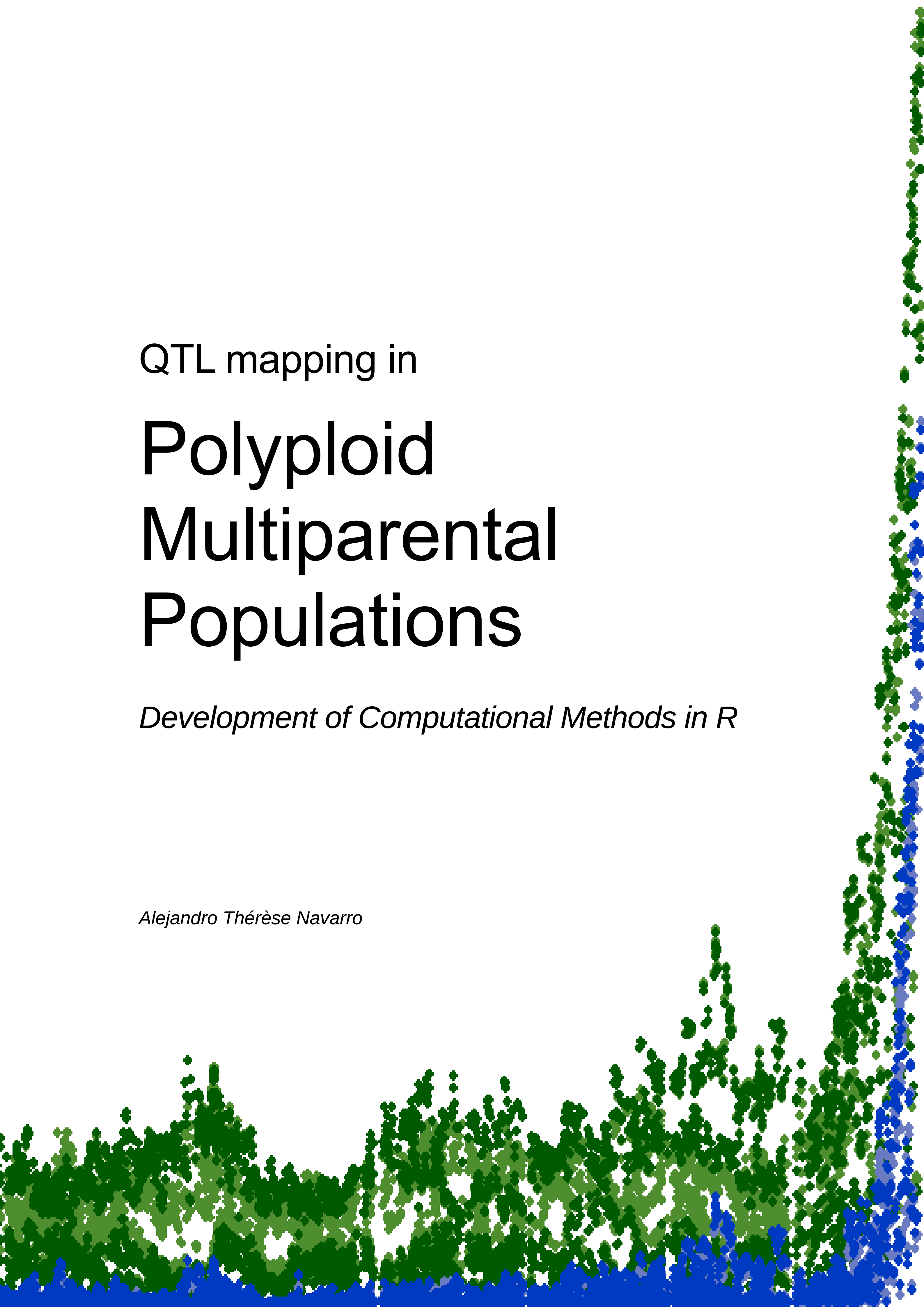
QTL mapping in

Polyploid Multiparental Populations

Development of Computational Methods in R

Alejandro Thérèse Navarro



QTL Mapping in Polyploid Multiparental Populations

Development of computational methods in R

MSc Thesis of

Alejandro Thérèse Navarro

with supervision from

Dr.Ir. Chris Maliepaard & Dr. Giorgio Tumino

April 2018, Wageningen, the Netherlands

This research was conducted as part of the Polyploid Project research group, within the Plant Breeding chair group of Wageningen Research.

ABSTRACT

One of the main challenges of classical breeding in autopolyploids is the complex segregation patterns that polysomic inheritance generates. For that reason, application of novel molecular approaches adapted to the particularities of such genetics has a great potential to improve breeding efforts in these species. Additionally, autopolyploids tend to present high heterozygosity, meaning that multiple alleles will be segregating within one single cross. The use of Multiparental Populations (MPPs) such as Nested Association Mapping (NAM) can help tackle this issue, by using information of multiple crosses in one single analysis to characterize allelic effects. In this thesis, simulated autotetraploid NAM populations with different degrees of genetic distance between the parents have been used to test the efficacy of models based on three types of markers: biallelic SNP markers, multiallelic ancestral markers and multiallelic parental chromosome markers. The best model in all scenarios in terms of detection power was the ancestral model, which used IBD alleles as predictor of the genetic effects. The power of this model was maximized when parental diversity was not extreme, neither too low nor too high, suggesting that for design of MPP populations the best approach in terms of QTL detection power is to sample multiple parents from the same ancestral group.

INDEX

1 INTRODUCTION

Plant breeding has been, and still is, key for agricultural development and for scientific research, allowing to unravel the biological mechanisms ruling traits of interest. Many different methodologies have been developed to tackle plant breeding problems, especially since the ongoing biotechnological revolution that started at the beginning of the last century.

The advent of high-throughput technologies has allowed us to study biological problems in a completely different dimension. This is the case for QTL studies, which, in its early days, relied on molecular markers such as RFLPs and SSRs but have, since then, adapted to take advantage of SNP arrays and sequencing. Thanks to the methodologies developed by a wide and diverse set of experts in statistics, computational sciences, molecular biology and genetics, among others; QTL methods can be used to identify, in many different crops, genomic regions associated with phenotypic traits. For instance, 72 resistance genes against wheat rust were identified by QTL analysis methods (Soriano and Royo, 2015).

Many relevant crops are polyploids, that is, they have more than two sets of their genome (see Table 1 of Sattler *et al.*, 2016 for an extensive list). Some of them present disomic inheritance (they follow Mendelian segregation, such as wheat or cotton), while others do not, generating a series of complications in their genetic analysis that have hindered the application of QTL techniques.

Recent developments in interpretation of SNP dosage data of polyploids (Voorrips *et al.*, 2011; Hackett *et al.*, 2013; Zheng *et al.*, 2016) have paved the way to the modernization of QTL detection and breeding in polyploid crops with the use of high-throughput techniques. In fact, some models for QTL detection in biparental crosses (Xu *et al.*, 2013; Hackett *et al.*, 2014) or GWAS analysis in a diverse line panel (Rosyara *et al.*, 2016) have already been developed for autotetraploids, and even applied successfully (Massa *et al.*, 2015 and Berdugo-Cely *et al.*, 2017, respectively).

Nevertheless, as described by Würschum (2012), mapping based on biparental populations, or in a diverse panel (family mapping vs. association mapping), represent two extremes of a gradient. An intermediate step, that combines the advantages and reduces the drawbacks of both methods, can be found in multi-parental populations (MPP), where a set of diverse related lines are crossed to study their genetic and phenotypic variation in order to perform mapping analysis. However, no methodology specific to polyploids has been developed for such studies. This prevents the utilisation of novel breeding strategies for finding and characterizing new interesting variation.

The objective of this thesis is to expand already-existing QTL analysis methods to multiparental polyploid populations, enlarging the toolbox with which to study the complex genetics that some of these organisms have. To that end, genotypic and phenotypic polyploid data were simulated and tested under six different QTL models, to study the effect of modelling on QTL detection. This toolset and preliminary study pave the way for further investigation of the effects of genetic parameters on polyploid MPPs and the possibilities of QTL modelling in such populations.

1.1 Genetic analysis in polyploids

To perform QTL analysis, advanced genetic knowledge of the species is required. The essential building blocks of these studies include: a linkage map, i.e. a set of molecular markers and their positions in the genome; genotype characterization at each marker for each individual and a set of phenotypic values.

The two first components are not trivially obtained in polyploids, particularly when using the mostly biallelic SNP markers. However, during recent years an increasing number of tools to study polyploids using SNP markers have been made publicly available. Let us discuss the main issues of polyploid genetic analysis as well as introduce some tools regarding these problems.

1.1.1 Meiosis

Polyploidization affects meiotic gamete production, requiring elementary modifications to function in polyploids. This affects expected allelic segregation in polyploids, which is require for multiple statistical methods. Understanding the types and modes of meiosis is essential to adequately understand polyploid populations. For a more detailed review of the processes underpinning polyploid meiosis refer to Zielinski and Mittelsten Scheid (2012).

Meiosis initiates when homologous chromosomes are paired, followed by recombination between non-sister chromatids due to programmed double-strand DNA breaks. In diploids, finding homologous chromosomes is a rather straightforward process: only one other chromosome is homologous. In contrast, polyploid recombination may occur between different pairs of chromosomes, requiring a different estimator for the recombination frequency.

Meiotic behaviour will vary depending on the origin of the polyploid. We recognise two broad categories: **allopolyploids** originated by hybridisation between closely related species coupled with genome duplication (i.e. at least two copies of similar chromosomes originating from different species, homoeologous chromosomes); and **autopolyploids,** caused by a whole-genome duplication (i.e. multiple copies of the same chromosome, homologous chromosomes). It is generally assumed that allopolyploids have strong preferential pairing, resulting in disomic inheritance (the case of wheat, cotton or straweberry). Autopolyploids are more complex. They are expected to form random (or slightly preferential) bivalents and to a lesser extent, multivalents (pairings of more than two chromosomes). In these cases, we expect polysomic inheritance, which breaks Mendelian segregation proportions (for instance in potato, chrysanthemum or leek). Importantly, different levels of multivalent formation can be observed depending on the degree of homology between individual chromosomes and even chromosome segments (Sybenga, 1996).

These assumptions regarding autopolyploids and allopolyploids were for long given for granted, however, there is not a clear association of polysomic and disomic inheritance respectively: newly formed autopolyploids present higher bivalent formation than expected, and allopolyploids may form multivalents to some extent (Ramsey and Schemske, 2002).

1.1.2 SNP arrays and genotyping in polyploids

Obtaining the marker genotypes from SNP arrays in polyploids is more challenging than one might expect. Initial techniques to genotype diploid SNPs were based on absence/presence of the alleles, which allowed to characterize homozygous AA (only A present), heterozygous AB (both A and B present) or homozygous BB (only B present). However, the same system applied to polyploids does not allow to differentiate heterozygotes (AAAB, AABB or ABBB).

Dosage information can be used instead, presented as a fluorescence intensity ratio of both alleles, generating expected values to range from 0 (AAAA), 0.25 (AAAB), 0.5 (AABB), 0.75 (ABBB) to 1 (BBBB). However, a continuous gradient of values can be observed, which complicates assigning ratios to specific genotypes. To that end, the R package FitTetra was developed (Voorrips *et al.*, 2011), using a **mixture model** approach to resolve the separation of SNP genotypes. By comparing multiple models under different assumptions, multiple genotype call probabilities are calculated. Finally, only genotypes assigned with a probability exceeding a certain threshold (0.99 by default) are accepted, generating a highly reliable genotype calling.

1.1.3 Phasing and Linkage mapping

Linkage maps, although not essential, are extremely useful in QTL analysis. Obtaining them requires understanding recombination frequencies between markers. In polyploids, the complexity of this problem does not only lie in the increased number of possible marker combinations that need to be taken into account, but the fact that different marker types have different amount of information: some marker segregations are clearly informative of what the linkage was in the parents, while others are ambiguous and are possible in multiple scenarios (Hackett *et al.*, 1998; Luo *et al.*, 2001).

Many methods exist to resolve phasing and linkage mapping in multiallelic markers, and in biallelic markers for some phase configurations (e.g, Hackett *et al.*, 1998); however, only two workflows have been developed that can use SNP dosage data to obtain linkage maps and resolve phasing of all markers (Hackett *et al.*, 2013; Zheng *et al.*, 2016). Hackett's method can estimate the **linkage map** (ordering and distance of markers) based solely on SNP dosage information, although assuming random bivalent pairing (no preferential or multivalent pairing). On the other hand, Zheng's approach depends on a previously defined linkage map, but can reconstruct SNP phasing of each homologue of every individual, obtaining what is known as an **integrated linkage map.** Together, both models can be pipelined to obtain integrated linkage maps only from SNP dosage data.

1.2 Population structure

Population design for mapping genomic regions is as diverse as the purposes they are designed for. Two main approaches encompass most methods: biparental crosses and diversity panels (Würschum, 2012).

Biparental crosses were originally designed to contrast the genes of pure lines: when crossed, a homogeneous F1 is obtained that can be analysed if crossed further. Some cross schemes include: F2 populations; Recombinant Inbred Lines, where an F2 is crossed further with itself to reduce linkage disequilibrium; Backcrosses or Near Isogenic Lines, which allow to introgress genetic regions to a specific genetic background, and more (Sehgal *et al.*, 2016). Biparental crosses, however, have a great drawback: the lack of diversity. Not only will be less alleles characterized, but those regions for which the parents do not differ in their genetic effect, will not be detected in the study; even if they contain relevant genes to understand or improve a certain trait. This still holds even in biparental autopolyploid populations, which are not expected to be fully homozygous.

