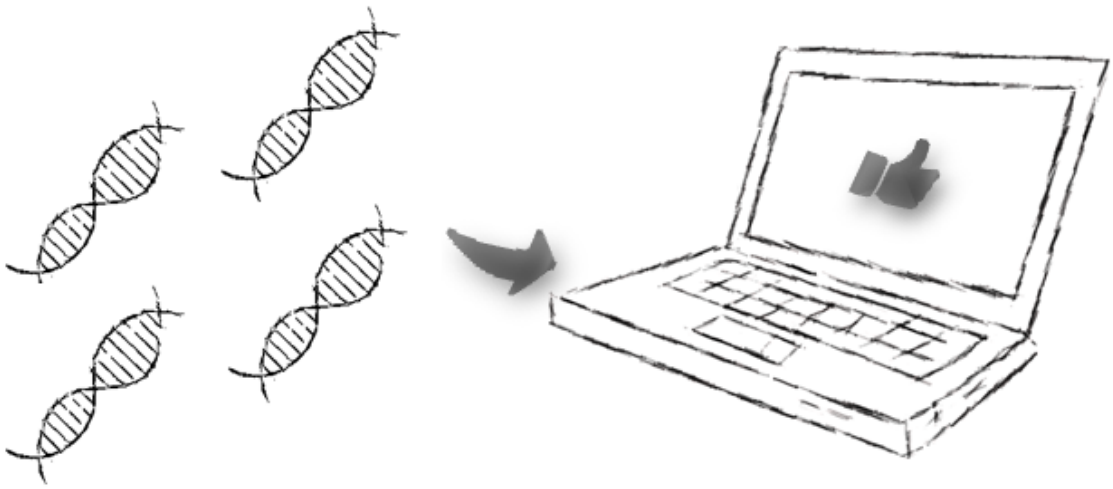
The feasibility of applying bi-allelic marker assisted selection in polyploids ---- a simulation study in autotetraploid populations.



The feasibility of applying bi-allelic marker assisted selection in polyploids ---- a simulation study in autotetraploid populations.

Master thesis

Zhou, Heming
951107987050
Master of Plant Sciences
Plant breeding department
PBR-80436
Maliepaard, Chris; Bourke, Peter
27th, Feb 2020

Wageningen University & Research
Wageningen, Netherlands

Table of contents

Abstract

In this simulation study we performed marker assisted selection (MAS) on an autotetraploid population to test its selection efficiency. We programmed interested QTL-marker linkages on two sets of virtual homologous chromosomes. First, we tested how these linkages broke over generations of MAS. Next, we investigated difference of the number of inherited QTLs and genotypic values between MAS and phenotypic selection. Our results showed that MAS outperformed phenotypic selection in polyploids if selected traits have low heritability. MAS efficiency is determined by QTL-marker linkage distance and phase, and the number of linked QTLs. QTL-marker linkages in coupling phase with short genetic distance greatly increased MAS efficiency in polyploid breeding.

Key words: polyploid, simulation, markers, selection efficiency, dosage.

Introduction

Polyploid crops are very important in agricultural production. Leek, wheat, roses, and potatoes are common, widely planted polyploid crops. High ploidy offers many advantages in plant production. The gigas effect is one of the most distinctive features of polyploids. Plants with high ploidy usually grow larger and yield higher than diploid ones (Sattler et al., 2016). Polyploids also have an advantageous feature called genome buffering which means that deleterious alleles can be masked by other alleles and some hidden alleles may even develop novel functions (Adams et al., 2005). Heterosis and heterozygosity are important features in polyploid crops as well. With increasing ploidy, a plant can achieve high heterozygosity and multiple distantly related sub-genomes in allopolyploids can greatly boost the heterosis effect (Comai, 2005). An uneven ploidy level usually leads to low fertility in plants and thus seedless fruit is produced (Sattler et al., 2016). Banana (*Musa acuminata*) as triploid is a typical fruit using the uneven ploidy advantage. Polyploids are usually specified as autopolyploids and allopolyploids based on the composition of genomes. Autopolyploids have duplicated or very similar genomes while allopolyploids have differential originated subgenomes in one set of homoelogues. Thus, different composition of genomes results in different chromosomal pairing patterns during meiosis (Stebbins, 1947; Sybenga, 1996). Autopolyploids exhibit complete random pairing between homologous chromosomes during meiosis whereas allopolyploids show non-random pairing patterns among homoeologues (Stebbins, 1947; Bourke et al., 2015). Another kind of polyploids called ‘segmental allopolyploids’ refer to polyploids that exist somewhere between these two extremes, with a “homologues preferential pairing rate” between 0~100 % (Sybenga, 1996). The boundary among many polyploids is still hard to be distinguished. In general, most of allopolyploid manifests disomic inheritance thus can be regarded as diploid during breeding research (Bourke et al., 2015). Polyploids mentioned later refer more to autopolyploid.

Research in polyploid breeding has seen some progress recently, but in general lags behind that of diploids. The application of molecular markers in the breeding of autopolyploids is relatively less common compared to diploid crops where marker-assisted-selection (MAS) is routinely executed in modern breeding programs, such as in tomato breeding (Yamamoto et al., 2016; Sim et al., 2012). Although markers have been used in potato breeding to select for late blight resistance (Colton et al., 2006), markers that are linked to multi-locus quantitative traits are less researched compared to diploids. Recently, the research concerning quantitative traits using markers has much progress (Hackett et al., 2013; Bourke et al., 2015), but the continuous application is rarely seen in polyploid breeding industry, at least from scientific studies. In practice, potato breeding has seen limited progress for yield and yield itself is a very complex trait that is most likely quantitatively inherited (Bradshaw et al., 1994). MAS is not yet popular in autopolyploids

due to several practical difficulties and cost constraints. One distinctive feature of polyploids is multi-allelism. Since alleles have more than two copies in polyploids compared to diploids. Thus bi-allelic markers in polyploids manifest the number of copies of the allele carried at a certain locus and we called this feature 'dosage'. Dosage corresponds to the allele copy number of bi-allelic single nucleotide polymorphisms (SNP) markers (Bourke et al., 2019). Dosage changes from 0 to A levels if the ploidy level is A (for example in a tetraploid, dosage can vary from 0 to 4). Varying dosage levels of markers makes the genetic analysis more complicated as marker segregation ratios become more complex. Scientists essentially rely on expected segregation ratios in offspring to determine parental genotype and do the linkage analysis. But determining these ratios using data acquired from a less researched autopolyploid population (e.g. leek, alfalfa) is not simple (Bourke et al., 2018). These features lead to a result that bi-allelic markers are less efficient in polyploids as compared to diploids. Bi-allelic markers in polyploids can have more than two different phases linked with quantitative trait loci (QTL). The situation in which a marker allele is physically linked to a QTL allele on the same chromosome is called coupling phase. If the marker allele is linked to a QTL allele on one of the other homologous chromosomes it is called repulsion phase. Some markers are linked in mixed phases with QTL. Because of this, markers linked in different phases to a QTL show different prediction ability. Preferential pairing between homologous chromosomes is another feature. Some level of preferential pairing was observed in tetraploid rose cross and this phenomenon was found to be unevenly distributed across chromosomes, which suggests that non-random homologous chromosome pairing in polyploids does exist (Bourke et al., 2017). Apart from preferential pairing, double reduction is also a unique phenomenon in autopolyploids. Double reduction leads to a situation whereby one gamete receives a double copy of part of the same parental chromosome (Fisher, 1947). In the end, the offspring inherits unusual combinations of parental chromosomes, which may appear impossible compared to diploid. For example, an autotetraploid with allele dosage of 'AAAa' can produce gametes carrying 'aa'. Previous research showed two interesting results: 1) that double reduction happens at very low frequency and 2) that the frequency increases as the distance from the locus to the centromere increases (Bourke et al., 2015). But the least study has suggested that ignoring scenarios such as multivalents and double reduction in polyploid research has marginal influence compared to another factors like QTL detection power and marker coverage on chromosomes (Bourke et al., 2019).

Another difficulty of MAS in plant breeding is the expenses. Theoretically, employing MAS in the breeding process can facilitate the selection of targeted genes, thus reducing the need to grow all individuals to maturity to be tested, and helping to speed up the whole process (Slater et al., 2014). However, it is important to consider cost-effectiveness. Previous research has suggested that mapping repulsion linkage caused many false

results (Qu et al., 2001). If a targeted trait has multiple underlying loci (QTL), then selection based on an insufficient number of markers linked in repulsion phase to a fraction of loci may not provide more effective results than phenotypic selection. Moreover, when trait heritability is very high, the necessity to apply MAS drops as phenotypic selection is equally efficient. Phenotypic selection is more tangible and cheaper than MAS. But one research has also pointed out that MAS is very cost-effective in terms of reducing breeding cycles and labour cost of phenotyping (Slater et al., 2013).

Traditional methods of polyploid breeding involve many in the field phenotypic selections on offspring. For example, potato breeding and selection is dependent on large progeny because of the current routine is to cross known varieties. Based on a large amount of offspring breeders hope to find one group of new variety out of the reshuffling genomes. However, traditional methods gradually become less efficient in plant breeding as to non-visual selection or quantitative traits, at least in diploids (Dreher et al., 2002). Future plant breeding demands a fast and accurate selection process within limited time. Polyploid crops usually have a high genetic load, carrying many deleterious alleles masked by genome buffering effects. These deleterious alleles may hide in the selection process and merge after combination among genomes. To get rid of such a problem through traditional methods requires tedious work. The traditional methods should be improved with the help of modern statistics and software. Simulation study has been given attention recently in polyploid breeding because of its growing reliability and reduction of huge amount of time compared to in the field experiments (Bourke et al., 2017; Hackett et al., 2017). One goal of *in silico* breeding simulation is to offer guidance for breeders in how to apply markers in their breeding programs. A certain superior genotype can be virtually designed by software and become the ideal goal for breeders (Peleman et al., 2003). According to research, applying marker-assisted selection helps a lot by reducing the number of field tests and individuals selected for each generation (Varshney et al., 2005; Slater et al., 2014). However, many breeders choose traditional phenotypic selection after considering the trade-off between the cost of novel techniques with their returns and traditional methods. Thus, it is vital to develop accurate and low-cost tools to assist polyploid breeding and simulation study is a good answer. However, one risk that should always be borne in mind that a simulation study does not fully guarantee to reflect real biological behavior. All parameters in a simulation are based on assumptions and defined by humans. If there are unique features in a certain crop, then the simulation will never predict those features as the program does not include those parameters in the first place. So, after simulation, breeding research still needs to be carried out in the field test for confirmation.

As for polyploid genetic research, there are several main obstacles. Many polyploid crops lack high-resolution linkage maps and genome information. Their germplasm is usually

highly heterozygous, and many crops have inbreeding depression. Developing homozygous lines through selfing is biologically impossible for many polyploids. To specifically design a breeding scheme and predict breeding results, therefore, is a challenge in polyploid breeding. Another problem is that marker application in polyploid breeding is, yet, at an early stage. Many obstacles hinder the application of markers in polyploid breeding. Positioning QTLs on a linkage map is a challenging task and there is quite some uncertainty with positions on a linkage map, therefore it cannot often be assured whether markers on the linkage map are on the flanking side or the same side of the region with the LOD score peak. Preferential pairing and multivalent formation during meiosis in polyploids also add more difficulties in predicting accurate genotypes (Zheng et al., 2016). Some simulations demonstrate that, even when such selection is quite effective, the markers utilized by selection are not necessarily the most tightly linked to the QTLs controlling the trait. Moreover, markers associated with the additive effect alleles may not accurately reflect the contributions to the trait by the most tightly linked QTLs (Gimelfarb et al., 1995). Marker dosages are important since many genetic analysis tools rely on these dosages to denote the genotype information. However, to use these dosages in real plant breeding is quite difficult compared to research. A solution is to develop multi-allelic markers for more precise genotype calling. In a bi-parental tetraploid population, the most ideal situation is that every allele on homologous chromosomes is represented by distinct markers. As for now, the accuracy of genotyping and in particular dosage determination remains an obstacle to deploying multi SNP reads in polyploids (Bourke et al., 2018). Current research has provided lot of progress concerning high-density linkage analysis in tetraploid potato (Hackett et al., 2013; Bourke et al., 2015; Endelman et al., 2018). Such high-resolution mapping information contributes to the foundation of MAS in tetraploid potato breeding.

Previous studies have developed many models and software trying to determine polyploid inheritance behaviors. Simulation software for polyploid research has been developed recently. TetraploidSNPmap is a new software that quickly orders many SNP using dosage information (Hackett et al., 2017). PedigreeSim V2.0 (Voorrips & Maliepaard, 2012) is a powerful software that simulates pedigrees and cross populations in polyploids with changeable parameters including multivalent rates and preferential pairing rates. Simulation study becomes gradually popular as it saves huge amount of time and produces reliable and instructive results. These results are derived from many scenarios and show the ‘truth’ in that ‘truth’ is a logical result of a pre-defined program. The corresponding results derived from experiment in the actual breeding program may neglect such ‘truth’ because of noises. Traditional research methods without the assistance of software and statistical model cannot handle the complicated relations between QTL and markers. Hence, building models and conducting *in silico* analysis should be utilised in polyploid breeding studies. However, some biological barriers like

inbreeding depression and self-incompatibility in polyploids are not programmed in simulation and simulation study cannot replace in the field experiment as the experiment produces the real variety. To improve polyploid breeding in the future, testing hypotheses in simulations first provides theoretical basis, and future breeders are able to use the simulation results to guide future breeding schemes.

