Aus dem Institut für
Pflanzenbau und Pflanzenzüchtung II
der Justus-Liebig-Universität Gießen
Professur für Biometrie und Populationsgenetik
Prof. Dr. Matthias Frisch

Computer simulations to optimize the design of marker-assisted backcrossing for high-throughput marker systems

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften
im Fachbereich
Agrarwissenschaften, Ökotrophologie und Umweltmanagement
Justus-Liebig-Universität Gießen

von
Eva Herzog
aus Aschaffenburg

Gießen, 28. Februar 2014

Contents

[1]Herzog, E, & Frisch, M. 2011. Selection strategies for marker-assisted backcrossing with high-throughput marker systems. *Theor Appl Genet*, **123**(2), 251-260.

[2]Herzog, E, & Frisch, M. 2013. Efficient marker-assisted backcross conversion of seed-parent lines to cytoplasmic male sterility. *Plant Breeding*, **132**(1), 33-41.

[3]Herzog, E, Töpfer, R, Hausmann, L, Eibach, R, & Frisch, M. 2013. Selection strategies for marker-assisted background selection with chromosomewise SSR multiplexes in pseudo-backcross programs for grapevine breeding. *Vitis*, **52**(4), 193-196.

[4]Herzog, E, Falke, KC, Presterl, T, Scheuermann, D, Ouzunova, M, & Frisch, M. 2014. Selection strategies for the development of maize introgression populations. *PLOS ONE*, **9**(3), e92429.

Abbreviations

AFLP	amplified fragment length polymorphism
cM	centimorgan
CMS	cytoplasmic male sterility
DArT	diversity array technology
DH	doubled haploid
DNA	desoxyribonucleic acid
HT	high-throughput
JKI	Julius Kühn Institute
KASP	Competitive Allele Specific PCR
MABC	marker-assisted backcrossing
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism
SM	single marker
SNP	single nucleotide polymorphism
SSR	simple sequence repeat

Chapter 1

General introduction

Marker-assisted backcrossing in plant breeding

Marker-assisted backcrossing (MABC) is one of the most successful applications of DNA markers in plant breeding. It is now routinely applied in gene and transgene introgression, seed parent line conversion to cytoplasmic male sterility (CMS) and the development of introgression populations for QTL detection and pre-breeding (Semagn *et al.*, 2006; Xu & Crouch, 2008). A typical application in public and commercial plant breeding is the introgression of single or multiple resistance genes to biotic or abiotic stresses, *e.g.*, in the major cereals maize (Willcox *et al.*, 2002), rice (Datta *et al.*, 2002; Neeraja *et al.*, 2007) and wheat (Liu *et al.*, 2000; Wilde *et al.*, 2008). The importance of MABC is underlined by the fact that in 2013 over 90% of the total acreage of maize, soybean and cotton in the U.S. was planted with varieties that were developed with breeding schemes using MABC for trait introgression (National Agricultural Statistics Service, 2013).

A typical backcross scheme for the introgression of a dominant target allele from a donor into the genome of a recipient line is shown in Figure 1.1. At the target locus, the donor parent P_1 carries the target allele Q in homozygous state. The recipient P_2 carries the allele q in homozygous state. The donor is crossed with recipient to create a heterozygous F_1 population

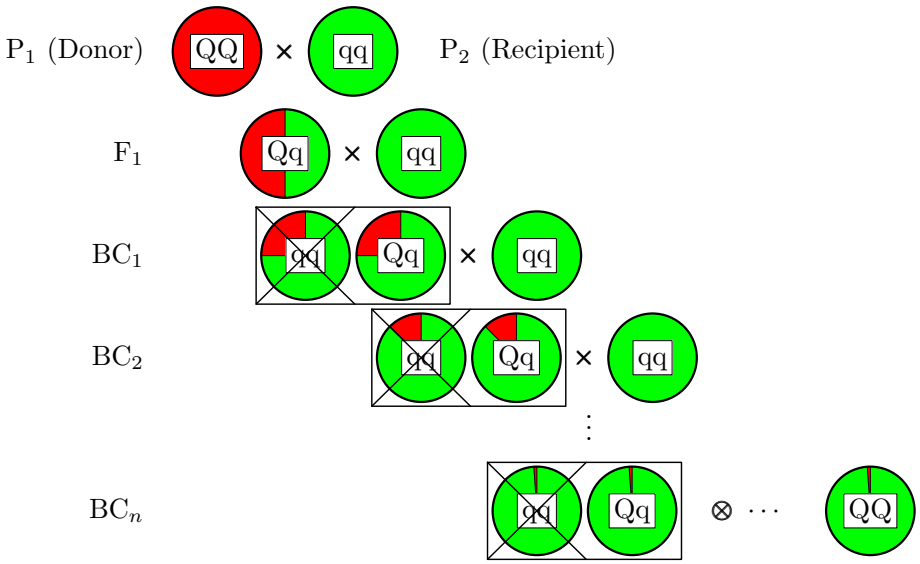

Figure 1.1. Schematic representation of a backcross program for gene introgression with n backcross generations. The donor parent P_1 carries the target allele Q at the target locus. The recipient parent P_2 carries the allele q at the target locus. The genome of the donor parent P_1 is displayed in red. The genome of the recipient parent P_2 is displayed in green. Modified from Becker (2011, p. 198f).

with genotype Qq at the target locus. The F_1 is backcrossed to the recipient to create a BC_1 population. From this BC_1 population, heterozygous carriers of the target allele with genotype Qq are again backcrossed to the recipient, while BC_1 individuals with genotype qq are discarded. This process is repeated for n backcross generations. To obtain a homozygous carrier of the target allele with genotype QQ, this process is followed by one or several generations of selfing. Individuals carrying the target allele Q can be selected with markers linked to or located in the target gene. This process is called foreground selection (Hospital & Charcosset, 1997).

Beside foreground selection, fast and complete recovery of the genotype of the recipient is the major objective of MABC. Without selection for the recipient genome (in Figure 1.1 displayed in green), the donor genome proportion (in Figure 1.1 displayed in red) is per expectation reduced by 50% in every backcross generation. The average recipient genome proportion in

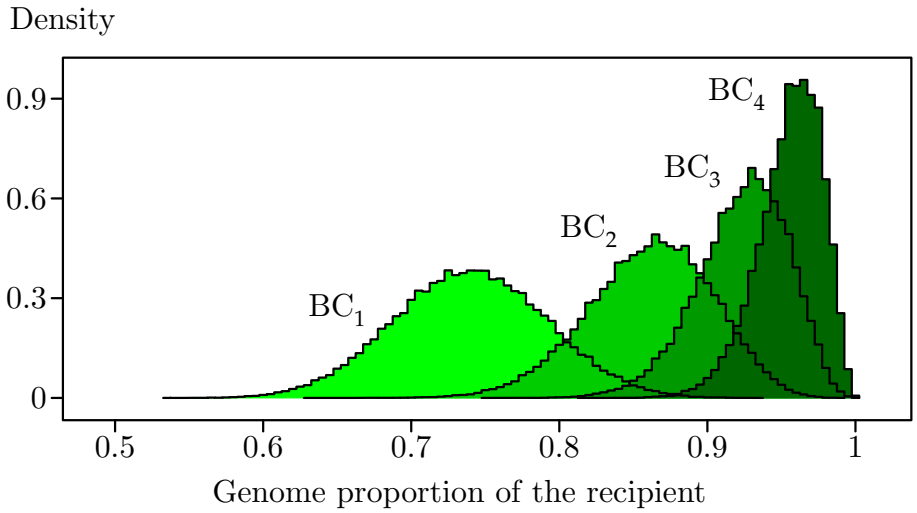

Figure 1.2. Distribution of the recipient genome proportion in backcross populations of generations BC_1-BC_4. Data generated with simulations based on a published linkage map of maize (Schön *et al.*, 1994).

generation n is thus $(2^{n+1} - 1)/2^{n+1}$. For example, in generation BC_1 of a gene introgression program in maize, the average recipient genome proportion is 75% (Figure 1.2). However, the actual recipient genome proportion of the individuals in the BC_1 population ranges around this expected value from about 60% to 90%. The possibility to select individuals from the upper tail of the distribution, having a recipient genome proportion of about 90%, leads to considerable gains in recipient genome in generation BC_2 compared to no selection.

The actual recipient genome proportion of a backcross individual can be estimated by genotyping background markers which cover the entire genome and allow to distinguish between alleles from the donor and the recipient (Figure 1.3). The individuals with the highest proportion of the green recipient alleles at the background marker loci will be selected as non-recurrent parents for the following backcross generation. This process was described by Tanksley et al. (1989) and is referred to as background selection (Hospital & Charcosset, 1997). Background selection has the potential to speed up the restoration of the recipient genotype by several generations.

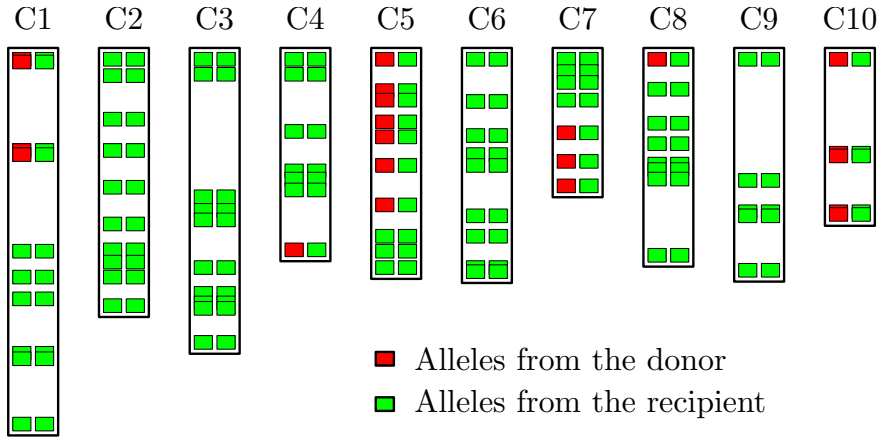

Figure 1.3. Graphical genotype of a diploid backcross plant with ten chromosomes (C1-C10) derived from simulations with a published linkage map of maize (Schön *et al.*, 1994). The donor alleles are displayed in red. The recipient alleles are displayed in green.

Since the advent of molecular markers in the 1980's, it has constantly been hypothesized that new developments in marker technology will improve the speed and efficiency of marker-assisted selection (Collard *et al.*, 2005; Ragot & Lee, 2007). However, during the 1990's, the large-scale implementation of marker-assisted background selection in breeding programs proceeded slowly and to a much lower extent than expected. The reason was that the effort in the laboratory was very high with the then available marker systems. Moreover, the analysis of a high number of molecular markers was very expensive.

Different types of molecular markers

During the 1980's, restriction fragment length polymorphisms (RFLPs) emerged as the first system of DNA markers that was suitable for widespread use in genetic studies (Botstein *et al.*, 1980; Beckmann & Soller, 1986). During the 1990's, amplified fragment length polymorphism (AFLP) and simple sequence repeat (SSR) markers began to replace RFLPs as the markers of choice in plant breeding (Vos *et al.*, 1995; Zietkiewicz *et al.*, 1994). SSR markers have up to now been very useful as they are abundant in the genome,

highly informative and transferable between crop species and their wild relatives. Nevertheless, the described marker types generally provide information only about one locus per assay (Collard & Mackill, 2008). This type of marker assay is therefore referred to as single marker (SM) assay.

From the early 2000's on, single nucleotide polymorphisms (SNPs) began to arise which are the most abundant source of genetic variation in the genome (Gupta *et al.*, 2001). For SSRs and SNPs, several different high-throughput (HT) assays have been developed which allow to multiplex several SSRs in one polymerase chain reaction (PCR), or to genotype ten-thousands of SNPs with one chip or microarray (Syvänen, 2005; Appleby *et al.*, 2009). HT assays have considerably reduced the cost and effort of marker analysis, but their relative efficiency compared to SM assays in background selection has not yet been determined.

An important difference between SM and HT assays with respect to background selection is that with SM assays only those markers which have not yet been fixed for the recipient alleles have to be analyzed in advanced backcross generations. They are therefore very flexible to use, but have the disadvantage that the analysis of the single locus is comparatively expensive. For HT assays such as SNP chips, the complete set of markers included in the assay has to be analyzed in every analysis run. With this type of assay, the analysis of the single locus is cheaper than with SM assays, but the complete HT assay is expensive. HT assays are therefore less flexible than SM assays and only cost-efficient if used at or near full capacity.

Due to their different characteristics, SM and HT assays are suitable for different applications within a MABC program. However, as they usually are employed for different types of molecular markers, *e.g.,* SSRs and SNPs, it is not always possible to combine them efficiently. With Competitive Allele Specific PCR (KASP) assays, a type of SNP assay has recently emerged which is suitable for genotyping small subsets of SNP markers originally developed for analysis with HT assays (Chen *et al.*, 2010; Mammadov *et al.*, 2012). With this type of assay, new possibilities for the combination of HT and SM assays have arisen which have not yet been investigated.

The theoretical framework of MABC

With the multitude of marker types and marker assays that is now available, innovative strategies are required to apply them efficiently in the breeding process (Septiningsih *et al.*, 2013). However, gathering expertise through field experiments is time-consuming and very costly. Even though the costs of molecular markers have constantly been decreasing since the beginning of the millennium, they are still the main factor which limits the implementation of marker-assisted selection in practice. Beside the financial component, utilizing molecular markers increases the complexity of breeding programs, as it requires additional steps of analysis, interpretation and decision-making within a limited timeframe (Eathington *et al.*, 2007). It is therefore crucial to develop tools and guidelines beyond "gut-instinct" which help breeders to decide whether the way in which they plan to incorporate markers in their breeding programs is likely to be cost-effective (Morris *et al.*, 2003).

An important step towards this goal was the development of a theoretical framework for MABC by mathematical modelling (Stam & Zeven, 1981; Hillel *et al.*, 1990; Hospital *et al.*, 1992; Hill, 1993; Visscher, 1996; Markel *et al.*, 1997; Hospital & Charcosset, 1997; Frisch *et al.*, 1999b; Ribaut *et al.*, 2002) that was built on classical population-genetical investigations (Bartlett & Haldane, 1935; Hanson, 1959). However, the numerical solutions presented in these studies are only valid for unselected populations in which the donor genome proportions of the individuals are stochastically independent. They are consequently of little use for small backcross populations under marker-assisted selection. In 1999, Visscher showed in a simulation study that marker-assisted selection significantly reduces the variance of the donor genome proportion compared to the theoretical estimates which assumed the absence of selection (Visscher, 1999).