In the second population type, diversity panels, a sample is taken from a defined population (be it wild accessions, commercial cultivars, samples from different genetic pools…). Typically, to reduce unwanted association between alleles, unrelated individuals are chosen. However, because of the evolutionary forces that drive these populations, complex genetic structures that affect frequency and association of genetic variants is unavoidable, and will cause bias if not taken into account (Yu *et al.*, 2006). Although methodologies for correcting structure are diverse, the principle behind them is common: to identify expected similarities between individuals. An accurate structure correction method will be crucial: the efficacy of the QTL mapping experiment will depend on the correction precision, as well as the frequency of the segregating alleles.

1.2.1 Multiparental Populations

An alternative intermediate between both possibilities are multiparental populations. As in biparental mapping, a controlled offspring is used for QTL study. While there is an increased variation to characterize than in biparental mapping, the final structure and spurious associations are much less complex than in diversity panels. Some multi-parental crossing schemes include Nested Association Mapping (NAM) (McMullen *et al.*, 2009), diallel crosses (Hayman, 1954), factorial designs or Multi-Parent Advanced Generation Inter-Cross populations (MAGIC) (Cavanagh *et al.*, 2008). The choice of crossing scheme will depend on the final objective of the study.

In this thesis, we will explore the possibilities of an adapted scheme of the NAM proposed by McMullen *et al.* (2009), but without the subsequent inbreeding process. In it, a central parent is crossed with multiple peripheral parents, allowing to explore the interaction of the genetic background from the central parent, with each of the peripherals. This crossing scheme is in fact a series of ***related biparental crosses***, with a central parent being the connection between all the crosses.

As in diversity panels, genetic structure must be considered in QTL mapping. Individuals in a NAM population are either half-siblings, only sharing the central parent, or full-siblings, sharing both parents. The correlations between individuals can be broken down in two different components: **recent relatedness**, due to the family pedigree; and **ancestral relatedness**, due to the similarity between peripheral parents. Those crosses where the peripherals are more similar will yield more similar offspring, increasing their level of relatedness.

A key factor in MPPs is the genetic diversity of the population, which is a direct consequence of parent similarity. With higher diversity, linkage disequilibrium between markers will decrease, improving the accuracy of the QTL study as the spurious association between markers diminishes. On the other hand, increasing diversity may lead to a higher **multiallelism**, that is, increased number of expected alleles at any given locus. The effect of increased multiallelism on statistical power is not evident: while it will increase phenotypic segregation, which might heighten the contrast between genotypes; it might generate confounding situations, where combinations of different alleles lead to the same phenotype.

Studying the effect of diversity in the QTL study is one of the key elements of this thesis, we can expect that under different circumstances, and different objectives, the desired level of parental diversity will vary. To characterize this relationship is therefore crucial, particularly in autopolyploid outcrossing organisms where diversity is already expected to be higher than in diploids.

1.3 Statistical framework

Many different approaches have been proposed to study QTL regions, which differ both in underlying concepts and methodology. Ultimately, all of them aim to analyse the association between genomic regions (represented by markers, or using markers as covariates), and phenotype variation. It is not the purpose of this thesis to review all possible approaches to QTL mapping, so for further reference be sure to read the extensive review by Würschum (2012).

In the early genetic linkage studies, which are the first attempts to associate genetic and phenotypic variation, the few molecular markers available were individually tested for association with discrete phenotypes. This was the first example of single marker methods, exemplified in the linkage maps of Thomas Hunt Morgan and his pupil Alfred Sturtevant. Single marker methods have one important assumption: that at least one marker is in very tight linkage disequilibrium with the causative QTL. When marker densities were still too low, interval mapping approaches were developed to overcome this limitation (Lander and Botstein, 1989), and improved to take into account multiple and linked QTLs (Jansen, 1993; Zeng, 1994; Kao *et al.*, 1999).

Interval mapping approaches have several advantages: they have increased accuracy, can estimate QTL interval position and drop the marker linkage assumption. However, with the advent of omics technologies, high density marker maps are available, making the initial assumption of single marker methods not unrealistic anymore. Moreover, with a threshold definition, single marker methods can also establish a QTL location. The increase on detection power of interval methods might be less relevant nowadays than during their development almost 30 years ago and come at the expense of a higher mathematical complexity. For these reasons, single marker methods have been considered more appropriate for this thesis.

1.3.1 Mixed model for QTL analysis

A strong method to correct for correlation between observations due to genetic structure, is the **mixed model**. Yu *et al.* (2006) defined a “unified mixed model”, also known as the $Q + K$ model (Rosyara *et al.*, 2016), that includes both a kinship matrix (K) and a population structure matrix (Q):

$$y = X\boldsymbol{\alpha} + Q\boldsymbol{v} + \underline{Z\boldsymbol{u}} + \underline{\boldsymbol{\varepsilon}} \qquad Var(\boldsymbol{u}) = K\sigma_G^2 \quad Var(\boldsymbol{\varepsilon}) = R\sigma_\varepsilon^2 \tag*{[1]}$$

Where $X\boldsymbol{\alpha}$ represents the marker effects (SNP effect in Yu *et al.* (2006)) and incidence matrix; $Q\boldsymbol{v}$ are the population structure matrix and vector, respectively; $\underline{Z\boldsymbol{u}}$ are incidence matrix and vector of genetic background effects (polygene effects in Yu *et al.* (2006)); and $\underline{\boldsymbol{\varepsilon}}$ is the residuals vector. The variances of the random effects, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are also defined: where K is the kinship matrix and σ_G^2, the genetic variance; R is a matrix with off-diagonal numbers being 0, and the diagonal is the reciprocal of the number of observations underlying each genotype estimation, and σ_ε^2 is the residual variance.

Kinship matrix calculation

Typically, pedigree-based methods were used for genetic relatedness calculation, but marker-based distances have gained popularity nowadays, especially when pedigree information is incomplete or absent. Marker-based methods are not new (Morton *et al.*, 1971), and there exist multiple ways of calculating them (Lynch and Walsh, 1998; Taylor, 2015), including for polyploid species (Pembleton *et al.*, 2013; Huang *et al.*, 2014; Rosyara *et al.*, 2016). However, no measure is best in all situations, they depend on factors such as level of relatedness, number of alleles considered or population structure. As it remains unclear which would suit best the purpose of the K matrix estimation in this experiment, various methods have been applied and compared.

1.3.2 Genetic models

A mixed-model is no less than a statistical approach. However, to perform a QTL analysis, a genetic model must be defined to link genotypic effects and phenotypes. Polyploid genetic models have existed for a long time (Kempthorne, 1957), and have inspired more recent versions applied to SNP data (Hackett *et al.*, 2001; Luo *et al.*, 2005). Similarly, this thesis has been influenced by two pieces of work: the *autopolyploid GWAS* approach applied in the GWASpoly R package (Rosyara *et al.*, 2016), and an adaption of Garin *et al.* (2017) models to *pure line diploid* MPPs.

It is important to realize that, unlike in diploid models in which a heterozygote might only present one dominance or interaction effect, in polyploids the landscape is more complex, and up to tetra-allelic interactions can be defined. Rosyara *et al.* (2016) tackled this issue and implemented it in their GWASpoly R package, not for all possible scenarios, but for a few most representative ones. The simplest of them is that of total additivity, where the genetic effect granted from one specific allele is equivalent to the dosage times the genetic effect, $\delta_j \alpha_j$. For the sake of simplicity, this model variation was the only one taken into account.

In an MPP population, genetic diversity will affect the number of expected alleles, which we can reflect in the number of modelled genetic effects. In that sense, Garin *et al.* (2017) used three types of parametrizations for diploid pure lines, which were based in multiple models found in literature. In our simulation, we will find a way to generate a design matrix(X in Yu's model) for each situation.

Biallelic model

This model, equivalent to model B in Würschum (2012) and the association mapping model in Liu *et al.* (2012), considers the SNP state as equal to the QTL allele, with only two genetic effects defined. We can understand it as an **IBS** model, or a linear regression of the dosage of one of the alleles.

When the QTL alleles and the SNP alleles are well associated, one can expect the biallelic model to be the most effective. Furthermore, its simplicity grants it a higher statistical power due to a reduced number of parameters, increasing the accuracy of the QTL analysis.

Ancestral model

A more realistic scenario would be that which uses **IBD** information, rather than IBS. We can expect that in NAMs with low parental diversity, there will be a low number of alleles segregating, and as the diversity increases, more alleles come into play. Then, IBD estimation can give us an idea of the allelic effects to be modelled in that particular locus. This model corresponds to linkage disequilibrium and linkage analysis (LDLA) models in Bardol *et al.* (2013) and Giraud *et al.* (2014).

A possibility to estimate IBD alleles is the concatenation of consecutive SNPs, generating a set of haplotypes. In polyploids, finding the haplotypes from sheer SNP data is quite complex, but if adequate estimations would be achieved, the number of haplotypes could serve to estimate the number of alleles present in the NAM population at hand.

Parental model

Lastly, an extreme scenario of allelic diversity, would be that in which each **parental chromosome** is contributing a different allele with a different effect to the NAM population, equivalent to the connected model in Blanc *et al.* (2006). Even in an extremely diverse panel, such a situation is highly unlikely. Since we can expect different alleles to have the same genetic effect, it is an unrealistic assumption that the number of parental chromosomes will be equal to the number of alleles. This being the least parsimonious model, it is also the most inefficient in terms of statistical power usage.

1.4 Objectives and Research Questions

The study of polysomic inheritance (typically autopolyploid inheritance) was conceptualized long ago (Kempthorne, 1957), but it remained theoretical until recent times due to a lack of scientific methods to tackle their many challenges. However, new developments in molecular and computational tools have allowed the introduction of autopolyploid organisms into the modern breeding strategies. Additionally, new and more complex population schemes have been designed, such as MPPs, which require more sophisticated statistical methods to be analysed. A software has been recently developed which allow GWAS analyses in autopolyploid crops, using biallelic markers (Rosyara *et al.*, 2016). However, the use of multiallelic markers for QTL studies in populations with complex structure is still very limited and dedicated software is lacking.

The objectives of this thesis are: i) to develop a statistical tool for QTL analyses in multiparental populations or diversity panels of polyploid species, allowing for the use of either biallelic or multiallelic markers; ii) to explore its performances on simulated data under different scenarios, varying the main factors affecting power of QTL detection.

To that end, a three-step process has been followed:

1. First, we simulated a set of polyploid NAM populations with different **levels of genetic diversity** and a set of phenotypic data, that includes the possibility of **multiallelism**.
2. Secondly, various models of QTL analysis specific to polyploid MPP populations have been developed based on the general framework provided by Yu's **unified mixed model**. Particularly, we adapted genetic models developed for **diploid MPPs** in order to be extended to polyploid genetics. From a statistical point of view, the main concern is the increased number of parameters (allele effects) needed for polyploids, which might result in overfitting the data and lowering the detection power.
3. QTL analyses under different scenarios have been carried out using all the proposed QTL models, in order to i) *characterize the effect of factors such as genetic diversity or heritability* and ii) to *understand the strengths and weaknesses of the QTL models*.

Particularly interesting it would be to compare performance of biallelic and multiallelic markers. The use of multiallelic markers adds an extra layer of analysis, as usually they are obtained by further processing of SNP data. Theoretically, it could be possible to estimate multiallelic markers with increased linkage between marker alleles and causal alleles, therefore increasing QTL detection power. However, the higher number of model parameters might reduce detection power of the QTL models. Importantly, the higher the diversity the higher the number of alleles segregating per locus in the population, and therefore the more number of multiallelic alleles we could expect. The trade-off between number of genetic effects and marker-causal allele increase is what will determine the effectiveness of multiallelic markers in our study.

2 MATERIALS & METHODS

A standard method to validate statistical models is data simulation with controlled parameters. Coupled with real-data evaluation, it offers a complete overview of statistical methodologies. Due to the lack of real data, only simulated data was used in this thesis. Before explaining the methodology, it is relevant to introduce some tools used and assumptions for different processes.

2.1 Genotype Simulation

A polyploid population simulator has been developed that allows to generate individuals from a specified set of parents and pedigrees: PedigreeSim V2.0 (Voorrips and Maliepaard, 2012). Let us give an overview of this program (for more detailed information be sure to consult the PedigreeSim Manual (Voorrips, 2014)).

Six files are needed to perform a simulation:

- `.par`: the **parameters** file, which contains option specifications to run the program.
- `.ped`: the **pedigree** file, which defines the crossing scheme.
- `.chrom`: the **chromosome** file, containing a list of chromosomes, their length (in cM), centromere positions, and preferential pairing and quadrivalent probabilities.
- `.map`: the **genetic map** file, a table specifying marker name, position and chromosome.
- `.gen`: the **alleles** file, with the alleles that each parental individual has at each marker and each homologue.

Using the information contained in these files, a meiotic simulation is performed, obtaining genotypes of the offspring individuals. As output, five different files are generated. Of those five, only two are interesting for our work: **alleledose** and **founderallele** files. The first specifies the dosage of the (alphabetically) highest marker for each individual, while the second specifies the parental allele of each chromosome from each individual, which are equivalent to Identity By State (IBS) and Identity By Descent (IBD) files.

Similar information can be estimated from real data using FitTetra (Voorrips *et al.*, 2011) for IBS and TetraOrigin (for tetraploids only) (Zheng *et al.*, 2016) for IBD. However, the estimations are qualitatively different than the PedigreeSim data. On the one hand, FitTetra output is dependent on probabilistic assignment of genotype, and therefore has a specific error chance. Similarly, the TetraOrigin package returns IBD *probabilities*. In contrast, PedigreeSim files are fixed and certain data, meaning that the results obtained in this thesis will be more accurate than those obtained with real data.

In the parameters file of PedigreeSIM, the following options were chosen in all simulations: HALDANE mapping function, meaning that no chiasma interference was simulated; and NATURALPAIRING=1, meaning that the fraction of quadrivalents arises automatically from the pairing process at the telomeres. The `.map` file was based on the oversaturated genetic map (multiple markers at the same cM position) of *Solanum tuberosum* as published in Bourke *et al.* (2016). Centromeres could not be specified as those defined in the same article, due to the inability of PedigreeSim to handle chromosomic regions as centromeres; they must be specific positions. Instead, the half-length of each chromosome was set as the centromeric position.

