilayer microwave absorbers using fast Pareto genetic algorithm. *Science in China Series E: Technological Sciences, 52*(9), 2749-2757. doi:10.1007/s11431-009-0145-x

Jiang, J., Ma, L., Prakapenka, D., VanRaden, P. M., Cole, J. B., & Da, Y. (2019). A Large-Scale Genome-Wide Association Study in U.S. Holstein Cattle. *Frontiers in Genetics, 10*, 412-412. doi:10.3389/fgene.2019.00412

Kaivanto, K. (2008). Maximization of the sum of sensitivity and specificity as a diagnostic cutpoint criterion. *J Clin Epidemiol, 61*(5), 517-518. doi:10.1016/j.jclinepi.2007.10.011

Kaler, A. S., & Purcell, L. C. (2019). Estimation of a significance threshold for genome-wide association studies. *BMC Genomics, 20*(1), 618. doi:10.1186/s12864-019-5992-7

Kantorovich, L. V. (1939). Mathematical Methods of Organizing and Planning Production. *Translated in Management Science (1960), 6*(4), 363-505.

Kantorovitch, L. (1958). On the Translocation of Masses. *Management Science, 5*(1), 1-4. doi:10.1287/mnsc.5.1.1

Khattri, S. K., & Witkowski, A. (2012). Euler's Number and Some Means. *Tamsui Oxford Journal of Information and Mathematical Sciences, 28*(4), 369 - 377.

Kijas, J. W., Porto-Neto, L., Dominik, S., Reverter, A., Bunch, R., McCulloch, R., . . . Consortium, t. I. S. G. (2014). Linkage disequilibrium over short physical distances measured in sheep using a high-density SNP chip. *Animal Genetics, 45*(5), 754-757. doi:https://doi.org/10.1111/age.12197

Kinghorn, B. P. (2000). The tactical approach to implementing breeding program. In B. P. Kinghorn, J. H. J. Van der Werf, & M. Ryan (Eds.), *Animal Breeding – use of new technologies*: Universtiy of Sydney Veterinary Post Graduate Foundation.

Kinghorn, B. P. (2011). An algorithm for efficient constrained mate selection. *Genetics Selection Evolution, 43*(1), 4. doi:10.1186/1297-9686-43-4

Klein, R. J. (2007). Power analysis for genome-wide association studies. *BMC Genetics, 8*(1), 58. doi:10.1186/1471-2156-8-58

Kooperberg, C., LeBlanc, M., & Obenchain, V. (2010). Risk prediction using genome-wide association studies. *Genetic epidemiology, 34*(7), 643-652. doi:10.1002/gepi.20509

Koopman, B. O. (1936). On Distributions Admitting a Sufficient Statistic. *Transactions of the American Mathematical Society, 39*(3), 399-409. doi:10.2307/1989758

Kraft, P., & Hunter, D. J. (2009). Genetic risk prediction--are we there yet? *N Engl J Med, 360*(17), 1701-1703. doi:10.1056/NEJMp0810107

Kremelberg, D. (2011). Practical Statistics: A Quick and Easy Guide to IBM® SPSS® Statistics, STATA, and Other Statistical Software. In. Thousand Oaks,

California: SAGE Publications, Inc. Retrieved from https://methods.sagepub.com/book/practical-stats. doi:10.4135/9781483385655

Kuiper, N. H. (1960). *Tests concerning random points on a circle.* Paper presented at the Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen, Series A.

Lanzante, J. R. (2021). Testing for differences between two distributions in the presence of serial correlation using the Kolmogorov–Smirnov and Kuiper's tests. *International Journal of Climatology, 41*(14), 6314-6323. doi:https://doi.org/10.1002/joc.7196

Laodim, T., Elzo, M. A., Koonawootrittriron, S., Suwanasopee, T., & Jattawa, D. (2019). Genomic-polygenic and polygenic predictions for milk yield, fat yield, and age at first calving in Thai multibreed dairy population using genic and functional sets of genotypes. *Livestock Science, 219*, 17-24. doi:https://doi.org/10.1016/j.livsci.2018.11.008

Lawlor, G. R. (2020). l'Hôpital's Rule for Multivariable Functions. *The American Mathematical Monthly, 127*(8), 717-725. doi:10.1080/00029890.2020.1793635

Legendre, A. M. (1805). *Nouvelles méthodes pour la détermination des orbites des comètes*. Paris: Firmin Didot.