A general selection theory for the recipient genome proportion in MABC, which in contrast to standard normal distribution selection theory takes the

reduced variance of the donor genome proportion under marker-assisted selection into account, has more recently been published (Frisch & Melchinger, 2005). Compared to the previously presented numerical solutions, this approach had the advantage that individuals that were used as non-recurrent parents were subject to background selection. Nevertheless, it was limited to a specific marker score as selection index and did not allow a comparison of alternative selection strategies. Moreover, this approach was developed for only one backcross generation. Hence, whereas the presented analytical approaches greatly improved the understanding of the underlying principles of population genetics and are generally applicable, they are not sufficient for planning practical breeding programs.

The role of simulations

Simulations are often more powerful than analytical approaches, as they can be tailored to be closer to real conditions of selection (Moreau *et al.*, 1998). For more complex breeding designs which take into account many different parameters and even interactions thereof, numerical approaches are not straightforward, and sometimes no exact solution is available (Hospital *et al.*, 1992). A great advantage of simulations is therefore that they are comparatively easy-to-use tools which allow breeders to evaluate the efficiency of alternative crossing and selection schemes over all generations of a backcross program without the need to conduct expensive field experiments (Moreau *et al.*, 1998; Frisch & Melchinger, 2005).

As the scale and complexity of breeding programs increases, the optimization of breeding designs by computer simulations and the development of decision support tools for breeders is gaining importance for the successful application of marker-assisted selection (Xu & Crouch, 2008). Its efficient implementation in practical large-scale plant breeding programs requires, among other logistical and genetical prerequisites, the design of optimal

breeding systems by simulation analysis, and the development of decision support tools for breeders (Eathington *et al.*, 2007).

Validation studies have shown that simulations are effective and robust tools to improve the planning process of practical breeding programs (Kuchel *et al.*, 2007; Prigge *et al.*, 2008; Randhawa *et al.*, 2009). They have been recognized as useful and integral parts of efficient plant breeding in scientific literature (Utomo *et al.*, 2012; Septiningsih *et al.*, 2013). Guidelines for optimizing MABC designs have also been used as a basis for more sophisticated models (Tesfaye *et al.*, 2013; Peng *et al.*, 2014a) and have found their way into practical breeding programs (Timonova *et al.*, 2013).

To date, a broad range of simulation studies is available which cover important applications of MABC. They provide guidelines for different aspects of the introgression of single dominant or recessive target genes (Hospital *et al.*, 1992; Frisch *et al.*, 1999a; Frisch & Melchinger, 2001b; Prigge *et al.*, 2009), the combination of two genes (Frisch & Melchinger, 2001c) or several genes (Ribaut *et al.*, 2002; Servin *et al.*, 2004) and quantitative trait loci with estimated positions (Hospital & Charcosset, 1997). More recent studies have dealt with the development of introgression libraries (Falke *et al.*, 2009) and multiple integration of transgenic traits (Peng *et al.*, 2014a; Peng *et al.*, 2014b). These studies have focused on optimizing the use of SM assays. However, none of these studies has provided guidelines for the efficient application of HT assays.

Objectives

The aim of my Ph.D. project was to employ computer simulations for the development of efficient strategies for MABC with HT assays. Guidelines should be derived for a wide range of applications of MABC and should be implementable in practical breeding programs. The thesis project was divided into the following four sub-projects:

(1) Application of marker-assisted background selection for gene introgression is still limited by the high costs of marker analysis. HT assays promise to reduce these costs, but new selection strategies are required for their efficient implementation in breeding programs. The objectives of the first study were to investigate the properties of HT assays compared to SM assays, and to develop optimal selection strategies for marker-assisted gene introgression with HT assays in maize (Chapter 2).

(2) For many crops, efficient conversion of seed-parent lines to CMS is a cornerstone of hybrid production. In contrast to gene introgression, no target genes have to be considered in CMS conversion programs. The optimal selection strategies for CMS conversion will consequently differ from those for gene introgression and have not yet been investigated. The objectives of the second study were to evaluate and optimize the resource requirements of CMS conversion programs in rye, sugarbeet, sunflower and rapeseed, and to determine the most cost-effective use of SM and HT assays (Chapter 3).

(3) Organizing SSR markers located on the same chromosomes into PCR multiplexes has the potential to reduce the costs of marker analysis and constitutes an HT assay with a level of throughput between SM assays and SNP chips. The objectives of the third study were to develop selection strategies for gene introgression in grapevine with chromosome-wise SSR multiplexes (Chapter 4).

(4) Introgression populations are valuable resources for QTL detection and breeding, but their development is costly and time-consuming. Selection strategies for the development of introgression populations with a limited number of individuals and HT marker assays are required. The objectives of the fourth study were to design and compare selection strategies for the development of maize introgression populations with limited resources for different doubled haploid (DH) and S_2 crossing schemes (Chapter 5).

Chapter 2

Selection strategies for marker-assisted backcrossing with high-throughput marker systems[1]

[1]Herzog, E, & Frisch, M. 2011. Selection strategies for marker-assisted backcrossing with high-throughput marker systems. *Theor Appl Genet*, **123**(2), 251-260.

Theor Appl Genet (2011) 123:251–260
DOI 10.1007/s00122-011-1581-0

ORIGINAL PAPER

Selection strategies for marker-assisted backcrossing with high-throughput marker systems

Eva Herzog · Matthias Frisch

Received: 10 September 2010 / Accepted: 23 March 2011 / Published online: 8 April 2011

Abstract Application of marker-assisted backcrossing for gene introgression is still limited by the high costs of marker analysis. High-throughput (HT) assays promise to reduce these costs, but new selection strategies are required for their efficient implementation in breeding programs. The objectives of our study were to investigate the properties of HT marker systems compared to single-marker (SM) assays, and to develop optimal selection strategies for marker-assisted backcrossing with HT assays. We employed computer simulations with a genetic model consisting of 10 chromosomes of 160 cM length to investigate the introgression of a dominant target gene. We found that a major advantage of HT marker systems is that they can provide linkage maps with equally spaced markers, whereas the possibility to provide linkage maps with high marker densities smaller than 10 cM is only of secondary use in marker-assisted backcrossing. A three-stage selection strategy that combines selection for recombinants at markers flanking the target gene with SM assays and genome-wide background selection with HT markers in the first backcross generation was more efficient than genome-wide background selection with HT markers alone. Selection strategies that combine SM and HT assays were more efficient than genome-wide background selection with HT assays alone. This result was obtained for a broad range of cost ratios of HT and SM assays. A further considerable reduction of the costs could be achieved if the population size in the first backcross generation was twice the population size in generations BC_2 and BC_3 of a three-generation backcrossing program. We conclude that selection strategies combining SM and HT assays have the potential to greatly increase the efficiency and flexibility of marker-assisted backcrossing.

Communicated by Y. Xu.

E. Herzog · M. Frisch (✉)
Institute of Agronomy and Plant Breeding II,
Justus Liebig University, 35392 Giessen, Germany
e-mail: matthias.frisch@uni-giessen.de

Introduction

Marker-assisted backcrossing is used for transferring genes which are responsible for favorable agronomic traits from a donor line into the genome of a recipient line. Using molecular markers for selection against the genetic background of the donor can reduce the time and resources required for gene introgression. Although background selection has become a standard tool in plant breeding, the high costs of marker analysis still limit its use in practice and are the crucial factor for the experimental designs of gene introgression programs (Collard and Mackill 2008). These designs depend on the number of target genes to be transferred, the employed marker map, and the number of generations available for the gene introgression. Computer simulations are a robust tool for optimizing the design parameters of a marker-assisted backcrossing program before implementing it in practice (Prigge et al. 2008).

The design of marker-assisted backcrossing programs was studied with respect to the introgression of single dominant and recessive genes (Hospital et al. 1992; Frisch et al. 1999a, b; Frisch and Melchinger 2001a), two genes (Frisch and Melchinger 2001b), and favorable alleles at quantitative trait loci (Hospital and Charcosset 1997; Bouchez et al. 2002). More recently, marker-assisted backcrossing for developing libraries of near-isogenic lines was studied (Peleman and van der Voort 2003; Falke et al. 2009; Falke and Frisch 2011). These studies have mainly

focused on optimizing the number of genotyped individuals as well as the positions and density of background selection markers with respect to the required number of marker data points. The optimizations have been carried out assuming marker systems in which each marker locus is analyzed in a separate assay (cf. Prigge et al. 2009). We refer to such systems as single-marker (SM) systems. Typical examples are the simple sequence repeat (SSR) and the restriction fragment length polymorphism (RFLP) marker systems.

Recently, high-throughput (HT) marker systems based on single nucleotide polymorphisms (SNPs) have been developed. Due to the high level of automation of systems such as DNA chips, they allow for cheap and fast analysis of hundreds of marker loci in a single analysis step (Gupta et al. 2001; Syvänen et al. 2005). HT marker systems have been developed for crops (Ragot and Lee 2007) and are becoming the marker systems of choice in commercial breeding programs of many economically important crops.

The crucial difference between HT and SM marker systems is that with SM marker systems, only those markers are analyzed in advanced backcross generations which were not already fixed for the recipient alleles in earlier generations. In contrast, with HT marker systems, the entire panel of markers used in a gene introgression program needs to be analyzed also for individuals of advanced backcross generations, even if 80 or 90% of these markers have already been fixed for the recipient alleles. To our knowledge, no study investigating the implications of this property on the efficiency of marker-assisted backcrossing is available. The combination of SM marker systems for the reduction of the chromosome segment attached to the target gene and HT markers for genome-wide background selection promises to further enhance selection efficiency in marker-assisted backcrossing and is not yet investigated.

The objectives of our simulation study were to (1) compare the relative costs of genome-wide background selection with SM and HT marker systems for different cost ratios of HT:SM markers, (2) compare the efficiency of equally spaced and randomly distributed markers with respect to the recovery of the recipient genome, (3) develop selection strategies combining SM and HT assays, which are more efficient than genome-wide background selection with SM or HT assays alone.

Simulations

A genetic model with ten equally sized chromosomes of 160 cM length was used for the simulations. Its genome length of 1,600 cM is similar to that of published linkage maps of maize (cf. Schön et al. 1994). Markers for genome-wide background selection were assumed to be (a) randomly distributed in the genome or (b) equally spaced. Average marker distances (randomly distributed markers) or marker distances (equally spaced markers) between two adjacent marker loci of $\delta_{GW} = 2$, 5, 10, 20 cM were investigated. For equally spaced markers, two markers were located at the telomeres of each chromosome. One dominant target gene to be introgressed was located on Chromosome 1. It was 81, 82.5, 85, and 90 cM distant from the telomere for linkage maps with $\delta_{GW} = 2$, 5, 10, 20 cM, respectively. Flanking markers for selection against the donor chromosome segment attached to the target gene were located on both sides of the target gene. The distances between target gene and each flanking marker were $\delta_F = 5$, 10, 20, 30, 40 cM.

The investigated breeding scheme started with the cross of two homozygous parents (donor and recipient), which were polymorphic at all loci. The recipient carried the desirable alleles at all loci of the genome except for the target locus, while the donor carried the desirable allele at the target locus. The donor and recipient were crossed to create an F_1 individual, which was backcrossed to the recipient. From the BC_1 population of size n_1, one individual was selected with two- or three-stage selection, as described below, and backcrossed to the recipient. This procedure was repeated for t backcross generations.

Two-stage selection consisted of pre-selection of carriers of the target gene in the first selection step. The pre-selected individuals were subjected to genome-wide background selection in the second step. A selection index $i = \sum_m x_m$ was constructed, where summation is over markers and $x_m = 1$ if a marker is homozygous for the recipient allele. A plant with the highest value of i was selected and backcrossed to the recipient. Two-stage selection was carried out with SM and HT assays. For SM assays, only those markers were analyzed in advanced backcross generations which were not yet fixed for the recipient allele in the non-recurrent parent.

Three-stage selection combined selection for recombinants between the target gene and its two flanking markers, genotyped with SM assays, and genome-wide background selection with HT assays. It consisted of (1) selection for the target gene followed by (2) pre-selection with flanking markers and (3) genome-wide selection with background markers. For selection step (2), a selection index f was created, which took the values 0, 1, or 2, depending on whether recombination occurred between the target gene and none, one, or both flanking markers, respectively. On the basis of f, pre-selection of individuals was carried out according to one of two decision rules. Either (a) individuals with $f \geq 1$ were selected, or (b) all individuals having the maximum observed score of f ($f = \max$) were selected.

Four series of simulations were carried out with software Plabsoft (Maurer et al. 2008), assuming no interference in crossover formation. Each simulation was replicated 10,000 times in order to reduce sampling effects and to obtain results with high numerical accuracy and a small standard error. The 10% quantile (Q10) of the distribution of recipient genome (in percent) was determined in the last backcross generation to measure the success of a marker-assisted backcrossing program with respect to restoring the genome of the recipient. The number of SM and HT assays was determined as a measure for the costs of a marker-assisted backcrossing program.

In the first series of simulations, the population size n_t (constant across all backcross generations BC_t, $t = 1, \ldots, 3$) and the number of marker assays were determined which were required to reach Q10 values of 93, 94, 95, 96, 97, 98%, respectively. For 93–96%, we investigated two-generation backcrossing programs, and for 96–98% three-generation backcrossing programs. Two-stage selection with either SM or HT assay or a combination of both systems (HT in backcross generation BC_1 and SM in the following backcross generations) was carried out for linkage maps with $\delta_{GW} = 5, 10, 20$ cM.

In the second series of simulations, two-stage selection with HT assays was carried out. Background selection markers were either equally spaced or randomly distributed with $\delta_{GW} = 2, 5, 10, 20$ cM. We considered three backcross generations and constant values of n_t ranging from 40 to 200 individuals.

In the third series of simulations, three-stage selection was carried out either in backcross generation BC_1 or BC_3. In the remaining two generations, two-stage selection with HT assays was carried out. The flanking markers for three-stage selection had distances of $\delta_F = 5, 10, 20, 30, 40$ cM from the target gene and individuals with $f \geq 1$ were selected for genome-wide analysis with HT assays. Distances between genome-wide background selection markers were $\delta_{GW} = 5$ cM. In the generations with two-stage selection, we investigated population sizes from $n_t = 40$ to 200. In the generation with three-stage selection, these population sizes were multiplied by a factor $m = 1, 2, 5$.

In the fourth series of simulations, three-stage selection was carried out in backcross generations BC_1 and BC_2. Marker distances of $\delta_{GW} = 5$ cM and $\delta_F = 20$ cM were employed. Individuals with $f \geq 1$ were pre-selected for genome-wide analysis in backcross generation BC_1, while only individuals having the highest observed number of recombinations between target gene and flanking markers ($f = \max$) were pre-selected in backcross generation BC_2. In backcross generation BC_3, two-stage selection was carried out with HT assays. We investigated population sizes from $n_t = 40$ to 200 for generations BC_2 and BC_3. In backcross generation BC_1, these population sizes were multiplied by the factor $m = 1, 2, 5$.