The major research aim is to discover the selection efficiency of marker assisted selection (MAS) in polyploid breeding based on bi-allelic QTL-marker linkage. In this MSc thesis report, we tried to ultimately compare the potential of marker assisted selection (MAS) versus phenotypic selection through various aspects such as genotypic values and inherited the number of alleles. Actual breeding program involves screening in many generations. So, we first investigated the breakage of QTL-marker linkage through many generations. We examined how this breakage changed with the different distances and phase of QTL-marker linkage without environmental noise. In the next step, we constructed populations with phenotypes to compare MAS and phenotypic selection methods through the fraction of QTL inherited individuals. Finally, we put the comparison under multiple dosage scenarios and evaluate the feasibility of MAS through multiple standards.

Experiment settings & Methods

Software and programs

We carried out the simulations using the software PedigreeSim version 2.0 (Voorrips & Maliepaard, 2012), a Java- program that simulates meiosis in polyploid species. (acquired from http://www.plantbreeding.nl/UK/software_pedigreeSim.html). The software can read input files in which parameters like population size, ploidy levels, and genetic mapping functions are pre-defined. Researchers can specify ploidy levels, the number and length of chromosomes and centromere positions for parents in the initial files. The simulation is started by entering instructions in a system command prompt (Voorrips & Maliepaard, 2012). Some advanced settings including multivalent forming ratios can also be introduced. However, in our simulations these parameters were neglected.

However, PedigreeSim itself is not able to simulate selection and breeding. Therefore, scripts of R version 3.5.2 (R Core Team, 2018) were written to instruct the software to simulate selection processes as well as data analysis. All the selection process written in R scripts did not interfere settings during PedigreeSim running. These commend lines only make changes in input and gather data from output files.

Environment settings and parameters

The choice of simulation parameters for PedigreeSim followed two principles: to be representative and non-redundant. The settings should be close to real world polyploids and simple at the same time.

First, multivalent formation and preferential pairing were not considered in the simulations. This is not only in order to simplify the situations to autotetraploids, but also since some research has pointed out that their influence is not so significant compared to other aspects such as marker coverage (Bourke et al, 2015). Second, to test whether population size can influence the selection process, three population sizes were chosen based on previous population size used in research (Bourke et al., 2015; Hackett et al., 2013) and actual breeding progeny size. Third, marker distances and linkage phase were decided. The detailed parameters were described in Table 1.

Table 1. Basic parameters for experiment. If not mentioned specifically, all parameters in the following simulation are unchanged.

Ploidy level	4
Population sizes	50, 200, 500
No. of chromosomes	1 or 2
Length of Chromosomes	100cM
Multivalent ratio	0%
Centromere position	50cM
QTL-Marker distance	1, 5 or 10cM
Marker types	Single, two and flanking, two on same side
Preferential pairing within homologous	0%

Research Outline

To tackle the research questions mentioned, three steps of simulation from basic to complicated were designed: First, to determine when QTL will be lost in subsequent rounds of marker assisted selection. Second, to compare selection efficiency between traditional phenotypic selection and MAS. Third, to investigate phenotypic selection and MAS performance under scenarios with different dosage of functional genes and markers. In each step, sub-research scenarios such as the effectiveness of repulsion linkage were also simulated. Every simulation was replicated 100 times to obtain mean values which were close to real value, variance within values and to mitigate random errors.

a. QTL-marker linkage break period

In this section we explored when QTL-linkage broke if selection was only based on marker results. We limited the simulation to at most 50 generations, where each generation resulted from a cross to a parent nulliplex for the marker allele. The parent 1 carried one copy of QTL and marker allele on the homologous (simplex). The parent 2 carried no such alleles (so no copies of the selection allele of the marker).

Experimental design

In this very first step, the simulation was aimed to discover the generation time when QTL were lost if MAS was continuously conducted as well as testing simulated recombination ratios against theoretical predictions. Here we introduced the term 'simplex'. Simplex refers to a polyploid individual carrying one copy of the allele and nulliplex as zero copies of the allele, but in this section, we called parents simplex when having one copy of both the marker and the favourable QTL allele (Figure 1). In the following sections we used SxN (simplex crossed with nulliplex) and DxD (duplex crossed with duplex) to specify such cross events. For each allele, a SxN by expectation produces 50% individuals that carry the allele and 50% that do not (Figure 2).

In the first generation, parents were crossed using PedigreeSim to create progeny populations of sizes 50, 200, and 500 individuals. Then, an R script was used to read the output dosage files and select all progeny that carried single or double markers. Double markers included flanking markers and same side markers. 'Flanking markers' means a situation where there are markers at both sides of the QTL. 'Same side markers' means that both markers were at the same side of a QTL allele. Individuals that carried double markers were selected only when both markers were present. Flanking markers had little recombination with QTL. If the marker-marker genotype is non-recombinant, the probability of a recombination between the first marker and second marker with the QTL, given the observed non-recombinant marker-marker genotype, is very small, especially when QTL-marker distance is very small. Same side markers were simulated because some same side markers were mistakenly ordered on maps in research and were treated as flanking markers. Thus, such a mistake brought disastrous consequence for breeding. From these selected progenies one individual was randomly selected to become the parent for the next generation and crossed with the other nulliplex parent. If the selected parent was found to have lost the QTL, then the simulation was aborted, and its generation number recorded. At most 50 generations were allowed, to prevent extremely long runs (in this simulation, we were most interested in determining conditions under which a marker-QTL association is lost). In each round of simulation, the number of recombinants and their genotypes were recorded. The process was inspired by a backcross breeding scheme.

Marker distance, types and linkage with QTL

The markers simulated were bi-allelic single nucleotide polymorphism (SNP) markers. The distance and linkage between QTL and markers were certain. We neglected any possibility of insertion and deletion events (or uncertainty about the distance). The QTL regions that carried one copy of QTL and marker alleles were represented by number 1 in PedigreeSim. The alternative alleles in QTL region were considered as non-functional and represented by number 0. The QTL-marker genetic distances were 10cM, 5cM and 1cM. Each QTL had symmetric markers both sides on the same chromosome. Thus, markers types were 1) single, 2) two, flanking and 3) two, same side. Markers and QTL could be linked in coupling phase or in repulsion phase. However, having noticed that repulsion linkage markers were not efficient for detecting QTL in previous studies (Wu et al., 1992; Ripol, 1999), only single marker linked in repulsion phase with QTL was simulated (Figures 1, 2).

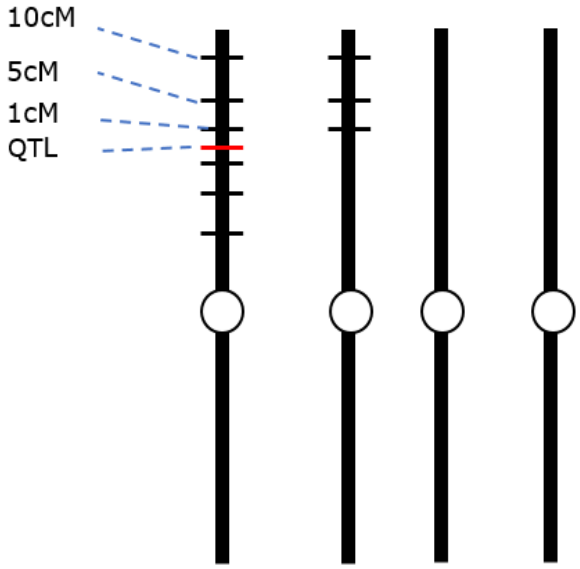


Figure 1. Representation of QTL-marker linkage in an autotetraploid (h=4). This is an example of an autotetraploid situation with four homologous chromosomes. The chromosomes (black lines) are 100cM long and the centromeres (open circles) are situated at 50cM at the center of the chromosome. The red dash representing the QTL position is at 30cM away from the centromeres. Black dashes on the chromosomes represent marker alleles and are located on both sides of the QTL, symmetrically. The QTL-marker distances are indicated. At the second chromosome repulsion phase linkage markers are indicated.

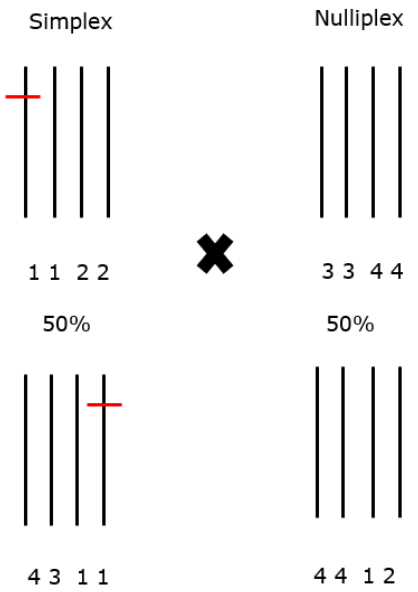


Figure 2. SxN theoretical results. The initial parent 1 carries one allele and parent 2 does not. This cross produce, by expectation 50% progeny individuals that carry the allele and 50% that do not. Numbers below represents the homologous chromosomes from different parents. Since it is autotetraploid case and pairing among the chromosomes are random, the offspring receives two homologous chromosomes from parent 1 (1 and 2) and two from parent 2 (3 and 4).The position of the allele on which homolog may change due to recombination. The ratio of nulliplex(0) and simplex(1) is 1:1.

Repulsion linkage

Previous research has suggested that repulsion linkage QTL gives far less information than coupling linkage (Ripol, 1999). Thus, for repulsion phase we only used single QTL-marker linkage in our simulations. Testing repulsion linkage marker was aimed to confirm repulsion markers were not useful in MAS.

b. Comparison of the efficiency of phenotypic selection and MAS (SxN)

MAS largely boosted the selection accuracy on certain QTL. However, in the reality breeders must consider trade-offs between selection efficiency and cost of selection. In this step of simulation, MAS was compared with traditional phenotypic selection in an autotetraploid bi-parental cross with parents that were simplex, and nulliplex, respectively (SxN) for both the marker and the QTL. The offspring population size was 500.

Scenario settings

This simulation included two scenarios that one individual contained either one QTL or two QTL on non-homologues chromosomes (i.e. unlinked). Three variables were QTL effect size, environmental variance (EV) and QTL-marker distance. The QTL effect size governed the magnitude of additive effects over the basic trait mean value μ. The

environmental variance (EV) was represented by the standard deviation of a randomly generated normal distribution. When we input EV, R randomly generated a list of 500 number from normal distribution based on every EV. These numbers were treated as environment effects(E) and added later to construct phenotypes. The QTL effect size changed from 1 to 10 and for each given QTL effect size and EV changed from 1 to 10 as well. To simplify the comparison, Two QTL had the same effect size in the second scenario when effect sizes changed. Single markers at distance 10cM, 5cM, and 1cM were chosen to simulate MAS in both scenarios.

Simulation on marker assisted selection and phenotypic selection

For every combination of QTL effect size and EV, 100 simulation runs were executed. In each simulation both phenotypic selection and MAS were performed. Phenotypic selection was based on phenotypic value of every offspring. We selected 10% individuals that had highest phenotypic value. How phenotype constructed was shown below:

$$\text{Genotypic value(G)} = \mu + \text{QTL effect size}$$
$$\text{Phenotypic value(G)} = \text{Genotypic value(G)} + \text{Environmental effect (E)}$$

After phenotypic selection, MAS was implemented based on whether marker was carried by an individual. In MAS we selected the same number of individuals as in phenotypic selection. Considering that approximately 50% of the offspring had the markers yet the selected offspring consisted of only 10%, therefore we randomly sampled from the marker carrying offspring the same number of selected individuals as was done for phenotypic selection. Both approaches did not change the total population size. Individuals which were chosen by phenotypic selection could also be selected by MAS. In the end we acquired 50 selected individuals for MAS and phenotypic selection. The selection efficiency was evaluated by the fraction of selected individuals that carried the QTL. The efficiency was plotted to discover the difference between MAS and phenotypic selection. In the end, we simulated single, flanking, and same side markers for 1, 2 and 5 QTL to discover how MAS efficiency changes as the number of QTLs increases.

c. Selection efficiency comparison under duplex dosage levels (DxD)

In the final step we performed a cross between duplex parents (DxD). In a set of homologous chromosomes, two chromomeres carried one allele at the same position, respectively. The copy numbers of QTL and linked markers of both parents were in double dose and represented by number 2. In total we tested two independent QTL, QTL1 and QTL2. In an autotetraploid, this cross provided all marker dosage situations (0~4) in the offspring. This section was aimed to compare MAS efficiency with traditional phenotypic selection in polyploid breeding by various standards including genotypic values, QTL copy numbers and the fraction of highest QTL dosage carrying individuals. Dosage levels were

very important features. Except of dosage levels changed, other settings followed the previous experiment. The QTL effect on trait value was simple additive, which meant there was no interaction between different alleles. Besides, we did not consider negative effect QTL.