Li, C. C., & Horvitz, D. G. (1953). Some methods of estimating the inbreeding coefficient. *American journal of human genetics, 5*(2), 107-117.

Llinares-López, F., Grimm, D. G., Bodenham, D. A., Gieraths, U., Sugiyama, M., Rowan, B., & Borgwardt, K. (2015). Genome-wide detection of intervals of genetic heterogeneity associated with complex traits. *Bioinformatics, 31*(12), i240-i249. doi:10.1093/bioinformatics/btv263

Lloyd-Jones, L. R., Zeng, J., Sidorenko, J., Yengo, L., Moser, G., Kemper, K. E., . . . Visscher, P. M. (2019). Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nature Communications, 10*(1), 5086. doi:10.1038/s41467-019-12653-0

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography, 17*(2), 145-151. doi:https://doi.org/10.1111/j.1466-8238.2007.00358.x

Lohn, J. D., Linden, D. S., & Hornby, G. (2008). *Advanced Antenna Design for a NASA Small Satellite Mission.* Paper presented at the 22nd Annual AIAA/USU Conference on Small Satellites, Logan.

Lukacs, E. (1942). A Characterization of the Normal Distribution. *The Annals of Mathematical Statistics, 13*(1), 91-93. doi:10.1214/aoms/1177731647

Lynch, M., & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sunderland: Sinauer Associates, Inc.

Magori, K., & Gould, F. (2006). Genetically engineered underdominance for manipulation of pest populations: a deterministic model. *Genetics, 172*(4), 2613-2620. doi:10.1534/genetics.105.051789

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature, 456*(7218), 18-21. doi:10.1038/456018a

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., . . . Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature, 461*(7265), 747-753. doi:10.1038/nature08494

Marasco, I., Niro, G., Mastronardi, V. M., Rizzi, F., D'Orazio, A., De Vittorio, M., & Grande, M. (2022). A compact evolved antenna for 5G communications. *Scientific Reports, 12*(1), 10327. doi:10.1038/s41598-022-14447-9

Mason, D. M., & Schuenemeyer, J. H. (1983). A Modified Kolmogorov-Smirnov Test Sensitive to Tail Alternatives. *The Annals of Statistics, 11*(3), 933-946.

McQuillan, R., Leutenegger, A.-L., Abdel-Rahman, R., Franklin, C. S., Pericic, M., Barac-Lauc, L., . . . Wilson, J. F. (2008). Runs of homozygosity in European populations. *American journal of human genetics, 83*(3), 359-372. doi:10.1016/j.ajhg.2008.08.007

Mendel, G. (1865). *Versuche über Pflanzen-Hybriden* (Vol. 4). Brno: Verlage des Vereines.

Meuwissen, T. H. (1997). Maximizing the response of selection with a predefined rate of inbreeding. *Journal of animal science, 75*(4), 934-940. doi:10.2527/1997.754934x

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics, 157*(4), 1819-1829.

Meuwissen, T. H. E., & Luo, Z. (1992). Computing inbreeding coefficients in large populations. *Genetics Selection Evolution, 24*(4), 305. doi:10.1186/1297-9686-24-4-305

Misztal, I., Lawlor, T. J., & Gengler, N. (1997). Relationships among estimates of inbreeding depression, dominance and additive variance for linear traits in Holsteins. *Genetics Selection Evolution, 29*(3), 319. doi:10.1186/1297-9686-29-3-319

Mitchell, M., & Forrest, S. (1994). Genetic Algorithms and Artificial Life. *Artificial Life, 1*(3), 267-289. doi:10.1162/artl.1994.1.3.267