For comparing the costs of marker-assisted backcrossing programs with different selection strategies, linkage maps, and population sizes, the numbers of SM and HT assays required for the entire backcrossing program were assessed. For SM analyses, only those markers not yet fixed for the recipient allele in the non-recurrent parent of a backcross population were considered. For HT analyses, the number of assays was the same as the number of individuals subjected to genome-wide background selection. Calculation of costs was based on five cost ratios of one HT assay (corresponding to all HT marker loci on the linkage map) compared to one SM assay (corresponding to one SM locus). Cost ratios of HT:SM of 200:1, 100:1, 50:1, 20:1, 10:1 were investigated. For example, a cost ratio HT:SM of 100:1 corresponds to a price of 200€ for analyzing all SNP background marker loci with a DNA chip, and 2€ for analyzing one SSR marker locus. Comparisons were carried out to compare (a) the costs of two-stage selection with HT assays to those of two-stage selection with SM assays, (b) the costs of two-stage selection with HT assays in generation BC_1 and SM assays in BC_2 and BC_3 to those of two-stage selection with HT assays in all backcross generations, (c) the costs of three-stage selection in BC_1 to those of two-stage selection with HT assays in all generations. For (a) the costs of SM assays were set 1 and the relative costs of HT assays were determined, for (b) the costs of using HT assays in all backcross generations were set 1 and the relative costs of the strategy combing HT and SM were determined, and for (c) the costs of two-stage selection were set 1 and the relative costs of three-stage selection were determined.

Results

For two-stage selection, HT assays were considerably more expensive (up to factor 4.77) than SM assays for scenarios with high relative costs of HT markers (200:1, 100:1, and 50:1) in combination with large marker distances and/or large attempted Q10 values (Table 1). For scenarios with small marker distances and/or low relative cost ratios of HT:SM assays and low attempted Q10 values, HT assays were cheaper. To reach a Q10 value of 96% in two generations, the number of required marker assays was 9–14 times greater than those required to reach the same Q10 value in three generations. The increase in the required number of marker assays, which accompanied the shortening of a backcrossing program from three to two generations, was greater for SM than for HT marker systems.

For high cost ratios of HT:SM markers (200:1, 100:1, and 50:1) and large marker distances, combining HT assays

Table 1 Relative costs of a gene introgression program using HT assays in generations BC_1 to BC_3 (HT[BC_{1-3}]) compared to using SM assays in BC_1 to BC_3 (SM[BC_{1-3}]) depending on the cost ratio of HT:SM assays

δ_{GW}	Q10 (%)	No. of BC generations	n_t	No. of assays		Cost ratio HT:SM				
						Relative costs				
				HT[BC_{1-3}]	SM[BC_{1-3}]	200:1	100:1	50:1	20:1	10:1
20 cM ($n_m = 90$)	93	2	44	44	2,643	3.33	1.66	0.83	0.33	0.17
	94	2	72	72	4,260	3.38	1.69	0.85	0.34	0.17
	95	2	133	133	7,737	3.44	1.72	0.86	0.34	0.17
	96	2	291	291	16,583	3.51	1.75	0.88	0.35	0.18
	96	3	17	26	1,158	4.40	2.20	1.10	0.44	0.22
	97	3	30	45	1,975	4.56	2.28	1.14	0.46	0.23
	98	3	70	105	4,401	4.77	2.39	1.19	0.48	0.24
10 cM ($n_m = 170$)	93	2	39	39	4,442	1.76	0.88	0.44	0.18	0.09
	94	2	62	62	6,960	1.78	0.89	0.45	0.18	0.09
	95	2	110	110	12,141	1.81	0.91	0.45	0.18	0.09
	96	2	222	222	24,050	1.85	0.92	0.46	0.18	0.09
	96	3	16	24	2,070	2.32	1.16	0.58	0.23	0.12
	97	3	26	39	3,258	2.39	1.20	0.60	0.24	0.12
	98	3	53	80	6,382	2.49	1.25	0.62	0.25	0.12
5 cM ($n_m = 330$)	93	2	38	38	8,406	0.90	0.45	0.23	0.09	0.05
	94	2	60	60	13,077	0.92	0.46	0.23	0.09	0.05
	95	2	104	104	22,292	0.93	0.47	0.23	0.09	0.05
	96	2	206	206	43,361	0.95	0.48	0.24	0.10	0.05
	96	3	15	23	3,780	1.19	0.60	0.30	0.12	0.06
	97	3	25	38	6,094	1.23	0.62	0.31	0.12	0.06
	98	3	50	75	11,719	1.28	0.64	0.32	0.13	0.06

Two-stage selection, n_m equally spaced background selection markers with distances δ_{GW}, and population sizes n_t were used to recover Q10 target values of 93–98% in two or three backcross generations

in generation BC_1 with SM assays in generations BC_2 and BC_3 for genome-wide background selection was cheaper (up to 60%) than using HT assays alone (Table 2). This cost reduction was more pronounced for three-generation than two-generation backcross programs.

To reach a given Q10 value with randomly distributed background selection markers, linkage maps with two to four times more markers are required than with equally spaced markers of marker distances $\delta_{GW} = 20$ or 10 cM (Table 3). With equally spaced markers and $\delta_{GW} = 5$ cM, approximately the same Q10 values were reached as with randomly distributed markers and $\delta_{GW} = 2$ cM. A decrease in the distance between equally distributed markers from $\delta_{GW} = 10$ to 5 cM resulted in only marginally greater Q10 values in generation BC_3. No difference in the Q10 values was observed for $\delta_{GW} = 5$ and 2 cM.

With three-stage selection combining SM and HT assays in generation BC_1, the flanking marker distance δ_F had only marginal influence on the recovered genome-wide Q10 values (Table 4). For population sizes $n_2 = n_3 < 100$ in generations BC_2 and BC_3, a substantial increase of the Q10 values was observed, if in generation BC_1 larger populations $n_1 > n_2 = n_3$ were employed. Doubling the population size in generation BC_1 ($n_1 = mn_2 = mn_3$, $m = 2$) had approximately the same effect on the Q10 values as increasing a constant population size by about 20 individuals ($n_1' = n_2' = n_3' = n_2 + 20$). The combination of doubled population sizes in generation BC_1 and small flanking marker distances δ_F resulted in less required HT assays at the expense of more required SM assays to reach a certain Q10 value, compared to backcrossing programs with constant population sizes across generations.

Three-stage selection in generation BC_3 recovered similar Q10 values as three-stage selection in generation BC_1 for all combinations of n_t and m. However, more HT assays were required (data not shown).

Three-stage selection in generations BC_1 and BC_2 required more SM assays but less HT assays compared to three-stage selection only in generation BC_1 for all combinations of n_t and m (Table 5). For population sizes smaller than 100, slightly lower Q10 values were recovered.

Three-stage selection combining SM and HT assays in generation BC_1 of a three-generation backcrossing program was cheaper than two-stage selection with HT assays for all

Table 2 Relative costs of a gene introgression program using HT assays in backcross generation BC_1 and SM assays in backcross generations BC_2 and BC_3 (HT[BC_1], SM[$BC_{2,3}$]) compared to using HT assays in all backcross generations (HT[$BC_{1–3}$], data presented in Table 1) depending on the cost ratio of HT:SM assays

δ_{GW}	Q10 (%)	No. of BC generations	n_t	No. of assays		Cost ratio HT:SM				
						Relative costs				
				HT[BC_1]	SM[$BC_{2,3}$]	200:1	100:1	50:1	20:1	10:1
20 cM ($n_m = 90$)	93	2	44	22	664	0.58	0.65	0.80	1.25	2.01
	94	2	72	36	1,019	0.57	0.64	0.78	1.21	1.92
	95	2	133	67	1,749	0.57	0.64	0.77	1.16	1.82
	96	2	291	146	3,490	0.56	0.62	0.74	1.10	1.70
	96	3	17	9	393	0.42	0.50	0.65	1.10	1.86
	97	3	30	15	624	0.40	0.47	0.61	1.03	1.72
	98	3	70	35	1,250	0.39	0.45	0.57	0.93	1.52
10 cM ($n_m = 170$)	93	2	39	20	1,130	0.66	0.80	1.09	1.96	3.41
	94	2	62	31	1,686	0.64	0.77	1.04	1.86	3.22
	95	2	110	55	2,787	0.63	0.75	1.01	1.77	3.03
	96	2	222	111	5,183	0.62	0.73	0.97	1.67	2.83
	96	3	16	8	712	0.48	0.63	0.93	1.82	3.30
	97	3	26	13	1,051	0.47	0.60	0.87	1.68	3.03
	98	3	53	27	1,880	0.46	0.57	0.81	1.51	2.69
5 cM ($n_m = 330$)	93	2	38	19	2,129	0.78	1.06	1.62	3.30	6.10
	94	2	60	30	3,194	0.77	1.03	1.56	3.16	5.82
	95	2	104	52	5,138	0.75	0.99	1.49	2.97	5.44
	96	2	206	103	9,359	0.73	0.95	1.41	2.77	5.04
	96	3	15	8	1,300	0.63	0.91	1.48	3.17	6.00
	97	3	25	13	1,969	0.60	0.86	1.38	2.93	5.52
	98	3	50	25	3,479	0.57	0.80	1.26	2.65	4.97

Two-stage selection, n_m equally spaced background selection markers with distances δ_{GW}, and population sizes n_t were used to recover Q10 target values of 93–98% in two or three backcross generations

Table 3 Q10 values recovered in generation BC_3 for constant population sizes n_t in generations BC_1 to BC_3 and equally spaced or randomly distributed markers ($\delta_{GW} = 2, 5, 10, 20$ cM) applying two-stage selection with HT assays

δ_{GW} (cM)	Generation	Equally spaced markers, n_t					Randomly distributed markers, n_t				
		40	80	120	160	200	40	80	120	160	200
20	BC_1	79.7	81.4	82.4	83.0	83.4	78.0	79.6	80.5	80.9	81.4
	BC_2	92.8	94.2	94.9	95.3	95.6	91.3	92.6	93.2	93.6	94.0
	BC_3	97.4	98.1	98.4	98.6	98.7	96.4	97.0	97.3	97.4	97.5
10	BC_1	79.9	81.7	82.7	83.3	83.8	78.8	80.5	81.3	81.9	82.3
	BC_2	93.0	94.5	95.2	95.6	95.9	91.9	93.4	94.1	94.4	94.8
	BC_3	97.6	98.4	98.7	98.9	99.0	97.0	97.8	98.1	98.3	98.4
5	BC_1	80.0	81.7	82.7	83.4	83.9	79.3	81.0	81.9	82.5	83.0
	BC_2	93.1	94.5	95.3	95.7	96.0	92.4	93.8	94.4	94.8	95.1
	BC_3	97.8	98.5	98.8	99.0	99.1	97.1	97.9	98.3	98.4	98.6
2	BC_1	80.0	81.8	82.8	83.4	83.8	79.8	81.5	82.5	83.1	83.7
	BC_2	93.2	94.6	95.3	95.7	96.0	93.0	94.4	95.1	95.5	95.9
	BC_3	97.8	98.5	98.8	99.0	99.1	97.7	98.5	98.7	98.9	99.1

investigated combinations of n_t with $m = 1$ and $m = 2$ (Fig. 1). The costs were ranging between 75.3–83.0% ($m = 1$) and 57.1–89.7% ($m = 2$) of the costs of two-stage selection. For $m = 5$, three-stage selection was only cheaper for cost ratios of HT:SM from 200:1 to 50:1. Three-stage selection with doubled population size ($m = 2$) in generation

Table 4 Q10 values recovered in generation BC_3 and number of required SM/HT assays for increased population sizes $n_1 = mn_t$ ($m = 1, 2, 5$; $t = 2, 3$) in generation BC_1 and equally spaced markers ($\delta_{GW} = 5$ cM) applying three-stage selection ($\delta_F = 5, 10, 20, 30, 40$ cM; $f \geq 1$) in generation BC_1 and two-stage selection in generations BC_2 and BC_3

m	δ_F (cM)	n_t								
		40	60	80	100	120	140	160	180	200
		Q10 (%) in generation BC_3								
1	40	97.8	98.2	98.5	98.7	98.8	98.9	99.0	99.0	99.1
	30	97.8	98.2	98.5	98.7	98.8	98.9	99.0	99.0	99.1
	20	97.8	98.2	98.6	98.7	98.9	99.0	99.0	99.1	99.1
	10	97.6	98.2	98.5	98.7	98.9	99.0	99.1	99.1	99.2
	5	97.4	98.0	98.3	98.6	98.8	98.9	99.0	99.1	99.1
2	40	98.0	98.4	98.6	98.8	98.9	99.0	99.0	99.1	99.2
	30	98.0	98.4	98.6	98.8	98.9	99.0	99.1	99.1	99.2
	20	98.0	98.5	98.7	98.8	99.0	99.1	99.1	99.2	99.2
	10	97.9	98.4	98.7	98.8	99.0	99.1	99.2	99.2	99.3
	5	97.7	98.2	98.6	98.8	98.9	99.1	99.1	99.2	99.3
5	40	98.2	98.6	98.8	98.9	99.0	99.1	99.1	99.2	99.2
	30	98.2	98.6	98.8	98.9	99.0	99.1	99.2	99.2	99.2
	20	98.2	98.6	98.8	99.0	99.1	99.1	99.2	99.2	99.3
	10	98.2	98.6	98.9	99.0	99.1	99.2	99.3	99.3	99.4
	5	98.1	98.6	98.8	99.0	99.1	99.2	99.3	99.3	99.4
		No. of required SM/HT assays								
1	40	40/49	60/73	80/98	100/123	120/148	140/172	160/197	180/222	200/246
	30	40/47	60/71	80/95	100/119	120/143	140/167	160/191	180/215	200/239
	20	40/45	60/68	80/91	100/114	120/137	140/160	160/183	180/206	200/229
	10	40/43	60/64	80/86	100/108	120/129	140/151	160/173	180/195	200/216
	5	40/44	60/64	80/84	100/104	120/125	140/146	160/166	180/187	200/208
2	40	80/58	120/88	160/117	200/146	240/176	280/205	320/235	360/265	400/294
	30	80/55	120/83	160/111	200/139	240/167	280/195	320/223	360/251	400/279
	20	80/51	120/77	160/103	200/129	240/155	280/181	320/207	360/233	400/259
	10	80/46	120/69	160/93	200/116	240/140	280/163	320/187	360/210	400/234
	5	80/44	120/65	160/86	200/108	240/130	280/152	320/174	360/196	400/218
5	40	200/86	300/130	400/174	500/218	600/261	700/305	800/349	900/393	1,000/437
	30	200/79	300/119	400/159	500/199	600/239	700/279	800/319	900/359	1,000/399
	20	200/69	300/104	400/140	500/175	600/210	700/245	800/280	900/315	1,000/350
	10	200/56	300/85	400/114	500/142	600/171	700/200	800/228	900/257	1,000/285
	5	200/48	300/73	400/98	500/122	600/147	700/172	800/196	900/221	1,000/245

BC_1 was the optimal selection strategy for reaching Q10 values of 98 and 99%. The only exception was the combination of a cost ratio of HT:SM assays of 10:1 and a desired Q10 value of 99%. In this case, constant population size over generations ($m = 1$) was optimal.