Dosage simulation

In this simulation experiment, both QTL and markers were represented by numbers 0,1,2,3,4 for copy numbers in an individual and the alternative alleles were always nonfunctional. By assigning marker dosage with simple numbers simplified our questions and calculation process. DxD produced a population with segregating patterns at 1:8:18: 8:1 corresponding to dosage 0,1,2,3,4.

Scenarios settings

In this scenario the additive effect was the putative effect of QTL. There were 3 parameters altered in the simulation: QTL-marker distance, QTL effect size for QTL1 and QTL2, and environmental variance (EV).

We tested two QTL-marker distances, 1 cM and 10 cM. For QTL effect size we assigned 3 combinations for single QTL effect size(sQef): sQef1 >> sQef2, sQef1 > sQef2, sQef1 = sQef2. Because both QTL1 and 2 acted at the same, there was no necessity to test a symmetric scenario like sQef1 < sQef2. Since in this section we only considered simple additive effect, so the total QTL effect size was the sum of both QTL's single allele effect multiplied by its copy numbers. For environmental effects, three levels were set: EV was equal to smaller sQef, EV was equal to larger sQef, and EV was much larger than both sQef. EV was represented by standard deviation (sd) in the normal distribution which added to individual genotype to simulate phenotypic scores. The phenotypic scores were regarded as continuous quantitative traits such as plant height. Thus, we presented the following formula to construct genotype and phenotype in our simulations:

$$\text{Genotypic value(G)} = \mu + \text{sQef1}^* \text{ QTL1 dosage} + \text{sQef2} * \text{QTL2 dosage}$$
$$\text{Phenotypic value(G)} = \text{Genotypic value(G)} + \text{Environmental effect(E)}$$

The population's genotype and phenotype values were recorded to calculate variances. Based on these variances, the heritability (h^2) was calculated:

$$H^2 = \frac{\mathrm{Var}(G)}{\mathrm{Var}(P)}$$

Heritability acted as an important and clear indicating factors for selection standards. Out of 18 simulated scenarios that had different heritability, we chose 6 scenarios in Table 2 to compare multiple aspects about MAS in more detail.

Table 2. DxD cross settings for phenotypic selection and MAS analysis.

distance marker1 (cM)	distance marker2 (cM)	sQef1	sQef2	EV
10	10	10	1	1
10	10	10	1	10
10	10	10	1	20
1	1	10	1	1
1	1	10	1	10
1	1	10	1	20

marker 1 and 2 are linked with QTL1 and QTL2, respectively. 'sQef' means single QTL effect size. For QTL1 each dose effect is 10 and for QTL is 1. EV represents environment variance and is related to standard deviation of random normal distribution. EV is only an input parameter and not equal to environment effect that formed individuals' phenotype.

Simulation on marker assisted selection and phenotypic selection

After setting up the scenarios, R scripts were written to conduct the simulations with respect to the crosses and the selection process. For each scenario, 100 replicate simulations were performed. For each cross, we first performed phenotypic selection. 10% selection strength was implemented based on phenotypic values. The 10% individuals with the highest phenotypic values were selected, and their corresponding genotypic values and QTL copy numbers were extracted to a single file. Then, the MAS process was simulated by calculating the weighted marker scores based on the given QTL effect size (so assumed to be known or correctly estimated). The marker linked to the QTL with larger effect size was given a higher weight. The formula was:

$$\text{Marker1 weight} = \text{sQef1}/ (\text{sQef1} + \text{sQef2})$$
$$\text{Marker2 weight} = \text{sQef2}/ (\text{sQef1} + \text{sQef2})$$

So, the marker scores for an individual were:

$$\text{Marker scores} = \text{Marker1 weight} * \text{dosage} + \text{Marker2 weight}* \text{dosage}$$

After the weighted marker scores were calculated, we sampled the same number of individuals as were previously selected by phenotypic selection using these marker scores. In the MAS approach, the individuals with the highest weighted marker scores were selected. Genotypic values and QTL copy numbers were also collected for later

analysis. Both selection processes were independently carried out on the same simulations, meaning that the same individuals could be selected by both methods.

After getting data through phenotypic selection and MAS, genotypic values and inherited QTL numbers were two major measures to compare the two selection procedures.

Apart from weighted marker score method, an alternative strategy to execute MAS was to perform unweighted selection:

Marker scores = marker 1 highest dosage + marker 2 highest dosage

Via second method the simulation produced the same format results as the first method did. This method was meant to test whether different decision for MAS changed its performance against phenotypic selection (and against weighted selection).

Results

QTL-marker linkage break period in relations to recombination ratios

The Haldane mapping function is to correct for multiple crossover events between two genes (Haldane, 1930). So, we can calculate theoretical recombination ratio(r) based on Morgan genetic distance(d). The equation is:

$$d = -\ln(1 - 2r)$$

$$r = 0.5*(1-e^{(-2d)})$$

So, when d = 1cM,

$$r=0.5*(1-e^{(-2*0.01)}) = 0.00990$$

Then theoretical results from distance 1cM to 10cM were calculated. We first verified that the recombination ratios generated by PedigreeSim were consistent with theoretical expectations (Appendix 1). The simulated results from population sizes 50 and 200 were also much the same to theoretical results. But smaller population size produced more random error than a larger one. The following experiments only applied 500 as default population size. Although using much larger population size such as 50000 further lowers random error and gives more credibility to simulation, it is very time consuming to compute such a large population and unrealistic for breeding process, while results are not very different.

The recombination ratio of both flanking QTL-marker linkage was equal to each QTL-marker recombination ratio multiplies (Table 3). The theoretical results showed that flanking markers were very reliable at ensuring a QTL was selected. If the future experiment is aimed to test these double recombination ratios, the progeny population must be much larger.

Table 3. Theoretical recombination ratios and probability of selection of a QTL using flanking markers in a single generation

Flanking marker	r_1*r_2	QTL selected ratio(1-*r*)
10 and 10cM	0.008208	0.9918
5 and 10cM	0.004312	0.9957
5 and 5cM	0.002265	0.9977
1 and 5cM	0.000471	0.9995
1 and 1cM	0.000098	0.9999

The first column refers to double markers distance in relations to QTL position. The second column refers to flanking QTL-marker linkage recombination ratio. The third column is the success rate selecting offspring carrying QTL.

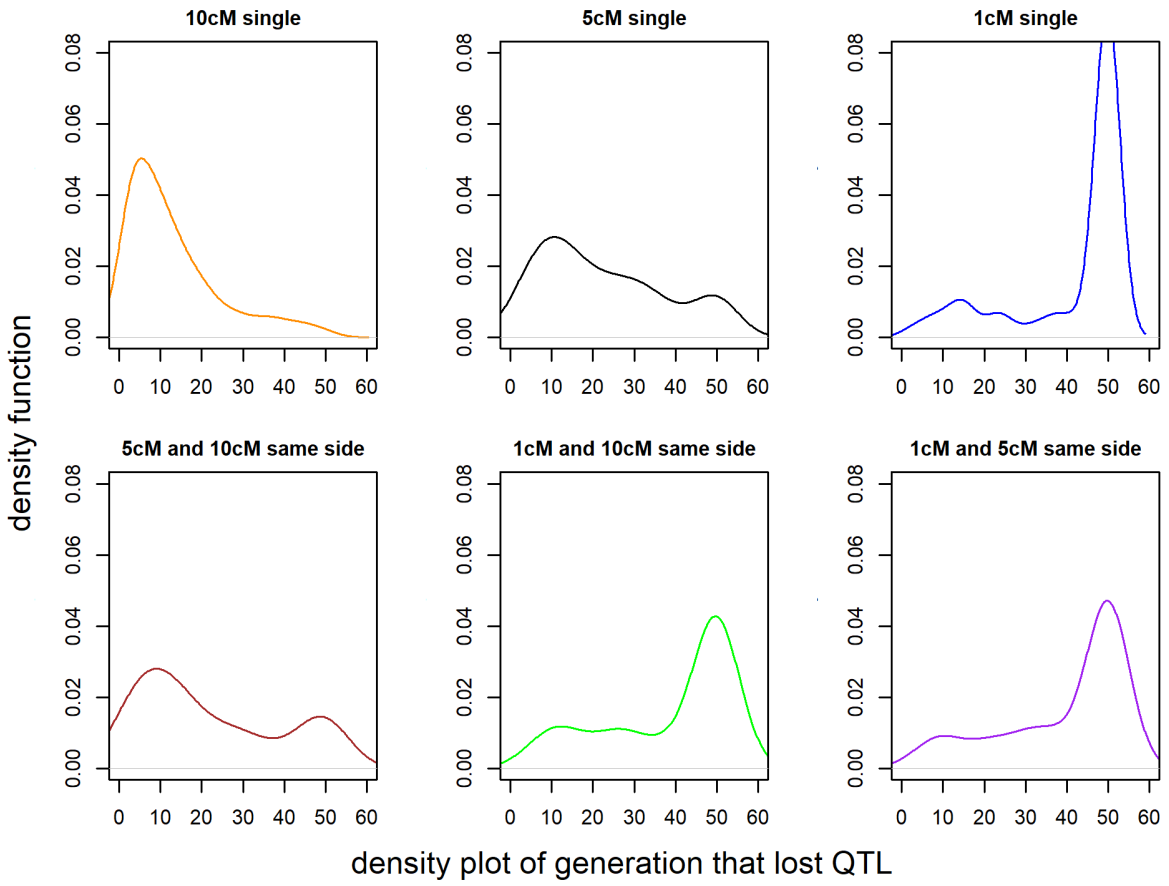
Same side markers theoretically functioned as single markers. Only the closer QTL-marker linkage influenced the recombination ratios. However, we still needed to validate this in the simulation.

Next, we analyzed the distributions and occurrences of QTL-marker linkage break period. Although in each round selection on next generation parent was random, we found these distributions are not random. As replication repeated after 100 times, the distributions of when the linkage breaks had revealed patterns in relations to marker distance and types.

As the QTL-marker distance decreased, QTL-marker linkage break events took place in later generations and the frequency of linkage break events decreased. The 10cM single marker linkage showed that most of the break events happened before the 10^{th} generation while for 1cM the break events happened near the 50^{th} generation. The graph also showed that same side markers show a similar generation break distribution as the single marker. This showed that in QTL-same side markers linkage, only the closest linked marker affects the selection results and the more distant one does not show any extra influence. For example, the 5 and 10cM same side markers had similar distribution as the 5cM single marker (Figure 4). However, the approach with two same side markers performed much worse than the corresponding situation with flanking markers. The 5 and 10cM same side markers manifested very different patterns compared to 5 and 10cM flanking markers.

We were curious about whether flanking markers and same side markers perform differently compared to single marker that had similar recombination ratio. Multiple comparisons have revealed 10 and 10cM flanking markers, 1 and 10cM same side and 1 and 5cM same side markers show no significant difference as to 1cM single marker. But 5 and 5cM,1 and 10cM, and 1 and 1cM flanking markers performed significantly better than 1cM single marker. Not only the linkage break period happened in later generations, but also the frequency of such events in 100 times run occurred less. For 1 and 1cM flanking marker case no break events ever happened in all replications. The density shifted from around 10^{th} generation to 50^{th} generation as the QTL-marker distance changed from 10cM to flanking 1 and 1cM case (Figure 3). In other words, the QTL was less likely to be lost even after 50 rounds of crossing. Hence, as QTL-marker linkage recombination ratios dropped, the generation of QTL-marker linkage break was postponed as well.

However, these generation break events only show trends as to which distance markers performed better during selection. Meticulous relations between these break events distribution and marker distance cannot be determined using simple mathematics calculation. Thus, simulation provided trends rather than absolute equations.