Moghaddar, N., Khansefid, M., van der Werf, J. H. J., Bolormaa, S., Duijvesteijn, N., Clark, S. A., . . . MacLeod, I. M. (2019). Genomic prediction based on selected variants from imputed whole-genome sequence data in Australian sheep populations. *Genetics Selection Evolution, 51*(1), 72. doi:10.1186/s12711-019-0514-2

Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., & Visscher, P. M. (2015). Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model. *PLOS Genetics, 11*(4), e1004969. doi:10.1371/journal.pgen.1004969

Mueller, J. C. (2004). Linkage disequilibrium for different scales and applications. *Brief Bioinform, 5*(4), 355-364. doi:10.1093/bib/5.4.355

Mühlenbein, H. (1992). Parallel Genetic Algorithm in Combinatorial Optimization. In O. Balci, R. Sharda, & S. Zenios (Eds.), *Computer Science and Operations Research* (pp. 441 - 456). New York: Pergamon Press.

Mun, J. (2012). *Advanced Analytical Models: Over 800 Models and 300 Applications from the Basel II Accord to Wall Street and Beyond*. Hoboken: John Wiley & Sons, Inc.

Naaman, M. (2021). On the tight constant in the multivariate Dvoretzky–Kiefer–Wolfowitz inequality. *Statistics & Probability Letters, 173*, 109088. doi:https://doi.org/10.1016/j.spl.2021.109088

Nayeri, S., Sargolzaei, M., Abo-Ismail, M. K., May, N., Miller, S. P., Schenkel, F., . . . Stothard, P. (2016). Genome-wide association for milk production and female fertility traits in Canadian dairy Holstein cattle. *BMC Genet, 17*(1), 75. doi:10.1186/s12863-016-0386-1

Nazir, H. Z. (2014). *Robust control charts in statistical process control.* (PhD Thesis), University of Amsterdam,

Nehrenberg, D. L., Wang, S., Buus, R. J., Perkins, J., de Villena, F. P., & Pomp, D. (2010). Genomic mapping of social behavior traits in a F2 cross derived from mice selectively bred for high aggression. *BMC Genet, 11*, 113. doi:10.1186/1471-2156-11-113

Newberry, M. G., McCandlish, D. M., & Plotkin, J. B. (2016). Assortative mating can impede or facilitate fixation of underdominant alleles. *Theoretical Population Biology, 112*, 14-21. doi:https://doi.org/10.1016/j.tpb.2016.07.003

Nielsen, H. M., Sonesson, A. K., & Meuwissen, T. H. E. (2011). Optimum contribution selection using traditional best linear unbiased prediction and genomic breeding values in aquaculture breeding schemes. *Journal of animal science, 89*(3), 630-638. doi:10.2527/jas.2009-2731

Nishino, J., Ochi, H., Kochi, Y., Tsunoda, T., & Matsui, S. (2018). Sample Size for Successful Genome-Wide Association Study of Major Depressive Disorder. *Frontiers in Genetics, 9*, 227-227. doi:10.3389/fgene.2018.00227

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American journal of human genetics, 74*(4), 765-769. doi:10.1086/383251

O'Connor, L. J. (2021). The distribution of common-variant effect sizes. *Nature Genetics, 53*(8), 1243-1249. doi:10.1038/s41588-021-00901-3

Orr, A. H. (1999). The evolutionary genetics of adaptation: a simulation study. *Genetics Research, 74*(3), 207-214. doi:10.1017/S0016672399004164

Panagiotou, O. A., & Ioannidis, J. P. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol, 41*(1), 273-286. doi:10.1093/ije/dyr178

Park, J.-H., Gail, M. H., Weinberg, C. R., Carroll, R. J., Chung, C. C., Wang, Z., . . . Chatterjee, N. (2011). Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants. *Proceedings of the National Academy of Sciences, 108*(44), 18026-18031. doi:10.1073/pnas.1114759108

Park, J.-H., Wacholder, S., Gail, M. H., Peters, U., Jacobs, K. B., Chanock, S. J., & Chatterjee, N. (2010). Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genetics, 42*(7), 570-575. doi:10.1038/ng.610