Discussion

HT marker systems

HT marker systems are expected to increase the cost-efficiency of marker-assisted backcrossing programs (Ragot and Lee 2007; Collard and Mackill 2008). However, previous studies on the efficiency of gene introgression programs have rarely taken differences between marker systems into account (Ribaut et al. 2002). In this study, we investigated the different properties of SM and HT marker systems and their effect on the efficiency of gene introgression. The simultaneous analysis of a large number of marker loci at comparatively low cost per individual marker locus is made feasible in HT assays (Syvänen et al. 2005). They, therefore, promise to be a powerful tool for marker-assisted background selection, especially when the expected number of required marker analyses is high. However, HT assays do not provide the possibility to

Table 5 Q10 values recovered in generation BC_3 and number of required SM/HT assays for increased population sizes $n_1 = mn_t$ ($m = 1, 2, 5$; $t = 2, 3$) in generation BC_1 and equally spaced markers ($\delta_{GW} = 5$ cM) applying three-stage selection ($\delta_F = 5, 10, 20, 30, 40$ cM) in generations BC_1 ($f \geq 1$) and BC_2 ($f = \max$) and two-stage selection in generation BC_3

m	δ_F (cM)	n_t								
		40	60	80	100	120	140	160	180	200
		Q10 (%) in generation BC3								
1	20	97.3	98.0	98.4	98.7	98.8	98.9	99.0	99.1	99.1
2	20	97.6	98.3	98.6	98.8	99.1	99.0	99.1	99.2	99.2
5	20	98.0	98.5	98.8	99.0	99.1	99.1	99.2	99.3	99.3
		No. of required SM/HT assays								
1	20	58/30	86/45	115/61	143/77	172/93	200/109	228/125	256/141	285/157
2	20	97/36	146/55	194/73	242/92	291/111	338/131	387/151	436/169	484/189
5	20	217/54	325/82	433/111	542/138	650/168	758/196	866/224	974/252	1082/281

selectively analyze individual markers. In contrast to SM assays, all markers on the linkage map need to be analyzed for every backcross individual, even if a large proportion of markers has already been fixed for the recipient alleles, as is the case in advanced backcross generations.

Comparing two-generation with three-generation gene introgression programs showed that SM marker systems require relatively less assays in three-generation programs than HT assays. For example, in a two-generation gene introgression program with distances of genome-wide background selection markers of $\delta_{GW} = 20$ cM, both 44 HT and 2,643 SM assays resulted in a Q10 value of 93%, whereas in a three-generation program, 45 HT or 1,975 SM assays resulted in a Q10 value of 97% (Table 1). This effect is expected to be even more pronounced for background selection in higher backcross generations, and when background selection is carried out in selfing generations or during doubled haploid production. In line, using HT assays for genome-wide background selection in the first backcross generation, and SM assays in advanced backcross generations reduced the costs of marker analysis compared to using HT assays in all backcross generations (Table 2). Only 5–9% of all marker analyses in a three-generation backcross program fell upon backcross generation BC_3. The cost reduction compared to using HT assays in all backcross generations was consequently greater for three-generation than for two-generation programs. We conclude that HT assays are particularly suited for short gene introgression programs, while SM assays are efficient for marker-assisted background selection when in advanced generations already large percentages of the markers have been fixed for the recipient alleles.

Marker distance and distribution for genome-wide background selection

HT systems based on SNP markers are often analyzed with techniques employing marker numbers that are multiples of 96. We did not limit our investigations to these marker numbers for two reasons. Firstly, usually not all markers of such a set are polymorphic for a certain cross. Moreover, reduced representation sequencing approaches have recently emerged and a trend towards genotyping by sequencing can be observed. For these systems, fixed marker numbers are less relevant. Therefore, we focused in our study on marker distances δ_{GW}, but not on the fixed marker numbers employed by a certain marker technology. The results discussed below can be regarded as thresholds, which, if they are surpassed for two parental lines and a certain HT markers system, result in the presented Q10 values.

SNPs occur in abundance in plant genomes. Dense linkage maps with marker distances below 5 cM can consequently be established at reasonable costs. However, the effect of such dense markers on the recipient genome recovery has not yet been investigated. Decreasing the marker distances δ_{GW} below 10 cM had only marginal effect on the recipient genome recovery (Table 1). An explanation for this result is that on expectation one crossover per meiosis and chromatid occurs on a chromosome segment of length 1 M. In two- or three-generation backcrossing programs, the number of recombination events resulting in chromosome segments of different parental origin is therefore limited. To detect these chromosome segments and to efficiently identify the backcross individuals with the smallest percentage of donor genome, a marker distance of $\delta_{GW} = 10$ cM is sufficient. Smaller marker distances are not required, because the factor limiting selection response is not the precise estimation of the donor genome percentage, but the limited number of crossovers.

The difference in the Q10 values between equally spaced and randomly distributed markers was considerable for all marker distances δ_{GW} except 2 cM. Less than half the markers were required to reach a certain Q10 value with equally spaced markers compared with randomly

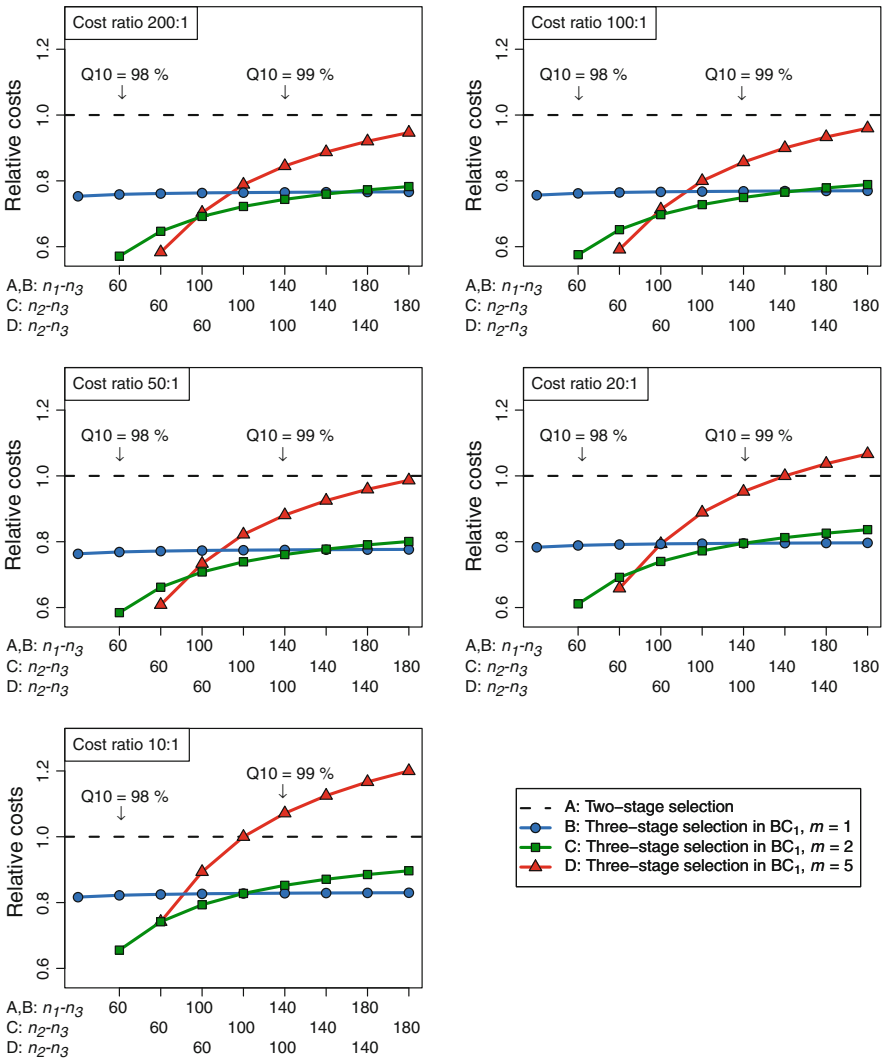

Fig. 1 Relative costs of three-stage selection with $m = 1, 2, 5$ in generation BC_1 and two-stage selection in generations BC_2 and BC_3 compared to two-stage selection in generations BC_1 to BC_3 for cost ratios for HT:SM assays of 200:1, 100:1, 50:1, 20:1, and 10:1

distributed markers (Table 3). This difference can be explained by the fact that, with random marker distribution, occasionally the distance between adjacent markers can get quite large, resulting in random gaps in the marker coverage. The recipient genome content of the chromosome regions in these gaps is not assessed and, therefore, the correlation of the marker estimate of the recurrent parent genome contribution and the true recurrent parent genome contribution is lower than for equally spaced markers. This results in a smaller response to marker-assisted background selection for randomly distributed compared to equally spaced markers.

We conclude that the possibility to generate linkage maps with equidistant marker distribution is a major advantage of HT marker systems, while the possibility to establish linkage maps with marker distances below 10 cM is only of secondary importance for gene introgression programs.

Pre-selection with flanking markers

In three-stage selection, the pre-selection of backcross plants showing recombination between the target gene and flanking markers allows an efficient control of the donor chromosome segment attached to the target gene. This reduces the probability of introducing negative alleles linked to the target gene into the genome of the recipient. Further, three-stage selection reduces the number of

backcross plants subjected to genome-wide background selection and, therefore, reduces the number of required marker assays (Frisch et al. 1999a). To take advantage of these favorable properties of three-stage selection, a pre-selection for recombination between the target gene and flanking markers analyzed with SM assays can be combined with genome-wide background selection on the basis of HT assays. The design decisions required to implement such a selection strategy are discussed in the following.

Distances of flanking markers

Tightly linked flanking markers result in short donor chromosome segments attached to the target gene. However, they also result in a greater reduction of the number of individuals subjected to genome-wide background selection than loosely linked flanking markers. This reduced selection intensity can result in a decline of the genome-wide recovery of the recurrent parent genome. Therefore, the smallest δ_F that has no negative effect on the genome-wide response to selection can be regarded as an optimal flanking marker distance.

In backcrossing programs with constant ($m = 1$) population sizes $\leq$60, marker distances $\delta_F = 20$ cM between each flanking marker and the target gene resulted in high overall Q10 values while minimizing the number of HT assays required for background selection (Table 4). For larger populations, $\delta_F = 10$ was optimal. With $\delta_F = 5$ cM, controlling the donor genome segment attached to the target gene resulted in a decrease of the overall Q10 values. For such tightly linked flanking markers, only few recombinations do occur in a backcross population (see Frisch et al. 1999a, b for theoretical results) and, hence, only few plants are pre-selected and subjected to genome-wide background selection. This small number of individuals available for genome-wide background selection results in a smaller response to selection compared with less tightly linked flanking markers. We conclude that for gene introgression programs with constant population sizes, an optimum exploitation of the advantages of three-stage selection is reached with flanking marker distances of $\delta_F = 20$–10 cM, and that with smaller flanking marker distances, controlling the donor segment attached to the target gene is only possible at the cost of a lower overall Q10 value.

Generation of three-stage selection

Carrying out pre-selection for recombinants at markers flanking the target gene in only some, but not all generations of a gene introgression program can considerably reduce the logistic effort required for the marker analysis. A comparison of three-stage selection in generations BC_1 and BC_3 showed similar genome-wide Q10 values, but three-stage selection in generation BC_3 required more HT marker analyses (results not shown). Therefore, carrying out three-stage selection in generation BC_1 can be regarded as superior to three-stage selection in generation BC_3.

Three-stage selection in generations BC_1 and BC_2 required less HT assays but more SM assays than three-stage selection in generation BC_1 (Tables 4, 5). For population sizes below 100 individuals, this was accompanied by smaller genome-wide Q10 values. For population sizes greater than 100, employing three-stage selection in generations BC_1 and BC_2 provides a means to reduce the number of required genome-wide HT assays, by increasing the number of required SM analysis. Depending on the actual costs of SM and HT analysis and the work flow in the lab, this strategy can be used to shift the number of required marker analyses from HT to SM assays.

Large population sizes in the first backcross generation

As pre-selection with SM assays reduces the number of required HT assays, it provides a means to handle larger populations without necessarily increasing the cost of marker analysis. Increasing the population size in the generation where pre-selection with flanking markers is carried out increases the chance to find an individual with a small donor chromosome segment attached to the target gene, which has in addition a high proportion of recurrent parent genome (Frisch et al. 1999b). This theoretical consideration can serve as a rationale for using large population sizes in generations with three-stage selection.

We investigated backcrossing programs with three-stage selection in BC_1 populations that had $m = 1$, 2, or 5 times the size of the BC_2 and BC_3 populations in which two-stage selection was employed (Table 4). The Q10 values reached with $m = 1$ were comparable to those reached with two-stage selection for constant population sizes across generations (Table 3). Doubling the population size for three-stage selection in generation BC_1 ($m = 2$, $n_1 = mn_2 = mn_3$) resulted in Q10 values that were comparable to those reached with constant population sizes but using 20 more individuals per generation ($n_1' = n_2' = n_3' = n_2 + 20$). Using $m = 2$ required more SM but less HT assays than $m = 1$. A similar effect was observed for $m = 5$ and $n_1' = n_2' = n_3' = n_2 + 40$. However, here the increase in the number of required SM assays was considerable, while the reduction in the number of required HT assays was only small.

In conclusion, three-stage selection can be employed to put a stronger emphasis on the reduction of the donor segment attached to the target gene, and using two times larger population sizes in generation BC_1 ($m = 2$) than in BC_2 and BC_3 allows to shift the effort in the lab from HT to

SM assays compared to constant population size in all backcross generations ($m = 1$). These effects can be exploited without a reduction in the overall Q10 values. However, neither genetic advantages nor a reduction in the required marker assays supported employing five times larger populations in generation BC_1 ($m = 5$) than in generations BC_2 and BC_3.

