博士論文

Evaluation of the potential of genomic selection in plant breeding using simulations and field trials

（シミュレーションと栽培試験を用いた
植物育種におけるゲノミックセレクションの
可能性評価）

矢部 志央理

Evaluation of the potential of genomic selection in plant breeding using simulations and field trials

by

Shiori Yabe

A thesis submitted
for the degree of Doctor of Philosophy
in the Department of Agricultural and Environment Biology of the
Graduate school of Agricultural and Life Sciences

The University of Tokyo

Research Supervisor Dr. Hiroyoshi Iwata

March 2015

Abstract

Genomic selection (GS) is a promising method for animal and plant breeding. Using a model relating phenotypes of target traits to marker genotypes, GS enables breeders to predict the genotypic potential of selection candidates based on their marker genotypes. Although GS is expected to streamline and accelerate plant breeding, studies on GS may not be enough to achieve the routine use of GS in breeding programs. In plant breeding, the performance of GS is difficult to evaluate generally because it is affected by among-species variations in mating systems, ways to propagation, population structures, and conventional breeding strategies. In this dissertation, I evaluated the performance of GS through simulation studies and field trials in consideration of the among-species variations. In addition, I developed a novel simulation language for breeders to evaluate the potential of planned breeding schemes.

1. Simulation study of genomic selection in allogamous plants

Mass selection is an important method for the breeding of allogamous crops. This method, however, has drawback, i.e., low efficiency of genetic improvement, because it requires a single-plant evaluation. GS enables higher reliability of a single plant evaluation than phenotypic selection (PS) because GS can be performed on marker genotypes. On the other hand, GS may not work well in an allogamous species with a large effective population size, which results in a low level of linkage disequilibrium (LD), because GS utilizes LD between QTLs and markers. In this study, on the assumption that an annual allogamous plant species has a very low level (i.e., close to linkage equilibrium) of LD, I conducted breeding simulations for two types of target traits, a trait expressed before pollination and a trait expressed after pollination. Especially for a trait expressed after pollination, in which pollen parents cannot be selected before crossing in PS, GS had a larger genetic gain than PS. For a trait expressed before pollination, I compared GS with PS and conventional marker-assisted selection (MAS) in the simulations and evaluated the performance of GS under various scenarios. Results showed that GS attained higher genetic gain than either PS or MAS. GS with a larger population size and more cycles attained higher genetic gain except when the population size

was small. The cost efficiency of GS was higher than that of PS only when the genotyping cost was lower than about one-fourth of the phenotyping cost. To evaluate the performance of GS in a trait expressed after pollination, I compared GS and PS in traits expressed before and after pollination. Results show that GS showed almost identical genetic gain genetic gain in both traits except when GS was conducted once per year (i.e., same as PS), while PS in a trait expressed after pollination showed much lower genetic gain than PS in a trait expressed before pollination. High efficiency of GS in a trait expressed after pollination was attributable to the high selection accuracy of chromosomes derived from a pollen parent at GS steps immediately after model updating steps, at which pollen parents were not selected because model updating requires phenotype evaluation. It resulted in increased population size and prevented depletion of genetic variation in a breeding population. This study indicated that GS has a great potential to improve the efficiency of mass selection of allogamous crops in particular when a target trait is expressed after pollination.

2. Simulation evaluation of island-model genomic selection in an autogamous plant

In the breeding of autogamous crops, population breeding and pedigree method, which utilize inbred lines, are commonly used in breeding programs. This situation results in the issue of a lack of new combinations of genes in a breeding population. Recurrent selection can be used to create recombination in a population, but it requires single-plant evaluation, which is generally inaccurate. GS may have high reliability of single-plant evaluation and would be effective in recurrent selection of an autogamous species. Additionally, the concept of "island model" inspired from population genetics and evolutionally algorithms may be useful to maintain genetic variation through the breeding process. I conducted GS simulations using a real marker genotypic data of rice cultivars to evaluate the efficiency of recurrent selection and the island model in an autogamous species. Results suggested that recurrent selection could attain higher gain than a conventional method using inbred lines. In the recurrent selection, an initial population derived from multiple bi-parental crosses showed larger genetic gain than a population derived from a single bi-parental cross, suggesting the importance of genetic variation in an initial population. The island-model GS could attain higher gain than the bulked GS in later generations because the island-model GS could maintain larger genetic variation than the bulked GS and improve the genetic potential of the whole subpopulations. Because of the ability of the bulked GS to attain gain rapidly in early generation, it is suggested that breeders should choose a suitable breeding scheme according to their required time.

3. Simulation of the impact of mis-labeling on genomic selection in cassava

In actual plant breeding, humans make mistakes unlike simulations. Especially in GS breeding, humans would tend to make mistakes because GS involves more steps than PS does. To implement GS breeding in actual, the effect of human mistake should be taken into account to consider the level of restriction to prevent human mistakes. In plant breeding, controlling mistakes too strictly may not be cost effective if the mistakes do not have a large impact. I evaluated the impact of mis-labeling, in which a plant happens to be swapped for another one, in cassava breeding using simulation. As simulation results, all scenarios with six levels of mis-labeling (5, 10, 20, 30, 40, and 50%) attained a certain genetic gain because of the relationship between the genetic variance and the prediction accuracy. The higher mis-labeling rate became in a breeding population, the lower selection intensity the breeding population experienced at selection cycles. This situation made the genetic variance in a population with mis-labeling high, and made the response to selection high. The increased genetic variance observed under mis-labeling led to sufficiently improve the accuracy, at least for low mis-labeling rate (10% or less). It is suggested that the large scale of mis-labeling should be prevented, but that preventing small scale of mis-labeling is not cost effective in plant breeding.

4. Field trial of genomic selection using common buckwheat

A field experiment of GS breeding was performed with a real breeding population in common buckwheat. I compared the efficiency of GS with that of PS for improvement of seed yield per unit area in the two years of field trial. To select seed yield per unit area, which cannot evaluated in a single plant, I built a selection index that predict performance of each plant in seed yield based on other seven traits (main stem length, number of nodes, flowering of the first flower, number of flower clusters, number of primary branches, 1000 seed weight, and test weight) that can be evaluated in a single plant. This index was used throughout the selection cycles. In GS breeding, two selection cycles were conducted in each year, and the prediction model was updated every year by using 14,598 to 50,000 markers. In PS breeding, selection was conducted once per year. To verify the difference in performance between GS breeding and PS breeding, a field test was conducted in 2013 after the two years of breeding. In the test, 48 plants from each generation were cultivated. The selection index, seven traits composing the selection index, the number of seed set of a plant, and the number of secondary branches were evaluated in the field test. Through two years of selection, GS breeding attained 49% higher

gain in number of flower clusters and number of seed set than the base population. For the selection index, which was used in selection directly, GS breeding attained 15% higher value than the base population. In PS breeding, selection index increased 4% from the base population, while it was not statistically significant. These results show that GS has higher performance than PS in the genetic improvement in common buckwheat. I compared two prediction models built at the first and second years through the evaluation of their prediction accuracy at the second year, and found that the former showed lower accuracy than the latter, suggesting the importance of model updating in GS breeding. The superiority of GS over PS might be resulted from the effects of accelerating generations using offseason nursery. In the field trial, the efficiency of GS based on the selection index was suggested to improve yield related traits simultaneously.

5. Development of a simple language to script and simulate breeding schemes: the breeding scheme language

It is difficult for plant breeders to determine the optimal scheme under conditions of a target species and target traits because there are a number of possible breeding schemes. Although simulation study is useful to help choose a better (or the best) breeding scheme, it is difficult for breeders to take the first step in conducting breeding simulation because of the complexity to build a simulation platform or even to using a simulation tools. In the present study, I developed a simple and flexible simulation platform, breeding scheme language. Users can define their target species and breeding schemes by utilizing the language. This language might be useful for breeders to evaluate breeding schemes and to choose a breeding scheme among a number of possible schemes.

I demonstrated the high potential of GS by using simulations and field trials. Through these studies, it is suggested that there are factors that affect on GS gain as well as factors that have less effects on GS gain. Update of a prediction model is essential for GS breeding even with a cost especially when a breeding population has low levels of LD. The result of the simulations was coincided with that of the field trial, suggesting the properness of the simulations performed in this study. In the future, the collateral implementation of breeding simulations and a field trial might enable us to improve the efficiency of plant breeding by reflecting the current situation to the simulations and choosing suitable selection strategy at each step of breeding on the basis of the results of the simulations.

Table of contents

Chapter 1

Introduction

One of the serious problems that humankind faces today is a huge increase in human population. The world's production of wheat (*Triticum aestivum* L.), rice (*Oryza sativa* L.) and maize (*Zea mays* L.), which provides half of the world's diet and two-thirds including feed for livestock and poultry, is linearly increasing (Tweeten and Thompson, 2008). To meet the recent estimated world's population in 2050, we need to improve the annual increase of the cereal food supply by 38% (i.e., from 32 metric tons per year to 44 metric tons per year) in the next 40 years (Tester and Langridge, 2010). In spite of the requirement of the rapid increase of yield in not only cereals but also other crops, the amount of land required to provide food for the world's population is decreasing due to the growth of population (Phillips, 2010). Nowadays, in addition to the demand of food crops, the demand of biofuels also grows because of renewable energy demands, and is competing with food production directly in crops used both for food and fuel, and indirectly on the utilization of arable land and other resources (Tilman et al., 2009). The demand of biofuel faces the necessity to develop new crop varieties and species for biofuel that can grow in unsuitable land for growing food crops. Due to the double demand from food supply and biofuel, a rapid increase in the production of crops is strongly required.

To fulfill the targeted level of improvement in genetic performance of crops, it is absolutely necessary to develop and utilize breeding technologies using genomic information and tools such as DNA sequencing methods and analysis approach which can handle the data obtained from the new technologies. A recent remarkable revolution in genetic knowledge enables us to utilize these biological technologies. One of the active technologies using genome information is production of genetically modified (GM) crops. GM technologies are used in many plant species because of its usefulness owing to the ability to introduce useful genes beyond the species. However, the production and development of GM crops caused many problems because of the concern about gene drift, herbicide-resistant super-weeds, biodiversity

and the unknown long-term consequences (Gaskell et al., 2004). Another method for plant breeding is marker-assisted selection (MAS). MAS uses the information derived from the linkage mapping and quantitative trait loci (QTL) analysis, i.e., the estimation of positions, number and effects of QTL, for the target traits in target species, and uses molecular markers that are closely linked to QTL in order to select favorable QTL alleles indirectly (Ukai, 2000). With MAS, breeders may be able to develop desired cultivars in short time span because of the non-necessity of phenotypic evaluation associating with field trials. As discussed later, however, MAS has several drawbacks: extra labor for QTL mapping and development of markers close to QTL, and inefficiency in the improvement of traits controlled by a number of genes. Genomic selection (GS; Meuwissen et al., 2001) is a method that uses markers distributing over the whole genome. In GS, we build a model for predicting genotypic values based on marker genotype data by using phenotype and marker genotype data of a number of individuals/lines. Then, we use the model for predicting genotypic values of selection candidates based on their marker genotype data, and select good individuals/lines based on their predicted genotypic values. The use of a large number of high-density markers is one of crucial features of GS so that at least one marker is expected to be in linkage disequilibrium (LD) with QTL controlling the target trait in the entire target population in the modeling (Desta and Ortiz, 2014). GS has a potential to improve genetic ability rapidly in both animal and plant breeding, and has been put into practical use in animal breeding. GS is attracting great expectations also in plant breeding, but is currently still in the research phase.

GS is expected to attain higher genetic improvement in a unit time and cost than phenotypic selection (PS) (Wong and Bernardo, 2008), especially when one genotype is evaluated on the basis of a single plant. Since the dawn of history, PS has been the most common breeding strategy. In PS, breeders evaluate each individuals/lines based on their phenotypic values. When we evaluate phenotypes based on a single plant and/or a trial in a single environment, PS is largely affected by environmental errors and results in inefficiency of the selection. GS can predict genotypic values of selection candidates on single plant basis and under any environments because the selection is performed on marker genotype data. In PS, it takes for years to evaluate the genetic potential of individuals/lines because of field trials over multiple environments. Especially, it takes a long time to evaluate genetic performances for perennial plant species such as forest and fruit trees. For example, it takes long time to evaluate traits relating to fruit quality because trees do not produce fruit in the juvenile phase (Ukai, 2003). By combining GS with accelerating generations using offseason nursery, breeders can

conduct more than one cycle of GS per one cycle of PS. From these factors, GS is expected to have a greater potential than PS especially for accelerating the rate of genetic improvement.

MAS is also the selection method that uses genetic markers. Because MAS requires no field trial at the selection step, as is the case of GS, breeders can implement MAS several cycles per one ordinal growth cycle by using offseason nursery. Thus, MAS has high potential in short-term breeding (Edwards and Page, 1994), and has an advantage especially when it is difficult to evaluate phenotype of target trait (Ribaut et al., 1997). However, MAS has two major drawbacks: the necessity of QTL mapping and development of markers that are closely linked to the QTL, and the difficulty of improvement of traits controlled by a large number of genes. The QTL mapping step requires extra materials, time, and labor for breeding. In GS, the extra labor is also required to obtain the training data to build a prediction model, but it may be possible to use historical breeding records as the training data of GS. In MAS, it might be possible to use QTL information that was obtained previously. However, it is reported that QTL detected in a mapping population might not be responsible for genetic variation existing in a breeding population (Strauss et al., 1992). Overestimated QTL effects (Hoeschele and VanRaden, 1993; Lande and Thompson, 1990; Melchinger et al., 1998) and failing to detect small-effect QTL might also make MAS inefficient. Although MAS is effective when a target trait is controlled by a small number of QTL (Bernardo, 2001), gene pyramiding (i.e., aggregation of favorable alleles of QTL) is difficult especially when a trait is controlled by a large number of QTL. In general, many quantitative traits of breeding target are controlled by a large number of QTL (Hayes and Goddard, 2001; Kearsey and Farquhar, 1998). In gene pyramiding, careful planning of crosses are required to obtain an ideal genotype even when the target trait is controlled by a small number of QTL (Servin et al., 2004). GS is suitable especially for improving a trait dominated by a large number of genes, because GS does not hold any threshold to select markers to be incorporated into a prediction model, and thus, all of the genetic variation may be captured by whole-genome markers (Hayes et al., 2013). Therefore, GS is expected to have possibilities to overcome the limitations of MAS (Heffner et al., 2010).

GS has the potential to improve the efficiency of plant breeding dramatically because of its advantages over other breeding systems mentioned above. However, there are still several problems for the practical use of GS in plant breeding. The potential of GS itself is not well evaluated in plant breeding in the first place. Additionally, a single breeding program has its specific restriction on breeding scheme. This situation requires us to search the most efficient breeding scheme under the current restriction each time when we start a breeding program. In

plant breeding, we have to improve various plant species, e.g., allogamous species and autogamous species. Restrictions for breeding may be largely different between species, depending on its breeding system and breeding history. Moreover, there are different types of traits to be improved even for a single plant species. We should evaluate the potential of GS and search a suitable breeding scheme under various situations. Furthermore, to evaluate the potential of GS, empirical studies have been lacking. Empirical evidences are essential to put GS to practical use. I proposed to conduct simulation studies and field trials in the present study for solving the problems above. These studies would reveal the potential of GS under various restrictions imposed in an actual breeding program. It is also important to develop a system that enables us to evaluate the potential of GS under the current restriction. In the present study, thus, I created a novel system to evaluate easily the potential of GS via simulations.

Simulation study is one of the best ways to verify the potential of GS and choose a suitable breeding scheme on ahead of the real-world GS breeding under the restrictions in a real breeding programs. Despite of the great expectations for GS, there are many issues that need to be addressed for implementing GS in a real plant-breeding program. GS has various factors to affect the efficiency of selection, such as marker density, a structure of training population, and a prediction model. Because of the factors, it is difficult for breeders to decide an optimal breeding scheme. Moreover, breeders have to decide everything in view of their research budget. Simulation is a good tool to find the best design of GS. Hickey et al. (2014) simulated genomic prediction to see the difference of prediction accuracy between multiple training population designs and genotyping strategies. Iwata and Jannink (2011) also simulated genomic prediction to compare statistical methods used for prediction models on the basis of a real marker data of barley. These studies, however, try to evaluate the accuracy of GS at the one cycle of selection. To evaluate the potential of each breeding scheme from a long-term perspective, simulations of whole breeding generations are necessary. Iwata et al. (2011) conducted GS simulations of 60 years in forest tree species, and discussed about the timing of updating a prediction model. Bernardo (2009) suggested that the rank of genetic gain between PS and GS changed during some generations. Jannink (2010) evaluated the GS methods in an effort of long-term improvement. These simulation studies showed the factors that that are important for long-term breeding. It is necessary to conduct simulation studies including several cycles of selection ahead of practice of GS breeding with repeated selections and crosses. In plant breeding, additionally, differences of mating and reproduction systems remain as unsolved issues. Most of cereal crops, supplying a large percentage of the world's food supply, are autogamous. On

another front, there are allogamous plant species that have an important meaning to improve genetic ability for food supply and for exploitation of new plant species for biofuels. Suitable breeding schemes are different according to the types of cultivars released to market. For example, most of autogamous crop species are released as pure lines. Crops that have ability of vegetative propagation require only a single good genotype. Vegetables such as tomato require a good F_1 genotype to release to market. Maize is also released as F_1 between two inbred lines as a result of heterosis breeding. Common buckwheat and a number of forage crops require a genetic improvement as an outcrossing population. Thus, simulation studies should be conducted according to mating system, life cycle, and target type of cultivar of each target plant species. As for breeding simulations, another problem remains. Generally, breeding simulations are conducted assuming that no human mistake happens. However, GS involves many steps in both field and laboratory. The more steps the procedure requires, the more human mistakes may happen. Ly et al. (2013) reported the possibility of mis-labeling in their training population for GS. All events relating to humans have possibility of error and mistakes. Before an actual GS is performed, the effects of human mistakes should be evaluated.

It is essential to verify the potential of GS in field-experimental study because simulation studies may have unnoticed pitfalls that cause discrepancy between conditions assumed in simulations and ones realized in field trials. Almost all empirical studies of GS just evaluated the prediction accuracy in a population. As far as I know, two papers reporting empirical studies that evaluated the potential of GS have been published so far. Massman et al. (2013) showed the result of three cycles of GS in maize breeding, and suggested the advantage of using all available markers instead of using only markers having significant effects. Asoro et al. (2013) also showed the results of GS, whose target trait was β-glucan of oat (*Avena sativa* L.). They compared the efficiency of GS with MAS and pedigree-based selection, and showed the superiority of GS for selection of the polygenic trait. However, these two studies are specified to their target plants (i.e., maize and oat). To verify the efficiency of GS in plant breeding, more empirical studies involving breeding process are required. And these empirical studies can be compared with the results of simulations assuming the same situations. Breeding simulations are conducted on the basis of a number of assumptions (e.g., absence of epistatic and dominance effects, number of QTL, size of QTL effects, levels of LD in a breeding population, and same fertility in selected parents), to which actual plant breeding might not follow. Comparing simulation results with actual breeding results, we can examine the properness of the assumptions in simulations, and can discuss the issues that need to be addressed for improving

the efficiency of GS in actual breeding programs.

Breeding simulation is one of the good tools to guess the consequences of different breeding schemes. Thus, a system that enables breeders to try different breeding schemes easily via simulations will help them to choose a suitable breeding scheme. To implement GS in breeding programs, professionals of many field (e.g., breeders for field work, experimenters for laboratory work, and statisticians for genomic prediction) should corporate each other for the implementation because it is difficult that one person can do and understand all process in GS breeding. Brown and Caligari (2008) mentioned that plant breeders would require knowledge in many subjects relating plant breeding. However, it is usually quite difficult to get knowledge in many topics. Thus, plant breeders tend to be conservative, and do not want to change their breeding schemes if they know the traditional one works. This is not only for GS breeding, but also for other breeding schemes. If breeders can try their planned breeding schemes by computer simulations by themselves, the results help them introduce new breeding schemes. However, it is difficult for breeders to learn how to compose breeding simulations. A simple and flexible simulation platform is required so that non-professional people can conduct breeding simulations.

In this Ph. D. thesis, I report the researches of GS from some aspects. In Chapter 2, the basic information and some previous researches about of GS are introduced. In Chapter 3, I conducted GS simulations assuming an annual allogamous crop. There, GS was performed on the basis of the strategy of mass selection, in which selection and crossing were repeated based on a single plant evaluation. In the simulations, I compared GS with PS and MAS. The effects of many factors relating to the outcome of GS were evaluated, such as number of markers, mode of inheritance of markers, statistical methods of the prediction model, population size, and levels of promoting generations. On the basis of the suitable design, the cost efficiency was evaluated according to population size and the number of generations. Additionally, the efficiency of genetic improvement in different types of traits, i.e., traits expressed before pollination and after pollination, was evaluated under GS and PS breeding. In Chapter 4, I conducted GS simulations assuming an autogamous crop by using an actual marker data of rice. In breeding of autogamous crops, inbred lines are generally used. In this study, I used recombinant inbred lines as a training population and parents of breeding population. An important advantage of inbred lines is the possibility of repeated measurements of traits by using almost identical genotypes, which increases selection accuracy of the traits. On another front, they have a problem in exhaustive use of genetic resource because of their strong linkage

brock and limited resources. I evaluated the efficiency of GS in an autogamous crop with proposing a new breeding strategy, island model GS. In Chapter 5, GS breeding simulations including human error (i.e., mis-labeling) were conducted to evaluate the effect of human mistakes, which are unavoidable in plant breeding involving many people. I assumed breeding of cassava (*Manihot esculenta* Crantz) as an example. In this situation, each genotype were propagated by using clones and contributed to the training population. The impact of mis-labeling was thought to remain over generations in GS. In Chapter 6, on the basis of simulation studies in Chapter 3, the result of two-year field trials of GS in common buckwheat (*Fagopyrum esculentum* Moench) is reported. In that program, the target trait was seed yield, which was a complex trait and could not be evaluated with a single plant. To implement GS efficiently, a selection index with which the seed yield per unit area can be estimated from other traits was built and used as a target for the improvement. The efficiency of GS was compared with that of PS. By posteriori analysis, the result of the field trial was compared with the simulations. The common and different points were examined between them. In Chapter 7, a simple computer language that I developed to simulate breeding schemes is introduced. This computer language should be useful for plant breeders to use to decide the breeding scheme. As a summary, I overview all results in this dissertation, and discussed about the remaining problems and future perspective in Chapter 8.

Chapter 2

Overview of genomic selection

2-1. Basis of genomic selection

Meuwissen et al. (2001) proposed the idea of GS, selection based on genetic potential predicted by using genome-wide markers, for accelerating genetic improvement of quantitative traits controlled by a number of genes. Using genome-wide dense markers, we can expect that some markers will be close to the QTL and in LD with the QTL (Fig. 2.1), and use such genome-wide dense markers to estimate the marker effects in place of true QTL effects that are linked to their neighbor markers. The genotypic values are predicted in the breeding population by using the estimated marker effects. GS, in which we can predict abilities of selection candidates based on their genome-wide DNA marker genotypes, is conducted on the basis of the predicted genotypic values instead of phenotypic values. Consequently, the most valuable plants can be selected as parents for the next generation without actually testing them in the field. Moreover, rapid identification at the seedling stage or even at the seed stage can shorten the breeding cycle by obviating the time necessary for field-testing.

GS is a good method to select quantitative traits controlled by a number of genes, as mentioned above. Important traits in plant breeding, such as grain, fruit or tuber yield, biomass yield, end-use quality, are sometimes controlled by a large number of genes. Some other traits are also influenced by these traits. Benefits from the conventional MAS based on a small number of markers are limited especially in such traits controlled by a number of genes. MAS might be effective and has been used for the improvement of traits controlled a few genes, such as disease resistance (e.g., the traits reviewed by Collard and Mackill (2008)). GS is expected to attain higher efficiency than MAS for selection of traits controlled by a number of genes.

GS is a novel selection method, which enables us to estimate the detailed composition of genotypic value of a genotype by estimating marker effects instead of QTL effects. In the past, geneticists have been trying to represent genotypic values of candidate genotypes. The recent

development of the technology of genotyping has enabled us to do that in detail. Genome-wide association study (GWAS) is also a technique utilizing dense markers distributing over the whole genome. While both of GS and GWAS are the genomic information based strategies for crop improvement, the purposes are different. The purpose of GWAS is to identify DNA marker alleles that are associated with a quantitative trait. The purpose of GS is to predict breeding value of selection candidates on the basis of marker data, and thus causal loci are not necessarily identified in GS (Hamblin et al., 2011). Through GS breeding programs, both phenotype and marker genotype data will be accumulated. The accumulated data can be used for genetic dissection of complex traits, e.g., gene/QTL discovery. Thus, we will have reward for genetics from the GS breeding.

The main process of GS is calculating genotypic values for selection candidates, which have only marker genotype data. By using a prediction model that was trained from individuals with both phenotype and genotype data sets, genotypic values are calculated by applying the marker genotype of selection candidates to the model (Fig. 2.2; Heffner et al., 2009). The population that have both phenotype and genotype data and is used for model building is called "training population". Selection is performed on the basis of the predicted genotypic values among selection candidates (i.e., plants in a breeding population).

To attain high accuracy of prediction, a training population must be representative of a population to which selection candidates belong (Heffner et al., 2009). In fact, some empirical study suggested that the prediction accuracy was higher when a training population was highly related to a breeding population than when a training population was no representative of a breeding population (e.g., Albrecht et al., 2011; Legarra et al., 2008; Ly et al., 2013). It might be better to use a breeding population directly as a training population and use fitted genetic breeding values for selection (i.e., estimate breeding values for selection candidates by using the models trained from the selection candidates). Simulation study showed that this method attained high selection accuracy (Iwata et al., 2011). However, it is noteworthy that the training of a prediction model from a breeding population has time penalty, in which it takes the same time as PS to conduct GS because it is the necessary to collect phenotypic data for the training (Hickey et al., 2014; Iwata et al., 2011). Figure 2.2 starts from a training population and a breeding population, and then, the next generation of breeding population succeeds. It is also possible to update the prediction model along with the selection cycles. It is suggested that updating of the prediction model is necessary (e.g., Iwata et al., 2011; Jannink, 2010). Additionally, the size of a training population is important for GS (Hayes et al., 2009). However,

in general, it is difficult to collect data from a large training population that is closely related to the breeding population. This problem is obvious when historical records of breeding lines are used for the training of a prediction model. The reason is that we cannot control genetic structure of a training population included in historical records, whilst we can control genetic structure in a purpose-built training population to a certain extent. Nevertheless, some GS studies used historically collected information to build a prediction model (Lin et al., 2014). The methods to select a suitable set of genotypes for a training population have been developed to build a prediction model using existing information (e.g., Rincent et al., 2012).

Another important matter is the level and range of LD in a training population and a breeding population. For GS to work, at least the single markers must have sufficiently high LD with QTL controlling the target trait (Hayes et al., 2009). Meuwissen et al. (2001) concluded that the marker should be close to QTL so that the level of LD between adjacent markers and QTL could be maintained at the level where the product of the effective population size and the recombination rate was more than 2. The relationship between genetic and/or physical distances on one hand and the degree of LD on the other is different among populations (Hamblin et al., 2011). Flint-Garcia et al. (2003) reported the large difference of the extension of LD among plant species. They mentioned that the mating system of species (e.g., selfing and outcrossing) influences pattern of LD strongly. Gupta et al. (2005) reviewed studies that analyzed LD in various plant species and showed that the levels of LD were different not only between species but also within species. Remington et al. (2001) showed that the range of extension of LD was different among genes even in the same population. It is suggested that different plant species and populations that have different effective number of independent chromosome segments have different levels of prediction accuracy (Lin et al., 2014). Therefore, the levels of LD should be taken into account when we conduct GS for the target species (or the target breeding population). Heterogeneity of LD within species also decreased the degree of relationship between the training population and the breeding population. A training population should be similar to the pattern of LD in a breeding population (Nakaya and Isobe, 2012).

The prediction accuracy also depends on the heritability (Luan et al., 2009) of the target trait and the distribution of QTL effects (Hayes et al., 2009). In contrast to the factors mentioned in the previous paragraphs, these two factors are not under control. However, we may able to manage these two factors by having a large training population size and by selecting an optimal statistical method for building a prediction model. For a breeding population with low heritability, the larger the training population size is, the higher the

prediction accuracy becomes, when the observed heritability and the number of loci involved are fixed (Daetwyler et al., 2008). For the distribution of QTL effects, it might be better to choose a suitable statistical method according to the number of QTL, the size of QTL effects, and the levels of non-additive genetic effect (i.e., dominance effects and epistatic effects). For breeding populations with different genetic architecture, a statistical method represented different prediction accuracies (e.g., Daetwyler et al., 2013). I mention about the choice of the statistical methods in Section 2-2 in this chapter.

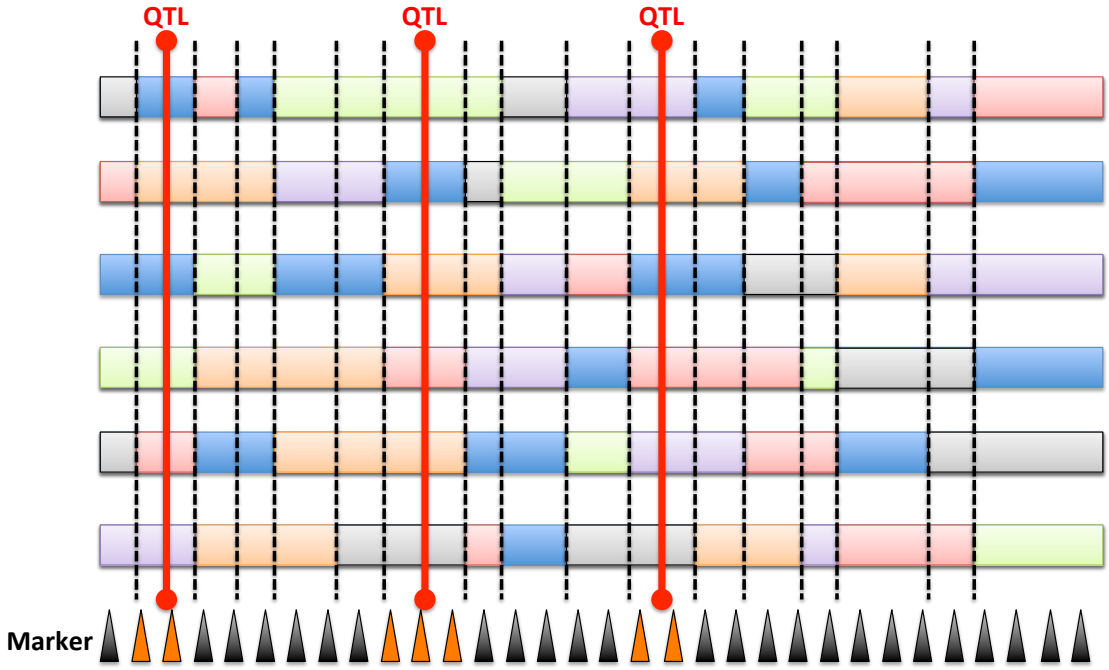


Figure 2.1. LD between QTL and markers in GS. Colored bars represent chromosome segments and different colors represent linkage block derived from different ancestors. Dashed lines show the breakpoints of linkage blocks in the population. Red lines represent the positions of QTL. Triangles in the bottom were the positions of markers used in GS, and orange triangles represent the markers in LD with QTL.

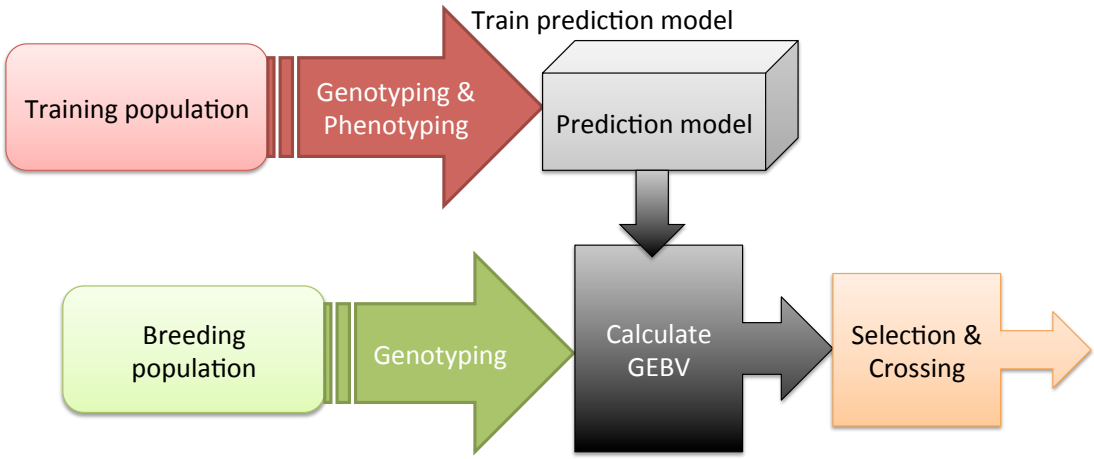


Figure 2.2. GS process starting from the training population and breeding population (i.e., selection candidates). The prediction model will be used in following generations derived from the breeding population shown here. It is also possible to update the prediction model along the way of selection cycles. This figure is modified from Heffner et al. (2009).

2-2. Statistical methods

The challenge of GS is to estimate a large number of marker effects (i.e., predictor effects, *p*) from a small number of phenotypes (i.e., observations, *n*), that is known as "large *p*, small *n*" problem (Lorenz et al., 2011). In this situation, all predictor effects cannot be estimated in ordinal least squares estimation. Even if all marker effects can be estimated, the over-fitted model will be built because of high multicollinearity among the predictors. Meuwissen et al. (2001) proposed best linear unbiased prediction (BLUP) using marker allelic effects and two Bayesian methods to solve the "large *p*, small *n*" problem, and concluded that these methods were effective to estimate marker effects.

In this section, I introduce some statistical methods that are used in GS. Although there are a number of statistical methods used for genomic prediction, I focus mainly on the methods that were used in this dissertation.

2-2-1. Ordinal least squares estimation

The relation between the marker genotype and phenotype has the following form of linear regression:

$$y_i = \mu + \sum_{j=1}^{p} m_{ij}\alpha_j + \varepsilon_i \qquad [2.1]$$

where y_i denotes the phenotypic value of plant i (i = 1, 2, …, n), μ stands for the overall mean, and α_j represents the genetic effect of the marker j (j = 1, 2, …, p). In addition, m_{ij} denotes the genotype of marker j for plant i. ε_i represents the model residuals assumed to follow N (0, σ_ε^2). In matrix notation, the equation [2.1] can be expressed as:

$$y = M\alpha + \varepsilon \qquad [2.2]$$

where $\mathbf{y} = \{y_i\}$ is a vector of phenotypes, $\mathbf{M} = \{\mathbf{1}, \mathbf{m}_1, \ldots, \mathbf{m}_p\}$ is an incidence matrix for the vector of regression coefficients, $\boldsymbol{\alpha} = (\mu, \alpha_1, \ldots, \alpha_p)'$, and $\boldsymbol{\varepsilon} = \{\varepsilon_i\}$ is a vector of residuals.

The estimated $\boldsymbol{\alpha}$ is obtained by solving the optimization problem to minimize the residual sum of squares. The estimated coefficients, $\hat{\alpha}$, is expressed as:

$$\hat{\alpha} = [M'M]^{-1}M'y \qquad [2.3]$$

However, when the number of predictors (p) is large relative to the size of observations (n), a high degree of multicollinearity among predictors exists, and over-fitted model is produced. Moreover, when $p > n$, all coefficients cannot be estimated enough because **M'M** is singular. Although the equation [2.3] can be solved by using a generalized inverse of **M'M**, such model has a similar problem to the model with large number of predictors.

2-2-2. Ridge regression

Ridge regression (Hoerl and Kennard, 1970) is also known as random regression best linear unbiased prediction (RR-BLUP). Ridge regression is one of the shrinkage estimation methods. In general, shrinkage estimation involves a penalty term. The optimization problem is expressed as:

$$\hat{\alpha} = \underset{\text{argmin}}{} L_{data}(y,\alpha) + \lambda L_{\text{mod}el}(\alpha) \qquad [2.4]$$

where $L_{data}(\mathbf{y}, \boldsymbol{\alpha})$ is the loss function depending on the residuals of data, $L_{model}(\boldsymbol{\alpha})$ is the regularization term that infers the model complexity, and λ is the regularization parameter that controls the trade-off between $L_{data}(\mathbf{y}, \boldsymbol{\alpha})$ and $L_{model}(\boldsymbol{\alpha})$. In ridge regression, $L_{data}(\mathbf{y}, \boldsymbol{\alpha})$ is the residual sum of square as is the case with ordinal least squares estimation, and $L_{model}(\boldsymbol{\alpha})$ is the sum of squares of the regression coefficient. Therefore, the equation [2.4] can be expressed as:

$$\hat{\alpha} = \underset{\text{argmin}}{} (y - M\alpha)'(y - M\alpha) + \lambda\alpha'\alpha \qquad [2.5]$$

By solving this optimization problem, the estimated coefficients is expressed as:

$$\hat{\alpha} = [M'M + \lambda I]^{-1} M'y \qquad [2.6]$$

where **I** is an identity matrix.

To chose a suitable λ, cross-validation is occasionally performed to search λ that produces the minimum error of model. Another common method to chose λ is using the ratio of the residual variance and the common marker effect variance, $\lambda = \sigma_\varepsilon^2 / \sigma_\alpha^2$. This is the equivalent to solve BLUP of **α** assuming that the regression coefficients are independently derived from a common normal distribution with mean of zero, $\alpha_j \sim N(0, \sigma_\alpha^2)$.

By adding the penalty term, the estimated coefficients shrinks toward zero, meaning that the estimated coefficients have bias. However, it solves the "large *p*, small *n*" problem and works better than ordinal least squares estimation.

2-2-3. Least absolute shrinkage and selection operator

Least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) is another penalized estimator that uses the equation [2.4]. In LASSO, while $L_{data}(\mathbf{y}, \boldsymbol{\alpha})$ is the residual sum of square in common with ordinal reast squares estimation and ridge regression, $L_{model}(\boldsymbol{\alpha})$ is the L1 penalty. Thus, the equation [2.4] can be expressed as:

$$\hat{\alpha} \underset{\text{argmin}}{=} (y - M\alpha)'(y - M\alpha) + \lambda \sum_{j=1}^{p} |\alpha_j| \qquad [2.7]$$

To choose a suitable λ, cross-validation is performed to search λ that produces the minimum error of model.

The estimated coefficients shrinks toward zero like ridge regression. In LASSO, the coefficients that are close to zero are estimated as zero, thus LASSO is occasionally used for variable selection in high dimensional feature space. However, for the variable selection, the bias approximately of size λ might be problematic (Fan and Lv, 2010).

2-2-4. Elastic net

Elastic net (Zou and Hastie, 2005) is the method that tries combining the good features of ridge regression and LASSO. In elastic net, $L_{model}(\boldsymbol{\alpha})$ possess the weighted average of the L1 and L2 penalty as:

$$L_{model}(y, \alpha) = \gamma \sum_{j=1}^{p} |\alpha_j| + (1 - \gamma) \sum_{j=1}^{p} \alpha_j^2 \qquad [2.8]$$

for $0 \leqq \gamma \leqq 1$.

Elastic net can be used for variable selection just like LASSO, but elastic net works better than LASSO when $p > n$ owing to the feature of ridge regression.

2-2-5. G-BLUP

For the form as [2.2], if it is assumed that $\alpha_j \sim N(0, \sigma_\alpha^2)$, $\boldsymbol{\alpha}$ can be solved by maximizing the joint probability of $\mathbf{y}$ and $\boldsymbol{\alpha}$. Therefore, the optimization problem can be described as:

$$\hat{\alpha} \underset{\text{argmax}}{=} \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2}(y - M\alpha)'(y - M\alpha) - \frac{1}{2\sigma_\alpha^2}\alpha'\alpha \right\} \qquad [2.9]$$

by:

$$\begin{cases} y | \alpha, \sigma_\varepsilon^2 \sim N(M\alpha, I\sigma_\varepsilon^2) \\ \alpha | \sigma_\alpha^2 \sim N(0, I\sigma_\alpha^2) \end{cases} \qquad [2.10]$$

If the equation [2.9] is solved, the same solution as [2.6] is obtained. On another front, the joint density of $\mathbf{y}$ and $\boldsymbol{\alpha}$ is represented as:

$$\begin{bmatrix} y \\ \alpha \end{bmatrix} = MVN\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} MM'\sigma_\alpha^2 + I\sigma_\varepsilon^2 & M\sigma_\alpha^2 \\ M'\sigma_\alpha^2 & I\sigma_\alpha^2 \end{pmatrix} \right] \qquad [2.11]$$

The conditional mean vector of **α** can be obtained as:

$$E\left[\alpha\middle|y\right]=0+M'\sigma_\alpha^2\left(MM'\sigma_\alpha^2+I\sigma_\varepsilon^2\right)^{-1}\left(y-0\right) \quad [2.12]$$

according to the property of multivariate normal density (e.g., Sorensen and Gianola, 2002). Therefore, the BLUP of **α** is:

$$E[\alpha]=M'[MM'+\frac{\sigma_\varepsilon^2}{\sigma_\alpha^2}I]^{-1}y \quad [2.13]$$

This solution is equivalent to the solution described in the equation [2.6]. It is noteworthy that n × n inverse matrix must be calculated to solve [2.11] although p × p inverse matrix is involved in [2.6]. In the context of GS, the dimension of the matrix to calculate inverse is crucial because GS occasionally involved several hundreds of thousands of markers despite several hundreds samples.

G-BLUP can be solved in a different form as below in the context of GS:

$$y=X\beta+Zu+\varepsilon \quad [2.14]$$

where **β** is a vector of fixed effects, **u** is a vector of random genotypic values with $Var[\mathbf{u}] = \mathbf{K}\sigma_u^2$. **K** represents the realized additive relationship matrix calculated from marker genotypes. **X** and **Z** represent design matrices for the fixed effects and the random effects, respectively. **ε** is a vector of the error deviations, and its variance is $Var[\boldsymbol{\varepsilon}] = \mathbf{I}\sigma_e^2$. This model can be solved in similar way to the solution of G-BLUP described in [2.11]. In additional, VanRaden (2008) discussed some methods of G-BLUP.

2-2-6. Bayesian methods

To solve the "large *p*, small *n*" problem, the shrinkage estimations are used commonly. In fact, in this dissertation, these shrinkage estimation methods were used. On another front, Bayesian methods are used in many studies. In Bayesian methods, the levels of shrinkage are controlled by the prior distributions that belong to marker effects. In Bayesian ridge regression and Bayesian LASSO (Park and Casella, 2008), the prior of marker effects is normal distribution and a double-exponential distribution, respectively. In additional, there are other Bayesian methods called as Bayesian alphabets, such as Bayes A (Meuwissen et al., 2001), Bayes B (Meuwissen et al., 2001), Bayes C and Bayes Cπ (Habier et al., 2011). In later three methods, variable selection mechanism is implemented in a Bayesian way.

2-2-7. Choice of methods

There are many statistical methods for GS including the methods that are not mentioned in this Chapter. We must choose one statistical model from them according to the condition of target plants and traits.

Ridge regression is well suited to the traits that are controlled by many genes with small effects because ridge regression assumes all markers possessing effects. In ridge regression, the estimated effects are distributed in whole genome, that is, the effects are scattered overall. So, if breeders know on ahead that the target traits are controlled by several genes having relatively large effects, it is better to choose LASSO or Bayesian methods with variable selection. Elastic net can be used to balance between the shrinkage and variable selection. Reproducing kernel Hilbert spaces (RKHS; Gianola et al., 2006) is one of the semi-parametric regression methods, which is effective to catch non-additive gene effects (i.e., dominance and epistatic effects) because of its flexibility. However, the choice of the kernel is one of the challenges (de los Campos et al., 2010). Machine-learning methods, such as random forest, are good for capturing the epistatic effects because they can build a non-linear prediction model (Jannink et al., 2010). As the methods to reduce dimension of predictors, partial least squares regression (PLSR) and principal component regression (PCR) are known. They search the latent variables to represent the variation of the original variables and the relationship between variables and response in PCR and PLSR, respectively.

In animal breeding, the pedigree information is recorded in detail especially for beef cattle, dairy cattle and horse. Misztal et al. (2009) proposed a prediction method that utilizes pedigree information in addition to phenotypic and genomic information. Aguilar et al. (2010) showed the efficiency of this method in Holsteins. Chen et al. (2011) suggested the efficiency of this method by involving all individuals' information even when some individuals do not have genotypic information. In plant breeding, this method also has a possibility to represent higher accuracy than the ordinal ways if the breeding population has their pedigree information.

There are some studies that compared the efficiency among several statistical methods for GS. Moser et al. (2009) compared five methods (i.e., ordinal least squares estimation, ridge ridression, Bayes A, support vector regression, and PLSR) for prediction of protein percentage and Australian selection index in dairy bulls and showed that least squares estimation worked less accurately than the other methods and that support vector regression attained highest accuracy among other four methods. Heffner et al. (2011) compared prediction accuracy in several models for 13 traits of wheat and showed the similar prediction accuracy among all

methods. Heslot et al. (2012) compared several statistical methods in wheat, barley and maize data sets and analyzed the resemblance among the methods. They discussed that there are differences between these models in the levels of over-fitting and computational times even when the prediction accuracy is similar among them. Consequently, they recommended to use Bayesian LASSO, weighted Bayesian shrinkage regression (Hayashi and Iwata, 2011), and random forest. However, they indicated the efficiency of ridge regression by taking into account of general condition.

In this dissertation, I used ridge regression (G-BLUP) and LASSO. The various statistical methods may not cause much difference in selection accuracy as mentioned above, because I assumed only additive genetic effects in the simulations. Moreover, I focused on the breeding process rather than the evaluation of one cycle of selection in this dissertation. Therefore, I chose these methods that are used commonly in GS studies.

2-3. Genomic selection in plant breeding

2-3-1. Current situation

After the introduction of GS by Meuwissen et al. (2001), the advantages of GS have been shown in breeding of livestock, especially in dairy cattle breeding. In dairy cattle breeding, the traditional breeding scheme is time consuming and cost intensive because it depends on pedigree and phenotypic information of the sire and his relatives, thus the potential of GS was investigated in detail shortly after the introduction of GS (Jonas and de Koning, 2013). It is shown that GS has a high potential for dairy cattle breeding. Empirical studies suggested that GS attained high efficiency even for the selection of complex traits such as protein yield, milk yield, fat yield, and selection indices (Hayes et al., 2009; Luan et al., 2009; VanRaden et al., 2009). For other livestock species, strategies of GS breeding are still being examined, because breeding systems differ between livestock animals (Jonas and de Koning, 2013).

Also for plant breeding, simulation and empirical GS studies have been reported. As I mentioned in Chapter 1 of this dissertation, each plant species has a suitable breeding scheme according to a large number of factors such as mating and propagation systems and the types of cultivars released to market. In this section, I introduce some of the previous simulation studies of GS in plant breeding. By reviewing the previous studies, the difference of assumptions about breeding populations and breeding schemes according to the targeted plant species will be confirmed.

There are a number of simulation studies that were conducted on the basis of the possibility of utilizing inbred lines or pure lines (or doubled haploids). Jannink (2010) simulated GS breeding of barley, which is an autogamous species, by using a real marker genotype data. This simulation study assumed that an initial breeding population was also used as a training population and that a prediction model was updated by using doubled haploids derived from individuals in the previous generation. This timing of model updating was based on the necessary time to create and evaluate the doubled haploids. This study suggested the necessity of updating a prediction model and the difficulty of long-term selection. Bernardo and Yu (2007) and Mayor and Bernardo (2009) conducted GS simulation assuming maize breeding. For maize, which is an allogamous species, phenotypic selection is generally conducted based on testcross between inbred lines and a tester. They followed this basic way. In their simulations, two parental inbred lines were assumed at first. After crossing these inbred lines, the simulated F_2 or doubled haploids are evaluated by testcross. This testcross result was used as a training data to build a prediction model. These studies showed that GS was more effective than marker

assisted recurrent selection (MARS) that used only significant markers for selection. The advantage of GS over MARS was obvious when the number of QTL controlling the target trait was large. McClosky et al. (2013) also simulated GS breeding in a bi-parental breeding population and evaluated the efficiency of GS. After crossing two parental inbred lines, several generations of selfing or creating doubled haploids were simulated. They assumed high levels of LD in the breeding population because the population was derived from a bi-parental cross. As the result, the breeding population that experienced selfing more times attained higher response to selection, because selfing increased the genetic variance in the breeding population. The number of selfing that a training population experienced did not affect to the prediction accuracy.

For the allogamous species in which selfing is not assumed, GS simulations followed the different types of breeding schemes from the simulations mentioned above. Iwata et al. (2011) simulated GS in forest tree species *Cryptomeria japonica*. They simulated an initial population that had a low level of LD to represent the situation of allogamous species with large population size. They suggested that GS could attain higher gain than PS in early generations but that updating of a prediction model was required to maintain the efficiency of GS. Wong and Bernardo (2008) simulated GS in oil palm and suggested the advantage of GS over PS and MARS considering the gain per unit time and cost. However, they assumed that the training and initial population were derived from inbred lines. This assumption was not realistic for breeding of oil palm. In these two simulation studies, the common point was that GS showed superiority to PS owing to the possibility of accelerating generations in GS breeding. For tree breeding, one of the problems in phenotypic selection is the necessity of a long time to evaluate phenotype from the seedling period.

Most of empirical studies of GS in plant breeding are based on evaluation of prediction accuracy through cross-validation. Resende Jr et al. (2012) evaluated the prediction accuracy of GS using marker and phenotypic data of *Pinus taeda*. Their target traits were diameter and height, which were evaluated in four environments and at multiple ages. They showed that phenotypic data obtained in the different places from the target place worked worse as the training population than the data obtained in the identical place, and that the early generation's phenotypic data did not work well to predict breeding values in later generation. They suggested the advantage of GS over PS considering the effects of promoting generations in GS. It would be possible only if the prediction model could work during selection cycles, but which might be difficult. Resende et al. (2012) also evaluated the prediction accuracy of GS in a forest tree

species, *Eucalyptus*, using two unrelated elite breeding populations. They showed that a prediction model built using data of one population predicted the ability of individuals in another population poorly although the genomic regions explaining trait variation coincided between these populations. For fruit tree, Kumar et al. (2012) evaluated the prediction accuracy of six fruit quality traits in apple using progenies derived from four female parents and two pollen parents. They suggested the superiority of GS breeding over BLUP-based PS breeding per unit time considering the effect of accelerating generations in GS, as is the case of Resende Jr et al. (2012). For allogamous species other than forest and fruit trees, the efficiency of GS has been investigated in empirical studies. de Oliveira et al. (2012) evaluated the prediction accuracy of GS using 358 accessions of cassava. They also suggested the importance of increasing gain per unit time by accelerating generations in GS, as is the case of tree breeding. Ly et al. (2013) also evaluated the prediction accuracy in cassava, and showed the difficulty to predict breeding values in a breeding population using a prediction model built using a training population that was not closely related to the breeding population. Even in breeding of allogamous plant species, the potential to create hybrids is evaluated by testcrosses in maize breeding. Thus, in the evaluation of the prediction accuracy of GS, the performance of inbred lines was evaluated by testcrosses. Albrecht et al. (2012) evaluated prediction accuracy using doubled haploids of maize. Phenotypic values were evaluated by testcrosses for grain dry matter yield and grain dry matter content. They suggested the efficiency of GS, but showed that the prediction accuracy was lower when the training population was from the same family as the breeding population than when these populations were derived from different populations. Riedelsheimer et al. (2012) generated 570 testcrosses by crossing the 285 maize inbred lines from Dent heterotic pool with two Flint testers, and evaluated the prediction accuracy of GS using the phenotype data of the 570 testcrosses marker genotype data of 285 inbred lines. On the other hand, Zhao et al. (2012) evaluated the prediction accuracy of GS in maize using six subpopulations derived from half-diallel crosses between four Dent inbred lines. GS for autogamous species is mainly applied in wheat so far. Crossa et al. (2010) used 599 wheat lines and evaluated the prediction accuracy. They showed that a GS model could attain higher prediction accuracy than a pedigree-based model. Heffner et al. (2011) compared the prediction accuracy of GS with those of PS and MAS for 13 traits in winter wheat. They showed that GS could attain slightly higher accuracy than MAS and a little lower accuracy than PS on average of the 13 traits, suggesting that GS might increase genetic gain per unit time and cost. Rutkoski et al. (2012) compared the prediction accuracy between statistical methods of GS for Fusarium

head blight resistance traits in wheat. They suggested that GS was more efficient than MAS for most of traits examined in their study, but that selection based on the combination of QTL and the information of other traits was more efficient than GS for the content of deoxynivalenon, which is toxic to humans and is induced by the fungal pathogen that causes Fusarium head blight.

Only by reviewing the previous studies, the differences in breeding schemes and assumed levels of LD in breeding populations are recognized for each targeted plant species. These studies suggested that the impact of breeding schemes (e.g., timing of model updating, the number of selfing in a breeding population, design of a training population, and the genetic relationship between a training population and a breeding population) on the outcomes of GS breeding was large. The results underscore the importance of simulation studies to analyze an outcome of planned scheme and to find a suitable scheme under the constraints supposed in a targeted plant species. The empirical studies suggested that it is important to choose a training population and a statistical method to attain high prediction accuracy in GS.

2-3-2. Challenges

There are some problems to apply GS breeding to an actual breeding program. The efficiency and feasibility of GS in plant breeding can be problems in the first place. As pointed out above, we should consider the situation of each breeding population and plant species under the constraint supposed in the breeding program. It is, however, difficult to verify the efficiency of GS in each plant species for supposed target traits and breeding populations via an actual breeding program. Even if GS is a promising approach in plant breeding, an efficient breeding scheme using GS is not clear. Because GS breeding involves various factors (e.g., number of markers, a statistical method used in a prediction model, a generation interval via acceleration of generations, and a population size of a breeding population) that may have a large impact on the efficiency of GS breeding, it is difficult to choose the most efficient scheme for a supposed breeding program. For the feasibility, the cost of genotyping is a major problem under present circumstances though it has declined rapidly, suggesting that the evaluation of cost efficiency is essential for practical use of GS. On the other hand, the cost of phenotyping remains as a large problem, and is increasing due to labor and land-use expenses (Desta and Ortiz, 2014). Thus, the size of breeding population, a way for resource allocation in a breeding population, and the size and constitution of a training population are important for considering the feasibility.

The lack of empirical studies of GS breeding involving selection trials is one of the great

problems. Even if simulation studies demonstrate the efficiency of GS, there is no absolute guarantee of success of GS breeding in a field trial, because simulations are generally conducted under the assumption of some simplified processes. Empirical studies are required to verify the actual efficiency of GS, and detect new problems that are difficult to be revealed in simulation studies.

The most important thing is the prior decision of the goal of a breeding program. How much gain do we expect? How much potential loss can the breeding program absorb? We should answer the questions prior to the application of GS to various breeding schemes in plants (Jonas and Koning, 2013). To achieve a higher goal with the feasible plan, the advantages and disadvantages of GS should be clarified in a knowledgeable way. The knowledge helps breeders to set up a feasible goal and to develop a plan to achieve that. It is essential to evaluate the potential of GS in plant breeding across various situations in order to put it into practical use.

Chapter 3

Simulation study of genomic selection in allogamous plants

3-1. Introduction

Mass selection is an important breeding method for open-pollinated plant species especially at the initial stage of genetic improvement. Mass selection provides the following attractive features: (i) the practical simplicity of the procedure, (ii) selection that is applicable in each generation, and (iii) the non-straightforward relationship between a high selection intensity and a low effective population size. Mass selection, however, has an important shortcoming, low efficiency of genetic improvement (Bos and Caligari, 2008). Single-plant evaluation is one of the reasons of the low efficiency of mass selection because of the low accuracy of the phenotypic evaluation of a single plant, which is caused by a large environmental error affecting on individual plants. Mass selection provides good long-term response but limited short-term response for all these factors. The improvement in the accuracy of single-plant selection has possibilities that mass selection would become a more efficient and attractive breeding method for genetic improvement of allogamous crops. There are many allogamous plant species depending on mass selection, such as forage crop species and common buckwheat (*Fagopyrum esculentum* Moench). Genetic improvement of biomass crops, such as *Erianthus*, *Miscanthus*, switchgrass (*Panicum virgatum* L.), and Guinea grass [*Megathyrsus maximus* (Jacq.) B. K. Simon & S. W. L. Jacobs], is important as a solution for the world' energy issue and competition of resource with food crops. Most of the biomass crops that have been developed recently are also allogamous species, suggesting that mass selection is expected to play an important role in biomass crop breeding. Mass selection with improved efficiency will have a strong impact on allogamous crop breeding.

GS (Meuwissen et al., 2001) may enable us to evaluate a single plant with high accuracy because the selection is not influenced by the local environment around the selection candidates. The prediction accuracy of GS, the Pearson's correlation between the selection criterion and the true breeding value, depends on the narrow sense heritability, the number of individuals in the training population, and the number of independent loci affecting a trait for a continuous phenotype (Daetwyler et al., 2008). A large training population can provide high prediction accuracy even when the phenotype is measured with single-plant evaluation. These factors might raise the mass selection efficiency. Moreover, rapid identification at the seedling stage or even at the seed stage can shorten the breeding cycle by obviating the time necessary for field-testing. The conventional MAS is also an efficient breeding technology enabling selection without field-testing. However, QTL detected in a mapping population might not be responsible for variation in a breeding population (Strauss et al., 1992). GS is anticipated as a method that might compensate for the weakness of MAS (Heffner et al., 2010), especially for improving a trait dominated by a number of QTL.

GS may be inefficient when the LD between markers and QTL is low. The degree of LD in a breeding population is directly related to the GS accuracy. Because the degree of LD in a randomly mating population is inversely proportional the population size (Sved, 1971), lower LD are expected in an allogamous plant species population with a larger effective population size (Auzanneau et al., 2007; Fiil et al., 2011; Isobe et al., 2009; Rafalski and Morgante, 2004). This situation may cause less accurate GS in allogamous plant species. For recently developed crops, their short history of breeding might also lead to low levels of LD in breeding populations (Flint-Garcia et al., 2003; Gupta et al., 2005). Therefore, when evaluating the efficiency of mass selection with GS in an allogamous species, breeders should assume low levels of LD in an initial breeding population.