Patron, J., Serra-Cayuela, A., Han, B., Li, C., & Wishart, D. S. (2019). Assessing the performance of genome-wide association studies for predicting disease risk. *PLOS ONE, 14*(12), e0220215-e0220215. doi:10.1371/journal.pone.0220215

Pearson, K. (1924). Historical Note on the Origin of the Normal Curve of Errors. *Biometrika, 16*(3/4), 402-404. doi:10.2307/2331714

Pearson, T. A., & Manolio, T. A. (2008). How to interpret a genome-wide association study. *Jama, 299*(11), 1335-1344. doi:10.1001/jama.299.11.1335

Pegolo, S., Cecchinato, A., Savoia, S., Di Stasio, L., Pauciullo, A., Brugiapaglia, A., . . . Albera, A. (2020). Genome-wide association and pathway analysis of carcass and meat quality traits in Piemontese young bulls. *Animal, 14*(2), 243-252. doi:10.1017/s1751731119001812

Pengelly, R. J., Tapper, W., Gibson, J., Knut, M., Tearle, R., Collins, A., & Ennis, S. (2015). Whole genome sequences are required to fully resolve the linkage disequilibrium structure of human populations. *BMC Genomics, 16*(1), 666. doi:10.1186/s12864-015-1854-0

Picoli, S., Mendes, R. S., Malacarne, L. C., & Santos, R. P. B. (2009). q-Distributions in Complex System: A Brief Review. *Brazilian Journal of Physics, 39*(2a), 468-474.

Pillai, N. S., & Meng, X.-L. (2016). An unexpected encounter with Cauchy and Lévy. *The Annals of Statistics, 44*(5), 2089-2097, 2089.

Plackett, R. L. (1958). Studies in the History of Probability and Statistics: VII. The Principle of the Arithmetic Mean. *Biometrika, 45*(1/2), 130-135. doi:10.2307/2333051

Porto-Neto, L. R., Lee, S. H., Sonstegard, T. S., Van Tassell, C. P., Lee, H. K., Gibson, J. P., & Gondro, C. (2014). Genome-wide detection of signatures of selection in Korean Hanwoo cattle. *Animal Genetics, 45*(2), 180-190. doi:https://doi.org/10.1111/age.12119

Posbergh, C. J., & Huson, H. J. (2021). All sheeps and sizes: a genetic investigation of mature body size across sheep breeds reveals a polygenic nature. *Animal Genetics, 52*(1), 99-107. doi:https://doi.org/10.1111/age.13016

Pryce, J. E., Bolormaa, S., Chamberlain, A. J., Bowman, P. J., Savin, K., Goddard, M. E., & Hayes, B. J. (2010). A validated genome-wide association study in 2 dairy cattle breeds for milk production and fertility traits using variable length haplotypes. *Journal of Dairy Science, 93*(7), 3331-3345. doi:https://doi.org/10.3168/jds.2009-2893

Purfield, D. C., Berry, D. P., McParland, S., & Bradley, D. G. (2012). Runs of homozygosity and population history in cattle. *BMC Genetics, 13*(1), 70. doi:10.1186/1471-2156-13-70

Quaas, R. L. (1976). Computing the Diagonal Elements and Inverse of a Large Numerator Relationship Matrix. *Biometrics, 32*(4), 949-953. doi:10.2307/2529279

Ramsey, J. B., Harvill, J. L., & Newton, H. J. (2002). The elements of statistics with applications to economics and the social sciences. In. Stamford, Conn.: Duxbury Thompson Learning.

Ren, X., Yang, G.-L., Peng, W.-F., Zhao, Y.-X., Zhang, M., Chen, Z.-H., . . . Li, M.-H. (2016). A genome-wide association study identifies a genomic region for the

polycerate phenotype in sheep (Ovis aries). *Scientific Reports, 6*(1), 21111. doi:10.1038/srep21111

Robertson, A. (1970). Some optimum problems in individual selection. *Theoretical Population Biology, 1*(1), 120-127. doi:https://doi.org/10.1016/0040-5809(70)90045-6

Ryder, M. L. (1973). The Use of the Skin and Coat in Studies of Changes Following Domestication. In J. Matolcsi (Ed.), *Domestikationsforschung und Geschichte der Haustiere* (pp. 163-168). Budapest: Akadémiai Kiadó.