2.1.1 Ancestral Populations

Simulating NAM populations is relatively easy with PedigreeSim, only requiring to specify a NAM pedigree in the `.ped` file. However, the program can also be used to simulate the different levels of ancestral relatedness: by generating multiple ancestral groups (AG), from which the parents descend. Between the parents from the same AG, a higher level of relatedness will be present than between parents from different AGs. In a NAM population, when all parents come from different AGs, the total diversity will increase; and if all originate from the same AG, the opposite holds.

In total, 10 ancestral populations were generated, each of them in the same manner (Fig. 1). A set of 10 founders (generation 0, G0) were randomly assigned marker states at each chromosome. That is, `.gen` files were randomly generated, randomly specifying the SNP alleles (A or B) of the founders at each marker.

The subsequent generations were obtained by randomly mating parents from the previous generation, without selfing. Each of these generations (G1 to G49) was formed by 100 individuals.

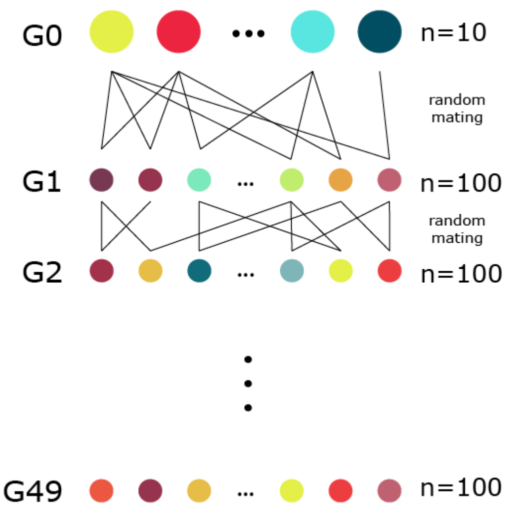


Figure 1: *Pedigree scheme of an ancestral population. The generation G49 is the parental generation, that will be used to simulate NAM crosses in step two of the genotype simulation process.*

2.1.2 NAM crosses

To generate a NAM population, a series of parents must be chosen from a set of AGs. That is not a straightforward process, as this is in fact a partitions problem (how many ways can we sum k with n non-negative numbers). For instance, if a NAM is composed of 9 crosses, and therefore of 10 parents, when choosing from 2 AGs there are 5 different possibilities: 9+1, 8+2, 7+3, 6+4 and 5+5. All samplings represent 2 AGs, but the obtained NAMs are hardly comparable.

If we think about this problem in terms of ancestral representation in the NAM population we see that there is a clear gradient between 1 AGs, and the most balanced situation of 2 AGs. Imagine a cross with 10 parents from 1 AG: 100 offspring belong to that 1 AG. If 2 AGs are present, 9+1 is equivalent to 90 from one group and 10 from the other, 8+2 is 80 and 20, and so on until 50 and 50. The 9+1 situation is relatively close to the single ancestral situation; we can reduce the variability of NAMs by choosing only the most ***balanced*** situations. That is, 5+5 for 2 AGs, 4+3+3 for 3, and so on. There can still be certain variability in that, depending on where does the central parent fall. For instance, when choosing from 2 AGs, in the 9 crosses, one AG will be present in 5 and the other in 4.

In total, 1000 individuals, 100 from each AG, were obtained. These will be referred to as ***parents***. To simplify downstream processes, the name of each parent was coded as *A3P49* to refer to parent 49 from AG 3 (AGs from 0 to 9, and parents from 00 to 99). To study the effect of parental relatedness, different NAM cross structures were defined, with 1 to 10 AGs represented. Choosing the most balanced configuration 100 crosses were simulated for each number of AGs from 1 to 10, which we will denominate as NAM1 to NAM10. Each consisted in 9 crosses from 10 parents, each cross with 50 individuals as offspring, summing to a total of 460 individuals per NAM population, including the parents.

2.2 Phenotype Simulation

Simulating phenotypes can be quite challenging, particularly because the assumptions underlying real phenotypes vary depending on the species, populations and traits. Such can be seen in Garin *et al.* (2017): multiple QTL models (with different assumptions) are developed and applied to different maize traits in the same NAM population. No model performs best in all scenarios; the underlying explanation is that due to different assumptions holding in each trait, a different model is best.

Because of this inherent limitation in phenotype simulation, we will choose assumptions based on what we consider relevant examples. We will imagine that *the purpose of this NAM is to cross a central elite cultivar, with high effect QTLs, with other peripheral parents which might have other QTLs in other regions of the genome*, *or different lower-effect alleles in the same QTL.*

Phenotypes were simulated using the **ancestral alleles**, which, to our understanding, is the most realistic way to simulate the relationship between SNPs states and QTL alleles (or the relationship between IBS and IBD). The ancestral alleles are defined as the IBD alleles provided by each founder of the AGs, which can be found in the `founderallele.dat` file from the AGs simulations.

To identify the ancestral alleles from each of the different AGs, the following labelling was followed: for A0, the alleles spanned from 0 to 39, for A2, 40 to 79 and so on until allele 400 from A9. To summarize this, a single file containing all chromosomes of all parents was generated, where the ancestral alleles were coded as described. This “Total Population” file is crucial in the phenotype simulations, as it allows to translate parental alleles from a NAM population into the corresponding ancestral alleles.

2.2.1 Single QTL simulation

Genetic Model

The first step in phenotype simulation is to define a relationship between genotype and phenotype, to be able to generate phenotypic values. Let us consider the standard genetic model:

$$y_i = G_i + E_i \qquad E_i \sim N(0, \sigma_E^2) \qquad [2]$$

Where y_i is the phenotype of individual i, with a genetic effect G_i and an environmental effect E_i, which is normally distributed with a variance of σ_E^2 and an expectation of 0. If we would assume the genetic effect to be defined additively by one single QTL, we could write this model as:

$$y_i = \sum_{j=1}^{k} \delta_{ij}\alpha_j + E_i \qquad [3]$$

Where δ_{ij} is the dosage of the j^{th} QTL allele, and α_j is the genetic effect of such allele. In this scenario, the broad-sense heritability H^2 is equal to the narrow sense heritability h^2, as there are no dominance or epistatic effects. Our interest in simulating phenotypes is to be able to control the heritability of the simulated trait, given a specified environmental variance σ_E^2 and a set of genetic effects α_j. The genetic component of each individual, and the mean genetic component are:

$$G_i = \sum_{j=1}^{k} \delta_{ij}\alpha_j \quad \text{and} \quad \bar{G} = \sum_{j=1}^{k} \bar{\delta}_j\alpha_j \qquad [4]$$

We can express the genetic variance σ_G^2, which in our case is equal to the additive variance σ_A^2, as a function of the dosages and allelic effects:

$$G_i - \bar{G} = \sum\nolimits_{j=1}^{k} \delta_{ij}\alpha_j - \sum\nolimits_{j=1}^{k} \bar{\delta}_j\alpha_j = \sum\nolimits_{j=1}^{k} \alpha_j(\delta_{ij} - \bar{\delta}_j)$$

$$\sigma_A^2 = \frac{\sum_{i=1}^{n}(G_i - \bar{G})^2}{n-1} = \frac{\sum_{i=1}^{n}\left[\sum_{j=1}^{k}\alpha_j\left(\delta_{ij} - \bar{\delta}_j\right)\right]^2}{n-1} \qquad [5]$$

Heritability control

As QTL analysis is largely influenced by the heritability, we are interested in controlling it. To that end, we can rescale the size of the genetic variance, multiplying it by a constant such that:

$$h^{2*} = \frac{R\sigma_A^2}{R\sigma_A^2 + \sigma_E^2} \qquad [6]$$

Where h^{2*} is the controlled heritability, and R is a rescaling constant. Our only method to control the size of the genetic variance before the phenotype simulation, is to apply the rescaling to the genetic effects. If we define the controlled genetic effects as $\alpha_j^* = c\alpha_j$ we can show:

$$\text{if} \quad \alpha_j^* = c\alpha_j \quad \text{then} \quad \sigma_A^{2*} = \frac{c^2\sum_{i=1}^{n}\left[\sum_{j=1}^{k}\alpha_j\left(\delta_{ij} - \bar{\delta}_j\right)\right]^2}{n-1} \quad \text{and} \quad R = c^2 \qquad [7]$$

Although the defined α_j will no longer be the original genetic effects, the proportionality between them will still hold. If we isolate this constant c, we can show that:

$$c = \sqrt{\frac{h^{2*}\sigma_E^2(n-1)}{(1-h^{2*})\sum_{i=1}^{n}\left[\sum_{j=1}^{k}\alpha_j\left(\delta_{ij} - \bar{\delta}_j\right)\right]^2}} = \sqrt{\frac{h^{2*}\sigma_E^2}{\left(1-h^{2^*}\right)\sigma_A^2}} \qquad [8]$$

Therefore, applying a rescaling constant c to all genetic effects, can allow us to simulate traits with specific heritabilities.

Genetic effects

Typically, genetic effects for phenotype simulations are obtained by sampling left-skewed distributions, with limit at 0, where genetic effects around and average are most probable, and high genetic effects are increasingly improbable. For instance, a gamma distribution with $\alpha = 2$ and $\lambda = 1$, where the average is α/λ. However, with such small number of genetic effects to sample, the simulations become too variable. For that reason, a fixed formula was used to determine them.

The genetic effects were defined based on the number of alleles present at a given locus in each population. Because different numbers of alleles can be expected between populations, a formula was derived to find equivalent **genetic configurations** between populations:

$$\alpha_j = \frac{(k_l - j - 1)^b}{\max((k_l - j - 1)^b)} \qquad [9]$$

Where j is the allele index, k_l is the number of alleles at that locus, and b is a number obtained empirically. This formula has the property that if the same b is applied, similarly structured genetic effects are obtained for different k_l. To vary between a multiallelic or a biallelic situation, it is enough to define multiple b's that increase in exponential steps. When $b = 1$, the allelic effects decrease linearly as the allele index increases. If b is a very large number, only the first allelic effect will not be 0, generating a biallelic situation (see Figure 3b of section 3.2.1).

Multiallelism Experiment

A first experiment was conducted to study the effect of different allele numbers on detection power of our QTL models. To reduce the number of variables affecting the power of the analysis, single QTL traits were simulated. The following parameters were used:

- Heritabilities of 0.2, 0.5 and 0.8.
- Four genetic configurations with $b = 3.3^0, 3.3^1, 3.3^2, 3.3^3$
- Three QTL positions at a chromosome extreme, PotVar0120075; at a high marker density region, solcap_snp_c2_53034 and at a low marker density region, PotVar0100045.

In total 36 phenotypes per NAM1 population were simulated.

2.2.2 Multiple QTL simulation

Genetic model

To control the total heritability of the trait, the same procedure can be used as in section 2.2.1, if we expand the model to a multiQTL situation. Let us define a series of locus, $l = 1 \ldots L$, and a number of alleles per locus k_l. In an additive situation, we can define the phenotype as:

$$y_i = \sum_{l=1}^{L} \left(\sum\nolimits_{j=1}^{k_l} \delta_{ijl} * \alpha_{jl} \right) + E_i \qquad [10]$$

Where δ_{ijl} is the dosage of individual i, of allele j from locus l; and α_{jl} is the genetic effect of allele j from locus l. Similarly as before, if we define all genetic effects from all loci as $\alpha_{jl}^* = c\alpha_{jl}$, where c is the same constant for all effects, it can be shown that the same equation holds:

$$c = \sqrt{\frac{h^{2*}\sigma_E^2}{\left(1 - h^{2^*}\right)\sigma_A^2}}$$

An alternative question is whether we can control the **partial heritabilities** of each QTL independently, to simulate different-size QTL scenarios. Let us rename the genetic effect of each individual at each locus such that $\Delta_l = \sum_{j=1}^{k_l} \delta_{ijl} * \alpha_{jl}$. The genetic variance of a NAM3 population with 3 QTLs segregating can be written as:

$$\begin{aligned}\sigma_G^2 = var(\Delta_1 + \Delta_2 + \Delta_3) &= var(\Delta_1) + var(\Delta_2) + var(\Delta_3) \\ &+ 2cov(\Delta_1, \Delta_2) + 2cov(\Delta_1, \Delta_3) + 2cov(\Delta_2, \Delta_3)\end{aligned} \qquad [11.1]$$

However, we can also express σ_G^2 as the sum of a variance matrix Q**:**

$$Q = \begin{bmatrix} var(\Delta_1) & cov(\Delta_1, \Delta_2) & cov(\Delta_1, \Delta_3) \\ cov(\Delta_1, \Delta_2) & var(\Delta_2) & cov(\Delta_2, \Delta_3) \\ cov(\Delta_1, \Delta_3) & cov(\Delta_2, \Delta_3) & var(\Delta_3) \end{bmatrix} \qquad [11.2]$$

To define the partial heritabilities of each QTL, we might be tempted to use an adaption of the classical formula $h_l^2 = var(\Delta_l)/\sigma_P^2$, which is problematic. With this definition, the sum of all partial heritabilities is not equal to the total heritabilities, as the covariances between loci are not taken into account. It is more appropriate to define the partial heritability of a QTL, for instance QTL 1, as:

$$h_1^2 = \frac{var(\Delta_1) + \sum\left(cov(\Delta_1, \Delta_{l'})\right)}{\sigma_P^2} \qquad l' \neq 1 \tag{12}$$

Using this formulation, the sum of partial heritabilities is equal to the total trait heritability, and the term $\sum\left(cov(\Delta_1, \Delta_{l'})\right)$ acts as the structure-bias for QTL heritability. In biparental crosses where the two parents contribute a different QTL each, the covariance between the genetic effects can be expected to be 0, showing how under a random segregation of two unlinked QTLs this definition of partial heritability is equal to the traditional formula: $h_l^2 = var(\Delta_l)/\sigma_P^2$.