Relative costs of three-stage selection

To compare the costs of three-stage selection in generation BC_1 with those of two-stage selection, we assumed cost ratios of 200:1 to 10:1 for the costs of one HT assay (comprising all marker loci on the linkage map) in relation to one SM assay (for one SM locus). First, the number of marker assays required to reach a given Q10 value with three-stage selection was determined from the simulations presented in Table 4, and the number of marker assays required to reach this Q10 value with two-stage selection was determined from the simulations presented in Table 3. Then the costs required with three-stage selection were determined with the above cost ratios and were set in relation to the costs that were required with two-stage selection (Fig. 1). For example, with a cost ratio of 200:1 for HT:SM assays (first diagram in Fig. 1) reaching the Q10 value of 99% with three-stage selection and $m = 5$ required 0.85 times the costs that were required to reach the Q10 value of 99% with two-stage selection. Three-stage selection with $m = 1$ required 0.77, and three-stage selection with $m = 2$ required 0.74 times the costs of two stage selection.

From the cost comparisons, we conclude that three-stage selection reaches a given Q10 value with less cost than two-stage selection, regardless of the cost ratio of HT:SM assays. If the aspired Q10 values are 99% or less, then doubling the population size in generation BC_1 provides a means to further reduce the costs required for the marker analyses.

Acknowledgements We thank the anonymous reviewers and the editor for their helpful suggestions. In particular, we gratefully acknowledge the comments of one reviewer that considerably improved the manuscript. We thank Gregory Mahone for proof-reading the manuscript.

References

Bouchez A, Hospital F, Causse M, Gallais A, Charcosset A (2002) Marker-assisted introgression of favorable alleles at quantitative trait loci between maize elite lines. Genetics 162:1945–1959

Collard BCY, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Phil Trans R Soc 363:557–572

Falke KC, Frisch M (2011) Power and false positive rate in QTL detection with near-isogenic line libraries. Heredity 106:576–584

Falke KC, Miedaner T, Frisch M (2009) Selection strategies for the development of rye introgression libraries. Theor Appl Genet 119:595–603

Frisch M, Melchinger AE (2001) Marker-assisted backcrossing for introgression of a recessive gene. Crop Sci 41:1485–1494

Frisch M, Melchinger AE (2001) Marker-assisted backcrossing for simultaneous introgression of two genes. Crop Sci 41:1716–1725

Frisch M, Bohn M, Melchinger AE (1999) Comparison of selection strategies for marker-assisted backcrossing of a gene. Crop Sci 39:1295–1301

Frisch M, Bohn M, Melchinger AE (1999) Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. Crop Sci 39:967–975

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535

Hospital F, Charcosset A (1997) Marker-assisted introgression of quantitative trait loci. Genetics 147:1469–1485

Hospital F, Chevalet C, Mulsant P (1992) Using markers in gene introgression breeding programs. Genetics 132:1199–1210

Maurer HP, Melchinger AE, Frisch M (2008) Population genetic simulation and data analysis with Plabsoft. Euphytica 161:133–139

Peleman JD, van der Voort JR (2003) Breeding by design. Trends Plant Sci 7:330–334

Prigge V, Maurer HP, Mackill DJ, Melchinger AE, Frisch M (2008) Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. Theor Appl Genet 116:739–744

Prigge V, Melchinger AE, Dhillon BS, Frisch M (2009) Efficiency gain of marker-assisted backcrossing by sequentially increasing marker densities over generations. Theor Appl Genet 119:23–32

Ragot M, Lee M (2007) Marker-assisted selection in maize: current status, potential, limitations and perspectives from the private and public sectors. In: Guimaraes EP, Ruane J, Scherf BD, Sonnino A, Dargie JD (eds) Marker-assisted selection. Current status and future perspectives in crops, livestock, forestry and fish. FAO, Rome, pp 117–150

Ribaut JM, Jiang C, Hoisington D (2002) Simulation experiments on efficiencies of gene introgression by backcrossing. Crop Sci 42:557–565

Schön CC, Melchinger AE, Boppenmaier J, Brunklaus-Jung E, Herrmann RG, Seitzer JF (1994) RFLP mapping in maize: quantitative trait loci affecting testcross performance of elite European flint lines. Crop Sci 34:378–389

Syvänen AC (2005) Toward genome-wide SNP genotyping. Nat Genet 37:S5–S10

Chapter 3

Efficient marker-assisted backcross conversion of seed parent lines to cytoplasmic male sterility[1]

[1]Herzog, E, & Frisch, M. 2013. Efficient marker-assisted backcross conversion of seed-parent lines to cytoplasmic male sterility. *Plant Breeding*, **132**(1), 33-41.

Plant Breeding **132**, 33–41 (2013)
doi:10.1111/pbr.12021

Efficient marker-assisted backcross conversion of seed-parent lines to cytoplasmic male sterility

Eva Herzog[1] and Matthias Frisch[1,2]

[1]Institute of Agronomy and Plant Breeding II, Justus Liebig University, D-35392, Giessen, Germany; [2]Corresponding author, E-mail: matthias.frisch@uni-giessen.de

With 2 figures and 5 tables

Received July 2, 2012 / Accepted September 29, 2012
Communicated by H.-P. Piepho

Abstract

For many crops, cytoplasmic male sterility (CMS) is a cornerstone of hybrid production. Efficient conversion of elite lines to CMS by marker-assisted backcrossing is therefore desirable. In contrast to gene introgression, for which donor segments around target genes have to be considered, background selection for CMS conversion focuses solely on recovery of the recurrent parent genome. The optimal selection strategies for CMS conversion will consequently differ from those for gene introgression and have not yet been investigated. The objectives of our study were to evaluate and optimize the resource requirements of CMS conversion programmes and to determine the most cost-effective use of single-marker (SM) and high-throughput (HT) assays for this purpose. We conducted computer simulations for CMS conversion of genetic models of sugar beet, rye, sunflower and rapeseed. CMS conversion required fewer resources than gene introgression with respect to population size, marker data points and number of backcross generations. Combining HT assays in early backcross generations with SM assays in advanced backcross generations further increased the cost-efficiency of CMS conversion for a broad range of cost ratios.

Key words: cytoplasmic male sterility — simulation study — high-throughput markers — hybrid breeding — marker-assisted backcrossing

Cytoplasmic male sterility (CMS) in plants is a maternally inherited condition, which inhibits the production of functional pollen. It is mediated by plant mitochondrial genomes and the interaction of mitochondrial and nuclear genes (Chase 2007). In seed crops such as rye, sunflower, rice and rapeseed, CMS plus nuclear restoration of male-fertility in F_1 progeny is essential for large-scale production of hybrid seeds. CMS is a mainstay for hybrid breeding and seed production in sugar beet and rye (Hagihara et al. 2005, Tomerius et al. 2008). For some crops such as *Brassica oleracea*, where the use of CMS in hybrid breeding is a comparatively new system, conversion of existing elite lines to CMS is required. For rapeseed (*Brassica napus*), in which the genetic basis of adapted germplasm is relatively narrow (Gehringer et al. 2007), CMS conversion of newly developed lines is used after the introduction of new genetic variation into the breeding pool. Moreover, it has been recognized in maize and rice that cytoplasmic uniformity can lead to vulnerability to pathogens (Pring and Lonsdale 1989, Dalmacio et al. 1995). For such crops, it may be important to convert existing lines to newly identified CMS systems to reduce maternally inherited disease susceptibility.

New CMS donors used in early cycles of hybrid breeding programmes are often poorly adapted or wild relatives of cultivated crops (Hanson and Bentolila 2004). Complete recovery of the converted elite genotypes is therefore desirable. Typically, elite lines are selected as fertile maintainers and converted to CMS by backcrossing. As thousands of lines often are to be converted, breeders will seek to devote as little resources as possible to the conversion of a single line.

In commercial breeding programmes, dense marker maps are available for major crops. In combination with high-throughput (HT) marker systems based on single nucleotide polymorphisms (SNPs), they can speed up the backcross process by marker-assisted background selection (Gupta et al. 2010).

In the field of single-marker (SM) assays, the Competitive Allele Specific PCR (KASPar) assay has quite recently emerged. KASPar is a SNP detection system, which is cost-effective for genotyping small subsets of SNP markers. It enables the combined use of HT and SM assays for SNP genotyping at different stages in marker-assisted breeding programmes, given that a SNP set exists which is inter-convertible between KASPar and HT marker platforms (Chen et al. 2010, Mammadov et al. 2012). An advantage of HT assays is fast and cost-effective screening of large populations with a high number of marker data points. However, while with HT assays such as SNP chips, all markers need to be analysed in every backcross generation, SM assays allow for analysing only those marker loci which are not yet fixed for the desired alleles in advanced backcross generations. A combination of HT assays in early backcross generations with SM assays in advanced backcross generations has the potential to increase the cost-effectiveness of background selection for gene introgression (Herzog and Frisch 2011).

For gene introgression, background selection focuses on both reduction of donor segments around target genes and recurrent parent genome recovery. In contrast, in CMS conversion programmes, background selection solely focuses on fast and complete recurrent parent genome recovery. Moreover, as no preselection for target genes is conducted, all individuals from a backcross are subjected to background selection. This results in higher selection intensity and hence a greater selection response per backcross generation. However, it will also substantially increase the number of required marker data points. The optimal strategies for using molecular markers for CMS conversion will consequently differ from those for gene introgression and have not yet been investigated for major CMS crops. Depending on the genome size of a crop species, population size, marker density and use of HT and/or SM marker systems need to be optimized.

The goal of our study was to investigate, with computer simulations, CMS conversion in sugar beet, rye, sunflower and rape-

seed with two to four backcross generations. In particular, our objectives were (i) to assess recurrent parent genome recovery with different marker densities and to investigate the effect of increasing population size per backcross generation, (ii) to evaluate the resource requirements for recovering varying target levels of recurrent parent genome while minimizing the number of marker data points, and (iii) to determine the most efficient use of SM and HT assays for different cost ratios of HT/SM.

Material and Methods

Simulations were conducted assuming no interference in crossover formation. Each simulation was replicated 10 000 times to reduce sampling effects and to obtain results with high numerical accuracy and a small standard error. The 10% quantile (Q10), the arithmetic mean and the standard deviation of the probability distribution of the proportion of recipient genome in the entire genome of selected individuals (in percentage) were determined in every backcross generation to measure recurrent parent genome recovery.

Q10 values were included as they allow inferences about the probability to reach a certain level of recurrent parent genome. For example, a Q10 value of 96% can be interpreted as 'with a probability of 0.9 a recurrent parent genome proportion >96% can be achieved'. The arithmetic mean does not allow such probability inferences in advanced backcross generations, when the distribution of recurrent parent genome is getting more skewed.

We investigated four different genetic models that represent different crop species for which CMS is used in hybrid seed production. Model 1 represented sugar beet (*Beta vulgaris*) and cabbage (*B. oleracea*) and had n = 9 chromosomes of 100 cM length (cf. Weber et al. 1999, cf. Iniguez-Luy et al. 2009). Model 2 represented rye (*Secale cereale*) and had n = 7 chromosomes of 100 cM length (cf. Gustafson et al. 2009). Model 3 represented sunflower (*Helianthus annuus*) and had n = 17 chromosomes of 80 cM length (cf. Tang et al. 2002). Model 4 represented rapeseed (*B. napus*) and had n = 19 chromosomes of 140 cM length (cf. Piquemal et al. 2005). These models are hereafter referred to as sugar beet, rye, sunflower and rapeseed, respectively.

Markers for genome-wide background selection were assumed to be equally spaced. We considered different marker densities: two markers per chromosome (2M/chr), three markers per chromosome (3M/chr), as well as marker distances between two adjacent loci of 20, 10, 5 and 2 cM. For 2M/chr and 3M/chr, markers divided the chromosomes in 3 or 4 equal parts, respectively. For marker densities of 20, 10, 5 and 2 cM, the first marker and last marker of each chromosome were placed on the telomeres.

Each backcross scheme started by crossing two homozygous parents (CMS donor and recipient), which were polymorphic at all loci. The CMS recipient carried the desirable alleles at all loci of the genome, while the donor carried no desirable alleles. The CMS recipient was assumed to be a fertile maintainer. An F_1 individual was created by crossing CMS donor and recipient. This F_1 individual was backcrossed to the recipient to create n_1 BC_1 individuals. The n_1 BC_1 individuals were subjected to genome-wide background selection. A selection index $i = \Sigma_m x_m$ was constructed, where summation is over markers and x_m is the number of recurrent parent alleles at the mth marker. The plant with the highest value of i was selected and backcrossed to the recipient. For each of the four genetic models, we investigated two to four backcross generations t and constant population sizes of n_t ranging from 10 to 200 individuals.

For calculating the relative costs of different marker strategies, the resource requirements for target Q10 values of 96% in generation BC_2 and 99% in generation BC_3 with a marker density of 10 cM were determined. One HT assay included genotyping one individual for all markers on the linkage map. One SM assay corresponded to one locus and thus one marker data point. For estimating the total number of required marker data points for SM assays, only marker loci not yet fixed for the recipient allele were analysed in advanced backcross generations. We took into account cost ratios of HT : SM of 200 : 1, 100 : 1, 50 : 1, 20 : 1 and 10 : 1. To give an example in absolute costs, a cost ratio of HT/SM of 100 : 1 corresponded to costs of € 50 for analysing all SNP background marker loci with a SNP chip, and € 0.5 for analysing one SNP marker locus with a KASPar assay. We compared the costs of using only HT assays in all generations of the backcross conversion programme (strategy HT) to the costs of using only SM assays in all generations of the backcross conversion programme (strategy SM). In this case, the costs for strategy SM were set to 1.

In addition, for two-generation programmes, we compared the costs of a combined strategy that relied on HT assays in generation BC_1 and SM assays in generation BC_2 (strategy Combined A) to the costs of strategy SM and strategy HT. In this case, the costs of strategy SM and strategy HT were set to 1, respectively. For three-generation programmes, we compared the costs of a strategy using HT assays in generation BC_1 and SM assays in generations BC_2 and BC_3 (strategy Combined B) to the costs of strategy HT. We also compared the costs of a strategy using HT assays in generations BC_1 and BC_2, and SM assays in generation BC_3 (strategy Combined C) to the costs of strategy HT. In both cases, the costs for strategy HT were set to 1.

Results

For a marker density of 20 cM and constant population sizes of n_t = 40, 80, 120, 160, 200 individuals per backcross generation, the Q10 values recovered in generations BC_1 and BC_2 were higher for genetic models with shorter genomes (Tables 1–4). Q10 values for rye were 2.9–3.8% higher than for rapeseed, while for sugar beet and sunflower, intermediate Q10 values were recovered. The differences in Q10 values between the genetic models that were observed in generation BC_2 diminish in advanced backcross generations.