Ryder, O. A., & Wedemeyer, E. A. (1982). A cooperative breeding programme for the Mongolian wild horse Equus przewalskii in the United States. *Biological Conservation, 22*(4), 259-271. doi:https://doi.org/10.1016/0006-3207(82)90021-0

Savitzky, A., & Golay, M. J. E. (1964). Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Analytical Chemistry, 36*(8), 1627-1639. doi:10.1021/ac60214a047

Schafer, R. W. (2011, 4-7 Jan. 2011). *On the frequency-domain properties of Savitzky-Golay filters.* Paper presented at the 2011 Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE).

Scheper, C., Emmerling, R., Götz, K.-U., & König, S. (2021). A variance component estimation approach to infer associations between Mendelian polledness and quantitative production and female fertility traits in German Simmental cattle. *Genetics Selection Evolution, 53*(1), 60. doi:10.1186/s12711-021-00652-z

Schisterman, E. F., Perkins, N. J., Liu, A., & Bondell, H. (2005). Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology, 16*(1), 73-81. doi:10.1097/01.ede.0000147512.81966.ba

Schlie, A. (1967). *Der Hannoveraner: Geschichte und Zucht des Edlen Hannoverschen Warmblutpferdes*. Munich: BLV Bayerischer Landwirtschaftsverlag.

Schmüdgen, K. (2020). Ten Lectures on the Moment Problem. *arXiv preprint arXiv:2008.12698*.

Scott, D. W. (1985). Averaged Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions. *The Annals of Statistics, 13*(3), 1024-1040, 1017.

Sevinga, M., Vrijenhoek, T., Hesselink, J. W., Barkema, H. W., & Groen, A. F. (2004). Effect of inbreeding on the incidence of retained placenta in Friesian horses1. *Journal of animal science, 82*(4), 982-986. doi:10.2527/2004.824982x

Shafquat, A., Crystal, R. G., & Mezey, J. G. (2020). Identifying novel associations in GWAS by hierarchical Bayesian latent variable detection of differentially misclassified phenotypes. *BMC Bioinformatics, 21*(1), 178. doi:10.1186/s12859-020-3387-z

Shen, X., & Carlborg, Ö. (2013). Beware of risk for increased false positive rates in genome-wide association studies for phenotypic variability. *Frontiers in Genetics, 4*(93). doi:10.3389/fgene.2013.00093

Shepherd, R., & Kinghorn, B. (1998). *A tactical approach to the design of crossbreeding programs.* Paper presented at the Proceedings of the Sixth World Congress on Genetics Applied to Livestock Production: 11-16 January; Armidale.

Signer-Hasler, H., Flury, C., Haase, B., Burger, D., Simianer, H., Leeb, T., & Rieder, S. (2012). A Genome-Wide Association Study Reveals Loci Influencing Height and Other Conformation Traits in Horses. *PLOS ONE, 7*(5), e37282. doi:10.1371/journal.pone.0037282

Simard, R., & L'Ecuyer, P. (2011). Computing the Two-Sided Kolmogorov-Smirnov Distribution. *Journal of Statistical Software, 39*(11), 1 - 18. doi:10.18637/jss.v039.i11

Simes, R. J. (1986). An Improved Bonferroni Procedure for Multiple Tests of Significance. *Biometrika, 73*(3), 751-754. doi:10.2307/2336545

Singer, J. M., & Andrade, D. F. (2010). Large-sample Statistical Methods. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education (Third Edition)* (pp. 232-237). Oxford: Elsevier.