Partial heritability control

Following the reasoning used in section 2.2.1, we can aim to find a series of rescaling factors, such that $\alpha_{jl}^* = \beta_l \alpha_{jl}$, where β_l is the rescaling factor common to all allelic effects from locus l. Through this rescaling, we aim to modify the variance matrix Q such that the true heritability of each locus explains a specific proportion of the total heritability. The rescaled variance matrix Q^*, in an example with three QTLs, can be shown to be equal to:

$$Q^* = \boldsymbol{\beta^T} Q \boldsymbol{\beta}$$

$$Q^* = \begin{bmatrix} \beta_1^2 var(\Delta_1) & \beta_1\beta_2 cov(\Delta_1, \Delta_2) & \beta_1\beta_3 cov(\Delta_1, \Delta_3) \\ \beta_1\beta_2 cov(\Delta_1, \Delta_2) & \beta_2^2 var(\Delta_2) & \beta_2\beta_3 cov(\Delta_2, \Delta_3) \\ \beta_1\beta_3 cov(\Delta_1, \Delta_3) & \beta_2\beta_3 cov(\Delta_2, \Delta_3) & \beta_3^2 var(\Delta_3) \end{bmatrix} \tag{13}$$

Where $\boldsymbol{\beta}$ is the vector of β_l's. If we define a series of controlled partial heritabilities, h_l^{2*}, we can define the following system of non-linear equations:

$$0 = \frac{\beta_1^2 var(\Delta_1) + \sum\left(\beta_1\beta_{l'} cov(\Delta_1, \Delta_{l'})\right)}{\sigma_P^2} - h_1^{2*} \qquad 0 = \frac{\beta_2^2 var(\Delta_2) + \sum\left(\beta_2\beta_{l'} cov(\Delta_2, \Delta_{l'})\right)}{\sigma_P^2} - h_2^{2*}$$
$$0 = \frac{\beta_3^2 var(\Delta_3) + \sum\left(\beta_3\beta_{l'} cov(\Delta_3, \Delta_{l'})\right)}{\sigma_P^2} - h_3^{2*} \tag{14}$$

In this system, all variables except the different β_l's are known, and we can expect it to have a solution. However, the methods to solve this kind of non-linear equation systems escape the scope of this thesis.

An alternative is to find a numerical solution. Defining the partial heritabilities as functions of all betas, $f(\beta)$, adding up the square of each function we obtain $\sum_{l=1}^{L} f_l(\beta_l)^2$. Optimizing it to a minimum, which should be 0, would yield a possible solution if there was any. This method was tested using the `optim()` function of R, but as it remained inaccurate it was discarded as part of the experiment and the partial heritabilitites were left as non-controlled variables.

Genetic effects

In the multiple-QTL simulations, only one allele per QTL was used, as suggested by the results of the Multiallelism experiment and each AG was considered to contribute one QTL locus; all QTLs unlinked. To exemplify, in a NAM3 population 3 QTLs were simulated; at locus 1 only one allele from the first AG is active, at locus 2 one allele from the second AG, and at locus 3 one allele from the last AG. The size of the genetic effects, previous to rescaling, was defined as:

$$\alpha_l = L - (l - 1) \qquad [15]$$

Such that if $L = 3$, the active allele at locus 1 would have effect 3, the active allele at locus 2, effect 2 and the active allele at locus 3, effect 1.

MultiQTL experiment

Using the methods defined in section 2.2.2, we simulated three phenotypes on NAM3, NAM7 and NAM10 populations; with heritabilities of 0.2, 0.5 and 0.8. Due to computational constraints, only 25 populations of each NAM were used, with QTLs simulated at the positions specified in table 1. The unlinked QTL locations were randomly selected, yielding the locations specified in Table 1

Table 1: *QTL positions used for multi-QTL experiment.*

	marker	*chromosome*
QTL1	PotVar0052284	12
QTL2	PotVar0079428	5
QTL3	solcap_snp_c2_32381	2
QTL4	PotVar0044823	1
QTL5	PotVar0129023	10
QTL6	PotVar0039687	6
QTL7	solcap_snp_c2_42306	3
QTL8	PotVar0130503	11
QTL9	solcap_snp_c2_3962	9
QTL10	solcap_snp_c2_2746	8

2.3 QTL Model

Adaptions of the model proposed in Yu *et al.* (2006) were used. To test the relevance of the different correction terms, models were tested with and without the Q structure term, under linear models and under mixed models using a variance structure as defined by the kinship matrix K. Our maximal model, with both Q and K can be expressed as:

$$y = X\beta + Q\boldsymbol{v} + \underline{Z\boldsymbol{u}} + \underline{\boldsymbol{\varepsilon}} \qquad var(u) = K\sigma_G^2 \quad var(\varepsilon) = I\sigma_\varepsilon^2$$

In this notation, underlined components are considered random, while the rest are fixed. $X\beta$ represents the $n \times b$ incidence matrix, and the $1 \times b$ allelic effects vector, respectively. The number of genetic effects modelled, b, will depend on the allele parametrization, as described below. $Q\boldsymbol{v}$ defines the population incidence matrix and population effect vector, respectively. $\underline{Z\boldsymbol{u}}$ are $n \times n$ incidence matrix and $1 \times n$ vector of polygenic effects; and $\underline{\boldsymbol{\varepsilon}}$ is the $1 \times n$ residuals vector. The variances of the random effects, $\boldsymbol{u}$ and $\boldsymbol{\varepsilon}$ are also defined: where K is the kinship matrix and σ_G^2 the genetic variance; I is an identity matrix of $n \times n$ and σ_ε^2 is the residual variance.

2.3.1 Fixed term: allele parametrization

Biallelic model

In a biallelic model, the SNP dosages are used to predict genetic effects, giving the $X\beta$ term the following form:

$$X_b\beta = \begin{bmatrix} 1 & \delta_1 \\ 1 & \delta_2 \\ \vdots & \vdots \\ 1 & \delta_n \end{bmatrix} \begin{bmatrix} \mu \\ \beta \end{bmatrix} \qquad [16]$$

Where δ_i are the dosages of one of the SNP alleles, μ is the intercept and β the genetic effect of that SNP allele. We denote the design matrix as X_b for this modelling strategy.

To produce the X_b matrix, the `genotypes.dat` output file of PedigreeSim was used, as it defines the dosage of the alphabetically highest allele (A in our case).

Ancestral model

Under the ancestral model, the dosage of each ancestral allele in the NAM population is used to estimate genetic effects. The shape of the $X\beta$ term takes then the form:

$$X_a\beta = \begin{bmatrix} 1 & \delta_{11} & \delta_{12} & & \delta_{1k} \\ 1 & \delta_{21} & \delta_{22} & & \delta_{2k} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \delta_{n1} & \delta_{n2} & & \delta_{nk} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \qquad [17]$$

In this case, the dosages of all alleles *except one* are specified. Therefore, k is the number of alleles -1. Each β represents the additive genetic effect of each ancestral allele, expressed as a difference from the effect of the reference ancestral allele, represented by μ.

In our dataset, the *real* **ancestral IBD** information can be obtained if the `founderallele.dat` output file is joined with the ancestral allele table mentioned at the beginning of section 2.2.

Parental model

The $\boldsymbol{X}\beta$ term would be similar to the ancestral model term, but in this case the number of genetic effects would be the number of parental chromosomes -1. Having 10 parents per NAM population, this would mean:

$$\boldsymbol{X_p} = \begin{bmatrix} 1 & \delta_{1,01} & \delta_{1,02} & & \delta_{1,39} \\ 1 & \delta_{2,01} & \delta_{2,02} & & \delta_{2,39} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & \delta_{n,01} & \delta_{n,02} & & \delta_{n,39} \end{bmatrix} \begin{bmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{39} \end{bmatrix} \qquad [18]$$

The information transmitted by this design matrix, is the number of copies of each parental chromosome that each individual has. This is equivalent to **parental IBD**, which in our dataset can be obtained from the `founderallele.dat` output file. An estimation of such parental IBD can be obtained from SNP data with the TetraOrigin package from Zheng *et al.* (2016).

2.3.2 Random term: genetic distances

The polygenic term, the random part of the model, $Z\boldsymbol{u}$, corrects for spurious associations between other markers at the rest of the genome and the marker under test. In case we would have multiple observations from the same genotype, the Z matrix would assign the same random effect $\boldsymbol{u_i}$ to these individuals. With the NAM structure that we followed, there are no such individuals and therefore the Z design matrix is equal to an identity matrix of size $n \times n$, n being the number of individuals and of random effects.

Additionally, a variance structure must be estimated using some genetic distance measures. As suggested by Taylor (2015), multiple distances were used and tested before choosing the most adequate measure to define the K variance structure matrix.

Euclidean distance

A simple, non-genetic method of distance calculation, is the basic Euclidean distance. Using dosage scores as a vector, the distance between the vector of each individual can be easily calculated. Unlike genetic distances, Euclidean distances do not have a specific scale, and are therefore not comparable between populations.

Realized Relationship

A standard method in breeding values computation was developed by VanRaden (2008), and is applied also in the GWASpoly package (Rosyara *et al.*, 2016), named as *realized relationship matrix.* Its formula is quite simple:

$$K = MM^T \qquad [19]$$

Where M is the marker dosage matrix, each marker being centred by subtracting its average dosage. Moreover, the matrix was rescaled so that the average distance of each individual with itself was 1, by dividing the K matrix by the average of its diagonal. This measure has the property that it has no limits at 0 or at 1, but the measures of an individual with itself are a normal distribution around 1, while unrelated individuals present a normal distribution around 0.

Maximum likelihood estimator

The software PolyRelatedness (Huang *et al.*, 2014), implements a maximum-likelihood (ML) estimator for relatedness between individuals, from 0 to 1, that supports multiple levels of ploidy. The estimator is based on probability of sharing certain number of alleles, given a set of allelic frequencies in the population. To use the software, allelic frequencies must be specified. We used the ancestral allelic frequencies to that end, as they represent the original allelic frequencies our NAM populations were derived from.

StAMPP method

In the R package StAMPP (Pembleton *et al.*, 2013), multiple genetic distance measures are calculated. Between them, there is a genomic relationship matrix calculation, based on a diploid method (Yang *et al.*, 2010); although it is extendable to polyploids by adapting dosages to values between 0 and 2. Unlike traditional relatedness measures that range from 0 (totally unrelated) to 1 (identical individuals), this genomic relationship score is set such that 0 is the average relatedness for a given population, and 1 is the value for an individual with itself, if consanguinity is 0.

2.3.3 Implementation

In total, twelve types of models have been tested, with combinations of Q and K corrections; and with each of the three fixed term parametrizations. To obtain a solution of the models, that is: p-values and estimates of allelic effects; two different approaches were used.

For linear models, the standard Least Squares Estimation was used, as implemented in the `lm()` function of R (Wilkinson and Rogers, 1973; Cleveland *et al.*, 1991). The p-values were computed as an F-change test between a model with or without marker fixed effects.

Mixed models

Solving mixed models required a more complex approach, as most mixed model solvers implemented in R are not designed to deal with complex variance structures. To that end, the ***ridge regression*** method was used, which in this context is equivalent to Best Unbiased Linear Predictor (BLUP) (Whittaker *et al.*, 2000; Meuwissen *et al.*, 2001). The R package rrBLUP (Endelman, 2011) has been developed to perform ridge regression on genomic data, and is the core statistical package for the GWASpoly R package (Rosyara *et al.*, 2016). Although rrBLUP itself was not used, the solution algorithm was based completely on the `mixed.solve()` function present in that package. P-values were obtained as an F-test of the fixed effects, as they are calculated in the rrBLUP package.

An important issue when estimating the variance components in QTL analysis, is the iteration of large matrix multiplications and inversions, which can increase considerably computation time. To reduce the computational burden, variance components can be approximated only once, and recycled at each marker position; which is known as the EMMAX or P3D approach (Kang *et al.*, 2010 and Zhang *et al.*, 2010 respectively). This algorithm was applied successfully, calculating the variance components in a model without allelic effects, and applying those variance components to the estimation of fixed effects and p-values at each marker, with each possible model.

P-value correction for multiple testing

Li and Ji (2005) defined a method to define a corrected threshold for multilocus tests that takes into account test correlation due to marker linkage. This method was applied to correct for type I error both in the linear and mixed model scenarios.

3 Results & Discussion

During this thesis research, a single outline was followed, under different scenarios. A set of NAM populations with multiple diversity levels was generated. Based on these genotypes, a set of phenotypes that followed specific conditions was obtained. Lastly, multiple QTL models were used to analyse the association between genotypes and phenotypes, characterizing the models and offering insights into the different scenarios that had been simulated.

3.1 Genetic Distances

Euclidean distances, Relaized Rleationship matrix (Rosyara *et al.*, 2016), PolyRelatedness ML estimator (Huang *et al.*, 2014) and the StAMPP genetic matrix (Pembleton *et al.*, 2013) were used to evaluate distances between parents from different AGs, as well as between individuals from NAM populations. All measures were able to distinguish parents from the same or different ancestral groups, as well as individuals from the same cross (full siblings), from related crosses (related half-siblings) and unrelated crosses (unrelated half-siblings) (Fig. 2). However, calculation times between these methods varied significantly. The ML estimator of PolyRelatedness was extremely slow (~20 minutes per population) and cumbersome to implement as it required generation of specific tabular files; the StAMPP method was much faster, but was also somewhat cumbersome, requiring specific objects for each distance calculation. Finally, both Euclidean and the Realized Relationship (RR) matrix were the fastest and easiest to obtain. As the RR matrix is more comparable between populations and is closer to standard genetic distance measures, it was chosen as the distance measure to build the kinship matrix for the subsequent association analyses.