The potential of GS in allogamous crops has been studied mainly in the context of tree breeding (Grattapaglia and Resende, 2011; Iwata et al., 2011; Wong and Bernardo, 2008). Nevertheless, these studies examined perennial plant breeding. It is necessary to evaluate the potential of GS in annual plant breeding. GS has a profound effect on the acceleration of selection cycles (Grattapaglia and Resende, 2011; Iwata et al., 2011; Wang and Bernardo, 2008). Therefore, the advantage of GS over PS might become much smaller in annual plant breeding than in perennial plant breeding. Moreover, for small annual plants such as forage crops and buckwheat, many individuals can be tested in experimental fields at low cost every year. Annual and perennial plant species have different factors affecting the efficiency of GS in mass

selection of allogamous species. Therefore, this study was conducted to clarify the potential of GS breeding in annual allogamous crops.

For GS in allogamous crops, another problem is the time point at which the target trait is expressed, i.e, before or after pollination. In selection of traits expressed before pollination (e.g., plant height and tiller number), selection can be implemented before pollination by separating selected plants from other plants in a breeding population to conduct pollination only among selected plants. On the one hand, the efficiency of improvement is lower in selection of traits expressed after pollination (e.g., seed yield and seed quality) than in traits expressed before pollination. Traits expressed after pollination cannot be evaluated and selected before pollination. Therefore, plants have already been pollinated with unselected pollen parents at the time of selection. This uncontrolled pollination is insufficient for the selection of traits expressed after pollination. For example, the main target traits in breeding of common buckwheat are related to the quality and quantity of seeds, which cannot be measured before pollination. The selection after pollination limits the genetic improvement in breeding of such kind of crops. Progeny testing may avoid this inefficiency, and can be effective for both traits expressed before and after pollination (Bos and Caligari, 2008). However, it is difficult to apply a progeny test to annual plant species because of their limited capacity to maintain their fertility. Moreover, a progeny test diminishes the attractive features of mass selection (i.e., simplicity and one-generation selection cycle). GS has a possibility to improve the selection efficiency in breeding of traits expressed after pollination. Because the prediction requires only marker genotypes when the prediction model has been already built, GS can be done before pollination even for the traits expressed after pollination. Consequently, GS may have great potential to improve the efficiency of mass selection of traits expressed after pollination.

GS for traits expressed after pollination, however, has a great concern. Jannink (2010) and Iwata et al. (2011) described low prediction accuracy at GS cycles without building a prediction model in their simulation study. For this reason, in breeding of small annual plant species in which phenotyping can be conducted relatively easily, breeders can assume that the prediction model is updated every year by measuring trait phenotypes and marker genotypes of plants in a breeding population. The scheme would improve prediction accuracy because the prediction model can correspond to a changing LD pattern through repeated selections. However, these updating steps require trait-phenotyping processes for which breeders cannot select pollen parents for traits expressed after pollination. Therefore, the updating steps possess both benefits and shortcomings for improvement of traits expressed after pollination.

The present study investigated the potential of GS in mass selection breeding of annual allogamous crops by simulations. I simulated two kinds of target traits, one was the trait expressed before pollination and another was the trait expressed after pollination. In this study, I assumed breeding under the extreme situation of initial linkage equilibrium. I compared the efficiency of GS breeding with the efficiency of conventional PS or MAS breeding. For the trait expressed before pollination, I evaluated the degree to which the efficiency of GS breeding was influenced by each of the following: (1) the number of genome-wide markers, (2) the mode of inheritance of markers, (3) statistical model for building a prediction model, (4) the breeding population size, (5) the number of selection cycles per year, and (6) cost efficiency. I sought to detect the appropriate GS method for mass selection breeding of annual allogamous crops on ahead. After that, I compared GS gain between traits expressed before and after pollination, and understood the reasons of differences between these traits in GS breeding.

3-2. Methods

3-2-1. Simulated plant species and target traits

In the simulations, I assumed an annual allogamous plant species that had 10 pairs of chromosomes ($2n$ = 20). Each chromosome was 100 cM long. I assumed that the plants were complete out-crossers and monoecious (i.e., sex was not considered for selection).

As a target trait, I assumed one trait controlled by 300 QTL that had only additive allelic effects. No dominant or epistatic effect was simulated in this study because currently it is not easy to fix the dominance effects in an open-pollinating population such as a synthetic cultivar in most forage crops and buckwheat. Consequently, the heritability described above was that in the narrow sense, and genetic variance attributable to dominance effects can be regarded as a part of the environmental variance in present simulations. Heritability of the trait (h^2) was assumed to be 0.5 (only for the trait expressed before pollination) or 0.2. Unless mentioned otherwise, I assumed that h^2 = 0.2 throughout this study. The QTL positions were decided randomly on the chromosomes. The frequencies of QTL alleles were sampled independently from the uniform distribution ranging from 0 to 1. It is noteworthy that the allele frequencies were completely independent among all combinations of loci (i.e., marker loci and QTL) because linkage equilibrium was assumed in the base population (mentioned later). The effects of QTL were sampled from the gamma distribution, of which the shape parameter (k) was 0.4 and the scale parameter (θ) was 0.13. The parameter setting was based on Meuwisses et al. (2001) in which it is suggested to sample QTL effects from a gamma distribution with $k = 0.4$ so that QTL were composed of a few QTL with large effects and of many QTL with small effects. Each QTL effect (α_l: effect of lth QTL; l = 1, 2, …, 300) was sampled from the gamma distribution, with mean value of $k\theta$ and variance of $k\theta^2$. Then,

$$E\left(\alpha_l^2\right) = k\theta^2 + k^2\theta^2 \qquad [3.1]$$

When the allele frequency of a single locus is p, it can be derived that

$$E(2p(1-p)) = 1/3. \qquad [3.2]$$

Because the genetic variance of a single locus is

$$\sigma_g^2 = 2p(1-p)\alpha_l^2 \qquad [3.3]$$

(Falconer and Mackay, 1996), the expected value of the variance becomes

$$E(\sigma_g^2) = k\theta^2(1+k)/3. \qquad [3.4]$$

In the simulations, the overall genetic variance was set to 1.0 in the base population. Because

the expected value of genetic variance of single locus can be derived from dividing overall genetic variance by the number of QTL,

$$E(\sigma_g) = \frac{1}{300}. \quad [3.5]$$

Consequently, the scale parameter (θ) was calculated as

$$\theta = \frac{1}{10\sqrt{k(1+k)}}. \quad [3.6]$$

The environmental error was calculated as

$$\mathrm{var}(\varepsilon) = \frac{1-h^2}{h^2} \quad [3.7]$$

because the overall genetic variance was set to 1.0 and was assumed to have only additive effect. The size of environmental variance was kept throughout breeding cycles.

In this study, I simulated two types of traits, trait expressed before pollination and trait expressed after pollination. The difference between them, however, exists only in breeding simulation procedures.

3-2-2. Breeding simulations for trait expressed before pollination

For the present study, I assumed annual plant species, and simulated six years of three breeding procedures: phenotypic selection (PS breeding), genomic selection (GS breeding), and marker-assisted selection (MAS breeding). As assumed in a study by Bernardo and Yu (2007), I assumed up to three breeding cycles per year by using offseason nursing. The first cycle of each year was evaluated during the regular growing season when phenotypic measurements were meaningful. The second and third cycles of each year were conducted in a greenhouse or a year-round nursery, where phenotypic evaluations were not meaningful. Consequently, GS and MAS were conducted up to three cycles per year at a maximum, whereas PS only had one cycle per year. The information of each breeding strategy is summarized in Table 3.1 (a).

In PS breeding, I conducted one breeding cycle per year for the reason described above (Fig. 3.1). In each cycle, phenotyping, selection, and crossing were conducted. At the phenotyping step, except as otherwise noted, 600 plants were grown and phenotyped in the field. At the selection step, the top 10% of the 600 plants were selected according to their phenotypic values. At the crossing step, the 60 selected plants were intermated. Then 10 seeds were collected from each plant. Plants germinating from the 600 collected seeds served as plants in a breeding population (i.e., selection candidates) in the subsequent breeding cycle.

In GS breeding, I conducted up to three breeding cycles per year (Fig. 3.1). At the first cycle of each year, I built a prediction model based on phenotype and marker genotype data of all plants in a breeding population. That is, I tested plants in the field once a year also in GS breeding. At the first cycle, selection was not performed directly on the phenotypic values of plants observed in the field but on the expected breeding values calculated using the prediction model obtained at that cycle. At the second and third breeding cycles of each year, I used the prediction model obtained in the first cycle of the year for predicting breeding values from marker genotypes. In each breeding cycle, I selected the top 10% of plants as parents and intermated them to produce the next generation population in the same way as PS. The breeding population size was set as 50, 200, or 600. The training population size was equal to that of the breeding population. I set the size as less than or equal to the population size in PS to reduce the GS population size when the budgets are equal for GS and PS. The number of markers used in GS was set 100, 500, or 5000. In the present study, no marker was located at the exact positions of QTL. Except as otherwise noted, I performed the simulations under the following setting: the breeding population size was set at 200, the number of markers was 500, the mode of inheritance of the markers was co-dominant, and the number of breeding cycles per year was three.

In MAS breeding, I assumed that the true positions and true effects of five QTL with the largest effects had been known. I therefore assumed that I could know the true genotypic effects attributed to the five QTL for every plant exactly when I selected plants. When I selected the plants, QTL were weighted according to their effects. In MAS, I had three cycles of selection per year. The breeding population size was set to 200. I selected the top 10% of plants as parents for PS and GS.

In all breeding procedures, the base breeding population size was 600, which was identical to the population size of PS. To simulate low levels of linkage disequilibrium in an initial breeding population, mimicking allogamous crops with a large population size (e.g., buckwheat), I performed one cycle of PS on the base population and generated an initial breeding population by intermating the selected parents. I then started a breeding program from this initial breeding population, which was common among all breeding procedures. When the breeding population sizes were set respectively to 50 and 200, 50 and 200 founder plants were selected randomly from 600 plants in the initial breeding population.

I first compared GS breeding with MAS and PS breeding by conducting simulations under the basic simulation setting (i.e., the number of markers was 500, the mode of inheritance of

markers was co-dominant, the breeding population size was 200, the number of breeding cycles per year was three in GS breeding), with two levels of heritability (i.e., h^2 = 0.2 or 0.5). Then I conducted simulations under various conditions to evaluate the influences of (1) the number of markers, (2) the mode of inheritance of markers, (3) statistical methods for building a prediction model, (4) the breeding population size, and (5) the number of breeding cycles per year, on the efficiency of GS. The simulation was repeated 100 times for each simulation setting.

3-2-3. Breeding simulations for trait expressed after pollination

For the trait expressed after pollination, I assumed that plants were selected after pollination at PS steps and the step of selection based on marker genotypes and phenotypes (genomic and phenotypic selection: GPS). At these steps, all plants in a breeding population contributed as pollen parents, while only selected plants contributed as seed parents. Actually, GS requires no measurement of phenotypic data. Therefore, the plants were assumed to be selected before pollination at GS step. Moreover, only selected plants were assumed to contribute to the next generation's population as both pollen and seed parents (Fig. 3.2). In PS breeding, only PS was implemented. In GS breeding, GPS and GS were implemented. The information of each breeding strategy is tabulated in Table 3.1 (b).

Almost all the simulation settings for a trait expressed after pollination were similar to those for the trait expressed before pollination, which was explained above. To compare GS gains between traits requiring different timing of selection, I used one basic set of simulation settings and conducted 12 years of simulations. I adopted the following settings as simulations to compare gains between two traits expressed at different timings. I assumed that we used 500 markers in GS and GPS. I produced a base population with 200 plants, assuming that the population was in linkage equilibrium between every pair of markers. I built an initial population for PS and GS breeding by conducting one cycle of PS on the base population.

For PS breeding, I conducted PS with one cycle per year because the measurements of the target trait were always required for PS. In each cycle, phenotyping, selection, and crossing steps were done. At the phenotyping step, 200 plants were grown. Their phenotypes were measured in the field. At the selection step, the top 10% of the 200 plants were selected according to their phenotypic values. At the crossing step, in the trait expressed before pollination, the 20 selected plants were intermated randomly. In the trait expressed after pollination, the 20 selected plants were crossed randomly with all 200 plants including unselected plants in the population because pollination took place before selection in PS of the

trait expressed after pollination. Then, in both traits, 10 seeds were collected from each plant of the 20 selected plants. Plants germinating from the 200 collected seeds served as plants in a breeding population (i.e. selection candidates) in the subsequent breeding cycle.

For GS breeding, the breeding population size was 200. GPS was conducted at the first cycle of each year to renew a prediction model from phenotype and genotype data. Up to three breeding cycles per year were performed. The selection intensity was 10% (20 plants). In the trait expressed after pollination, at GPS, I selected the top 10% as seed parents based on the expected genotypic values. The selected 20 plants were crossed randomly with all 200 plants because of the same reason of PS. In the trait expressed before pollination, the selected 20 plants were intermated randomly at GPS. For GS, I selected the top 10% of plants based on predicted values calculated from marker genotypes with the prediction model renewed at the latest GPS. For GS, mating took place randomly only between selected plants in traits of both types because GS did not require the phenotyping step.

3-2-4. Prediction model for GS

I used two statistical methods, ridge regression and LASSO, to estimate the genetic effect coefficients $\boldsymbol{\beta}$ in the equation [2.2]. In ridge regression, I used two procedures to optimize ridge ridge penalty parameter λ_R in the equation [2.5] and [2.6]. One is based on the ratio of estimated genetic and environmental variance (i.e., $\lambda_R = \mathrm{var}(\varepsilon)/\mathrm{var}(\beta)$) (Iwata and Jannink, 2011; Piepho, 2009). The other is based on ten-fold cross-validation. Except as otherwise noted, I used the former procedure in ridge regression. In LASSO, I optimized lasso penalty parameter λ_L in the equation [2.7] based only on ten-fold cross-validation. In both regression methods, a prediction model was obtained using R (R Development Core Team, 2014). To estimate the ratio of genetic and environmental variances for optimizing the ridge penalty parameter, I used the R package efficient mixed-model association ("emma") (Kang et al., 2008). To perform ridge regression and LASSO with penalty parameters optimized via ten-fold cross-validation, I used the R package of "glmnet" (Friedman et al., 2010). Unless mentioned otherwise, I assumed that ridge regression was conducted for prediction.

3-2-5. Summarization of simulation results

At each selection cycle, I calculated the mean genotypic value of plants in a breeding population. The values were averaged over 100 simulations performed with a single simulation setting. To measure the genetic gain from each breeding procedure, I subtracted the mean genotypic values

of the initial breeding population from the mean genotypic values at each selection cycle so that the values of the initial breeding population were adjusted to zero. To test the significance of the difference in genetic gain between different breeding procedures, I conducted a matched-pairs Wilcoxon test by treating 100 simulations as replications.

To ascertain the efficiency of selection, I also calculated the genetic variance and prediction accuracy at each breeding cycle. The genetic variance and selection accuracy were averaged over 100 simulations. Selection accuracy was calculated as the Pearson's correlation coefficient between the phenotypic values and the true genotypic values in PS and as the correlation coefficient between the expected genotypic values and the true genotypic values in GS.

To evaluate the prediction accuracy for the genetic potential harbored by chromosomes derived separately from the seed parent and the pollen parent, the true genotypic value of chromosomes derived from the seed parent was calculated as

$$g_{is} = \sum_{l=1}^{300} z_{ils}\alpha_l \qquad [3.8]$$

where α_l stands for the true effect of QTL l and where z_{ils} represents the allele count of QTL l in chromosomes derived from the seed parent of plant i. The value of z_{ils} can be 1 or 0, corresponding to the number of QTL allele harbored by the chromosomes. The predicted genotypic value of chromosomes derived from the seed parent was calculated as

$$\hat{g}_{is} = \mu + \sum_{j=1}^{500} x_{ijs}\beta_j \qquad [3.9]$$

where μ stands for the overall mean of genetic values of all parents, β_j represents the estimated effect of marker j, and x_{ils} denotes the allele count of marker j harbored by chromosomes derived from the seed parent of plant i. The accuracy of prediction for chromosomes derived from the seed parent was therefore calculated as Pearson's correlation coefficient between g_{is} and $\hat{g}_{is}$. The accuracy for chromosomes derived from the pollen parent was calculated similarly.

To evaluate the cost efficiency of GS and PS breeding, I calculated the genetic gain per unit cost. For the calculation, I assumed that the respective costs of phenotyping and genotyping for one plant were equal to 1 and x. Then, the costs of phenotyping C_p and genotyping C_g were calculated as

$$C_p = n_p t \qquad [3.10]$$

and

$$C_g = x n_g t, \qquad [3.11]$$

where n_p stands for the number of plants phenotyped in one year, n_g signifies the number of plants genotyped in one year, and t denotes the number of years spent for genetic improvement. For example, six years of PS with the population size of 600 required phenotyping cost C_p = 3,600 and genotyping cost C_g = 0 (n_p = 600, n_g = 0, t = 6), and six years of GS selecting 600 plants per cycle and three cycles per year required phenotyping cost C_p = 3,600 and genotyping cost C_g = 10,800x (n_p = 600, n_g = 1,800, t = 6). The genetic gain per cost was calculated as

$$G = \frac{\Delta_g}{C_p + C_g}, \qquad [3.12]$$

where Δ_g stands for the difference of mean genotypic values between the 0th year (Fig. 3.1) and current breeding populations. The cost for breeding increases linearly according to increasing population size, whereas the gain might show diminishing returns from the increasing population size. Therefore, breeding with smaller population size could be more advantageous. Consequently, for evaluation of cost efficiency, I compared GS breeding only to PS breeding with identical population size to ensure a fair comparison. The results were averaged over 100 simulations.

To ascertain the change in LD pattern through breeding cycles, I calculated the measure of LD, r^2, between QTL and their adjacent polymorphic markers. LD (r^2) between loci A and B, of which the alleles are represented as A, a, B and b, was defined as

$$r^2 = \left(P_{AB} - p_A p_B\right)^2 \Big/ \left(p_A q_a p_B q_b\right), \qquad [3.13]$$

where P_{AB} was the haplotype frequency of AB, p_A and q_a respectively denote allele frequencies of A and a, and p_B and q_b respectively represent allele frequencies of B and b (Hartl and Clark, 2007). I calculated r^2 for chromosomes derived separately from seed parents and pollen parents.

Table 3.1 Information of PS, GS, and MAS breeding in a trait expressed before pollination (a) and a trait expressed after pollination (b).

(a)

	PS	GS	MAS
Number of cycles per year	1	3 (1, 2)	3
Breeding population Size	600 (50, 200)	200 (50, 600)	200
Selected size	10% of population	10% of population	10% of population
Type of selection	Phenotypic value	Predicted value by 500 (100, 5,000) markers	Genotypic value by 5 QTL

(b)

	PS	GS
Number of cycles per year	1	3
Breeding population Size	200	200
Selected size	10% of population	10% of population
Type of selection	Phenotypic value	Predicted value by 500 markers

The number in the bracket represents the number used only for comparison of the impacts of the difference of these numbers.

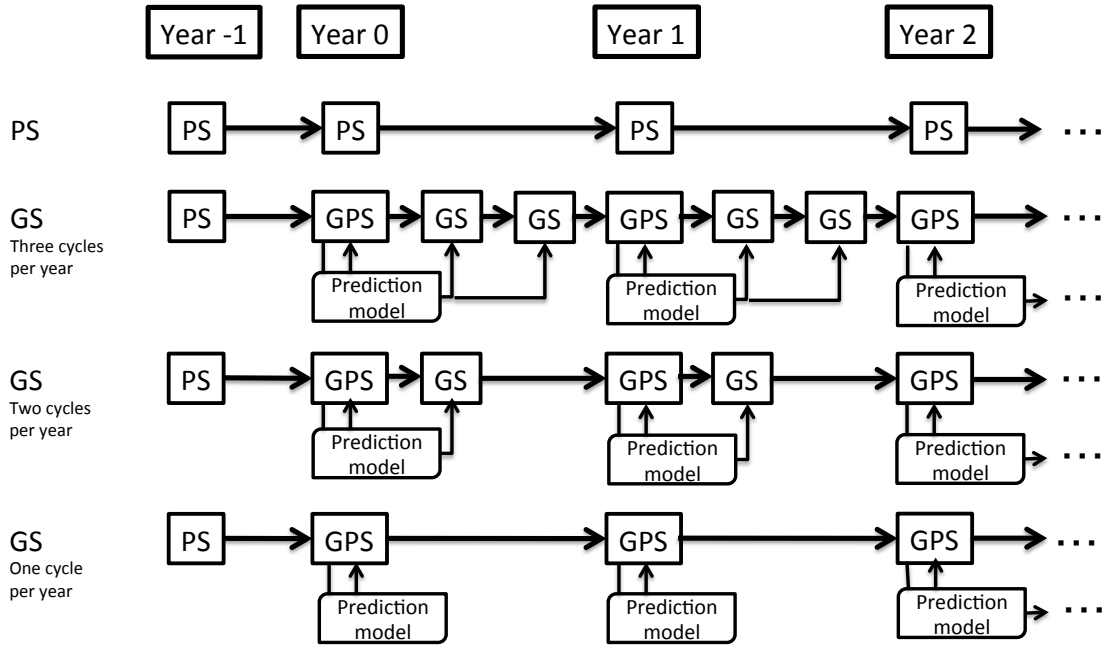


Figure 3.1. Breeding procedures compared in the simulations. In GS breeding, the prediction model built at the first cycle of each year was used in the other cycles of the year. PS: phenotypic selection, GS: genomic selection, GPS: genomic and phenotypic selection.

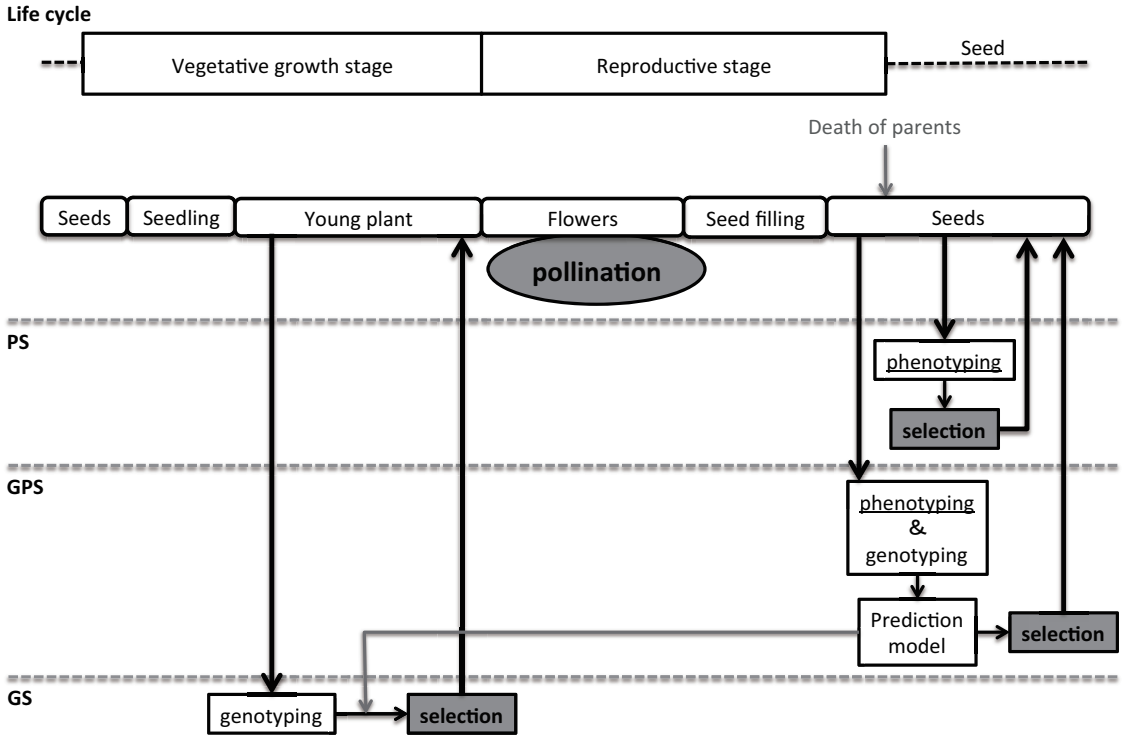


Figure 3.2. Life cycle of plants (first line) and breeding operations in the breeding simulations when the trait is expressed after pollination. PS: phenotypic selection, GS: genomic selection, GPS: genomic and phenotypic selection.

3-3. Results

3-3-1. Comparison among breeding strategies

Figure 3.3 shows the mean genotypic values of plants in a breeding population through six years of breeding under different breeding schemes (e.g., GS, MAS, or PS breeding) when $h^2 = 0.5$ (Fig. 3.3a) and 0.2 (Fig. 3.3b) for the trait expressed before pollination. The relative efficiency of different breeding schemes was similar between two different settings of heritability. Therefore, I present the results of simulations only under low heritability (i.e. $h^2 = 0.2$), except for comparison among GS, MAS, and PS breeding. The conventional MAS performed better than GS and PS during the first year but attained lower gain than the others in subsequent years. In the initial population, genetic variance explained by the five target QTL in MAS were 38.4±10.9 (%) and 41.8±12.7 (%) of total genetic variations, in average over the 100 replications of the simulation assuming that $h^2 = 0.2$ and 0.5, respectively.

3-3-2. Number of markers for genomic selection

Figure 3.4 presents the genetic gain of breeding populations for GS breeding using different number of markers (i.e., 100, 500, or 5,000 markers) for the trait expressed before pollination. GS breeding using the larger number of markers attained higher genetic gain in general. The difference between GS based on 500 and 5,000 markers at the sixth year was small but significant ($p = 0.013$). The averaged genetic gain of GS using 100 markers were much lower than that of GS using a larger number of markers especially in later breeding cycles. This is partly because most of the 100 markers were fixed in the later breeding cycles. Through the first six years of GS breeding, all 100 markers were fixed in two cases out of 100 replications of the simulation.

3-3-3. Mode of inheritance of markers

Figure 3.5 presents the genetic gain of breeding populations when I used marker systems with different modes of inheritance for GS (i.e., dominant or co-dominant markers) for the trait expressed before pollination. The genetic gain attained by GS with dominant markers was lower than that with co-dominant markers. The difference between the two marker systems was significant ($p < 0.01$) at the sixth year. The difference between the two marker systems was, however, less apparent than the difference between PS and GS breeding, and remained almost constant throughout the six years of breeding.

3-3-4. Statistical models for prediction

Figure 3.6 presents the genetic gain of breeding populations across years of selection using different statistical methods for building a prediction model for the trait expressed before pollination. For ridge regression, I used two different procedures for optimizing the penalty parameter λ_R (i.e., the procedure based on the ratio of estimated genetic and environmental variance and the procedure based on the ten-fold cross-validation). The difference in genetic gain between the two procedures was not significant at 5% level at the sixth year ($p = 0.07$). LASSO performed worse than ridge regression throughout the six years of breeding, and the difference between LASSO and ridge regression based on both procedures was highly significant at the sixth year ($p < 0.001$).

3-3-5. Breeding population size and number of selection cycles per year

Figures 3.7 shows the genetic gain of breeding populations when I had different breeding population sizes (i.e., 50, 200, and 600 plants) in GS breeding for the trait expressed before pollination. The figure also represents the gain when I had a different number of GS cycles per year. In all situations, a breeding population in PS breeding involved 600 individuals per cycle.

GS with larger population size attained higher genetic gain throughout six years of breeding. However, GS with the population size of 50 attained lower genetic gain than PS in later generations of breeding, although it remained almost comparable to PS. In GS with population size of 50, all 500 markers were fixed in 11 out of 100 replications through the six years of GS breeding. Figure 3.8 (a) shows the change of genetic variance through the six years of breeding programs under PS breeding and the different population sizes for GS breeding. The genetic variance decreased more rapidly in GS breeding with a smaller population size. Figure 3.9 (a) presents the change in prediction accuracy through the six years of breeding programs under PS and GS. The prediction accuracy in GS cycles for which I did not build a prediction model was much lower than the accuracy at GPS cycles for which I built a prediction model and used fitted values for selection. The accuracy decreased more rapidly in GS breeding with smaller population size.

When I had 200 plants in all GS breeding schemes, the difference between the genetic gains of GS breeding with three and two cycles per year was not significant at the 1% level ($p = 0.03$ at the sixth year) (Fig. 3.7b). Results show that GS breeding with one cycle per year had lower genetic gain than GS breeding with more cycles per year, but still attained significantly higher genetic gain than PS breeding ($p < 0.001$) (Fig. 3.7b). Using 600 plants per cycle, GS breeding

with a larger number of selection cycles attained higher genetic gain, and GS breeding always (i.e., under any number of cycles per year) attained higher genetic gain than PS did (Fig. 3.7c). When I had 50 plants per selection cycle, PS breeding and GS breeding with one or three cycles per year attained comparable genetic gain in early generations, but GS breeding eventually showed lower genetic gain than PS breeding did ($p < 0.001$), irrespective of the number of GS cycles per year (Fig. 3.7a). Figures 3.8 (b) and 3.9 (b) respectively show changes of genetic variance and selection accuracy for different numbers (1, 2, or 3) of GS cycles per year when the population size was 200. The selection accuracy decreased more rapidly in GS breeding with the larger number of cycles per year. As presented in Fig. 3.9 (a), the prediction accuracy for GS cycles in which a prediction model was not built was much lower than the accuracy for GPS cycles. GS breeding with smaller number of cycles per year tented to maintain higher prediction accuracy.

3-3-6. Cost efficiency

Figure 3.10 shows the cost efficiency of GS breeding relative to PS breeding with the identical population size among GS breeding and PS breeding for the trait expressed before pollination. When the number of cycles per year was three (Fig. 3.10a), GS breeding with the population size of 600 was more cost-effective than GS breeding with population sizes of 200 and 50. The genetic gain per unit cost of GS breeding with the population sizes of 600 and 200 was equal to that of PS breeding when the genotyping cost was 25 and 16 percent of the phenotyping cost, respectively (Fig. 3.10a). GS breeding with the population size of 50 was always surpassed by PS breeding in all range of the genotyping cost (Fig. 3.10a). When the population size was 600 (Fig. 3.10b), GS breeding with three cycles per year was most cost-effective when the genotyping cost was less than 17% of the phenotyping cost. GS breeding with two or one cycles per year was most cost-effective when the genotyping cost was in the range of 17% to 47% or higher than 47% of the phenotyping cost, respectively (Fig. 3.10b). GS breeding with these combinations of the population size (i.e., 600) and the number of cycles were most cost-effective among GS breeding the other combinations of population sizes (i.e., 200 or 50) and the numbers of cycles over all range of genotyping cost (Fig. 3.10c). The genetic gain per unit cost of GS breeding surpassed that of PS breeding when the genotyping cost was less than 27% of the phenotyping cost (Fig. 3.10b).

3-3-7. Timing of expression of traits

In simulations comparing the efficiency of GS for the traits expressed before and after pollination, I conducted all breeding simulations by using breeding population size of 200. By using this identical population size, I can compare breeding scenarios in an equal status without considering breeding budget.

Figure 3.11 portrays the genetic gain obtained through PS breeding and GS breeding, which were averaged over 100 replications of simulations performed for each setting. Black and gray lines represent the result of the trait expressed after pollination and the trait expressed before pollination, respectively. As expected, GS breeding showed superiority over PS breeding in a trait expressed after pollination than in a trait expressed before pollination. In PS breeding, the genetic gain was much lower in a trait expressed after pollination than in a trait expressed before pollination, depending on the possibility of pollen parent selection. When the number of breeding cycles was two or three per year, GS breeding attained almost identical genetic gain between a trait expressed after pollination and a trait expressed before pollination. At later stages, GS breeding attained higher genetic gain in a trait expressed after pollination than in a trait expressed before pollination. In a trait expressed after pollination, GS breeding attained more than twice the genetic gain of PS breeding at the end of sixth year. The difference in genetic gain between GS breeding with two cycles per year and three cycles per year was small. When the breeding cycle occurred once per year, the genetic gain of GS breeding was low. It was comparable to that of PS breeding.

Prediction accuracy of GS breeding in a trait expressed after pollination was high at GPS steps, where a prediction model was built. It decreased linearly as running over GS, where selection was conducted using the prediction model built at the last GPS (Fig. 3.12b). However, the prediction accuracy of GS in a trait expressed before pollination decreased drastically at GS immediately after GPS (Fig. 3.12a), similar to the simulations assuming six years breeding (Fig. 3.9). It is noteworthy that, in a trait expressed after pollination, pollen parents cannot be selected even though the prediction accuracy was high at GPS. Figures 3.12 (c) − 3.12 (e) show the prediction accuracy of GS breeding for chromosomes derived from a seed parent and a pollen parent in a trait expressed after pollination. For GS breeding with two or three cycles per year, the prediction accuracy for chromosomes derived from a seed parent decreased drastically at GS immediately after GPS, although the accuracy for chromosomes derived from a pollen parent was high at GS immediately after GPS (Figs. 3.12c and 3.12d). At the second GS of each year in GS breeding with three cycles per year, no difference in the prediction accuracy was found

between chromosomes derived from a seed parent and chromosomes derived from a pollen parent (Fig. 3.12c). When the breeding cycle occurred once per year, the prediction accuracy was invariably lower for chromosomes derived from a pollen parent than for chromosomes derived from a seed parent (Fig. 3.12e).

LD increased rapidly with the increase of breeding cycles (Fig. 3.13). Figure 3.13 (a) shows the generational differences of r^2 in GS breeding with three cycles per year in a trait expressed before selection. The LD was calculated for chromosomes derived separately from a seed parent and a pollen parent. GS breeding for a trait expressed before pollination showed no difference in LD pattern between those of chromosomes derived from a seed parent and a pollen parent. Figures 3.13 (b) − 3.13 (d) show LD of GS breeding for a trait expressed after pollination. When the breeding cycles were two or three per year, LD in chromosomes derived from a seed parent changed drastically at GS immediately after GPS, although LD in chromosomes derived from a pollen parent did not change greatly (Figs. 3.13b and 3.13c). When breeding cycles were one per year, LD in chromosomes derived from a seed parent was always higher than LD in chromosomes from a pollen parent (Fig. 3.13d) because the number of plants which contributed as pollen parents was much larger than those which contributed as seed parents. Given a larger population, the levels of LD are lower for the same distances of loci (Hill and Weir, 1988; Sved, 1971). Consequently, LD was low in chromosomes derived from a seed parent with a smaller effective population size. The levels of LD decreased greatly after the average distance between QTL and adjacent polymorphic (i.e., unfixed) markers became greater than 10 cM (Fig. 3.14).

Figure 3.15 (a) shows the proportion of QTL fixed in a breeding population. For a trait expressed before pollination, 64.8, 87.1, and 94.7% of QTL were fixed at the end of the twelfth year of selection, when the breeding cycles were one, two, and three per year, respectively. Among the fixed QTL, 42.0, 43.1, and 44.9% of QTL were fixed to unfavorable alleles (Fig. 3.15b). For a trait expressed after pollination, the fixation rates of QTL were 14.0, 69.6, and 87.4%, respectively, in GS breeding with one, two, and three cycles per year. They were lower than for a trait expressed before pollination (Fig. 13.5a). Among the fixed QTL, 44.9, 41.8 and 42.4% of QTL were fixed to an unfavorable allele (Fig. 3.15b). The proportion of the fixed markers used for genomic prediction followed the same patterns of fixed QTL (Fig. 3.15c). However, the proportion of fixed markers was lower than that of fixed QTL.

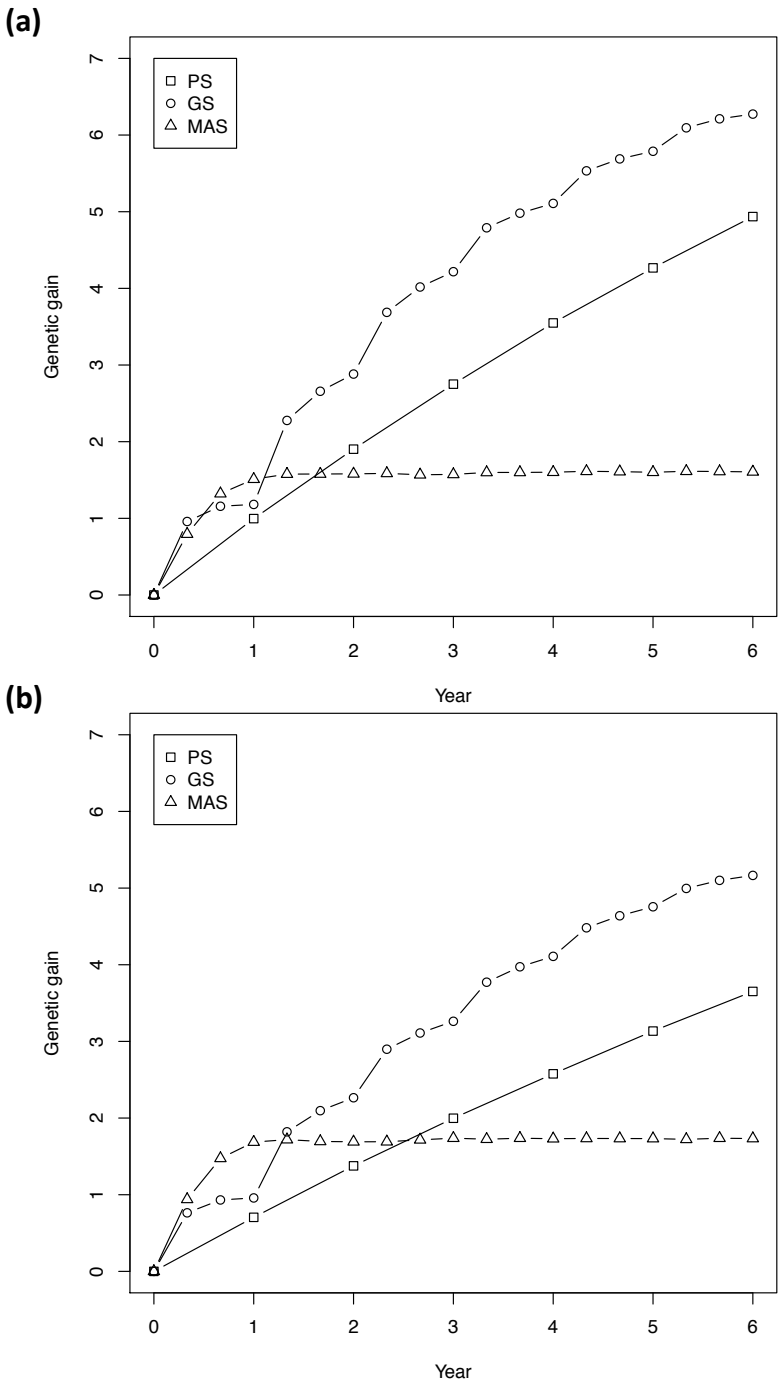


Figure 3.3. Genetic gain averaged over 100 simulations for three kinds of breeding procedures: phenotypic selection (PS), genomic selection (GS) and conventional marker-assisted selection (MAS), when h^2 = 0.5 (a) and h^2 = 0.2 (b). In PS, the breeding population size was 600. In GS, the population size was 200, and 500 co-dominant markers were used for genomic prediction. In MAS, the population size was 200, and the true effects and genotypes of largest five QTL were known.

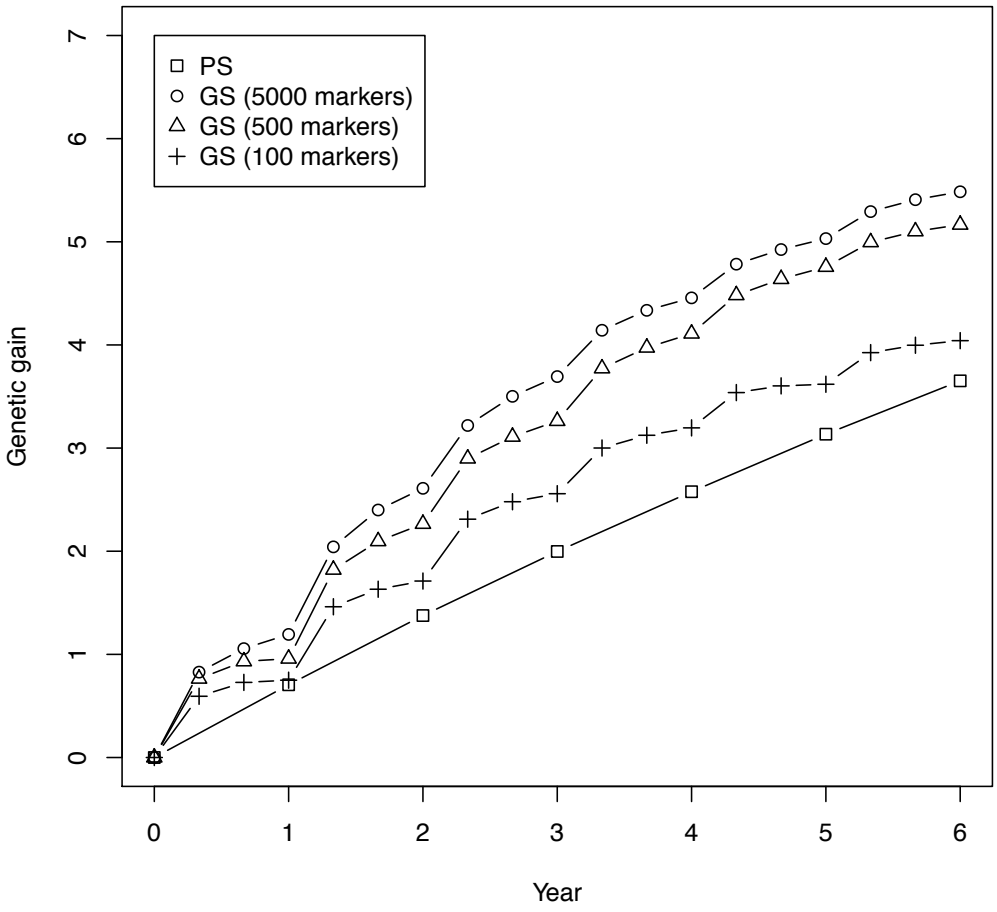


Figure 3.4. Genetic gain averaged over 100 simulations for two kinds of breeding procedures: phenotypic selection (PS) and genomic selection (GS) with three cycles per year, when the numbers of markers used for GS were 5,000, 500 and 100. The heritability of a target trait was 0.2 at the base population.

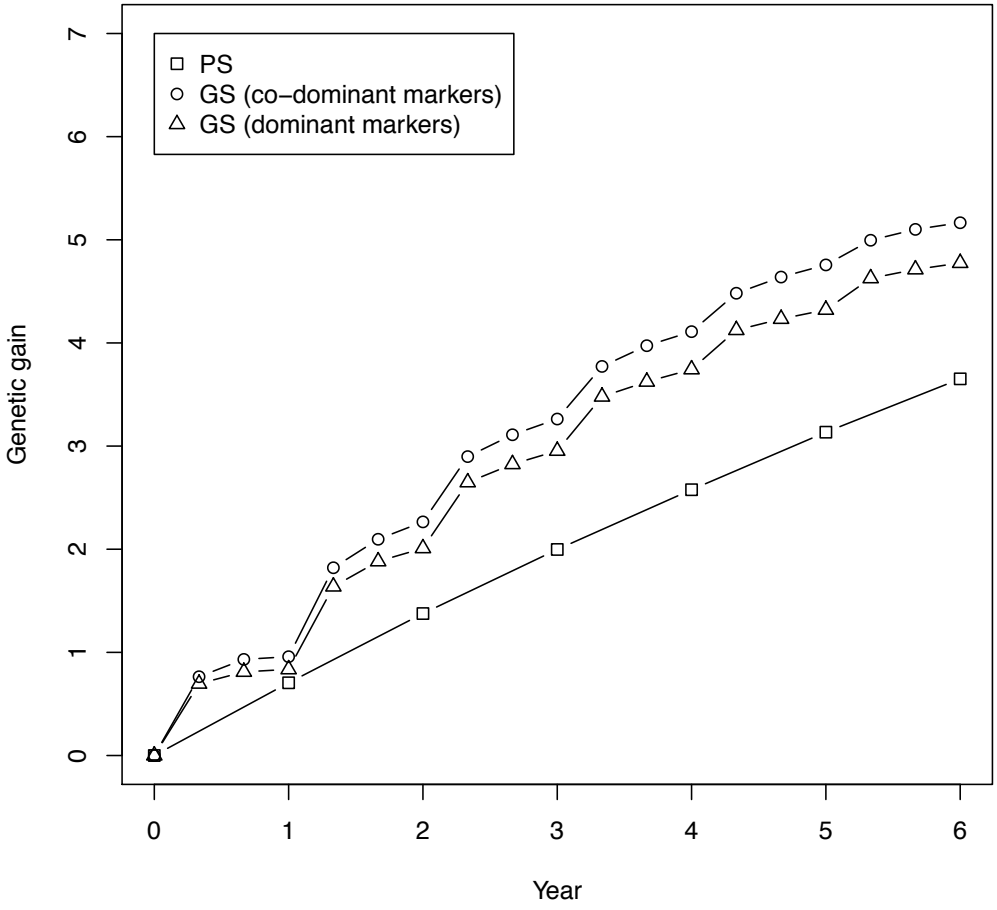


Figure 3.5. Genetic gain averaged over 100 simulations for two kinds of breeding procedures: phenotypic selection (PS) and genomic selection (GS), when the numbers of markers used for GS had two kinds of modes of inheritance: co-dominant markers or dominant markers. The heritability of a target trait was 0.2 at the base population.

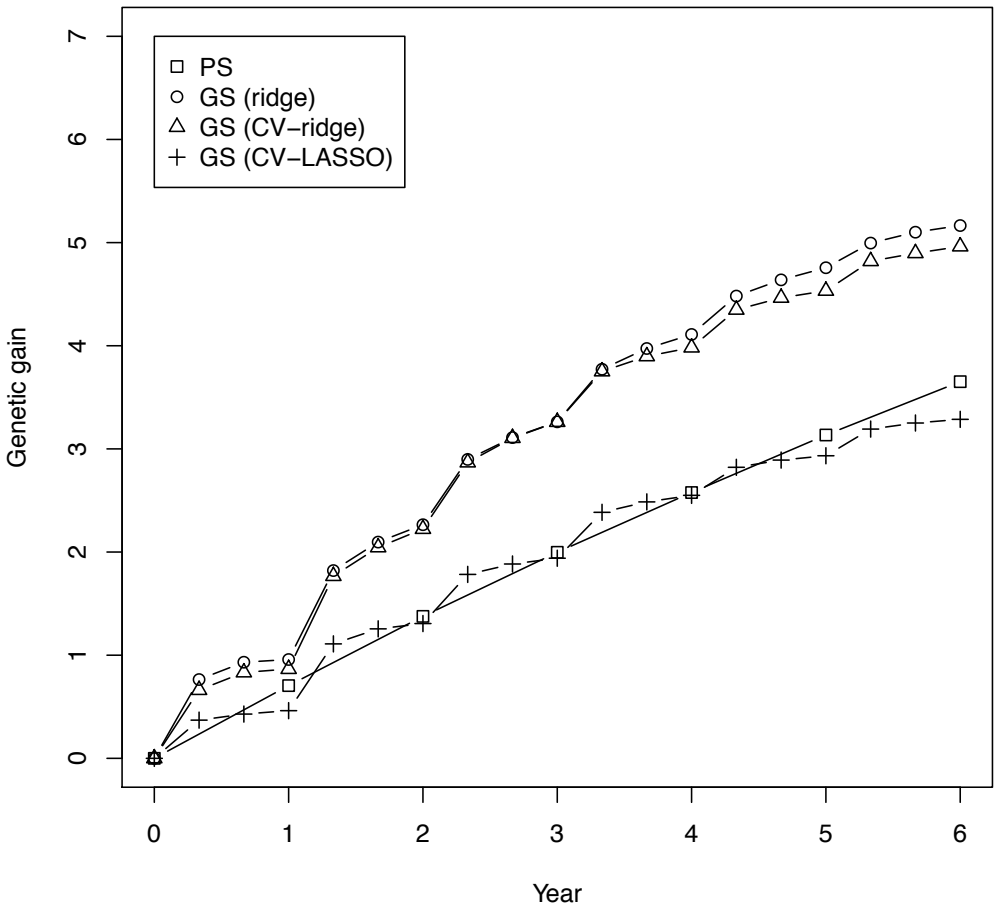


Figure 3.6. Genetic gain averaged over 100 simulations for two kinds of breeding procedures: phenotypic selection (PS) and genomic selection (GS), when the statistical methods for GS were ridge regression using the ratio of the estimated genetic and environmental variances, ridge regression using 10-fold cross validation (CV), and LASSO using 10-fold CV. The heritability of a target trait was 0.2 at the base population.

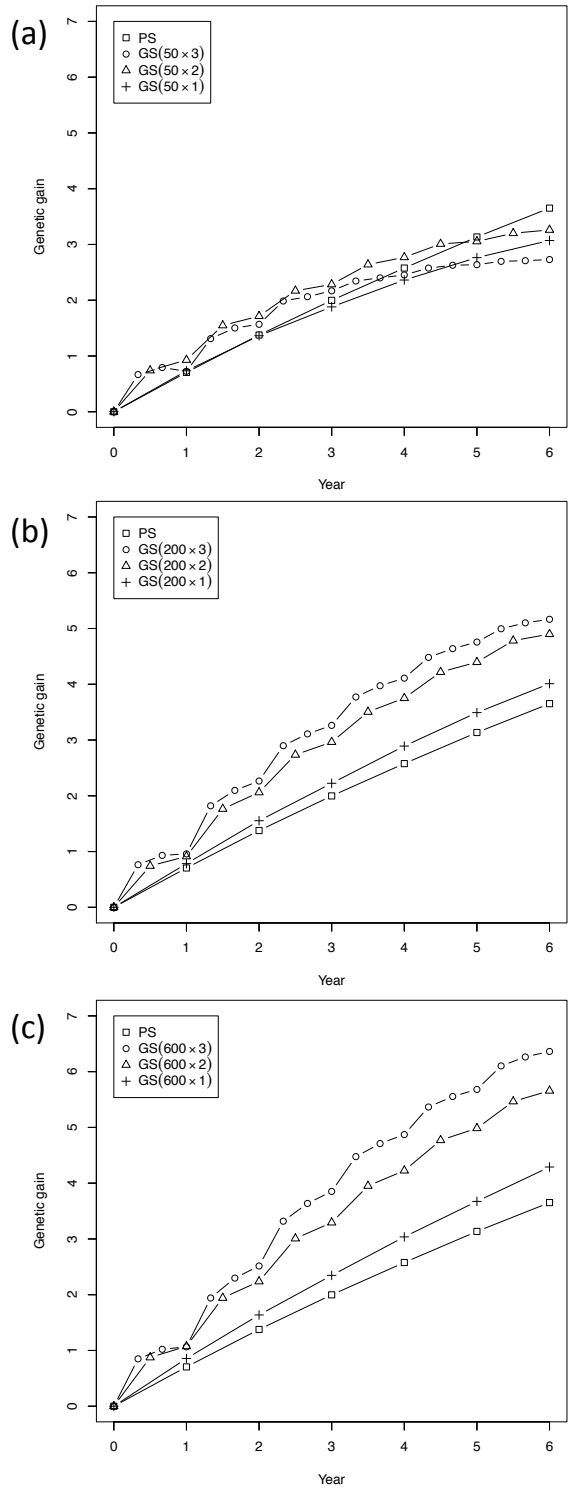


Figure 3.7. Genetic gain averaged over 100 simulations for two kinds of breeding procedures: phenotypic selection (PS) and genomic selection (GS), when the sizes of breeding population in GS were 50 (a), 200 (b) or 600 (c). GS breeding conducted one, two or three cycles per year. The heritability of a target trait was 0.2 at the base population.

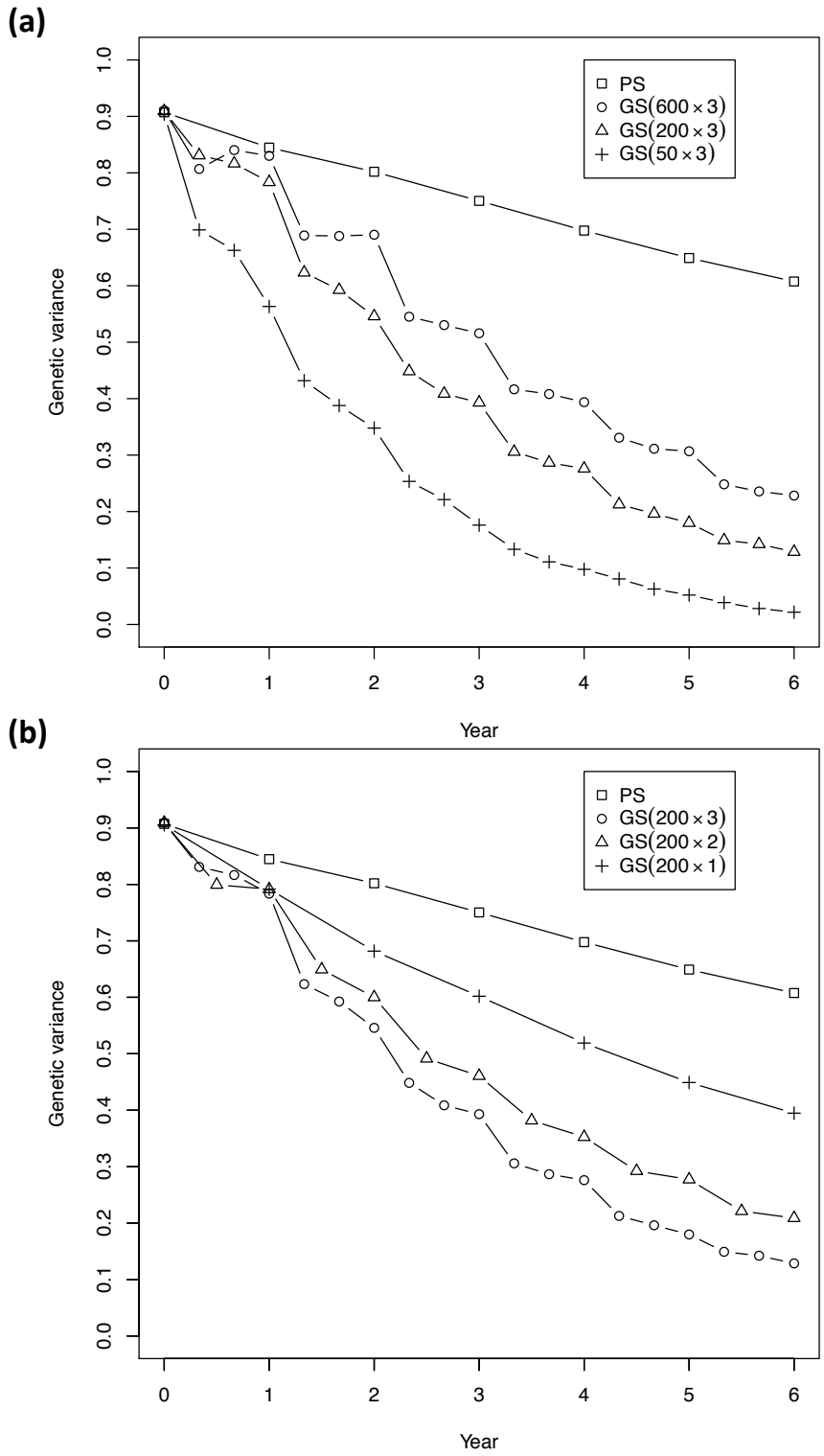


Figure 3.8. Genetic variance existing in a breeding population when the heritability of a target trait was 0.2 at the base population. (a) Comparison of the different breeding population size. (b) Comparison of the different number of selection cycles per year.

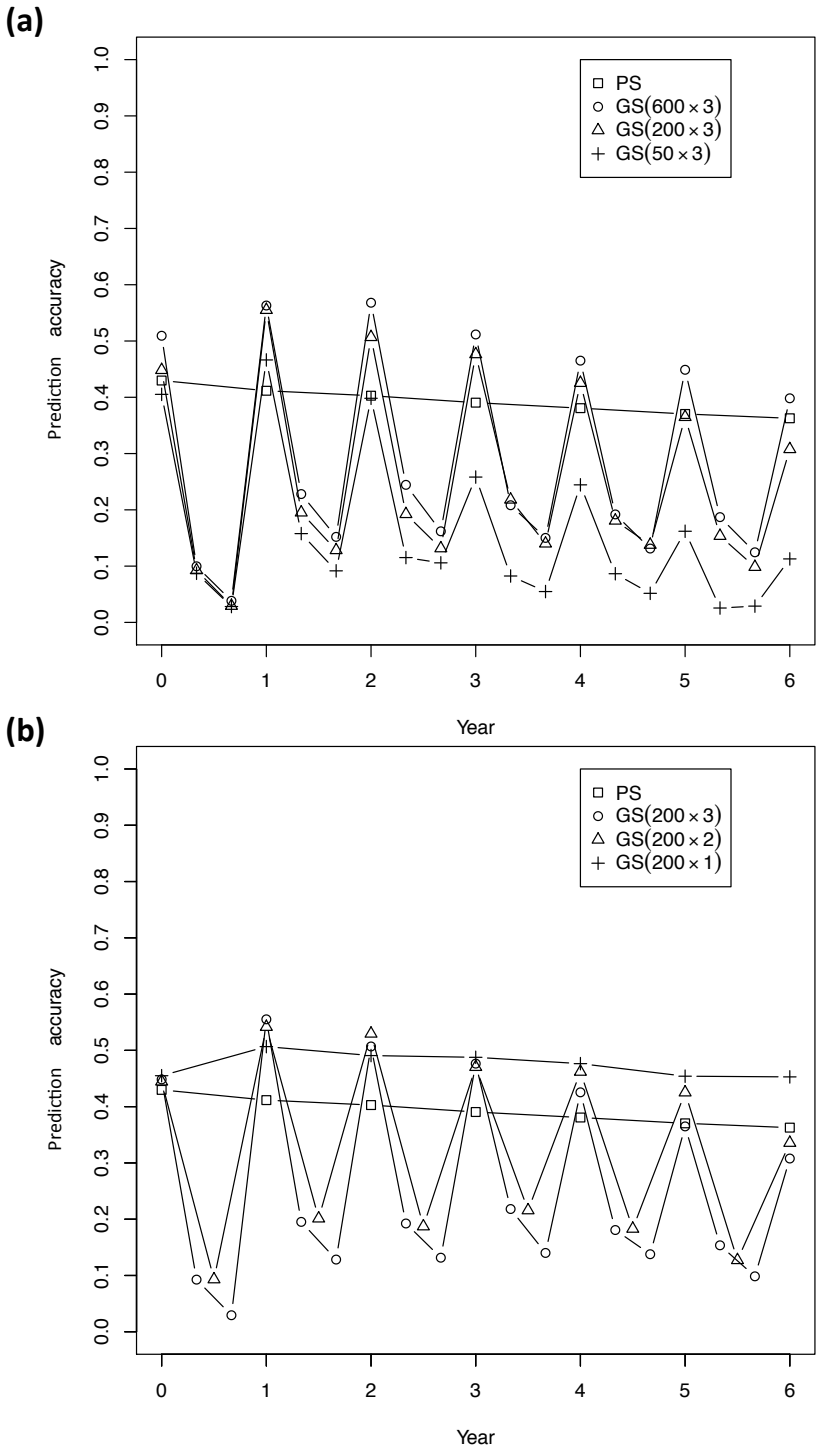


Figure 3.9. Prediction accuracy measured as the Pearson's correlation coefficient between true genotypic values and predicted genotypic values at each selection cycle. The heritability of a target trait was 0.2 at the base population. (a) Comparison of the different breeding population size. (b) Comparison of the different number of selection cycles per year.