10cM single
5cM single
1cM single
5cM and 10cM same side
1cM and 10cM same side
1cM and 5cM same side
density function
0.00
0.02
0.04
0.06
0.08
0
10
20
30
40
50
60
density plot of generation that lost QTL

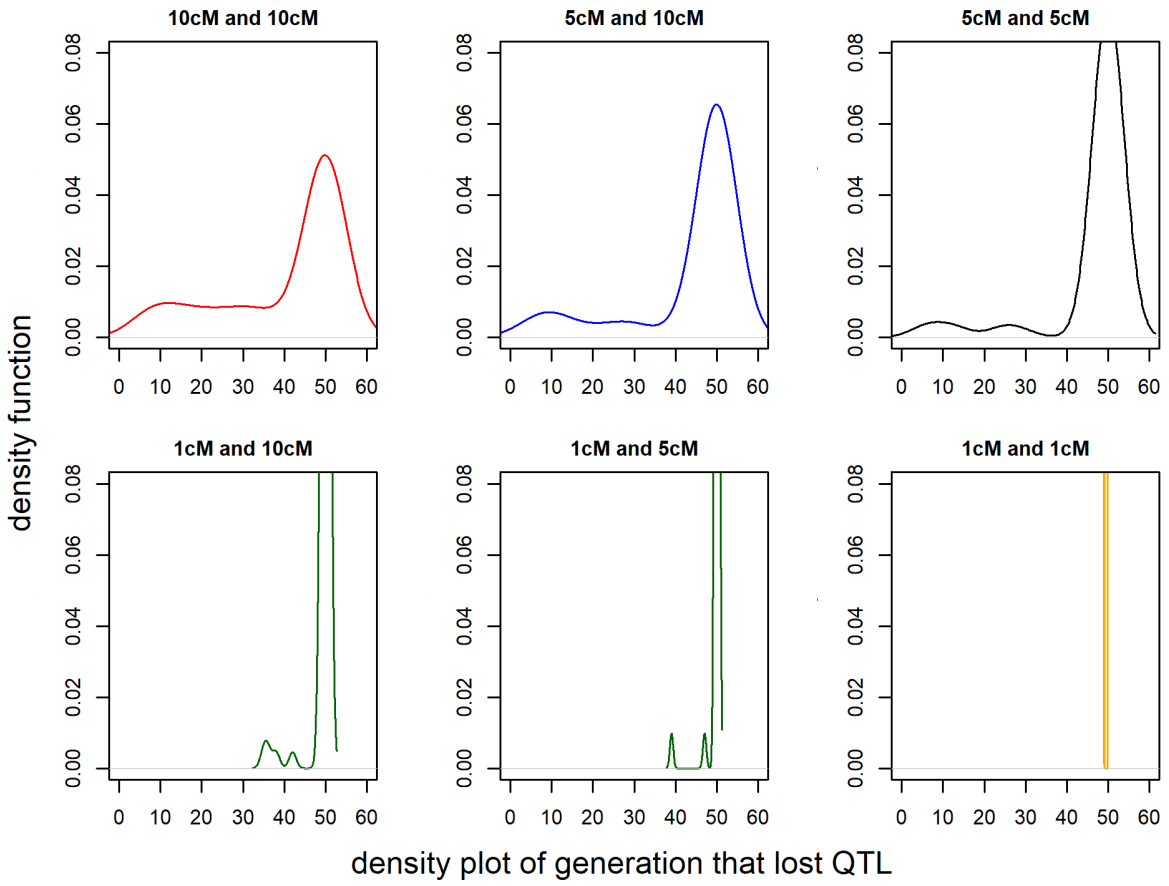


Figure 3. Density plot of QTL-marker linkage break period applying single markers, same side markers, and flanking markers in the selection process. The generation limit is 50. In the event of no breakage in linkage between marker(s) and QTL, the number of generations of maintained linkage was recorded as 50.

Repulsion linkage marker simulations

Theoretically, SxN provided the same result as the TxQ (3x4) SxQ (1x4) and TxN (3x0) because these crosses can be treated as selecting on alternative alleles. We tested the TxQ for 10cM, 5cM and 1cM single marker in repulsion linkage. The TxQ results showed the same recombination ratios as the SxN. Its QTL-marker linkage break period had further indicated that TxQ was fundamentally the same as SxN (Figure 4).

However, the repulsion linkage SxN produced very different results. According to Wu et al. (1992), if an individual with ploidy level h and carries only single dose two alleles A and B in repulsion phase linkage, then the expected frequency of its gametes that carry both AB alleles f(AB) is:

$$f(AB) = 0.25*(1-w) + 0.5w*r$$

w=1/(h-1), h = number of homologous chromosomes.

While the expected frequency of gametes that carry only A f(A-) or B f(B-) is:

$$0.25*(1-w) + 0.5w*(1-r)$$

Because in a SxN cross, nulliplex parent carried no allele A or B and act as a blank background, different offspring frequencies were equal to the simplex parent's gamete frequencies. In an autotetraploid, h = 4. Supposing the recombination ratio is 0 for both coupling and repulsion linkage, the fraction of coupling and repulsion AB linkage is 0.5 and 0.25*(1-⅓) =⅙, respectively. If we set A as marker and B as a QTL, the goal of MAS in coupling phase was to select f(AB). While as to repulsion phase linkage it was quite different. There were two possible ideas to carry out MAS. One way was to select progeny based on marker presence. In such a case the recombinants that carry QTL and marker only took a very small fraction of marker carrying individuals. Suppose no recombination happened, then QTL-marker f(AB) type only takes ⅓ in all markers fraction f(A-)+f(AB):

$$f(AB) = 0.25*(1-⅓) +0.5*⅓*0 =$$
$$f(A-) + f(AB)= ⅙ + 0.25*(1-⅓) + 0.5*1/3 *(1-0) = ½$$

So, if 10cM, 5cM, and 1cM markers were in repulsion phase linked with QTL, the QTL-marker fractions were:

$$f(10cM\ AB) =0.25*(1-⅓) + 0.5*0.0906*⅓ = 0.182$$
$$f(5cM\ AB) =0.25*(1-⅓) + 0.5*0.0476*⅓ = 0.175$$
$$f(1cM\ AB) =0.25*(1-⅓) + 0.5*0.0099*⅓ = 0.168$$

This suggests that to select QTL based on a repulsion linked marker is very risky, unless large amount of progeny was generated. Distantly linked markers in repulsion phase counter-intuitively provided better results (Although the frequency does not increase much.). Simulated success rates to select for a QTL based on marker presence for 10cM, 5cM and 1cM repulsion linked marker were 0.362, 0.351 and 0.337. Simulation had proven this hypothesis since the simulated results are very close to theoretical prediction.

Another way is to 'deselect' a marker, by excluding individuals that carry markers. This method had approximately ⅔ probability to select QTL carrying progeny. However, neither ways can provide accurate selection results. With higher ploidy, *w* becomes smaller. The MAS accuracy gradually approached 50%, which was not even better than a random guess. Thus, there was no vital reasons to apply repulsion linkage marker for breeding. Hence, simulation results have demonstrated the futility of applying repulsion-phased markers in polyploid breeding.

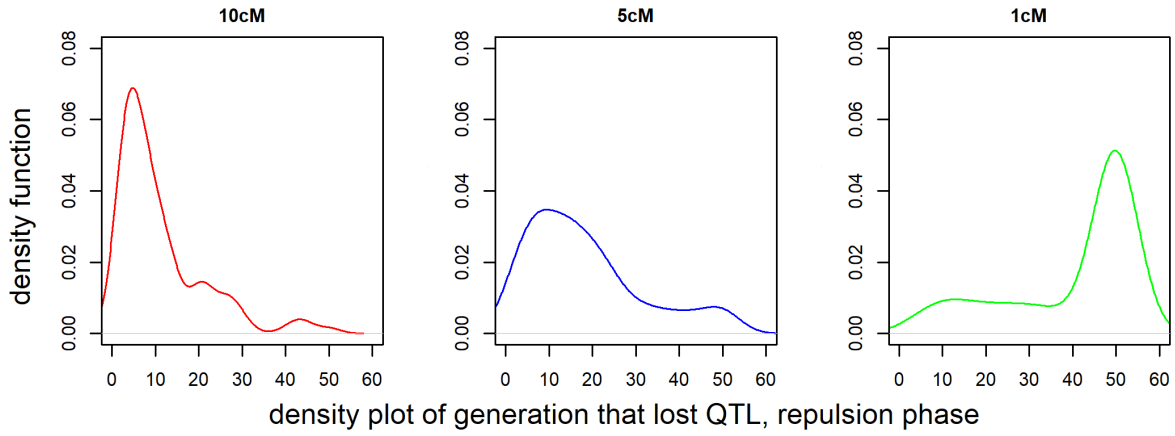


Figure 4. Density plot of QTL-marker linkage break period applying single markers in the selection process, repulsion phase. The generation limit is 50. In the event of no breakage in linkage between marker(s) and QTL, the number of generations of maintained linkage was recorded as 50.

Multiple QTL-marker linkage break period

In the end of the first section, we simulated two unlinked QTL using setting mentioned above. These results confirmed conclusions derived from previous single QTL-marker linkage results. With more closely linked markers the frequency of lost QTL events decreases, and the linkage break generation is higher.

Table 4. the QTL linked with more distant marker contributes more to the QTL-marker linkage break period.

	QTL1 lost*	QTL2 lost
Q1:10cM flanking, Q2:10cM flanking	36	25
Q1:10cM flanking, Q2:10cM single	11	88
Q1:10cM flanking, Q2:1cM single	32	26
Q1:10cM single, Q2:1cM single	91	11
Q1:10cM single, Q2:10cM single	65	37
Q1:1cM flanking, Q2:10cM flanking	0	40

*: occurrence of QTL lost in 100 times replication

The occurrence of QTL-marker linkage break period here means only the counting results of such event rather than which generation. Flanking markers are situated both sides of the QTL.

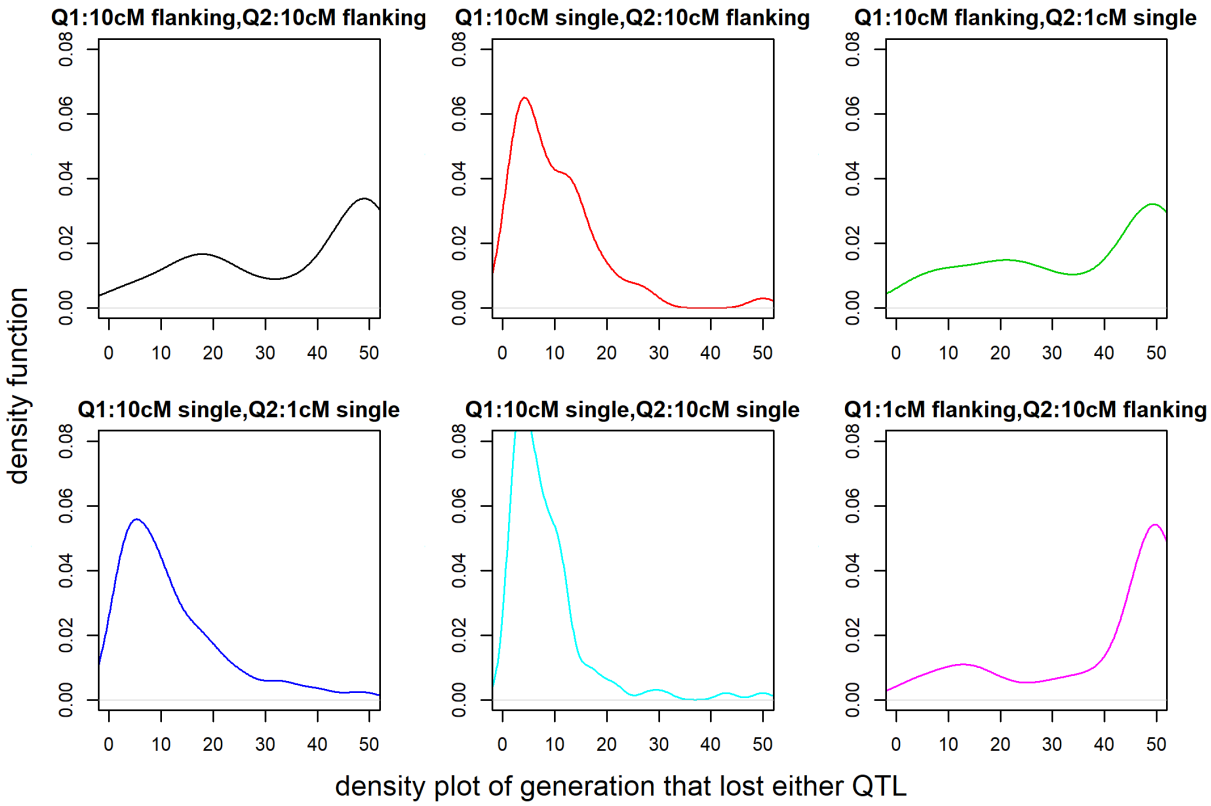


Figure 5. **Density plot of either two QTL-marker linkage break period applying single markers and flanking markers in the selection process.** The generation setting is the same as Figure 3&4.

Phenotypic value distribution

In the second step of simulation, we compared distributions of progeny phenotype values across different simulation scenarios. These distributions change drastically when different QTL effect sizes and EV are given. When EV is small compared to QTL effect size, the distribution shows distinct peaks with different phenotypic values. These distinct peaks are very close to genotypic values. But when EV is very large compared to QTL effect size, the distribution is very close to a normal distribution and phenotypic selection targets become very unclear. This graph has indicated phenotypic selection becomes inefficient as the environmental influence grows.