Slowik, A., & Kwasnicka, H. (2020). Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications, 32*(16), 12363-12379. doi:10.1007/s00521-020-04832-8

Smołucha, G., Gurgul, A., Jasielczuk, I., Kawęcka, A., & Miksza-Cybulska, A. (2021). A genome-wide association study for prolificacy in three Polish sheep breeds. *Journal of Applied Genetics, 62*, 323-326. doi:10.1007/s13353-021-00615-6

Sørensen, M. K., Sørensen, A. C., Baumung, R., Borchersen, S., & Berg, P. (2008). Optimal genetic contribution selection in Danish Holstein depends on pedigree quality. *Livestock Science, 118*(3), 212-222. doi:https://doi.org/10.1016/j.livsci.2008.01.027

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009). Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLOS Genetics, 5*(5), e1000477. doi:10.1371/journal.pgen.1000477

Srinivas, M., & Patnaik, L. M. (1994). Adaptive probabilities of crossover and mutation in genetic algorithms. *IEEE Transactions on Systems, Man, and Cybernetics, 24*(4), 656-667. doi:10.1109/21.286385

Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov, Cramer-Von Mises and Related Statistics Without Extensive Tables. *Journal of the Royal Statistical Society. Series B (Methodological), 32*(1), 115-122.

Steri, R., Moioli, B., Catillo, G., Galli, A., & Buttazzoni, L. (2019). Genome-wide association study for longevity in the Holstein cattle population. *Animal, 13*(7), 1350-1357. doi:https://doi.org/10.1017/S1751731118003191

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(3), 479-498. doi:10.1111/1467-9868.00346

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q - value. *Ann. Statist., 31*(6), 2013-2035. doi:10.1214/aos/1074290335

Storn, R., & Price, K. (1997). Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *J. of Global Optimization, 11*(4), 341–359. doi:10.1023/a:1008202821328

Streiner, D. L., & Cairney, J. (2007). What's under the ROC? An Introduction to Receiver Operating Characteristics Curves. *The Canadian Journal of Psychiatry, 52*(2), 121-128. doi:10.1177/070674370705200210

Sun, C., VanRaden, P. M., O'Connell, J. R., Weigel, K. A., & Gianola, D. (2013). Mating programs including genomic relationships and dominance effects1. *Journal of Dairy Science, 96*(12), 8014-8023. doi:https://doi.org/10.3168/jds.2013-6969

Sved, J. A. (1971). Linkage disequilibrium and homozygosity of chromosome segments in finite populations. *Theoretical Population Biology, 2*(2), 125-141. doi:https://doi.org/10.1016/0040-5809(71)90011-6

Taherdangkoo, M., Paziresh, M., Yazdi, M., & Bagheri, M. (2013). An efficient algorithm for function optimization: modified stem cells algorithm. *Open Engineering, 3*(1), 36-50. doi:doi:10.2478/s13531-012-0047-8

Tam, V., Patel, N., Turcotte, M., Bossé, Y., Paré, G., & Meyre, D. (2019). Benefits and limitations of genome-wide association studies. *Nature Reviews Genetics, 20*(8), 467-484. doi:10.1038/s41576-019-0127-1

Tchebichef, P. (1874). Sur les valeurs limites des intégrales. *Journal de mathématiques pures et appliquées 2e série, 19*, 157-160.

Tchebychef, P. L., Markoff, A. A., & Sonin, N. (1907). Sur deux théorèmes relatifs aux probabilités. In *Oeuvres de P.L. Tchebychef, Tome II* (pp. 481-491). St. Petersburg: Commissionaires de l'Académie impériale des sciences.

Tian, R., Cox, S. H., & Zuluaga, L. F. (2017). Moment Problem and Its Applications to Risk Assessment. *North American Actuarial Journal, 21*(2), 242-266. doi:10.1080/10920277.2017.1302805

Tippett, M. K., Lepore, C., & Cohen, J. E. (2016). More tornadoes in the most extreme U.S. tornado outbreaks. *Science (New York, N.Y.), 354*(6318), 1419-1423. doi:doi:10.1126/science.aah7393

Toro, M., & Pérez-Enciso, M. (1990). Optimization of selection response under restricted inbreeding. *Genetics Selection Evolution, 22*(1), 93. doi:10.1186/1297-9686-22-1-93