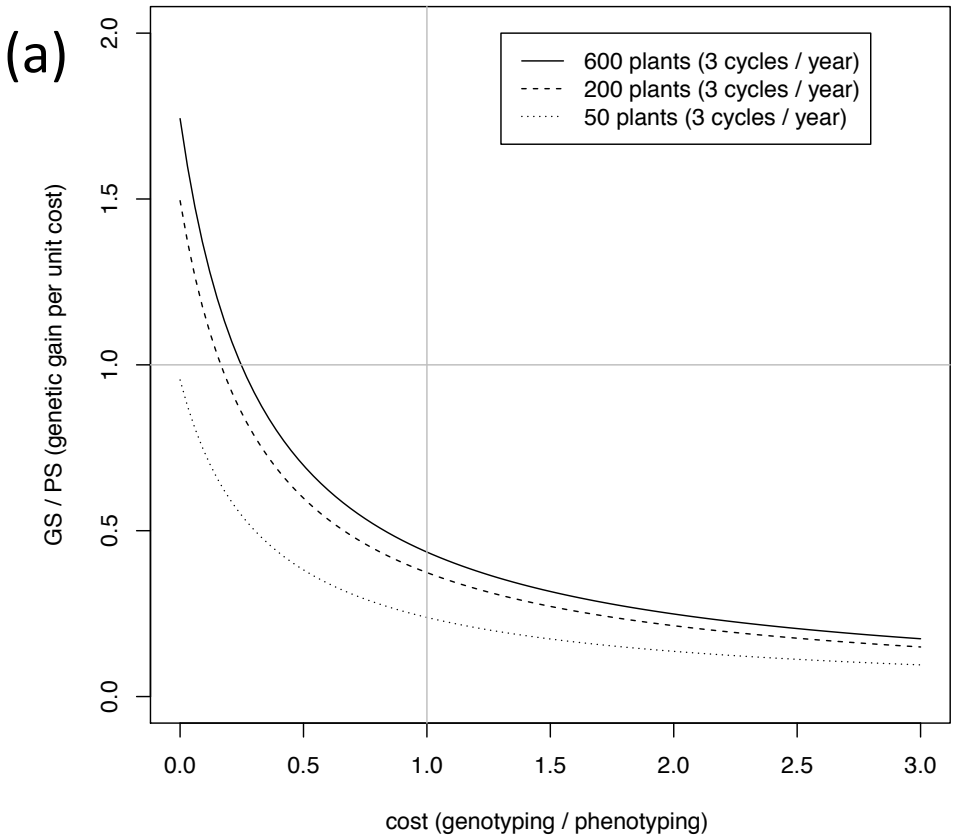
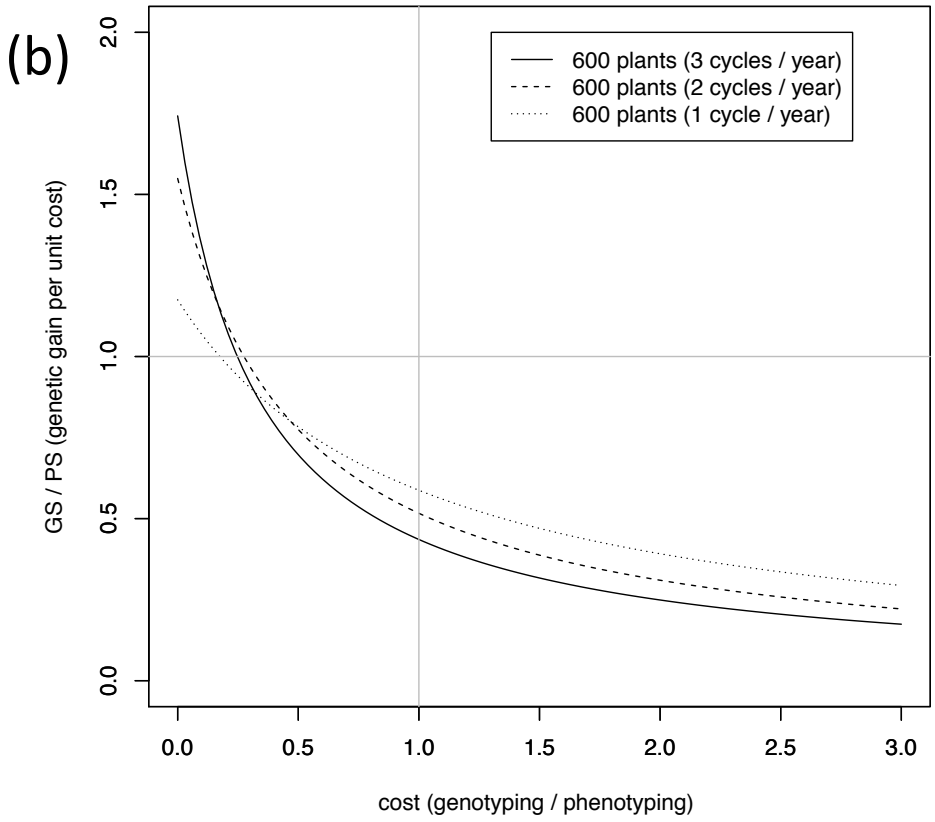


Figure 3.10. Genetic gain per unit cost at the end of the six-year breeding program for a comparison of the population size (a) and a comparison of the number of selection cycles per year (b), by using the identical population size among PS and GS. For (a) and (b), the number of cycles per year and the population sizes were fixed, respectively, at three and 600. For (c), the all kinds of scenarios in this study (i.e., three kinds of population sizes and three kinds of cycles per year) were represented. The horizontal axis shows how many times greater the cost for genotyping was than that for phenotyping. The vertical axis shows the ratio of genetic gain per unit cost of GS to that of PS with identical population size to GS. Vertical and horizontal gray lines in the figures respectively show the point at which the cost for genotyping equals the cost for phenotyping and that at which the genetic gain per unit cost of GS equals that of PS. The heritability of a target trait was 0.2 at the base population.



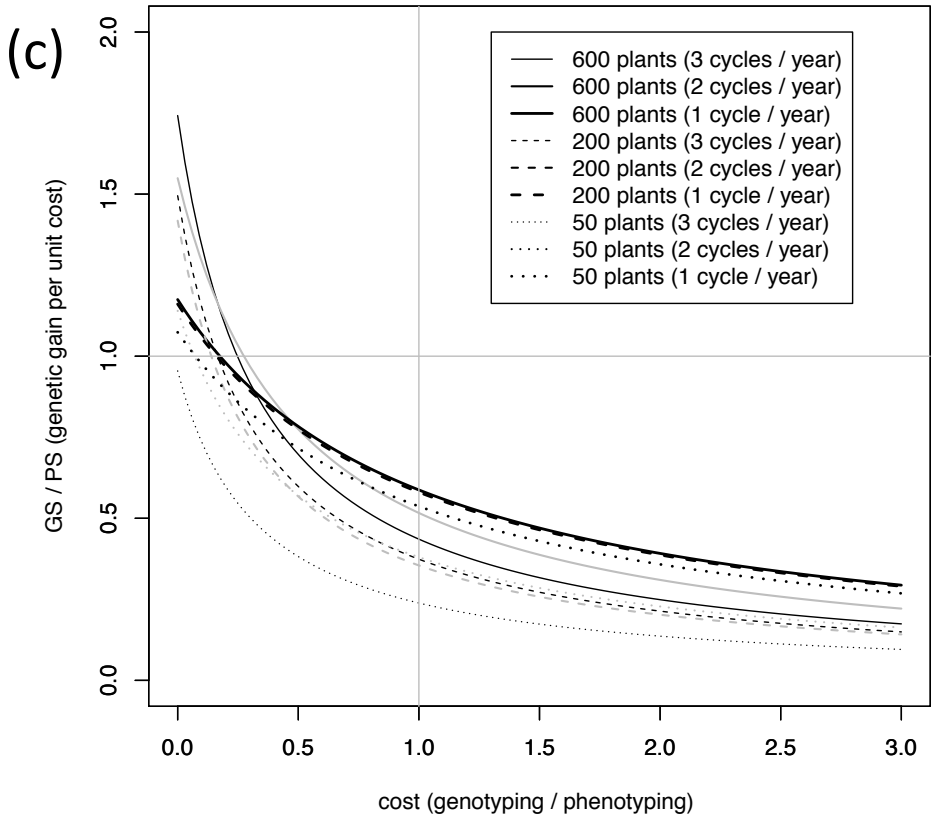


Figure 3.10. (Continued)

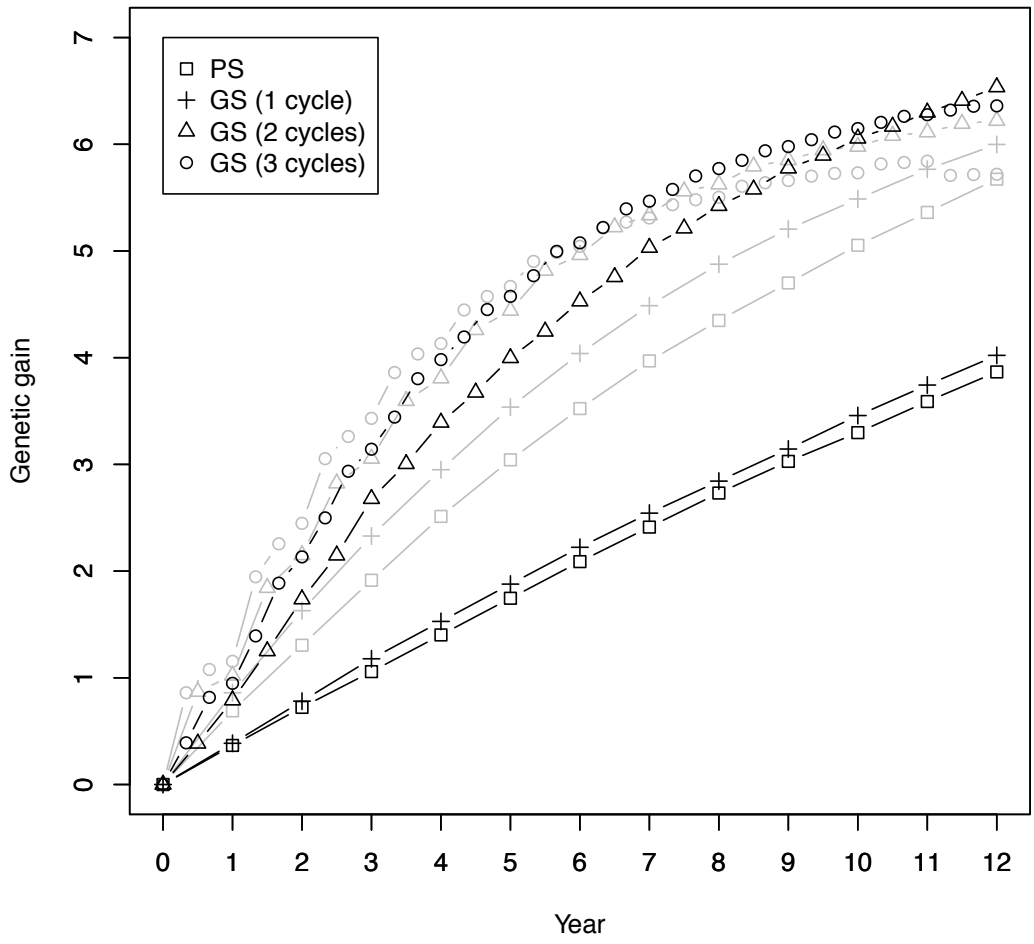


Figure 3.11. Mean genetic gains of plants in a breeding population at each selection cycle for various breeding procedures and traits. Black lines show the values of a trait expressed after pollination. Gray lines show the values of a trait expressed before pollination. Squares represent results of PS breeding. Crosses, triangles, and circles respectively represent the result of GS breeding with one, two, and three cycles per year. In both PS and GS, the breeding population size was 200. In GS, 500 co-dominant markers were used for genomic prediction. The heritability of a target trait was 0.2 at the base population.

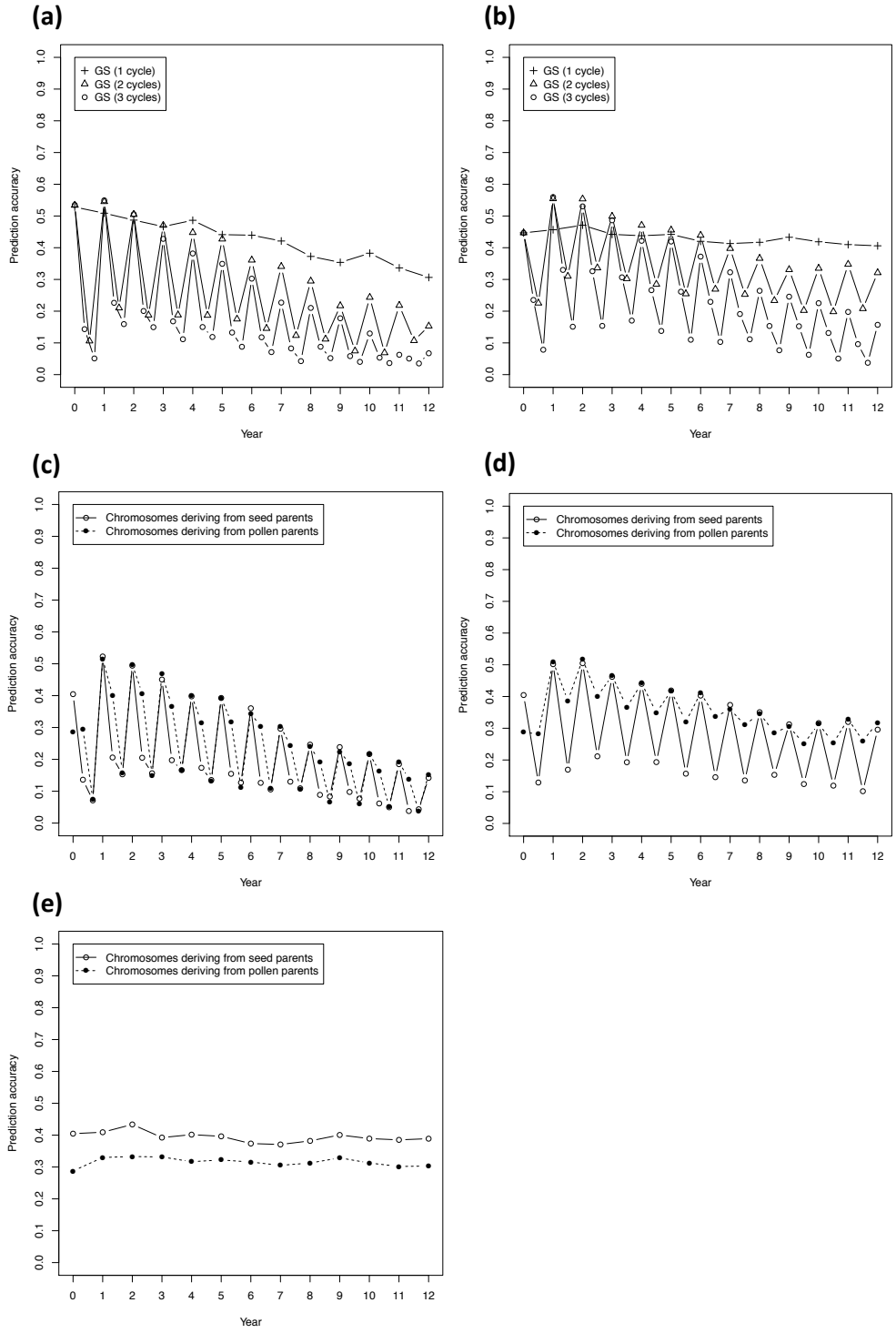


Figure 3.12. Prediction accuracy measured as a correlation coefficient between true genetic values and predicted values for a trait expressed before pollination (a) and for a trait expressed after pollination (b). Figures (c) − (e) show the prediction accuracy of chromosomes derived from seed and pollen parents, separately, in GS breeding of a trait expressed after pollination with three, two, and one cycles per year. Only seed parents were selected at GPS.

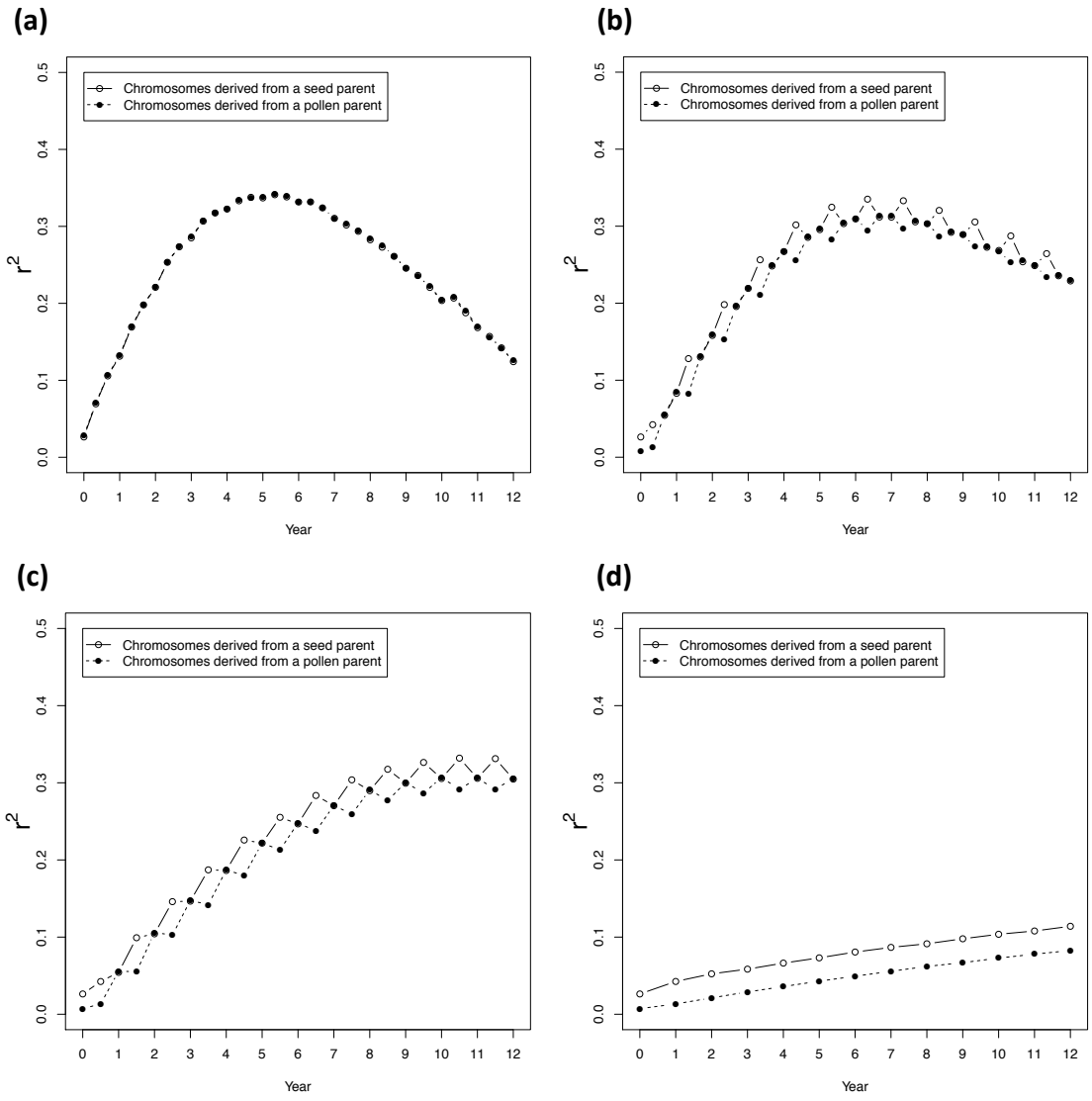
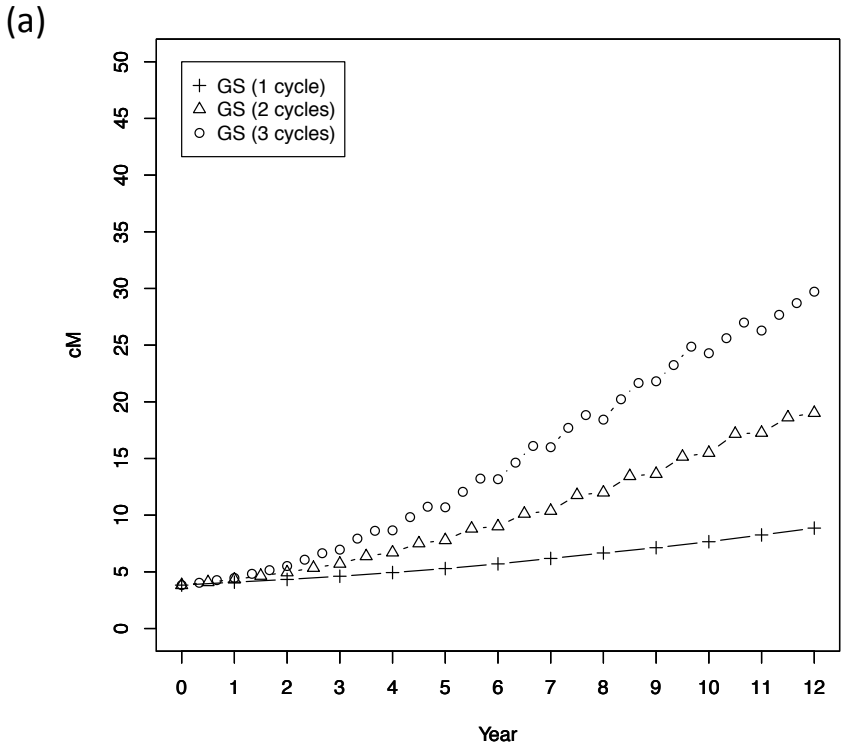


Figure 3.13. Figure (a) portrays the generational change of LD (r^2) between QTL and their adjacent polymorphic markers in GS breeding with three cycles per year in a trait expressing before pollination. Figures (b) − (d) show LD between QTL and their adjacent polymorphic markers of GS breeding of a trait expressed after pollination when three, two and one selection cycles were held per year. The LD values were calculated separately for chromosomes derived from seed and pollen parents.



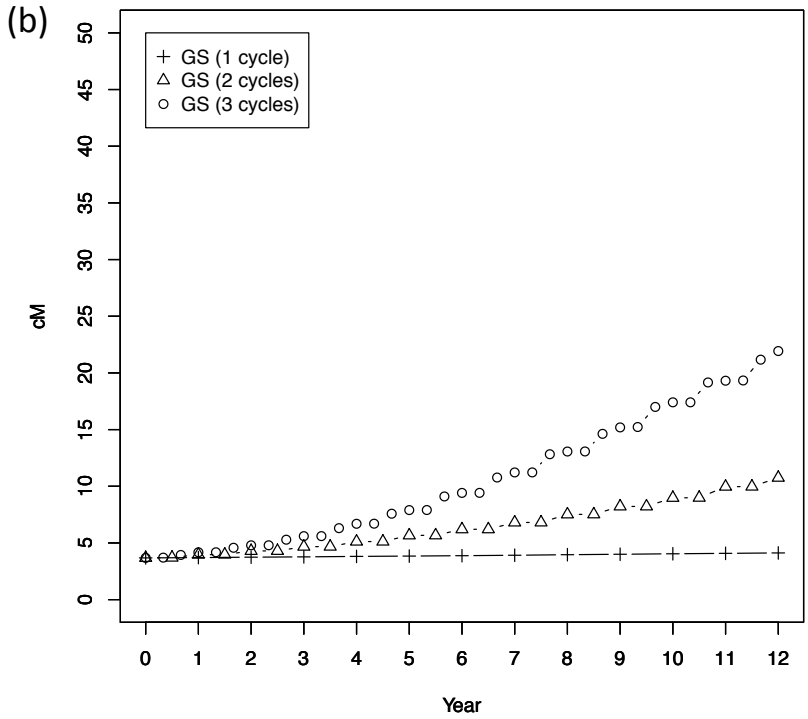


Figure 3.14. The average distance between QTL and the adjacent polymorphic markers for a trait expressed before pollination (a) and for a trait expressed after pollination (b). Crosses, triangles, and circles respectively represent the result of GS breeding with one, two, and three cycles per year.

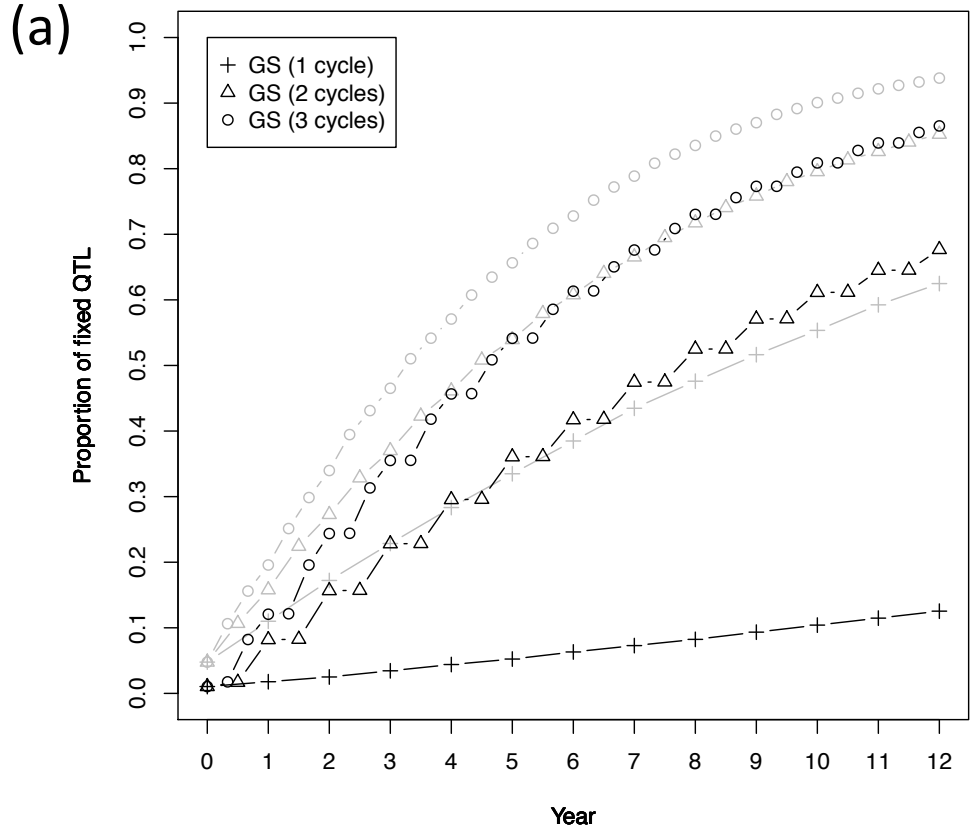
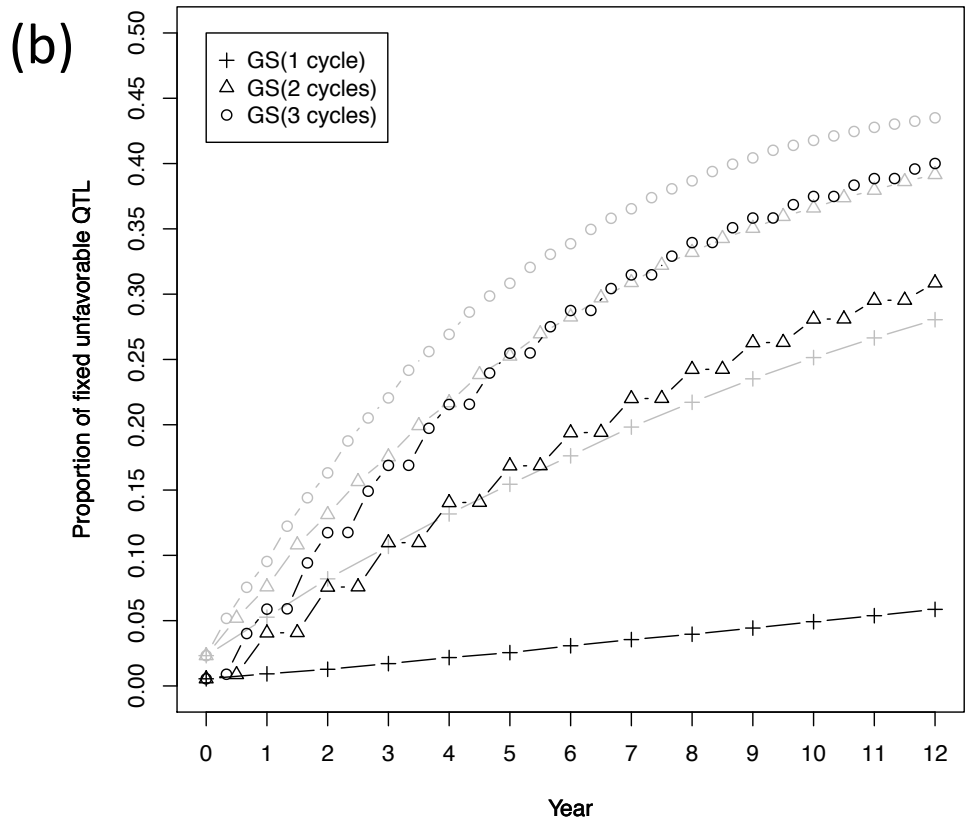


Figure 3.15. Figure (a) represents the proportions of QTL that were fixed in the breeding population when the target trait was expressed after pollination (black lines) and when the target trait was expressed before pollination (gray lines). Figure (b) represents the proportions of QTL that were fixed to the unfavorable allele in the breeding population when the target trait was expressed after pollination (black lines) and when the target trait was expressed before pollination (gray lines). Figure (c) represents the proportions of markers used for genomic prediction that were fixed in the breeding population when the target trait was expressed after pollination (black lines) and when the target trait was expressed before pollination (gray lines).



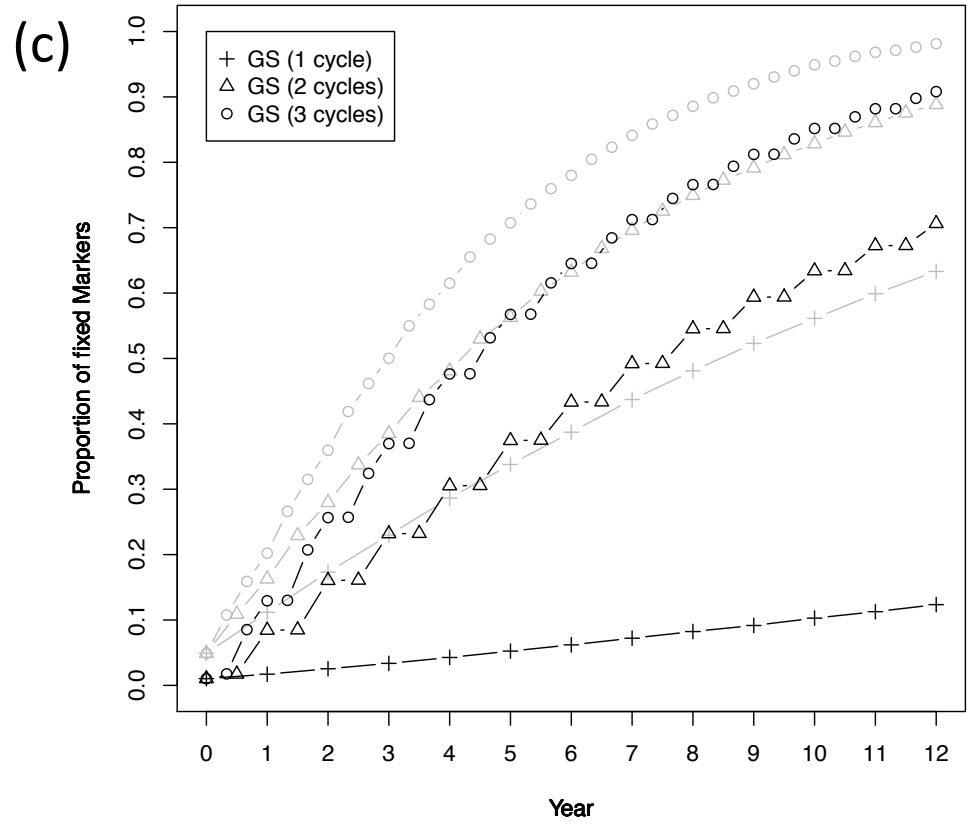


Figure 3.15. (Continued)

3-4. Discussion

3-4-1. Genomic selection in the context of mass selection

It was shown that the genetic gain attained by GS is greater than or equal to the gain by PS in all simulation settings assessed in this study (Figs. 3.3 – 3.7 and 3.11). GS was able to improve especially the short-term response of mass selection, suggesting that GS can compensate for the major shortcoming of mass selection (i.e., low efficiency of genetic improvement) and that it can improve the efficiency of mass selection in an annual allogamous crop.

Compared to GS, MAS was able to attain rapid genetic improvement in earlier generations (during one year of breeding), but it was unable to yield further improvement in later generations because of the fixation of all QTL targeted by MAS (Fig. 3.3). In this study, genetic variance explained by the five target QTL in MAS was assumed as around 40% in the initial population, which might limit the effectiveness of MAS. Even if I could fix the five QTL targeted in MAS, I could treat a part of variation in a breeding population in MAS breeding. For simulations, I assumed that I would know true genotypic effects for every plant and that I would be able to identify all QTL genotypes exactly with no errors. Actually, however, breeders should infer QTL genotypes from the genotypes of closely linked markers in MAS breeding, which makes the MAS accuracy lower than that estimated in the present simulations. Bernardo and Yu (2007) and Mayor and Bernardo (2009) showed that the efficiency of GS was higher than that of marker-assisted recurrent selection (MARS) in their simulation studies. Mayor and Bernardo (2009) also reported that the relative efficiency of GS to MARS became higher when the number of QTL was larger. The results obtained in the present study agree with their results. It is suggested that the accumulation of a number of small-effect or medium-effect QTL is necessary to improve quantitative traits, in particular, when the target trait are controlled by numerous QTL as is the case in this simulation study. In the present study, I found that GS based on a LASSO model was better than MAS (Figs. 3.3b and 3.6). Although LASSO is intended to explain the variation in phenotypic values with fewer markers than ridge regression, LASSO remained better than MAS. This result suggests that LASSO employed information from more numerous QTL in the prediction than MAS by using the genome-wide marker. GS breeding based on a LASSO model, however, was less efficient than a ridge regression model (Fig. 3.6). It is suggested that the accumulation of small or medium QTL effects, which are explainable by ridge regression but not by LASSO, was important to improve quantitative traits controlled by a number of QTL.

It is noteworthy that GS breeding in our simulation was based on a prediction model built

based on a single plant phenotype. Therefore, the heritability we assumed for this study (i.e., h^2 = 0.5 or 0.2) might be higher than the heritability expected under a single-plant basis evaluation. The relative efficiency of GS and PS when h^2 = 0.5 or 0.2 (Fig. 3.3) suggests that GS might attain higher gain than PS even when a heritability of a breeding population is much lower than the setting of this study. To improve the absolute potential of GS breeding, nevertheless, it is important to solve the problem caused by the low heritability of a single-plant-based evaluation. The progeny testing can be a solution of the problem, but it may cancel the benefits of mass selection (i.e., simplicity and selection that is applicable in each generation) and makes the model-updating process time-consuming.

3-4-2. Effective strategies for genomic selection

It was revealed that genetic gains were not significantly different between GS with 500 markers and with 5000 markers (Fig. 3.4). Under the present simulation settings, the number of markers necessary for efficient GS was a realistic level even in allogamous crops with low levels of LD. This result was consistent with those reported by Iwata et al. (2011), where they conducted simulations of GS breeding under low levels of LD in the forest tree conifer (*Cryptomeria japonica* (L. f.) D. Don). However, GS with lower marker density (i.e., with 100 markers in the present study) was not efficient, especially in later generations (Fig. 3.4), because many markers were going to fixation before the fixation of QTL located nearby. Iwata et al. (2011) and Grattapaglia et al. (2011) reported that the accuracy of GS improved when a higher density (e.g. larger number) of genetic markers was used in GS. Grattapaglia et al. (2011), however, showed that no increase in accuracy occurred when the marker density was rather high and that the accuracy of GS came to a plateau with lower marker density when the effective population size was smaller (i.e., LD was higher). The results suggest that the optimal marker density for GS can be decided according to the range of LD in a breeding population. It suggests that the optimal marker density was defined according to the range of LD in the breeding population, which indicates that preliminary genotyping of a target breeding population is necessary to determine the optimal marker density in practical breeding programs. In the present study, I assumed linkage equilibrium in the initial population. Consequently, LD increased only through breeding operations. This assumption widened the range of LD and decreased the number of markers necessary for GS. If historical LD (LD generated in past demographic history) exists in the initial population, then the LD might be narrower in the range than that assumed in the simulations. In such a case, the GS accuracy can be improved further using numerous markers.

Even when the mode of inheritance of markers is dominant, the efficiency of GS does not decrease greatly (Fig. 3.5). This suggests a possibility that many kinds of markers can provide high gains in GS, even when the markers are not suited to detect heterogeneous genotypes.

In the simulations, response to GS reached a plateau in later generations, especially when the number of selection cycles per year was large and the breeding population was small, whereas PS breeding displayed moderate but consistent long-term improvement (Fig. 3.7). GS breeding with a greater number of cycles per year attained higher genetic gain in earlier generations of breeding but reached the plateau earlier. GS with a larger number of cycles per year caused a rapid decrease in the genetic variance in earlier generations (Fig. 3.8b), mainly because of the additional GS cycles. Moreover, selection accuracy was low at GS cycles without updating a prediction model (Fig. 3.9b). Low accuracy of GS cycles without building a prediction model suggested by Jannink (2010) and Iwata et al. (2011) was also found in the present simulations. The rapid decrease of genetic diversity decreased the selection accuracy further in later generations. If the genetic diversity declined rapidly, high gain would be attained with high prediction accuracy. Consequently, the rapid decrease of genetic diversity and the lower accuracy in GS cycles without an updated prediction model would cause this plateau. As mentioned by Iwata et al. (2011) and Bernardo (2009), it is suggested in the present study that it is important to build (i.e., update) a prediction model periodically to adapt the model to current circumstances. In the present simulations, LD of the breeding population was quite low and increased only through breeding operations. Consequently, the periodical update of a prediction model is prerequisite for GS breeding to fit the model to LD increase. GS breeding with a smaller breeding population size reached the plateau earlier (Fig. 3.7). This result was also caused by a rapid decrease of genetic variance (Fig. 3.8a).

GS can be implemented in more cycles per unit time even when the prediction accuracy is low, and thereby attain higher genetic gain than PS (Heffner et al., 2010). Results obtained in the present study show agreement with results of the previous study. As suggested by Bernardo and Yu (2007) and Mayor and Bernardo (2009), the main advantage of selection using genetic markers is the gain per unit time rather than gain per cycle. Consequently, technologies for shortening generation time engender higher efficiency in GS breeding, as suggested by Grattapaglia et al. (2011) and Iwata et al. (2011).

3-4-3. Cost efficiency of genomic selection

The cost efficiency of GS breeding was lower than that of PS breeding, unless the genotyping

cost was lower than 27% of the phenotyping cost (Fig. 3.10). A GS breeding population had to be phenotyped at the first cycle of each year to update a prediction model. Therefore, the genotyping cost of GS breeding became purely extra cost when the cost of GS breeding was compared to that of PS breeding with the identical population size. Under this condition, GS breeding must be expected to attain high genetic gain to compensate for the extra genotyping cost. Even under such a disadvantageous condition, however, the cost efficiency of GS breeding could surpass that of PS breeding when the genotyping cost was lower than about one-fourth (i.e., 27%) of the phenotyping cost (Fig. 3.10). When the genotyping cost was low, GS breeding with the combination of larger population size and a larger number of cycles per year would be more cost-effective. The advantage of GS breeding with larger population size mainly came from the fact that GS breeding could attain higher accuracy than PS breeding at the model updating steps (Fig. 3.9), because a prediction model built by large population's information has a possibility to attain higher accuracy than PS (Daetwyler et al., 2008).

In this scenario, I assumed only the genotyping cost as the extra cost of GS breeding comparing with PS breeding. It means that GS breeding will become more cost-efficient when the genotyping cost becomes cheaper. In fact, the genotyping cost is one of the important factors that determine the cost efficiency. However, GS breeding requires additional effort and cost for plant breeders to extract DNA from plants, grow plants using offseason nursing. To manage the extra work, it is also required to employ additional labors. In this study, I eliminated this kind of costs for the sake of simplicity. In the actual breeding, we must consider these extra costs to calculate the cost efficiency of GS breeding.

3-4-4. Trait expressed before pollination and after pollination

The genetic gain of GS breeding was almost equal for a trait expressed before pollination and a trait expressed after pollination, except when the breeding cycle occurred once per year (Fig. 3.11). The genetic gain of PS breeding was much lower for a trait expressed after pollination than gain for a trait expressed before pollination (Fig. 3.11). This result of PS breeding gave agreements with the selection theory that gain by one cycle of selection with pollen control would be twice the gain by one cycle of selection without pollen control (Hallauer and Miranda, 1981). These results emphasize the high efficiency of GS breeding for a trait expressed after pollination via pollen control at GS, even though pollen parents cannot be selected at GPS. Results for traits expressed before pollination show that the prediction accuracy of GS after GPS was reduced severely (Figs. 3.9 and 3.12a) and that the low accuracy produced only slight

genetic improvement at GS (Figs. 3.3 – 3.7; gray lines in Fig. 3.11). In a trait expressed after pollination, however, high prediction accuracy at GS immediately after GPS (Fig. 3.12b) raised the efficiency of GS breeding even without pollen control at GPS (Fig. 3.11).

For traits expressed after pollination, the increased efficiency, which results from the improved prediction accuracy at GS immediately after GPS, is mainly attributable to the fact that the prediction accuracy of chromosomes derived from a pollen parent was high at GS immediately after GPS (Figs. 3.12c and 3.12d). The prediction accuracy of chromosomes derived from a pollen parent exhibited a small reduction at GS immediately after GPS, although that of chromosomes derived from a seed parent decreased greatly (Figs. 3.12c and 3.12d). For a trait expressed after pollination, selection occurs only among seed parents at GPS steps because of the requirements of phenotyping. In chromosomes derived from a seed parent, therefore, the LD pattern changed drastically from GPS to GS immediately after GPS because of a genetic bottleneck attributed to the selection among seed parents (Figs. 3.13b and 3.13c). In contrast, the LD pattern changed little in chromosomes derived from a pollen parent (Figs. 3.13b and 3.13c) because these chromosomes did not experience the selection bottleneck at GPS. This changing pattern of LD gave the genetic potential of chromosomes derived from a pollen parent as quite predictable using the model at GS immediately after GPS. The results suggest that GS after GPS is important to raise the efficiency of breeding in the selection of a trait expressed after pollination. Although GS immediately after GPS is important to make GS breeding effective for the selection of a trait expressed after pollination, the second round of GS had only a slight impact because of the low prediction accuracy in that round (Fig. 3.12b). Results of this simulation study suggests that one GS step after GPS might be better to avoid less efficient selection with less accurate prediction model because of the changing LD pattern for a trait expressed after pollination.

I anticipated the low selection efficiency because of the impossibility of pollen control at GPS step, which resulted in low efficiency of selection at GPS step actually. Because of the high accuracy of GS step, however, GS breeding for a trait expressed after pollination showed high efficiency for the whole selection cycles. Conventionally, mass selection with progeny test (i.e., half-sib selection) is used for breeding of allogamous crops (Ukai, 2003). This method achieves higher selection accuracy than ordinal mass selection, because it evaluates a half-sib population instead of a single plant. On another front of this merit, this method is based on maternal-line selection, which is the same situation as PS breeding of a trait expressed after pollination simulated in my simulation. In mass selection with progeny test, not only is excess

time for progeny test required, but maternal-line selection is also implemented at every cycle. GS breeding, in which pollen control and acceleration of breeding cycles are possible, would be more effective for breeding of allogamous crops, considering the gain per unit time.

In GS breeding of both traits expressed before and after pollination, prediction accuracies decreased over time (Figs. 3.12a and 3.12b). As shown in simulations assuming a trait expressed before pollination to compare many strategies in GS breeding (Figs. 3.7 – 3.9), the decline of genetic variance in the breeding population caused a decrease of a selection accuracy, resulting in the increase of the ratio of the environmental variance to the genetic variance ($var(e) / var(g)$) at the model-updating cycles, which reduced the accuracy of the prediction model. Moreover, the fixation of markers caused the decay of marker–QTL associations. The distance between QTL and adjacent polymorphic marker lengthened with selection cycles (Fig. 3.14). Owing to this situation, the decay in QTL–marker LD started at the later cycle of selection (Figs. 3.13a and 3.13b). Jannink (2010) reported that one cause of decreased accuracy of GS predictions is the decay of marker–QTL association caused by the fixation of the markers, which agrees well with the results in the present study. Although GS is useful especially for the genetic improvement of traits expressed after pollination, the appropriate number of selection cycles differs according to the LD pattern in the selected population. When a breeding population has low levels of LD, the prediction model becomes inaccurate in early generations because of a rapid change in the LD pattern (Fig. 3.13).

3-4-5. Long-term selection

PS breeding showed genetic improvement linearly through selection cycles (Figs. 3.3 and 3.11), suggesting that PS breeding (i.e., conventional mass selection) is better at long-term response than GS breeding. It revalidated that mass selection is attractive owing to its good long-term response. On another front, GS reached a plateau in later generations even though it could compensate for the limited short-term response of mass selection. The difficulty of long-term improvement using GS has been discussed previously (Goddard, 2009; Jannink, 2010). The present study also revealed that the advantage of GS over PS became smaller in later generations especially when the target trait is expressed before pollination (Figs. 3.3 – 3.7 and Fig. 3.11). Goddard (2009) introduced a new selection criterion by which minor alleles are assigned larger weights to prevent the loss of the alleles from a breeding population. This approach is expected to enable long-term improvement using GS. Jannink (2010) conducted simulations with Goddard's selection criterion improved to take account of allelic effects and

showed that the criterion improved the long-term genetic gain using GS. In the present study, I also used the criterion suggested by Jannink (2010) to improve the long-term genetic gain using GS in the situation similar to simulations comparing GS strategies for a trait expressed before pollination, but it did not improve the long-term gain significantly (data not shown).

In the present study, GS breeding without pollen control at GPS attained similar levels of genetic gain to those of GS breeding with pollen control at GPS. In fact, GS breeding without pollen control at GPS eventually produced a higher genetic gain than GS breeding with pollen control at GPS at the later stage of selection, except when the breeding cycle occurred once per year (Fig. 3.11). It was also apparent that GS breeding without pollen control at GPS reached a plateau at a later time than GS breeding with pollen control at GPS steps (Fig. 3.11). When I perform selection after pollination, all plants contribute as pollen parents at GPS. This situation prevented breeding population from experiencing a severe genetic bottleneck. GS breeding without pollen control at GPS can achieve high selection efficiency and can prevent loss of genetic diversity. Additionally, maintaining genetic diversity in a breeding population results in high accuracy in the breeding scheme I used (i.e., updating a prediction model every year by using breeding population data). The results of GS breeding for a trait expressed after pollination might provide clues to elucidate long-term GS.

3-4-6. Suggestion for actual breeding

In this study, I assumed a single target trait for each breeding scheme. Selection on a single trait, however, can be regarded as selection on multiple traits summed up with certain weights to a selection index, as is often used in practical breeding (e.g., VanRaden et al., 2009). For the present study, I simulated a single trait controlled by 300 QTL. Because the number of QTL in the simulations was large, it might be justified to consider the simulated trait as a multi-trait selection index. Actually, however, selection on multiple traits might be more complicated because of the tradeoffs among traits (e.g., negative correlations between yield and quality traits; Ivkovich and Koshy, 2002).

In the simulations, I only assumed additive QTL effects. Actually, however, non-additive QTL effects, i.e., dominance and epistatic effects, might contribute significantly to the phenotypic variation of quantitative traits. When the non-additive effects of QTL are large, the accuracy of GS might decrease greatly because the non-additive effects bias estimated marker effects in the GS prediction model, assuming only additive effects. Even when the contribution of dominance effects is large, however, it is not easy to fix the dominance effects in an

open-pollinating population such as most forage crops and buckwheat. Therefore, only additive effects can be used for the genetic improvement of the open-pollinated crop population.

Results of this study suggests that mass selection with GS has potential to yield benefits in practical use that are greater than those of traditional mass selection with PS or mass selection with MAS in annual allogamous crops, even when the degree of LD is low in a breeding population. The necessary number of markers for mass selection breeding with GS is practical even in current situations. Elshire et al. (2011) described that the cost of genotyping one sample would become less than $20 using their genotyping-by-sequencing (GBS) technology and that it would be reduced further to $5 or less in the near future. When the genotyping cost becomes lower, the genetic gain per unit cost became greater in GS breeding than in PS breeding. Even when the genetic gain per unit cost of GS breeding is lower than that of PS breeding, however, GS breeding might be more advantageous than PS breeding because the reduction of time required for developing one variety usually has a large economic impact (Brennan, 1989; Morris et al., 2003; Pandey and Rajatasereekul, 1999). Although GS breeding might reach a plateau in long-term selection, GS breeding is superior to PS for short-term and medium-term selection. Moreover, the scale of GS (the number of markers and the breeding population size, etc.) can be chosen according to available budgets and resources in the breeding program. Consequently, GS can be an efficient and practical breeding method for allogamous crop breeding even when LD is low in an initial population. These points encourage us to use GS in the mass selection of annual allogamous crop breeding.

Iwata et al. (2011) and Bernardo (2009) described the importance of building (i.e., updating) a prediction model periodically to adapt the model to current circumstances. Iwata et al. (2011) described the sudden decrease of selection accuracy in GS breeding without updating a prediction model, although Bernardo (2009) reported that a prediction model of 7 or 8 cycles could be used. The difference might derive from the difference of LD patterns in the breeding population they simulated. Iwata et al. (2011) assumed extremely low levels of LD in their simulation of conifer. However, the maize population used by Bernardo (2009) had a high and wide range of LD. These results suggest the importance of updating the prediction model in response to the changing LD pattern. They also suggest that the optimal frequency of the updating depends strongly on the degree of LD in a breeding population. Updating a prediction model costs much money in the current situation, which makes it difficult to update a prediction model frequently. However, update of a model is essential to attain enough level of genetic gain through GS breeding. Even when the cost of genotyping is high, update of a prediction model

should be conducted at the suitable timing according to LD of a breeding population. In the near future, the time required for genotyping would become more of problem than the cost of genotyping. While the cost of genotyping is decreasing because of the wide spread use of high-throughput genotyping system, it is still difficult to complete the genotyping of all selection candidates in a breeding population in a short time. This problem has arisen in Chapter 6 of this dissertation in fact.

The results of GS simulations without pollen control at GPS are highly suggestive, even in the genetic improvement of traits expressed before pollination. Results suggest that it was not necessary to select pollen parents at GPS. Even for the improvement of traits expressed before pollination, it is possible to conserve effort and to achieve the same levels of genetic gain by implementing GS breeding without pollen control at GPS. Even with recent progress in genotyping technologies, it might not be easy to determine the genome-wide SNP genotypes of numerous plants within a vegetative growth stage at GPS especially in crops with a short lifetime (e.g. a few months). In such crops, it might be important to save effort in genotyping plants before pollination to select pollen parents at GPS. Not only for genotyping effort, but also for labor for control cross, it might be beneficial to perform selection after pollination at GPS. The strategy also has effect to conserve genetic diversity in a breeding population. Moreover, traits expressed in and after the reproductive phase, such as characteristics of flowers and seeds, might be important for producing next-generation seeds. For example, in forage crops, main breeding-target traits are herbage production, quality, and resistance to biotic and abiotic stresses, but a regular supply of commercial quantities of seeds is also an important trait that is necessary to develop a new variety (Wilkins, 1991; Walter et al., 2012).

As is often the case in forage crops, the selection of target traits, such as yield and persistence, usually requires population-based evaluation. For the traits evaluated as a population, it is often difficult to improve the genetic potential of a population through selection based on single-plant evaluation (Connolly, 2001). Hayes et al. (2013) proposed a breeding scheme of GS in breeding of forage plant species to solve this problem. Further studies must be conducted to improve the potential of GS breeding in traits evaluated as a population.

Chapter 4

Simulation evaluation of island-model genomic selection in an autogamous plant

4-1. Introduction

Cereals account for a large proportion of food supply for human (Tweeten and Thompson, 2008). Maize, rice and wheat dominate the largest amount of production among the cereals (FAOSTAT, 2014), suggesting that an increase in yield of these cereals apparently leads to a stable supply of food in the world. The progress in the production of total cereals is stable in these 20 years. To meet the huge increase in human population, more rapid improvement is required than the ongoing progress rate.

In this Chapter, I evaluated the efficiency of genomic selection in rice. Rice occupies the second largest production among cereals in the world (FAOSTAT, 2014). In autogamous cereal plants such as rice, wheat and barley, population breeding and pedigree method are commonly used in breeding programs. In population breeding, cultivation is conducted in bulk with repeated self-pollination until F_5. After that, selection based on a single plant evaluation is conducted at F_6 generation, and then, selection based on a single line is conducted after F_6 generation. In pedigree method, selection based on single-plant evaluation is conducted at F_2 generation, and then, a number of single plants are selected from better lines that are selected based on single-line evaluation after F_2 generation. The selection is performed based on a single plant, preliminary yield trials or comparative yield trials (Brown and Caligari, 2008; Ukai, 2003). These methods can raise the selection accuracy by using the averaged value of plants that belong to a single line, while large environmental effects make it difficult to evaluate the genetic ability accurately only according to phenotype of a single plant. In addition, for selection of traits that cannot be evaluated with a single plant, such as yield per unit area, the measurement of a group of plants that belong to a single line is required. In these methods,

because we repeat selfing of each line, the size of linkage blocks tends to maintain larger than one expected in a random mating, which results in the low ability to generate new combinations of genes in a breeding population. The expected levels of breakup of initial linkage blocks through infinite times of self-pollination are the same level of two to three cycles of random mating (Hanson, 1959). Recurrent selection, in which selection and crossing of selected individuals are performed repeatedly, has been proposed to solve this issue (e.g., Fujimaki 1979), but it has also some extra issues, one of which is low selection accuracy in single plant evaluation, in its realization. The problem of recurrent selection may be solved effectively by using GS that enables breeders to evaluate plants according to their marker genotypes at the single plant basis. Rutkoski et al. (2011) mentioned about the efficiency of recurrent selection using GS in stem rust resistance in wheat. They emphasized that the increase of recombination events facilitates combining favorable alleles, and that it would make greater gain than conventional bulk breeding methods. GS based on recurrent selection should be efficient for autogamous plants breeding. Moreover, GS enables us to skip phenotyping at each selection cycle and thus to implement rapid-cycle genetic improvement by accelerating generation advancement. In this study, I evaluated the efficiency of recurrent selection using GS in autogamous plant breeding with a simulation study. GS may break through the restrictions of conventional PS, and allow us to realize more flexible styles of plant breeding than conventional breeding.

In the present study, I propose “island-model GS” as a new breeding strategy for autogamous plants. The term “island model” is originally derived from the field of population genetics, meaning that a large population is split into multiple subpopulations and that each subpopulation receives migrants from the others. In a natural population, individuals undergo selection to adapt themselves to their local environment using various loci and alleles. Wright (1932) considered the case of a widely distributed species that is subdivided into many small local races. While selection is taken place within each race, crossbreeding occasionally happens between races. The selection strategy derived from this kind of consideration is called island model. Migration among subpopulations tends to counteract the dispersion of allele frequency (Hartl and Clark, 2007).

The concept of island model described above has inspired global optimization problems in the computational science research field. In the optimization problem, an objective function (i.e., function to be optimized) sometimes has a large number of parameters. When the objective function is non-linear and non-differentiable, it is difficult to search global optimum because of

the existence of a number of local optima. Evolutionary algorithms (EAs) are frequently used to solve the issue. EAs are heuristic algorithms that use a system of natural selection in their system to find the optimal solution. In EAs, individuals are constructed by the parameters in question, which are assumed as genes of living organisms, and selection and crossing are repeated like selection in a natural population. The concept of island model is utilized in the field of EAs. That is, in the island model of EA, individuals are split into subpopulations, and selection and crossing are repeated between subpopulations. Whitley et al. (1998) reported that the island model showed better search performance than a single population model in some cases. One reason of this efficiency is that various islands (i.e., subpopulations) maintain some degree of independence and thus explore different regions of the parameter space. The success of the island model of EA suggests that the concept of island model can be effective not only in natural selection but also in artificial selection.

The concept of island model may also be efficient in plant breeding. As mentioned above, recurrent selection might be efficient for genetic improvement of autogamous plants. For selection strategies, it is important not only to select candidates harboring good genetic ability accurately, but also to maintain genetic variation in a breeding population (Chapter 3 in this dissertation). Especially for an autogamous plant population, which has large linkage blocks, it is difficult to maintain genetic variation in a breeding population with selection even if recurrent selection is conducted. Moreover, when among-cultivar diversity is large as is often the case in autogamous crops, genetic difference among segregating families becomes large. When we consider segregating families as islands, the islands maintain some degree of independence and can be considered to explore different regions of the parameter space for "optimizing QTL genotypes". The concept of island model might work effectively in this kind of breeding populations.

However, it has some risks to consider that the algorithms in EAs are also effective in plant breeding just because of the resemblances between EAs and plant breeding. There are some differences between EAs and plant breeding. First, mutations occur frequently in EAs while they were quite limited in a breeding population because of the time scale of a breeding program. Second, EAs generally simulates a large population, while the size of a breeding population is limited in plant breeding. Third, EAs can gain high ability through a number of selection cycles in short time by using computers, while plant breeding requires long time to evaluate and select plants. Although O'Hagan et al. (2012) conducted breeding simulations by using some concepts in EAs, they assumed large population size and high mutation rate

expecting radiation dose. It was not realistic for implementing their algorithms directly in a real plant breeding program. To examine the more empirical potential of the algorithms in EAs, it is required to take into account the restrictions in a real plant breeding program and evaluate the efficiency of the algorithms under the restriction.