Genetic models with shorter genomes had fewer and shorter fragments of donor genome in generations BC_1 and BC_2 (Tables 1–4). In generations BC_1 and BC_2, the length of donor fragments is decreasing to a greater extent with increasing n_t in genetic models with shorter genomes. The average length of donor fragments is decreasing by about 39% in sugar beet, 30% in rye, 28% in sunflower and 20% in rapeseed if n_t is increased from 40 to 200 individuals in generation BC_2. The average length of donor fragments ranged between 32% and 46% of marker distance for rye and 88–110% for rapeseed in generation BC_2. In advanced backcross generations, the differences in the length of donor fragments between population sizes and genetic models diminish considerably.

Genetic models with shorter genomes required fewer marker data points (Tables 1–4). For a population size of n_t = 200, rapeseed required about four times as many marker data points as rye, about three times as many marker data points as sugar beet, and about twice as many marker data points as sunflower. For all genetic models, the major proportion of marker data points had to be analysed in generations BC_1 and BC_2. For example, for sugar beet, 98.2–99.8% of marker analyses had to be conducted in generations BC_1 and BC_2. From generation BC_3 –BC_4, marker data points were no longer or only marginally increasing, indicating complete fixation. This also held true for higher marker densities of 10, 5 and 2 cM (data not shown).

For all four genetic models, Q10 values of >90% could be recovered in generation BC_2 with low marker densities of 2M/chr or 3M/chr and n_t = 10–20 individuals per backcross generation (Fig. 1). Q10 values increased considerably for all four investigated genetic models when population size was increased from n_t = 10 to n_t = 40–50 individuals per backcross generation, irrespective of marker density.

Table 1: Sugar beet: recovered proportion of recurrent parent genome (Q10, $\bar{x}$, s_x), required number of marker data points (MDP) for single-marker assays, number of donor fragments ($\bar{x}$, s_x) and length of donor fragments in cM ($\bar{x}$, s_x) in generations BC_1–BC_4 with genome-wide background selection with constant population sizes n_t = 40, 80, 120, 160, 200 and equally spaced markers (marker density 20 cM) (Note that the number of required high-throughput assays can be easily obtained by multiplying n_t by the number of backcross generations)

		Recurrent parent genome (%)				No. of donor fragments		Length of donor fragments (cM)	
n_t	BC_t	Q_{10}	$\bar{x}$	s_x	MDP	$\bar{x}$	s_x	$\bar{x}$	s_x
40	BC_1	84.56	88.08	2.81	2160	6.48	1.71	33.10	28.16
	BC_2	97.23	98.71	1.07	2682	2.00	1.31	11.60	10.20
	BC_3	99.32	99.80	0.32	2732	0.59	0.78	6.01	4.49
	BC_4	99.57	99.90	0.23	2732	0.32	0.58	5.70	4.36
80	BC_1	86.30	89.60	2.61	4320	6.06	1.68	30.90	26.79
	BC_2	98.19	99.27	0.75	5233	1.44	1.16	9.19	7.75
	BC_3	99.34	99.82	0.31	5276	0.54	0.74	6.08	4.56
	BC_4	99.62	99.91	0.22	5276	0.28	0.55	5.79	4.38
120	BC_1	87.32	90.34	2.45	6480	5.86	1.65	29.65	25.95
	BC_2	98.58	99.46	0.62	7754	1.20	1.09	8.06	6.63
	BC_3	99.38	99.83	0.30	7790	0.50	0.72	6.02	4.55
	BC_4	99.65	99.92	0.21	7790	0.27	0.53	5.64	4.38
160	BC_1	87.98	90.89	2.36	8640	5.71	1.64	28.71	25.25
	BC_2	98.81	99.55	0.55	10245	1.10	1.05	7.34	5.91
	BC_3	99.40	99.83	0.30	10273	0.51	0.73	5.89	4.42
	BC_4	99.65	99.91	0.21	10273	0.27	0.54	5.59	4.36
200	BC_1	88.46	91.30	2.29	10800	5.60	1.63	27.96	24.76
	BC_2	98.92	99.60	0.51	12723	1.01	1.01	7.07	5.58
	BC_3	99.40	99.84	0.29	12745	0.49	0.71	5.93	4.47
	BC_4	99.65	99.92	0.21	12745	0.27	0.53	5.65	4.35

Table 2: Rye: recovered proportion of recurrent parent genome (Q10, $\bar{x}$, s_x), required number of marker data points (MDP), number of donor fragments ($\bar{x}$, s_x) and length of donor fragments ($\bar{x}$, s_x) in generations BC_1–BC_4 with genome-wide background selection with constant population sizes n_t = 40, 80, 120, 160, 200 and equally spaced markers (marker density 20 cM) (Note that the number of required high-throughput assays can be easily obtained by multiplying n_t by the number of backcross generations)

		Recurrent parent genome (%)				No. of donor fragments		Length of donor fragments (cM)	
n_t	BC_t	Q_{10}	$\bar{x}$	s_x	MDP	$\bar{x}$	s_x	$\bar{x}$	s_x
40	BC_1	85.75	89.73	3.15	1680	4.71	1.51	30.50	26.45
	BC_2	98.05	99.28	0.87	2031	1.11	1.04	9.11	7.80
	BC_3	99.31	99.82	0.35	2047	0.42	0.66	5.93	4.47
	BC_4	99.66	99.91	0.24	2047	0.22	0.48	5.46	4.31
80	BC_1	87.79	91.35	2.85	3360	4.32	1.45	28.01	24.64
	BC_2	98.79	99.60	0.58	3953	0.78	0.89	7.23	5.79
	BC_3	99.35	99.84	0.33	3961	0.38	0.63	5.88	4.43
	BC_4	99.69	99.92	0.23	3961	0.20	0.46	5.57	4.34
120	BC_1	88.82	92.21	2.67	5040	4.14	1.41	26.34	23.47
	BC_2	98.93	99.66	0.51	5844	0.72	0.87	6.59	4.99
	BC_3	99.38	99.85	0.33	5847	0.36	0.62	5.89	4.50
	BC_4	99.71	99.92	0.23	5847	0.20	0.46	5.53	4.37
160	BC_1	89.55	92.78	2.56	6720	3.98	1.42	25.41	22.59
	BC_2	99.01	99.69	0.47	7715	0.68	0.83	6.40	4.72
	BC_3	99.39	99.85	0.32	7716	0.36	0.61	5.87	4.49
	BC_4	99.73	99.93	0.22	7716	0.18	0.45	5.60	4.31
200	BC_1	90.14	93.18	2.44	8400	3.88	1.38	24.58	21.87
	BC_2	99.02	99.70	0.46	9576	0.65	0.82	6.39	4.67
	BC_3	99.37	99.85	0.32	9577	0.35	0.60	6.09	4.54
	BC_4	99.72	99.93	0.22	9577	0.18	0.43	5.76	4.38

For all genetic models, Q10 values of $\geq$96% could be reached in generation BC_2. Minimum required marker densities for a Q10 value of 96% were 3M/chr for sugar beet and rye, 2M/chr for sunflower and 20 cM for rapeseed. For sugar beet and rye, there was a limit of recurrent parent genome that could be recovered, indicated by a plateau in the Q10 curves for marker densities of 2M/chr, 3M/chr and 20 cM. The population sizes per backcross generation for which the limit was reached depended on marker density and lay between n_t = 70–200 for sugar beet and between n_t = 50–150 for rye. For sunflower and rapeseed, the plateau was not reached with the highest investigated population size of n_t = 200.

The differences in Q10 values between marker densities were bigger in sugar beet and rye than in sunflower and rapeseed (Fig. 1). For example, for n_t = 100 individuals per backcross generation, the differences in Q10 values between a marker density of 3M/chr and 20 cM were 1.7% for sugar beet, 2.3% for rye, 0.4% for sunflower and 1.0% for rapeseed. The maximum

Table 3: Sunflower: recovered proportion of recurrent parent genome (Q10, $\bar{x}$, s_x), required number of marker data points (MDP), number of donor fragments ($\bar{x}$, s_x) and length of donor fragments ($\bar{x}$, s_x) in generations BC_1–BC_4 with genome-wide background selection with constant population sizes n_t = 40, 80, 120, 160, 200 and equally spaced markers (marker density 20 cM) (Note that the number of required high-throughput assays can be easily obtained by multiplying n_t by the number of backcross generations)

		Recurrent parent genome (%)				No. of donor fragments		Length of donor fragments (cM)	
n_t	BC_t	Q_{10}	$\bar{x}$	s_x	MDP	$\bar{x}$	s_x	$\bar{x}$	s_x
40	BC_1	82.19	85.06	2.31	3400	11.81	2.17	34.39	26.21
	BC_2	95.46	97.08	1.24	4424	4.94	1.83	16.11	14.47
	BC_3	99.20	99.69	0.36	4620	1.23	1.14	6.84	5.53
	BC_4	99.57	99.87	0.21	4626	0.59	0.80	5.75	4.42
80	BC_1	83.61	86.23	2.12	6800	11.35	2.13	32.99	25.73
	BC_2	96.45	97.84	1.04	8692	4.17	1.69	14.12	12.53
	BC_3	99.37	99.77	0.29	8976	1.01	1.03	6.24	4.63
	BC_4	99.59	99.89	0.20	8976	0.53	0.75	5.82	4.41
120	BC_1	84.37	86.88	2.03	10200	11.05	2.13	32.30	25.48
	BC_2	96.94	98.20	0.94	12904	3.76	1.64	13.01	11.44
	BC_3	99.40	99.78	0.28	13249	0.98	1.00	6.10	4.49
	BC_4	99.60	99.89	0.19	13249	0.52	0.75	5.63	4.35
160	BC_1	84.93	87.36	1.98	13600	10.84	2.11	31.72	25.24
	BC_2	97.29	98.45	0.86	17076	3.47	1.60	12.15	10.65
	BC_3	99.41	99.79	0.27	17464	0.94	1.00	6.06	4.59
	BC_4	99.62	99.90	0.19	17464	0.49	0.71	5.75	4.45
200	BC_1	85.30	87.68	1.95	17000	10.72	2.07	31.26	25.02
	BC_2	97.52	98.61	0.81	21244	3.24	1.55	11.64	10.04
	BC_3	99.41	99.79	0.27	21670	0.93	0.98	6.04	4.51
	BC_4	99.61	99.89	0.19	21670	0.49	0.72	5.81	4.45

Table 4: Rapeseed: recovered proportion of recurrent parent genome (Q10, $\bar{x}$, s_x), required number of marker data points (MDP), number of donor fragments ($\bar{x}$, s_x) and length of donor fragments ($\bar{x}$, s_x) in generations BC_1–BC_4 with genome-wide background selection with constant population sizes n_t = 40, 80, 120, 160, 200 and equally spaced markers (marker density 20 cM) (Note that the number of required high-throughput assays can be easily obtained by multiplying n_t by the number of backcross generations)

		Recurrent parent genome (%)				No. of donor fragments		Length of donor fragments (cM)	
n_t	BC_t	Q_{10}	$\bar{x}$	s_x	MDP	$\bar{x}$	s_x	$\bar{x}$	s_x
40	BC_1	81.07	83.37	1.89	6080	19.60	2.87	45.13	38.58
	BC_2	94.19	95.65	1.11	8116	10.57	2.60	21.92	20.56
	BC_3	98.68	99.29	0.44	8644	3.69	1.86	10.27	9.21
	BC_4	99.64	99.86	0.16	8712	1.28	1.17	5.75	4.46
80	BC_1	82.19	84.36	1.78	12160	19.16	2.86	43.45	37.65
	BC_2	95.12	96.39	0.99	15992	9.59	2.53	20.00	18.78
	BC_3	99.07	99.54	0.34	16861	2.88	1.71	8.42	7.34
	BC_4	99.66	99.87	0.15	16926	1.20	1.13	5.74	4.39
120	BC_1	82.86	84.89	1.69	18240	18.87	2.87	42.59	37.14
	BC_2	95.62	96.79	0.91	23797	9.09	2.46	18.77	17.61
	BC_3	99.24	99.64	0.30	24955	2.53	1.64	7.47	6.38
	BC_4	99.67	99.88	0.15	25011	1.15	1.11	5.70	4.36
160	BC_1	83.26	85.25	1.64	24320	18.71	2.84	41.93	36.68
	BC_2	95.91	97.04	0.86	31551	8.64	2.38	18.23	17.00
	BC_3	99.34	99.70	0.26	32970	2.29	1.55	6.96	5.68
	BC_4	99.68	99.88	0.15	33014	1.09	1.09	5.77	4.40
200	BC_1	83.58	85.51	1.58	30400	18.58	2.84	41.50	36.50
	BC_2	96.16	97.22	0.82	39284	8.39	2.37	17.64	16.35
	BC_3	99.40	99.73	0.24	40941	2.18	1.53	6.63	5.33
	BC_4	99.68	99.88	0.14	40974	1.09	1.10	5.69	4.36

Q10 values recovered in generation BC_2 depended on genome length and were 99.60% for sugar beet, 99.99% for rye, 98.09% for sunflower and 96.50% for rapeseed. Increasing marker density from 10 to 5 or 2 cM did not substantially increase Q10 values. This held true for all investigated genetic models. Moreover, marker densities of 5 and 2 cM incurred very high numbers of marker data points (data not shown).