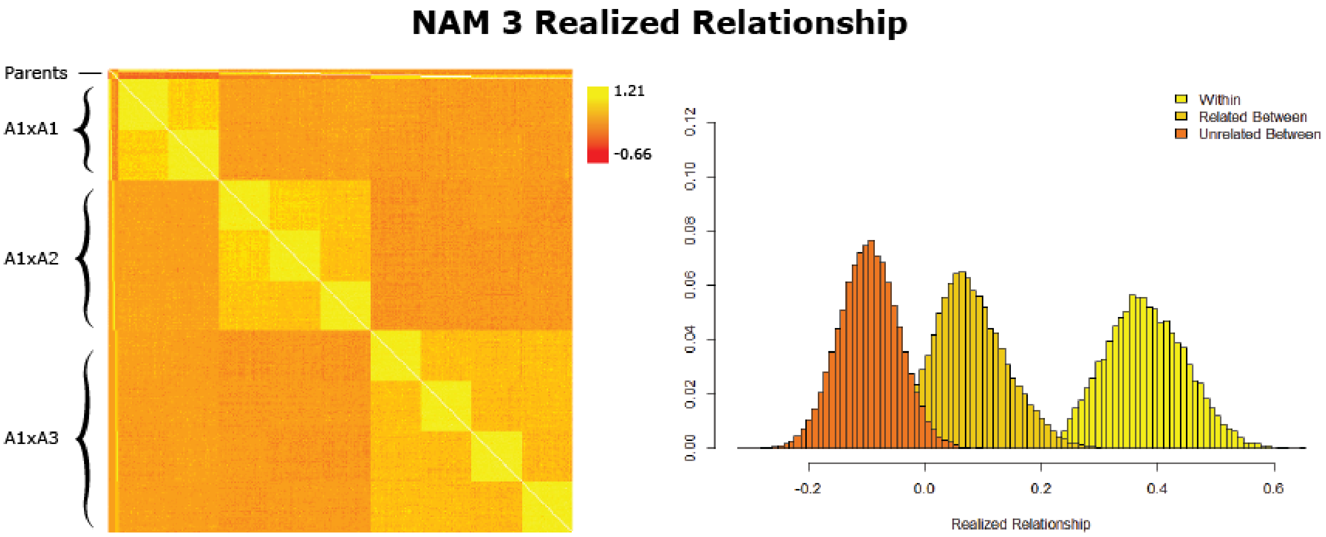


Figure 2. Left: *heatmap representation of the Realized Relationship matrix as described in section 2.3.2. Crosses 1 and 2 belong to the A1xA1 group, where individuals descend from an A1 central parent, and an A1 peripheral parent. Crosses 3, 4 and 5 belong to A1xA2, where the peripheral parent originates from AG 2. Similarly, in crosses 6 to 9 the peripheral parent is from AG3.* ***Right:*** *overlapped frequency distributions of genetic distance types (within crosses, between related crosses and between unrelated crosses). The y axis represents frequencies within each distance type. If a single frequency distribution plot was made, the "within" and "related between" would be much shorter, as most distances are "unrelated between".*

Similar results to those shown in Figure 2 were obtained for other NAM populations. In NAM1 and NAM10 populations, no additional structure was observed besides the within-cross relatedness, as all peripheral parents are equally distant between each other. However, in NAM1 populations peripheral parents are more closely related (all originate from the same AG) than in NAM10 populations (all parents from different AGs), meaning differentiation between crosses due to genetic structure is stronger in NAM10 populations. Other NAM populations displayed more complex structures, as shown in Figure 2 for NAM3, where genetic distances between individuals of related crosses is higher due to the similarity between their parents (they originate from the same AG).

The results shown in Figure 2 confirm that the AG simulation design generated different levels of parental relatedness between NAMs, which was reflected in the genetic distance between the F1s of each cross. We see how "between unrelated" distances and "between related" distances are overlapping (right panel, Fig. 2), while "within" distances are clearly distinguishable from the rest. However, the fact that the distributions are mostly separated from each other suggests that such distance provide a good estimation of the actual relatedness between individuals — there is a clear enough characterization of full siblings, related half-siblings and unrelated half-siblings. Similarly, offspring from a peripheral parent of an AG were more genetically similar to any other parent of the same AG.

3.2 Alleles and Phenotypes

Breeding requires utilizing genetic diversity to discover new genotype combinations that improve the desired qualities. In modern breeding set-ups, it is common to use germplasms from diverse origins, which likely contain different allelic variants, to obtain combinations of traits that result interesting for the breeder. The purpose of our AG simulation design was to represent such a scenario, where a NAM population could be set up using multiple parents with different degrees of genetic diversity. Two questions arose within this scenario:

- What is the effect of higher number of alleles, therefore a higher number of genetic effect parameters, in terms of QTL detection?
- What is the effect of genetic similarity (relatedness) between individuals on QTL detection?

These two questions are strictly related in our case. When total relatedness decreases in a NAM, it is due to a higher number of AGs being present, and therefore more allelic variants being present. So, allelic diversity and genetic similarity are indirectly related. It is reasonable to assume that such relationship also exists in real diversity panels: less related individuals have higher chances of contributing a non-represented allele to the population.

In terms of our simulation, we needed to establish a genetic relationship between AGs; that is, whether most of the allelic variation is present within AGs or between AGs. Should we simulate many different alleles in each AG, but keep the average genetic affect constant between AGs? Should we establish different average genetic effects between AGs, but have similar genetic effect distributions? Was it more reasonable to variate the genetic effects mostly between AGs? Since no assumption was clear, and studies on genetic structure pointed to a more complex scenario, dependent on trait and species (D'hoop *et al.*, 2010; Berdugo-Cely *et al.*, 2017; Garin *et al.*, 2017), the study of multiallelism was limited to a NAM1 population. Afterwards, the effect of different relatedness within the NAMs was studied using multiple biallelic QTL loci (only one ancestral allele with an effect, while all other had effect 0).

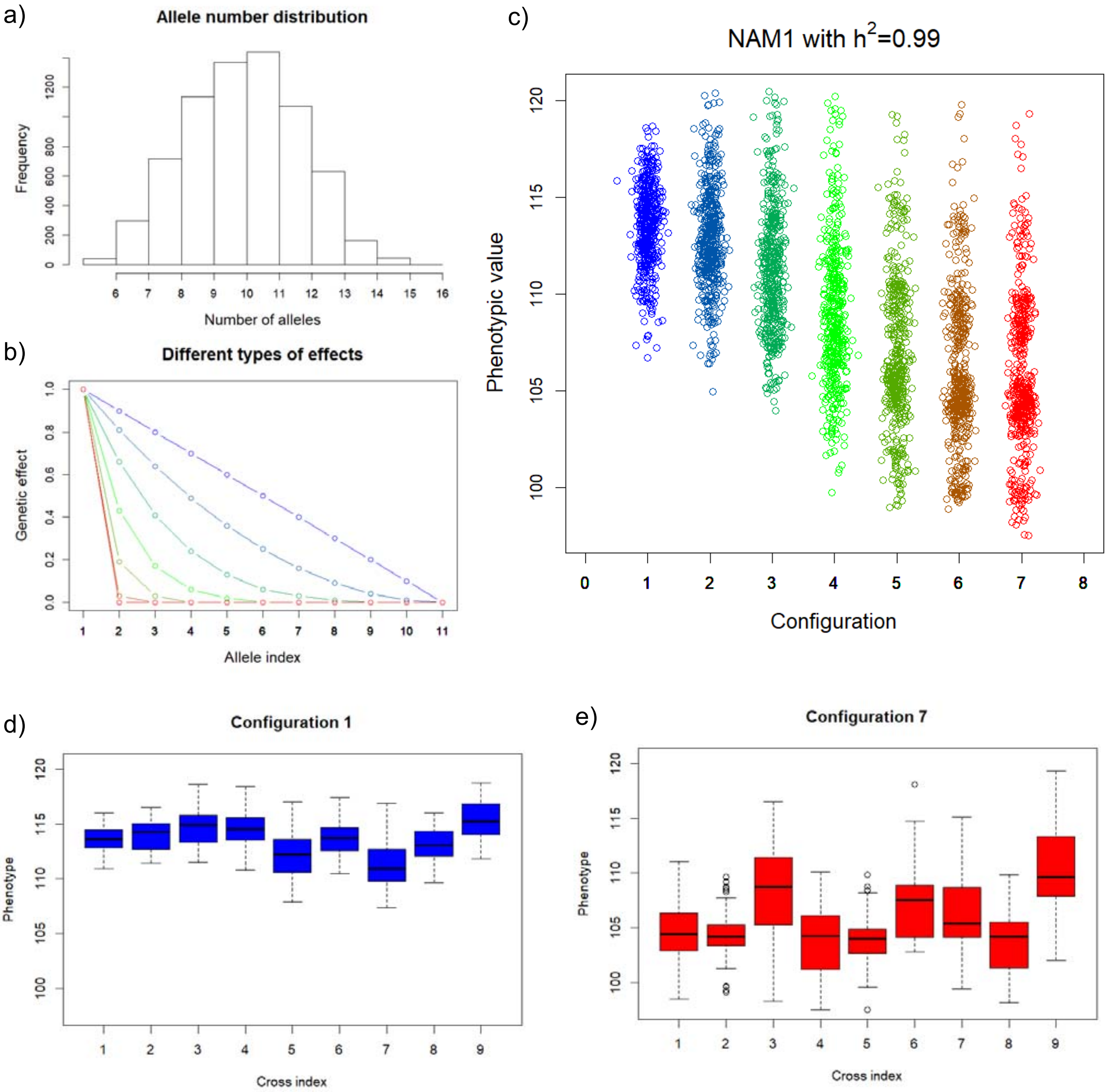


Figure 3: Genetic effects and phenotypes under multiallelism. a) *Number of alleles present at each locus for a NAM1 cross.* ***b)*** *genetic effect configurations for the 11 alleles present at marker 90, solcap_snp_c2_27878.* ***c)*** *Phenotype distribution for configurations 1 to 7. Colours correspond to the distribution of genetic effects displayed on figure 3b.* ***d)*** *boxplot of phenotypes grouped by cross of the data shown in figure 3c, for configuration 1.* ***e)*** *Same boxplot as d), but for data from configuration 7.*

3.2.1 Multiallelism

In any given NAM1 population, there might be between 5 and 16 ancestral alleles at a specific locus (Fig. 3a). Formula 9 (section 2.2.1) was designed to obtain equivalent genetic effect distributions under different number of segregating alleles. This concept is explained in Figure 3b. In blue, a situation where all alleles but the last have an effect (of different sizes); in red, a situation where only one allele has an effect, generating a biallelic locus.

Changing the configurations has a direct effect on the phenotype segregation of the NAM population (Fig. 3c). With a multiallelic locus (blue, configuration 1), the whole population displays a high phenotype; as all individuals carry active alleles, all tend to high phenotypes thanks to the additivity of the trait. On a biallelic situation (red, configuration 7), five classes are detected (dosages 0 to 4 of the active allele) and the total phenotypic variation of the population is wider. Moreover, this change between multiallelic and biallelic situations also has a direct effect in cross differentiation: under the multiallelic scenario, crosses display similar ranges of phenotypes and close means (Fig. 3d) while under a biallelic segregation, crosses have a wider variation within cross and between cross (Fig. 3e).

This lower contrast between crosses when more alleles are present can be traced back to the additive model proposed plus the fact that all genetic effects given were positive. As in configuration 1 there are more alleles having a positive effect, the average trait value is increased. Such situation would not happen if dominance or interaction between alleles was modelled. Moreover, if negative effects are also included, the variation range and differentiation between crosses is increased, up to levels similar to the biallelic scenario (data not shown).

Another key point is the effect of dosage on population segregation. In a biallelic scenario, a diploid organism would present three distinct classes segregating under a Mendelian pattern. Under polyploid genetics, such distinction becomes fuzzier. Indeed, classes can be recognized, but under a heritability of 0.99! When heritabilities have a more realistic value, environmental variance alone erases the clear category division generating an almost continuous range of phenotypic variation. This effect is increased as the number of segregating alleles rises, even if no additional loci determine the trait.

These findings suggest that, in NAM1, under multiallelic scenarios, QTL analysis will be less powerful, as differences between individuals will be lower; a conclusion that will be confirmed by the QTL analysis results.

3.2.2 MultiQTL experiment

Selection of genetic effects for NAMs with multiple AGs was done following formula 15 (section 3.2.2). This procedure generates a dataset with three important properties:

1. Each AG contributes with one active QTL allele at a different QTL location. Therefore, there will segregate as many QTLs as AGs present in the NAM.
2. Each additional QTL effect is smaller than the present ones, meaning that the central parent will carry the strongest QTL effect, and the last peripheral will harbour the smallest effect.
3. The major QTL, from the AG of the central parent, will segregate in all crosses; while the peripheral QTLs will segregate in a limited number of crosses.

This scheme generates a trade-off between diversity and statistical power: in higher diversity NAMs, more QTLs will segregate, each present in a smaller number of crosses. Therefore, we can expect that under higher diversity levels the power to detect peripheral QTLs will diminish.

Correlation between genetic structure and phenotype is an important confounder of QTL analysis. To investigate this point, we checked this correlation in our simulation (Fig 4). In Figure 4 we see the phenotypes present in each cross (points) grouped by the ancestral group of the peripheral parent (boxes). We observe how, in a NAM3, there is a clear decreasing tendency between AGs; the first crosses segregate for bigger effect QTLs than the last crosses. Such pattern is maintained in the NAM 7 plot (right panel), but only on the first two crosses. The last four crosses all maintain the same mean, which translates as a lower correlation between genetic structure and phenotype.