In this study, I conducted breeding simulations with a real marker genotype data of cultivars in Asian cultivated rice, *Oryza sativa* L. To take advantage of existing materials and their information, I assumed to use recombinant inbred lines (RILs) derived from crosses between the existing cultivars as a training population and as an initial breeding population. I focused on the following three points in the simulations. The first is the efficiency of recurrent selection in an autogamous plant. To evaluate the efficiency, I compared genetic values attained by recurrent GS with those attained by RILs that were used as a breeding population. Second is the suitable constituent of an initial breeding population. I compared genetic gains from recurrent GS between initial populations derived from a single bi-parental cross and derived from multiple bi-parental crosses. Third is the efficiency of the island-model GS, which is proposed for plant breeding in this study. I evaluated genetic gains for "island-model GS", in which the initial breeding population were divided into subpopulations, and GS starting from a single population that was composed of the same genotypes as the initial population of the island-model GS.

4-2. Materials and methods

4-2-1. Marker data and position estimation

The dataset consisting of marker genotype of 3,102 loci in 112 rice cultivars (Table 4.1) was used to implement breeding simulation. The 112 cultivars represented a geographical and historical diversity of rice cultivars developed mainly in Japan. The marker genotype data was offered by Dr. Yamasaki (Food Resources Education and Research Center, Graduate School of Agricultural Science, Kobe University) and Dr. Ebana (National Institute of Agrobiological Sciences). Among 3,102 markers, 3,071 were single nucleotide polymorphism (SNP) markers developed from the sequence of Japanese cultivars (Nagasaki et al., 2010; Yamamoto et al., 2010), 31 were simple sequence repeat (SSR) markers (Yamasaki and Ideta, 2013). The physical map positions were detected in rice genome build 4. The marker positions on the linkage map were necessary to simulate recombination process that occurred in meiosis. The estimation of genetic map position was done by a polynomial regression of the genetic position on the physical position by using the information of rice genetic linkage map among F_2 plants derived from a single cross between the *japonica* variety Nipponbare and the *indica* variety Kasalath (Kurata et al., 1994; Harushima et al., 1998 and its updated information (Cheng et al., 2001). Imputation of missing marker genotypes was held using fastPHASE version 1.3 (Scheet and Stephens, 2006). Alleles at each imputed locus were imputed alternatively as one of two homozygous genotypes according to the proportion among 100 times replications of imputations.

4-2-2. Simulation settings

The 100 markers out of 3,102 markers were assumed as QTL controlling a target trait in each simulation trial. The proportion of phenotypic variance explained by each QTL (i.e., the heritability of each QTL) was set to follow the equation proposed by Lande and Thompson (1990) in the population constructing of the 112 cultivars. The effective number of QTL was set as 40. The sum of heritability of all QTL was set as 0.6. The genetic variation was explained only in additive way (i.e., no dominance and no epistatic effects affected the trait). Genotypic values were simulated using these simulated QTL effects. Phenotypic values were calculated by adding simulated environmental deviations into those genotypic values. Phenotypic variance of population composed of the 112 cultivars was standardized to be 1.0. 100 replications of simulation were implemented for each breeding procedure.

4-2-3. Breeding schemes

First, seven varieties, Koshihikari, Yumeakari, Hitomebore, Hatsushimo, Hinohikari, Nanatsuboshi, and Asahinoyume, were selected from 112 rice cultivars. They represent the genetic diversity in the 112 cultivars well. Second, six F_1 lines derived from six bi-parental crosses were made. Koshihikari, which was a predominant variety in Japan, was used as a common parent for the six bi-parental crosses. Starting from the six F_1 lines, six F_6 populations were simulated with the repeated selfing and the single seed decent (SSD) procedure. These simulated populations were used as initial populations for GS breeding and as training population for building a GS prediction model. Each F_6 population was constructed of 180 lines (i.e., 1,080 F_6 lines in total).

In all GS breeding schemes, 20 cycles of GS were conducted. A prediction model was built from phenotypic values and marker genotypes of the initial populations (i.e., six F_6 populations constructed 1,080 lines) and used throughout the 20 cycles. Selection intensity was set as 10%. Instead of random mating, I employed a single-round robin (Verhoeven et al., 2006), in which crosses were conducted as a chain, i.e., plant1 × plant2, plant2 × plant3, …, plant *S* × plant1 among *S* plants, as a rule for crossing selected plants for the next generation because of the difficulty of random mating in rice (i.e., autogamous species) population. This GS scheme used G-BLUP by R package "rrBLUP" (Endelman, 2011) for genomic prediction. This statistical method was introduced in the section 2-2-5 in this dissertation. In this simulation study, *X* was a vector of 1's as the intercept of the model with the length of the number of observation, and *Z* was a matrix that an identity matrix (number of observations × number of observations) was combined with 0's matrix (number of observation × number of prediction) by columns. Each marker genotype is defined as 1, 0, or -1 when the number of the considered allele contained is two, one, or zero, respectively.

First, to evaluate the efficiency of recurrent selection, I compared the outcomes of GS breeding with those of RILs by using the same breeding population derived from a single bi-parental cross. Second, to evaluate the impact of genetic architecture of a breeding population in GS breeding, two types of breeding populations were compared: (a) six breeding populations each of which was derived from a single bi-parental cross (discrete GS; Fig. 4.1a), and (b) a breeding population derived from multiple bi-parental crosses (bulked GS; Fig. 4.1b). For the former, one breeding population was constructed of 180 plants derived from one bi-parental cross. For the latter, six F_6 populations were gathered to construct one breeding population, in which 30 lines were came from each population. In the discrete GS, six breeding populations

with 180 genotypes, which were derived from a single bi-parental cross, experienced GS. The one population that showed the best outcome was selected as the result (Fig. 4.1a). In the bulked GS, only one breeding population with 180 genotypes, which were derived from multiple bi-parental crosses, was created and experienced GS (Fig. 4.1b). Third, to evaluate the efficiency of the island-model GS breeding, I compared two types of GS designs: (1) the bulked GS and (2) the island-model GS. In the bulked GS, GS was performed on a single breeding population derived from multiple bi-parental crosses (Fig. 4.1b). In the island-model GS (Fig. 4.1c), breeding was conducted based on six equal-sized subpopulations that connected to each other with a small amount of migration. The initial state of each subpopulation was F_6 population derived from a single bi-parental cross. To make genetic migration between subpopulations, one of selected plants was exchanged between subpopulations every cycle, and mating was held after exchange among each subpopulation. All simulations were done using R (R Development Core Team, 2014).

4-2-4. Summarization of results

From RILs composing the initial breeding population of GS breeding, the best line (i.e., a line with the highest genotypic value) was selected. The genotypic value of the best line was used as a standard for comparing the efficiency of recurrent selection with that of breeding utilizing inbred lines.

In GS breeding, an attained genotypic value was represented as the maximum of the true genotypic values among selected plants (i.e., upper 10% of plants selected based on predicted values) at each selection cycle. Here, I assumed a breeder could detect the best plant from the selected plants through field trials prior to variety release. The average breeding values of plants belonging to a single population and the distribution of breeding values are examined just to compare attained genotypic values (i.e., the maximum of the true genotypic values among selected plants) with the population mean.

The accuracy of genomic prediction was measured by Pearson's correlation coefficient between predicted values and true genotypic values.

The proportion of fixed loci (QTL + markers) was calculated as [number of fixed loci] / [total number of loci]. Figures were shown as the averaged value of 100 simulations in each breeding procedure.

Table 4.1. Rice cultivars used in the simulation

Name	Name (Japanese)
kirara397	きらら397
hoshinoyume	ほしのゆめ
yukihikari	ゆきひかり
hayamasari	はやまさり
hatsushizuku	初雫
yu-kara	ユーカラ
tsugaruroman	つがるロマン
yumeakari	ゆめあかり
mutsuhomare	むつほまれ
hukei175	ふ系175号
akihikari	アキヒカリ
reimei	レイメイ
tousaka5	藤坂5号
toyonishiki	トヨニシキ
ouu197	奥羽197号
chiyohonami	チヨホナミ
hitomebore	ひとめぼれ
manamusume	まなむすめ
akitakomachi	あきたこまち
okiniiri	おきにいり
hukuhibiki	ふくひびき
himenomochi	ヒメノモチ
kinuhikari	キヌヒカリ
dontokoi	どんとこい
itadaki	いただき
nourin1	農林1号
hounenwase	ホウネンワセ
todorokiwase	トドロキワセ
koshihikari	コシヒカリ
hanaechizan	ハナエチゼン
yukinosei	ゆきの精
gohyakumanngoku	五百万石
koganemochi	こがねもち
koshijiwase	越路早生
goropikari	ゴロピカリ
husaotome	ふさおとめ
hatsushimo	ハツシモ
aichinokaori	あいちのかおり
matsurihare	祭り晴

Name	Name (Japanese)
asanohikari	朝の光
tsukinohikari	月の光
daichinokaze	大地の風
mineasahi	ミネアサヒ
nourin29	農林29号
akebono	アケボノ
mirenishiki	ミレニシキ
kantou209	関東209号
otomemochi	オトメモチ
yamabiko	ヤマビコ
nakateshinsenbon	中生新千本
kinmaze	金南風
yamadanishiki	山田錦
nourin22	農林22号
asahi	朝日
reihou	レイホウ
hiyokumochi	ヒヨクモチ
houyoku	ホウヨク
jikkoku	十石
nourin18	農林18号
hinohikari	ヒノヒカリ
nishihomare	ニシホマレ
koganemasari	コガネマサリ
yumetsukushi	夢つくし
aikoku	愛国
asahi	旭
ooba	大場
kameji	亀治
kamenoo	亀の尾
jinriki	神力
takenari	竹成
hutaba	双葉
nipponbare	日本晴
nihonmasari	ニホンマサリ
kihou	喜峰
koganebare	黄金晴
yukimaru	ゆきまる
nanatsuboshi	ななつぼし
taichuu65	台中65号

Name	Name (Japanese)
EIKOU	栄光
akage	赤毛
hujihikari	フジヒカリ
benisengoku	ベニセンゴク
yumehitachi	ゆめひたち
asahinoyume	あさひの夢
akaneiro	あかね空
hatsuhoshi	初星
chiyonishiki	チヨニシキ
menkoina	めんこいな
notohikari	能登ひかり
hohohonoho	ほほほの穂
yumehikari	ユメヒカリ
natsuhikari	ナツヒカリ
haenuki	はえぬき
domannaka	どまんなか
haruru	晴るる
nourin6	農林6号
nourin8	農林8号
moritawase	森田早生
RIKUU132	陸羽132号
joushuu	上州
kiichi	撰一
ginbouzu	銀坊主
RIKUU20	陸羽20号
kamenoo4	亀の尾4号
omachi	雄町
sekitori	関取
bouzu	坊主
shirosenbonn	白千本
hatsunishiki	ハツニシキ
yamasenishiki	ヤマセニシキ
sasashigure	ササシグレ
sasanishiki	ササニシキ

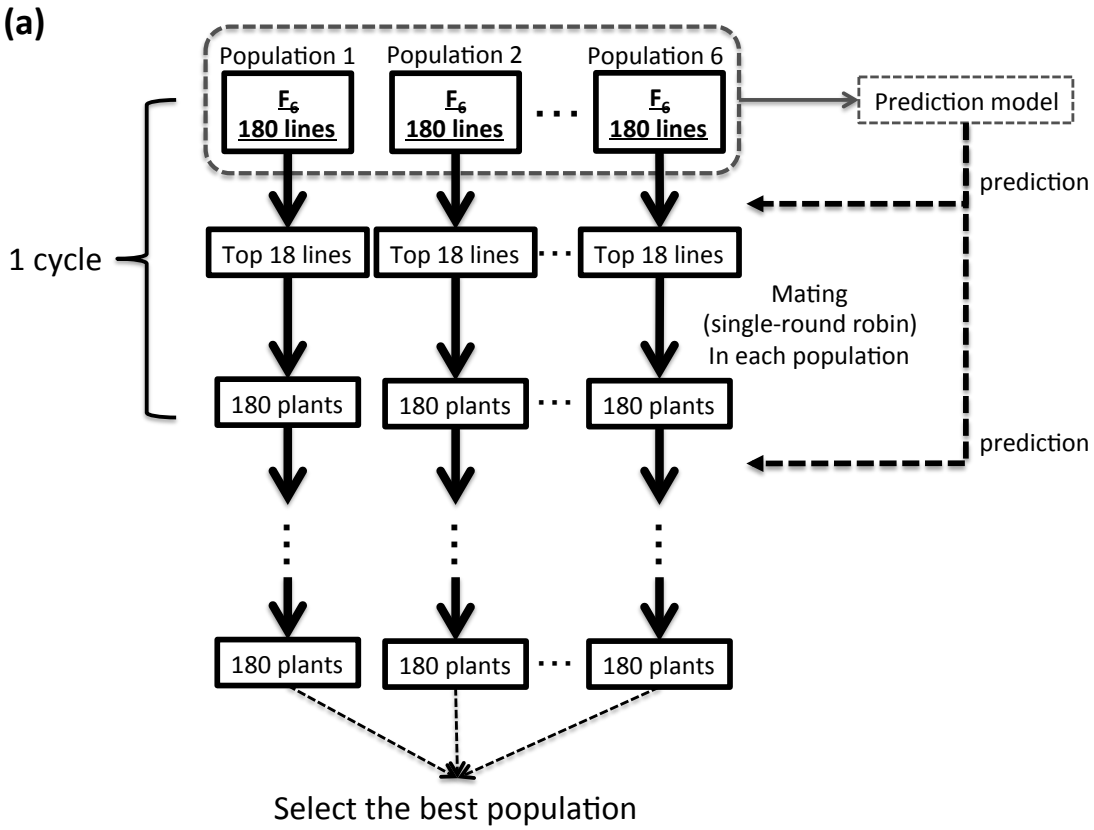


Figure 4.1. The GS breeding scheme: (a) discrete GS, which improves six breeding populations each of which is derived from a single bi-parental cross, (b) bulked GS, which improves a breeding population derived from multiple bi-parental crosses, and (c) island-model GS, which improves six breeding populations connected each other. Each of six populations is derived from multiple bi-parental crosses.

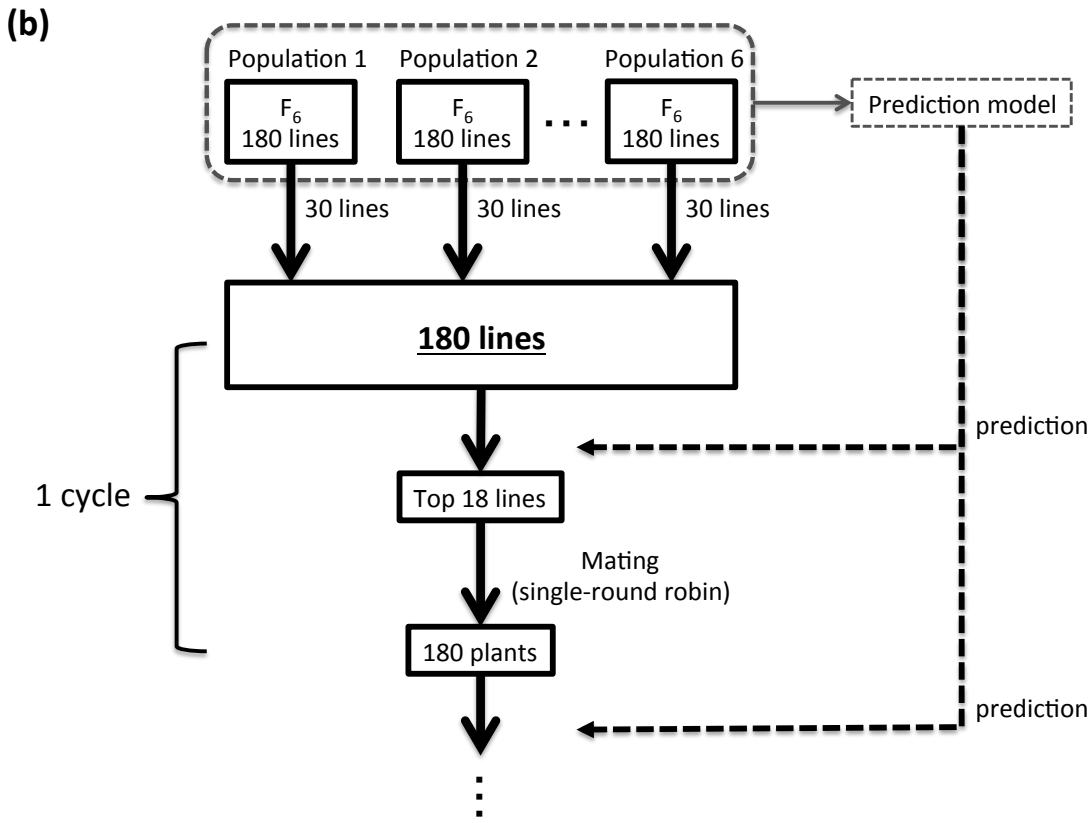


Figure 4.1. (Continued)

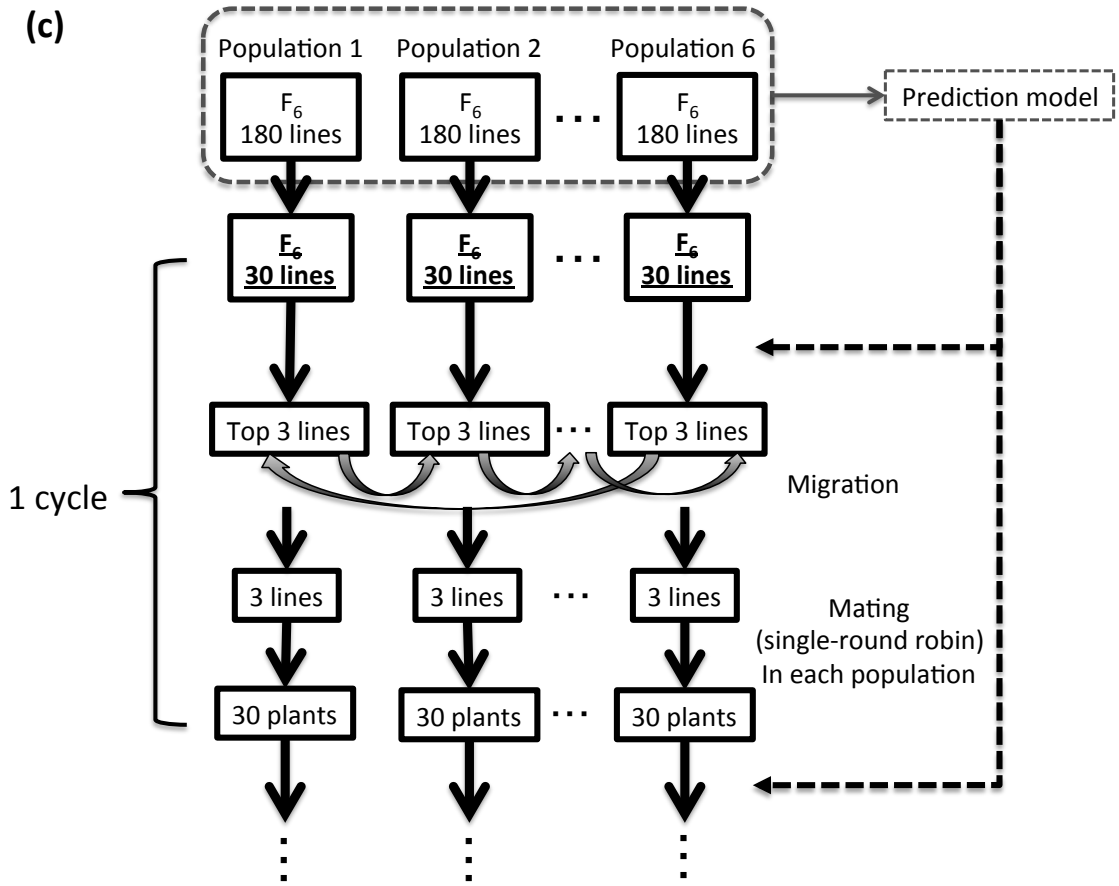


Figure 4.1. (Continued)

4-3. Results

4-3-1. Cultivars in the simulation

Figure 4.2 shows the result of cluster analysis of the 112 rice cultivars by 3,102 markers. The distance matrix was calculated as the Euclidean distance, in which two SNP genotypes are treated as 0 or 1. Because the cultivars are all homozygous, they have homozygous genotype for all SNPs. The cluster was made by Ward's method. The identical cultivar for bi-parental cross, Koshihikari, was in the bottom left of the figure. The remaining cultivars used in my breeding simulation were selected to cover the genetic diversity of the 112 cultivars well.

4-3-2. GS breeding designs

Figure 4.3 (a) shows the attained genetic value in the discrete GS. The dashed line shows the maximum value of genotypic values in F_6 population derived from a single cross. Black lines represent the result of Koshihikari × Hatsushimo population, which attained the highest genotypic value among the six breeding populations on average (52 out of 100 replications of simulation in GS breeding). Gray lines represent the mean value of the best population in each simulation. The discrete GS attained higher genotypic value than the maximum of F_6 population. For the result of Koshihikari × Hatsushimo population, the discrete GS exceed the maximum of F_6 in the 85 out of 100 trials after two cycles of selection, then attained higher genotypic value than the maximum of F_6 in all 100 trials after six cycles of selection. It suggests the superiority of GS breeding to the conventional methods using inbred lines. The discrete GS reached plateau after 7 – 8 cycles of selection.

Figure 4.3 (b) shows the genotypic values through selection cycles of the discrete GS and the bulked GS. The bulked GS, in which the breeding population derived from multiple bi-parental crosses, attained higher genotypic values than the discrete GS using the population derived from a single bi-parental cross. The bulked GS showed rapid genetic improvement during the early cycles and reached plateau after 12 – 13 cycles, while the discrete GS reached the plateau after 8 – 9 cycles of selection.

I implemented the simulation of the island model GS only for the population derived from multiple bi-parental crosses because the population derived from multiple bi-parental crosses might have higher potential than the population derived from a single bi-parental cross (Fig. 4.3b). Figure 4.3 (c) showed the genotypic value of the island-model GS using six subpopulations. The island-model GS attained lower genotypic values than the bulked GS until seventh cycle of selection for the averaged value over 100 trials (Fig. 4.3c). At the end of sixth

cycle of selection, however, the island-model GS exceed the bulked GS in 57 out of 100 simulation replications. In the later selection cycles, the island-model GS attained higher genotypic values than the bulked GS. The island-model GS exceed the bulked GS in 83 out of 100 simulation replications after 12 cycles of selection. The island-model GS did not seem to reach a plateau in my GS cycles (i.e., 20 cycles). Through the 20 cycles of selection, genetic ability of all subpopulations converged to the same level even though the initial ability was different from each subpopulation.

In the island-model GS simulation, I generally assumed that (i) the number of migration individuals was set to 1 in each subpopulation, (ii) the exchange interval was one (i.e., breeder should exchange selected individuals every cycle), and (iii) one population derived from a single bi-parental cross constructed one subpopulation. These assumptions attained better results than the other assumptions, as derived below. That is, the island-model GS simulation, in which two individuals were exchanged from each subpopulation, resulted in the similar genetic ability to the island-model GS with one individuals' exchange (Fig. 4.4a). The simulation with different migration intervals (i.e., exchanging event was conducted every 2 – 5 cycles) resulted in lower genetic gain than the island-model GS in which individuals were exchanged every cycle (Figs. 4.4b – 4.4e). The simulation with randomly separated initial population represented lower gain than the simulation with the initial population separated according to their parents (Fig. 4.4f).

In all GS breeding, decreases of the genotypic values were observed after the first selection cycle (Fig. 4.3). Figure 4.5 shows an example of simulations during the first three cycles of bulked GS in the population derived from multiple bi-parental crosses. The distribution of genotypic values of breeding population showed the improvement of average genotypic values through the selection cycles.

4-3-3. Genetic variance and prediction accuracy

The all GS breeding schemes show the similar trends in the genetic variance (Fig. 4.6). The variance decreased after the first selection and increased a little after the second selection in all three kinds of GS breeding. Then, after the third selection, the genetic variance decreased gradually. For the discrete GS and the bulked GS, although both breeding strategies showed similar levels of decline of variance at the first selection, the level of increase was larger in the bulked GS than in the discrete GS at the second selection. Because of this difference of increment of variance, the breeding population of the bulked GS maintained higher variance

than another in early generations (Figs. 4.6a and 4.6b). The island-model GS showed different levels of genetic variances among the initial subpopulations. At the first selection, the island-model GS showed the lower variance than the others, and increased variance more largely than the others. The subpopulations in island-model GS did not converge until the 5^{th} selection cycles, while they almost converged after the 6^{th} selection cycle (Fig. 4.6c).

In the bulked GS, almost all loci (99.03% on average) were homozygous in the initial population (Fig. 4.7a). In the next generation, breeding population had many heterozygous loci (23.98%). The proportion of fixed loci increased rapidly with repeated selections. Figure 4.7 (b) shows the proportion of fixed loci of the bulked GS. It was 19.63% in the initial population. And, the proportion increased to 60.48%, 75.25%, 85.90% and 92.36% in fifth, tenth, fifteenth and twentieth generation, respectively.

The prediction accuracy represented the same trend as that of genetic variance in the all GS breeding schemes (Fig. 4.8). The first selection, at which the training population included the predicted candidates, attained the highest prediction accuracy. Then, it decreased at the first selection and increased a little at the second selection. From the third selection, the prediction accuracy declined gradually. The decline of accuracy at the first selection was smaller in the bulked GS than in the discrete GS (Figs. 4.8a and 4.8b). The island-model GS showed much smaller decrease of accuracy at the first selection than the others. The prediction accuracy varied among subpopulations over the 20 selection cycles (Fig. 4.8c).

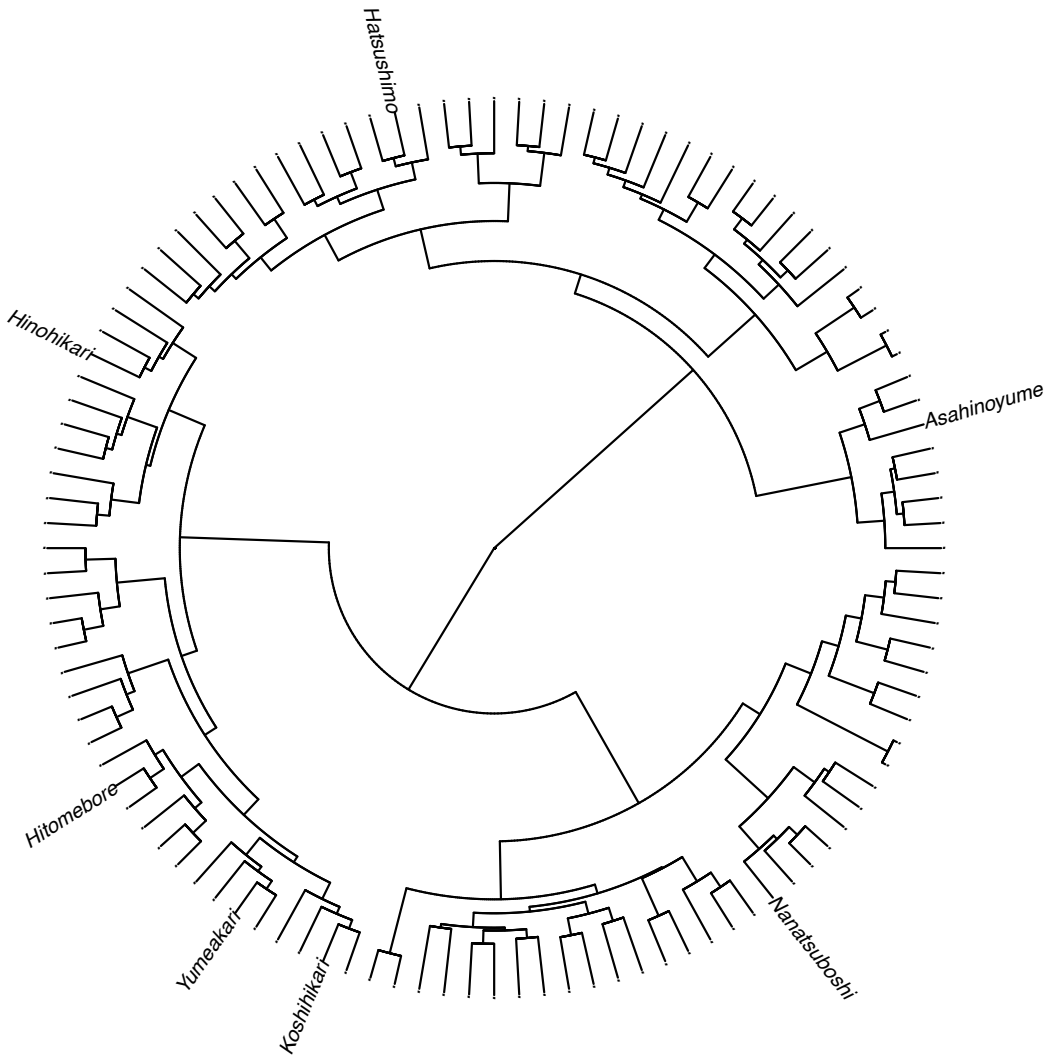


Figure 4.2. Cluster analysis of the 112 rice cultivars by 3,102 markers.

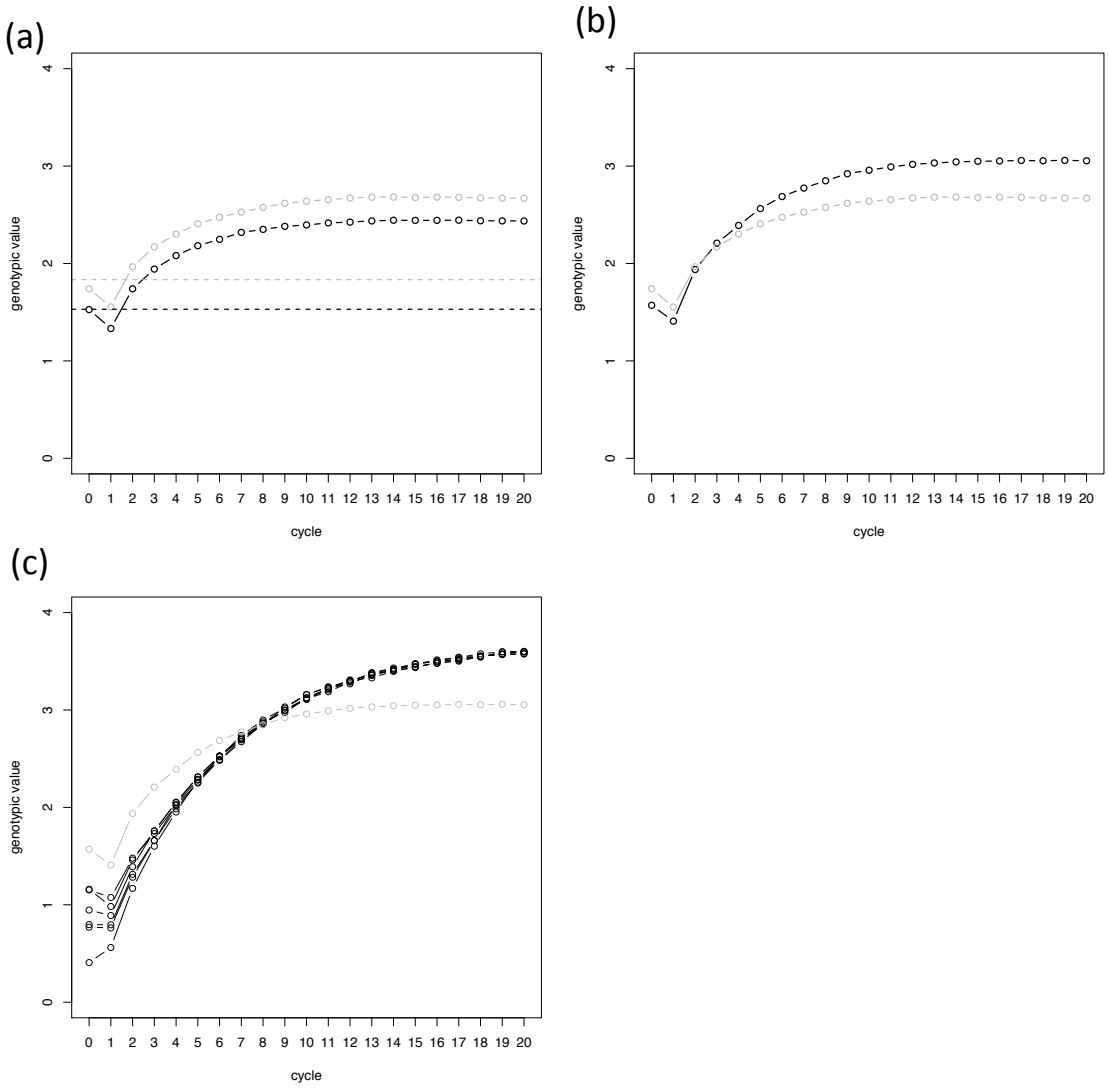


Figure 4.3. Genotypic value attained through selection cycles. (a) The solid lines show the results of the discrete GS breeding. The dashed lines represent the largest genotypic values of lines in the six RILs populations used as a breeding population. Black and gray lines represent the results of Koshihikari × Hatsushimo and the results of the best population in each simulation trial, respectively. (b) Black line shows the result of the bulked GS. Gray line represents the result of the discrete GS that is the same as a gray line in (a). (c) Black lines represent the results of the each subpopulation in the island-model GS in which the exchange interval was 1 and the number of exchanged individual was 1 in each subpopulation. Gray line shows the result of the bulked GS.

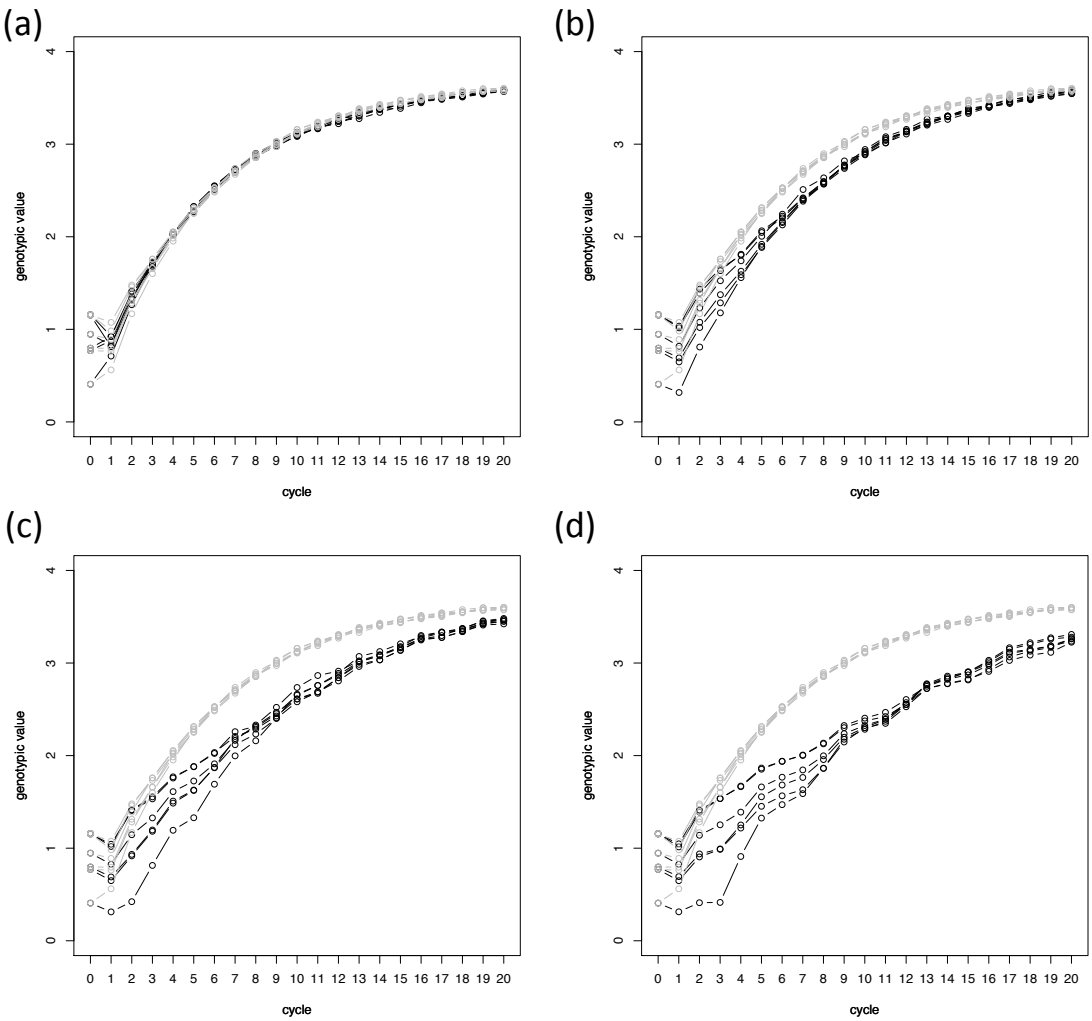
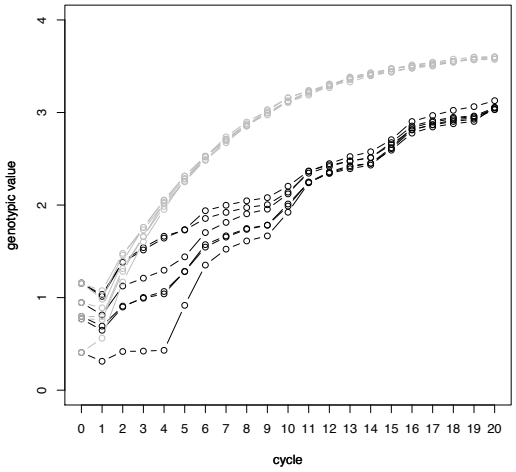


Figure 4.4. Genotypic values attained through selection cycles in the island-model GS. Gray lines represent the genotypic values attained when the migration interval was 1 and the number of exchanged individual was 1. Black lines show the results of following settings: (a) the migration interval was 1 and the number of exchanged individual was 2, (b) the migration interval was 2 and the number of exchanged individual was 1, (c) the migration interval was 3 and the number of exchanged individual was 1, (d) the migration was 4 and the number of exchanged individual was 1, (e) the migration interval was 5 and the number of exchanged individual was 1, and (f) the migration interval was 1 and the number of exchanged individual was 1 while the initial subpopulations were constructed randomly regardless of their parents.

(e)



(f)

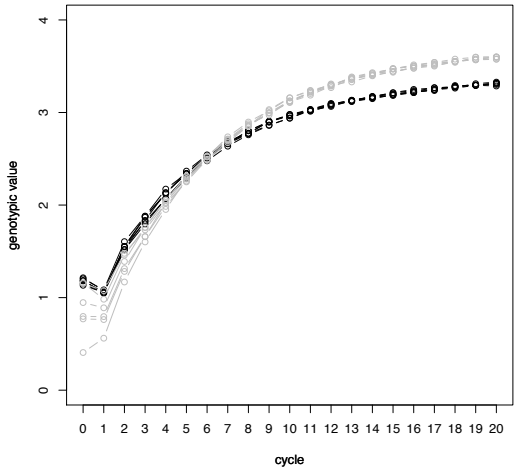


Figure 4.4. (Continued)

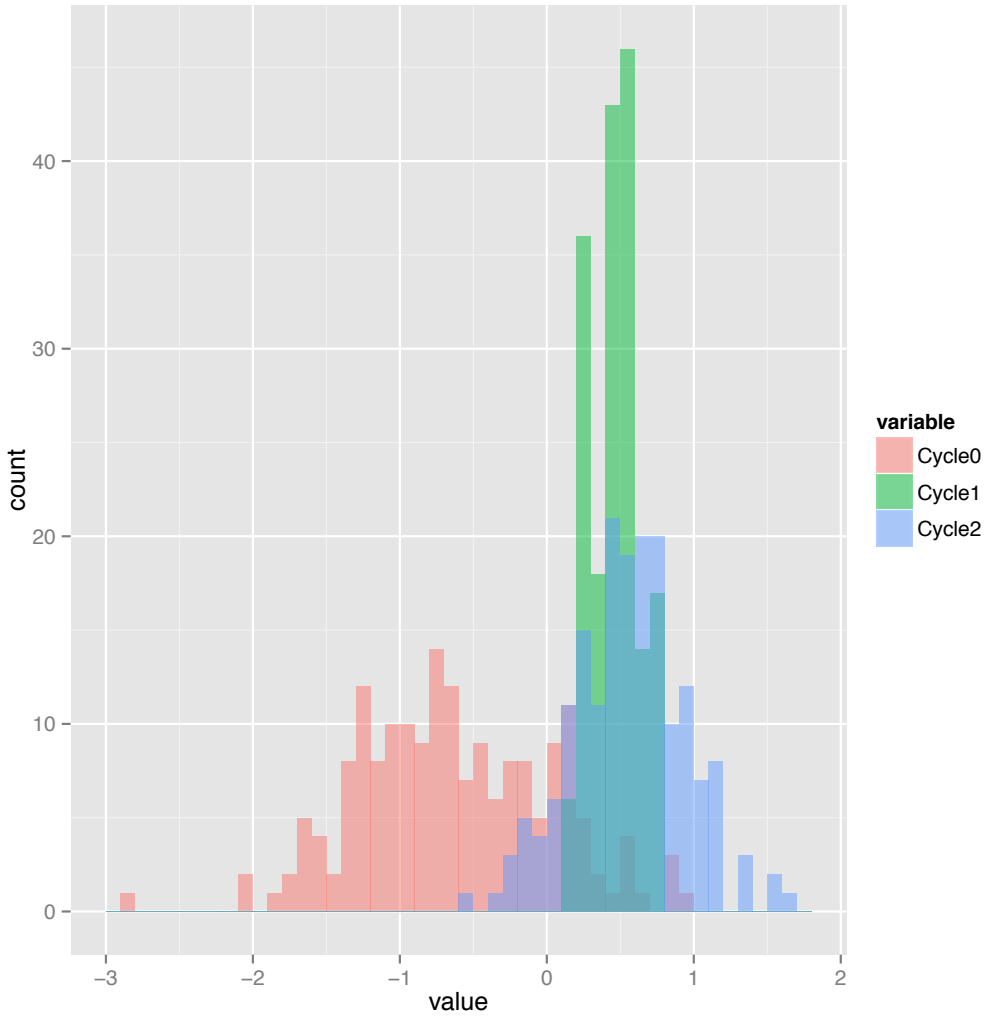


Figure 4.5. The distribution of genotypic values at one simulation trial in the bulked GS. Red, greed and blue distribution show the values of the initial population, the population experienced one selection cycle, and the population experienced two selection cycles, respectively.

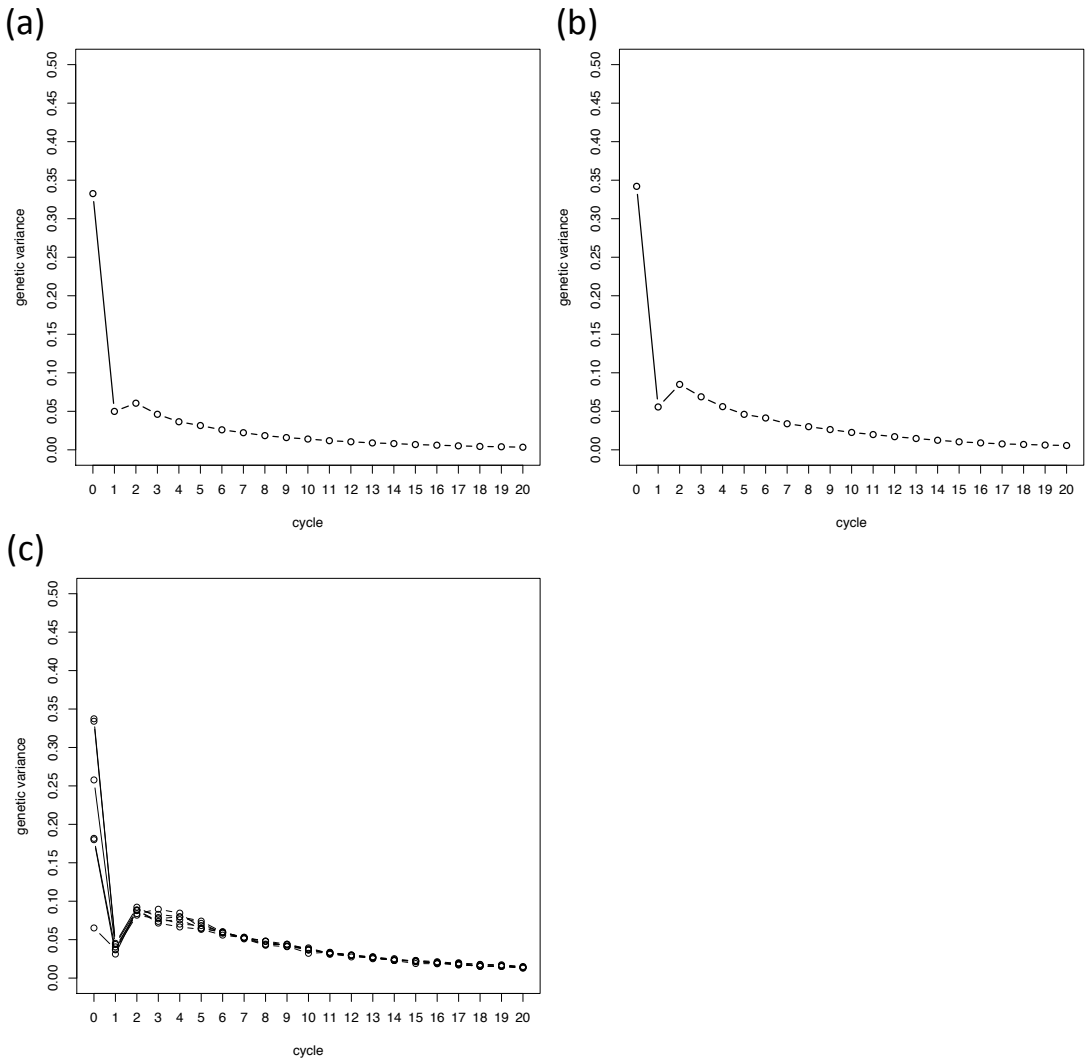


Figure 4.6. Genetic variance shown through selection cycles. (a) The results of the discrete GS in a population derived from a single bi-parental cross between Koshihikari and Hatsushimo. (b) the result of the bulked GS. (c) The results of the each subpopulation in the island model GS.

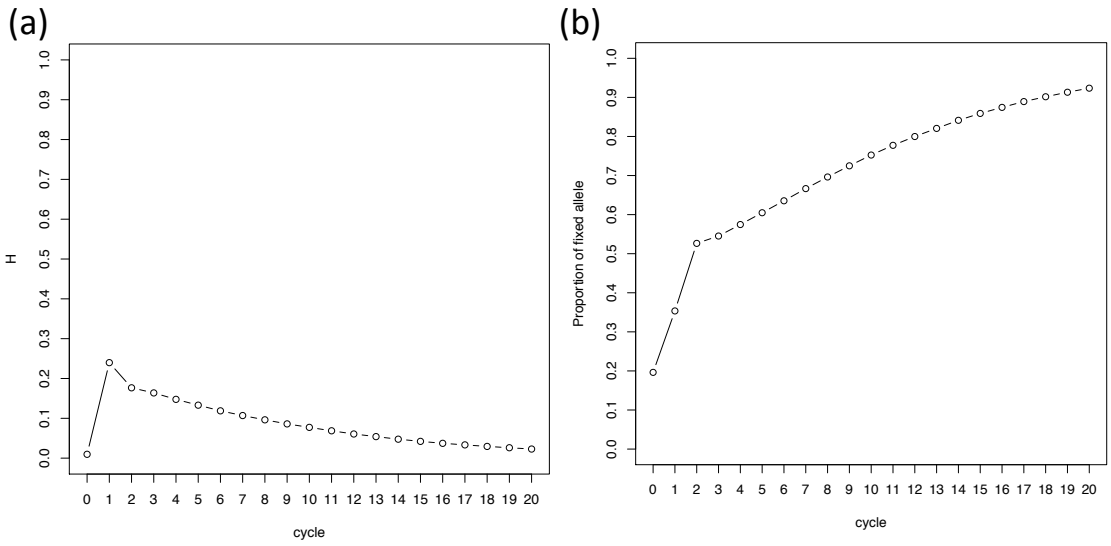


Figure 4.7. (a) The proportion of heterozygous loci among the all SNPs in the bulked GS. (b) The proportion of fixed loci among the all SNPs in the bulked GS.

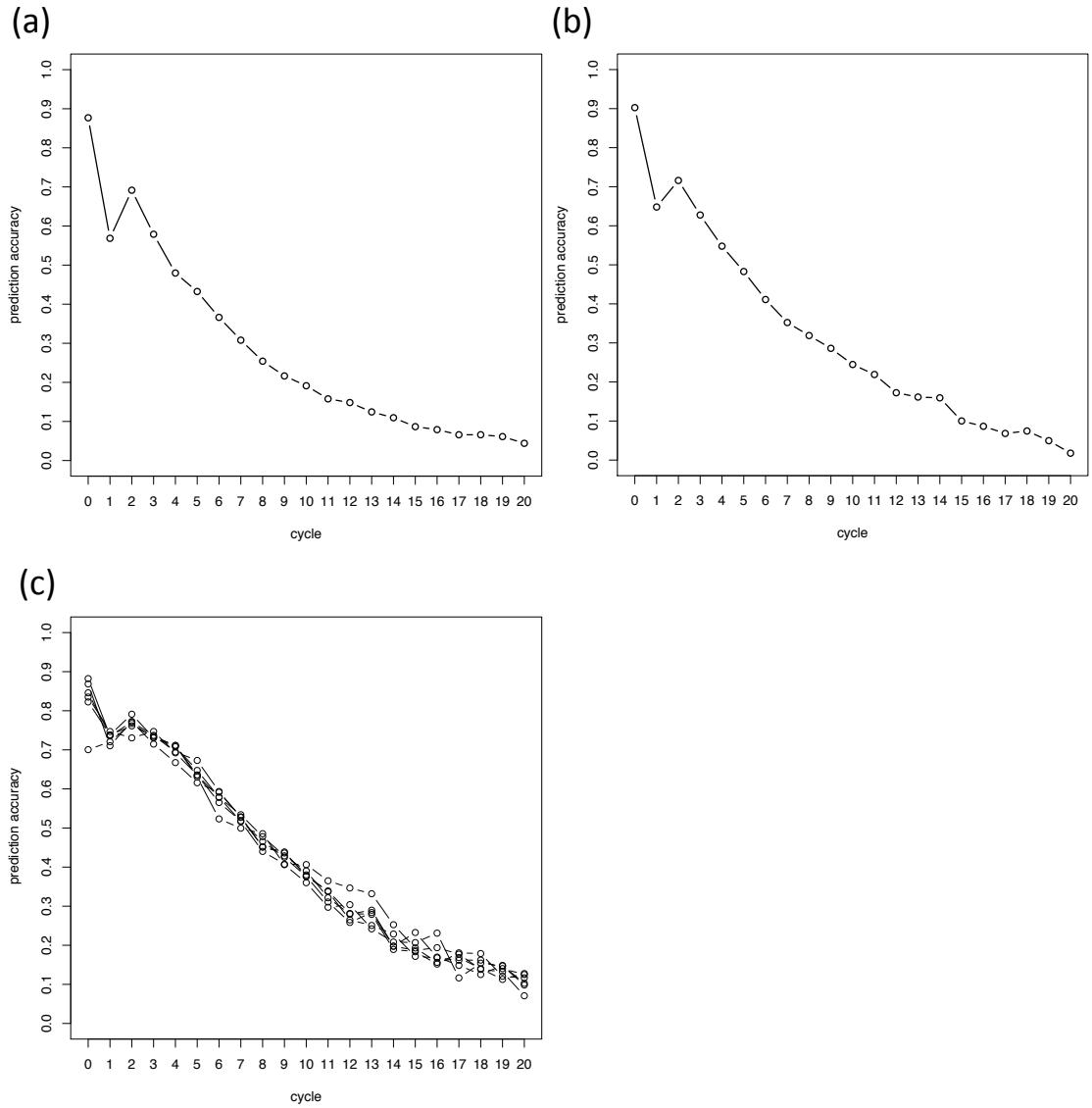


Figure 4.8. Prediction accuracy attained through selection cycles. (a) The results of the discrete GS in a population derived from single cross between Koshihikari and Hatsushimo. (b) the result of the bulked GS. (c) The results of the each subpopulation in the island-model GS.

4-4. Discussion

4-4-1. Structure of breeding population

This study performed GS simulations in rice, which is an autogamous crop. By using a real marker data of rice varieties, population structure existing among the rice varieties could be involved in the simulations. First, I compared the potential of GS breeding with the maximum potential of RILs (Fig. 4.3a). The superiority of GS breeding suggested the efficiency of recurrent selection, in which selection and cross are repeated, even if the target crop was an autogamous crop and single-round robin was applied instead of random mating. This suggested that recurrent selection with repeated selections and crosses can make various combinations of genes in the breeding population and gradually fix favorite alleles through selection.

The reason of the decline of genotypic values at cycle 1 (Fig. 4.3) was the relationship between genetic variance and heterozygosity. It is well known that self-fertilization increases genetic variance between lines (Falconer and Mackay 1996). The initial population for GS breeding was F_6 population, which experienced repeating self-fertilization and harbored a large number of homozygous loci. After the first selection cycle, the proportion of heterozygous loci increased because of outcrossing (Fig. 4.7a). It resulted in the rapid decrease of genetic variance (Fig. 4.6b). This explained why the genotypic values of good plants in the second generation were lower than those of the initial population (Fig. 4.3) while the mean of breeding population improved (e.g., Fig. 4.5).

For recurrent selection involving GS, the initial population should have high genetic variation because of its ability to contribute to genetic improvement. It is natural that a breeding population derived from multiple bi-parental crosses is more efficient than a population derived from a single superior bi-parental cross when we assume the same breeding population size. The comparison between these breeding populations (i.e., comparison between the discrete GS and the bulked GS) proved that a breeding population derived from multiple bi-parental crosses could attain higher genotypic values than that derived from a single bi-parental cross, while the population size was same between the two populations (Fig. 4.3b). In the discrete GS, however, I selected one best population out of six breeding populations as the outcome. It means that the size of the initial breeding population was 1,080 in total in the discrete GS, which was six times lager than that in the bulked GS. This result suggests the importance of mixing a large genetic diversity in one population to create new combinations of genes. A population derived from multiple bi-parental crosses had higher genetic diversity in one initial population, and showed more moderate decline of genetic variance than a population derived from a single bi-parental

cross (Figs. 4.6a and 4.6b), resulting in more rapid genetic improvement and slower attainment to a plateau in the bulked GS (Fig. 4.3b). Higher genetic variance among the breeding population contributed to higher potential of GS owing to higher potential for genetic improvement involving a large number of related loci. And, as mentioned above, the genetic variation should be mixed via crossing among genotypes. From the result of the comparison of the discrete GS with the bulked GS, it is suggested that population derived from multiple bi-parental crosses can involve a large amount of genetic variation in an initial population, resulted in large improvement of genetic ability through GS breeding.

4-4-2. Island-model GS

In the present study, I proposed the island-model GS that was inspired by the island model in EAs, and evaluated the efficiency of the island-model GS by using computer simulation reflecting the situation of plant breeding. The island-model GS was first proposed by the present study. The island-model GS had uncertainty about its success because of some differences between plant breeding and EAs: (i) mutations occur frequently in EAs while they are limited in an actual plant population, (ii) EAs generally simulates a large population although the size of breeding population is limited in plant breeding, and (iii) EAs can gain high ability through a large number of selection cycles in short time by using computer although plant breeding requires much time to evaluate and select individuals. The present study conducted simulations reflecting actual rice breeding situation to evaluate the efficiency of the island-model GS in autogamous plant breeding. I used a population derived from multiple bi-parental crosses as an initial breeding population in the simulation of the island-model GS. In the simulation, the island-model GS showed a good performance particularly in the later cycles (Fig. 4.3c). The result suggests that the small population size (i.e., 180 individuals in this study) is enough to implement the island-model GS in actual breeding. Single-round robin was applied as a rule for making crossing because of the difficulty of random mating in autogamous species, causing no problem in implementing the island-model GS. These results suggest that the island-model GS is efficient as a selection strategy even though there are differences between actual breeding and EAs. In plant breeding, however, the time for breeding is limited. Thus, breeders should select a proper breeding scheme according to their purpose. If we can conduct only a few cycles of selection, we should select the bulked GS. On the other hand, if we have several years and can have a larger number of GS cycles by accelerating generations, we should select the island-model GS.

It is important for all breeding program to maintain genetic diversity in a breeding population. For the bulked GS, because the initial breeding population was composed of multiple families derived from a single bi-parental cross, a particular family might be preferably selected at the first selection cycle. Actually, in 99 out of 100 simulation trials, one or more family disappeared at the first selection cycle. For 75 out of 100 trials, the selected plants showed the selection bias (chi-square test showed significant gaps from the equal proportion for each family; p-value < 0.05). In the island-model GS, the selection bias was prevented by assuming that each family derived from a single bi-parental cross as an initial subpopulation. That is, it was important to separate each family into an initial subpopulation. The island-model GS attained lower gain when the initial subpopulations were separated randomly regardless of their parents than when the initial subpopulations were separated according to their parents in later generations (Fig. 4.4f). This result also suggests the importance of separating each family into an initial subpopulation. Thus, the genetic heterogeneity among subpopulations is an important factor that makes the island-model GS advantageous over the bulked GS. Moreover, the island-model GS reached a plateau slower than the bulked GS (Fig. 4.3c). Improvement in the bulked GS was rapid in the initial cycles (Fig. 4.3a) because selection bias at the first cycle favored populations derived from a specific cross with high ability. This also led the decline of genetic variance and fast plateau of improvement in the bulked GS. In the island-model GS, different genetic variations were conserved in each subpopulation, resulting in maintenance of genetic diversity in a whole population. Migration of genes that was attained by exchanging parents (i.e., selected plants) between subpopulations conserved genetic diversity in subpopulations and improved the genetic potential of the whole subpopulations. The issue of maintenance of genetic diversity in a breeding population also related to migration rates. In the present study, I assumed migration size as one, which seemed a limit when the selected size was three in each subpopulation. This migration size worked as efficient migrations in the simulations. When the migration size was two, the attained genotypic value was not much different, but the characters of subpopulations were unified earlier (Fig. 4.4a). It is because the large number of exchanged parents resulted in the early assimilation between subpopulations, then the same situation as when the initial subpopulations were made randomly. The island-model GS with the migration interval one (i.e., migration occurred every generation) represented the best results among the intervals of 1 – 5 cycles with same size of exchanged individuals in the simulation (Figs. 4.4b – 4.4e). In the simulations of the island-model GS, small population size and strong selection intensity were assumed. Selecting only three plants in

one subpopulation occurred a severe genetic bottleneck in the subpopulation. Because of the severe bottleneck, genetic variance in each subpopulation decreased drastically in two or three cycles of selection without migration. Although the drastic decline of genetic variation in each subpopulation resulted in the lower genetic gain through 20 selection cycles of the island-model GS with longer migration intervals, they did not show the tendency to reach a plateau through the 20 cycles (Fig. 4.4). The reason for this may be the maintenance of different genetic variations in subpopulations. Wright (1943) mentioned that random differentiation has a tendency to cause different adaptive trends and different process of selection in different subpopulations even under uniform environmental conditions. My simulations of the island-model GS, in which I imposed selection to the identical direction for all subpopulations, might follow this situation. The island-model GS with longer migration intervals promotes the utilization of large genetic variations in a whole population. In a breeding program, it might be better to conduct the island-model GS with the shorter migration interval because of the demand of release of new cultivars in a shorter period of time. However, if breeders can spend much time or hope long-term selection, the migration interval should be decided on the basis of the balance between the pace of improvement and the pace of reaching a plateau.