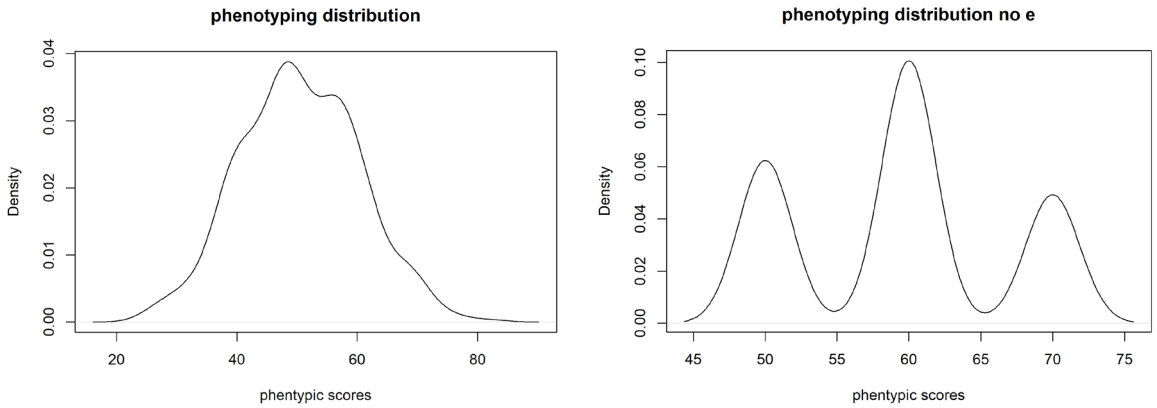


Figure 6. **Phenotypic value distribution as a result of SxN progeny population that carry two independent QTL.** The QTL are both additive. Graph A is when QTL total effect size = 2 and EV =10. Graph B is when QTL total effect size = 20 and EV = 0.

MAS and phenotypic selection efficiency comparison

The phenotypic selection efficiency is strongly determined by QTL effect size and EV. On the one hand, as the EV increases when QTL effect size is small, the selection efficiency drops sharply at first and gradually slows down. When certain trait is affected by environment strongly, phenotypic selection usually fails to select the offspring that inherit the gene (here we neglect GxE effect). On the other hand, when the QTL effect size grows relative to environmental effects, phenotypic selection efficiency would gradually increase and eventually reaches 100%.

However, MAS is independent from QTL effect size and environmental effect. We found MAS efficiency was only determined by QTL-marker distance and the number of QTL. In the second graph 10cM, 5cM and 1cM marker selection efficiency is closed to 0.82, 0.91, and 0.98. These results are very close to the mathematical expectation $(1-r)^{\#QTL)}$. The comparison plot has suggested two MAS characters which were both consistent with expectations. At least 80% of scenarios showed phenotypic selection had a lower efficiency than 10cM MAS (Figure 7). But how to choose between MAS and phenotypic selection under different QTL effect size and EV combinations requires further investigation.

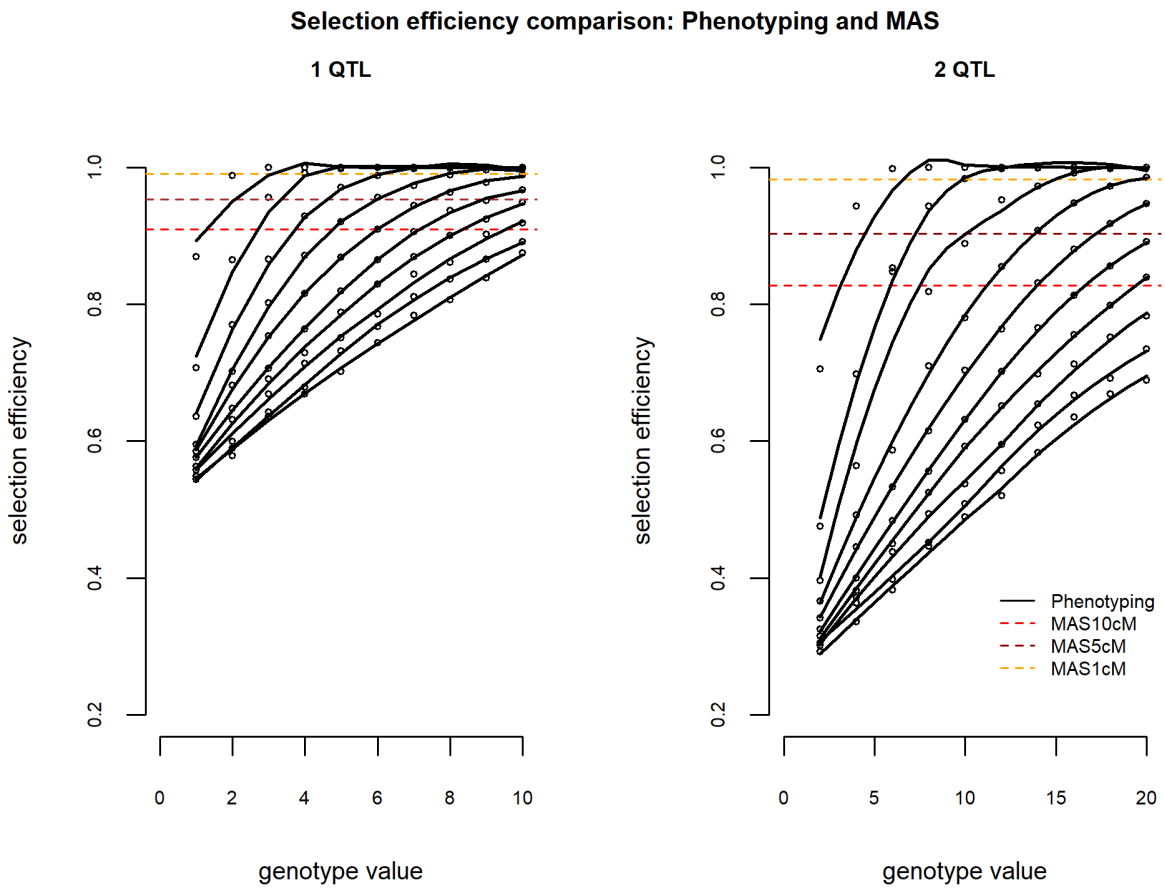


Figure 7. Comparison between different MAS and phenotypic selection about selection efficiency. The first graph contains single QTL. The second graph contains two QTL and both QTL have the same QTL effect size. The lines are created by loess regression using selection efficiency against genotypic value with changing EV from 1 to 10 and span=0.95. The highest lines to lowest lines represent EV =1 and EV =10. The horizontal dash lines represent mean values of MAS efficiency. When EV is low, the selection efficiency through phenotypic selection greatly increases with rising genotypic values and then reaches a plateau near 100%. When EV is large, its selection efficiency grows gradually but much lower than MAS efficiency. 80% of phenotypic selection efficiency is lower than 10cM MAS.

MAS efficiency in relation to number of QTL and QTL-marker linkage

We also simulated the theoretical selection efficiency with increasing numbers of QTLs. The average number of that contributed to certain trait ranged from 1 to 10 or even more. If all QTL were linked to markers in coupling phase, the theoretical selection efficiency should be $(1-r_1)*(1-r_2)*\ldots(1-r_n)$ in a SxN cross. The results were basically consistent with theoretical predictions.

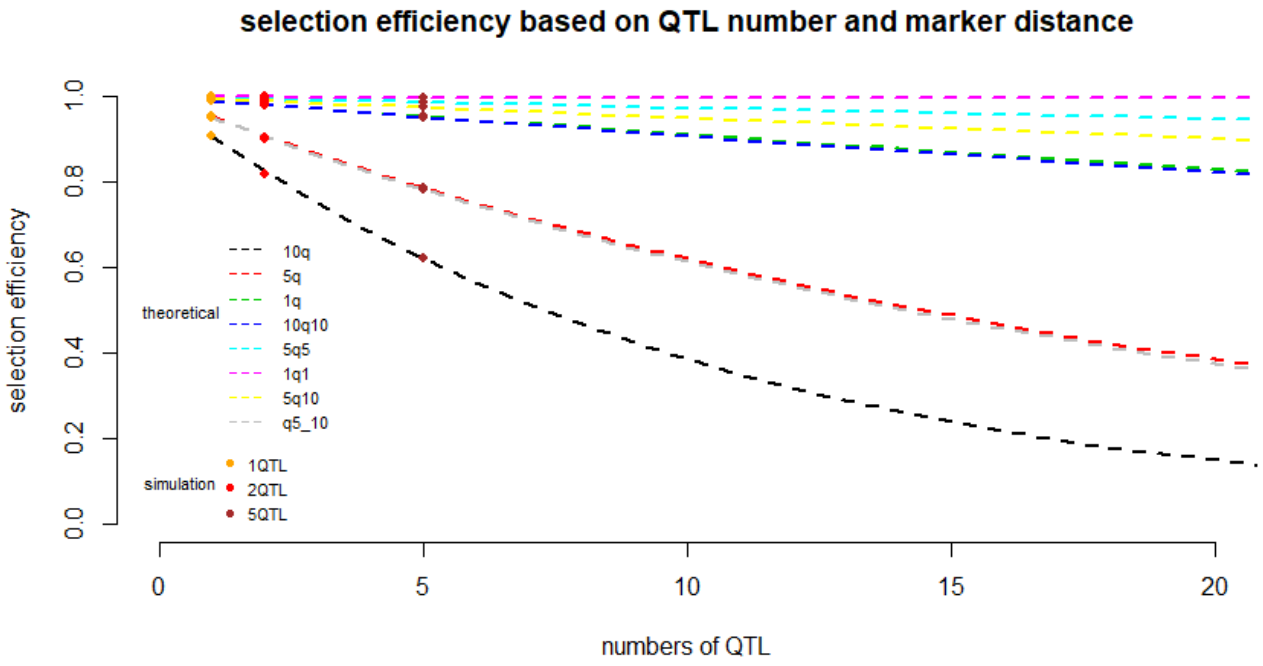


Figure 8. Theoretical predictions on selection efficiency of MAS under different numbers of QTL. '10q' means QTL-marker distance is 10cM. '10q10' means flanking QTL-marker distances are 10 and 10cM. 'q5_10' means same side QTL-marker distances are 5 and 10cM. The dots represent selection efficiency summarized by MAS in the simulation. These dots are fully corresponded to theoretical prediction.

The phenotype value ranges in a DxD cross

Based on previous defined phenotype equation, theoretically, the highest genotypic value was:

$$\mu + sQef1^* \text{ maximum dosage} + sQef2 * \text{maximum dosage} =$$
$$50 + 10*4 + 1*4 = 94$$

For the lowest genotypic value:

$$\mu + sQef1^* \text{ minimum dosage} + sQef2 * \text{minimum dosage} =$$
$$50 + 0*10 + 0*1 = 50$$

In order to calculate the theoretical highest and lowest phenotypic values in such a population, random normal distributions applying the same EV settings in simulation were generated. Each distribution generated a list of 500(same as population size) values. The calculation took 1% head and tail values from these normal distributions as the highest and lowest environmental effect (E) to form phenotypes Each highest and lowest values were the mean value of 100 times replication (Table 5).

Table 5. normal distribution and its environmental effect.

Settings	highest	lowest
N=500, μ=0, sd=1	2.70	-2.62
N=500, μ=0, sd=10	26.83	-26.07
N=500, μ=0, sd=20	52.87	-52.59

So, the highest theoretical phenotypic value was highest(G) + 1% highest (E):

$$94 + 2.70 = 96.7$$
$$94 + 26.83 = 120.83$$
$$94 + 52.87 = 146.87$$

The lowest theoretical phenotypic value was lowest(G) +1% lowest(E):

$$50 + (-2.62) = 47.38$$
$$50 + (-26.07) = 23.93$$
$$50 + (-52.59) \approx 0$$

(Phenotypic value cannot be negative)

Theoretical mean of genotypic value was the expectation of:

$$\mu + (sQef1+sQef2) * \text{mean dosage} =$$
$$50+ 2*10 + 1*2 = 72$$

The simulation result was very close to the theoretical phenotypic value range. As the heritability grew, the range of phenotype distribution became smaller and gradually close to genotype range. However, the simulated ranges were narrower than theoretical ones due to the rareness of extreme cases in simulation (Figure 9). The variation of phenotype values can be used as an indication to whether applies MAS in the future breeding process.