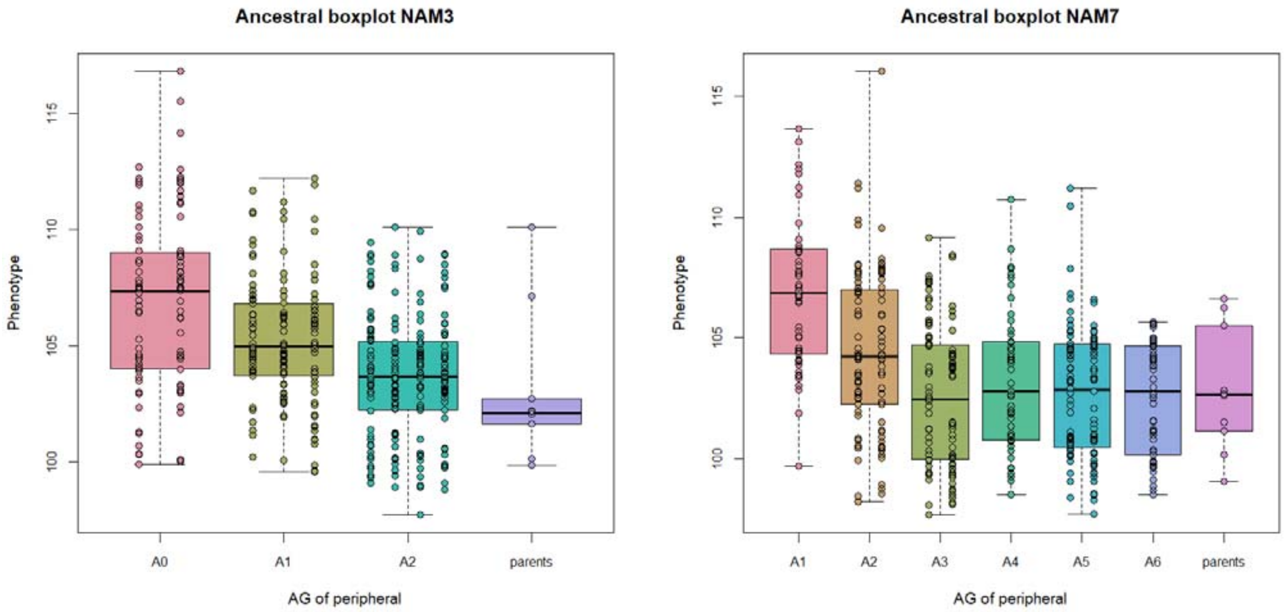


Figure 4: *Phenotype distribution of NAMs, with crosses grouped according to the ancestral group of the peripheral parent. A0 always represents the ancestral group of the central parent.*

Moreover, in both plots segregation within crosses is wider than the difference between crosses. Segregation in polyploid organisms generates a wide variation under additive traits and, even when QTL effects are very different in size, segregation still overlapped phenotypes of those crosses. For instance, in the right panel of Figure 4. The crosses representing A3, A4, A5 and A6 all present the same mean, even though they segregate for QTLs of size 4, 3, 2 and 1 as well as for the central QTL, of size 7. The segregation of the central QTL completely masks the effect of secondary QTLs with different sizes.

These plots suggest that our multiple QTL simulation procedure (section 2.2.2) is effective in generating a certain degree of correlation between genetic structure and phenotypes. There are differences between ancestral groups, although they are not bigger than the inner variation of each cross. Moreover, the simulation seems more effective in maintaining such structure in lower diversity levels (NAM3) than in higher diversity levels (NAM7), where central QTL segregation is accounting for most of the variation within each cross, effectively hiding the influence of secondary QTLs. As a consequence, we expect that in our association models structure corrections based on the fixed term Q, which corrects for different ancestral group means, will be more effective in lower diversity NAMs, where there is a clearer difference between ancestral group and cross means.

3.3 Statistical model evaluation

Association models were evaluated under the two scenarios described (multiallelism and multiQTL) to compare their detection power, varying the following factors: heritability (as defined in section 2.2), multiallelism (i.e. number of active alleles per QTL), and NAM structure (i.e. number of AGs represented in a NAM population).

Some general tendencies were consistent across all experiments. **An increase in heritability lead to higher detection power**. Such observation is not surprising, since with higher heritability, the correlation between genotype and phenotype is increased, and therefore p-values rise. In terms of structure correction, **K models** showed a better control of false positives. The Q correction never captured as much structure as the K matrix, and when combined (Q+K models) the Q parameters did not add any explanatory value to the model. This is due to the kinship matrix structure, which already identifies individuals from the same population, that is, the same cross (Fig 2, left panel). We can expect that this matrix is already absorbing all the explanatory value that a Q matrix could add. The Q correction was only useful in linear models, and only in NAM3 and NAM7 of the multiQTL experiment, not NAM10 and NAM1. This highlights the characteristics of the simulated phenotypes: as variation within crosses is higher than variation within crosses, the Q correction only has an effect when additional structure (ancestral relatedness structure) is present.

The Li and Ji p-value correction threshold yielded a similar corrected threshold for all populations, around $7.8 * 10^{-5}$, even though lower diversity NAMs have a higher expected correlation between markers. This apparent contradiction highlights that the increased linkage disequilibrium in low diversity NAMs plays a minor role compared to the linkage between markers due to recent relatedness. Therefore, the Li and Ji method does not vary its threshold even if there is an increase in the linkage disequilibrium.

3.3.1 Multiallelism

The results of this experiment concorded with what was expected according to the phenotype analysis: the models were only able to predict QTL positions under biallelism. Even under high heritabilities, multiallelic configurations led to no detection of the QTL. Under biallelic configurations, an increased heritability led to higher detection power. The same cannot be said when phenotypes

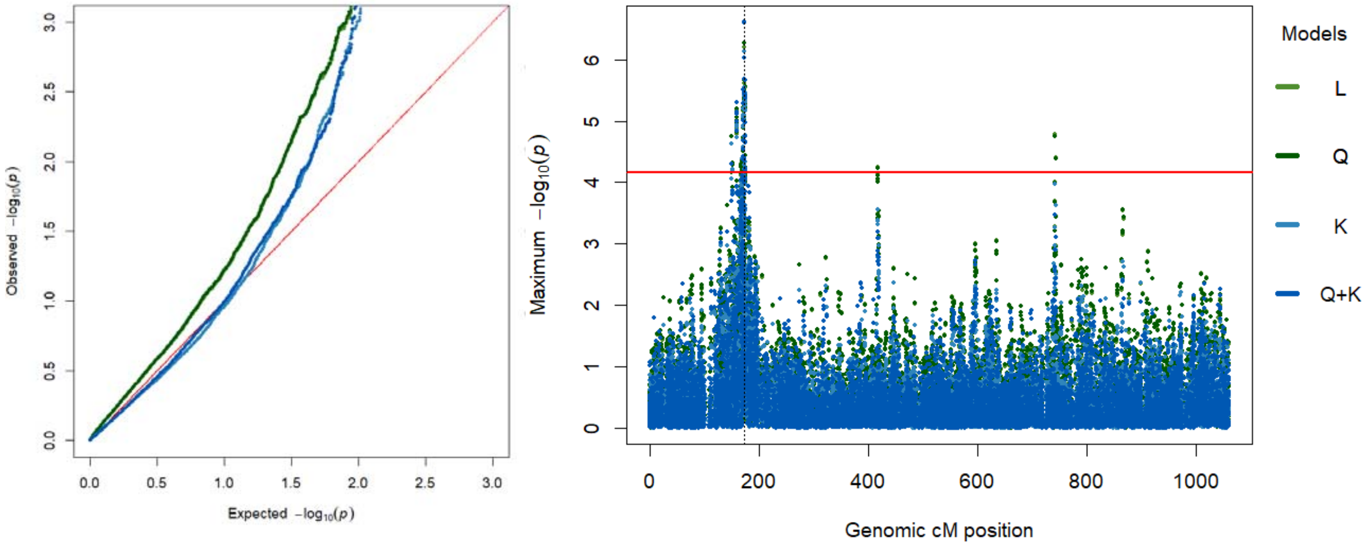


Figure 5: QQ-plot and Manhattan plot of the Ancestral models, under biallelic configuration with $h^2 = 0.8$. *The true QTL is marked with a dashed line and the horizontal red line represents the significance threshold for multiple tests.*

were generated with multiple active alleles; although there was a clear tendency to higher power with bigger heritabilities, some exceptions to the norm were observed.

In this experiment, the QQ-plots were significantly deviated in most cases, only with some biallelic models being close to the expectation line. Manhattan plots showed that all models delivered a similar set of p-value distributions, as can be seen in the right panel of Figure 5, where the points from different models cannot be clearly distinguished due to an overlap between p-values of different models. Moreover, many plots showed false positive associations, as peaks of adjacent significant p-values in regions were no QTL effect was simulated (see for example Figure 5, right panel). These peaks were consistent across different allele parametrizations, although were more abundant in higher-parameter models ($biallelic \ < ancestral \ < parental$).

The reason for these peaks was unclear. In GWAS panels, low allele frequencies tend to be a concern, leading to the common practice of removing alleles with >5% Minor Allele Frequency (MAF). It is possible that variation in frequencies of alleles can generate some spurious associations between low frequency alleles and the phenotype. This hypothesis was tested by comparing frequency patterns and the different Manhattan plots, yet no clear trend was observed. Moreover, correlation tests were performed between allele variance (variance of coded alleles) and p-value significance and, although generally there was a significant correlation between the two, they spanned values between -0.2 and 0.2, indicating a low relationship between the two measures.

Another explanation would be spurious association between phenotypes and alleles. Although possible, such scenario is unlikely. These correlations are rare, and although they would explain the observed peaks (linked regions will show similar associations, imitating a true positive peak) they would not appear repeatedly in similar regions under different scenarios.

3.3.2 MultiQTL experiment

With the use of higher diversity populations, the differences between biallelic, ancestral and parental models become clearer. In polyploids, the number of parental chromosomes segregating in a population will be equal to the number of parents times ploidy. In our case, 40 chromosomes. Thus, the parental model will present 40 genetic effect parameters, regardless of the diversity level. By contrast, the ancestral model only considers ancestral alleles, as defined by IBD. We can expect that in lower diversity MPPs, many parents might share the same IBD in their chromosomes, and thus the number of different alleles in the population is lower than the number of parental chromosomes. In high diversity populations, the ancestral and parental models will have a similar number of genetic effect parameters, but as diversity decreases, the ancestral models will reduce the number of estimated genetic effects. In contrast, the biallelic model will only estimate SNP effect, that is, it will always have one single genetic effect parameter (the difference between effect of SNP state A and SNP state B).

In Figure 6, the QQ-plots of each of the twelve models are shown, grouped by the type of marker used. We see how the mixed models (those including the K correction) present a better p-value distribution when compared to the linear models, which present overly inflated p-values. This is most likely due to correlated measures between individuals, which tend to increase significance of tests if not considered. In Figure 7 we can observe six Manhattan plots that summarize the twelve models applied to a NAM3 population with $h^2 = 0.5$. Each plot shows the linear or mixed model, with or without the Q correction term. It is clear that such term has an effect in linear models, but no effect in mixed models, as previously mentioned.