4-4-3. Suggestion for breeding of autogamous plants

The potential of the island-model GS in the breeding of an autogamous species was demonstrated by breeding simulations. For the actual plant breeding, distinctness, uniformity and stability are required to release of new cultivars (Brown and Caligari 2008). In autogamous crop species, pure lines are made as a new cultivar to realize its uniformity and stability. In general, cultivars in market experienced from 6 to 7 cycles of selfing (Broun and Caligari, 2008). Considering the rapid fixation of alleles in recurrent selection (Fig. 4.7), it might be required to perform a few cycles of self-pollination after the genetic improvement reached a plateau.

GS method in this paper focused on maintenance of genetic variation in a breeding population to select autogamous species. On the other hand, McClosky et al. (2013) suggested the GS process with selecting more homozygous individuals. They showed that a large genetic variance caused by self-fertilization (Falconer and Mackay, 1996) resulted in the improvement of genetic ability in GS. In their process, however, few new combinations of genes would appear in a breeding population. As suggested by the present study, GS with recurrent selection is a good strategy especially for a long span.

In the present study, GS was conducted to evaluate a single plant accurately. However, the prediction accuracy declined with repeated selection cycles. The main reason for the decline should be the increasing genetic distance between the training and breeding populations. In the simulation study of GS in barley (Jannink, 2010), prediction model was updated every cycle by making DHs after selection. In the present simulation, a prediction model was built based on 1,080 F_6 lines, and the prediction model was used throughout a breeding program. Selection accuracy decreased with repeated selections (Fig. 6.8). If the prediction model can be updated, selection accuracy would be improved (e.g., Iwata et al. 2011; Jannink 2010; Chapter 3 in this dissertation). Updating prediction model, however, requires much efforts and time. The present results show the potential to use one prediction model for a long time. Moreover, in the simulations, I assumed the prediction model was trained by using F_6 lines derived from multiple bi-parental crosses. The phenotype and marker genotype data of recombinant inbred lines and backcross inbred lines are usually collected in public and private sectors, suggesting that breeders can use existing data to build a prediction model. The optimal updating time of a prediction model should be considered based on time, cost, and effort for preparing a new training population as well as the accuracy of the updated model.

The present study assumed the initial population consisting of six families derived from six combinations of bi-parental cross. The seven varieties used as parents of the initial breeding and training population were selected to represent the genetic diversity in the 112 cultivars well on the basis of their marker genotypes. It may also be possible to choose parental varieties on the basis of phenotypic variation of target traits. We often conduct a breeding program under a certain restriction (e.g., a lower limit of quality of seed or fruit). In this case, it may be efficient to choose parents according to the phenotypic values, although we cannot incorporate the whole genetic diversity in parental candidates.

For island-model GS, it is possible that some public or private sectors of plant breeding work together to create a new cultivar. A breeding population has been selected according to the local adaptation and maintains different genetic diversity from other populations. We can utilize this situation in the island-model GS by assuming a breeding population in each region as a subpopulation. In each region, breeders can conduct their breeding programs by using their own population. They can occasionally exchange a part of their cultivars for cultivars of other regions to introduce new genetic variation into their population. In that case, the estimation of the suitable migration interval might be important.

Chapter 5

Simulation of the impact of mis-labeling on genomic selection in cassava

5-1. Introduction

Since the development of GS as a replacement for conventional MAS by Meuwissen et al. (2001), a number of simulation studies have been conducted to evaluate its efficiency for plant breeding. These simulation studies suggested that GS is effective for breeding (e.g., Bernardo and Yu, 2007; Heffner et al. 2010; Iwata and Jannink, 2011). The efficiency of GS is expected to be greatest for tree breeding, which requires long time per cycle for phenotypic selection, considering gain per unit time (Wong and Bernardo, 2008; Grattapaglia and Resende, 2011, Iwata et al., 2011). Such simulation results have created interest among breeders to realize GS. The revolution in sequencing technologies has enabled fast sequencing and inexpensive genome information (e.g., Elshire et al., 2011). This revolution makes it possible for breeders to conduct GS in actual breeding. Recently, some results of field trials of genomic selection have also been reported in the field of plant breeding (e.g., Asoro et al., 2013; Massman et al., 2013). It is believed that more and more results from field testing of genomic selection will be reported in the next few years.

Actual plant breeding in the field may differ in one important respect from simulations: humans make mistakes, and that should be taken into account. In PS, labeling errors may to some extent be self-correcting because breeders evaluate phenotype of candidate plants and select them according to the observations. Even when breeders happen to swap a selected plant for another one, this error would only persist to the next generation, when the inferior progenies in the generation would be removed by selection. In GS, however, mis-labeling may not be corrected so easily. If the connection of phenotype data and marker genotype data is wrong in the training population used to build a prediction model, the error may reduce the accuracy of

genomic prediction and will continue affecting selection until a new prediction model is built without the erroneous data. Switching or mis-labeling genotypes may happen in a number of ways. It is possible to extract DNA from wrong plants, swap the samples for wrong ones during transportation to the laboratory, misplace the DNA samples during laboratory work, or store the wrong marker genotypes or phenotypic values in the data sheet. The more steps the breeding scheme requires, the more mistakes opportunities there are for error. Moreover, it is difficult for breeders to notice these kinds of mistakes in GS because they cannot verify the selection results by looking at plants at the time of crossing. Ly et al. (2013) reported that they detected mis-labeling when they built a prediction model when they conducted GS in cassava (*Manihot esculenta* Crantz). They used historical phenotypic evaluation data for the training population to build a prediction model, and identified the potential of labeling error in 23 clones out of 626 clones. They mentioned the potential of the remaining clones to also be mis-labeling in ways that were not detected.

It is difficult to detect and prevent mis-labeling errors. This is not only for plant breeding but also for other fields. For example, mistakes by nurses are a critical problem in medical treatments. To prevent mistakes in nursing, many systems including education has been proposed (Philipsen, 2011). It costs money, time and effort to detect and prevent mistakes by humans. The cost may be beneficial for nursing, where mistakes can have direct consequences for patient health. In plant breeding, however, controlling mistakes too strictly may not be cost effective if the mis-labeling does not have a large impact. To determine the levels of control of mis-labeling, it is necessary to evaluate the impact of mis-labeling.

In the present study, I evaluated the impact of mis-labeling in cassava breeding using simulations. Cassava is produced mainly in Africa, and is the most produced crop there (FAOSTAT, 2014). In spite of the importance of cassava for the world's food supply, scientific research on cassava started later than for other crops because of its unfamiliarity outside the tropical and subtropical regions where it grows (Ceballos et al., 2004). Because the phenotypic selection cycle in cassava is lengthy, GS may be quite useful for this crop. However, once mis-labeling has happened in a breeding population or a training population in cassava breeding, it will have large impact on outcomes of GS breeding. Because genotypes are propagated vegetatively in cassava, wrong marker genotypes caused by mis-labeling are used continuously over generations even when prediction models are updated with new phenotypic data collected with the propagated clones. The simulation study of GS in cassava including mis-labeling may be useful to decide the levels of control the mis-labeling. Here, I assumed two types of

mis-labeling, in which marker genotypes and phenotypic values were mismatched. First, I assumed an excess of candidate progeny (e.g., backup for a breeding population) that may be genotyped instead of selection candidate mistakenly. Second, I assumed a breeding population and switching genotype data among selection candidates. To evaluate the impact of the levels of mis-labeling, I compared gain from selection, selection accuracies, and changes in genetic variance in scenarios with differing levels of mis-labeling.

5-2. Methods

5-2-1. Simulation settings

The simulated species had 18 pairs of chromosomes ($2n$ = 36) of 110 cM length each. The genotypes of the founder individuals in the base population were simulated by coalescent simulation using GENOME (Liang et al., 2007). In this coalescent simulation, one population was assumed whose population size and effective population size were 100. Recombination rates were set according to their genetic distances, and chromosomes were assumed divided into 11,000 segments each. All SNPs were defined to have minor allele frequencies of greater than or equal to 0.01. I simulated one polygenetic trait controlled by 100 causal QTL such as yield and retained 2,500 SNP markers for genomic prediction. I assumed only additive effects (i.e., no dominance or epistatic loci) for simplicity to discuss the impacts of breeding schemes and mis-labeling. These effects were sampled from a normal distribution with the mean of 0.0 and the variance of 1.0. Once created, the effect sizes were adjusted to make the initial genetic variance equal to 1.0. All genotypic and phenotypic values were calculated from these QTL effects in the simulations.

5-2-2. Breeding schemes

I simulated a scheme of cassava genomic selection with four cycles of selection (Fig. 5.1). The breeding program started from historical phenotypic evaluation data of 200 existing clones. Then, 40 individuals were selected on the basis of their predicted genotypic values using a prediction model built from the existing data. After random cross among the selected individuals, 600 seedlings were obtained. If seedling data was used to build the prediction model, the prediction model was updated with the historical data and the seedling data that was phenotyped before selection. If seedling data was not used to update the prediction model, the prediction model was not updated at the second selection. 40 seedlings were selected based on the predicted values by using the existing prediction model. After the second selection, the prediction model was updated at every cycle by using all available phenotypic data. The size of breeding population was 600, and 40 seedlings were selected every cycle. The seedlings that constituted a breeding population were separated into two parts, selected individuals and non-selected individuals. The non-selected individuals were propagated in a clonal evaluation trial (CET) to update the prediction model in the next generation. The selected individuals were propagated to a crossing nursery to create the next generation, which is why they were not included in the CET. All clones were subsequently propagated to a preliminary yield trial (PYT)

to update the prediction models after the next generation.

The error variance was 1.0 in the historical data so that the clone-mean heritability was 0.5. This low error variance reflected the assumption that founder clones would have been repeatedly evaluated in the past and would therefore be well-characterized. The error variances were 36.0, 16.0 and 9.0 for traits measured in seedling trials, CET and PYT, respectively. Breeders can evaluate only a single plant at the seedling stage, so I assumed a large error variance at this stage, consistent with a heritability lower than 3%. About 20 and 5 plants are commonly evaluated for each clone in PYT and CET, respectively, justifying lower error variances for these trials. Note that, for all types of trials, the error variance is inflated by GxE variance that cannot be statistically removed from a single-location trial. In the present simulation, I performed two types of selection: (i) using seedlings, and (ii) not using seedlings to build a prediction model. Seedlings were considered not to contribute to precise estimate or prediction because of their large error variances. If I could prove that seedlings were useless to predict the abilities of genotypes, breeders can cut off their effort and time to evaluate seedlings. Therefore, I conducted and compared the two types of breeding schemes in the simulation.

Each simulation scheme was repeated 100 times. All simulations were performed in R, version 3.1 (R Development Core Team, 2014).

5-2-3. Genomic prediction model

GBLUP, as implemented in the R package "rrBLUP" (Endelman, 2011), was used to estimate and predict the breeding values of accessions. I assumed different error variances at each phenotyping stage (i.e., historical data, seedlings, CET, and PYT). To take this situation into account, the function "kin.blup" in package rrBLUP was modified (Jeffrey B. Endelman, pers. Comm. 27 May 2014), and used the modified function for genomic selection. The equation [2.12] represents the simplest original expression to solve G-BLUP. In the situation that is considered here, the original equation is

$$\begin{aligned} y &= X\beta + Zu + \varepsilon \\ u &\sim N(0, K\sigma_u^2) \\ \varepsilon &\sim N(0, R\sigma_e^2), \end{aligned} \qquad [5.1]$$

where y is the phenotypic values, and β and u represent the fixed effects and random effects, respectively. X is a full-rank design matrix for β, and Z is the design matrix for u. The residual is shown as ε. K is a positive semi-definite matrix (i.e., kinship matrix calculated by pedigree or marker genotype), and R is a diagonal matrix proportional to the error variances of the

observations, *y*. The usual mixed model assumes that observations are distributed with constant variance, but the *R* matrix allows this assumption to be relaxed. To solve the problem, the equation was multiplied by $R^{-1/2}$ before solving the mixed model:

$$\tilde{y} = \tilde{X}\beta + \tilde{Z}u + \tilde{\varepsilon} \qquad [5.2]$$

$$u \sim N(0, K\sigma_u^2)$$
$$\tilde{\varepsilon} \sim N(0, I\sigma_e^2),$$

where $\tilde{y} = R^{-1/2}y$, $\tilde{X} = R^{-1/2}X$, $\tilde{Z} = R^{-1/2}Z$, and $\tilde{\varepsilon} = R^{-1/2}\varepsilon$. This modified mixed model can be solved in the ordinary way.

In this study, I made the *R* matrix by using the error variance values that were used in simulations. In an actual field trial, however, these values are not known. Thus, users should decide variance values a priori when they use this prediction model.

5-2-4. Mis-labeling

I considered two types of mis-labeling. First, I assumed an excess of candidate progeny that may enter the breeding population. The mis-labeling consisted of sampling DNA from one progeny for genotyping but mistakenly taking a different progeny, under the same label, for phenotyping (Fig. 5.2a). In this way, the genotyped individual did not enter the breeding population though its marker profile was analyzed under the label of different progeny and associated with the latter's phenotypes. Second, I assumed a fixed set of candidate progeny entering the breeding population. The mis-labeling consisted, in pairs of individuals, of associating the phenotype of one with the genotype of the other and vice versa (Fig. 5.2b). I assumed that breeders would perform marker genotyping only once per individual, thus the wrong genotypes would be used throughout the selection cycles and repeated model updating. Further, I assumed that if a mis-labeled individual was selected to become a parent of the next generation, the individual that was phenotyped was planted in the crossing nursery, rather than the individual that was genotyped.

5-2-5. Post-simulation analysis

The results were shown as the averaged value over 100 simulation replications in each scenario.

Genotypic values were represented as the improvement of population mean from Population 0 (i.e., an initial breeding population), thus the values at Population 0 were set to 0.0

in all breeding schemes. The genotypic values were compared in the last populations using the pairwise t-test with Bonferroni correction. Genetic variance was calculated as the variance of genotypic value in a breeding population. Prediction accuracy was calculated as Pearson's correlation coefficient between the true genotypic values and the predicted genotypic values in a breeding population. In scenarios with mis-labeling, the prediction accuracy was calculated only among individuals that were correctly labeled.

I considered the response to selection. In phenotypic selection, the response to selection is represented as

$$R = ih\sigma_A \quad [5.3]$$

where R is the response to selection, i is the selection intensity, h is square root of the narrow-sense heritability, and σ_A is square root of the variance of the additive genetic variance (Bulmer, 1980). In GS, it can be represented as

$$R = ir\sigma_A \quad [5.4]$$

where r is the prediction accuracy of GS model (i.e., correlation between true and predicted genotypic values). I used the true genotypic value to calculate R. When the breeding scenario included mis-labeling, mis-labeled individuals are selected at random such that the response among them is expected to be zero. Therefore, under mis-labeling, the response to selection should be

$$R = (1-e)ir\sigma_A \quad [5.5]$$

where e is the rate of mis-labeling in the training population.

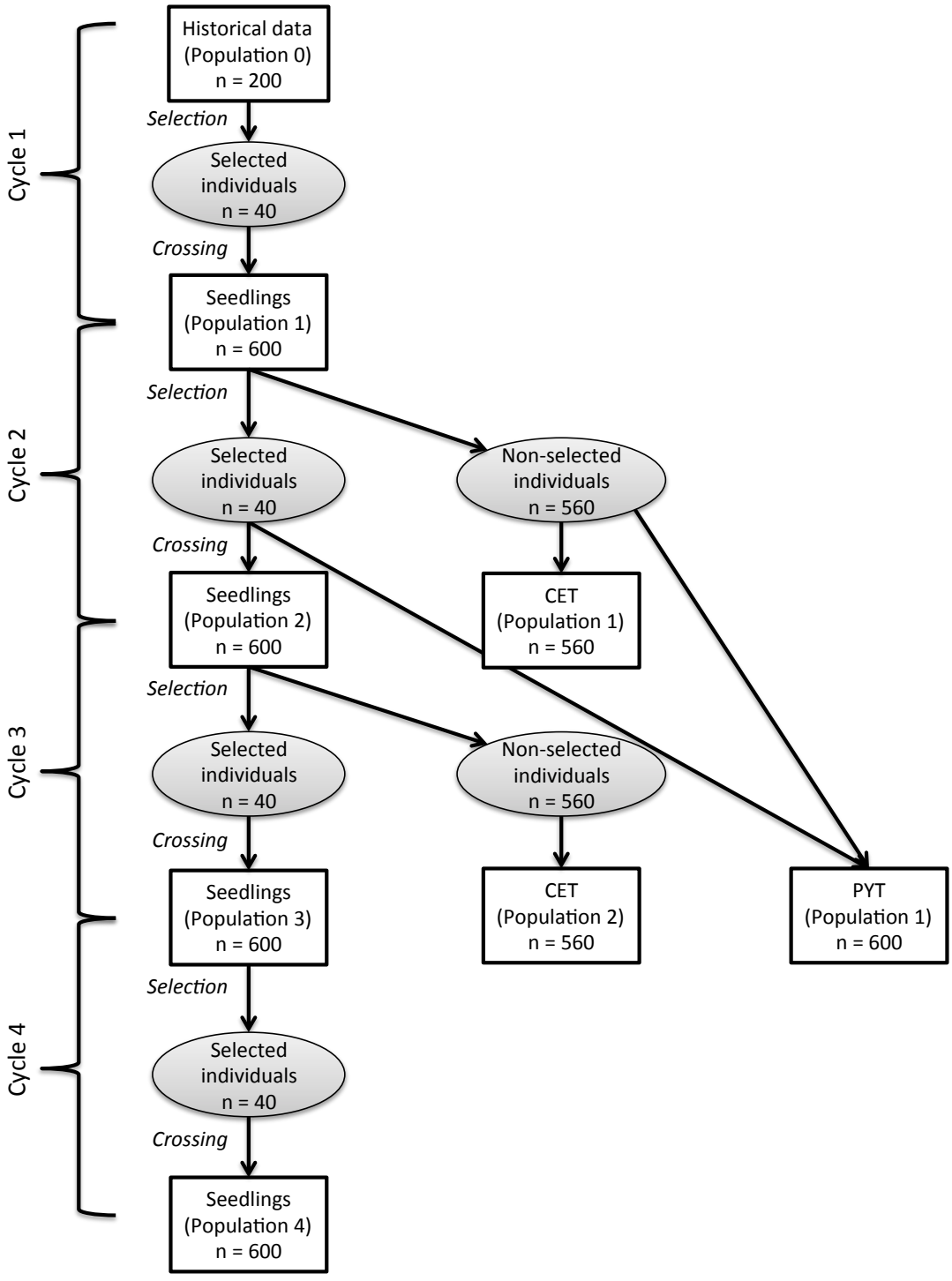


Figure 5.1. Breeding scheme of genomic selection in cassava. CET: clonal evaluation trial, PYT: preliminary yield trial.

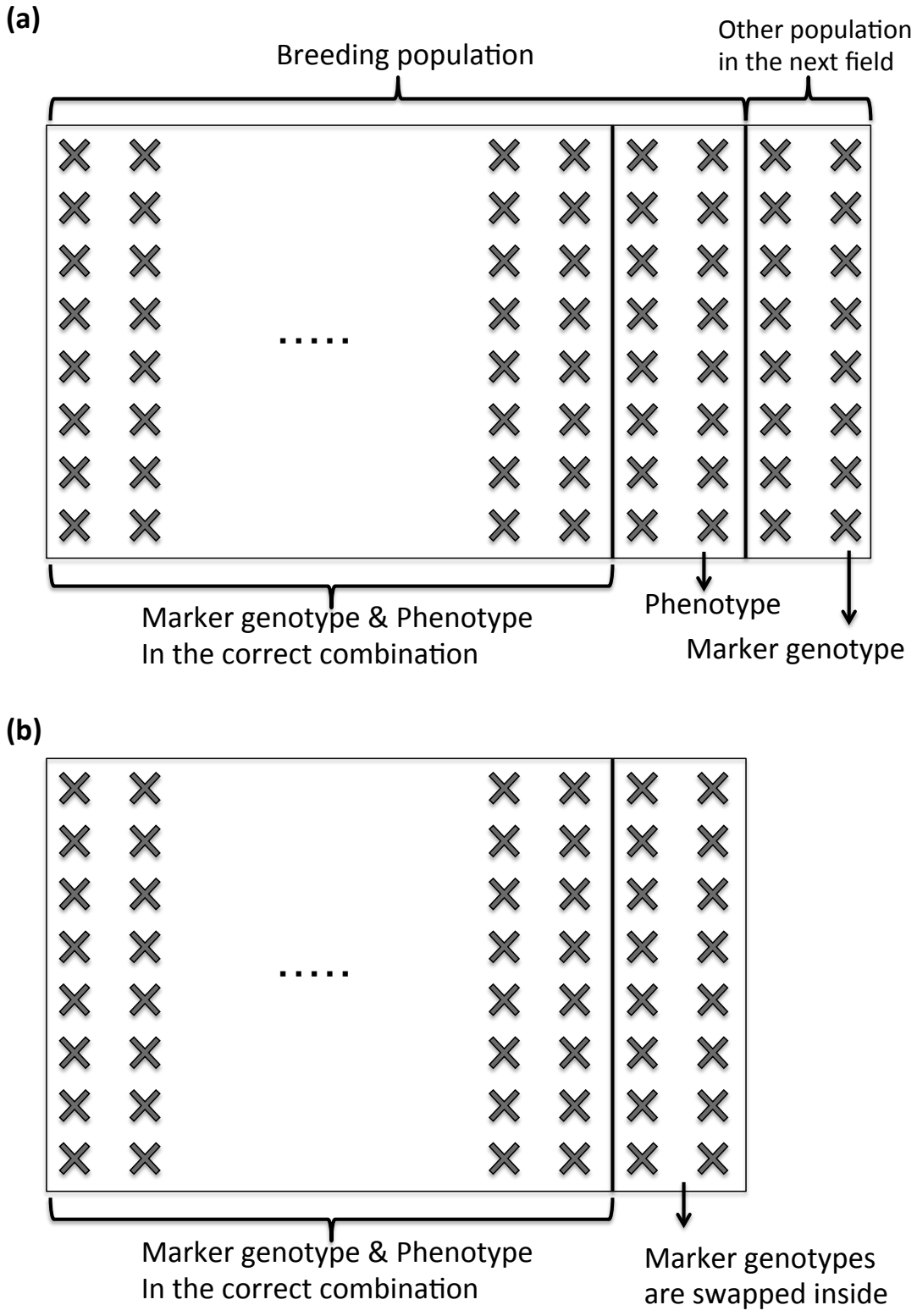


Figure 5.2. Mis-labeling in the training population. Swapping marker genotypes in the training population for wrong ones outside the population (a) and inside the population (b).

5-3. Results

5-3-1. Genetic gain

I evaluated genetic gain (i.e., improvement of genotypic values from an initial population) according to the population mean. Figure 5.3 shows the genetic gain through four selection cycles. In all scenarios, the first selection generated high genetic improvement, after-which stable though lower improvement was shown. The higher the rate of mis-labeling became, the lower genetic gain attained. When mis-labelings happened between populations, all mis-labeling rates showed significant (p-value < 0.05) difference in the process including the seedling data in training, while 0% and 5% mis-labeling rate did not show significant difference in the process excluding seedling data from training. When mis-labeling happened inside a population, 0%, 5% and 10% of mis-labeling rates did not showed significant differences at the last cycle. For both types of mis-labeling, the scheme in which the training population included seedling data attained higher gain than the scheme in which seedlings were excluded from the training population.

Figure 5.4 shows the coefficient of variation (c.v.) of population mean among 100 times of simulation trials in each breeding scenario. The higher the mis-labeling rate became, the higher the c.v. tended to attain. The c.v. increased with selection cycles. In many cases, the breeding scheme including seedlings to build a prediction model showed less variation than the breeding scheme excluding seedlings compared between the results of the same mis-labeling rate and selection stage.

5-3-2. Factors relating genetic gain

The results of the breeding scenario in which marker genotype data were swapped among selection candidates were similar to the results of the scenario in which genotype data were swapped between a breeding population and another population. Thus, here, I will show only the results of the scenario in which genotype data were swapped between populations.

Figure 5.5 shows the genetic variance through four cycles of selection. The decline of variance was more rapid in the scheme using seedlings for training than the other especially when the mis-labeling rate was low. The scenarios with higher error rates tented to maintain higher genetic variance in both breeding schemes. I can see the trend that the higher error rates made the higher genetic variance in all generations.

Figure 5.6 represents the prediction accuracy at four times of selections. At the first selection, the prediction accuracy decreased with the increase of mis-labeling rate. Prediction

accuracy declined drastically from the first selection to the second, as is the case in genetic variance. The accuracy in the strategy using seedlings for training was higher than another. Until the second selection, the scenario with higher mis-labeling rate tended to attain lower prediction accuracy. After that, however, the rank of accuracy changed among the scenarios with different rates of mis-labeling. Figure 5.7 represents the relationship between the prediction accuracy and the rate of mis-labeling from the second to the fourth selection. At the second selection (i.e., selection in cycle 1), the scenarios with higher mis-labeling rate showed lower prediction accuracy. After the second cycles, I cannot see any strong tendency in prediction accuracy to depend on the error rate.

I considered the observed and expected response to selection. The observed response to selection was calculated as the slope of genetic improvement (i.e., slope of Fig. 5.3). The expected response to selection was calculated as equation [5.5] by using the realized error rate, e, prediction accuracy, r, and square root of the genetic variance, σ_A. Figure 5.8 shows the proportion of the observed response to selection with mis-labeling assuming that the slope without the mis-labeling was 100%. Here, I calculated the observed response to selection from Cycle 2 to Cycle 4 by using the mean increment from Population 1 to Population 4 because the slope was stable enough to calculate as a whole (Fig. 5.3) and because the selection intensity was identical in the span of selection. The more the mis-labeling happened, the more the proportion moved away from the line whose slope was –1 and intercept was 100 (dashed line in Fig. 5.8). For 5 or 10% mis-labeling, the scenario using seedlings for training attained lower proportion than the other. For the higher mis-labeling, it was the opposite. Figure 5.9 shows the relationship between the observed and expected response to selection from Population 1 to Population 4 (i.e., from Cycle 2 to Cycle 4). Points seemed to follow the line that represented the same values in observed and expected response. At the Cycle 2 (i.e., selection at the Population 1), the observed response was higher than the expected response in all mis-labeling rates and both types of training population. And, the expected response became higher than the observed response in the later selection cycles.

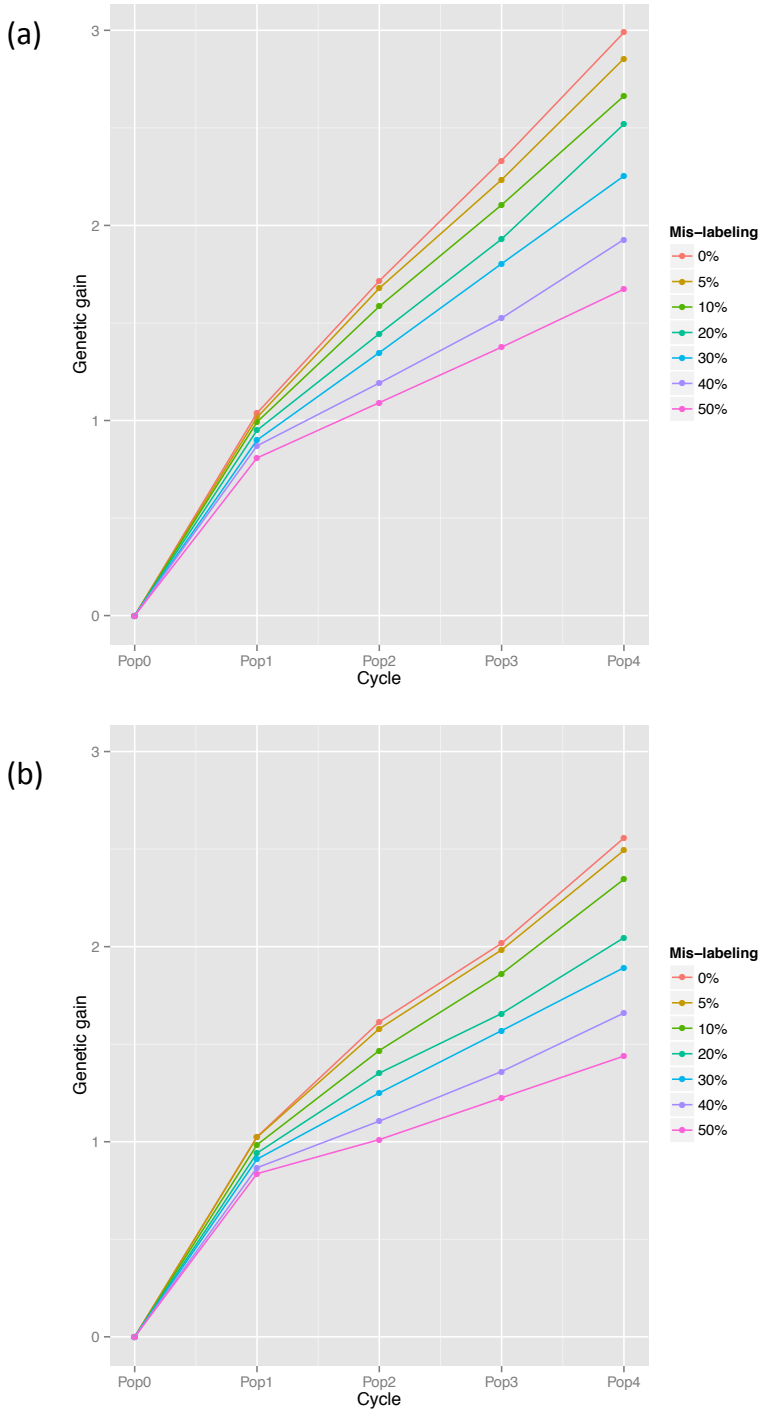
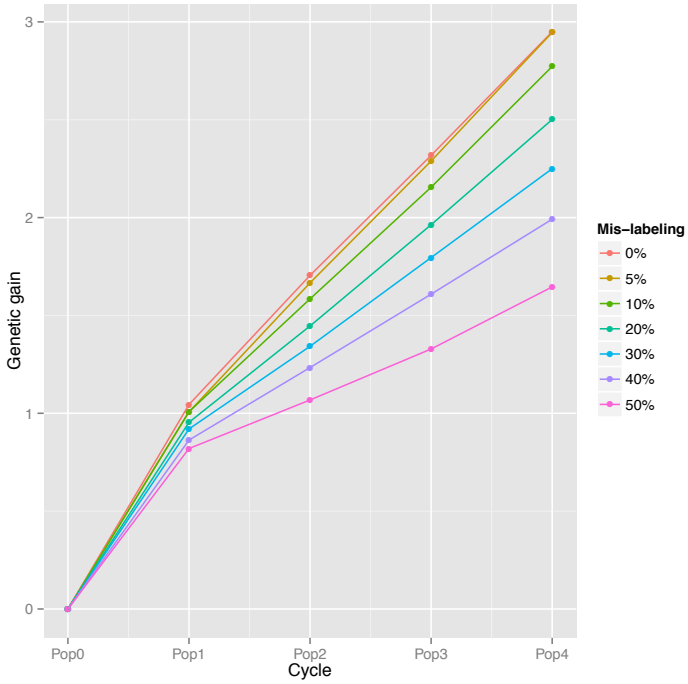


Figure 5.3. Genetic gain through four cycles of selection. (a) marker genotypes were switched between populations, and the training population included seedlings phenotypic data, (b) marker genotypes were switched between populations, and the training population excluded seedlings phenotypic data, (c) marker genotypes were switched inside the breeding population, and the training population included seedlings phenotypic data, (d) marker genotypes were switched inside the breeding population, and the training population excluded seedlings phenotypic data.

(c)



(d)

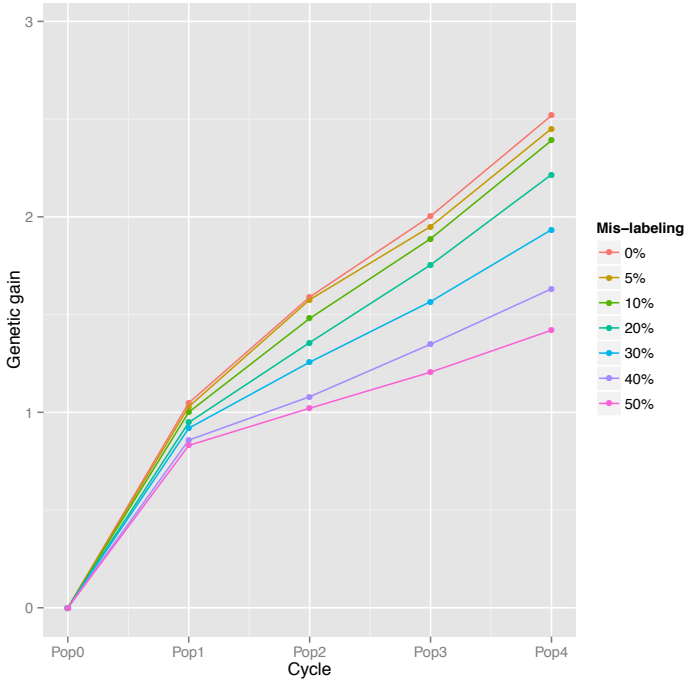


Figure 5.3. (Continued)

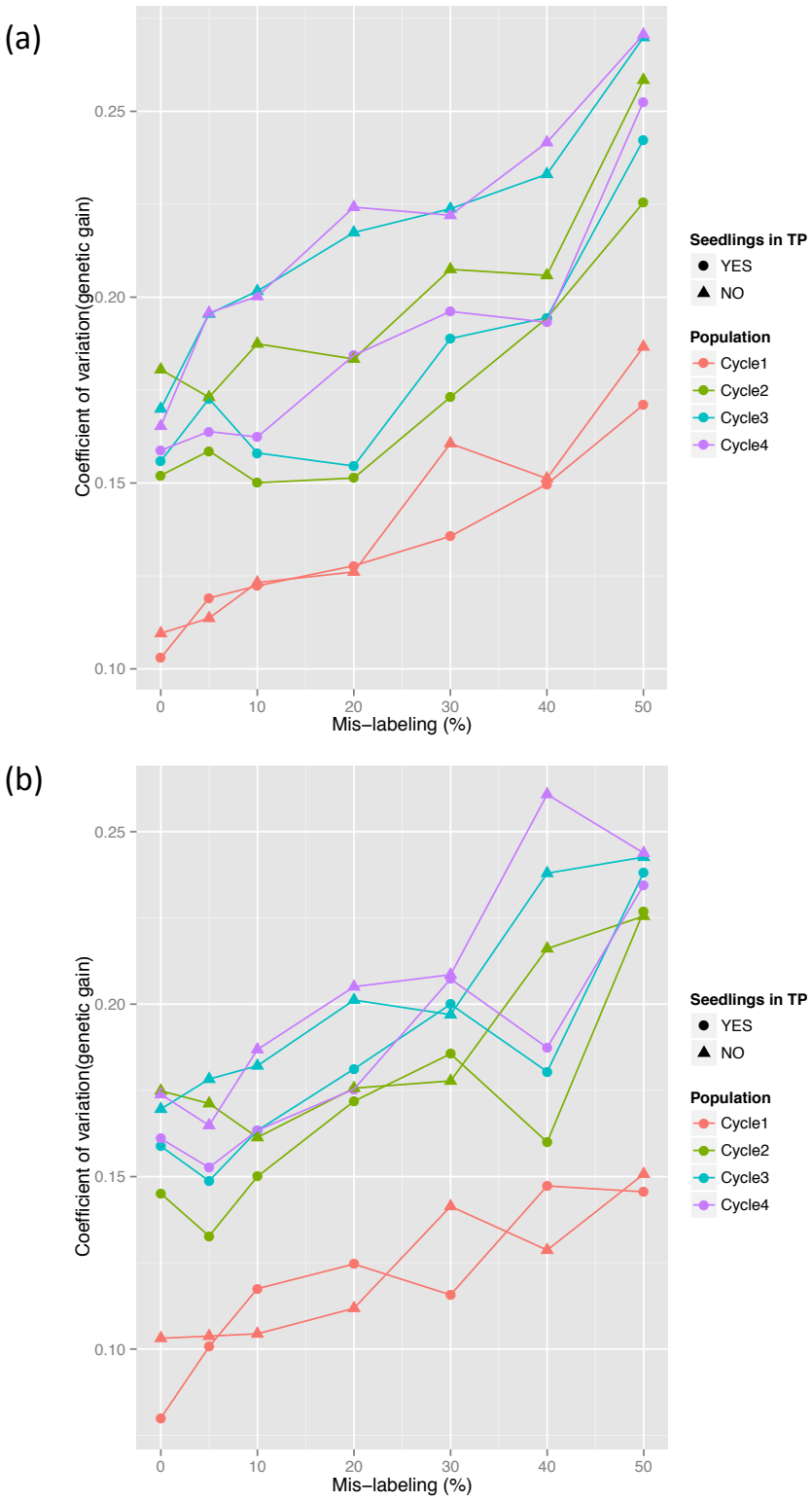


Figure 5.4. Coefficient of variation of genetic gain among 100 trials of simulation in each breeding scenario. (a) marker genotypes were switched between populations, and (b) marker genotypes were switched inside the breeding population.

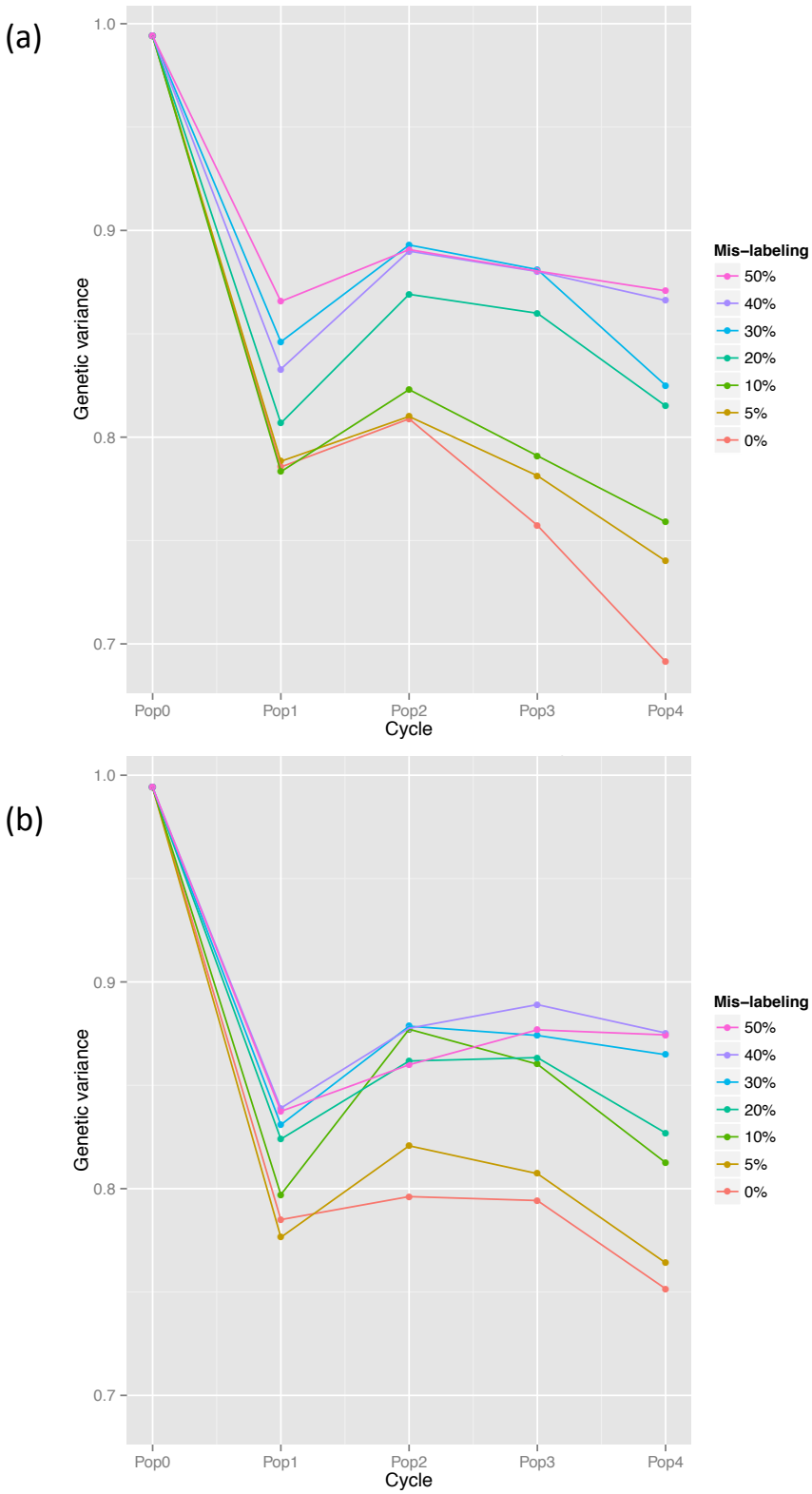


Figure 5.5. Genetic variance through four cycles of selection in the breeding scheme assuming that marker genotypes were switched between populations. (a) the training population included seedlings phenotypic data, and (b) the training population excluded seedlings phenotypic data.

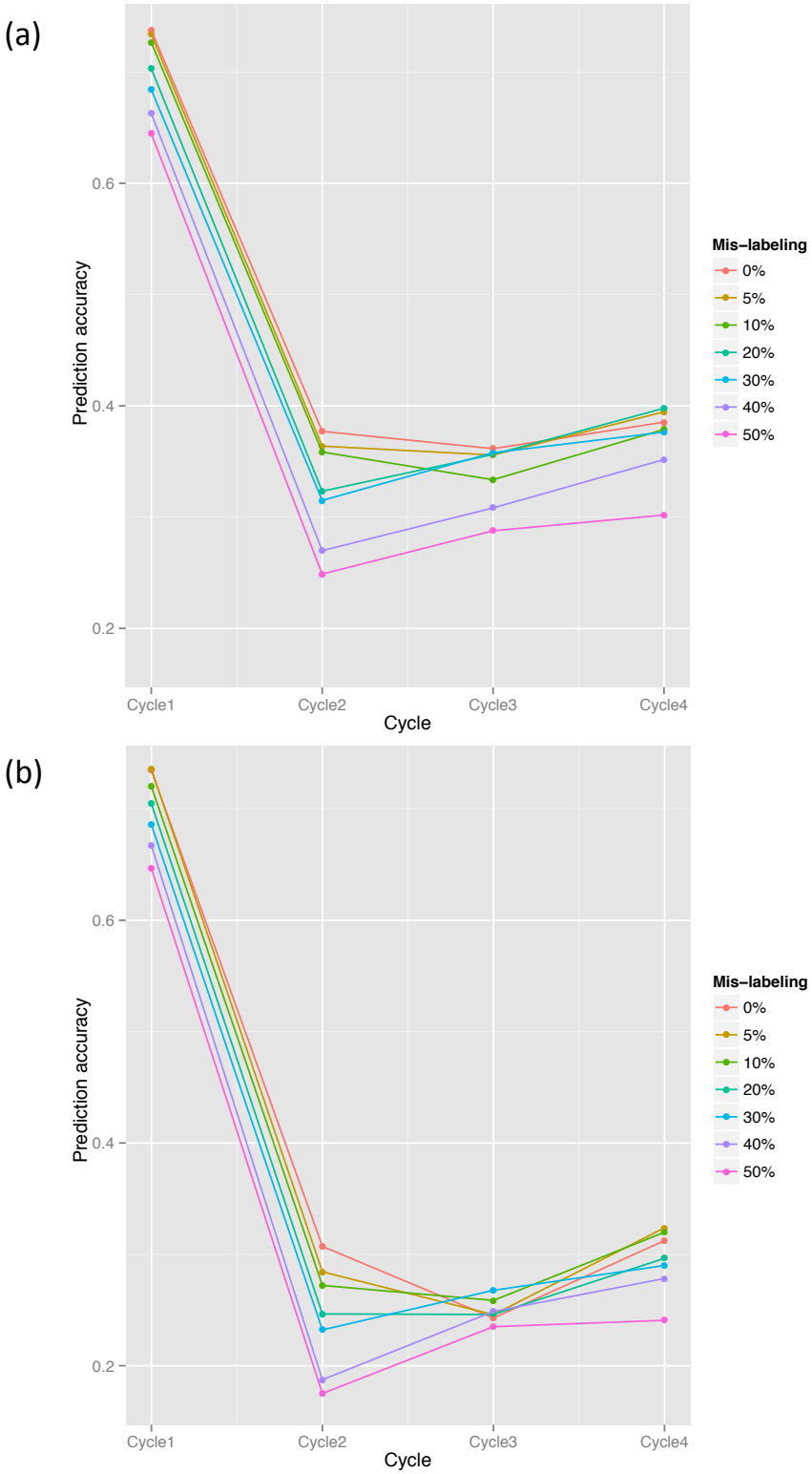


Figure 5.6. Prediction accuracy at four times of selections in the breeding scheme assuming that marker genotypes were switched between populations. (a) the training population included seedlings phenotypic data, and (b) the training population excluded seedlings phenotypic data.

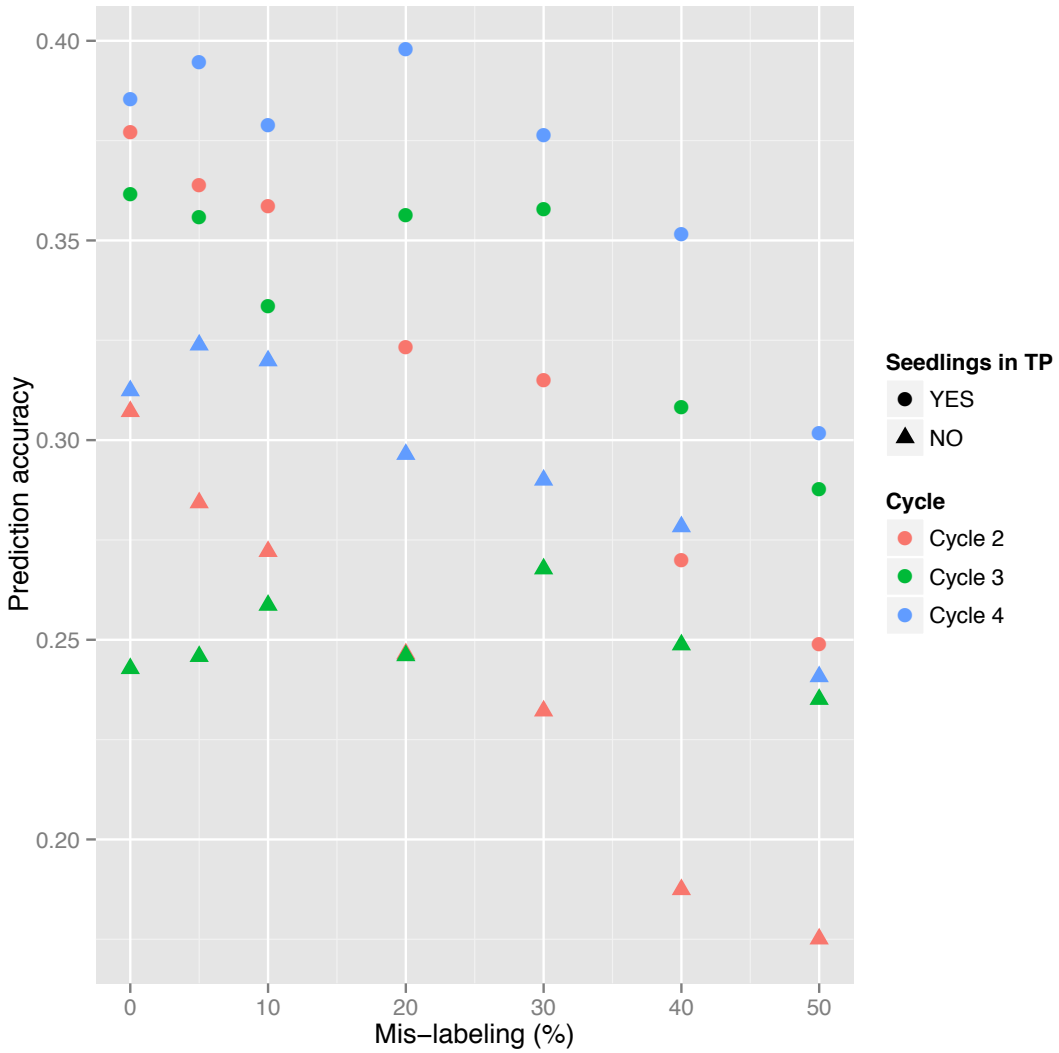


Figure 5.7. Relationship between prediction accuracy and human error from the second to the fourth selection in the breeding scheme assuming that marker genotypes were switched between populations.

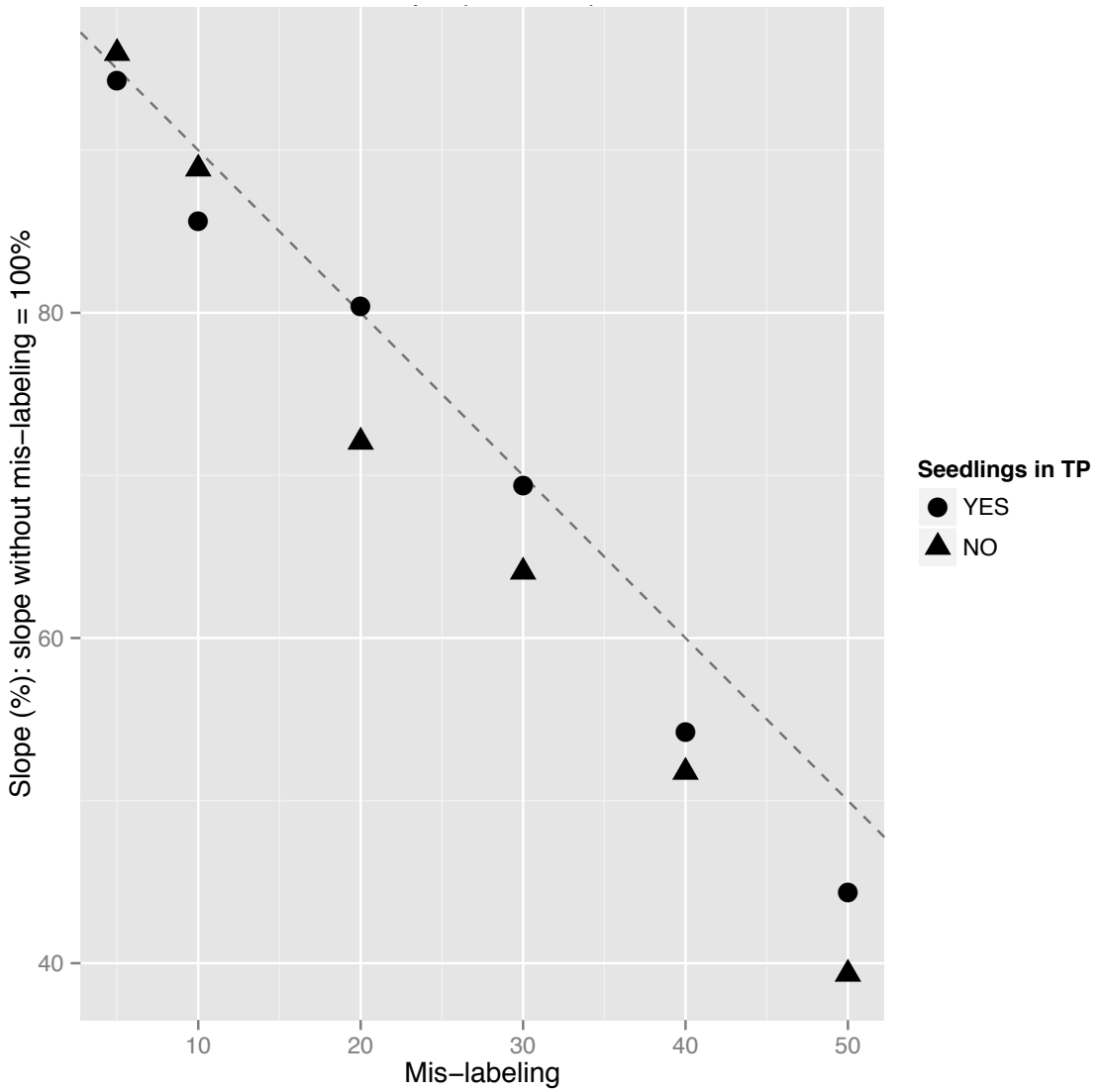


Figure 5.8. Observed response to selection from Cycle 2 to Cycle 4 in the breeding scheme assuming that marker genotypes were switched between populations. The vertical axis represents the percentage of the value of slope when the slope without human error was 100 % in each breeding scheme (i.e., the scheme using seedlings for training or not). The dashed line has the slope of –1 and the intercept of 100.

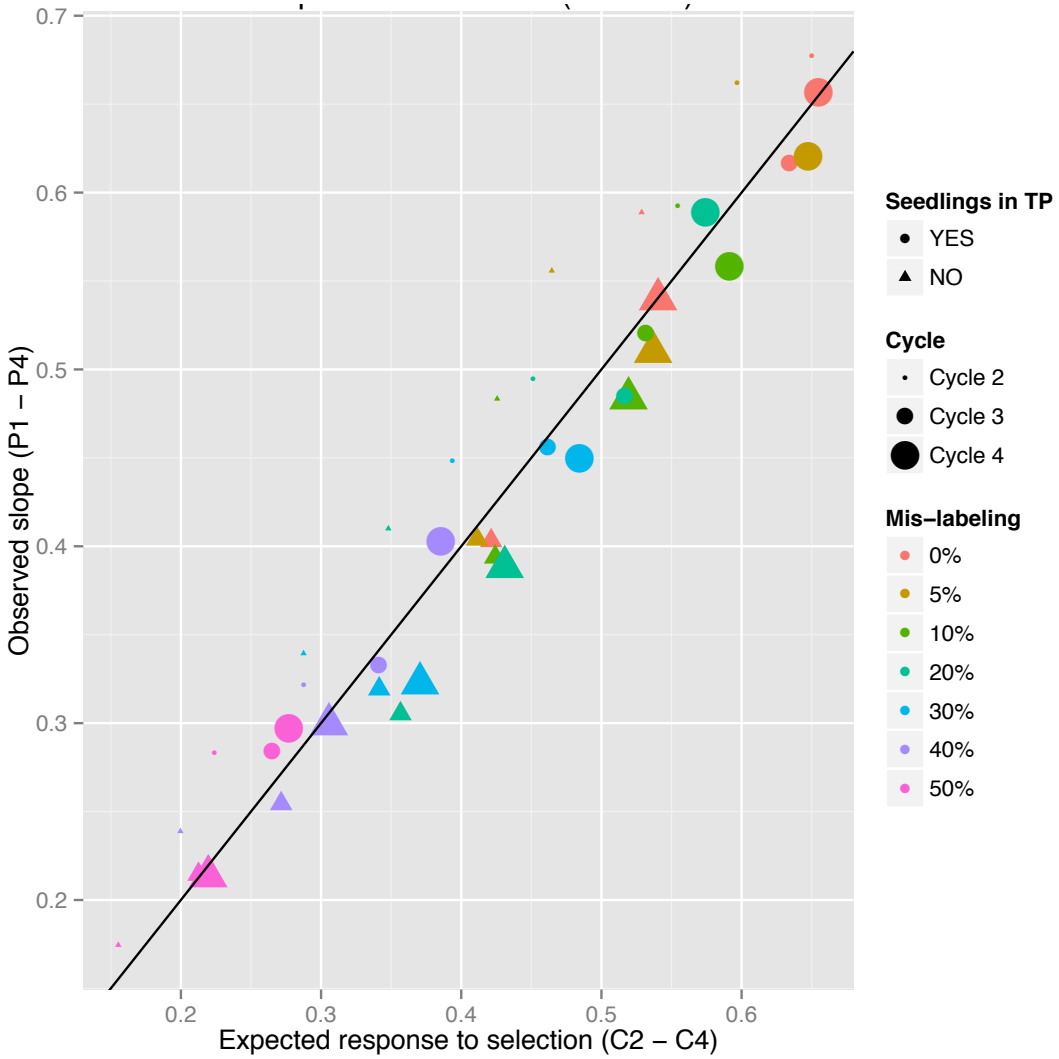


Figure 5.9. Relationship between the expected and observed response to selection from Cycle 2 to Cycle 4 in the breeding scheme assuming that marker genotypes were switched between populations. The vertical axis represents the observed response to selection, and the horizontal axis represents the expected response to selection. Observed response to selection is shown for each selection stage separately. The black line shows the slope of 1 and the intercept of 0.

5-4. Discussion

5-4-1. Breeding scheme

We compared two types of breeding schemes: collecting data on seedlings or not to build the prediction model. Simulation results show that the breeding scheme utilizing seedling data attained higher genetic gain compared among the simulations without (Fig. 5.3). This suggests that seedling data was useful to predict the genetic value of plants despite the low repeatability associated with data (i.e., the error variance was 36 times greater than the genetic variance). Thus, using a prediction model that can account for differences among trial types in the size of the error variance could work effectively our cassava breeding schemes. We verified that a prediction model able to account for differences in error variance was needed by comparing this model to a standard homogeneous variance model. The heterogeneous model indeed generated higher prediction accuracies (data not shown). In the present study, we assumed four types of phenotypic data, historical data, seedling data, CET data, and PYT data.