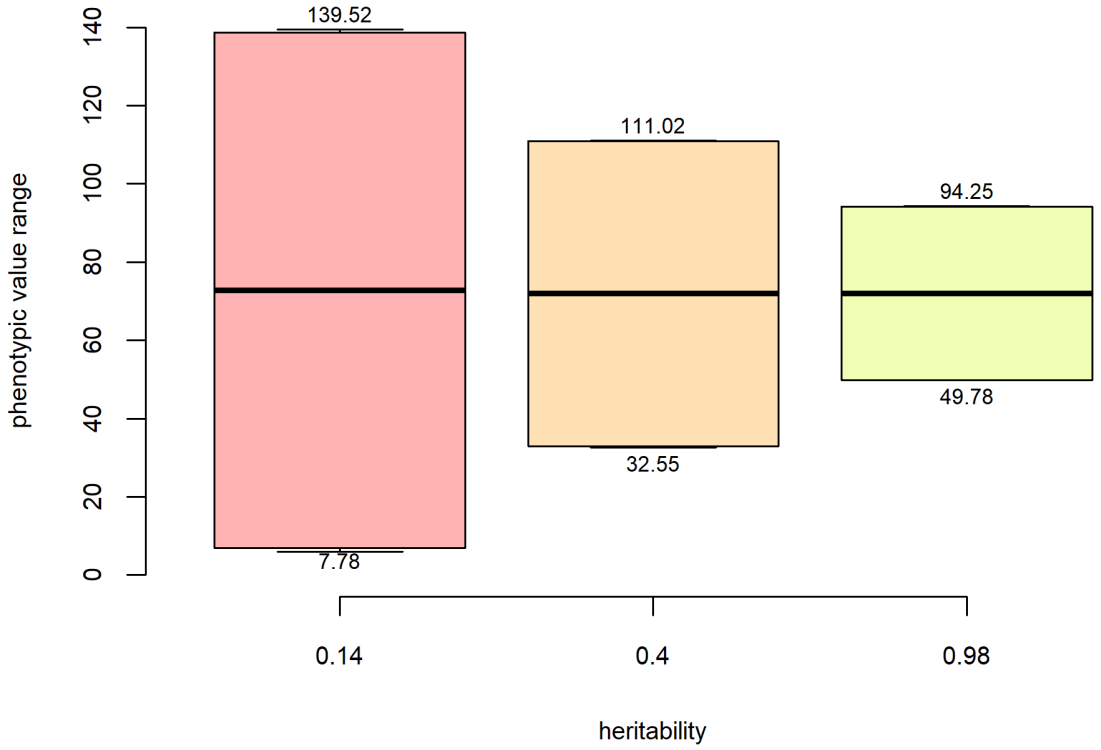


Figure 9. Simulation results of phenotype range. Heritability is calculated by var(G)/var(P). The means of phenotypes under 3 heritability are 72. As heritability increases, the phenotype range begins to shrink. This is ideal situation which genotypic variance is known for certain whereas in breeding program genotypic variance is hard to confirm.

Phenotypic selection and MAS efficiency comparison

In the previous section selection efficiency was defined as the fraction of QTL-inheriting progeny. It is a True or False standard (can be treated as 1 dosage additive effect or dominant effect). In this section, selection efficiency was evaluated by multiple standards, including genotypic values, inherited QTL copy numbers, and the proportion of individuals that carried highest QTL copy number. We discovered that MAS performs significantly better than phenotypic selection when the trait heritability is low. Furthermore, as the heritability became lower, the difference between MAS and phenotypic selection soared drastically as well (Figure 10).

We found the heritability is a vital indicator for whether MAS can outperform phenotypic selection. In a control group experiment where the EV was 0 and both QTL effect sizes

were quite large, the 10cM marker MAS had lower selection efficiency than phenotypic selection in terms of comparing genotypic value. This was a logical result of an ideal situation where phenotypic selection selected all QTL if there was no environmental effect on the trait. However, MAS loses its advantage because of recombination between marker and QTL.

After comparing genotypic values difference, we realized that more details and diverse aspects were needed to evaluate selection efficiency. The QTL copy number could be compared between MAS and phenotypic selection as another evaluation standard. The QTL with larger effect size is more likely selected. However, if the marker distance is very small and the EV is very large, then copy numbers of both independent QTL are selected more by MAS than phenotypic selection. This suggests again QTL-marker distance in coupling phase is essential to the efficiency of MAS.

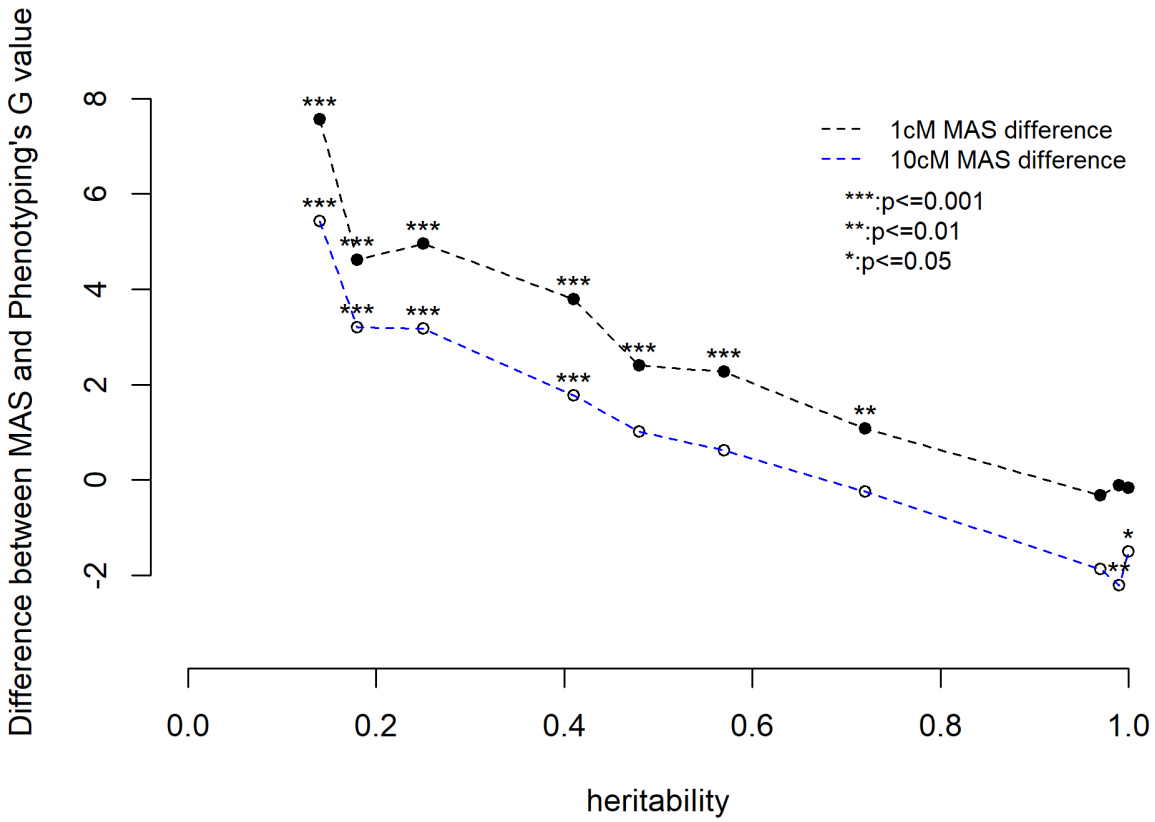


Figure 10. Mean genotype difference between 10cM, 1cM MAS and phenotypic selection selected individuals under various heritability. Significance is mean of p value deriving from 100 times t test results. These t tests compared between selected individual genotypic values between MAS and phenotypic selection.

First, under the same scenario settings, the 1cM MAS was always more effective than phenotypic selection, while this was not always the case for 10cM. Second, when both 10cM and 1cM MAS were more effective than phenotypic selection, the smaller distance marker selected higher genotypic values than the larger distance markers did. The test results showed that in all scenarios, the 1cM MAS produced significantly higher genotypic values than 10cM MAS. The performance between different markers MAS was exhibited in the graph directly (Figure 10).

Weighted MAS comparison with phenotypic selection

As we look more into other standards, we found MAS performed better than phenotypic selection in these standards if the heritability was not high. First, MAS manifested very stable results in terms of genotypic value and QTL copy numbers(Table 6 & Figure 11). These results were consistent with the second section result that MAS is independent from environment effect and QTL effect size. Even when heritability was not very low, MAS still had advantage of selecting more minor effect size QTL than phenotypic selection (Table 6). When EV=10, we found although 10cM MAS did not outperform phenotypic selection via genotypic value, MAS selected significantly more copies of QTL2. Second, 1cM MAS produced higher average genotypic values than 10cM. This improvement was largely determined by the effect size of the QTL to which a marker linked. The weighted marker method was powerful for selecting overall genotypes. But it possessed the risk of losing minor effect QTL in the subsequent selection process.

Table 6. Comparison between weighted marker score MAS and phenotypic selection via genotypic values and the copy number of QTL, based on a DxD cross.

Scenarios	h^2	No. of QTL1		No. of QTL2		G value	
Marker=10cM Effect,Q1=10,Q2=1		Pheno	MAS	Pheno	MAS	Pheno	MAS
EV=1	0.99	3.29*	3.06	2.4	2.5	85.31*	83.1
EV=10	0.41	2.91	3.05	2.11	2.52**	81.21	82.98
EV=20	0.14	2.55	3.05***	2.07	2.51**	77.54	82.97***
Marker=1cM Effect Q1=10,Q2=1							
EV=1	0.99	3.28	3.26	2.41	2.6*	85.25	85.2
EV=10	0.41	2.9	3.24**	2.1	2.59**	81.08	85***
EV=20	0.14	2.53	3.25***	2.02	2.62***	77.36	85.1***

* P ≤ 0.05
** P≤ 0.01
*** P ≤ 0.001

In this table, scenarios are initial input parameters including QTL-marker distance, single QTL effect size, and environment variance. No. of QTL represents the mean value of QTL dosage in a selected population. 'Pheno' represents results from phenotypic selection. In each round of replication, genotypic values, and QTL dosage were compared between MAS and phenotypic selection's results using two independent simples t test. The significance is the average p value derived from 100 replication runs.

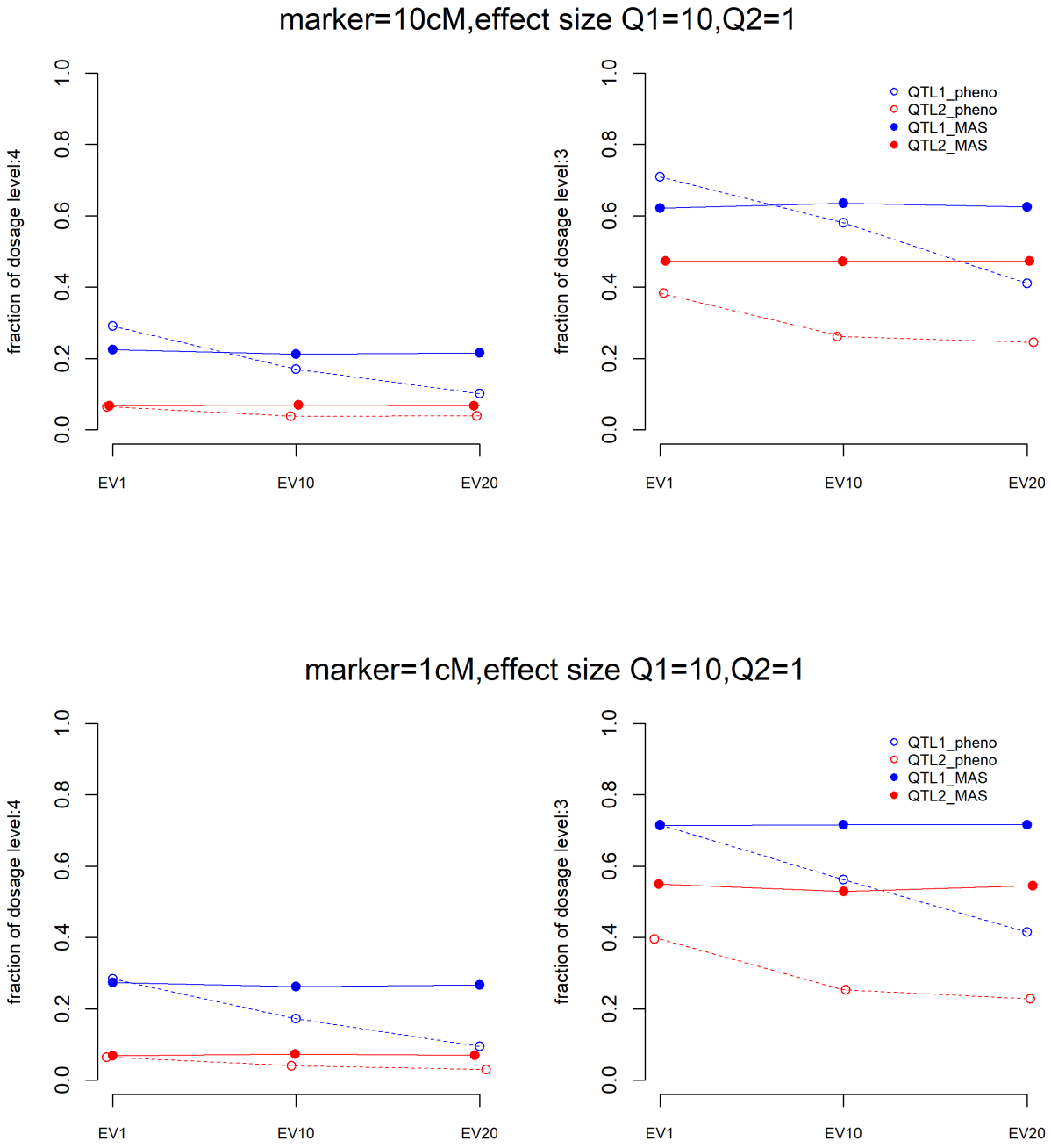


Figure 11. Fraction of individuals that carry highest QTL dosage >=3. On the x axis is EV representing standard deviation of normal distribution with μ=0. Solid circles represent individuals selected by MAS and hollow circles represent individuals selected by phenotypic selection. Because of different QTL effect size, The QTL1 (larger effect) were always selected more than QTL2 by both methods.