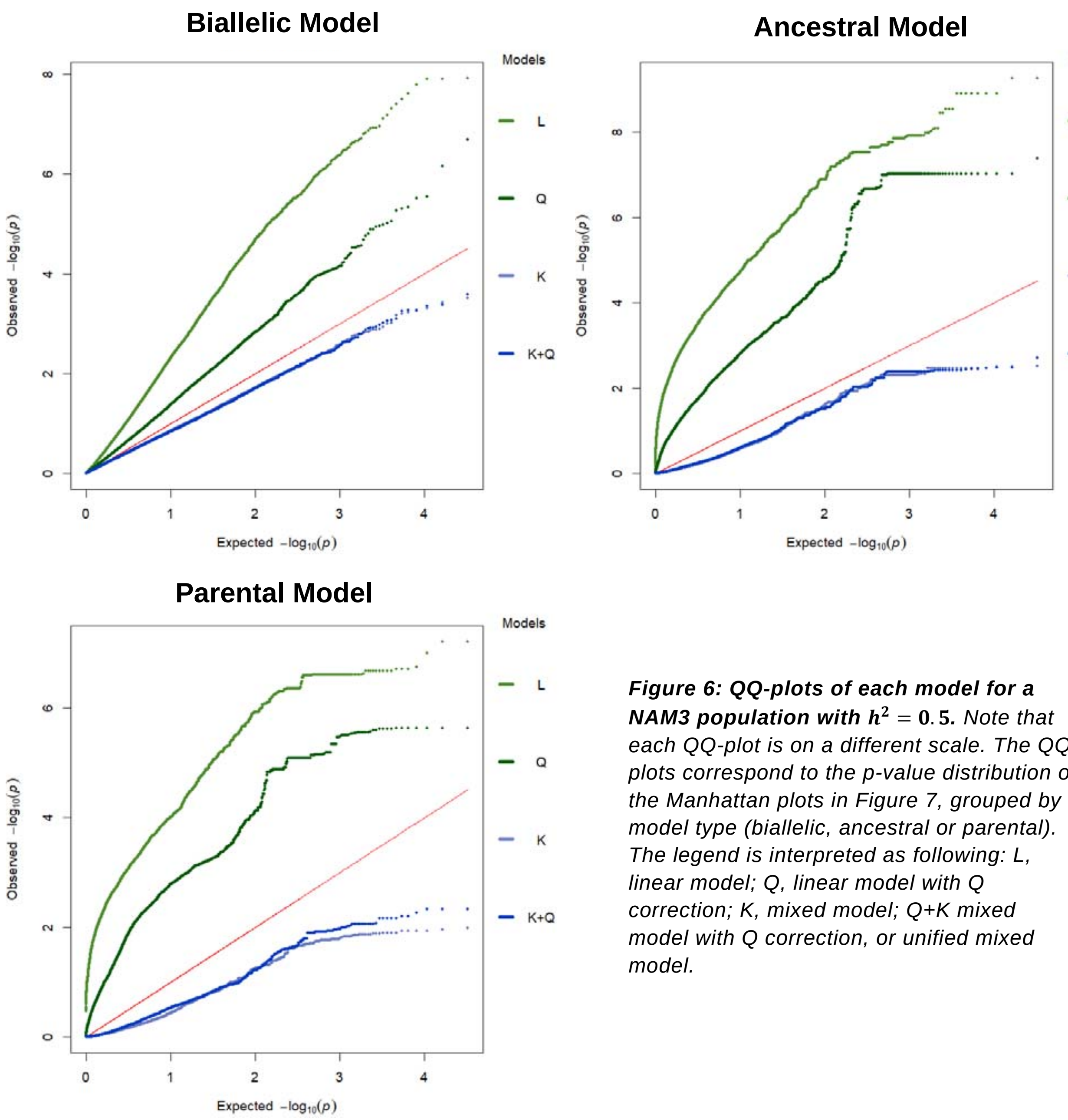
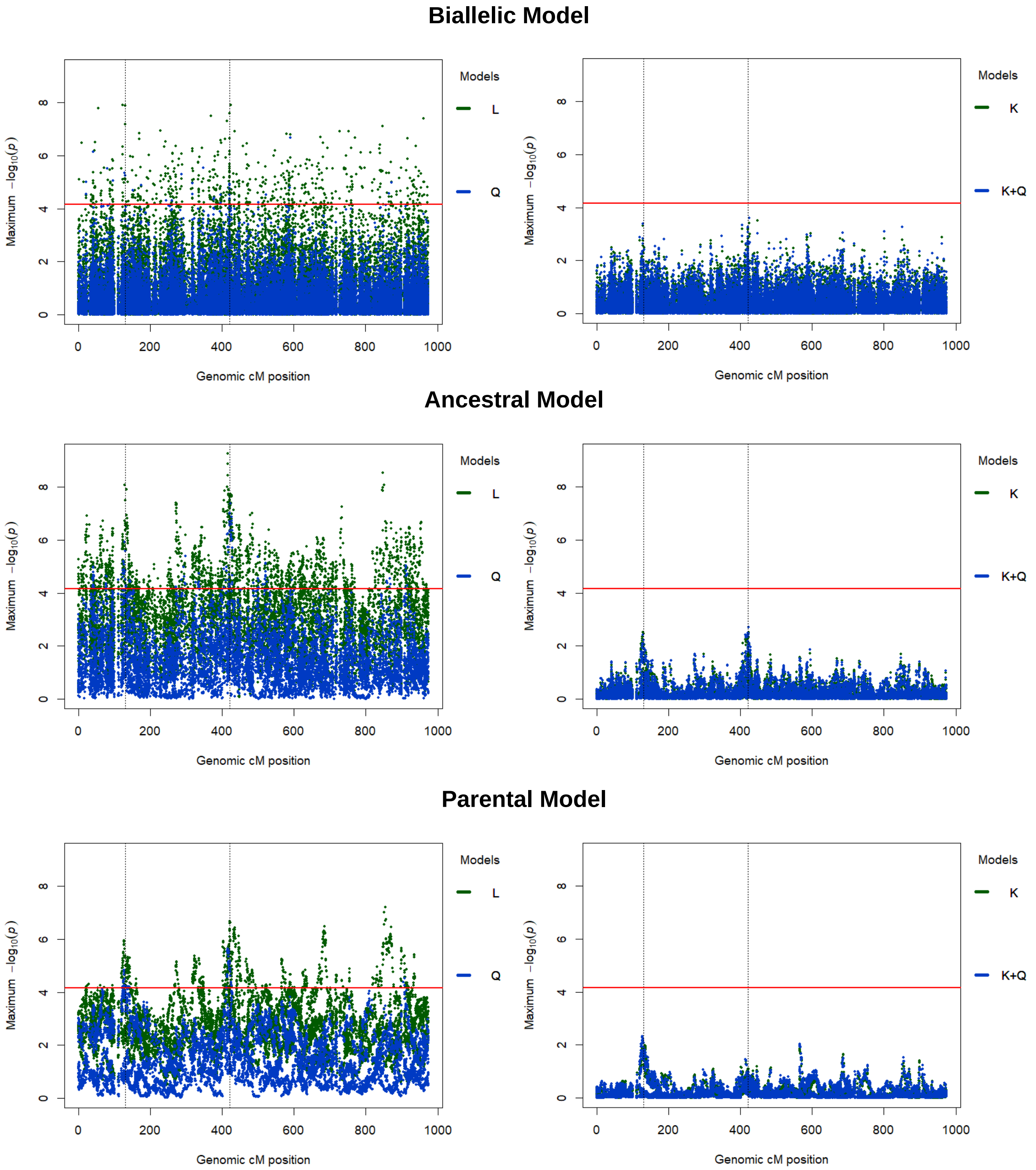


Figure 6: QQ-plots of each model for a NAM3 population with $h^2 = 0.5$. *Note that each QQ-plot is on a different scale. The QQ-plots correspond to the p-value distribution of the Manhattan plots in Figure 7, grouped by model type (biallelic, ancestral or parental). The legend is interpreted as following: L, linear model; Q, linear model with Q correction; K, mixed model; Q+K mixed model with Q correction, or unified mixed model.*

In terms of allele parametrization, we can see that **the best model is the ancestral K** model. The biallelic K shows a better QQ-plot distribution, but it is unfeasible to distinguish the peaks under the threshold, as many other regions display similarly high values. The increased number of parameters of the parental K model reduces the size of the peaks in the Manhattan plot, therefore reducing the power of the test. Moreover, the QQ-plot has a much worse distribution than the ancestral K. Lastly, we see that although the peaks are clear in the ancestral K model, and they are more significant than the false peaks along the genome, they remain under the Li & Ji threshold line, suggesting that such method might be too conservative for this approach.



Biallelic Model

Ancestral Model

Parental Model

Figure 7: PARTIAL Manhattan plots of a NAM3 population with heritability of 0.5. *The last markers of the genome were excluded as they contain the central QTL and the largest peak, which hampered appreciation of the tendencies. All plots have been obtained by overlapping the results of 15 NAM3 populations. Moreover, to ease interpretability, only the maximum p-value at each locus has been plotted. The dashed lines indicate positions of the simulated QTLs, and the red line indicates the Li&Ji threshold. To increase comparability between plots, all plots were drawn using the same scale.*

4 Conclusions

In section 1.4 the objectives and research questions of this thesis were outlined. The statistical methodology and computational tools were developed and successfully applied to the simulated datasets. The effect of diverse genetic parameters was characterized, and its impact on model performance has been assessed both theoretically and practically. By using the biallelic, ancestral and parental models we have assessed the effect of multiallelic markers. Let us review in further detail the lessons that can be extracted from this work.

Simulated populations

PedigreeSim has proven a very useful and powerful tool for simulating related populations by taking advantage of its pedigree specification options, allowing for the generation of ancestral genetic pools to obtain genetically structured polyploid populations. All distances used, including the geometric Euclidean distance, were powerful enough to distinguish such structure, although efficient implementation was not possible with PolyRelatedness software. Finally, the Realized Relationship matrix was chosen as the most efficient way of obtaining genetic distance estimations.

Simulated phenotypes

The methods proposed in section 2.2 to obtain additive polyploid phenotypes achieved to simulate a certain degree of correlation with the genotypic structure. The observations shown under 3.2 suggest that a different genetic effect selection method could improve the phenotype simulation. In the single QTL, multiallelic scenario, the addition of only positive genetic effects diminished the total phenotype variation in the offspring, consequently reducing statistical power during QTL analysis. In the multiple QTL experiment, the very low effects used in peripheral QTLs tend to become irrelevant when compared to the segregation of the main QTL. Generating phenotypes including negative genetic effects, or more balanced multiple QTL effects, might lead to different conclusions regarding the effect of multiallelism.

When additive traits are composed of a single QTL with multiple positive effect alleles (equivalent to configuration 1 of figure 3b) phenotype variation is decreased, thus diminishing the statistical power of the QTL analysis. It is easy to imagine that a breeder, willing to find the best combination of genes, would set up an MPP with the best available parents; yet our findings suggest that when no dominance or interaction is present, this would lead to a drop in QTL detection power. A “contrast approach” is therefore recommended to locate QTLs, where parents with opposite phenotypes are crossed, to generate a segregating population.

However, it seems also crucial to consider the size of QTL effects. Judging by the phenotyping results of the multiallelism experiment, when big effect QTLs are present in a cross, the segregation of this QTL plus the environmental variation can easily mask the effect of smaller size QTLs. Therefore, when trying to locate such small effect QTLs, polyploid segregation of big-effect QTLs should be avoided, for instance by crossing lines that have already introgressed the most interesting QTLs.

Model Selection

With all models, increased heritability lead to a higher detection power. Under all studied scenarios, the ancestral K model, using multiallelic markers and a mixed-model correction, had the highest detection power. The biallelic models showed better properties in terms of p-value distribution, but its power was limited to non-existent without a high heritability. These findings suggest that biallelic markers might conform better to model assumptions (normal distribution and homogeneity of variance) than multiallelic markers but are not sufficiently associated with phenotypes. We can conclude that in open-pollinated autopolyploids biallelic markers are inherently limited. The high genetic diversity unavoidably beers too far off the biallelic scenario that is generally applied to pure line crosses in diploid organisms. Even in biallelic QTL scenarios, unless there is a high degree of association between SNP and QTL segregation, biallelic markers are unable to adequately track QTL position.

On the other hand, the ancestral model loses detection power in highly diverse populations (with many AGs represented) and nested genetic effects (ancestral alleles only present in one cross). The reason is simple, with very high diversity and nested genetic effects, an ancestral model is equivalent to a parental model that would obtain parental chromosome effects. This directly translates to the design of the parental panel: rather than choosing a low diversity or high diversity panel, the crosses should be designed as “ancestral crosses” where multiple genetic pools are crossed by sampling individuals of each genetic pool. That is, having balanced groups of parents originating from the same genetic population. This design improved ancestral model performance under the additive scenario that we simulated.

Future prospects

This thesis proved to pose relevant questions regarding the statistical properties of MPP NAM populations and has generated methodologies for optimized QTL polyploid analysis using the unified mixed model framework. From this set of simulated data and statistical tools, much more can be explored. For instance, applying phenotype simulation procedures that include dominance and interaction to the same NAM populations might allow to tackle new questions regarding the genetics of autopolyploid organisms.

The proposed experiments allowed to answer some relevant questions, but additional answers could be obtained with other experimental designs. In the multiallelism experiment, genetic diversity and QTL number were linked; meaning that the effect of **QTL number** and **parental diversity** on detection power was not assessed separately. As a follow-up experiment, multi-QTL traits could be simulated with a fixed number of QTLs on NAMs with different diversity levels. Genetic effects could be designed as coming from equivalent distributions (same shape) but with different means for each ancestral group.

Additionally, QTL analysis can be strongly influenced by **allele frequencies,** which was taken into account during the thesis project. However, as it was not part of our research questions the effect of allele frequencies in this analysis has not been thoroughly explored. It is quite possible that exploring the effects of allele frequency in relation to genetic diversity, could yield further insights into the properties of MPPs and the statistical models developed.

Breeding in polyploid organisms is very unlikely to stop, and the prospect of using MPPs to characterize a wide diversity in a single experiment is very attractive. Clearly, a broader field of research can be explored regarding the genetic properties of polyploids. Hopefully, the tools and models developed in this thesis will help answer more questions regarding QTL detection and genetic diversity, as well as open new research possibilities.

Data accessibility

All data is available at the following address:
https://drive.google.com/open?id=1iCllHcXyjPgKAJ8ao3LxtYahpFvzWGOa

Two .rar files can be found in this drive: `mpQTL_scripts.rar`, containing only the script files and `mpQTL.rar`, which contains both scripts and the generated data and plots. In the latter, a series of folders are included. A short description of each of them:

- **Distances**: folder with PolyRelatedness program, frequency files used for the program, as well as storage of Euclidean and PolyRelatedness distance in text files.
- **PedigreeSIM:** contains all genetic information and files for simulating data, including the program itself. Within it, one can find some folders as well:
    - **NAM_crosses:** contains the simulated NAM populations, either as text files or compressed in a rar file.
    - **Parents:** contains the ancestral population data and text files with the parentals.
    - **Potato:** contains the genetic map information used for the simulation.
    - **Scripts:** contains code provided by Peter Bourke and Roeland Voorrips.
- **Storage**: rar file containing results of experiments 1 to 4 in RDS objects, as well as a number of plots comparing QQ-plots distributions, Manhattan plots and p-value distributions of various experiments.

All script files contain self-explanatory comments and should be understandable individually, but as a short summary:

- **mpQTL_fun.R** is an R file containing all functions used to obtain the results, including genetic simulation, PedigreeSim functions, PolyRelatedness functions, phenotype simulation, mixed model solving, data processing and generation of plots. Most functions are well documented and explained, except those for plot-making, as they were considered quite custom.
- **Old_base_QTL.R** old script for a Multiallelism experiment that does not include `map.QTL` function neither Q structure correction.
- **base_QTL2.R** and **base_MultiQTL.R** are scripts for performing multiallelism and multi QTL experiments.
- **Pedsim.R** contains the code for simulating the ancestral population and the NAM crosses.
- **QTLanalysis.Rmd** shows some characteristics of the phenotyping methods under the Multiallelism experiment, explaining the implications of different configurations and genetic effects as well.
- **Distances.Rmd** shows the distance comparison procedure in NAM1, NAM3, NAM7 and NAM10 for the different measures tested.
- **Herit_bias_multi.Rmd** shows the heritability bias explained on section 2.2.2 for a NAM3 population.

Self-Evaluation

During this thesis I have learnt much, particularly regarding optimization of computation and development. The need for iterating highly demanding operations inevitably led me to optimize code to ensure that it was not using any non-essential resources. That has finely improved my coding skills and my idea of programming, using a wider range of tools to attain high speed. However, it has also showed me that sometimes, such iterations are not necessary. At a certain point, repeating a test more times will not yield more information, but will significantly increase computation time and data complexity. Moreover, if the results go wrong (which they will), having to repeat another 100 tests will inevitably slow down the development process. Therefore, for big data, less is more and simple is powerful. As interesting and sophisticated some experiments might be, if they do not yield useful answers, they are little more than a waste of time and resources.