The optimum designs that minimized the required number of marker data points for target Q10 values of 96–99% in generation BC_2 employed marker densities of 2M/chr–10 cM for sugar beet and rye (Table 5). For the sunflower model, a Q10 value of 98% could only be reached with a marker density of 5 cM and 68 000 marker data points. For rapeseed, a Q10 value of 96% in generation BC_2 could only be reached with a marker density of 20 cM and about 33 000 marker data points. Higher target Q10 values could not be reached in generation BC_2 for this model. For all four genetic models, two-generation programmes incurred substantially more marker data points than

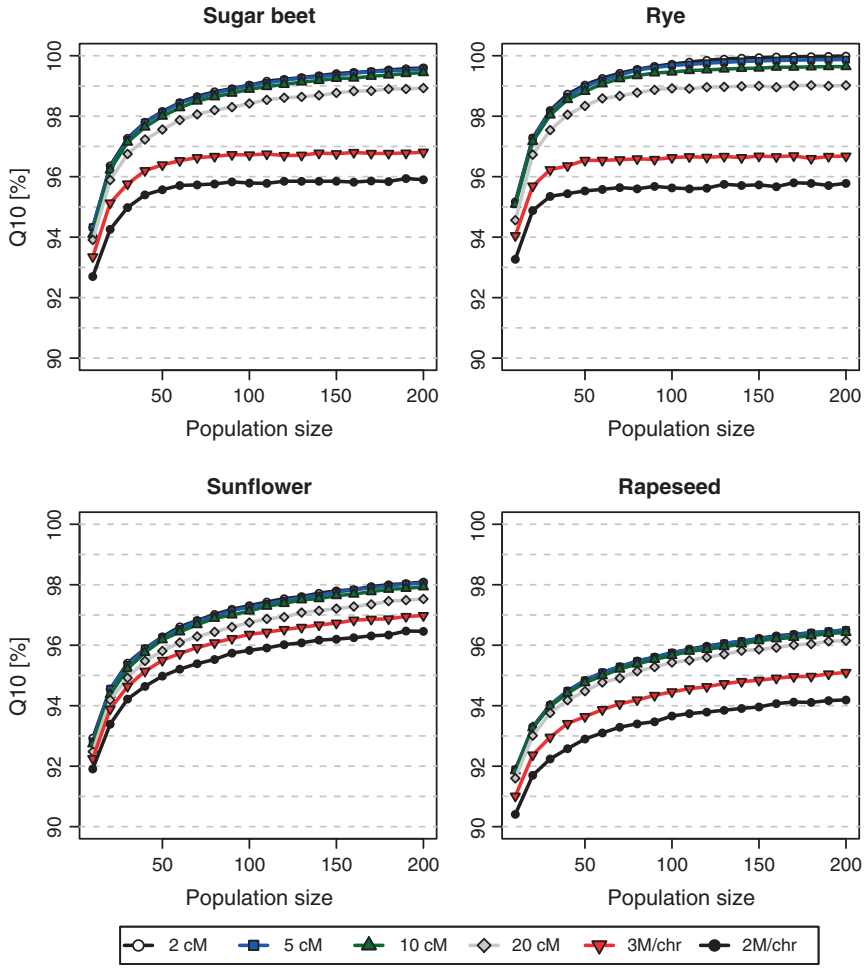

Fig. 1: Q10 values recovered in generation BC_2 with genome-wide background selection (marker densities 2M/chr, 3M/chr and 2, 5, 10, 20 cM) with constant population sizes n_t = 20–300 and equally spaced markers for four genetic models

three-generation programmes. The shorter genomes of sugar beet and rye required 3.6–11.2 times as many marker data points for two-generation programmes as for three-generation programmes. For sunflower and rapeseed, two-generation programmes required 28.8–60.2 times as many marker data points as three-generation programmes.

For two-generation programmes, strategy HT was 0.05–4.10 times as expensive as strategy SM for recovering a target Q10 value of 96%, depending on genetic model and cost ratio of HT : SM (Fig. 2a). Strategy Combined A was more cost-effective than strategy HT, indicated by the smaller range of relative costs and the smaller slopes of the cost curves (Fig. 2b). Which marker strategy was cheapest depended on the cost ratio of HT : SM and the genetic model. For sugar beet, strategy HT was the cheapest strategy for cost ratios of HT : SM of 10 : 1–35 : 1 (Fig. 2c). For cost ratios ranging between 35 : 1 and 100 : 1, strategy Combined A was cheapest (Fig. 2b). For cost ratios of HT : SM >100 : 1, strategy SM was cheapest. If the choice was between either strategy HT or strategy SM, strategy HT should be used for cost ratios of HT : SM of 10 : 1–60 : 1 (Fig. 2a). For longer genomes, using HT assays became relatively cheaper compared with SM assays. For rapeseed, strategy HT was the cheapest strategy for recovering a target Q10 value of 96% for cost ratios of HT : SM of up to 190 : 1 (Fig. 2a). For cost ratios >190 : 1, strategy Combined A was cheapest (Fig. 2b). Strategy SM was never cheaper than strategy Combined A.

For three-generation programmes with a target Q10 value of 99%, the use of HT assays became less efficient compared with SM assays (Fig. 2d), indicated by steeper cost curves. Strategy Combined C was equivalent to or cheaper than strategy HT for nearly all investigated scenarios (Fig. 2f).

Discussion

Genetic models

Computer simulations and model calculations are considered robust and useful tools for the optimization of breeding programmes (Prigge et al. 2008, Tomerius et al. 2008). However, the validity of simulated results for real breeding applications is influenced by the theoretical assumptions for the underlying genetic model.

We used a Poisson procedure for modelling crossover formation during meiosis, assuming no interference in crossover formation as

Table 5: Optimum designs for recovering Q10 values of 96–99% [marker density, population size, no. of backcross generations, no. of marker data points (MDP)] in two vs. three backcross generations with genome-wide background selection if the number of MDP is minimized (Note that the number of required high-throughput assays can be easily obtained by multiplying n_t by the number of backcross generations)

Genetic model	Q10 (%)	n_t	Marker density	No. of MDP
No. of BC generations = 2				
Sugar beet	96	40	3M/chr	1305
9 × 100 cM	97	40	20 cM	2684
	98	70	20 cM	4600
	99	120	10 cM	14 172
Rye	96	30	3M/chr	749
7 × 100 cM	97	30	20 cM	1541
	98	40	20 cM	2030
	99	60	10 cM	5475
Sunflower	96	120	2M/chr	5059
17 × 80 cM	97	130	20 cM	13 948
	98	190	5 cM	68 411
	99	–	–	–
Rapeseed	96	170	20 cM	33 476
19 × 140 cM	97	–	–	–
	98	–	–	–
	99	–	–	–
No. of BC generations = 3				
Sugar beet	96	–	–	–
9 × 100 cM	97	10	2M/chr	240
	98	10	20 cM	746
	99	20	20 cM	1421
Rye	96	–	–	–
7 × 100 cM	97	10	2M/chr	181
	98	10	20 cM	563
	99	10	10 cM	1028
Sunflower	96	–	–	–
17 × 80 cM	97	10	2M/chr	485
	98	20	3M/chr	1394
	99	30	20 cM	3513
Rapeseed	96	10	2M/chr	556
19 × 140 cM	97	20	3M/chr	1618
	98	20	20 cM	4448
	99	80	20 cM	16 863

proposed by Haldane (1919). This approach has the advantage of applicability for a broad range of scenarios, as has been discussed in detail in the study by Frisch and Melchinger (2001). Further necessary simplifications for the sake of generality include the assumptions of perfect fertility, no natural selection at gamete or zygote level, unchanged recombination frequencies and Mendelian segregation in any cross. This will not hold true in all cases, especially if CMS donors are unadapted wild relatives. For such wide crosses, the simulations might underestimate the actual resource requirements and/or overestimate recovered Q10 values. On the other hand, in advanced cycles of hybrid breeding programmes, adapted lines often are available as CMS donors, which might be similar to the recipient lines. In these cases, complete recovery of an elite genotype might be achieved with less resources or in shorter time.

The reader should be aware that the presented simulation approach does not cover every detail of the complex biological processes, which might underlie any specific cross. Conclusions drawn from simulated data should therefore be interpreted as guidelines and might require adjustment in specific breeding programmes.

Population size

In a simulation study on the introgression of one dominant target gene, Prigge et al. (2009) employed the same genetic model for sugar beet that was used in the present study. With a marker density of 20 cM and n_t = 40–200 individuals per backcross generation, they recovered Q10 values in generation BC_2 that were approximately 3–4% lower than in the present study (Table 1). The greater selection response in CMS conversion can be explained by the lack of preselection for the target gene and the lack of donor genome attached to the target gene. Consequently, CMS conversion required considerably smaller population sizes than gene introgression.

In generation BC_2, Q10 values increased considerably for all four genetic models when population size was increased from n_t = 10 to n_t = 40–50 individuals (Fig. 1). For sugar beet and rye, a plateau in the Q10 curves was observed. This limit of recurrent parent genome recovery is caused by the limited estimation accuracy of a given marker density. The wider adjacent markers are spaced, the more likely it is that segments of recurrent parent genome between markers go unnoticed. Sugar beet and rye had fewer and shorter donor fragments in generation BC_2, which were still considerably decreased with increasing population size n_t (Tables 1–4). For rye, for which the plateau is reached at n_t = 120 with a marker density of 20 cM, the average length of donor fragments is only 6.59 cM and consequently only about 33% of the distance between two adjacent markers (Table 2). As a consequence, the plateau is reached with smaller population sizes for lower marker densities. Increasing population size beyond the number of individuals for which the plateau is reached (Fig. 1) is not economic.

We conclude that recurrent parent genome recovery is maximized for all four genetic models with population sizes of $n_t \geq$ 40–50 individuals per backcross generation. For rye and sugar beet, population sizes should not exceed n_t = 50–150 and n_t = 70–200 individuals, respectively, depending on marker density. For sunflower and rapeseed, population sizes of $n_t > 200$ still have positive effects.

Marker density

It has been estimated for backcross programmes that a target Q10 value of at least 96% should minimize the risk of undesirable effects from unadapted donor genome (Prigge et al. 2009). For sugar beet and rye, a Q10 level of about 96% could be recovered in generation BC_2 with a marker density of 3M/chr and n_t = 40–60 individuals per backcross generation (Fig. 1). For sunflower, the Q10 value of 96% could be reached with a marker density of 2M/chr, indicating that two markers per chromosome are sufficient for controlling short chromosomes (Fig. 1). We therefore conclude that for CMS conversion, a threshold Q10 value of 96% in generation BC_2 can in most cases be reached with 2–3 markers per chromosome.

The differences in Q10 values between marker densities were bigger in sugar beet and rye than in sunflower and rapeseed (Fig. 1). For example, for a population size per backcross generation of n_t = 100, the differences in Q10 values between a marker density of 3M/chr and 20 cM were 1.7% for sugar beet, 2.3% for rye, 0.4% for sunflower and 1.0% for rapeseed. If marker density was increased from 3M/chr to 20 cM, the increase in the number of markers per chromosome was greater in genetic models with longer chromosomes, which partly accounts for the big gap in Q10 values. Moreover, increasing marker density shifted the frequency distribution of recurrent parent genome to the right and decreased the variance of the distribution in all four genetic models. The extent of these changes depended on chromosome number and length. The differences between marker densities were bigger for genetic models with a lower number of

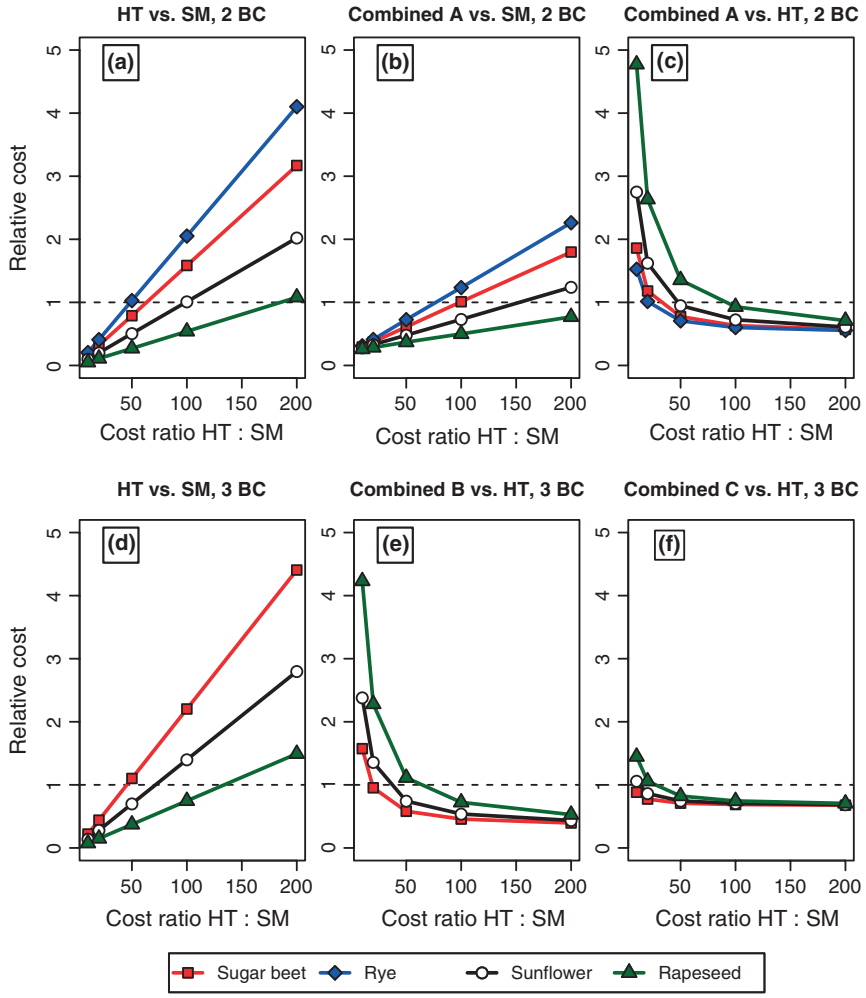

Fig. 2: Relative cost of different strategies of marker analysis plotted against the cost ratio of one high-throughput assays to one single-marker assay (Cost ratio HT : SM) for recovering Q10 values of 96% in generation BC_2, and Q10 values of 99% in generation BC_3 with a marker density of 10 cM HT, a strategy using only high-throughput assays in all backcross generations; SM, a strategy using only single-marker assays in all backcross generations; Combined A, a strategy using HT assays in generation BC_1 and SM assays in generation BC_2; Combined B, a strategy using HT assays in generation BC_1 and SM assays in generations BC_2 and BC_3; Combined C, a strategy using HT assays in generations BC_1 and BC_2 and HT assays in generation BC_3. HT, high-throughput; SM, single-marker

chromosomes. If chromosome number was comparable, the differences were bigger for genetic models with longer chromosomes. For sunflower, 2–3 equally spaced markers per chromosome seemed sufficient to get acceptable genome coverage for recurrent parent genome recovery. For rapeseed, sugar beet and rye, at least 6–8 equally spaced markers per chromosome, corresponding to a marker density of 20 cM, provide an adequate selection response.

Given that differences in Q10 values between marker densities were bigger (Fig. 1) and donor fragments on average shorter (Tables 1–4) in genetic models with shorter genomes, we conclude that it pays off more to invest in higher marker densities for sugar beet and rye than for sunflower and rapeseed.

For all four genetic models, hardly any differences in Q10 values could be observed between marker densities of 2, 5 and 10 cM (Fig. 1). However, marker densities of 5 and 2 cM incurred very high numbers of marker data points (data not shown). This was also observed in a previous simulation study on gene introgression (Herzog and Frisch 2011). The reason is that selection response is not limited by precise estimation of the genetic contribution of the recurrent parent, but by the limited number of recombination events that occur in two- or three-generation backcross programmes. We therefore conclude that it is not efficient to increase effective marker density beyond 10 cM, even though marker maps with higher density are available for SNPs.