Unweighted marker selection

The second method, with unweighted markers, produced similar results as the weighted marker method. Due to balanced choice on markers, the average copy number and the highest dosage fraction of both QTL selected by MAS were basically the same (Table 7.). However, in all comparisons, minor effect QTL were significantly selected more often by MAS than phenotypic selection (Table 7 & Figure 12). MAS always selected more individuals carrying QTL2 than phenotypic selection. But 1cM MAS did not obviously improve the genotypic values and highest dosage fraction compared to 10cM MAS (Table 7 & Figure 12). 1cM MAS only increased around 5% individuals carrying highest dosage of QTL than 10cM MAS. In general, MAS still outperformed phenotypic selection when heritability is not high.

Table 7. Comparison between unweighted marker score MAS and phenotypic selection via genotypic values and the copy number of QTL, based on a DxD cross.

Scenarios	h^2	No. of QTL1		No. of QTL2		G value	
Marker=10cM Effect Q1=10,Q2=1		Pheno	MAS	Pheno	MAS	Pheno	MAS
EV=1	0.98	3.27***	2.81	2.41	2.84**	85.11***	80.95
EV=10	0.4	2.9	2.84	2.09	2.83***	81.11	81.25
EV=20	0.14	2.53	2.83**	2.06	2.82***	77.34	81.12**
Marker=1cM Effect Q1=10,Q2=1							
EV=1	0.98	3.28*	2.99	2.41	2.99***	85.21	82.91
EV=10	0.4	2.89	2.97	2.1	3.02***	81.02	82.76
EV=20	0.14	2.53	3.01***	2.06	3***	77.41	83.08***

* P ≤ 0.05
** P≤ 0.01
*** P ≤ 0.001

This table followed settings in Table 6.

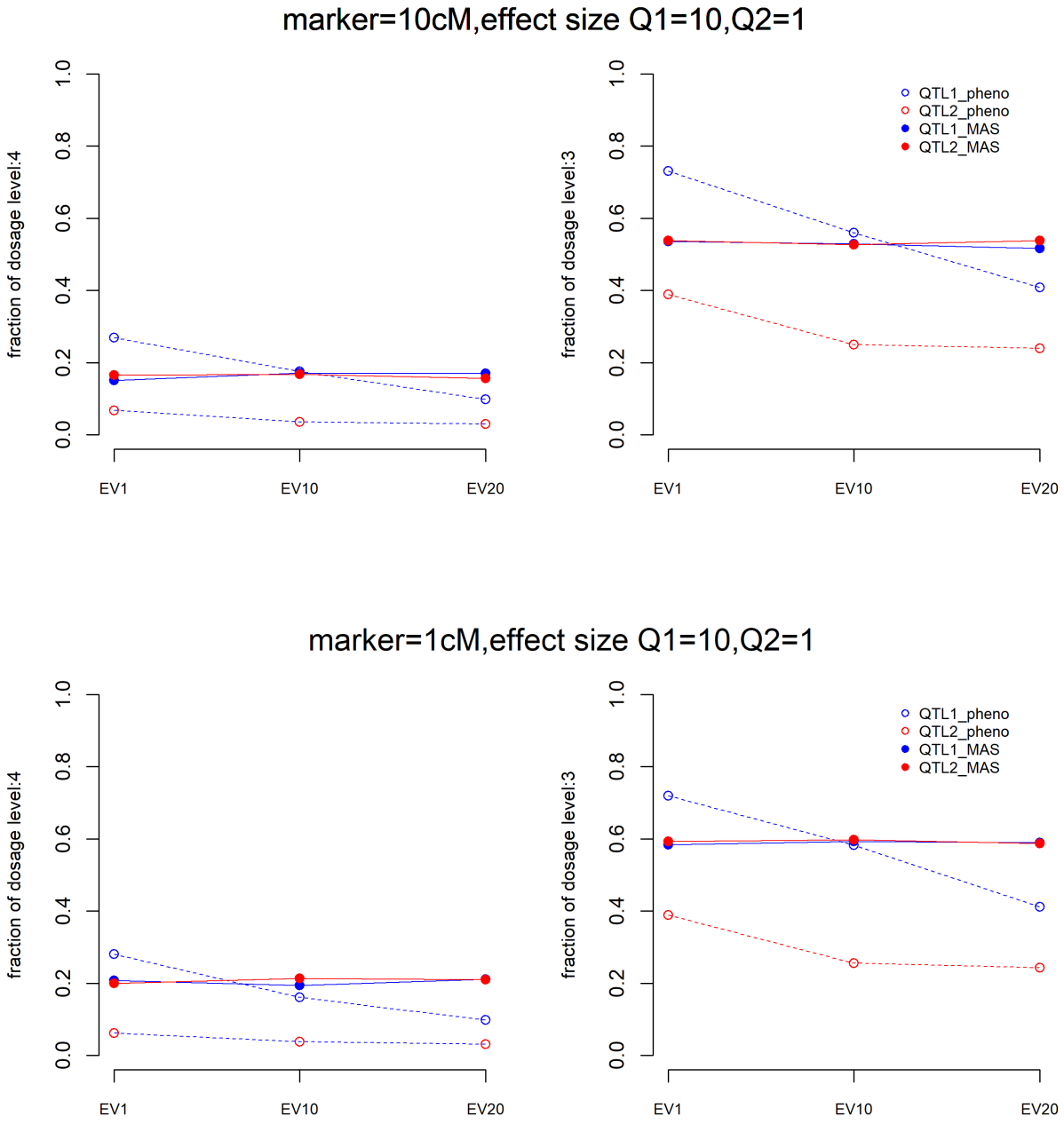


Figure 12. Fraction of individuals that carry QTL dosage >=3. On x axis are different scenarios for polyploids and markers. Solid circles represent ones selected by MAS and hollow circles represent selected by phenotypic selection. Due to balance selection on markers, the fraction of QTL1 and QTL2 in MAS stays the same in different scenarios. QTL2(minor effect) were always higher in MAS than phenotypic selection.

Discussion and conclusions

In this MSc Thesis project, we have demonstrated the feasibility of applying MAS in autotetraploid populations using simulations. The efficiency of MAS is determined by the linkage between markers and QTLs and by the number of QTLs. The linkage between QTL and markers and the heritability of traits are two determining factors on whether to favour MAS to phenotypic selection. Repulsion phase and distant coupling phase linked markers created severely insufficient selection efficiency while close linked coupling phase markers produced significantly better ones than phenotypic selection did. The selected offspring from MAS carried more alleles in terms of the number of different loci and copy numbers of each locus. In most of the scenarios, phenotypic selection showed no advantages over MAS. MAS shorten the breeding cycle through screening in seedling stages and it can be conducted in the first progeny generation. While conventional selection usually happens at second generation because the genetic composition is usually too heterozygous in first generation (Slater et al., 2014). Therefore, we propose that MAS can be applied in the breeding of polyploids given a sufficient amount of coupling marker linkage. We constructed an autotetraploid population with reasonable QTL positions and different markers type. These virtual chromosomes and marker positions are constructed based on many potato mappings results (Hackett et al., 2013). Although the number of QTL and chromosomes were limited, we think that the results are applicable for real chromosomes data in terms of length, centromeres positions and marker orders. Future simulation research can be completed by import other tetraploids like leek and alfalfa mapping data into PedigreeSim. Designing new polyploid breeding schemes is also a good direction for well-studied polyploids.

This simulation experiment had many advantages. One of the greatest advantages is that it drastically shortened time and effort needed for analysis compared to field tests. A similar experiment in the field requires a large amount of work and long time period, not to mention that specifying heritability for traits and controlling environment variance in the field experiment is very tedious. Field tests usually take years to finish whereas simulation only takes days. The second advantage is that it reasonably predicted results after many generations. In the first section, we have predicted till the 50^{th} generation. Testing such theory in the field test costs tremendous amount of time and its scientific gain is not great. Third, we constructed ideal individuals which are hard to acquire in the field test. In the third section, we assigned both parents with duplex markers and obtained a range of dosage from 0 to 4 in offspring population. Such parents are not easily found in nature as most of autotetraploids are genetically chaotic and the segregation patterns are not always distinguishable. In the simulation the effect of QTL accumulation is tested without noises. Additionally, the virtual environment had controlled inference, keeping unexpected errors to a minimum. 100 times of replication and large population size made the simulation results reliable compared to real values.

PedigreeSim is a software that simulates cross and selection process. We believe PedigreeSim performs more efficiently with higher ploidy levels in that high ploidy level structures are more complex to discover in the field test. With the help of PedigreeSim, breeders can easily design future breeding schemes for polyploids. PedigreeSim is very powerful in designing such schemes and each parent genotypes can be traced back while other software mainly focuses on genotype calling. TetraploidSNPmap (Hackett et al., 2017) is a new tool for linkage analysis in autotetraploids. The tool utilizes genotype data containing allele dosage information. This software gives potentials to precise marker information in polyploids. TetraOrigin is also helpful software that reconstructs haplotypes in outcrossing tetraploid using dosage data (Zheng et al., 2016). PedigreeSim is most suitable in this program because it simulates a cross population process and allows future researchers to design possible breeding schemes. Integrating other software into a breeding program is very useful. The current obstacle lies not in the software development but in the unclear genotypes for many polyploid crops. If markers are developed and their orders are precisely mapped, breeders can use the conclusion to guide breeding process. Potato (*Solanum tuberosum* L.) as an economic important autotetraploid crop, is ideal for testing MAS. Slater et al. (2014) demonstrated that combing MAS and estimated breeding values (EBVs) for simple and complex traits can lead to a significant reduction of time use for finding superior germplasm in a breeding program. Breeders would probably execute MAS first on potato traits that are either quantitatively inherited or heavily influenced by environmental noises, considering potato as an important crop and improving its many traits remains obstacles to overcome.

However, every simulation has a common limit that its parameters are predefined by humans, meaning certain results never occur if certain input parameters are not included. Situations including chromosome translocation, insertion and deletion never happen in PedigreeSim if developers did not program these settings in the first place. Second, some processes only theoretically happen in simulation while in the field test is impossible. For example, selfing in PedigreeSim is always possible so deriving homozygous lines in a simulated population is effortless, but because of inbreeding depression in many polyploid plants such selfing is not realistic (Rausch et al., 2005). Our simulation assumed 100% correctness for QTL-marker linkage in coupling phase. However, false-positive markers may be used during breeding. These markers can link in repulsion phase or random on non-homologues chromosomes but are regarded as linked in coupling phase. For a breeding program these markers are required to be validated before application. We have highlighted the unwanted consequences of using repulsion linked markers in the first chapter. Then we compared the selection results from flanking markers with same side markers and the difference was very significant. We also ran a controlled simulation where QTL and markers are not linked (on different homologues). Under this controlled

circumstance MAS efficiency dropped to 50% in a SxN. These situations stated the importance of marker quality in terms of linkage phase and genetic distances. But in the following results from DxD cross, we found the negative effect of QTL missing caused by recombination was mitigated by increasing dosages because it is almost impossible for both QTL-marker linkages to break in both parents at the same time. One other aspect is that improving QTL-marker linkage distance, to some extent, improving marker quality, becomes essential but has its plateau. We tested 1cM single marker for MAS and it manifested good performance. Any closer linked flanking markers, however, cannot improve the performance significantly more if the number of QTLs is limited.

Double reduction is a very interesting phenomenon and it is neglected along with multivalent forming in our simulation. Simulation could give different unexpected results like some individuals carrying two copies of alleles from one parent. But in general, double reduction happens with a low frequency in an outcross autopolyploid population (Bourke et al., 2015). Tetraploid potato the multivalent and univalent forming rates are very low compared to bivalent (Bradshaw, 2007). These research results suggested that ignoring multivalent and preferential pairing is reasonable. Newest research has also pointed out that multivalent and so-called double reduction phenomena are not as important as marker coverage on chromosomes in polyploids (Bourke et al., 2019). It is only worthwhile considering when population size is very small and multivalent pairing rate is high (Bourke et al., 2019). Thus, we think it is useful to include these parameters for more a precise model. But the key now is to develop more close linked markers and improve QTL detection precision (Bourke et al., 2019). For future application, developing high quality coupling phase markers is a major goal. Many recent kinds of research have argued the great focus on developing high quality markers (Bourke et al., 2019;). Despite insufficient efficiency demonstrated by phenotypic selection in our simulation, phenotypic selection still plays a vital role in breeding programs. One aspect is that some markers that are identified in diploid population cannot be used in autotetraploid breeding (Moloney et al., 2010). Our constructed phenotypes were ideal and simple whereas in the field phenotype values can range much differently for different traits and across various species. In the reality breeders can hardly determine genotypic variance for each trait. Plus, developing markers and run marker analyses are financially difficult for some breeders as well. Besides the financial aspects of MAS, there are a few situations that phenotypic selection outperforms MAS. Previous we set a controlled simulation with no environmental effect (EV=0). In such a case, phenotypic selection always selected QTL-carrying individuals but MAS failed because of some QTL lost via recombination. Some certain traits need no MAS, for example, disease resistance controlled by a single dominant gene. Other traits that are affected by epigenetics effect also need phenotypic selection to confirm. Flower colors changed by epigenetics in polyploids (Chen., 2007) cannot be selected based on marker results. The decision between MAS and phenotypic selection is not easy. MAS

can select traits with low heritability and shorten the breeding period for certain (Slater et al., 2014). But developing markers in the lab adds a lot of financial burden to MAS. Again, we proposed MAS in polyploid breeding, but many factors need to be considered.