Besides optimization, I have also seen that there is always something to learn from data, even if it is how *not* to perform a certain test. But more often than not, even if our original questions could not be answered, *something* can be extracted from the results. This can be a crucial part of a scientific process with big data, not only testing the original hypotheses, but also other alternative hypothesis that are suggested by the dataset itself.

However, this focus on the handling of complex dataset, optimization problems and result processing, has slowed down my statistical learning. The reader might notice a lack of proper statistical tests to compare models or to analyse phenotype and genotype structure correlation. Under the six months that took to generate this piece of work, optimization of computation took much of the time, unavoidably lagging the statistical part of the analysis. Nevertheless, the comparison methods used are based on solid statistical properties and should hold under more standard measures such as Akaiki Information Criterion tests.

Acknowledgements

This thesis would have been impossible without the insightful guidance of Chris Maliepaard and Giorgio Tumino, who had the patience and will to discuss and think aloud about the wide range of problems we tackled. Moreover, I am very grateful to Peter Bourke and Roeland Voorrips, from the Plant Breeding group, who helped me use some of the computational tools, contributed with very useful scripts and gave interesting ideas in the process. I would also like to thank Bas Engel and Vincent Garin from Biometris, who guided me through some mathematical problems when dealing with the statistical framework.

Aside from the academic support, this work would not exist if it was not for excellent teachers I had during my bachelor and master. I would specially like to thank Jesús Piedrafita, who introduced me into the wicked world of Quantitative Genetics and its wonders; Raquel Egea, who shaped the way I programmed and understood computing; and Cristian Dobre, whose insightful and entertaining classes spurred my interest for such a feared topic as statistics, which I have come to find so compelling.

And last but not least, thanks to my parents and family for their support, both emotional and economical, without whom this thesis and my master education would not have been possible.

To all, thanks.

REFERENCES

Bardol, N., Ventelon, M., Mangin, B., Jasson, S., Loywick, V., Couton, F., Derue, C., Blanchard, P., Charcosset, A. and Moreau, L. (2013) Combined linkage and linkage disequilibrium QTL mapping in multiple families of maize (Zea mays L.) line crosses highlights complementarities between models based on parental haplotype and single locus polymorphism. *Theoretical and Applied Genetics* **126**, 2717–2736. doi:10.1007/s00122-013-2167-9.

Berdugo-Cely, J., Valbuena, R. I., Sánchez-Betancourt, E., Barrero, L. S. and Yockteng, R. (2017) Genetic diversity and association mapping in the Colombian Central Collection of Solanum tuberosum L. Andigenum group using SNPs markers. *PloS one* **12**, e0173039. doi:10.1371/journal.pone.0173039.

Blanc, G., Charcosset, A., Mangin, B., Gallais, A. and Moreau, L. (2006) Connected populations for detecting quantitative trait loci and testing for epistasis: an application in maize. *Theoretical and Applied Genetics* **113**, 206–224. doi:10.1007/s00122-006-0287-1.

Bourke, P. M., Voorrips, R. E., Kranenburg, T., Jansen, J., Visser, R. G. F. and Maliepaard, C. (2016) Integrating haplotype-specific linkage maps in tetraploid species using SNP markers. *Theoretical and Applied Genetics* **129**, 2211–2226. doi:10.1007/s00122-016-2768-1.

Cavanagh, C., Morell, M., Mackay, I. and Powell, W. (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. *Current Opinion in Plant Biology* **11**, 215–221. doi:10.1016/j.pbi.2008.01.002.

Cleveland, W., Grosse, E., Shyu, W. and JM (1991) Linear Models, p. 608 *in* J. M. Chambers and T. J. Hastie (Ed.) *Statistical models in S*. Chapman & Hall/CRC.

D'hoop, B. B., Paulo, M. J., Kowitwanich, K., Sengers, M., Visser, R. G. F., van Eck, H. J. and van Eeuwijk, F. A. (2010) Population structure and linkage disequilibrium unravelled in tetraploid potato. *Theoretical and Applied Genetics* **121**, 1151–1170. doi:10.1007/s00122-010-1379-5.

Endelman, J. B. (2011) Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal* **4**, 250. doi:10.3835/plantgenome2011.08.0024.

Garin, V., Wimmer, V., Mezmouk, S., Malosetti, M. and van Eeuwijk, F. (2017) How do the type of QTL effect and the form of the residual term influence QTL detection in multi-parent populations? A case study in the maize EU-NAM population. *Theoretical and Applied Genetics* **130**, 1753–1764. doi:10.1007/s00122-017-2923-3.

Giraud, H., Lehermeier, C., Bauer, E., Falque, M., Segura, V., Bauland, C., Camisan, C., Campo, L., Meyer, N., Ranc, N., Schipprack, W., Flament, P., Melchinger, A. E., Menz, M., Moreno-Gonz??lez, J., Ouzunova, M., Charcosset, A., Sch??n, C. C. and Moreau, L. (2014) Linkage disequilibrium with linkage analysis of multiline crosses reveals different multiallelic QTL for hybrid performance in the flint and dent heterotic groups of maize. *Genetics* **198**, 1717–1734. doi:10.1534/genetics.114.169367.

Hackett, C. A., Bradshaw, J. E., Meyer, R. C., McNicol, J. W., Milbourne, D. and Waught, R. (1998) Linkage analysis in tetraploid species: a simulation study. *Genetical Research* **71**, 143–153. doi:10.1017/S0016672398003188.

Hackett, C. A., Bradshaw, J. E. and McNicol, J. W. (2001) Interval mapping of quantitative trait loci in autotetraploid species. *Genetics* **159**, 1819–1832.

Hackett, C. A., McLean, K. and Bryan, G. J. (2013) Linkage Analysis and QTL Mapping Using SNP Dosage Data in a Tetraploid Potato Mapping Population. *PLoS ONE* **8**,. doi:10.1371/journal.pone.0063939.

Hackett, C. A., Bradshaw, J. E. and Bryan, G. J. (2014) QTL mapping in autotetraploids using SNP dosage information. *Theoretical and Applied Genetics* **127**, 1885–1904. doi:10.1007/s00122-014-2347-2.

Hayman, B. I. (1954) The Theory and Analysis of Diallel Crosses. *Genetics* **39**, 789–809.

Huang, K., Guo, S., Shattuck, M., Chen, S., Qi, X., Zhang, P. and Li, B. (2014) A maximum-likelihood estimation of pairwise relatedness for autopolyploids. **114**,. doi:10.1038/hdy.2014.88.

Jansen, R. (1993) Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C. and Eskin, E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354. doi:10.1038/ng.548.

Kao, C. H., Zeng, Z. B. and Teasdale, R. D. (1999) Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–16.

Kempthorne, O. (1957) *An introduction to genetic statistics*. Shewhart, W. A. and Wilks, S. S. (Eds.) New York: John Wiley & Sons, Inc.; London : Chapman & Hall Ltd.

Lander, E. S. and Botstein, D. (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–99.

Li, J. and Ji, L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* **95**, 221–227. doi:10.1038/sj.hdy.6800717.

Liu, W., Reif, J. C., Ranc, N., Porta, G. Della and Würschum, T. (2012) Comparison of biometrical approaches for QTL detection in multiple segregating families. *Theoretical and Applied Genetics* **125**, 987–998. doi:10.1007/s00122-012-1889-4.

Luo, Z. W., Hackett, C. A., Bradshaw, J. E., McNicol, J. W. and Milbourne, D. (2001) Construction of a genetic linkage map in tetraploid species using molecular markers. *Genetics* **157**, 1369–1385.

Luo, Z. W., Maliepaard, C. A., Leach, L., Zhang, R., Bradshaw, J., Kearsey, M. and Newburg, L. (2005) Constructing Genetic Linkage Maps Under a Tetrasomic Model. *Genetics* **172**, 2635–2645. doi:10.1534/genetics.105.052449.

Lynch, M. and Walsh, B. (1998) *Genetics and analysis of quantitative traits*. Sinauer.

Massa, A. N., Manrique-Carpintero, N. C., Coombs, J. J., Zarka, D. G., Boone, A. E., Kirk, W. W., Hackett, C. A., Bryan, G. J. and Douches, D. S. (2015) Genetic Linkage Mapping of Economically Important Traits in Cultivated Tetraploid

Potato (Solanum tuberosum L.). *G3; Genes|Genomes|Genetics* **5**, 2357–2364. doi:10.1534/g3.115.019646.

McMullen, M. D., Kresovich, S., Villeda, H. S., Bradbury, P., Li, H., Sun, Q., Flint-Garcia, S., Thornsberry, J., Acharya, C., Bottoms, C., Brown, P., Browne, C., Eller, M., Guill, K., Harjes, C., Kroon, D., Lepak, N., Mitchell, S. E., Peterson, B., Pressoir, G., Romero, S., Rosas, M. O., Salvo, S., Yates, H., Hanson, M., Jones, E., Smith, S., Glaubitz, J. C., Goodman, M., Ware, D., Holland, J. B. and Buckler, E. S. (2009) Genetic Properties of the Maize Nested Association Mapping Population. *Science* **325**, 737–740. doi:10.1126/science.1174320.

Meuwissen, T. H. E., Hayes, B. J. and Goddard, M. E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829. doi:11290733.

Morton, N. E., Yee, S., Harris, D. E. and Lew, R. (1971) Bioassay of kinship. *Theoretical population biology* **2**, 507–24.

Pembleton, L. W., Cogan, N. O. I. and Forster, J. W. (2013) StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. *Molecular Ecology Resources* **13**, 946–952. doi:10.1111/1755-0998.12129.

Ramsey, J. and Schemske, D. W. (2002) Neopolyploidy in Flowering Plants. *Annual Review of Ecology and Systematics* **33**, 589–639. doi:10.1146/annurev.ecolsys.33.010802.150437.

Rosyara, U. R., De Jong, W. S., Douches, D. S. and Endelman, J. B. (2016) Software for Genome-Wide Association Studies in Autopolyploids and Its Application to Potato. *The Plant Genome* **9**, 0. doi:10.3835/plantgenome2015.08.0073.

Sattler, M. C., Carvalho, C. R. and Clarindo, W. R. (2016) The polyploidy and its key role in plant breeding. *Planta* **243**, 281–296. doi:10.1007/s00425-015-2450-x.

Sehgal, D., Singh, R. and Rajpal, V. R. (2016) Quantitative Trait Loci Mapping in Plants: Concepts and Approaches, pp. 31–59 *in Molecular Breeding for Sustainable Crop Improvement*. Springer International Publishing Switzerland doi:10.1007/978-3-319-27090-6_2.

Soriano, J. M. and Royo, C. (2015) Dissecting the Genetic Architecture of Leaf Rust Resistance in Wheat by\nQTL Meta-Analysis. *Phytopathology* **105**, 1585–1593. doi:10.1094/PHYTO-05-15-0130-R.

Sybenga, J. (1996) Chromosome pairing affinity and quadrivalent formation in polyploids: do segmental allopolyploids exist? *Genome* **39**, 1176–1184. doi:10.1139/g96-148.

Taylor, H. R. (2015) The use and abuse of genetic marker-based estimates of relatedness and inbreeding. *Ecology and evolution* **5**, 3140–50. doi:10.1002/ece3.1541.

VanRaden, P. M. (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414–4423. doi:10.3168/jds.2007-0980.

Voorrips, R. E. (2014) PedigreeSim 2.0 Manual. 1–12.

Voorrips, R. E. and Maliepaard, C. A. (2012) The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC bioinformatics* **13**, 248. doi:10.1186/1471-2105-13-248.

Voorrips, R. E., Gort, G. and Vosman, B. (2011) Genotype calling in tetraploid species from bi-allelic marker data using mixture models. *BMC bioinformatics* **12**, 172. doi:10.1186/1471-2105-12-172.

Whittaker, I. C., Thompson, R. and DENHAM, M. C. (2000) Marker-assisted selection using ridge regression. *Genetical Research* **75**, 249–252.

Wilkinson, G. N. and Rogers, C. E. (1973) Symbolic Description of Factorial Models for Analysis of Variance. *Applied Statistics* **22**, 392. doi:10.2307/2346786.

Würschum, T. (2012) Mapping QTL for agronomic traits in breeding populations. *Theoretical and Applied Genetics* **125**, 201–210. doi:10.1007/s00122-012-1887-6.

Xu, F., Lyu, Y., Tong, C., Wu, W., Zhu, X., Yin, D., Yan, Q., Zhang, J., Pang, X., Tobias, C. M. and Wu, R. (2013) A statistical model for QTL mapping in polysomic autotetraploids underlying double reduction. *Briefings in Bioinformatics* **15**, 1044–1056. doi:10.1093/bib/bbt073.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W., Goddard, M. E. and Visscher, P. M. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565–9. doi:10.1038/ng.608.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S. and Buckler, E. S. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics* **38**, 203–208. doi:10.1038/ng1702.

Zeng, Z. B. (1994) Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–68.

Zhang, Z., Ersoz, E., Lai, C.-Q., Todhunter, R. J., Tiwari, H. K., Gore, M. A., Bradbury, P. J., Yu, J., Arnett, D. K., Ordovas, J. M. and Buckler, E. S. (2010) Mixed linear model approach adapted for genome-wide association studies. *Nature Genetics* **42**, 355–360. doi:10.1038/ng.546.

Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J. and Bink, M. C. A. M. (2016) Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics* **203**, 119–131. doi:10.1534/genetics.115.185579.

Zielinski, M.-L. and Mittelsten Scheid, O. (2012) Meiosis in Polyploid Plants, pp. 33–55 *in Polyploidy and Genome Evolution*. Berlin, Heidelberg, Springer Berlin Heidelberg doi:10.1007/978-3-642-31442-1_3.