Note that the selection schemes presented, we assumed that seedling data was available to the breeder *before* the breeder needed to select among those seedlings. Thus, seedling data contributed to the training model prior to prediction. This data increased the prediction accuracy between 5% and 10% overall selection cycles and mis-labeling rates (Fig. 5.7). Furthermore, the variation in gain across repeated simulations was lower when seedling data was included than when it was not (Fig. 5.4). Ly et al. (2013) suggested that the training population with close relatives to the selection candidates attained high prediction accuracy in their study using real phenotypic and marker genotype data of cassava. The recommendation to update the prediction model during selection cycles, which was suggested in simulation studies (Iwata et al., 2011; Chapter 3), is also based on the idea to use the individuals close to the selection candidates as the training population. In the present simulations, seedlings were the selection candidates, thus they also work to improve the prediction accuracy even though they had large error variance. Moreover, using seedling data had a potential to maintain the level of genetic improvement when the mis-labeling happened (Fig. 5.8).

The superiority of using seedling data was shown in Fig. 5.3 in which the result was represented according to the selection cycles. Collecting seedling data, however, is not free. Thus, as for many choices breeders face, there is a tradeoff between greater gain from using seedling data versus lower cost from not using it. Breeders can choose their favorite strategy according to their required time and budget.

5-4-2. Mis-labeling

The first selection event generated greater gain than all future selection events across all simulation scenarios (Fig. 5.3). The reasons for the high prediction accuracy of this first selection event were the high genetic diversity in the breeding population (Fig. 5.5), high selection intensity and the usage of precise phenotypic data (i.e., historical phenotypic data) for training (Fig. 5.1). After the first selection step (i.e., steps from Cycle 2 to Cycle 4), the realized response to selection was lower but stable across cycles. In the present study, prediction models were updated each cycle, helping to maintain this stable gain.

To attain high genetic gain, considering the response to selection is crucial. The response to selection can be observed after selection by calculating the change from the parent generation to the offspring generation. It can be predicted at the parent generation by using information about the breeding population and selection (Falconer and Mackay, 1996). It depends on the selection intensity, prediction accuracy, and the square root of the additive genetic variance in the breeding population for GS. If mis-labeling happened, the response to selection may be reduced according to the rate of mis-labeling. Thus, the response to selection may be reduced by two factors, incomplete selection intensity and low prediction accuracy, when the mis-labeling is assumed. In the present study, we assumed two types of human mistakes, switching marker genotypes between populations or inside a population. The difference of types of mistakes resulted in the similar results (Fig. 5.3). It is because the mechanism to reduce the response to selection did not change among these mistakes. In both scenarios of human mistake, the mistake happened inside the same generations of progenies derived from the same parents. This situation caused a little difference of selection accuracy and genetic variance between these human mistakes.

Through the all selection cycles, all scenarios with six levels of mis-labeling attained a certain genetic gain (Fig. 5.3). The reason is the relationship between the genetic variance and the prediction accuracy. Among the breeding population having the same genetic variance, the prediction accuracy was depended just on the rates of mis-labeling (i.e., the accuracy at the first selection in Fig. 5.6). If the breeding population has different levels of genetic variance, however, the prediction accuracy depends not only on the mis-labeling but also on the genetic variance because a large genetic variance causes a high heritability. The higher the heritability is, the higher the prediction accuracy becomes in GS (Grattapaglia and Resende, 2011;Lorenz, 2013). The scenarios including higher rates of mis-labeling tended to maintain higher genetic variance in the breeding populations (Fig. 5.5) owing to weak genetic bottleneck caused by

incomplete selection. The prediction accuracy at the second selection (i.e., at the Cycle 2) declined with the increase of the rate of mis-labeling (Fig. 5.7). We updated prediction models at each selection, and included all available phenotypic data in the training population. The prediction model at the second selection largely depended historical phenotypic data. If the seedling data was used, it has only small effect to the prediction model because the model was adjusted by scale of environmental error variance. Therefore, the accuracy at the second selection depends mainly on the mis-labeling rate. At the third and fourth selections, however, the prediction accuracy was distributed in the same level especially when the mis-labeling was lower than 40% (Fig. 5.7). At these selection cycles, the prediction model included a number of individuals, which were in the low diversity populations. And these training data included CET and PYT data, which affected largely to the prediction models. It is suggested that the genetic improvement from Population 1 to Population 4 in the scenario with mis-labeling was realized by the trade-off between the mis-labeling and the genetic variance maintained in the breeding population. The observed response to selection showed that the genetic improvement with mis-labeling declined just following the mis-labeling rates when the mis-labeling was from 5 to 30% and seedling data was used (Fig. 5.8), suggesting that the effects of prediction accuracy and genetic variance got balanced out. When the mis-labeling was 40 or 50%, the observed response to selection was affected by the low prediction accuracy (Figs. 5.7 and 5.8).

Comparing the observed response to selection with the predicted response to selection calculated by equation [5.5], the observed response was higher than predicted response at the second selection (i.e., at the Cycle 2) in all mis-labeling rates (Fig. 5.9). The observed response moved to lower than expected one with selection cycles. The equation [5.5] was derived the idea that the individuals whose marker genotypes were swapped had prediction accuracy of 0. In the present study, however, we switched the marker genotypes, thus the phenotypic values were correct. This situation sometimes causes the non-zero accuracy in the individuals with wrong combinations of data, resulted in the high observed response at the second selection.

5-4-3. Suggestion for breeding

In the present study, we assumed cassava breeding by using genomic selection. Cassava is an allogamous species and propagated by using clones. Thus, the impact of mis-labeling is expected to be larger in cassava breeding than in other allogamous crops breeding, in which clonal propagation cannot be used, considering the continuing error in an existing genotype. The same situation will happen in breeding of autogamous plant species, in which inbred or

pure lines are used, because marker genotype data of lines tends to continue to be used. Although the breeding schemes in my simulations depended on the assumptions of cassava breeding, the results can be applied to the other plant species because the system of mis-labeling, prediction, and selection depends on more or less the same basis among plant species.

The mis-labeling caused the increase of genetic variance through selection cycles, resulted in the moderate scale of reduction of genetic gain. However, the reduction of gain was large when the rate of mis-labeling became large, suggesting that the large scale of mis-labeling would cause terrible reduction of genetic gain. At the outset of this study, I assumed that mis-labeling would cause decreases in gain through two mechanisms. First, for plants that are mis-labeled, selection occurs at random because the genotype selected does not correspond to the individual that will later be crossed. Thus, for these plants, the selection differential is zero. Second, the mis-labeled plants introduce error into the training population, which in turn will cause prediction accuracy to decrease. Thus, even for those plants that are not mis-labeled, response to selection will decrease. In fact, both plants (i.e., plants that were mis-labeled and plants that were not mis-labeled) were selected at each selection cycle in my simulations (data not shown). I did not anticipate the favorable effect of mis-labeling on genetic variance in the population. The increased genetic variance observed under mis-labeling led to sufficiently improve the accuracy that, at least for low mis-labeling rate (10% or less), accuracy was hardly affected (Figs. 5.6 and 5.7). It is suggested that the large scale of mis-labeling should be prevented, but that preventing small scale of mis-labeling is not cost effective in plant breeding. Careful statistical analysis to detect mis-labeling (e.g., detecting genotypes with extremely large prediction errors, i.e., discordance between phenotypes and marker genotypes, and excluding genotypes representing large difference between the genetic relationship matrix based on the marker data and the numerator relationship matrix based on the pedigree, i.e., discordance between marker genotypes and pedigree records) may be enough for preventing mis-labeling under these levels.

Chapter 6

Field trial of genomic selection using common buckwheat

6-1. Introduction

GS is highly expected to contribute to efficient and high-speed plant breeding (Heffner et al., 2009). The potential of GS in plant breeding has been mainly demonstrated via simulation studies. Most of empirical studies of GS in plant breeding were based on the verification of the accuracy of GS under various situations by cross validations (reviewed in Lin et al., 2014). Although the importance of field trials has been indicated, it is important to clarify the optimal design of a field trial of GS breeding to verify the efficiency of GS on ahead of implementing the field trial because field trials require much time and cost. By simulation studies in this dissertation in addition to the studies reported previously (e.g., Bernardo and Yu, 2007; Iwata et al., 2011; Jannink, 2010), it is suggested GS is efficient in plant breeding. It is required to implement field trials of GS breeding on the basis of the results from simulation studies, then to verify the efficiency and issues of GS in plant breeding through the field trials.

In general, simulation studies are performed on the basis of some simplified assumptions. For example, Wong and Bernardo (2008) assumed to use inbred lines in breeding of oil palm, which is an unrealistic situation. In most of simulation studies, non-additive genetic effects (i.e., dominance effects and epistatic effects) are not assumed (e.g., Bernardo and Yu, 2007; Iwata et al., 2011; Jannink, 2010). Also in simulation studies reported in Chapters 3, 4, and 5 of this dissertation, only additive effects are assumed as QTL effects. Because these unrealistic assumptions are not always suitable to an actual breeding in the field, it is thought that the results of field trials are not same as those of simulations. In field trials, factors that are not assumed in simulation studies will affect to the outcomes. In this case, it is important to compare the results of simulations and field trials to explain the reason of the differences

between them. By using the knowledge obtained from this comparison between simulations and field trials, an efficient way of GS breeding should be searched.

In this Chapter, a field trial was performed in common buckwheat, *Fagopyrum esculentum* Moench ($2n = 2x = 16$) to evaluate the efficiency of GS breeding in an annual allogamous plant, by comparing GS breeding with PS breeding over the two years of selection. It is relatively easy to grow common buckwheat because of its small size and the short generation time (i.e., 2 – 3 months per generation). These features are suitable for the model crop to verify the efficiency of GS breeding. The target trait was seed yield per unit area. It is not easy to improve this trait with mass selection because the trait cannot be evaluated on the single plant basis and is controlled by a number of genes. The scheme of this field trial was decided partly based on the results in simulation study in Chapter 3. Due to some restrictions on an actual field trial (e.g., time, cost, and labor), the breeding scheme could not follow the correct procedure recommended in Chapter3.

There were two main objectives in this study. The first is to evaluate the efficiency of GS via a field trial. I compared genetic gains of GS breeding with those of PS breeding to verify the advantages of GS, which has been suggested in my simulation study (Chapter 3). The second is to clarify important factors that affect the efficiency of GS breeding by comparing this field trial study with the simulation study described in Chapter 3. The breeding scheme employed in this field trial study was following a scheme simulated in Chapter 3, as mentioned above. However, there is some discordance between the empirical and simulation studies owing to the simplified assumptions of the simulation study. In Chapter 3, I simulated only additive effects (i.e., no dominant and epistatic effects) for QTL, which may not be realistic for an actual breeding population. The existence of non-additive effects may have a considerable impact on the prediction accuracy of GS model, which assumed only additive effects, and eventually on the genetic gains attained by GS breeding. As the mating system, random mating was assumed in simulation study of Chapter 3. Common buckwheat is a self-incompatible species that has heterostylous flowers, for which fertilization can be held only between flowers of different morphological types. The heterostyly is controlled by one gene, that is, *S*-locus (reviewed in Lewis and Jones, 1992). On the one hand, thrum-type flower occurs when the genotype of the *S*-locus is heterozygous (i.e., *S*/*s*); on the other hand pin-type flower occurs when the genotype is homozygous (i.e., *s*/*s*). Thus, the progenies with two types of flowers result from the mating at approximately the same frequency. Therefore, the influence from the discordance of mating system between the simulations and the field trial, i.e., random mating and outcrossing with

self-incompatibility, respectively, should be evaluated. It was necessary to verify that GS breeding could work well in reality as did in the simulation study. Discordance in the levels and range of LD in an initial breeding population may also affect the efficiency of GS through the impact on prediction accuracy. I assumed linkage equilibrium in the base population in Chapter 3 to reflect the low levels of LD in an allogamous plant with a large effective population size. In reality, however, because historical LD (LD generated in past demographic history) exists in a breeding population of common buckwheat, the LD might be higher and narrower than one assumed in the simulations. The high and narrow LD is more suitable to GS breeding than linkage equilibrium, suggesting the possibility of improvement of the prediction accuracy in the field trial. To confirm the levels of LD in the breeding population, I estimated between-marker LDs in the actual initial breeding population and compared them with genetic distance on the linkage map. In Chapter 3, the simulations underscored the importance of updating a prediction model. I verified if the importance of model updating was also suggested in the field trial study by an ex-post analysis of the accuracy of prediction models developed at different generations. Serious discordance between the simulation study and a field trial study may clarify factors that should be included in breeding simulations and that can improve the efficiency of GS in the real-world breeding.

6-2. Materials and methods

6-2-1. Linkage and QTL mapping in mapping population

92FE1-F4, a population produced by bulk crossing among 'Tempest', 'Kitawasesoba', 'Natsusoba', and 'Shinanonatsusoba', was employed. These cultivars are classified into a single agroecotype: summer type. A mapping population consisting of 178 F_1 progeny derived from a cross between P1 and P2 were developed. P1 and P2 were selected from 92FE1-F4 to make them represent large variations. Common buckwheat is a heteromorphic, self-incompatible species that has heterostylous flowers controlled by the *S*-locus (reviewed in Lewis and Jones, 1992). In the present study, P1 had thrum-type flowers and P2 had pin-type flowers, meaning that the genotype of the *S*-locus was heterozygous (i.e., *S*/*s*) in P1, and recessive homozygous (*s*/*s*) in P2. The F_1 progeny were sowed at a density of one seed per plastic pot (diameter, 24 cm; height, 24 cm) on 6th August 2012 and cultivated under natural conditions in an isolation chamber (L × W × H, 630 × 540 × 230 cm) at the University of Tsukuba (36°06′N, 140°05′E). Main stem length was measured as the trait for QTL mapping after harvesting on 9th October 2012. In this study, the main purpose of the QTL mapping was to evaluate the reliability and utility of the constructed linkage map. Therefore, I conducted the QTL mapping only for main stem length as one of the major traits relating to yield. These materials were provided by Dr. Takashi Hara and Professor Ryo Ohsawa in the University of Tsukuba. Total genomic DNA from both parental individuals and the F_1 progeny was extracted, and further was prepared from the mapping parents (P1 and P2) and from 40 plants of the 92FE1-F4 population (hereafter referred to as the 40-mix population). The procedure for the processing of row reads and the design of probes was similar to one provided in Iehisa (2014), in which the same genotyping system was applied to the linkage mapping of common wheat. All probes from P1, P2 and 40mix and control probes were synthesized in triplicate on a NimbleGen HD-2 135K × 12plex microarray (Roche Diagnostics, Madison, WI). Thereafter, genotypes of the parental plants, P1 and P2, and their 178 F_1 progeny were determined by using HD-2 135K × 12plex microarrays. DNA polymorphisms genotyped with this system were used as dominant markers. The DNA extraction and genotyping was conducted by by Ms. Mariko Ueno and Dr. Tasui Yasui in Graduate School of Agriculture, Kyoto University, Dr. Hiroyuki Enoki, Mr. Tatsuro Kimura, and Dr. Satoru Nishimura in Future Project Division, TOYOTA MOTOR CORPORATION on the basis of Enoki et al. (2012).

Linkage map of P1 and P2 were constructed by using the pseudo-testcross strategy (Grattapaglia and Sederoff, 1994). First, the deviation from the Mendelian segregation ratio was

tested for each marker by using the Chi-square test ($p < 0.01$; statistically significance). Markers segregating in a 1:1 ratio were used to construct the linkage map of P1 and P2. Markers were assigned to linkage groups by setting the recombination rate threshold at 0.3 and the threshold for the minimum number of markers at 3. Locus ordering was performed by using AntMap software (Iwata and Ninomiya, 2006) with a 50-run of an optimization process (i.e., maximization of the log-likelihood). The Kosambi mapping function (Kosambi, 1943) was used to calculate map distances. To connect linkage groups constructed in the P1 and P2 linkage map, recombination rates between markers segregating in a 3:1 ratio and markers represented in the P1 and P2 map were calculated. Among the markers segregating in a 3:1 ratio, those that had low recombination rates (<0.02) with both markers on the P1 and P2 map were used as "bridges" between the P1 and P2 map. When multiple markers on the P1 or P2 map had recombination rates <0.02 and were segregated in a 3:1 ratio, the centers of gravity of these multiple markers were connected by a bridge.

For the QTL analysis, 171 plants for which both marker data and phenotypic data (i.e., scores of main stem length [cm]) were available were used. Composite interval mapping (Zeng, 1993; Zeng, 1994) was performed by using the QTL Cartographer software ver. 1.17 (Basten *et al*. 2003). To analyze the pseudo-testcross data, I employed a model for inbred backcross design. In the analysis, I adopted the linkage phase estimated at the step of linkage map construction. A permutation test with 100 replicates was performed for each trait to estimate the empirical threshold value corresponding to the 5% significance level. The proportion of total (i.e., phenotypic) variance explained by each of detected QTL was calculated as ([residual variance under the null hypothesis] – [residual variance under the alternative hypothesis]) / [phenotypic variance].

6-2-2. Selection index

A target trait in this breeding experiment was seed yield per unit area (kg/10a). However, in mass selection, breeders cannot observe this target trait because it should be evaluated with multiple plants (i.e., plant population) grown in unit area. Because common buckwheat is an allogamous crop, there is genetic heterogeneity in the population and it is difficult to associate this target trait with genome-wide marker genotypes. In this study, I created a selection index that can be represent the yield per unit area, using the relationship between other yield related traits, which can be evaluated in each individual, and the target trait (i.e., yield per unit area).

Selection of good common buckwheat plants (i.e., genotypes) was conducted on the basis of magnitude of the selection index (i.e., larger is better).

Selection index was created based on the field trial data of 11 cultivars in 1993. The 11 cultivars are classified into a single agroecotype: summer type, which is the same agroecotype as the mapping population and breeding population. These cultivars were evaluated in a randomized block design with three replications. They were sowed at a density of 100 individuals per square meter (distance between rows, 60 cm; length of row, 3 m) on 25th August 1993. Fertilization was N:P:K = 5:20:20 and nitrogen was 4kg/10a. The yield and yield related traits were evaluated as population mean. These field trial data were provided by Professor Ryo Ohsawa.

The nine traits, seed yield, main stem length (cm), number of nodes, flowering of the first flower, number of flower clusters, number of primary branches, number of seed set in a plant, 1000 seed weight (g / 1000 seeds) and test weight (g / l), were used in the analysis. I used principal component regression (PCR) to build a multi variate regression model, in which seed yield was treated as the independent variable and other traits were treated as explanatory variables. R package "pls" (Mevik and Wehrens, 2007) was used for the analysis. To choose a number of principal components included into the regression model, I evaluated the root mean squared error of prediction (RMSEP) on the basis of leave-one-out cross-validation. To balance the number of principal components (i.e., complexity of regression model) with RMSEP (i.e., prediction accuracy), I selected the number of components that realized the smallest RMSEP in the cross-validation. The maximum number of components was set as seven. Because I used data of only 11 cultivars to create the selection index, the selection index may be inappropriate (inaccurate) to predict seed yield per unit area for the target breeding population. However, I decided to use the criterion as the breeding target and evaluated the efficiency of GS breeding by verifying whether GS could improve multiple traits simultaneously based on a selection index.

6-2-3. Genomic selection and phenotypic selection

The base population for breeding was set to a population produced by bulk crossing among 'Tempest', 'Kitawasesoba', 'Natsusoba', and 'Shinanonatsusoba', which was similar to the mapping population used in linkage and QTL mapping. This population was considered to have low and wide-ranging LD. To increase the level of LD of this base population, one cycle of random mating among 40 individuals was performed. With this cycle, the population

experienced genetic bottleneck and is expected to have increased LD. After this cycle, the population was used as the first generation of breeding population for both PS breeding and GS breeding. Two cycles of PS breeding and four cycles of GS breeding were conducted over the period of two years.

PS was conducted once per year in the regular growing season of summer type common buckwheat when phenotypic measurements were meaningful (Fig. 6.1). It was performed in August 2011 and 2012. The size of breeding population was 192, and 12 individuals were selected based on their observed values of selection index. These 12 individuals were crossed together by insect pollination of bee flies in a net to contribute the next generation.

GS was conducted twice per year (Fig. 6.1). At the selection in August (i.e., the regular growing season of summer type common buckwheat), phenotype was evaluated, and the prediction model was built (i.e., updated) with observed phenotypic values at these cycles (i.e., GS1 and GS3 in Fig. 6.1). Selection was conducted by applying the marker genotype data to the prediction model made of their own data in these cycles. The second cycle was conducted by using offseason nursing, where phenotypic evaluations were not meaningful (i.e., GS2 and GS4 in Fig. 6.1). Therefore, the prediction was conducted by using the model built at the previous cycle. Selection was done based on the predicted values of selection index. At the selection where the prediction model was built, the population size was 192, and 12 individuals were selected. On another hand, at the selection where the model was not built, 48 plants were grown and 12 plants were selected according to the predicted values. Fewer plants were grown in these cycles than when the prediction model was built to save time and effort. To conduct GS, G-BLUP was used by R package "rrBLUP" (Endelman, 2011). The prediction model was built for each trait included in the selection index. Prediction was conducted for each trait, and the predicted values of all the traits were summed up in the predicted selection index by using the weights of each trait for selection index.

GS breeding was conducted in parallel with making a selection of appropriate markers for GS. In addition, after the first selection, some individuals selected in the previous selection cycle were genotyped with the current breeding population to evaluate the reliability of marker genotype. At the first selection cycle (i.e., GS1 in Fig. 6.1), 274,303 candidate markers were genotyped. I selected 50,000 out of 274,303 markers according to their polymorphism, clarity, MAF, and linkage with other markers. GS1 was conducted on the basis of the prediction model built with the 50,000 markers. At the second selection (i.e., GS2 in Fig. 6.1), 45,000 candidate markers were genotyped, and 11,480 markers were selected according to their polymorphism,

linkage with other markers, and the degree of coincidence with marker genotype in GS1. At GS2, an adjusted prediction model was built with the 11,480 markers by using phenotype and marker genotype data of GS1. This adjusted prediction model was used at GS2. At the third and forth cycles (i.e., GS3 and GS4 in Fig. 6.1), 14,598 markers were genotyped. 6,373 and 6,225 markers were selected at GS3 and GS4, respectively, according to their polymorphism, linkage with other markers, and the degree of coincidence with marker genotype in the previous cycle.

All cultivation was conducted at the University of Tsukuba (36°06′N, 140°05′E) by Dr. Takashi Hara and Professor Ryo Ohsawa. The extraction of genomic DNA and genotyping were conducted by Ms. Mariko Ueno, Dr. Yasuo Yasui, Dr. Hiroyuki Enoki, Mr. Tatsuro Kimura, and Dr. Satoru Nishimura.

6-2-4. Linkage disequilibrium analysis

The levels of LD were evaluated in the breeding population. As a measure of LD, r^2 was calculated as equation 3.13. For the markers mapped on the linkage map, the pairs of markers within 50cM were selected in each linkage group. Marker haplotypes were unknown, and r^2 could not be calculated directly. To estimate r^2, I used the EM algorism proposed by Li et al. (2007) for the situation that genotyped markers were all dominant. The EM steps were repeated until the difference between the consecutive two estimated values attained smaller than 0.0001. To estimate the effective population size of the breeding population and the expected r^2, I used the method proposed by Weir and Hill (1986) and Hill and Weir (1988):

$$E(r^2) = \frac{10 + 4Nc}{(2 + 4Nc)(11 + 4Nc)}\left[1 + \frac{(3 + 4Nc)\{12 + 12 \times 4Nc + (4Nc)^2\}}{n(2 + 4Nc)(11 + 4Nc)}\right] \qquad [6.1]$$

where N is the effective population size, c is the recombination fraction between sites, and n represents the sample size.

6-2-5. Evaluation of breeding schemes

All generations of breeding populations that underwent PS breeding and GS breeding were evaluated in the field trial in August 2013. 48 seeds were sown from each population. Because the number of seeds of the initial population was not enough for the field trial, the seeds of base population (i.e., the population created by mixing four cultivars of summer type common buckwheat) were sown and evaluated instead. Nine traits (main stem length, number of nodes, flowering of the first flower, number of flower clusters, number of primary branches, number of seed set in a plant, 1000 seed weight, test weight, and number of secondary branches) were

evaluated at the harvest time. The field trial was conducted at the University of Tsukuba (36°06′N, 140°05′E) by Dr. Takashi Hara and Professor Ryo Ohsawa.

For the generations of GS1 and GS3, I performed leave-one-out cross-validation to calculate prediction accuracy of GS prediction model. The accuracy was measured with the Pearson's correlation coefficient between predicted values on one hand and expected values, which were obtained by fitting a prediction model to marker genotype data in a training data set (i.e., dataset used for model building) as well as observed values on one hand. Because we measured a single plant to get observed values, the values were affected by large environmental variation. Thus, I employed the correlation between predicted and expected values because the expected values were less affected by environmental variation than observed values. I also evaluated the levels and directions of changes in each trait and selection index throuout the breeding process. To compare generations derived from GS and PS breeding, Tukey test and Mann-Whitney U test with Bonferroni correlation for traits that had equal variances and unequal variances, respectively. The equality of variance was examined by Bartlett's test at the 10% significance level. The level was chosen to make type II error (i.e., error that fails to detect traits with unequal variance) smaller.

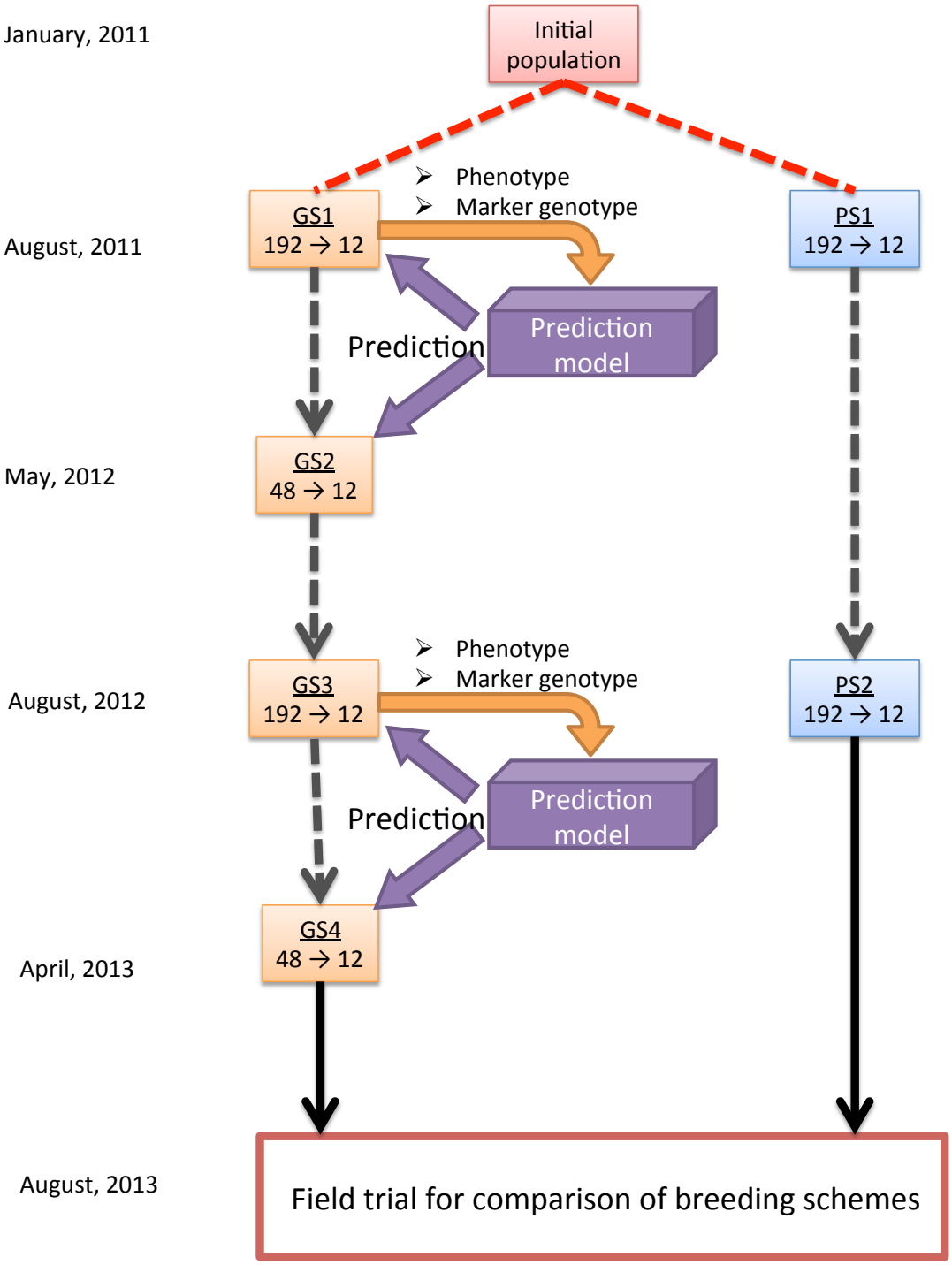


Figure 6.1. Breeding schemes for genomic selection (GS) and phenotypic selection (PS) in common buckwheat.

6-3. Results

6-3-1. Linkage and QTL mapping

DNA microarray genotyping was performed for 44,836 markers. From those markers, I used 16,841 markers that were heterozygous in one or both parents for further analysis. In the F_1 population of 178 plants, the numbers of markers segregating in 1:1 and 3:1 were 14,442 (P1: 6,875 markers; P2: 7,567 markers) and 2,399, respectively. A segregation ratio of 1:1 is expected when one parent has a heterozygous genotype and the other has a recessive homozygous genotype. A segregation ratio of 3:1 is expected when both parents have a heterozygous genotype. Of the markers with segregation ratio of 1:1 and 3:1, 9,112 (P1: 4,325 markers; P2: 4,787 markers) and 1,701 markers, respectively, showed a clear distinction between the two genotypes. Of all 10,813 markers, 1,339 markers contained one or more missing values. Of these, 824 contained missing values for less than four plants, and 106 contained missing values for more than 17 (10%) plants. Markers derived from an identical contig with the same segregation type (i.e., P1 or P2 markers with a 1:1 segregation ratio or bridging markers with a 3:1 segregation ratio) were expected to reflect genotypes at the location of the contig because of the close linkage between the markers. I therefore collected markers from a single contig to make consensus genotypes. The missing genotypes of the collected markers were imputed on the basis of the consensus genotypes. In a few cases, two or three patterns of marker genotypes were observed in one contig. In those cases, I treated the groups of markers having different genotypes as separate contigs (sep-contigs). When sep-contigs derived from a single contig belonged to different segregation types, I excluded them from the subsequent analysis. Sep-contigs from a single contig showing a low level of consensus (i.e., genotypes were discordant in more than 30% of plants) were also excluded. I used flower morphology (i.e., pin or thrum) controlled by the *S*-locus as the phenotypic marker. The marker was located on the P1 map because P1 was heterozygous (i.e., *S/s*) at the locus. In total, I had 1,455, 1,631, and 869 contigs for P1, P2, and bridging markers, respectively. Next, to analyze the pseudo-testcross data, I inverted the genotype data by duplicating and converting all of the contigs (i.e., homozygous genotypes were converted into heterozygous genotypes and vice versa). I grouped contigs that had an identical genotype into a single marker group and then separated those marker groups into three marker types—P1 markers, P2 markers, and bridging markers. Marker groups were grouped into two and four clusters for P1 and P2 type, respectively, via a single-linkage cluster analysis based on Manhattan distances among marker groups. The distances were calculated based on the genotypes of marker groups, which were

scored with 0 and 1. I used single-linkage clustering because the method is similar to the one used in the linkage grouping. In total, I attained 346, 410, and 360 marker groups, respectively, for P1, P2, and bridging markers. The genotypes of these marker groups were used to construct the linkage map.

By using a pseudo-testcross strategy, I constructed linkage map for P1 and P2, and connected them with bridging markers (Fig. 6.2). I mapped 346 loci on the P1 map, and 410 loci on the P2 map (Table 6.2). I used 283 groups segregating 3:1 (bridging markers), which represented small recombination rates (i.e., <0.02) with loci in both the P1 and P2 map, to combine the loci in the P1 and P2 linkage map. I used markers segregating 3:1 only for the P1 and P2 map bridging, not for mapping on the linkage map. Because the precision of the estimation of recombination rates between markers segregating 3:1 and between markers segregating 1:1 and 3:1 was low (Ritter *et al*., 1990), the inclusion of 3:1 markers in the linkage map was thought to affect the estimation of the linkage map positions of 1:1 markers. Thus, as shown in Fig. 6.2, I used bridging markers only to connect the linkage groups between the P1 and P2 linkage map. The phenotypic marker, flower morphology, was located on linkage group P1_3 ("*S*" in Fig. 6.2). After connecting the P1 and P2 linkage map, the number of linkage groups converged to eight, which is the basic chromosome number of common buckwheat. The eighth linkage group was divided into two groups of short length in the P1 map. The P1 and P2 linkage map covered 773.8 and 800.4 cM, and contained 1,455 and 1,631 contigs, consisting of 4,227 and 4,657 markers, respectively (Table 6.1). The means of the intervals between adjacent positions were 2.30 and 1.99 cM (Table 6.1) and the medians were 1.68 and 1.15 cM in the P1 and P2 linkage map, respectively (Fig. 6.3). Most (90%) adjacent positions had intervals shorter than 5.07 cM (Fig. 6.3). On the linkage map, one position (i.e., a single marker group) harbored a number of contigs, and one contig consisted of a number of markers. Figure 6.4 shows the number of contigs per loci (a) and the number of markers per contig (b). Among the 756 loci on the map, 555 loci consisted of more than one contig, and 492 loci consisted of less than 10 contigs. Among the 3,086 contigs, 1,140 contigs consisted of more than one marker, and among those, 1,036 contigs consisted of less than 15 markers and 13 contigs consisted of more than 35 markers.

I performed QTL analysis of main stem length to confirm the application of the linkage map constructed above. Phenotypic values for main stem length observed in the mapping population had a unimodal, continuous distribution (Fig. 6.5), suggesting that main stem length is controlled by multiple QTL and is influenced by environmental effects. For main stem length,

significant QTL were detected at map positions 9.3 cM on the P1-1 group, 49.0 cM on the P1-2 group, 9.0 cM on the P1-5 group, and 16.9 cM on the P2-4 group (Fig. 6.6 and Table 6.2). The four QTL accounted for 5.64% to 8.51% of the phenotypic variance observed in main stem length (Table 6.2).

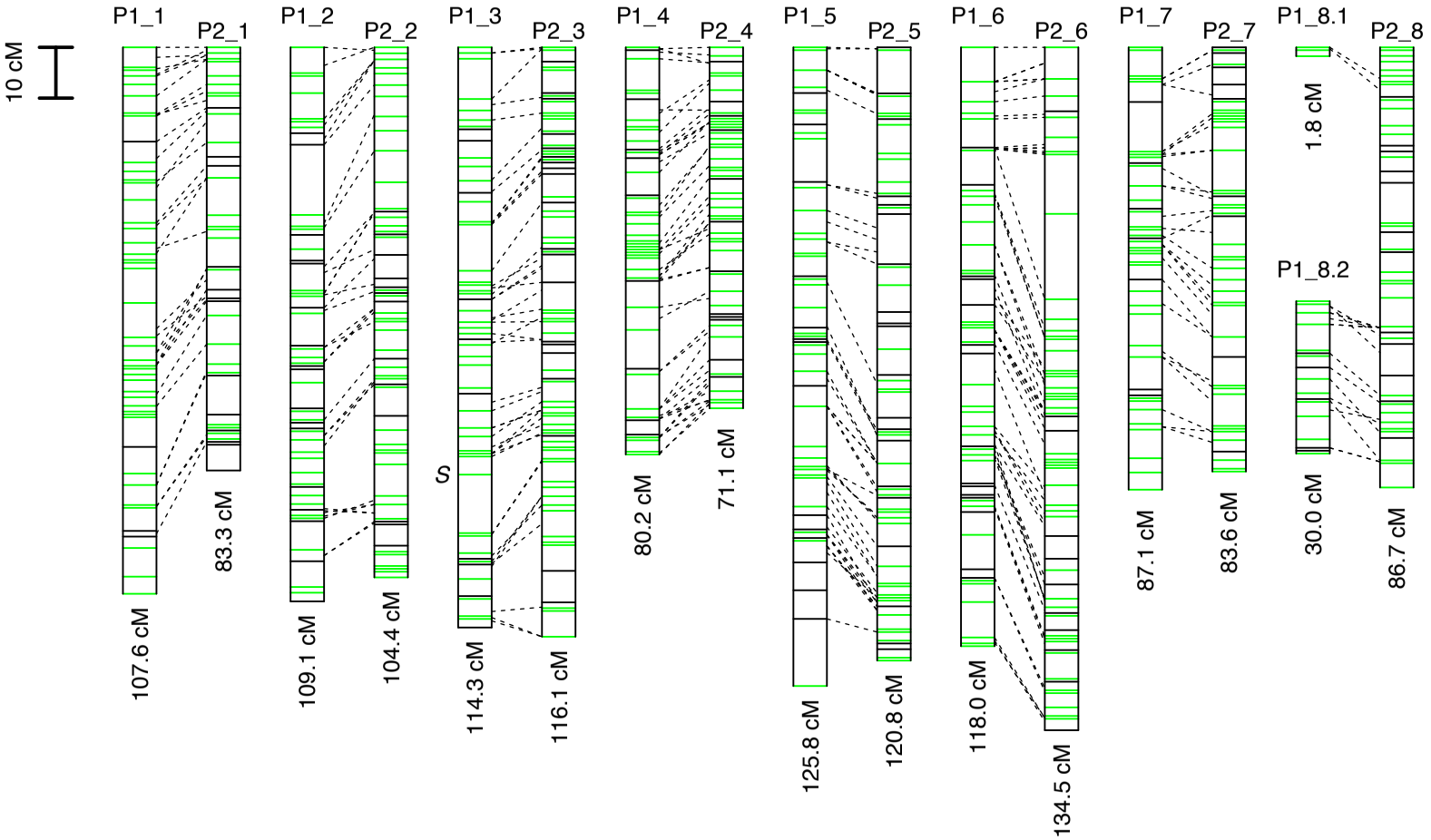


Figure 6.2. Linkage map for common buckwheat. Dashed lines represent “bridge” markers that connect marker groups between the linkage maps of P1 and P2. Phenotypic marker position (i.e., *S*-locus) is indicated by “*S*”. The loci genotyped in breeding population is colored in green.

Table 6.1. Summary information for the linkage map of common buckwheat.

Linkage group	No. of loci	Genetic length (cM)	Average interval of adjacent loci (cM)	No. of contigs	No. of markers
P1_1	45	107.63	2.45	198	674
P1_2	51	109.12	2.18	211	549
P1_3	51	114.27	2.29	223	691
P1_4	42	80.20	1.96	193	511
P1_5	47	125.79	2.73	203	626
P1_6	49	117.95	2.46	196	560
P1_7	41	87.12	2.18	161	453
P1_8.1	17	30.01	1.88	60	145
P1_8.2	3	1.75	0.88	10	18
P2_1	37	83.34	2.32	119	235
P2_2	54	104.40	1.97	206	621
P2_3	70	116.08	1.68	273	886
P2_4	49	71.09	1.48	184	498
P2_5	57	120.78	2.16	217	598
P2_6	57	134.50	2.40	224	695
P2_7	43	83.55	1.99	189	544
P2_8	43	86.69	2.06	219	580
P1	346	773.84	2.30	1,455	4,227
P2	410	800.43	1.99	1,631	4,657
All	756	1,574.27	2.13	3,086	8,884

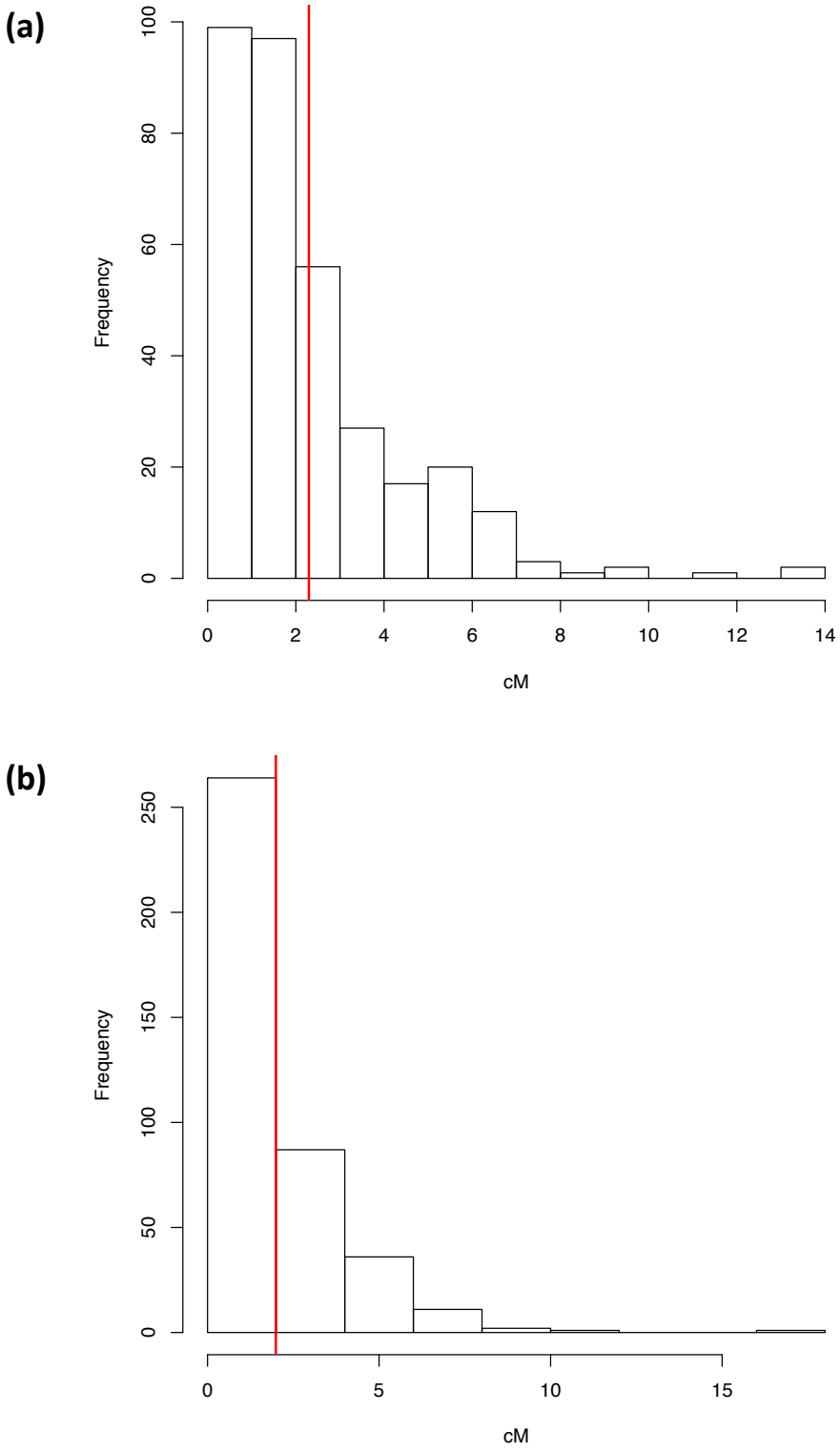


Figure 6.3. Distances between adjust loci on the linkage map of P1 (a) and P2 (b). Red lines represent the mean values of distances in each of maps.

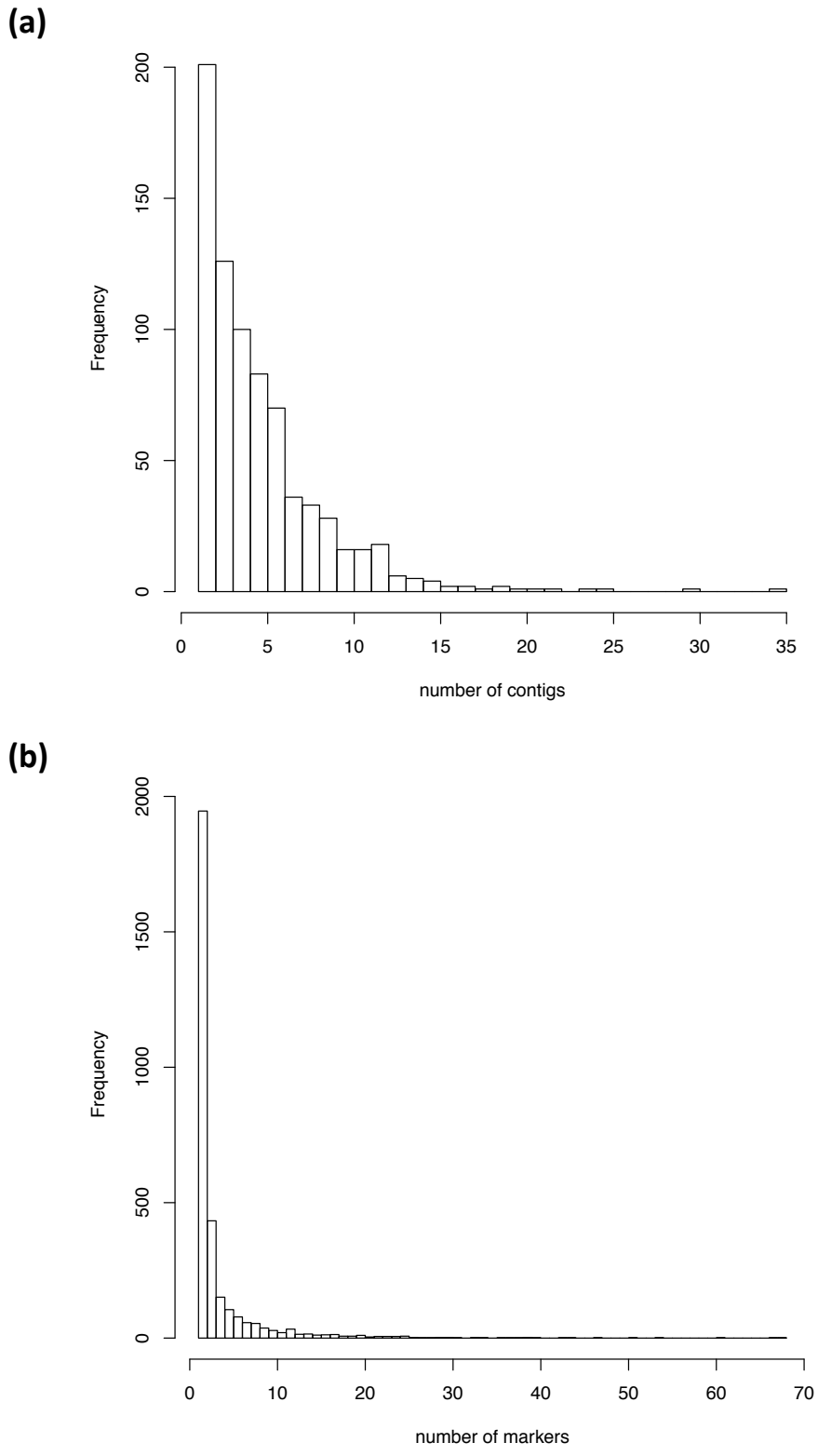


Figure 6.4. Distribution of the number of contigs per map position (a) and the number of markers per contig (b).

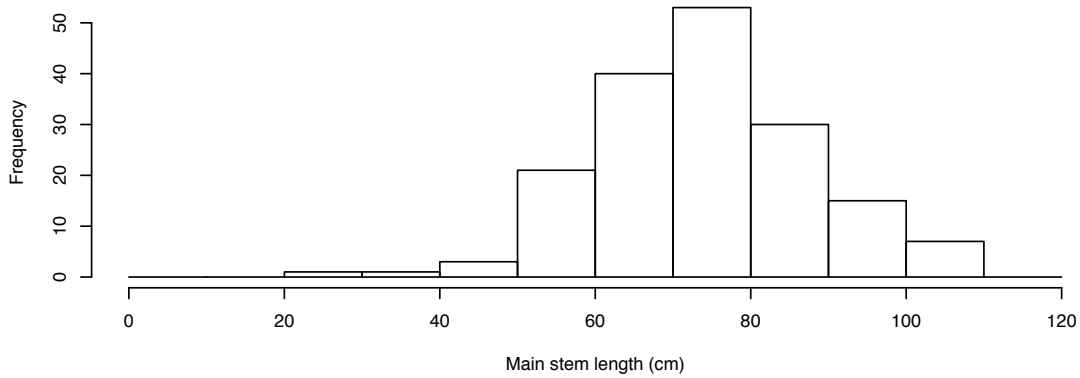


Figure 6.5. Distribution of phenotypic value for main stem length.

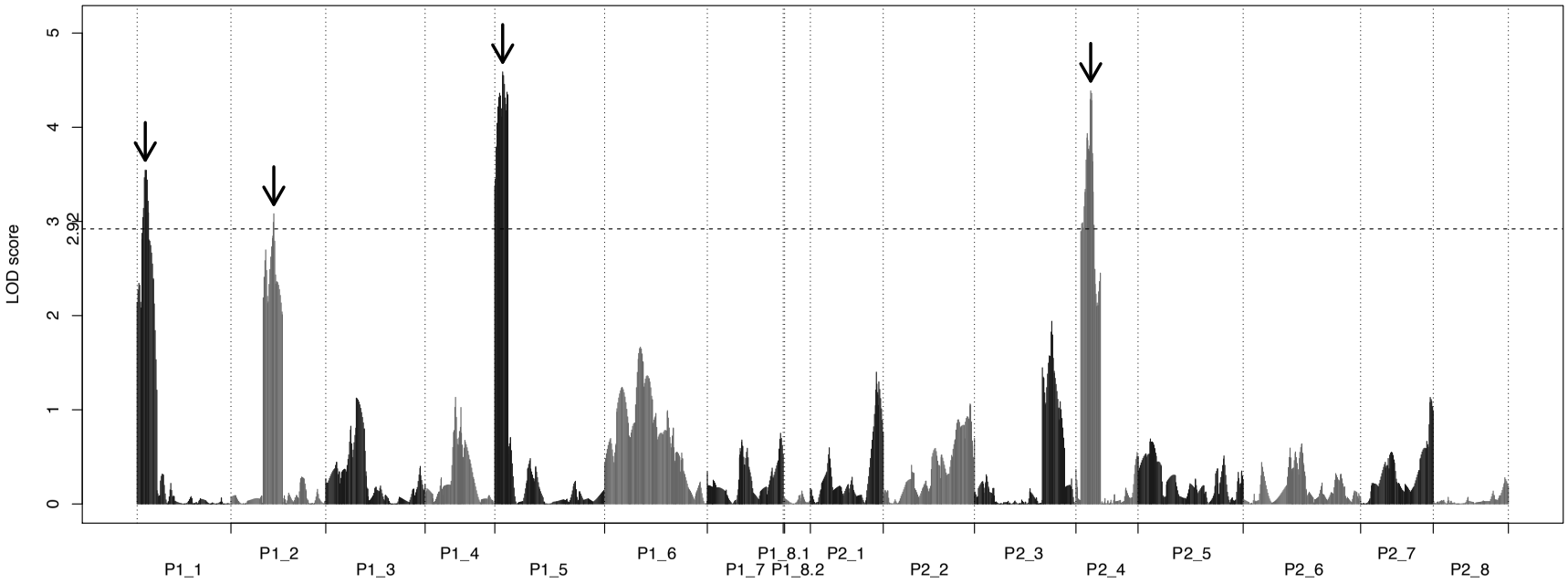


Figure 6.6. Logarithm of the odds (LOD) score profile of the quantitative trait loci analysis of main stem length. Map positions are shown on the *x*-axis. The horizontal dashed line shows the 5% threshold of LOD score (2.92) obtained from a permutation test. Arrows show the detected peaks of LOD scores.

Table 6.2. Positions, logarithm of the odds (LOD) scores, additive effects and explained phenotypic variations of the quantitative trait loci detected for main stem length.

Linkage group	Position (cM)	LOD	Additive effect	Explained phenotypic variance (%)
P1-1	9.3	3.55	−7.48	6.62
P1-2	49.0	3.08	−6.85	5.64
P1-5	9.0	4.59	−8.45	8.51
P2-4	16.9	4.39	8.29	8.18

6-3-2. Genomic selection and phenotypic selection

Green lines on the linkage map of common buckwheat in Fig. 6.2 represent the loci that were also genotyped in the breeding population at GS1 (i.e., in the initial population for PS breeding and GS breeding). In total, 1,484 markers on the linkage map were genotyped in GS breeding, and the markers were groups in 565 loci (P1: 265 loci; P2: 300 loci) on the linkage map. I plotted the degree of LD, r^2, against map distance for all markers combinations that co-located within 50cM (Fig. 6.7). In the initial population, the levels of LD between most pairs of markers were low over all range of distance while some pairs of close markers showed high levels of LD. The expected effective population size, which was calculated from the estimated r^2 in all the linkage groups, was 340.7, suggesting the large effective population size of the initial population.

I had scheduled to use all eight traits that were evaluated in both the breeding population and the agronomic trait evaluation data (i.e., main stem length, number of nodes, flowering of the first flower, number of flower clusters, number of primary branches, number of seed set in a plant, 1000 seed weight, and test weight) to create a selection index. However, at the first generation (i.e., GS1 and PS1), the seed shuttering rate was high, and phenotypic values of the number of seed set in a plant was not credible. Thus, I used seven traits, except the number of seed set, to use the selection index in PS breeding and GS breeding. As weights for the seven traits, I used the regression coefficients of PCR based on the first two principal components, which represented the smallest RMSEP (i.e., 13.3 by leave-one-out cross-validation among 11 varieties). The second-smallest RMSEP was 15.0 when I used the first three principal components. Table 6.3 shows the regression coefficients for the seven traits in the selection index. When these coefficients were used, the correlation coefficient between the observed yield and the predicted selection index was 0.73 in leave-one-out cross-validation in the 11 cultivars used in the agronomic trait evaluation data. Table 6.4 shows the correlation coefficients among seed yield and the seven traits used in the selection index for the agronomic trait evaluation data (above the diagonal) and the initial breeding population (below the diagonal). The traits representing high correlations with seed yield tended to obtain large coefficients in the selection index. The similar relationship among seven traits in both the agronomic trait evaluation data and the initial breeding population encouraged me to use the created selection index in the field trial. In the later breeding procedures, I used the coefficients shown in Table 6.3 to calculate the selection index.

Figure 6.8 shows the relationship between observed and expected (fitted) values of selection index at GS1 (a) and GS3 (b). At GS1, the correlation coefficient between observed and expected values was 0.92. On another hand, at GS3, the correlation coefficient was 0.71 and the heavy shrinkage of expected values to the mean value was shown.

Table 6.5 shows the prediction accuracy of the prediction model built at GS1 and GS3 in leave-one-out cross-validation. At GS1, the lowest prediction accuracy was 0.49 (flowering of the first flower) and the highest accuracy was 0.67 (test weight) among traits consisting of selection index. The prediction accuracy of selection index at GS1 was 0.71. At GS3, although the highest accuracy was 0.77 in flowering of the first flower, the lowest accuracy was –0.49 in test weight, suggesting that GS3 prediction model could not predict test weight at all. However, because there is no large variation in the expected values of test weight, it did not affect the rank of the expected selection index at GS3. Thus, I included test weight in the selection index at GS3. The prediction accuracy of selection index at GS3 was 0.67.

To evaluate whether a prediction model could keep accuracy in later generations, I conducted two types of analysis. First, I compared the prediction accuracies when the prediction model built at GS1 was used to predict genetic values at GS1 (i.e., the same way conducted in two years field trial) with when the prediction model built at GS1 was used to predict genetic values at GS3 (i.e., the way assuming that the prediction model was not updated at GS3). Blue bars in Fig, 6.9 shows the prediction accuracy at GS1, and red bars shows the prediction accuracy at GS3 in seven traits consisting of the selection index when the prediction model built at GS1 was used. The prediction accuracy at GS3 was lower than that at GS1 in all the traits. When the model built at GS1 was used to predict plants at GS3, the lowest accuracy was –0.07 in flowering of the first flower, and the highest accuracy was 0.54 in number of primary branches. Second, I compared the selected plants at GS3 based on the predicted values of the selection index when the prediction model built at GS3 was used (i.e., the same way conducted in two years field trial) with when the prediction model built at GS1 was used (i.e., the way assuming that the prediction model was not updated at GS3). Figure 6.10 represents the relationship the expected (i.e., values calculated using the updated model built at GS3) and predicted values (i.e., values calculated using the non-updated model built at GS1) of selection candidates at GS3. Only one plant (i.e., genotype) was selected by both prediction models, while the other 11 selected plants were different in the two situations, suggesting that the different genotypes would have contributed to the next generation if the prediction model had not been updated.

Figure 6.11 shows the change of population mean in nine traits through selection cycles of PS breeding and GS breeding, which was obtained from the field testing for comparison of populations. Note that the base population was not identical to the initial breeding population. For PS breeding, number of nodes showed the significant increase from the base population to the population that experienced two cycles of PS (p-value < 0.05 in pairwise Wilcoxon test). For GS breeding, significant increase through two years selection (i.e., four times of GS) was shown in main stem length (p-value < 0.01 in Tukey test), number of nodes (p-value < 0.05 in pairwise Wilcoxon test), flowering of the first flower (p-value < 0.01 in pairwise Wilcoxon test), number of flower clusters (p-value < 0.05 in pairwise Wilcoxon test), and number of seed set in a plant (p-value < 0.05 in Tukey test). In main stem length and number of flower clusters, GS breeding attained higher values than PS breeding significantly (p-value < 0.05 in Tukey test and pairwise Wilcoxon test, respectively).

Figure 6.12 represents the change of population mean of the selection index through two years of PS breeding and GS breeding. The population that experienced two years GS breeding attained 14.9% higher value than the base population (p-value < 0.01 in Tukey test). On the other hand, the population after PS breeding did not show the significant differences from any other populations while it attained 3.6% higher value than the base population.