Toro Ospina, A. M., Maiorano, A. M., Curi, R. A., Pereira, G. L., Zerlotti-Mercadante, M. E., dos Santos Gonçalves Cyrillo, J. N., . . . de V. Silva, J. A. I. (2019). Linkage disequilibrium and effective population size in Gir cattle selected for yearling weight. *Reproduction in Domestic Animals, 54*(12), 1524-1531. doi:https://doi.org/10.1111/rda.13559

VanRaden, P. (1992). Accounting for inbreeding and crossbreeding in genetic evaluation of large populations. *Journal of Dairy Science, 75*(11), 3136-3144.

VanRaden, P. M. (2008). Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science, 91*(11), 4414-4423. doi:https://doi.org/10.3168/jds.2007-0980

Vanvanhossou, S. F. U., Scheper, C., Dossa, L. H., Yin, T., Brügemann, K., & König, S. (2020). A multi-breed GWAS for morphometric traits in four Beninese indigenous cattle breeds reveals loci associated with conformation, carcass and adaptive traits. *BMC Genomics, 21*(1), 783. doi:10.1186/s12864-020-07170-0

Vaserstein, L. N. (1969). Markov Process over Denumerable Products of Spaces, Describing Large Systems of Automata. *Problemy Peredachi Informatsii, 5*(3), 64-72.

Visscher, P. M., Wray, N. R., Zhang, Q., Sklar, P., McCarthy, M. I., Brown, M. A., & Yang, J. (2017). 10 Years of GWAS Discovery: Biology, Function, and Translation. *American journal of human genetics, 101*(1), 5-22. doi:10.1016/j.ajhg.2017.06.005

Vitezica, Z. G., Reverter, A., Herring, W., & Legarra, A. (2018). Dominance and epistatic genetic variances for litter size in pigs using genomic models. *Genetics, selection, evolution : GSE, 50*(1), 71-71. doi:10.1186/s12711-018-0437-3

Vitezica, Z. G., Varona, L., Elsen, J.-M., Misztal, I., Herring, W., & Legarra, A. (2016). Genomic BLUP including additive and dominant variation in purebreds and F1 crossbreds, with an application in pigs. *Genetics Selection Evolution, 48*(1), 6. doi:10.1186/s12711-016-0185-1

Vitezica, Z. G., Varona, L., & Legarra, A. (2013). On the additive and dominant variance and covariance of individuals within the genomic selection scope. *Genetics, 195*(4), 1223-1230. doi:10.1534/genetics.113.155176

Wang, M., & Xu, S. (2019). Statistical power in genome-wide association studies and quantitative trait locus mapping. *Heredity, 123*(3), 287-306. doi:10.1038/s41437-019-0205-3

Wang, P., & Zhu, W. (2019). Replicability analysis in genome-wide association studies via Cartesian hidden Markov models. *BMC Bioinformatics, 20*(1), 146. doi:10.1186/s12859-019-2707-7

Wang, S.-B., Feng, J.-Y., Ren, W.-L., Huang, B., Zhou, L., Wen, Y.-J., . . . Zhang, Y.-M. (2016). Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. *Scientific Reports, 6*(1), 19444. doi:10.1038/srep19444

Wang, Y., Bennewitz, J., & Wellmann, R. (2017a). Novel optimum contribution selection methods accounting for conflicting objectives in breeding programs for livestock breeds with historical migration. *Genetics Selection Evolution, 49*(1), 45. doi:10.1186/s12711-017-0320-7

Wang, Y., Ding, X., Tan, Z., Ning, C., Xing, K., Yang, T., . . . Wang, C. (2017b). Genome-Wide Association Study of Piglet Uniformity and Farrowing Interval. *Frontiers in Genetics, 8*(194). doi:10.3389/fgene.2017.00194

Waples, R. K., Larson, W. A., & Waples, R. S. (2016). Estimating contemporary effective population size in non-model species using linkage disequilibrium across thousands of loci. *Heredity, 117*(4), 233-240. doi:10.1038/hdy.2016.60

Weerasinghe, W. M. S. P., Crook, B. J., Clark, S. A., Moghaddar, N., & Byrne, A. I. (2019). Genome-Wide Association Study of Carcase and Eating Quality Traits in Australian Angus Beef Cattle. *Proceedings of the Association for the Advancement of Animal Breeding and Genetics Conference, 23*, 266 - 269.