In the first section, generation limit was set to 50. This limit was enough for large distance markers linkage e.g. 10cM, 5cM to test linkage break period distribution. But for flanking markers linkage 50 generations were not enough to plot the actual distribution because of its extremely low recombination ratio with QTL. One way is to lengthen the generation limit, say, to 100 generations. Another way is to increase population size so more QTL losing progenies emerge, but computation time should be considered as both methods require increasing amount of time. In the third section, if we increase marker distance to e.g. 20cM and numbers of QTLs as well, one possible result could be that phenotypic selection outperforms MAS even if the heritability is not high. From that situation researchers must consider each marker linkage as more QTL leads to more segregation. The QTL with minor effect size is also a hard problem to tackle. We found through phenotypic selection the minor effect QTL is always selected less than the larger effect one. As the environmental effect becomes very large, the minor effect QTL is lost during the selection. There are probably many minor effect QTL contributing to one trait. We found MAS can select more minor effect alleles than phenotypic selection. If breeders would like to preserve minor effect QTL through MAS, one way is to give minor QTL more decision weight during MAS selection. This act can help accumulate enough minor effect alleles for the next generation selection. But we only tested additive effect for alleles. Models including dominance and other non-additive effects need to be tested in follow-up research.

Another new idea of applying markers is genomic prediction. Genomic prediction utilise all markers information on the entire genome regardless of their linkage with QTL to predict progeny genotypes. Genomic prediction has been proposed to boost polyploids breeding very recently (Endelman et al., 2018). One of the advantages is genomic prediction can ignore identifying candidate genes for one certain trait and possible mutations (Endelman et al., 2018). However, to correctly order markers across homologues remains difficult in genomic prediction. Even markers information is enough across genomes, how to optimize markers set to conduct selection is an important research focus (Bourke et al., 2019). Non-additive effect is important to many economic traits in potato (Endelman et al., 2018). Initial research has proven the feasibility of using genomic prediction in potato, provided enough nonadditive effect alleles in germplasm (Endelman et al., 2018). There is research suggesting genomic selection (GS) will substitute MAS in future polyploid breeding research, given enough marker coverage information (Slater et al., 2013). But for now, similar selection efficiency comparison is not confirmed by field experiment yet. Our conclusions still promote using MAS in polyploid

breeding as the marker development in some tetraploid has advanced a lot. In the future, we would like to witness more models and software for polyploid breeding with more realistic settings for breeders.

Acknowledgements

The quality of this report could not be good without the help of my supervisors, Chris and Peter. Simulation research is not an easy task, especially for people who have limited experience in programming. However, to design a simulation experiment and gradually debug one after another is great fun. In the entire process, Peter offered very useful and enlightening ideas which are often out of my regular thinking. Chris gave very critical comments that forced me to think thoroughly about details and the entire program. Thanks again for both of my supervisors.

References

Adams, K. L., & Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Current opinion in plant biology,* 8(2), 135-141.

Bourke, P. M., Voorrips, R. E., Visser, R. G., & Maliepaard, C. (2015). The double-reduction landscape in tetraploid potato as revealed by a high-density linkage map. *Genetics*, *201*(3), 853-863.

Bourke, P. M., Arens, P., Voorrips, R. E., Esselink, G. D., Koning-Boucoiran, C. F., van't Westende, W. P., ... & Visser, R. G. (2017). Partial preferential chromosome pairing is genotype dependent in tetraploid rose. *The Plant Journal*, *90*(2), 330-343.

Bourke, P. M., Voorrips, R. E., Visser, R. G., & Maliepaard, C. (2018). Tools for genetic studies in experimental populations of polyploids. *Frontiers in plant science*, *9*, 513.

Bourke, P. M., Hackett, C. A., Voorrips, R. E., Visser, R. G., & Maliepaard, C. (2019). Quantifying the power and precision of QTL analysis in autopolyploids under bivalent and multivalent genetic models. *G3: Genes, Genomes, Genetics*, g3-400269.

Bradshaw, J. E., & Mackay, G. R. (1994). Breeding strategies for clonally propagated potatoes.

Bradshaw, J. E. (2007). The canon of potato science: 4. Tetrasomic inheritance. *Potato Research*, *50*(3-4), 219-222.

Chen, Z. J. (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.*, *58*, 377-406.

Colton, L. M., Groza, H. I., Wielgus, S. M., & Jiang, J. (2006). Marker-assisted selection for the broad-spectrum potato late blight resistance conferred by gene RB derived from a wild potato species. *Crop Science*, *46*(2), 589-594.

Comai, L. (2005). The advantages and disadvantages of being polyploid. *Nature reviews genetics*, *6*(11), 836.

Dreher, K., Morris, M., Khairallah, M., Ribaut, J. M., Pandey, S., & Srinivasan, G. (2002, July). Is marker-assisted selection cost-effective compared to conventional plant breeding methods? The case of quality protein maize. In *Proceedings of the 4th annual conference of the international consortium on agricultural biotechnology research (ICABR'00)* (pp. 203-236).

Endelman, J. B., Carley, C. A. S., Bethke, P. C., Coombs, J. J., Clough, M. E., da Silva, W. L., ... & Holm, D. G. (2018). Genetic variance partitioning and genome-wide prediction with allele dosage information in autotetraploid potato. *Genetics*, *209*(1), 77-87.

Fisher, R. A. (1947). The theory of linkage in polysomic inheritance. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 233(594), 55-87.

Gimelfarb, A., & Lande, R. (1995). Marker-assisted selection and marker-QTL associations in hybrid populations. *Theoretical and Applied Genetics*, *91*(3), 522-528.

Hackett, C. A., McLean, K., & Bryan, G. J. (2013). Linkage analysis and QTL mapping using SNP dosage data in a tetraploid potato mapping population. *PloS one*, *8*(5).

Hackett, C. A., Boskamp, B., Vogogias, A., Preedy, K. F., & Milne, I. (2017). TetraploidSNPMap: software for linkage analysis and QTL mapping in autotetraploid populations using SNP dosage data. *Journal of Heredity*, *108*(4), 438-442.

Haldane, J. B. S. (1930). Theoretical genetics of autopolyploids. *Journal of genetics*, *22*(3), 359-372.

Moloney, C., Griffin, D., Jones, P. W., Bryan, G. J., McLean, K., Bradshaw, J. E., & Milbourne, D. (2010). Development of diagnostic markers for use in breeding potatoes resistant to Globodera pallida pathotype Pa2/3 using germplasm derived from Solanum tuberosum ssp. andigena CPC 2802. *Theoretical and applied genetics*, *120*(3), 679-689.

Peleman, J. D., & Van der Voort, J. R. (2003). Breeding by design. *Trends in plant science*, *8*(7), 330-334.

Qu, L. and J. Hancock (2001). "Detecting and mapping repulsion-phase linkage in polyploids with polysomic inheritance." Theoretical and applied genetics 103(1): 136-143.

Rausch, J. H., & Morgan, M. T. (2005). THE EFFECT OF SELF-FERTILIZATION, INBREEDING DEPRESSION, AND POPULATION SIZE ON AUTOPOLYPLOID ESTABLISHMEN. *Evolution*, *59*(9), 1867-1875.

R Core Team (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.

Ripol, M. I., Churchill, G. A., da Silva, J. A., & Sorrells, M. (1999). Statistical aspects of genetic mapping in autopolyploids. *Gene*, *235*(1-2), 31-41.

Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, *243*(2), 281-296.

Sim, S. C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., Van Deynze, A., ... & Francis, D. M. (2012). Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PloS one*, *7*(7), e40563.

Slater, A. T., Cogan, N. O., & Forster, J. W. (2013). Cost analysis of the application of marker-assisted selection in potato breeding. *Molecular breeding*, *32*(2), 299-310.

Slater, A. T., Cogan, N. O., Hayes, B. J., Schultz, L., Dale, M. F. B., Bryan, G. J., & Forster, J. W. (2014). Improving breeding efficiency in potato using molecular and quantitative genetics. *Theoretical and applied genetics*, *127*(11), 2279-2292.

Stebbins Jr, G. L. (1947). Types of polyploids: their classification and significance. In *Advances in genetics* (Vol. 1, pp. 403-429). Academic Press.

Sybenga, J. (1996). Chromosome pairing affinity and quadrivalent formation in polyploids: do segmental allopolyploids exist? *Genome*, *39*(6), 1176-1184.

Varshney, R. K., Graner, A., & Sorrells, M. E. (2005). Genomics-assisted breeding for crop improvement. *Trends in plant science*, *10*(12), 621-630.

Voorrips, R. E., & Maliepaard, C. A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. *BMC bioinformatics*, *13*(1), 248.

Wu, K. K., Burnquist, W., Sorrells, M. E., Tew, T. L., Moore, P. H., & Tanksley, S. D. (1992). The detection and estimation of linkage in polyploids using single-dose restriction fragments. *Theoretical and Applied Genetics*, *83*(3), 294-300.

Yamamoto, E., Matsunaga, H., Onogi, A., Kajiya-Kanegae, H., Minamikawa, M., Suzuki, A., ... & Miyatake, K. (2016). A simulation-based breeding design that uses whole-genome prediction in tomato. *Scientific reports*, *6*, 19454.

Zheng, C., Voorrips, R. E., Jansen, J., Hackett, C. A., Ho, J., & Bink, M. C. (2016). Probabilistic multilocus haplotype reconstruction in outcrossing tetraploids. *Genetics*, *203*(1), 119-131.

Appendix 1

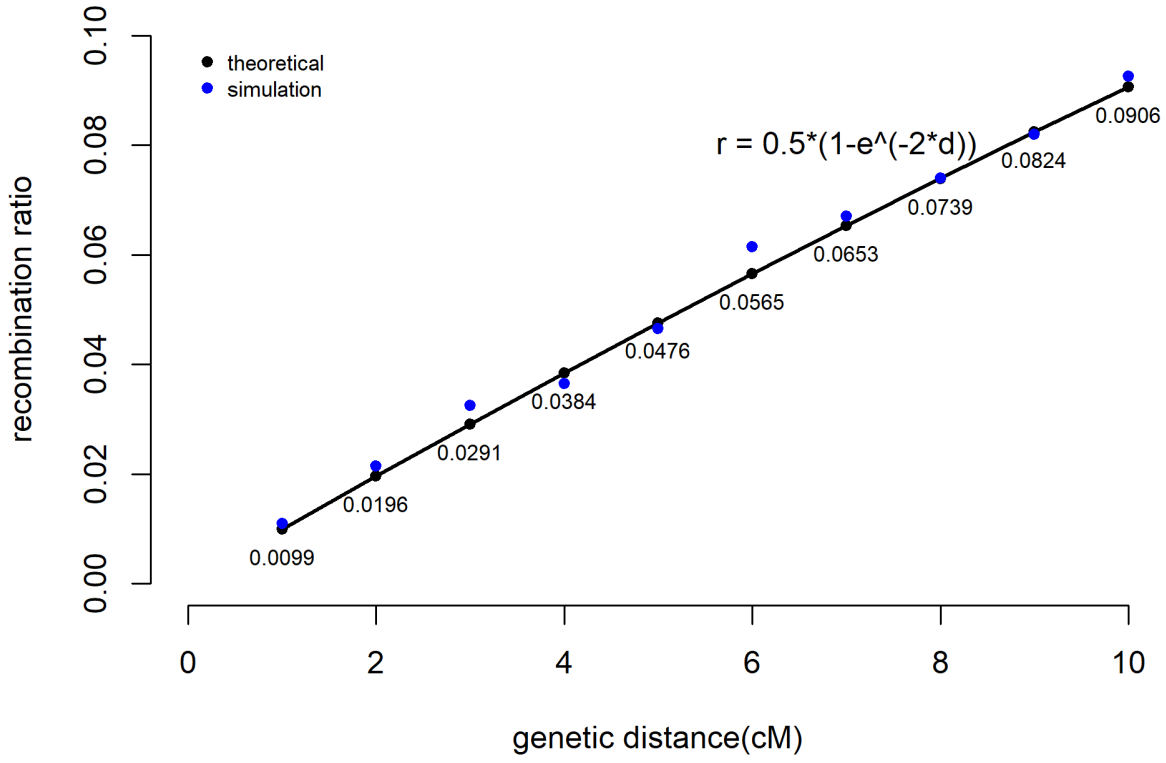


Figure 1, Appendix 1. Theoretical and simulated recombination ratios between single QTL-marker linkage from 1cM to 10cM using Haldane mapping function.