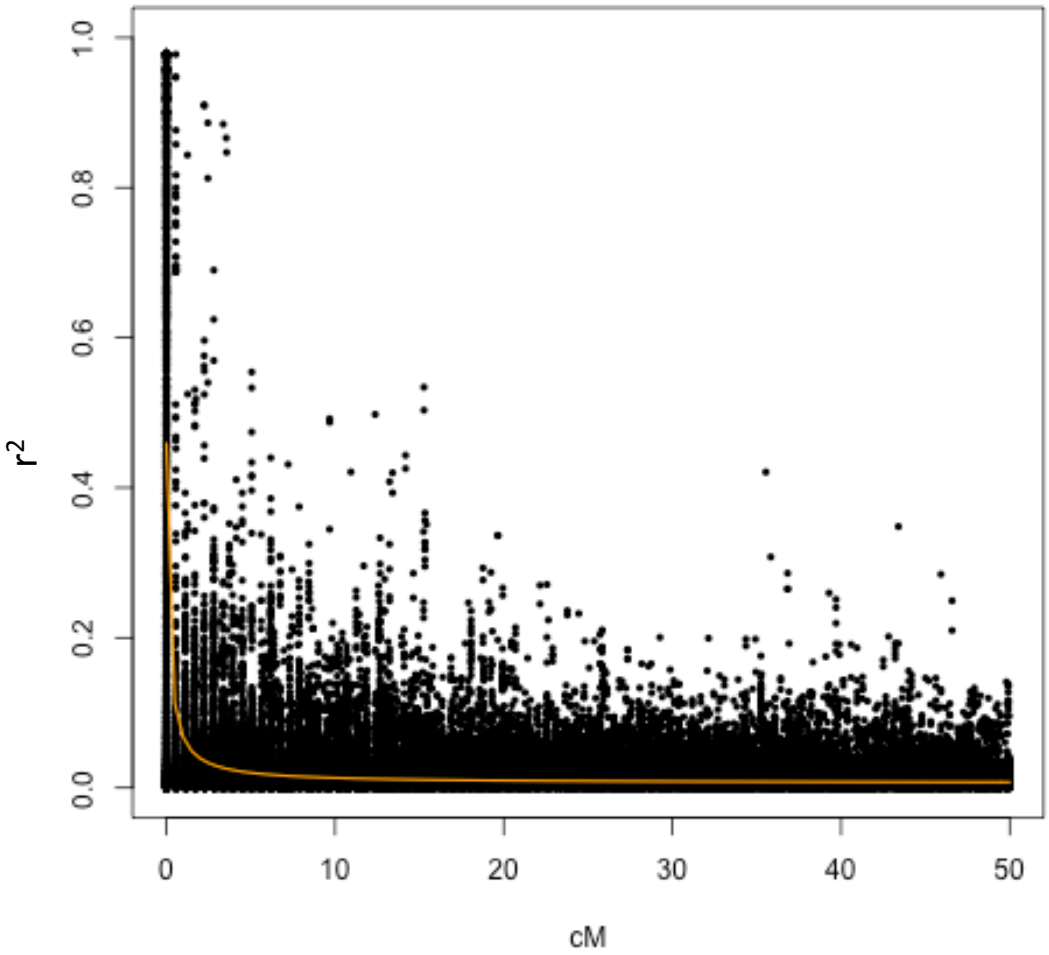


Figure 6.7. Linkage disequilibrium (r^2) in breeding population. The horizontal axis is the genetic distance between two candidate markers. The vertical axis is r^2. Orange line represents the expected r^2.

Table 6.3. The coefficients for calculation of the selection index.

Trait	Coefficient
main stem length	+0.550
number of nodes	+0.053
flowering of the first flower	+0.011
number of flower clusters	+0.728
number of primary branches	+0.015
1000 seed weight	-0.001
test weight	+0.306

Table 6.4. The correlation coefficients among seven traits and seed yield for the agronomic trait evaluation data (above the diagonal) and the initial breeding population (below the diagonal).

	Seed yield	Main stem length	Number of nodes	Flowering of the first flower	Number of flower clusters	Number of primary branches	1000 seed weight	Test weight
Seed yield		**0.74**	**0.85**	**0.05**	**0.66**	**0.52**	**−0.22**	**0.47**
Main stem length	—		**0.85**	**0.16**	**0.53**	**0.41**	**0.13**	**0.40**
Number of nodes	—	**0.56**		**0.12**	**0.75**	**0.64**	**−0.22**	**0.27**
Flowering of the first flower	—	**0.40**	**0.27**		**0.18**	**0.23**	**0.54**	**0.07**
Number of flower clusters	—	**0.36**	**0.46**	**0.26**		**0.73**	**−0.13**	**0.03**
Number of primary branches	—	**0.10**	**0.16**	**0.09**	**0.40**		**−0.37**	**−0.43**
1000 seed weight	—	**0.16**	**0.03**	**−0.04**	**0.04**	**−0.02**		**0.22**
Test weight	—	**0.13**	**0.26**	**0.09**	**0.31**	**0.19**	**0.24**	

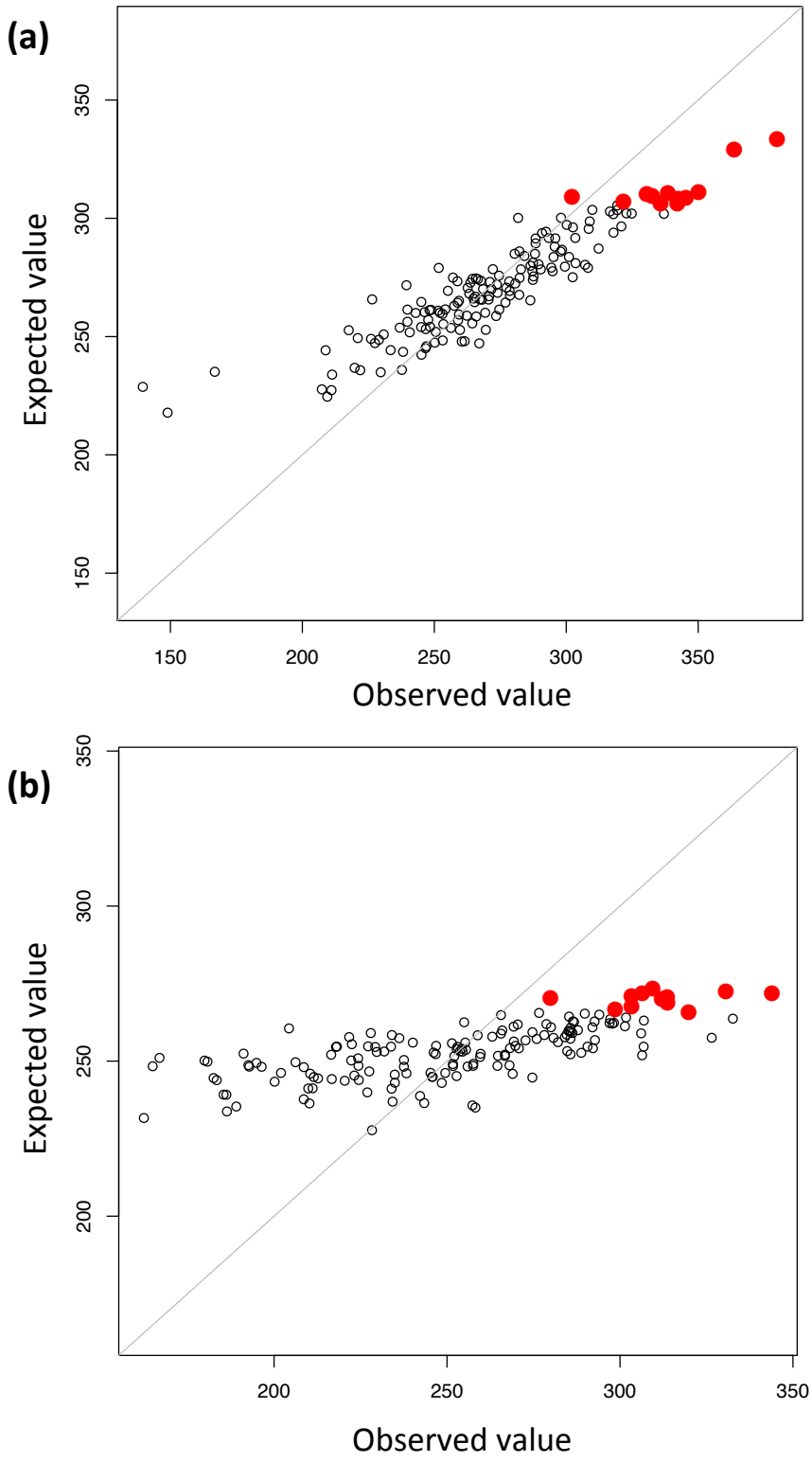


Figure 6.8. Relationship between observed and expected values of selection index at GS1 (a) and GS3 (b). Red points represent the individuals that were selected at the selection.

Table 6.5. Prediction accuracy at GS1 and GS3. The accuracy was calculated as Pearson's correlation coefficient between fitted values and predicted values in leave-one-out cross-validation.

	GS1	GS3
Main stem length	0.66	0.76
number of nodes	0.68	0.75
Flowering of the first flower	0.49	0.77
Number of flower clusters	0.58	0.73
Number of primary branches	0.65	0.73
1000 seed weight	0.61	0.74
Test weight	0.67	−0.49
Selection index	**0.71**	**0.67**

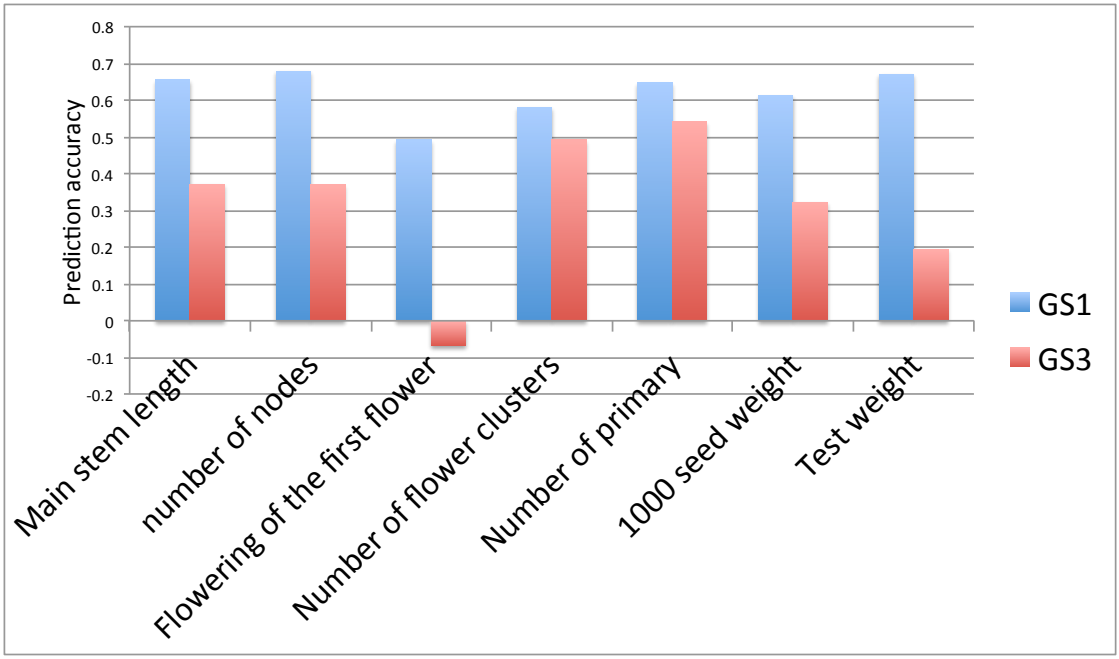


Figure 6.9. Prediction accuracy at GS1 and GS3 when the prediction model built at GS1 was used. The accuracy at GS1 was obtained by leave-one-out cross-validation.

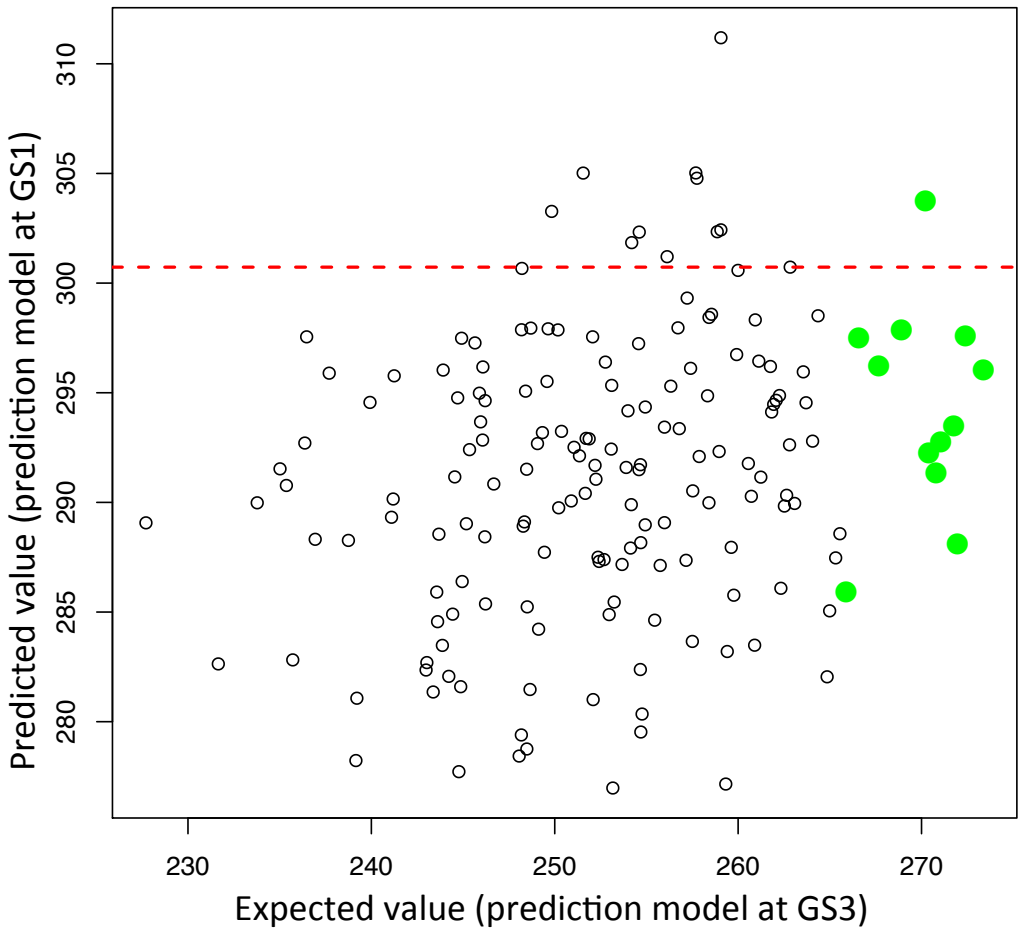


Figure 6.10. Expected values and predicted values at GS3. Expected values were calculated by using the prediction model built at GS3. Predicted values were calculated by using the model built at GS1. Greed points show the 12 individuals that were selected at GS3 in actual. Red line represents the value that should be the threshold for selection if the model built at GS1 was used.

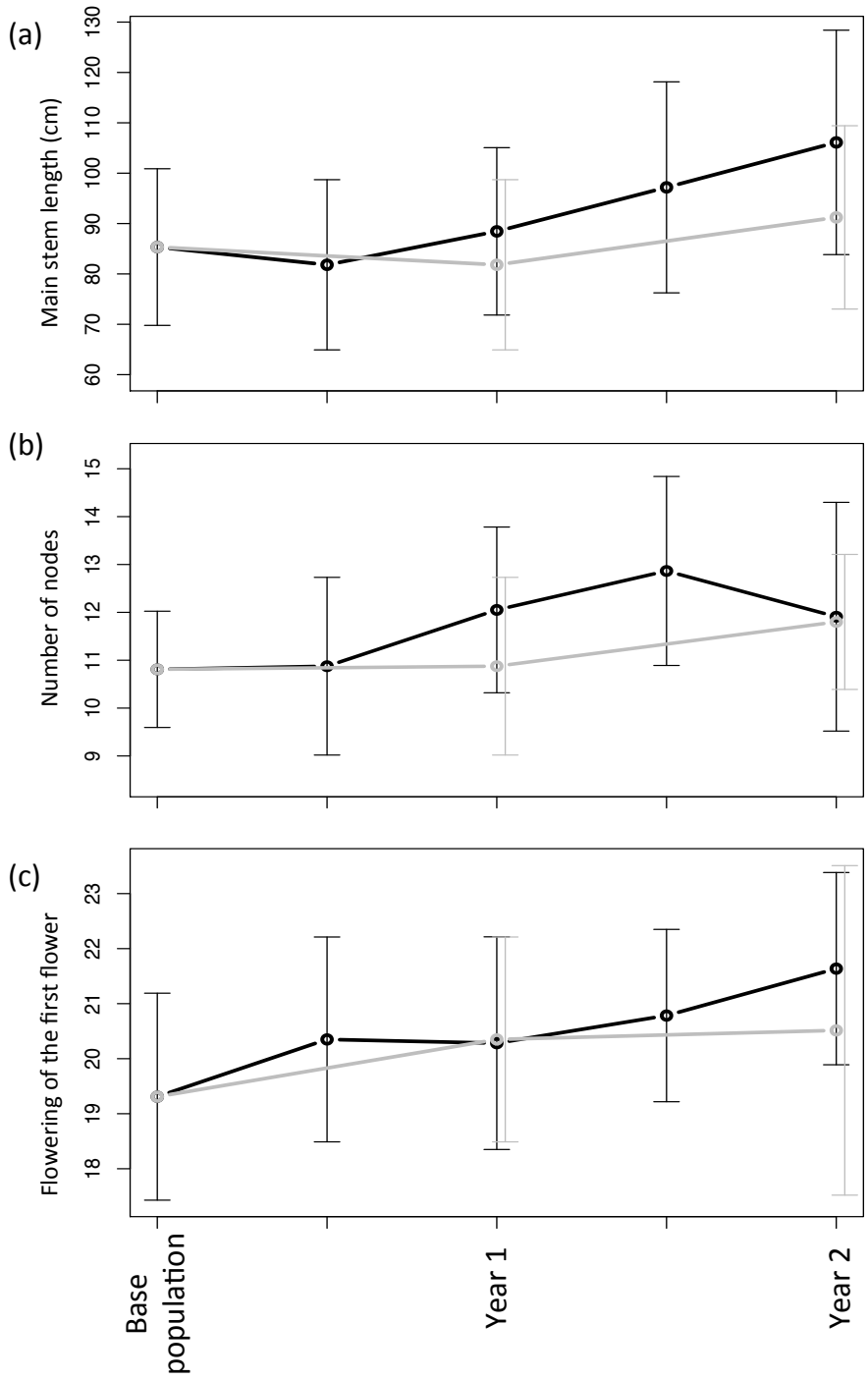


Figure 6.11. Change of population mean of nine yield relating traits through two years of PS breeding (gray) and GS breeding (black) for main stem length (cm: a), number of nodes (b), flowering of the first flower (days: c), number of flower clusters (d), number of primary branches (e), 1000 seed weight (g/1000 seeds: f), test weight (g/l: g), number of seed set in a plant (h), and number of secondary branches (i). The values were obtained in the field trial for comparison of breeding schemes.

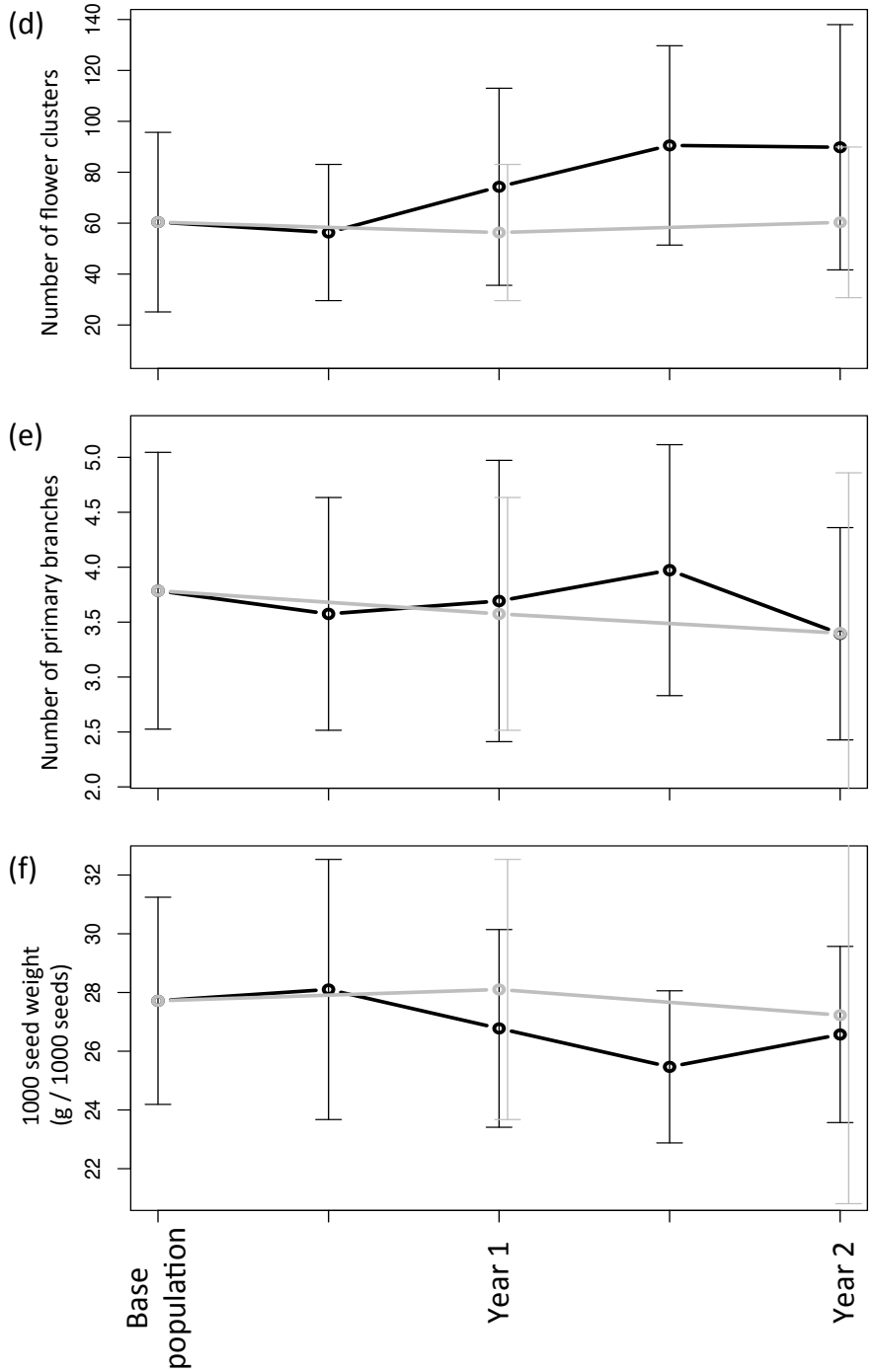


Figure 6.11. (Continued)

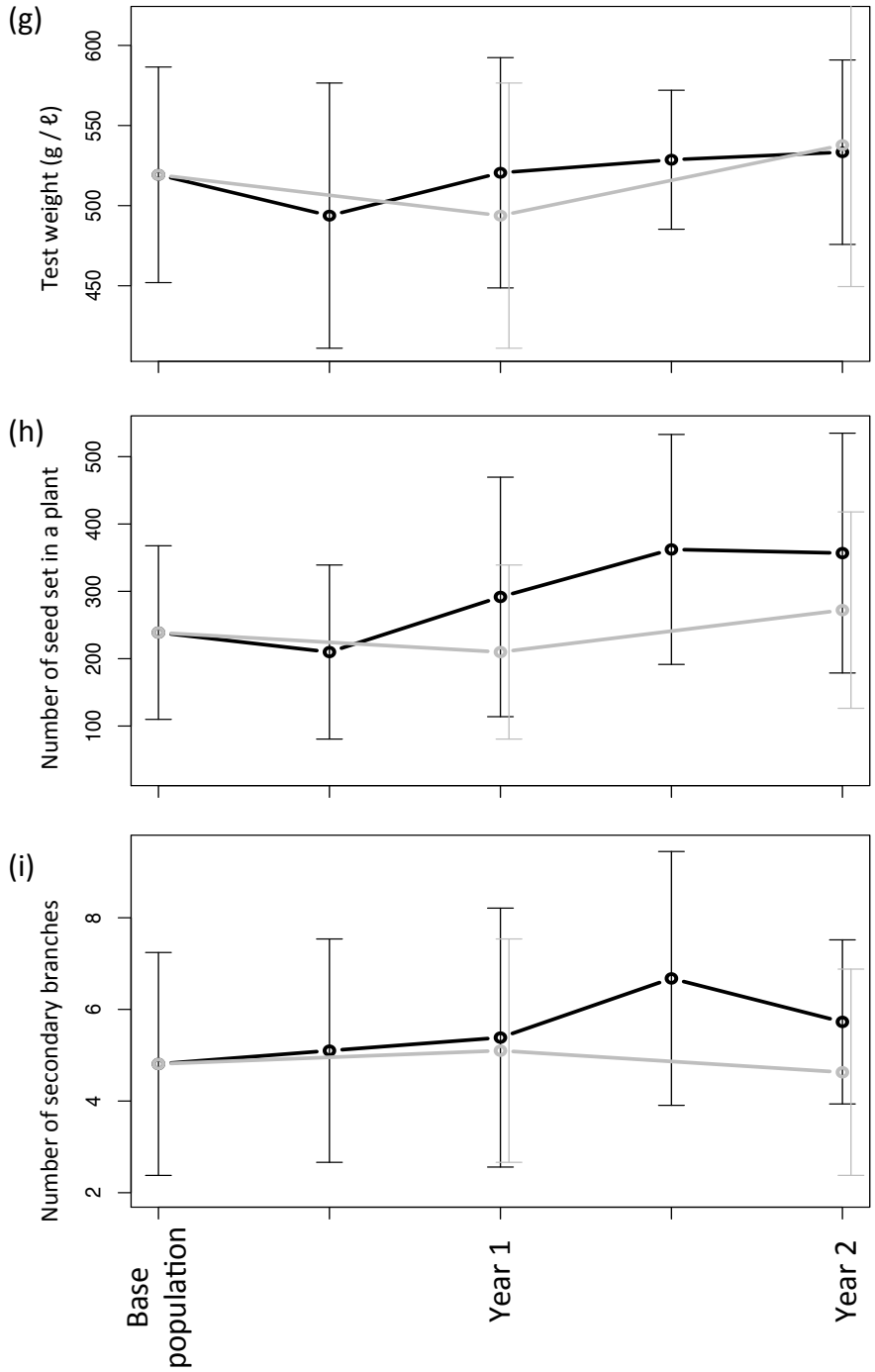


Figure 6.11. (Continued)

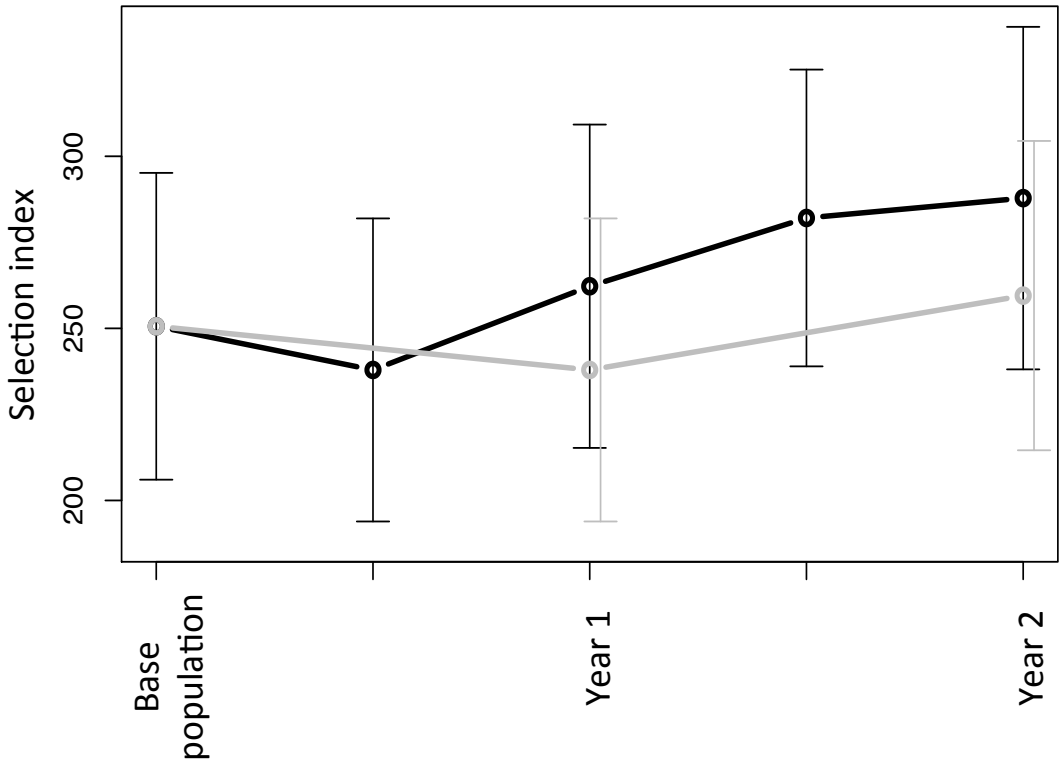


Figure 6.12. Change of population mean of selection index through two years of PS breeding (gray) and GS breeding (black). The values were obtained in the field trial for comparison of breeding schemes.

6-4. Discussion

6-4-1. Comparison of breeding schemes

I constructed a high-density linkage map for common buckwheat. The number of linkage groups converged to eight, which is the basic number of chromosomes in common buckwheat. The size of the linkage groups in the P1 and P2 linkage map were 773.8 and 800.4 cM, respectively (Fig. 6.2 and Table 6.1). Some linkage maps have previously been developed for common buckwheat by Hara et al. (2011), Konishi and Ohnishi (2006), Pan and Chen (2010), and Yasui et al. (2004). Compared with the previously constructed map, the map constructed in the present study had the largest number of markers on one side (i.e., P1 map or P2 map), and the average interval of adjacent markers was 2.13 cM, which is the smallest interval reported previously. Moreover, the map constructed in the present study converged to eight linkage groups including all markers used in the mapping analysis, whereas the others did not. Together, these results show that the map constructed in the present study will be suitable for use as a basic linkage map in the future studies of common buckwheat. The linkage map constructed in the present study had 756 independent loci and 8,884 markers (Fig. 6.2 and Table 6.1). Thus, the map harbored multiple markers at a single position (Fig. 6.4). This characteristic may contribute to the versatility of the linkage map. Common buckwheat has a high level of genetic variation; therefore, genetic composition differs by population. This high level of genetic variation makes it difficult to apply markers detected in one population to another population. However, if a single position has multiple markers, at least one of the markers may be polymorphic in a different target population. This allows the linkage-mapping step to be skipped for new target populations because the positions of the co-located polymorphic markers are already known. In actual, 1,484 polymorphic markers that represented 565 loci on the linkage map were detected in the breeding population of common buckwheat (Fig. 6.2). It suggests the usefulness of the linkage map constructed in the present study.

QTL were detected for main stem length (Fig. 6.6 and Table 6.2). QTL detected on the P1 map were not detected on the P2 map, suggesting that the QTL detected in one population (i.e., the loci heterozygous in one or both parents) may not be detected in another population. That is, QTL efficient in one population may not totally explain the genetic variation of the target trait in another population. This suggests that MAS may not be a suitable method for use in common buckwheat populations. This suggestion agrees with the previous study by Strauss et al. (1992), in which they suggested that QTL detected in a mapping population might not be responsible for variation in a breeding population. The result of QTL analysis implies the more

effectiveness of GS than MAS in plant breeding, in particular, when the target species is considered to have large variation in the population.

For the breeding population of common buckwheat in the present study, I calculated the level of LD among the pairs of polymorphic markers on the identical linkage group. The level of LD was low over the chromosome (i.e., linkage group), and the estimated effective population size was large (Fig. 6.7). It is suggested that the prediction accuracy of GS decreases in the population with large effective population size because the accuracy depends on the levels of LD (Resende et al., 2014). Calus et al. (2008) showed the dramatic improvement of prediction accuracy when the average r^2 of adjacent markers changed from 0.1 to 0.2 in their simulation study. Hayes et al. (2009) mentioned that r^2 should be greater than or equal to 0.2 on the basis of the previous studies. In the initial population of the present breeding program, although the expected r^2 was lower than 0.2 even between close markers, many pairs of markers represented the r^2 that were higher than 0.2 (Fig. 6.7). In the simulations in Chapter 3, additionally, I assumed an extreme situation, linkage equilibrium in the base population, and showed that GS breeding could work better than PS breeding even for such a breeding population with low levels of LD. From the information, it is inferred that GS breeding would work in the common buckwheat population under the field trial in the present Chapter.

In the field trial to compare GS breeding with PS breeding using common buckwheat, the target trait was seed yield per unit area, which it is required to evaluate with multiple plants. If measurements with multiple plants are performed, there is genetic heterogeneity in the common buckwheat population, and it is difficult to associate this target trait with genome-wide marker genotypes. To perform selection according to this target trait (i.e., yield per unit area), I created the selection index that can be represent the yield per unit area, using the relationship between other yield related traits, which can be evaluated in each individual, and the target trait (i.e., yield per unit area) (Table 6.3). All the selection steps were performed on the basis of this selection index. In GS breeding, the population mean of the selection index increased significantly by 14.9% from the mean of the base population through two years (four cycles) of selection. In PS breeding, this value increased 3.6% from the base population, while the increment was not significant (Fig. 6.12). This difference in improvement of the selection index between by GS breeding and by PS breeding suggests the advantage of GS breeding to improve genetic ability than PS breeding in common buckwheat. At the first GS and PS steps, the observed and expected values of the selection index showed high correlation, thus the selected 12 individuals (i.e., genotypes) were almost identical (Fig. 6.8a). And, the response to one cycle

of selection in GS breeding was similar to that of PS (Fig. 6.12). This suggests the efficiency of GS in the offseason nurseries that enable two selection cycles per year, which agree with the suggestion by the previous simulations in Chapter 3. In Chapter 3, GS breeding with three cycles of selection per year attained higher gain than GS with one or two cycles in short-term breeding program. In the field trial of this Chapter, my colleagues and I conducted GS breeding with only two cycles per year because of the speed of genotyping and analysis for GS. At the step where a prediction model was updated (i.e., GS1 and GS3), it was not necessary to complete three tasks, i.e., genotyping, phenotyping, and updating a prediction model, prior to flowering of the breeding population. At GS2 and GS4, however, we had to complete genotyping and selection of parents within a period from budding to flowering in the breeding population (i.e., about 40 days). Because of this tough labor, we have decided to conduct GS twice per year in this breeding program of common buckwheat. More cycles of GS would improve the efficiency of GS breeding in two years if we could do that. In this field trial, the selection index was observed only after pollination, thus the pollen parents were not controlled at all generations in PS breeding and at GS1 and GS3 (i.e., generations where the prediction model updated) in GS breeding. In GS breeding, however, pollen parents could be selected at GS2 and GS4 because only marker genotype data were used in selection at these generations. In Chapter 3, the efficiency of pollen control in GS breeding was suggested. In the field trial in the present Chapter, the possibility of the pollen control in GS breeding might work.

For the seven traits that consisted of the selection index, I observed the trend that traits weighted largely in the selection index were improved largely in GS breeding (Table 6.3 and Fig. 6.11). In GS breeding, number of seed set in a plant, which was not included in the selection index but is much important for yield per unit area, increased by 49.4% from the base population. Number of primary branches and 1000 seed weight, which had a small positive and a negative weight in the selection index, respectively, slightly decreased through two years of GS breeding. Although it is known that number of seed set in a plant and test weight show trade-off relationship, the trade-off was not shown in the field trial. Therefore, the selection index is thought to be effective to improve yield per unit area with keeping the balance between traits included in the index.

This field trial was conducted in the similar way to simulations in Chapter 3. The result of this field trial study was similar to that of the simulations in Chapter 3. In this study, I could not find the factors that caused serious discordance between the empirical study and the simulation study. I evaluated the selection accuracy to compare the result of the field trial with that of the

simulations. In the present study, the potential of the base population was evaluated instead of the potential of the initial breeding population. The base population is thought not to represent the genetic condition of the initial population because the initial population experienced a genetic bottleneck after the base population. Therefore, I could not analyze the effect of the prediction accuracy of GS to the degree of genetic improvement in this study. The prediction accuracy of GS3 was lower than that of GS1 for the selection index (Table 6.5). The reason of this decrease of the accuracy is the lower heritability at GS3 than at GS1 because the genetic variance became lower at GS3 than at GS1 trough selection. The prediction accuracy depends on the heritability with the assumption that the phenotypic variance is constant, and low heritability results in the low prediction accuracy (Daetwyler et al., 2008). In this field trial, the genetic improvement of the breeding population appeared to be lower than that expected in simulations. The reason might be the small breeding population size in the field trial, which resulted in strong genetic bottleneck in the breeding population. I, however, could not be convinced of this hypothesis, because the data of the initial population could not be used in this study. Moreover, the simulation result represented the mean value of 100 trials while the result of the present field trial was just one trial. To verify the difference between simulations and field trials and to clarify the reason behind it, further field trials and their detailed analysis might be required. In Chapter 3, the importance of updating a prediction model was suggested. It is true especially when a breeding population has low levels of LD because of the rapid change of patterns of LD in the breeding population through selection cycles. In the present study, the prediction model built at GS1 showed lower accuracy at GS3 than at GS1 (Fig. 6.9). If this previous model had been used at GS3, different individuals were selected from the individuals that were selected at GS3 actually. These results suggest that it is important to update the prediction model with selection cycles in GS breeding of an allogamous plant population with low levels of LD, as is the case of simulation study in Chapter 3.

6-4-2. Conclusion

GS was effective to improve the genetic ability of the breeding population in common buckwheat. This result can be applied to the other plant species because of the fundamentality of mass selection used in this study. The advantage of GS breeding over PS breeding was suggested in the field trial. The advantage of GS over MAS was implied by QTL analysis, in which it is suggested that QTL efficient in one population may not totally explain the genetic variation of the target trait in another population. The usage of a selection index might be

effective when it is difficult to evaluate the target trait in a single plant. In breeding of allogamous crops with low levels of LD, updating the prediction model is important to maintain the accuracy of GS, especially when selection and crossing are repeated among the selected plants (i.e. genotypes) as mass selection conducted in this field trial. The efficiency of GS mainly comes from the acceleration of generations by offseason nursing and the possibility of pollen control. In the current situation, however, the limiting factor of GS breeding with acceleration of generations might be the speed of genotyping, in particular, when the target species has a short generation time. To put GS into practical use, a cheap and rapid high-throughput genotyping system is required.

Chapter 7

Development of a simple language to script and simulate breeding schemes: the breeding scheme language

7-1. Introduction

The use of optimal breeding schemes is critical in plant breeding. Generally, however, there are a number of possible breeding schemes, and it is not easy to determine the optimal scheme under conditions of a target species and target traits. Plant breeders face the difficulty to decide a detailed breeding scheme. This situation makes an increased need for a system that helps breeders to find an optimal breeding strategy.

The difficulty of choosing a breeding scheme makes breeders conservative to try a new one. A new breeding strategy possesses the possibility of critical problems even if it looks like a good scheme because plant breeding is involving many systems to work. Once they are sure that a conventional breeding scheme works, they do not want to change their breeding scheme at a risk of failure. Thus, a system that helps breeders to try some breeding strategies is also necessary.

Simulation study is useful to help choose a better (or the best) breeding scheme. Wang et al. (2003) conducted a simulation study to compare the two breeding strategies in wheat, in which inbreeding were conducted. They chose one scheme according to the simulation results including genetic improvement and cost efficiency. They also evaluated the impact of genetic architecture. Yano et al. (2000) tried optimizing some factors in mass selection of outcrossing species by simulations. The above studies assumed PS, which included a lot of factors and systems in breeding. GS, in particular, more factors may affect to the outcome than PS because GS involves more factors than PS such as marker density, relationship between a training

population and a breeding population, LD patterns in these populations, and so on. Lorenz (2013) suggested the importance of resource allocation in training population in GS and evaluated the difference between different numbers of replications and population sizes by using simulations assuming maize. Hickey et al. (2014) also conducted simulations to decide training population designs and suggested that the best training population design depended on the marker density. As suggested by Hickey et al. (2014), simulation study is useful to detect unexpected outcomes before they perform an actual field trial. For example, Bernardo and Yu (2007) conducted simulation study to evaluate the prediction accuracy and the response to GS in maize. Massman et al. (2013) reported an actual field trial result, in which the breeding scheme was similar to that assumed in Bernardo and Yu (2007). The simulations showed GS has 18 to 43% larger gain than marker assisted recurrent selection when the target traits were controlled by 20, 40, and 100 QTL under the heritability heritability of 0.2, 0.5, and 0.8. This results were consistent with the field trials where GS showed 14 to 50% larger gain than marker assisted recurrent selection.

Some simulation platforms were created to evaluate and compare several breeding schemes or parameters (Sun et al., 2011). For GS, Riedelsheimer and Melchinger (2013) developed a calculation tool to optimize the allocation of resources in one cycle for biparental population. However, breeders should consider about the genetic improvement through the all selection cycles. It is difficult for breeders to take the first step in conducting breeding simulation because of the complexity to build a simulation platform or even to using a simulation tools. A simple and flexible simulation plat form is necessary for breeders to evaluate their planned breeding schemes.

In the present study, I created a breeding scheme language as a novel simulation platform by using R (R Development Core Team, 2014). Breeders can apply their planned genetic architecture and breeding schemes flexibly in the system. Moreover, it can be used as an education tool of breeding simulation because users can try various parameters and breeding schemes by themselves and understand how these parameters and schemes affect the results of the breeding simulation.

7-2. Description

The breeding scheme language is composed of some functions describing about founder, breeding, mating, and simulation result. The functions utilize the style of R function, thus users write the function name and describe parameters in the parenthesis after the function name. Users can simulate their planned breeding scheme by writing functions in order of implementation. Each function has the default inputs, and the default inputs are adopted when users describe no input in the function. The breeding scheme language is written in R (R Development Core Team, 2014).

Founder functions

To make an initial population for breeding, two functions are provided:

- defineSpecies(win = F, loadData = F, nSim = 1, nCore = 1, nChr = 7, lengthChr = 150, effPopSize = 100, nMarkers = 1000, nQTL = 50, propDomi = 0, nEpiLoci = 0)
- initializePopulation(nPop = 100, gVariance = 1)

where the parameters in the parenthesis represent the default inputs.

The function "defineSpecies" defines and creates species data. Each parameter means:

win	PC is Windows or not
loadData	load the last simulated species data or not
nSim	the number of simulation trials
nCore	simulation processed in parallel over this number of cores (If "win=T", this parameter is neglected and "nCore=1" is used.)
nChr	the number of chromosomes
lengthChr	the length of each chromosome in cM
effPopSize	the effective population size in the base population
nMarkers	the number of markers to use in GS
nQTL	the number of QTL controlling the target trait
probDomi	the probability of dominant QTL among the all QTL
nEpiLoci	the expected number of epistatic loci for each effect

To create species data, the function utilizes the software GENOME (Liang et al., 2007) in the present conditions. This software is unsupported by Windows, thus Windows users need to bring species data from another computer. And, I utilize R package "parallel" (Urbanek, 2013) to parallelize calculations, which is not supported in Windows. When the parameter "win" is TRUE, the parameter "nCore" is inputted as 1 automatically. By adopting

loadData=T, users can compare some breeding schemes by using the same data simulated in the platform.

The function “initializePopulation” creates an initial population for breeding. Parameters represent:

nPop	population size
gVariance	genetic variance in the initial population

The genetic variance is important when phenotyping is implemented (mentioned later) because it depends on the scale of heritability.

Breeding functions

Four functions are provided to implement actions in plant breeding:

- phenotype(errorVar = 1, popID = NULL)
- genotype()
- predict(popID = NULL, trainingPopID = NULL)
- select(nSelect = 40, random = F, popID = NULL)

where the parameters in the parenthesis represent the default inputs.

The function “phenotype” implements phenotyping of individuals in specified population. The parameters mean:

errorVar	environmental error variance
popID	population ID to be evaluated (Default is the latest population)

The function “genotype” implements acquires marker genotypes of all genotype included in the breeding scheme. It has no input.

The function “predict” performed genomic prediction by using phenotype data and genotype data obtained previously. Each parameter means:

popID	population ID to be predicted (Default is the latest population)
trainingPopID	population ID to be used to train a prediction model (Default is the all populations having phenotype data)

This function utilizes R package “rrBLUP” (Endelman, 2011).

The function “select” conducts selection in the defined population.

nSelect	the number of selected individuals
random	assuming random selection or selection according to their features

popID population ID to be selected (Default is the latest population when random=T. When random=F, default is the last evaluated population.)

This function can be used not only for selection of favorite genotype, but also for random selection. This is useful when breeders try producing genetic bottleneck or evaluating a portion of individuals among a population.

Mating functions

Three functions are provided for mating among parents and producing progenies:

- cross(nProgeny = 100, equalContribution = F, popID = NULL)
- selfFertilize(nProgenyPerInd = 1, popID = NULL)
- doubledHaploid(nProgeny = 100, popID = NULL)

where the parameters in the parenthesis represent the default inputs.

The "cross" function conducts random mating among parents. The function "selfFertilize" implements inbreeding, and the "doubledHaploid" function makes doubled haploids from assigned population. Parameters mean:

nProgeny the number of progenies

nProgenyPerInd the number of progenies derived from one parent

equalContribution If TRUE, all individuals are used the same number of times as parents. If FALSE, individuals are chosen at random to be parents.

popID population ID to be used as parents (Default is the latest population.)

In the function "cross", the number of progenies should be larger than the number of parents when users choose equalContribution=T.

Results functions

Two functions are provided to show and save the results of breeding:

- plotData(ymax = NULL, add = F)
- outputResults(summarize = T, directory = NULL)

where the parameters in the parenthesis represent the default inputs.

The function "plotData" shows the genotypic value trough generations of breeding in a figure. The figure shows population means of each simulation replication and the averaged value over simulation replications. By assigning the parameter "ymax", users can define the

maximum value of genotypic values in the figure. If user write add=T, the result obtained in the previous simulations were also drawn in the figure.

The function “outputResults” saves the results. If summarize=T, the results averaged over the all simulation replications are saved. If summarize=F, the data used in breeding simulations are saved directly. Users can assign the directory in which the data is saved. This parameter is useful especially when summarize=F because the data is too large to save in the current directory.

Population ID

In many functions mentioned above, the parameter “popID” is used. The population ID starts from 0 in the initial population. After the initial population, the population ID is updated when the following functions are represented in code:

1. select
2. cross
3. selfFertilize
4. doubledHaploid

The results are shown according to the population ID in the figure. However, the selected individuals belong to the previous population ID (i.e., population ID they belonged to before they were selected) in the result figure.

User-friendly tool

The whole simulation code was written in R. I used a package “shiny” (RStudio, Inc., 2014), which is a web application framework developed by RStudio, to create a user-friendly tool. By using a package “shinyAce” (Trestle Technology, LLC., 2013), users can describe their planned breeding scheme and look result in one screen (Fig. 7.1).

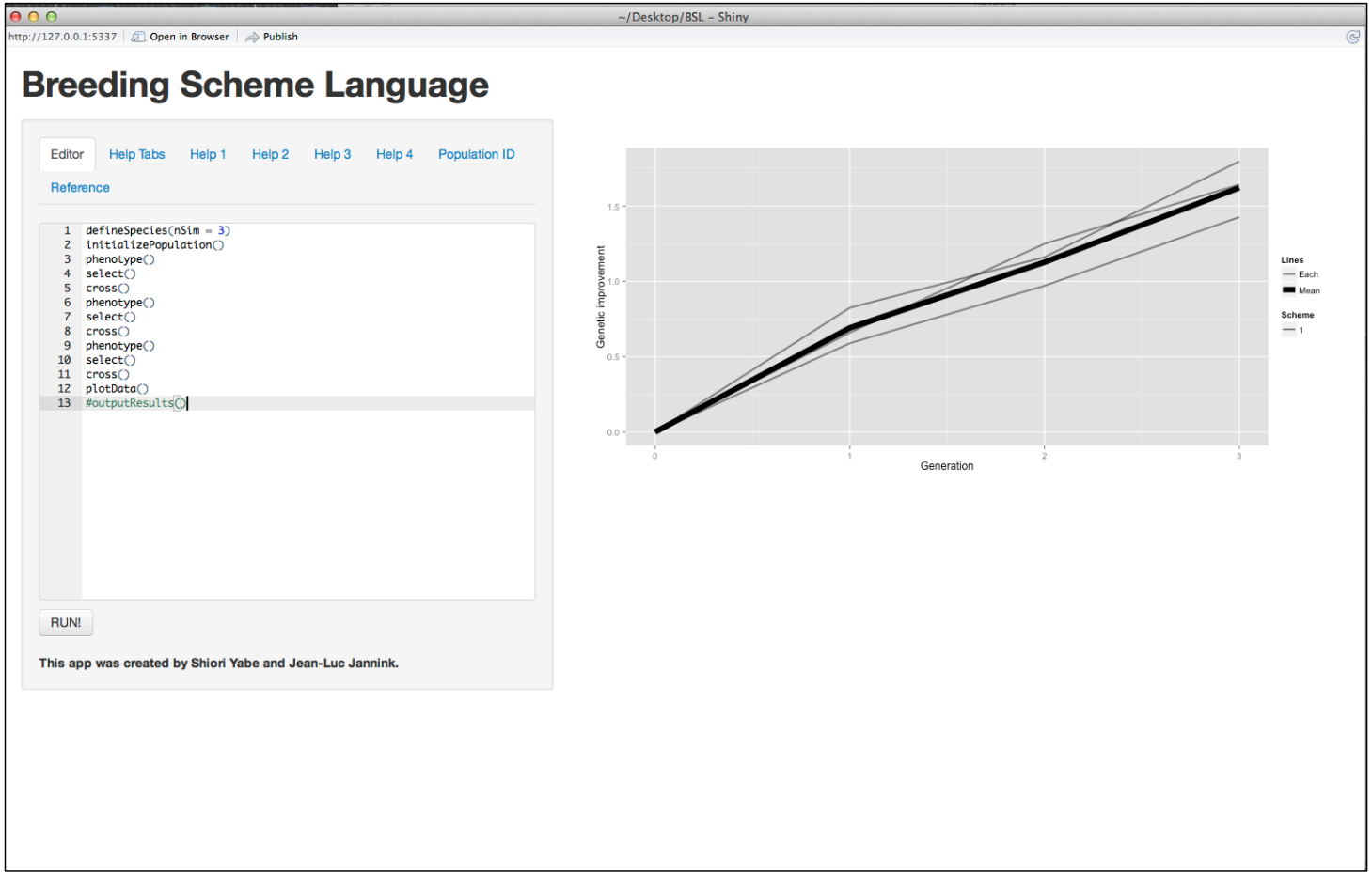


Figure 7.1 The screen of the simulation platform using breeding scheme language.

7-3. Computational problems and future plans

In the present conditions, the breeding scheme language creates an initial population ahead of breeding simulation. Moreover, the adopted software GENOME is not supported in Windows. In the future, I can rewrite the function to create an initial population by the other software and to import genome data users possess. It is useful that breeders import their data and conduct breeding simulation. In many cases, breeders want to try evaluating the efficiency of breeding schemes on the basis of the actual initial population data.

7-4. Examples

7-2-1. Example 1

As an example, I show the result of comparison PS with GS. At first, I conducted three times of simulation replications in which three times of PS were conducted (Fig. 7.2a). The executed code was:

```
defineSpecies(nSim = 3)
initializePopulation()
phenotype()
select()
cross()
phenotype()
select()
cross()
phenotype()
select()
cross()
plotData()
```

The obtained result was shown in Figure 7.3 (a). Gray lines represent results of each simulation replication, and black line represents the averaged value over all simulations. Then, the GS, in which the prediction model was updated each selection by using all available phenotypic data, was conducted (Fig. 7.2b). The implemented code was:

```
defineSpecies(nSim = 3, loadData = T)
initializePopulation()
phenotype()
genotype()
predict()
select()
cross()
phenotype()
genotype()
predict()
select()
cross()
```

```
phenotype()
genotype()
predict()
select()
cross()
plotData(add = T)
```

Here, I used the same initial population as the previous simulations of PS. And the result should be drawn with the previous one. Figure 7.3 (b) is the obtained result. Solid lines represent the result of GS, and dotted lines show those of PS. I can see that GS attained a little higher genetic gain than PS in the three simulation replications. After that, I tried GS in which the prediction model was not updated (Fig. 7.2c). The code was:

```
defineSpecies(nSim = 3, loadData = T)
initializePopulation()
phenotype()
genotype()
predict()
select()
cross()
genotype()
predict()
select()
cross()
genotype()
predict()
select()
cross()
plotData(add = T)
```

Figure 7.3 (c) shows the results of three kinds of breeding schemes mentioned above. Solid lines represent the result of GS without updating the prediction model. Dotted lines represent the GS with update of prediction models, and dashed lines represent the PS, which were implemented on ahead. Comparing in the selection cycles regardless the time necessary for breeding, GS without update the prediction model attained lower genetic gain than others. However, I need more replications of simulations to evaluate the potential of breeding schemes.

7-4-2. Example 2

I simulated the breeding schemes assumed in cassava breeding, which were similar to those simulated in Chapter 5. Here, however, I considered the time to implement the breeding schemes. Additionally, I took account of the cost and effort to propagate and evaluate genotypes. I assumed four years of breeding scheme after the first selection using the historical phenotypic data. The breeding scheme using seedlings assumed to take two years for one cycle because of the effort of evaluation seedlings. The scheme discarding seedling data takes one year for one selection cycle. Thus, the scheme using seedling data can implement two selection cycles after the first selection, while the other scheme implements four cycles. I selected 50 genotypes randomly to propagate for preliminary yield trials because it requires a lot of effort and land. The simulation of breeding scheme discarding seedling data (Fig. 7.4a) was written as:

```
defineSpecies(nSim = 6, nCore = 1, nChr = 18, lengthChr = 110, effPopSize = 100,
nMarkers = 2000, nQTL = 100, propDomi = 0, nEpiLoci = 0)
initializePopulation(nPop = 200, gVariance = 1)        # 0
phenotype(errorVar = 1, popID = NULL)
genotype()
predict(trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 1
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 2
genotype()
predict(trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 3
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 4
phenotype(errorVar = 16, popID = 2)
genotype()
predict(trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 5
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 6
phenotype(errorVar = 16, popID = 4)
select(nSelect = 50, popID = 2, random = T)        # 7
phenotype(errorVar = 9, popID = c(3, 7))
genotype()
```

```
predict(popID = 6, trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 8
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 9
phenotype(errorVar = 16, popID = 6)
select(nSelect = 50, popID = 4, random = T)        # 10
phenotype(errorVar = 9, popID = c(5, 10))
genotype()
predict(popID = 9, trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 11
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 12
plotData()
```

The number written after "#" is the population ID corresponding to the population derived by the function. The simulation of scheme using seedlings (Fig. 7.4b) was written as:

```
defineSpecies(loadData = T)
initializePopulation(nPop = 200, gVariance = 1)        # 0
phenotype(errorVar = 1, popID = NULL)
genotype()
predict(trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 1
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 2
phenotype(errorVar = 36, popID = NULL)
genotype()
predict(trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 3
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 4
phenotype(errorVar = 36, popID = NULL)
phenotype(errorVar = 16, popID = 2)
select(nSelect = 50, popID = 2, random = T)        # 5
phenotype(errorVar = 9, popID = c(3, 5))
genotype()
predict(popID = 4, trainingPopID = NULL)
select(nSelect = 40, popID = NULL, random = F)        # 6
cross(nProgeny = 600, equalContribution = F, popID = NULL)        # 7
```

```
plotData(add = T)
```

where the same initial populations as the previous simulations were used. Figure 7.5 shows the results of simulations described above. Solid lines represent the result of breeding scheme using seedling data, while dotted lines represent the other one. It suggests that both breeding schemes attained the same level of genetic gain in four years. I conducted six replications of simulations by using same initial populations for both of the breeding schemes. The expected genetic gain after the first selection was same in both schemes because the first selection used only historical data that was identical in both schemes. However, the results were different between these schemes because of the fluctuation in mating, suggesting the large fraction of each simulation replication. More simulation replications are required to evaluate the potential of breeding scheme.

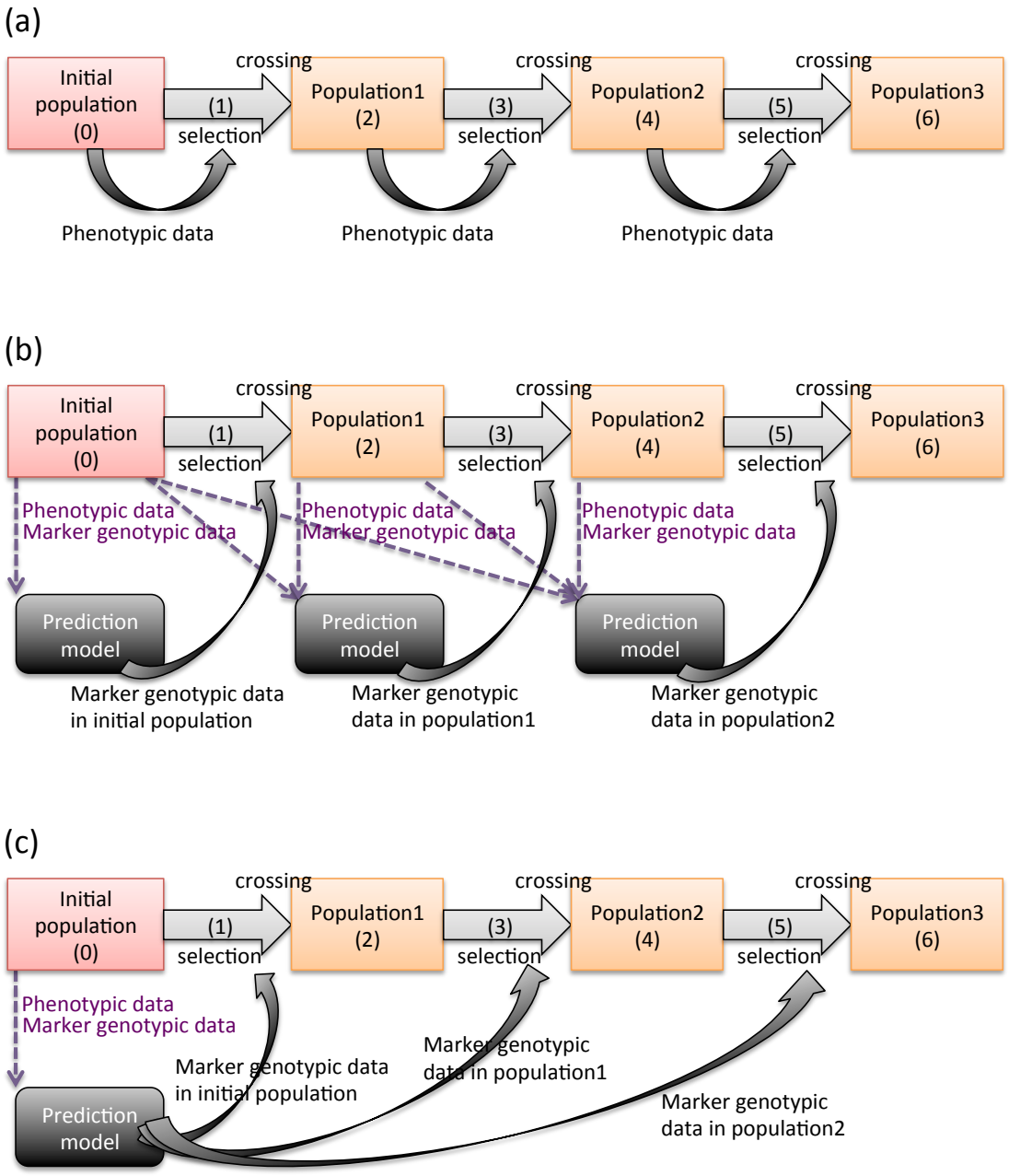


Figure 7.2 Breeding schemes evaluated in the example 1. Nnumbers in parenthesis are population IDs. Breeding scheme for PS (a), GS with model updates (b), and GS without model update (c).