Wilson, D. J. (2019). The harmonic mean *p*-value for combining dependent tests. *Proceedings of the National Academy of Sciences, 116*(4), 1195-1200. doi:10.1073/pnas.1814092116

Wray, N. R., & Goddard, M. E. (1994). Increasing long-term response to selection. *Genetics Selection Evolution, 26*(5), 431. doi:10.1186/1297-9686-26-5-431

Wright, S. (1921). Systems of Mating. I. the Biometric Relations between Parent and Offspring. *Genetics, 6*(2), 111-123.

Xia, J., Fan, H., Chang, T., Xu, L., Zhang, W., Song, Y., . . . Gao, H. (2017). Searching for new loci and candidate genes for economically important traits through gene-based association analysis of Simmental cattle. *Scientific Reports, 7*(1), 42048. doi:10.1038/srep42048

Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics, 163*(2), 789-801. doi:10.1093/genetics/163.2.789

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., . . . Visscher, P. M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics, 42*(7), 565-569. doi:10.1038/ng.608

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011). GCTA: a tool for genome-wide complex trait analysis. *American journal of human genetics, 88*(1), 76-82. doi:10.1016/j.ajhg.2010.11.011

Yin, T., & König, S. (2019). Genome-wide associations and detection of potential candidate genes for direct genetic and maternal genetic effects influencing dairy cattle body weight at different ages. *Genetics Selection Evolution, 51*(1), 4. doi:10.1186/s12711-018-0444-4

Zellinger, W., Grubinger, T., Lughofer, E. D., Natschläger, T., & Saminger-Platz, S. (2017). Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. *ArXiv, abs/1702.08811*.

Zeng, J., Toosi, A., Fernando, R. L., Dekkers, J. C. M., & Garrick, D. J. (2013). Genomic selection of purebred animals for crossbred performance in the presence of dominant gene action. *Genetics Selection Evolution, 45*(1), 11. doi:10.1186/1297-9686-45-11

Zeng, P., & Zhou, X. (2017). Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models. *Nature Communications, 8*(1), 456. doi:10.1038/s41467-017-00470-2

Zhang, L., Liu, J., Zhao, F., Ren, H., Xu, L., Lu, J., . . . Du, L. (2013). Genome-Wide Association Studies for Growth and Meat Production Traits in Sheep. *PLOS ONE, 8*(6), e66569. doi:10.1371/journal.pone.0066569

Zhang, Q., Calus, M. P. L., Guldbrandtsen, B., Lund, M. S., & Sahana, G. (2015). Estimation of inbreeding using pedigree, 50k SNP chip genotypes and full sequence data in three cattle breeds. *BMC Genetics, 16*(1), 88. doi:10.1186/s12863-015-0227-7

Zhang, Q., Privé, F., Vilhjálmsson, B., & Speed, D. (2021). Improved genetic prediction of complex traits from individual-level data or summary statistics. *Nature Communications, 12*(1), 4192. doi:10.1038/s41467-021-24485-y

Zhang, Y., Qi, G., Park, J.-H., & Chatterjee, N. (2018). Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nature Genetics, 50*(9), 1318-1326. doi:10.1038/s41588-018-0193-x

Zhu, Z., Bakshi, A., Vinkhuyzen, Anna A. E., Hemani, G., Lee, Sang H., Nolte, Ilja M., . . . Yang, J. (2015). Dominance Genetic Variation Contributes Little to the Missing Heritability for Human Complex Traits. *The American Journal of Human Genetics, 96*(3), 377-385. doi:https://doi.org/10.1016/j.ajhg.2015.01.001

Ziegler, H. (1981). Properties of Digital Smoothing Polynomial (DISPO) Filters. *Applied Spectroscopy, 35*(1), 88-92.