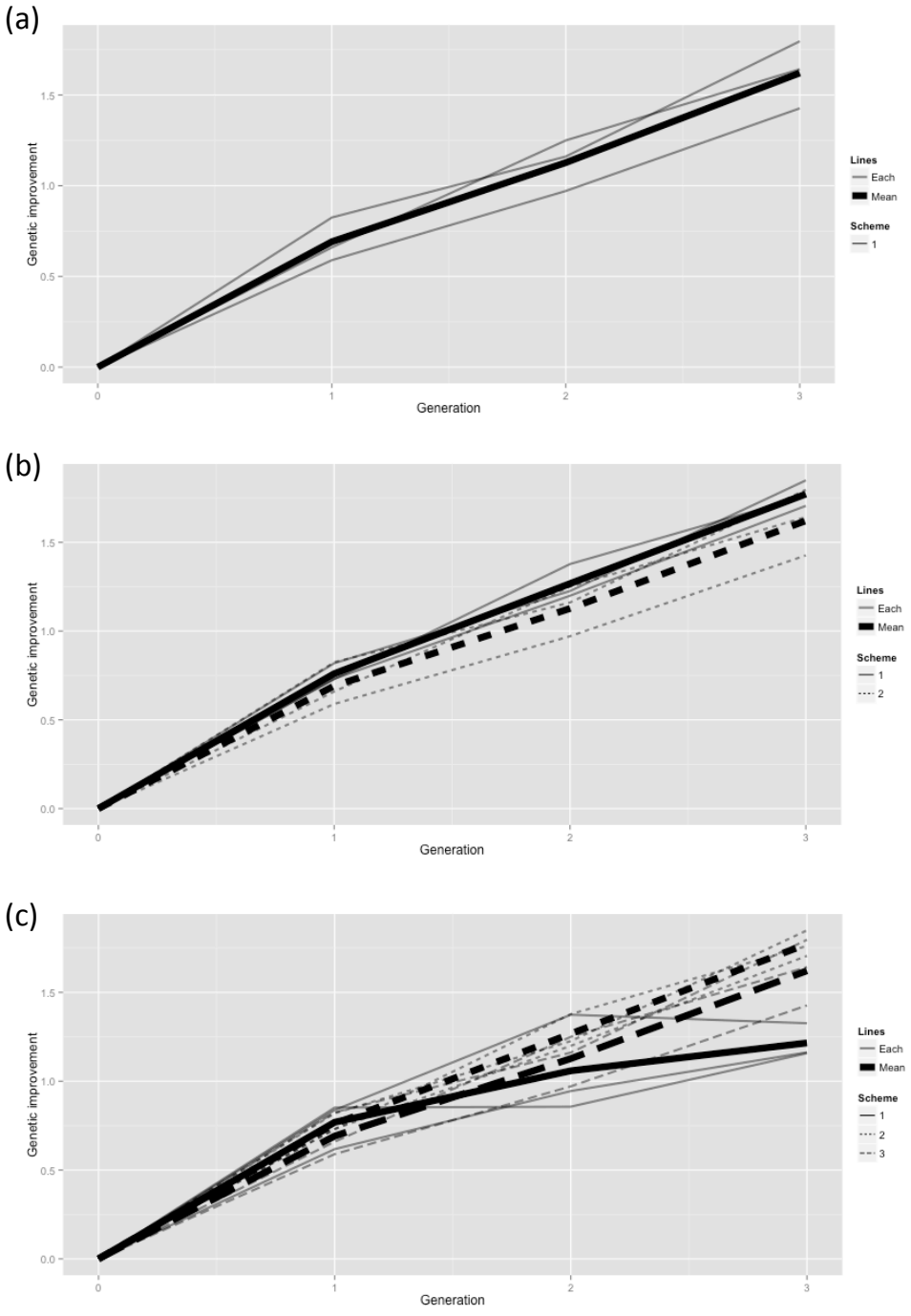


Figure 7.3 Comparison of genetic gains among three breeding schemes. Black lines represent the averaged genetic gain over the simulation replications. Gray lines show each simulation replication's genetic gain. (a) PS. (b) PS (dotted lines) and GS with model updates (solid lines). (c) PS (dashed lines), GS with model updates (dotted lines), and GS without model update (solid lines).

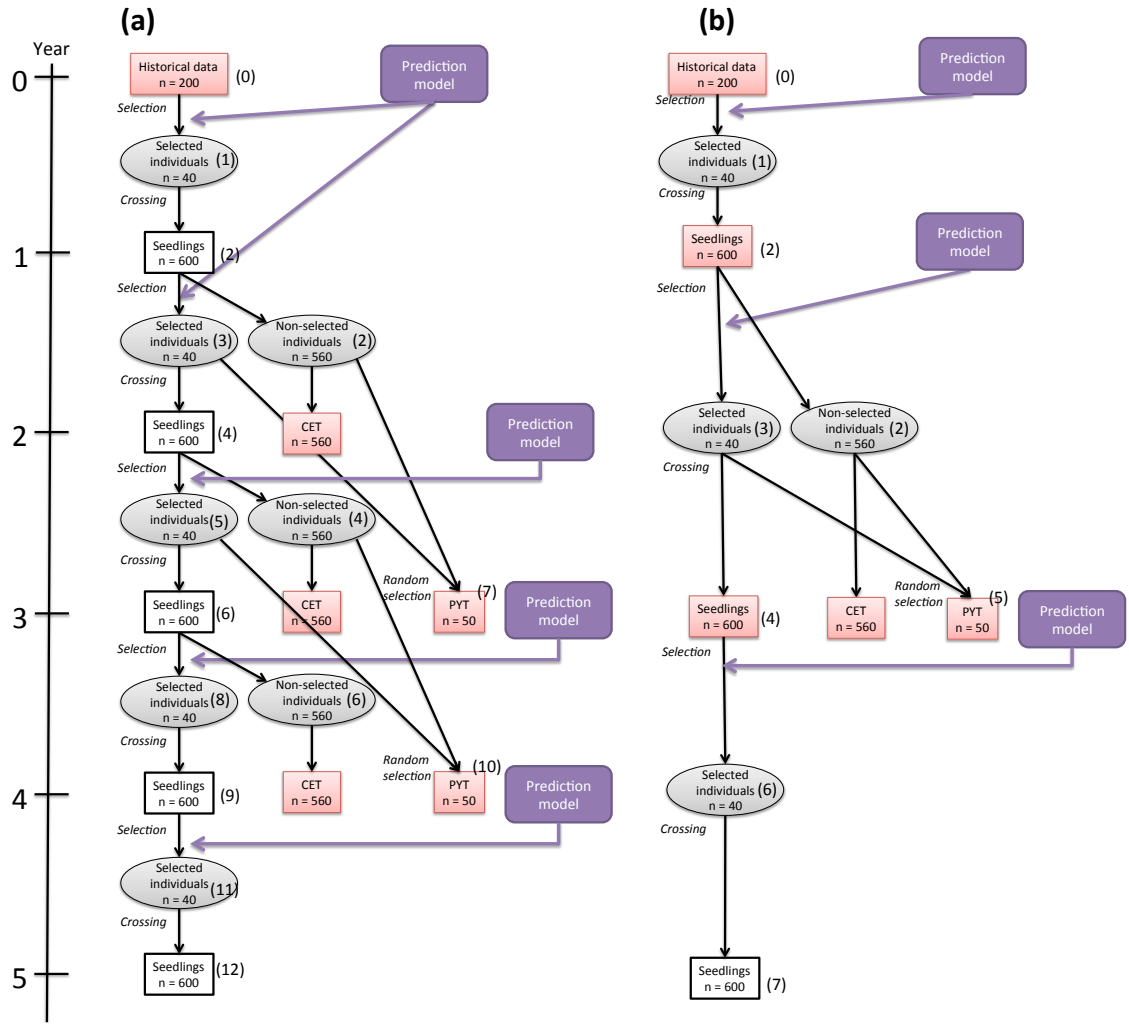


Figure 7.4 Breeding schemes evaluated in the example 2. Numbers in parenthesis are population IDs. To build a prediction model, all phenotypic data that were evaluated before model building process (phenotypic data for the populations shown as red boxes) are used. Breeding scheme discarding seedling data (a) and using seedling data (b).

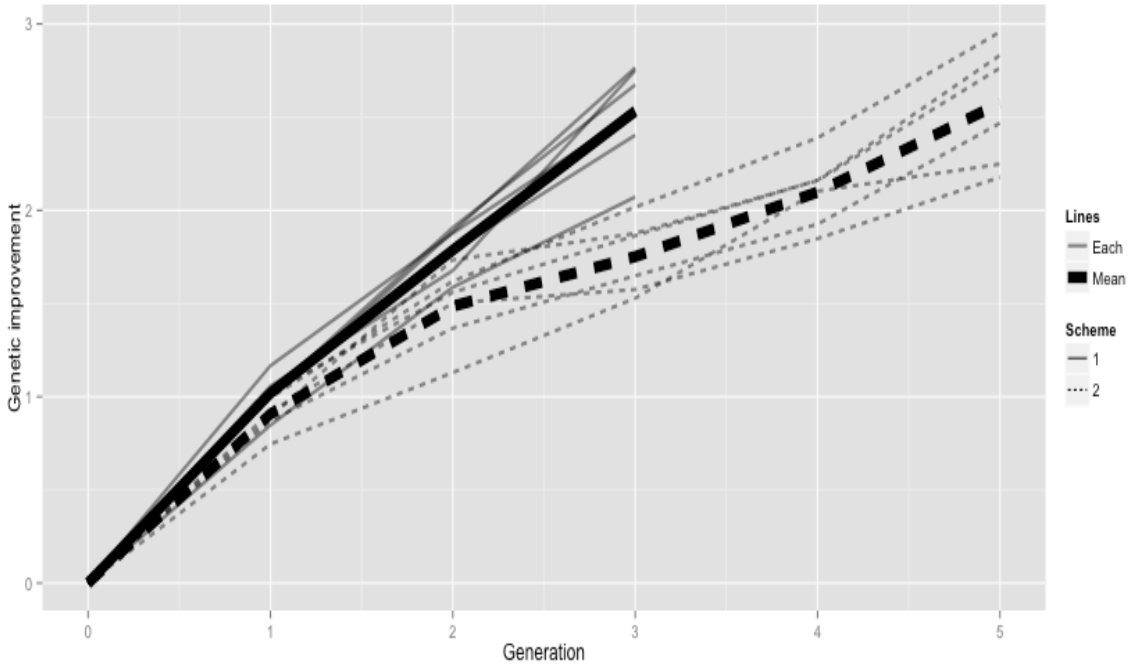


Figure 7.5 Comparison of genetic gains among two breeding schemes in cassava. Black lines represent the averaged genetic gain over the simulation replications. Gray lines show each simulation replication's genetic gain. Solid lines show the scheme using seedling data. Dotted lines show the scheme discarding seedling data.

7-5. Discussion

In the present study, I created a simple and flexible simulation platform, breeding scheme language. Users can define their target species and breeding schemes by the language. This language will be useful for breeders to evaluate breeding schemes and to choose a better (or the best) breeding scheme among a number of possible choices. The breeding scheme language can show a result figure in which users can compare various schemes. If users hope to analyze the data of breeding process in detail, they can see all results in each step of breeding process.

The breeding scheme language needs more improvement for flexible input data. Users may have marker genotype data of their breeding populations. In this case, it is better to use a real marker data than to use a virtual data so that the situation could reflect a real situation of a breeding population, e.g., the patterns of LD and the genetic relationship among genotypes compositing the population. Availability of the function to use an actual data would be helpful to choose an optimal breeding scheme based on the information reflecting users' target population.

This breeding scheme language is based on R, and the way of input to simulations is similar to the R's way. In this breeding scheme language, users are required to write functions in the correct order to represent their planned breeding scheme. By using this breeding scheme language, users can learn how to compose breeding simulation and how to write function in R. The breeding scheme language can be used not only to evaluate the planned breeding schemes, but also for education of simulation programming. Additionally, by conducting some simulation replications and looking the result of all replications, users can realize the fluctuation among the replications. However, there should be users who hope just to simulate breeding processes. For this type of users, an easier system may be more useful. For example, it is useful to develop a user interface that allows a user to compose their planed breeding scheme more easily than the present style.

In this language, it takes long time to finish all simulations with replications. This problem is crucial when users try to a number of replications in each setting or when the size of breeding population is large. By improving the way of data handling inside the simulation system, this problem will be solved in some degree. The improvements mentioned above will make this breeding scheme language more useful and more easy-to-use.

Chapter 8

Discussion

This dissertation shows that GS is effective and feasible in plant breeding. By comparison of genetic gains of GS breeding with those of PS breeding or MAS breeding, it was suggested that GS breeding could attain higher gains than the others on the basis of the simulations and the field trial. In the field trial, it was shown that GS could be implemented with acceleration of selection cycles. These results would encourage plant breeders to apply GS to their breeding programs.

At first, I demonstrated the high potential of GS on the basis of the simulation studies assuming three different types of plant species. In Chapters 3, 4, and 5, I conducted GS breeding simulations of an annual allogamous plant, an autogamous plant, and an allogamous plant with vegetative propagation, respectively. I performed these simulation studies under limitations and conditions imposed by reality in a breeding program of each plant species. In Chapter 3, the base population was in linkage equilibrium to reflect the situation of a species with a large effective population size. Although this situation was unsuited for GS because GS utilizes LD between QTL and markers, GS breeding could attain higher gain than PS breeding and MAS breeding. The success of GS in this breeding population was thought to be due to the LD that was formed through the selection cycles. In Chapter 5, I supposed that a target species was cassava, which is an allogamous species with the ability of vegetative propagation. The ability of vegetative propagation of cassava delivers the advantage that enables to estimate genotypic values of selection candidates precisely by repeated measurements of traits with clonally replicated genotypes. I assumed that we could update a prediction model by using CET and PYT data in addition to seedling data in the simulations, and showed that the prediction model utilizing seedling data showed higher performance than the model discarding seedling data. In both simulation studies, it is suggested that single-plant evaluation data, in which large environmental variation existed, was still useful to build a prediction model especially when the

evaluated plants (i.e., plants in a training population) are closely related (or identical) to selection candidates. In Chapter 4, I conducted GS simulations using marker genotype data of real cultivars in rice, which is an autogamous species. In this simulation study, I supposed that populations derived from crosses between two parental RILs were used as an initial population and a training population. I demonstrated that recurrent selection with GS could improve the population to have higher genotypic values than the highest genotypic value among all RILs. In this simulation study, I also showed the efficiency of the island-model GS, which was the novel strategy for GS breeding, which was proposed originally in this dissertation. The island-model GS requires moderately-differentiated subpopulations as initial subpopulations, and can utilize the genetic variation via selection conducted in each subpopulation. In autogamous crops, among-cultivar diversity is often large so that genetic difference among segregating families becomes large. Applying the island-model GS to such a breeding population can be efficient. Therefore, the island-model GS worked well in my simulation study assuming an autogamous crop. Through these simulation studies, the high potential of GS was suggested in plant breeding. At the same time, the importance of the selection of a suitable breeding scheme for each plant species is suggested.

Through the simulation studies of this dissertation, I found that some factors affected the efficiency of GS breeding, whereas some factors had only a small influence on the eficiency. In Chapter 5, it is suggested that small rates of mis-labeling (i.e., 10% or less) had an insignificant effect on the genetic gain in GS breeding. This result means that there is not much point in increasing cost to prevent breeders from making mis-labeling when the rate of mis-labeling is small. Carefulness of each person will be enough for preventing mis-labeling under these levels. On the other hand, breeders should update a prediction model even though it is fairly costly. In Chapter 3, it is suggested that a prediction model should be updated because prediction accuracy declined rapidly if one prediction model was used continuously. The same case also happened in Chapter 5 when a prediction model was not updated. In Chapter 4, on another front, one prediction model could be used in GS of an autogamous species with gradual decline in the prediction accuracy. The difference might result from the difference of LD patterns between breeding populations assumed in these simulations. If an initial population has low levels of LD, a prediction model should be updated to catch up the rapid change of LD patterns in a breeding population. In Chapter 4, a population of inbred lines tends to have high and wide range of LD, resulting in slower change of the patterns of LD than the case of populations that have low levels of LD initially. These results suggest that it is essential to update a prediction model in

GS breeding at a suitable timing based on the LD pattern in a breeding population. Although updating a prediction model is fairly costly, it is necessary because prediction accuracy affects on the outcomes largely. LD patterns in a breeding population are also important, considering the suitable number of markers as well as the timing of model updating. Thus, it is important to estimate LD patterns on ahead of breeding. To find critical factors in GS breeding (or plant breeding with other selection methods), it is useful to conduct simulations under various situations and breeding schemes. When a breeding population that contributes to a supposed breeding program is already defined, it can be effective to verify performance of GS breeding by using simulations with reflecting the situation of the breeding populations (e.g., allele frequencies, LD, and other aspects of population structure). If a proper level of LD is considered in simulations, an efficient timing of model updating and a suitable number of markers will be revealed. Simulations using an actual marker data of a breeding population would be one of the most efficient way in the similar way in Chapter 4 to find an efficient breeding scheme.

In Chapters 3, 4, and 5, the importance of maintenance of genetic variation in a breeding population was suggested if breeders hope long-term genetic improvement. There are a number of studies that discussed about the long-term selection in GS breeding. For example, Goddard (2009) and Jannink (2010) tried to keep rare allele in a breeding population by using a modified prediction model. Owing to the ability to keep genetic variation in a whole breeding population, the island-model GS worked well especially in later generations as suggested in Chapter 4. Chapter 3 suggested that occasional selection cycles without pollen control might be a clue for a long-term selection with the same levels of genetic gains in the earlier generations. And, low rate of mis-labeling affected less to the response to selection in Chapter 5 because mis-labeling resulted in weak genetic bottleneck at selection and maintained genetic variation at the high level. The maintenance of genetic diversity competes with rapid genetic improvement in a breeding population. It is, however, necessary to keep genetic variation in order to prevent the breeding population attained an incomplete genetic improvement and fixation, i.e., a local optimum.

In Chapter 6, the field trial of GS breeding was conducted in common buckwheat. This field trial was conducted according to the breeding scheme that was conducted in the simulations of Chapter 3. The field trial showed that GS was effective for genetic improvement of a real common buckwheat population. It was also showed that GS breeding in the empirical field trial worked as expected in the simulations. In the simulation study, I simulated only additive effects for QTL, which may be unrealistic in a real-world breeding. All simulation

studies and the field trial in this dissertation aimed to improve the genetic potential of a breeding population as a whole (i.e., the mean or top values of a breeding population). Thus, the difference of the existence of non-additive effects (i.e., dominance and epistatic effects) in the field trial might not influence largely on the outcomes of GS breeding. If the efficiency of GS were evaluated in the development of F_1 varieties, dominance effect would have a large impact on the outcome. In addition to the mode of inheritance of target traits, there is another discordance between the simulations and the field trial. Common buckwheat is a heteromorphic, self-incompatible species that has heterostylous flowers. Thus, in common buckwheat, mating can occur only between pin-type flowers and thrum-type flowers. This mating system was different from those assumed in simulation studies, i.e., random mating and single-round robin in an allogamous plant species and an autogamous species, respectively. Because the genetic improvement of a target trait was confirmed in the field trial of GS breeding as expected in the simulations, this restriction of mating in common buckwheat seems to have little influence on the outcome of GS breeding. Despite the discordance, the field trial pointed out almost the same factors as pointed out by the suggestions in Chapter 3 (e.g., importance of model updating in a breeding population with low levels of LD, the effect of selection for trait expressed after pollination, and the effect of offseason nursery in GS), suggesting that simulation studies are useful and reasonable to evaluate the performance of breeding strategies on ahead of actual breeding programs. By comparing the results of a field trial with the result of simulations, we may be able to find factors that cause discrepancy from the results expected in the simulations. These factors have potentials to permit more precise simulations, and to improve the gains in actual GS breeding closer to the expected gains in simulations.

All the simulations and field trials in this dissertation showed the difficulty of choosing an efficient breeding scheme and the importance of simulation studies prior to actual trials. In Chapters 3, 4, and 5, I conducted breeding simulations according to each appropriate population and schemes to the plant species, and searched the suitable breeding schemes. Chapter 6 verified that simulations could evaluate an actual efficiency of plant breeding, suggesting that a simulation study is useful for future GS breeding programs. Sun et al. (2011) suggests that breeding simulation can play an important role to decide a breeding scheme. However, most of breeders would not be so familiar to conducting simulations. In Chapter 7, I developed a simple and flexible breeding scheme language to script and simulate breeding schemes. The breeding scheme language might be useful for breeders to decide a breeding scheme. In the future, the systems to connect plant breeders working in the field and researchers are required to

implement plant breeding smoothly and effectively. This breeding scheme language, which was offered in this dissertation, would be one of the clues of such kind of systems.

To meet the high demand of food and biofuel crop production mentioned in Chapter 1, GS is one of promising methods. In this dissertation, I evaluated the efficiency of GS on the basis of simulations and a field trial, and concluded that GS was effective for genetic improvement in plant breeding. However, the studies in this dissertation focused on only narrow range of issues of GS breeding. I feel certain that there are at least three remaining issues of GS breeding: statistical methods, the relationship between a training population and a breeding population, and the existence of GxE. First, through all the studies, I conducted GS within the framework of mass selection and mainly used ridge regression (G-BLUP) as a statistical model for prediction. As I mentioned above, if target of breeding is the development of F_1 varieties, dominance effect would have a large impact on the outcome. As for epistatic effects, Huang et al. (2012) showed that epistasis was important as a principal factor that determined variation for quantitative traits. Thus, we should consider about the impact of non-additive effects and choose a statistical method that can take into account of non-additive effects. RKHS is effective to catch non-additive gene effects (Gianola et al., 2006). Denis and Bouvet (2013) showed the efficiency of a prediction model taking into account of dominance effect. In breeding expecting heterosis, these methods might work better than the methods assuming only additive effects. A prediction model taking into account of non-additive effects would be effective even in breeding expecting the improvement of a population as a whole (i.e., breeding strategies that were used in this dissertation). A model that can predict both additive and non-additive effects can be used for the selection based only on additive effects and may work efficiently to improve additive genetic variation in a breeding population. In the future, it might be necessary to develop the prediction model that can separate additive effects from all estimated effects, and to use that model in GS breeding even when we intend to improve only additive genetic variation in a breeding population. In addition to the mode of inheritance of target traits, there is another factor that might affect the choice of a statistical method. In the studies of this dissertation, the maximum number of markers was 50,000 (in Chapter 6). The recent development of genome technologies enables us to genotype millions of markers in a population. In some situations, in which we can use numerous markers, it might be better to choose LASSO or Bayesian methods with variable selection to prevent the entry of unnecessary markers for prediction. In Chapter 6, because G-BLUP showed the best accuracy among six statistical methods in cross-validation (data not shown), it was chosen as a method for the GS breeding. The other statistical method, however,

may be better than G-BLUP in longer-term selection. It would be meaningful to simulate long-term selection with various statistical methods to choose a suitable statistical method ahead of an actual breeding. Second, in all the studies in the dissertation, the training population was composed of a breeding population or ancestral population of a breeding population. And, my studies underscore the importance of updating a prediction model to catch the change of the genetic structure in a breeding population. In the real-world plant breeding, however, historical data exist and may be used as a training population that is not closely related to a breeding population. In this situation, the gain of GS is thought to be much lower than those expected in my studies, because the reliability of prediction depends on the distance between a training population and a selected population (Rincent et al. 2012). The long-term outcomes in this situation should be evaluated via simulations on ahead of real-world breeding. Third, my studies did not take into account of genotype × environment (GxE) interaction. Most GS methods treat GxE interaction as an error. Recently, the rapid climate change has become a problem. Chen et al. (2011) reported that many organisms are moving to other place with climate change. Iizumi et al. (2013) predicted the loss of food production because of the climate change in the world. If the present GS methods continue to be used in the rapidly changing environment, it cannot catch the environmental change and promote selection to the wrong direction. To fit GS breeding to such situation, it is necessary to take into account of response to environment. Heslot et al. (2014) proposed to involve crop modeling in GS model. They suggested the possibility to predict the performance of candidates according to future weather situations by using their model. To develop the strategies of GS breeding, it is required to use the model for the future performance.

In this dissertation, I conducted the simulations and the field trial. The result of the field trial was coincided with that of the simulations, suggested that the simulation study was appropriate to verify the potential of GS breeding. The success of GS breeding with mass selection in common buckwheat might encourage breeders to try GS breeding in other allogamous species including trees and other perennial species, in which it takes a long time to evaluate the potential of GS. I analyzed mainly the genetic improvement based on phenotypic values in the breeding population to compare the results of the simulations and the field trial. Further analysis of marker genotype data might help us to investigate a system of selection and a heritable structure in a target trait or a target species. The difference of allele frequency and the range under intense selection might produce a little difference of genetic gain and prediction accuracy in a short time, but might produce a large difference in long-term selection. By

conducting this analysis, the discrepancy that was not found in this study may be detected. GS compensates a limitation of MAS (i.e., the inefficiency for improvement of a trait controlled by a number of QTL) by using whole-genome markers, while MAS has an advantage to select several QTL with large effects. The effectiveness of using both information of whole-genome markers and markers that were detected in QTL analysis has been suggested (e.g., Rutkoski et al., 2012). When several genes that explain a large part of trait variation are known, it might be necessary to consider these genes in GS breeding. I believe strongly in the value of the collateral implementation of simulations and an actual breeding in the future. The efficiency of simulation study, which was suggested in this dissertation, means that simulation study can contribute to more detailed and targeted breeding planning. In this study, I conducted the simulations and the field trial separately, and compared the results of them after both studies have been completed. In the future, it might be possible to simulate breeding process at each step of GS breeding. By simulating breeding process on the basis of the current situation of a breeding population, we might be able to choose a suitable selection strategy and select parental genotypes contributing to the next generation with consideration of the expected situation on the basis of the current situation. Through this dissertation, the challenges for the future have been suggested. The studies provide a useful knowledge to improve GS breeding technologies for the current and future breeding.

Acknowledgements

I am deeply grateful to Dr. Hiroyoshi Iwata in Graduate School of Agricultural and Life Sciences of The University of Tokyo for his guidance and encouragement. In the past five years, I have learned a number of things from him not just about research, but social living. He gave me a lot of different experiences, and gave me a lot of opportunities to talk with many researchers in the world. It created an opportunity for me to give much thought to my career and life. When I lost my way, he pointed the way gently. Without his encouragement, this dissertation would not have materialized. Thank you again for his teaching during my student life.

I would like to offer my special thanks to Dr. Jean-Luc Jannink in College of Agriculture and Life Sciences of Cornell University for his warm encouragement. For six months, he taught me patiently even though we had many problems in communication each other. I appreciate that he gave me the chance to study in his laboratory. He also supported for my life in the United States. I am overwhelmed by the kindness that he has shown me.

I would like to express the deepest appreciation to Professor Hirohisa Kishino in Graduate School of Agricultural and Life Sciences of The University of Tokyo for his sincere encouragement since I entered the laboratory. I could learn the way to intellectualize thanks to his guidance.

I would like to express my gratitude to Dr. Hiroshi Omori in Graduate School of Agricultural and Life Sciences of The University of Tokyo. He gave me insightful comments since I entered the laboratory. I am indebt to him.

The materials used in Chapter 4 were contributed by Dr. Masanori Yamasaki in Graduate School of Agricultural Science of Kobe University and Dr. Kaworu Ebana in National Institute

of Agrobiological Science. Their support and suggestion were invaluable. I would like to show my greatest appreciation to them.

I received generous support from Dr. Takeshi Hayashi in National Agriculture and Food Research Organization. His statistical advice was so helpful that I could improve my research, and he gave me many suggestions about life to me. I would like to offer my special thanks to him.

Dr. Jeffrey B. Endelman in University of Wisconsin-Madison made enormous contribution to Chapter 5 by improving his R/package. I would like to thank him for his support.

The plant materials described in Chapter 6 were contributed by Dr. Takashi Hara and Professor Ryo Ohsawa in Graduate School of Life and Environmental Sciences of University of Tsukuba. Without their careful cultivation and assessment, this research would not have completed. They also gave me constructive comments from the viewpoint of plant breeders. I would like to express my deepest gratitude to them.

The genome extraction and formation for genotyping that were described in Chapter 6 were conducted by Ms. Mariko Ueno and Dr. Yasuo Yasui in Graduate School of Agriculture of Kyoto University. I received generous support from them. When I was in difficult time, they always encourage me. I owe my gratitude to them.

The genotyping system described in Chapter 6 was developed and conducted by Dr. Hiroyuki Enoki, Mr. Tatsuro Kimura, and Dr. Satoru Nishimura in TOYOTA MOTOR CORPORATION. Without their persistent support, this project would not have been possible. My heartfelt appreciation goes to them.

My heartfelt appreciation goes to Professor Seishi Ninomiya and Dr. Masaru Fujimoto in Graduate School of Agricultural and Life Sciences of The University of Tokyo. They gave me insightful comments and suggestions on this dissertation.

I have had the support and encouragement of the all member of the Laboratory of Biometry and Bioinformatics. For the six years that I have spent a time in the laboratory, I have shared many

scenes and feelings with them. Without their encouragement, I could not complete this dissertation. I would like to express the deepest appreciation to them. I appreciate to Dr. Hiromi Kanegae and Dr. Mai Minamikawa for their valuable comments on the manuscript.

I would like to show my greatest appreciation to the member of Dr. Jean-Luc Jannink's laboratory and Dr. Mark E. Sorrells's laboratory. They extended to me their hospitality when I visited them. Advice and comments given by them has been great help in my research.

References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta and T.J. Lawlor. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genomic evaluation of Holstein final score. Journal of Dairy Science 93: 743 – 752.

Albrecht, T., V. Wimmer, H.J. Auinger, M. Erbe, C. Knaak, M. Ouzunova, H. Simianer and C.C. Schön. 2011. Genome-based prediction of testcross values in maize. Theoretical and Applied Genetics 123: 339 – 350.

Asoro, F.G., M.A. Newell, W.D. Beavis, M.P. Scott, N.A. Tinker and J.L. Jannink. 2013. Genomic, marker-assisted, and predigree-BLUP selection methods for β-glucan concentration in elite oat. Crop Science 53: 1894 – 1906.

Auzanneau, J., C. Huyghe, B. Julier, and P. Barre. 2007. Linkage disequilibrium in synthetic varieties of perennial ryegrass. Theoretical and Applied Genetics 115: 837 – 847.

Basten, C.J., B.S. Weir and Z.B. Zeng. 2003. QTL Cartographer, version 1.17. A reference manual and tutorial for QTL mapping. Department of Statistics, North Carolina State University, Raleigh, NC 187.

Bernardo, R. 2001. What if we know all the genes for a quantitative trait in hybrid crops? Crop Science 41: 1 – 4.

Bernardo, R. and J. Yu. 2007. Prospects for genomewide selection for quantitative traits in maize. Crop Science 47: 1082 – 1090.

Bernardo, R. 2009. Genomewide selection for rapid introgression of exotic germplasm in maize. Crop Science 49: 419 – 425.

Bos, I. and P. Caligari. 2008. Selection methods in plant breeding, second ed. Springer, Dordrecht, The Netherlands, pp.104 – 106 & pp.341 – 343.

Brennan, J.P. 1989. An analysis of the economic potential of some innovations in a wheat breeding programme. Australian Journal of Agricultural Economics 33: 48 – 55.

Brown, J. and P. Caligari. 2008. An introduction to plant breeding. Blackwell Publishing Ltd, pp.1 – 3, p.11 & pp.34 – 41.

Bulmer, M.G. 1980. The mathematical theory of quantitative genetics. Clarendon Press, pp.144

– 147.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. de Roos and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. Genetics 178: 553 – 561.

Ceballos, H., C.A. Iglesias, J.C. Pérez and Alfred G.O.D. 2004. Cassava breeding: opportunities and challenges. Plant Molecular Biology 56: 503 – 516.

Chen, I.C., J.K. Hill, R. Ohlemüller, D.B. Roy and C.D. Thomas. 2011. Rapid range shifts of species associated with high levels of climate warming. Science 333: 1024 – 1026.

Chen, C.Y., I. Misztal, I. Aguilar, S. Tsuruta, T.H.E. Meuwissen, S.E. Aggrey, T. Wing and W.M. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: an example using broiler chickens. Journal of Animal Science 89: 23 – 28.

Cheng, Z., G.G. Presting, C. R. Buell, R. A. Wing and J. Jiang. 2001. High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics 157: 1749 – 1757.

Collard, B.C.Y. and D.J. Mackill. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philosophical Transactions of the Royal Society B 363: 557 – 572.

Connolly, V. 2001. Breeding improved varieties of perennial ryegrass. Taegasc, Crops Research Centre.

Crossa,, J., G. de los Campos, P. Pérez, D. Gianola, J. Burgueño, J.L. Araus, D. Makumbi, R.P. Singh, S. Dreisigacker, J. Yan, V. Arief, M. Banziger and H.J. Braun. 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186: 713 – 724.

Daetwyler, H.D., B. Villanueva and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. PloS One 3: e3395.

Daetwyler, H.D., R. Pong-Wong, B. Villanueva and J.A. Woolliams. 2013. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185: 1021 – 1031.

de los Campos, G, D. Gianola, G.J.M. Rosa, K.A. Weigel and J. Crossa. 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92: 295 – 308.

de Oliveira, E.J., M.D.V. de Resende, V.S. Santos, C.F. Ferreira, G.A.F. Oliveira, M.S. da Silva, L.A. de Oliveira, C.I. Aguilar-Vildoso. 2012. Genome-wide selection in cassava.

Euphytica 187: 263 – 276.

Denis, M. and J.M. Bouvet. 2013. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. Tree Genetics & Genomes 9: 37 – 51.

Desta, Z.A. and R. Ortiz. 2014. Genomic selection: genome-wide prediction in plant improvement. Trends in Plant Science 19: 592 – 601.

Edwards, M.D. and N.J. Page. 1994. Evaluation of marker-assisted selection through computer simulation. Theoretical and Applied Genetics 88: 376 – 382.

Elshire, R.J., J.C. Glaubitz, Q. Sun, J.A. Poland, K. Kawamoto, E.S.Buckler, and S.E. Mitchell. 2011. A robust, simple genotyping-bysequencing (GBS) approach for high diversity species. PloS One 6:e19379.

Endelman, J.B. 2011. Ridge regression and other kernels for genomic selection with R package rrBLUP. The Plant Genome 4: 250 – 255.

Endelman, J.B., G.N. Atlin, Y. Beyene, K. Semagn, X. Zhang, M.E.Sorrells and J.L. Jannink. 2014. Optimal design of preliminary yield trials with genome-wide markers. Crop Science 54: 48 – 59.

Enoki, H., S. Nishimura and A. Murakami. 2012. Method for designing probe in DNA microarray, and DNA microarray provided with probe designed thereby. United States Patent Application Publication; US2012190582 (2012.7.26).

Falconer, D.S. and T.F.C. Mackay. 1996. Introduction to Quantitative Genetics, 4th edn. Longman, Essex, UK, pp.125 – 131, pp.185 – 194 & pp.264 – 269.

Fan, J. and J. Lv. 2010. A selective overview of variable selection in high dimensional feature space. Statistica Sinnica 20: 101 – 148.

Fiil, A., I. Lenk, K. Petersen, C.S. Jensen, K.K. Nielsen, B. Schejbel, J.R. Andersen, and T. Lubberstedt. 2011. Nucleotide diversity and linkage disequilibrium of nine genes with putative effects on flowering time in perennial ryegrass (*Lolium perenne* L.). Plant Science 180: 228 – 237.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler IV. 2003. Structure of linkage disequilibrium in plants. Annual Review of Plant Biology 54: 357 – 374.

Food and Agriculture Organization of the United Nation Statistics Division (FAOSTAT). 2014. URL: http://faostat3.fao.org/faostat-gateway/go/to/home/E.

Friedman, J., T. Hastie, and R. Tibshirani. 2010. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software 33: 1 – 22.

Fujimaki, H. 1979. Recurrent selection by using genetic male sterility for rice improvement. JARQ (Tsukuba) 13: 153 – 156.

Gaskell, G., N. Allum, W. Wagner, N. Kronberger, H. Torgersen, J. Hampel and J. Bardes. 2004. GM foods and the misperception of risk perception. Risk analysis 24: 185 – 194.

Gianola, D., R.L. Fernando and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173: 1761 – 1776.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximization of long term response. Genetica 136: 245 – 257.

Grattapaglia, D. and R. Sederoff. 1994. Genetic linkage maps of Eucalyptus grandis and Eucalyptus urophylla using a pseudo-testcross: mapping strategy and RAPD markers. Genetics 137: 1121 − 1137.

Grattapaglia, D. and M.D.V. Resende. 2011. Genomic selection in forest tree breeding. Tree Genetics & Genomics 7.2: 241 – 255.

Gupta, P.K., S. Rustgi, and P.L. Kulwal. 2005. Linkage disequilibrium and association studies in higher plants: Present status and future prospects. Plant Mol Biol 57: 461 – 485.

Habier, D., R.L. Fernando, K. Kizilkaya and D.J. Garrick. 2011. Extension of the Bayesian alphabet for genomic selection. BMC Bioinformatics 12: 186.

Hallauer, A.R. and J.B. Miranda Filho. 1981. Quantitative genetics in maize breeding. Iowa State University Press Ames, pp. 168 – 170.

Hamblin, M.T., E.S. Buckler and J.L. Jannink. 2011. Population genetics of genomics-based crop improvement methods. Trends in Genetics 27: 98 – 106.

Hanson, W.D. 1959. The breakup of initial linkage blocks under selected mating systems. Genetics 44: 857 – 868.

Hara, T., H. Iwata, K. Okuno, K. Matsui and R. Ohsawa. 2011. QTL analysis of photoperiod sensitivity in common buckwheat by using markers for expressed sequence tags and photoperiod-sensitivity candidate genes. Breeding Science 61: 394−404.

Hartl, D.L. and A.G. Clark. 2007. Principles of population genetics, 4th ed. Sunderland, Massachusetts: Sinauer Associates, Inc. Publishers, pp.73 – 87 & pp. 295 – 311.

Harushima, Y., M. Yano, A. Shomura, M. Sato, T. Shimano, Y. Kuboki, T. Yamamoto, S.Y. Lin, B.A. Antonio, A. Parco, H. Kajiya, N. Huang, K. Yamamoto, Y. Nagamura, N. Kurata, G.S. Khush and T. Sasaki. 1998. A high-density rice genetic linkage map with 2275 marker using a single F_2 population. Genetics 148: 479 – 494.

Hayashi, T. and H. Iwata. 2011. EM algorithm for Bayesian estimation of genomic breeding

values. BMC Genetics 11: 1 – 9.

Hayes, B. and M.E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33: 209 – 229.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain and M.E. Goddard. 2009. Genomic selection in dairy cattle: progress and challenges. Journal of Dairy Science 92: 433 – 443.

Hayes, B.J., N.O.I. Cogan, L.W. Pembleton, M.E. Goddard, J. Wang, G.C. Spangenberg and J.W. Forster. 2013. Prospects for genomic selection in forage plant species. Plant Breeding 132: 133 – 143.

Heffner, E.L., J.L. Jannink and M.E. Sorrells. 2011. Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4: 65 – 75.

Heffner, E.L., A.J. Lorenz, J.L. Jannink and M.E.Sorrells. 2010. Plant breeding with genomic selection: gain per unit time and cost. Crop Science 50: 1 – 10.

Heffner, E.L., M.E. Sorrells and J.L. Jannink. 2009. Genomic selection for crop improvement. Crop Science 49: 1 – 12.

Heslot, N., D. Akdemir, M.E. Sorrells and J.L. Jannink. 2014. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. Theoretical and Applied Genetics 127: 463 – 480.

Heslot, N., H.P. Yang, M.E. Sorrells and J.L. Jannink. 2012. Genomic selection in plant breeding: a comparison of models. Crop Science 52: 146 – 160.

Hickey, J.M., S. Dreisigacker, J. Crossa, S. Hearne, R. Babu, B.M. Presanna, M. Grondona, A. Zambelli, V.S. Windhausen, K. Mathews and G. Gorjanc. 2014. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. Crop Science 54: 1476 – 1488.

Hill, W.G. and B.S. Weir. 1988. Variances and covariances of squared linkage disequilibria in finite population. Theoretical Population Biology 33: 54 – 78.

Hoerl, A.E. and R.W. Kennard. 1970. Ridge regression: applications to nonorthogonal problems. Technometrics 12: 55 – 67.

Hoeschele, I. and P.M. VanRaden. 1993. Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theoretical and Applied Genetics 85: 953 – 960.

Huang, W., S. Richardsb, M.A. Carbonea, D. Zhub, R.R.H. Anholtc, J.F. Ayrolesa, L. Duncana, K.W. Jordana, F. Lawrencea, M.M. Magwirea, C.B. Warnerb, K. Blankenburgb, Y. Hanb,

M. Javaidb, J. Jayaseelanb, S.N. Jhangianib, D. Muznyb, F. Ongerib, L. Peralesb, Y.Q. Wub, Y. Zhangb, X. Zoub, E.A. Stonea, R.A. Gibbsb, and T.F.C. Mackay. 2012. Epistasis dominatesthe genetic architecture of *Dorosophila* quantitative traits. Proceeding of the National Academy of Sciences 109: 15553 – 15559.

Iehisa, J. C. M., R. Ohno, T. Kimura, H. Enoki, S. Nishimura, Y. Okamoto, S. Nasuda and S. Takumi. 2014. A high-density genetic map with array-based markers facilitate structural and quantitative trait locus analyses of common wheat genome. DNA Research: dsu020.

Iizumi I., H. Sakuma, M. Yokozawa, J.J. Luo, A.J. Challinor, M.E. Brown, G. Sakurai and T. Yamagata. 2013. Prediction of seasonal climate-induced variations in global food production. Nature Climate Change 3: 904 – 908.

Isobe, S., R. Kolliker, H. Hisano, S. Sasamoto, T. Wada, I. Klimenko, K. Okumura, and S. Tabata. 2009. Construction of a consensus linkage map for red clover (*Trifolium pratense* L.). BMC Plant Biology 9: 57.

Ivkovich, M. and M. Koshy. 2002. Optimization of multiple trait selection in western hemlock (*Tsuga heterophylla* (Raf.) Sarg.) including pulp and paper properties. Annals of Forest Science 59: 577 – 582.

Iwata, H. and S. Ninomiya. 2006. AntMap: constructing genetic linkage maps using an ant colony optimization algorithm. Breeding Science 56: 371−377.

Iwata, H., T. Hayashi and Y. Tsumura. 2011. Prospects for genomic selection in conifer breeding: a simulation study of *Cryptomeria japonica*. Tree Genetics & Genomics 7: 747–758.

Iwata, H. and J.L. Jannink. 2011. Accuracy of genomic selection prediction in Barley breeding programs: a simulation study based on the real single nucleotide polymorphism data of barley breeding lines. Crop Science 51: 1915 – 1927.

Jannink, J.L. 2010. Dynamics of long-term genomic selection. Genetics Selection Evolution 42: 35.

Jannink, J.L., A.J. Lorenz and H. Iwata. 2010. Genomic selection in plant breeding: from theory to practice. Briefings in Functional Genomics 9: 166 – 177.

Jonas, E. and D.J. de Koning. 2013. Does genomic selection have a future in plant breeding? Trends in Biotechnology 31: 497 – 504.

Kang, H.M., N.A. Zaitlen, C.M. Wade, A. Kirby, D. Heckerman, M.J. Daly, and E. Eskin. 2008. Efficient control of population structure in model organism association mapping. Genetics 178: 1709 – 1723.

Kearsey, M.J. and A.G.L. Farquhar.1998. QTL analysis in plants; where are we now? Heredity 80: 137 – 142.

Konishi, T. and O. Ohnishi. 2006. A linkage map for common buckwheat based on microsatellite and AFLP markers. Fagopyrum 23: 1−6.

Kosambi, D. D. 1943. The estimation of map distances from recombination values. Annals of Eugenics 12: 172 − 175.

Kumar, S., D. Chagné, M.C.A.M. Bink, R.K. Volz, C.Whitworth, C. Carlisle. 2012. Genomic selection for fruit quality traits in apple (*Malus* × *domestica* Borkh.). PloS One 7: e36674.

Kurata, N., Y. Umehara, H. Tanoue and T. Sasaki. 1994. Physical mapping of the rice genome with YAC clones. Plant Molecular Biology 35: 101 – 113.

Lande, R. and R. Thompson. 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124: 743–756.

Legarra, A., C. Robert-Granié, E. Manfredi and J.M. Elsen. 2008. Performance of genomic selection in mice. Genetics 180: 611 – 618.

Lewis, D. and D.A. Jones. 1992. The genetics of heterostyly. *In*: Barrett, S. C. H. (ed.) Evolution and function of heterostyly. Springer, Berlin. pp.129–150.

Li, Y., Y. Li, S. Wu, K. Han, Z. Wang, W. Hou, Y. Zeng and R. Wu. 2007. Estimation of multilocus linkage disequilibria in diploid populations with dominant markers. Genetics 176: 1811 – 1821.

Liang, L., S. Zollner and G.R. Abecasis. 2007. GENOME: a rapid coalescent-based whole genome simulator. Bioinformatics 23: 1565 – 1567.

Lin, Z., B.J. Hayes and H.D. Daetwyler. 2014. Genomic selection in crops, trees and forages: a review. Crop and Pasture Science 65: 1177 – 1191.

Lorenz, A.J. 2013. Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: asimulation experiment. G3 3: 481 – 491.

Lorenz, A.J., S. Chao, F.G. Asoro, E.L. Heffner, T. Hayashi, H. Iwata, K.P. Smith, M.E. Sorrells and J.L. Jannink. 2011. Genomic selection in plant breeding: knowledge an prospects. Advances in agronomy 110: 77.

Luan, T., J.A. Woolliams, S. Lien, M. Kent, M. Svendsen and T.H.E. Meuwissen. 2009. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 183: 1119 – 1126.

Ly, D., M. Hamblin, I. Rabbi, G. Melaku, M. Bakare, H.G. Gauch Jr., R. Okechukwu, A.G.O. Dixon, P. Kulakow and J.L. Jannink. 2013. Relatedness and genotype × environment

interaction affect prediction accuracies in genomic selection: a study in cassava. Crop Science 53: 1312 – 1325.

MacClosky, B., J. LaCombe and S.D. Tanksley. 2013. Selfing for the design of genomic selection experiments in biparental plant populations. Theoretical and Applied Genetics 126: 2907 – 2920.

Massman, J.M., H.J.G. Jung and R. Bernardo. 2013. Genomewide selection versus marker-assisted recurrent selection to improve gain yield and stover-quality traits for cellulosic ethanol in maize. Crop Science 53: 58 – 66.

Mayor, P. and R. Bernardo. 2009. Genomewide selection and marker-assisted recurrent selection in doubled haploid versus F_2 population. Crop Science 49: 1719 – 1725.

Melchinger, A.E., H.F. Utz, and C.C. Schon. 1998. Quantitative trait locus (QTL) mapping using different testers and independent population samples in maize reveals low power of QTL detection and large bias in estimates of QTL effects. Genetics 149: 383– 403.

Meuwissen, T.H.E. 1997. Maximizing the response of selection with a predefined rate of inbreeding. Journal of Animal Science 75: 934 – 940.

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

Mevik, B.H. and R. Wehrens. 2007. The pls package: principal component and partial least squares regression in R. Journal of Statistical Software 18: 1 – 24.

Misztal, I., A. Legarra and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92: 4648 – 4655.

Morris, M., K. Dreher, J.M. Ribaut, and M. Khairallah. 2003. Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using conventional and marker-assisted selection. Molecular Breeding 11: 235 – 247.

Moser, G., B. Tier, R.E. Crump, M.S. Khatkar and H.W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution 41: 56.

Nagasaki, H., K. Ebana, T. Shibaya, J. Yonemaru and M. Yano. 2010. Core single-nucleotide polymorphisms – a tool for genetic analysis of the Japanese rice population. Breeding Science 60: 648 – 655.

Nakaya, A. and S.N. Isobe. 2012. Will genomic selection be a practical method for plant breeding? Annals of Botany: mcs109.

O'Hagan, S., L. Knowles and D.B. Kell. 2012. Exploiting genomic knowledge in optimizing molecular breeding programmes: algorisms from evolutionary computing. PloS One 7: e48862.

Pan, S. J. and Q. F. Chen. 2010. Genetic mapping of common buckwheat using DNA, protein and morphological markers. Hereditas 147: 27−33.

Pandey, S. and S. Rajatasereekul. 1999. Economics of plant breeding: The value of shorter breeding cycles for rice in Northeast Thailand. Field Crops Research 64: 187 – 197.

Park, T. and G. Casella. 2008. The Bayesian lasso. Journal of the American Statistical Association. 103: 681 – 686.

Piepho, H.P. 2009. Ridge regression and extensions for genomewide selection in maize. Crop Science 49: 1165 – 1176.

Philipsen, N.C. 2011. The criminalization of mistakes in nursing. Journal of Nurse Practitioners 7.9: 719 – 726.

Phillips, R.L. 2010. Mobilizing science to break yield barriers. Crop Science 50: S_99 – S_108.

R Development Core Team (2014) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Rafalski, A. and M. Morgante. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics 20: 10 – 111.

Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman and E.S. Buckler IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proceeding of the National Academy of Sciences 98: 11479 – 11484.

Resende, R.M.S., M.D. Casler and M.D.V. de Resende. 2014. Genomic selection in forage breeding: accuracy and methods. Crop Science 54: 143 – 156.

Resende, M.D.V., M.F.R. Resende Jr, C.P. Sansaloni, C.D. Petroli, A.A. Missiaggia, A.M. Aguiar, J.M. Abad, E.K. Takahashi, A.M. Rosado, D.A. Faria, G.J. Pappas Jr., A. Kilian and D. Grattapaglia. 2012. Genomic selection for growth and wood quality in *Eucalyptus*: capturing the missing heritability and accelerating breeding for complex traits in forest trees. New Phytologist. 194: 116 – 128.

Resende Jr, M.F.R, P. Muñoz, J.J. Acosta, G.F. Peter, J.M. Davis, D. Grattapaglia, M. D. V. Resende and M. Kirst. 2012. Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytologist 193: 617 – 624.

Ribaut, J.M., C. Jiang, D. Gonzalez-de-Leon, G.O. Edmeades and D.A. Hoisington. 1997. Identification of quantitative trait loci under draught conditions in tropical maize. 2. Yield components and marker-assisted selection strategies. Theoretical and Applied Genetics 94: 887 – 896.

Riedelsheimer, C., A. Czedik-Eysenberg, C. Grieder, J. Lisec, F. Technow, R. Sulpice, T. Altmann, M. Stitt, L. Willmitzer and A.E. Melchinger. 2012. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nature Genetics 44: 217 – 220.

Riedelsheimer, C. and A.E. Melchinger. 2013. Optimizating the allocation of resources for genomic selection in one breeding cycle. Theoretical and Applied Genetics 126: 2835 – 2848.

Rincent, R., D. Laloë, S. Nicolas, T. Altmann,†D. Brunel, P. Revilla, V. M. Rodríguez, J. Moreno-Gonzalez, A. Melchinger, E. Bauer, C-C. Schoen, N. Meyer, C. Giauffret, C. Bauland, P. Jamin, J. Laborde, H. Monod, P. Flament, A. Charcosset and L. Moreau. 2012. Maximizing the reliability of genomic selection by optimizing the calibration set of reference individuals: comparison of methods in two diverse groups of maize inbreds (*Zea mays* L.). Genetics 192: 715-728.

Ritter, E., C. Gebhardt and F. Salamini. 1990. Estimation of recombination frequencies and construction of RFLP linkage maps in plants from crosses between heterozygous parents. Genetics 125: 645−654.

RStudio, Inc. 2014. Web application framework for R. http://shiny.rstudio.com.

Rutkoski, J.E., J. Benson, Y. Jia, G. Brown-Guedira, J.L. Jannink and M.E. Sorrells. 2012. Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. The Plant Genome 5: 51 – 61.

Rutkoski, J.E., E.L. Heffner and M.E. Sorrells. 2011. Genomic selection for durable stem rust resistance in wheat. Euphytica 179: 161 – 173.

Scheet, P. and M. Stephens. 2006. A fast and flexible statistical model for large-scale population genotype data: application to inferring missing genotypes and haplotypic phase. The American Journal of Human Genetics 78: 629 – 644.

Servin, B., O. C. Martin, M. Mézard and F. Hospital. 2004. Toward a theory of marker-assisted gene pyramiding. Genetics 168: 513 – 523.

Sonesson, A.K., J.A. Woolliams and T.H.E. Meuwissen. 2012. Genomic selection requires genomic control of inbreeding. Genetics Selection Evolution 44: 27.

Sorensen, D. and D. Gianola. 2002. Likelihood, Bayesian, and MCMC methods in quantitative

genetics. Springer, pp.41 – 53.

Strauss, S.H., R. Lande, and G. Namkoong. 1992. Limitations of molecular-marker-aided selection in forest tree breeding. Canadian Journal of Forest Research 22:1050 – 1061.

Sun, X., T. Peng and R.H. Mumm. 2011. The role and basics of computer simulation in support of critical decisions in plant breeding. Molecular Breeding 28: 421 – 436.

Sved, J.A. 1971. Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theoretical Population Biology 2: 125 – 141.

Tester, M. and P. Langridge. 2010. Breeding technologies to increase crop production in a changing world. Science 327: 818 – 822.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B 58: 267 – 288.

Tilman, D., R. Socolow, J.A. Foley, J. Hill, E. Larson, L. Lynd, S. Pacala, J. Reilly, T. Searchinger, C. Somerville and R. Williams. 2009. Beneficial biofuels – the food, energy, and environment trilemma. Science 325: 270 – 271.

Trestle Technology, LLC. 2013. Ace editor bindings to enable a rich text editing environment within Shiny. http://cran.r-project.org/web/packages=shinyAce.

Tweeten, L. and S.R. Thompson. 2008. Long-term agricultural output supply-demand balance and real farm and food prices. Working Paper AEDE-WP 0044-08, Ohio State University, Columbus, OH.

Ukai, Y. 2000. Geneticanalysis at the genomic level: map and QTL (ゲノムレベルの遺伝解析：MAP と QTL). University of Tokyo Press, Tokyo, JP. (in Japanese), pp.310 – 312.

Ukai, Y. 2003. Plant breeding (植物育種学：交雑から遺伝子組み換えまで). University of Tokyo Press, Tokyo, JP. (in Japanese), pp.94 – 144, pp.178 – 189 & pp.199 – 203.

Urbanek, S. 2013. Package 'parallel'. https://stat.ethz.ch/R-manual/R-devel/library/parallel/doc/parallel.pdf.

VanRaden. 2008. Efficient methods to compute genomic predictions. Journal of Dairy Science 91: 4414 – 4423.

VanRaden, P.M, C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92: 16 – 24.

Verhoeven, K.J.F., J.L. Jannink and L.M. McIntyre. 2006. Using mating designs to uncover QTL and the genetic architecture of complex traits. Heredity 96: 139 – 149.

Walter, A., B. Studer, and R. Kölliker. 2012. Advanced phenotyping offers opportunities for

improved breeding of forage and turf species. Annals of Botany 110: 1271 – 1279.

Wang, J., M. van Ginkel, D. Podlich, G. Ye, R. Trethowan, W. Pfeiffer, I.H. DeLacy, M. Cooper and S. Rajaram. 2003. Comparison of two breeding strategies by computer simulation. Crop Science 43: 1764 – 1773.

Whitley, D., S. Rana, R.B. Heckendorn. 1999. The island model genetic algorithm: on separability, population size and convergence. J Computing and Information Technology 7:33 – 47.

Wilkins, P.W. 1991. Breeding perennial ryegrass for agriculture. Euphytica 52: 201 – 214.

Weir, B.S. and W.G. Hill. 1986. Nonuniform recombination within the human β-globin gene cluster. The American Journal of Human Genetics 38: 776 – 778.

Wong, C.K. and R. Bernardo. 2008. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. Theoretical and Applied Genetics 116: 815 – 824.

Wright, S. 1932. The roles of mutation, inbreeding, crossbreeding and selection in evolution. In Jones DF, editor, Proceedings of the Sixth International Conference of Genetics. Brooklyn Botanic Garden: 356 – 366.

Wright, S. 1943. Isolation by distance. Genetics 28: 114 – 138.

Yamamoto, E., H. Iwata, T. Tanabata, R. Mizobuchi, J. Yonemaru, T. Yamamoto and M. Yano. 2014. Effect of advanced intercrossing on genome structure and on the power to detect linked quantitative trait loci in a multi-parent population: a simulation study in rice. BMC Genetics 15: 50.

Yamamoto, T., H. Nagasaki, J. Yonemaru, K. Ebana, M. Nakajima, T. Shibaya and M. Yano. 2010. Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics 11: 267.

Yamasaki, M. and O. Ideta. 2013. Population structure in Japanese rice population. Breeding Science 63: 49 – 57.

Yano, K., T. Ishii, H. Ikehashi and K. Yonezawa. 2000. Optimization of the number of cycles and intensity of selection, and population size in mass selection: selection for single traits in outcrossing plant populations. Breeding Science 50: 37 – 43.

Yasui, Y., Y. Wang, O. Ohnishi and C. G. Campbell. 2004. Amplified fragment length polymorphism linkage analysis of common buckwheat (*Fagopyrum esculentum*) and its wild self-pollinated relative Fagopyrum homotropicum. Genome 47: 345−351.

Zhao, Y., M. Gowda, W. Liu, T. Würschum, H.P. Maurer, F.H. Longin, N. Ranc, J.C. Reif.

2012. Accuracy of genomic selection in European maize elite breeding populations. Theoretical and Applied Genetics 124: 769 – 776.

Zeng, Z.B. 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proceeding of the National Academy of Sciences 90: 10972 − 10976.

Zeng, Z.B. 1994. Precision mapping of quantitative trait loci. Genetics 136: 1457 − 1468.

Zou, H. and T. Hastie. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society. Series B 67: 301 – 320.