Marker fixation

For all four genetic models, the major proportion of marker data points was incurred in generations BC_1 and BC_2. From generation BC_3–BC_4, the number of required marker data points is only marginally increasing (Tables 1–4). Accordingly, the population sizes at which the plateau of recurrent parent genome recovery is reached are diminishing in generations BC_3 and BC_4 due to marker fixation (data not shown). As a consequence, dif-

ferences in Q10 values and average length of donor fragments between the four genetic models disappear in generation BC_4. This indicates that recurrent parent genome recovery was no longer controlled for by markers and resulted in a reduction in selection response.

For gene introgression, Prigge et al. (2009) reported that the optimum backcross designs were characterized by increasing marker densities and population sizes. Due to the faster rate of marker fixation in CMS conversion programmes, we conclude that keeping a constant population size in each backcross generation, or increasing population size in advanced backcross generations, is only efficient for CMS conversion if it is also accompanied by an increase in marker density. Additional markers could be placed between the original markers analysed in previous generations to increase the precision of selection. For sugar beet, rye and sunflower, marker densities of 3M/chr in generations BC_1 and BC_2, and 20 cM in advanced backcross generations could decrease the loss of selection response. For rapeseed, we suggest that CMS conversion programmes could start with 20 cM in generations BC_1 and BC_2, followed by 10 cM in advanced backcross generations.

CMS conversion designs for different genetic models

In the present study, Q10 values of 96–98% could be reached in generation BC_2 for sugar beet and rye with a marker density of 20 cM and n_t = 30–70 individuals (Table 5). We therefore conclude that for these crops, two-generation programmes are suitable for CMS conversion.

If Q10 values >96% were aimed for in generation BC_2, sunflower required n_t = 130–190 individuals per backcross generation and marker densities of 20–5 cM. Moreover, a target Q10 value of 98% in generation BC_2 required about 68 000 marker data points (Table 5). For rapeseed, a Q10 value of 96% in generation BC_2 could only be reached with n_t = 170 individuals per backcross generation, a marker density of 20 cM, and about 33 000 marker data points. We conclude that for target Q10 values of 96–99%, three-generation conversion programmes are required for the longer genomes of sunflower and rapeseed.

With the exception of a Q10 value of 98% for sunflower, all Q10 levels could be reached with marker densities of 2M/chr–10 cM. Increasing marker density beyond 10 cM incurs high numbers of marker data points, but will not help to save additional backcross generations (cf. Fig. 1). This confirms that a marker density of 10 cM is sufficient for almost all backcross designs, as has also been previously observed (Herzog and Frisch 2011).

For all four genetic models, two-generation programmes required considerably more marker data points than the three-generation programmes (Table 5). We therefore conclude that three-generation CMS conversion programmes are also advantageous for shorter genomes if the focus of cost reduction is on the cost of marker analysis.

Relative costs of HT and SM assays

Different strategies of using HT and SM assays for CMS conversion with a marker density of 10 cM were compared by calculating their relative costs for cost ratios of HT/SM ranging from 200 : 1 to 10 : 1 (Fig. 2). For a Q10 value of 96% in generation BC_2, the relative costs of strategy HT compared with strategy SM ranged from 0.10 to 2.02 for sunflower (Fig. 2a). In a gene introgression study on maize with the same parameters, the relative costs ranged from 0.09 to 1.85 (Herzog and Frisch 2011). These genetic models are comparable with respect to genome length (1360 vs. 1600 cM) and number of background markers. For a given population size, the number of SM and HT assays are approximately in the same ratio for gene introgression and CMS conversion in generations BC_1 and BC_2. It can therefore be assumed that the relative costs we determined in the present study are to a certain extent also valid for background selection in gene introgression programmes.

For sugar beet and a target Q10 value of 96% in generation BC_2, strategy HT was cheapest up to a cost ratio of HT/SM of 35 : 1 (Fig. 2c). From a cost ratio of HT/SM of 35 : 1–100 : 1, strategy Combined A was cheapest (Fig. 2b). For higher cost ratios of HT/SM, strategy SM was the cheapest option. If the choice is between either strategy HT or strategy SM, strategy HT should be used up to a cost ratio of HT/SM of 60 : 1 (Fig. 2a). For sunflower and rapeseed, strategies involving HT assays became relatively cheaper. We therefore conclude that the use of HT assays for background selection is cost-efficient for two-generation CMS conversion programmes and crops with long genomes such as sunflower and rapeseed.

For three-generation programmes, strategy HT became less efficient compared with strategy SM, indicated by steeper cost curves (Fig. 2d). For sugar beet and a target Q10 value of 99% in generation BC_3, strategy HT was only cheaper than strategy SM up to a cost ratio of HT/SM of 45 : 1. This can be explained by the fact that for sugar beet, 98–99% of marker data points are incurred in generations BC_1 and BC_2 and most markers are already fixed in generation BC_3 (Tables 1–4).

For three-generation programmes, strategy Combined C was equivalent to or cheaper than strategy HT for nearly all investigated scenarios (Fig. 2f). Combining HT and SM assays in one backcross programme can pose a challenge as HT and SM platforms often require different types of markers. Recently, KASPar assays have become available, which allow for inexpensive analysis of small sets of SNPs (Chen et al. 2010). It has been shown that SNP markers can be inter-converted between KASPar and HT assays (Mammadov et al. 2012). Combinations of HT and SM thus have the potential to make marker-assisted background selection more cost-effective. We conclude that for three-generation CMS conversion programmes, HT assays should be used in generations BC_1 and BC_2, and SM assays in generation BC_3 for all investigated genetic models.

Acknowledgements

We thank the anonymous reviewers and the editor for their helpful suggestions. We greatly appreciated the comments of one reviewer that considerably improved the manuscript. We thank Gregory Mahone for proof-reading the manuscript.

References

Chase, C. D., 2007: Cytoplasmic male sterility: a window to the world of plant mitochondrial-nuclear interactions. Trends Genet. **23**, 81—90.

Chen, W., J. Mingus, J. Mammadov, J. E. Backlund, T. Greene, S. Thompson, and S. Kumpatla, 2010: KASPar: a simple and cost-effective system for SNP genotyping. In: Final program, abstract and exhibit guide of the XVIII international conference on the status of plant and animal genome research, San Diego, CA, 9–13 Jan 2010.

Dalmacio, R., D. S. Brar, T. Ishii, L. A. Sitch, S. S. Virmani, and G. S. Khush, 1995: Identification and transfer of a new cytoplasmic male sterility source from *Oryza perennis* into indica rice (*O. sativa*). Euphytica **82**, 221—225.

Frisch, M., and A. E. Melchinger, 2001: Length of the intact donor chromosome segment around a target gene in marker-assisted backcrossing. Genetics **157**, 1343—1356.

Gehringer, A., R. Snowdon, T. Spiller, P. Basunanda, and W. Friedt, 2007: New oilseed rape (*Brassica napus*) hybrids with high levels of heterosis for seed yield under nutrient-poor conditions. Breed. Sci. **57**, 315—320.

Gupta, P. K., P. Langridge, and R. R. Mir, 2010: Marker-assisted wheat breeding: present status and future possibilities. Mol. Breeding **26**, 145—161.

Gustafson, J. P., X. F. Ma, V. Korzun, and J. W. Snape, 2009: A consensus map of rye integrating mapping data from five mapping populations. Theor. Appl. Genet. **118**, 793—800.

Hagihara, E., N. Itchoda, Y. Habu, S. Iida, T. Mikami, and T. Kubo, 2005: Molecular mapping of a fertility restorer gene for Owen cytoplasmic male sterility in sugar beet. Theor. Appl. Genet. **111**, 250—255.

Haldane, J. B. S., 1919: The combination of linkage values and the calculation of distance between the loci of linkage factors. J. Genet. **8**, 299—309.

Hanson, M. R., and S. Bentolila, 2004: Interactions of mitochondrial and nuclear genes that affect male gametophyte development. Plant Cell **16**, S154—S169.

Herzog, E., and M. Frisch, 2011: Selection strategies for marker-assisted backcrossing with high-throughput marker systems. Theor. Appl. Genet. **123**, 251—260.

Iniguez-Luy, F. L., L. Lukens, M. W. Farnham, R. M. Amasino, and T. C. Osborn, 2009: Development of public immortal mapping populations, molecular markers and linkage maps for rapid cycling *Brassica rapa* and *B. oleracea*. Theor. Appl. Genet. **120**, 31—43.

Mammadov, J., W. Chen, J. Mingus, S. Thompson, and S. Kumpatla, 2012: Development of versatile gene-based SNP assays in maize (*Zea mays* L.). Mol. Breeding **29**, 779—790.

Piquemal, J., E. Cinquin, F. Couton, C. Rondeau, E. Seignoret, I. Doucet, D. Perret, M. J. Villeger, P. Vincourt, and P. Blanchard, 2005: Construction of an oilseed rape (*Brassica napus* L.) genetic map with SSR markers. Theor. Appl. Genet. **111**, 1514—1523.

Prigge, V., H. P. Maurer, D. J. Mackill, A. E. Melchinger, and M. Frisch, 2008: Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. Theor. Appl. Genet. **116**, 739—744.

Prigge, V., A. E. Melchinger, B. S. Dhillon, and M. Frisch, 2009: Efficiency gain of marker-assisted backcrossing by sequentially increasing marker densities over generations. Theor. Appl. Genet. **119**, 23—32.

Pring, D. R., and D. M. Lonsdale, 1989: Cytoplasmic male sterility and maternal inheritance of disease susceptibilty in maize. Annu. Rev. Phytopathol. **27**, 483—502.

Tang, S., J. K. Yu, M. B. Slabaugh, D. K. Shintani, and S. J. Knapp, 2002: Simple sequence repeat map of the sunflower genome. Theor. Appl. Genet. **105**, 1124—1136.

Tomerius, A. M., T. Miedaner, and H. H. Geiger, 2008: A model calculation approach towards the optimization of a standard scheme of seed-parent line development in hybrid rye breeding. Plant Breeding **127**, 433—440.

Weber, W. E., D. C. Borchardt, and G. Koch, 1999: Combined linkage maps and QTLs in sugar beet (*Beta vulgaris* L.) from different populations. Plant Breeding **118**, 193—204.

Chapter 4

Selection strategies for marker-assisted background selection with chromosome-wise SSR multiplexes in pseudo-backcross programs for grapevine breeding[1]

[1]Herzog, E, Töpfer, R, Hausmann, L, Eibach, R, & Frisch, M. 2013. Selection strategies for marker-assisted background selection with chromosomewise SSR multiplexes in pseudo-backcross programs for grapevine breeding. *Vitis*, **52**(4), 193-196.

Vitis **52** (4), 193–196 (2013)

Selection strategies for marker-assisted background selection with chromosome-wise SSR multiplexes in pseudo-backcross programs for grapevine breeding

E. Herzog[1)], R.Töpfer[2)], L. Hausmann[2)], R. Eibach[2)] and M. Frisch[1)]

[1)] Institute of Agronomy and Plant Breeding II, Justus Liebig University, Giessen, Germany

[2)] Julius Kühn Institut (JKI), Federal Research Centre for Cultivated Plants, Institute for Grapevine Breeding Geilweilerhof, Siebeldingen, Germany

Summary

Organizing SSR markers located on one chromosome into PCR multiplexes has the potential to reduce the costs of marker analysis. The optimal selection strategies for such chromosome-wise multiplexes have not yet been investigated. We investigated with computer simulations three different selection strategies for gene introgression with a pseudo-backcross scheme and a marker density of one marker every 10 cM. Selecting individuals with the highest number of chromosomes carrying *V. vinifera* alleles at all background marker loci reduced the number of required multiplexes by 7.24-7.87 % in generations pBC_4-pBC_6 for population sizes n_t = 150-300 individuals per pseudo-backcross generation.

K e y w o r d s : gene introgression, simulation study, multiplex PCR, microsatellite, marker-assisted selection.

Introduction

American and Asian *Vitis* species carrying resistance genes against mildew disease have been employed in interspecific breeding programs (Di Gaspero and Cattonaro 2010, Töpfer *et al.* 2011). Along this line only one example has been described for a systematic development of introgression lines as described by the pioneering work of Alain Bouquet (Pauquet *et al.* 2001). This work turns out to be very time consuming as well as space and labor demanding. The development of molecular markers for early selection of seedlings with traits of agronomic interest is therefore of particular value (Eibach *et al.* 2007). Simple sequence repeats (SSRs) are useful genetic markers, as they are abundant in the genome, highly polymorphic and transferable between *V. vinifera* and related species (Salmaso *et al.* 2008, Vezzulli *et al.* 2008, Blanc *et al.* 2012). However, they have the disadvantage of low throughput compared to single nucleotide polymorphisms (SNPs), which limits their use in large-scale breeding programs. Organizing SSRs into PCR multiplexes considerably reduces the costs of marker analysis (Merdinoglu *et al.* 2005). Patocchi *et al.* (2009) suggested that organizing SSR markers located on one linkage group in one multiplex is applicable and advantageous. Efficient selection strategies for *V. vinifera* genome recovery in pseudo-backcross programs for gene introgression with such chromosome-wise multiplexes have not yet been investigated. The objective of our study was to compare with computer simulations different selection strategies for gene introgression with chromosome-wise multiplexes.

Material and Methods

Computer simulations were carried out assuming no interference in crossover formation. Each simulation was run 10,000 times in order to reduce sampling effects and to obtain stable results with small standard error.

The genetic model for grapevine consisted of 19 chromosomes (2 x 40 cM, 7 x 60 cM, 5 x 80 cM, 5 x 100 cM). This corresponded to a total genome length of 1400 cM. The marker for the dominant target gene was assumed to be a gene-based marker and was located on a 100 cM chromosome at 61 cM from the telomere. Background markers were equidistantly spaced with one marker every 10 cM, the first and last marker of each chromosome being placed on the telomeres.

A pseudo-backcross scheme with changing *V. vinifera* parents in every generation was investigated up to generation pBC_6. The goal was to recover as much *V. vinifera* genome as possible, irrespective from which of the parents. For chromosome-wise multiplexes it was assumed that one multiplex included genotyping all background marker loci located on one chromosome. This resulted in multiplexes comprising 5 to 11 SSRs. In advanced pseudo-backcross generations only those chromosomes were analyzed which did not yet carry *V. vinifera* alleles at all background marker loci.

The donor was heterozygous for the desired allele at the target locus. The *V. vinifera* parents could be distinguished from the donor at all marker loci. Initially, the donor and the first *V. vinifera* parent were crossed to produce n_{F1} F_1 individuals. From this F_1 population, one individual that carried the donor allele at the target locus was selected as parent for generation pBC_1. This individual was crossed to the second *V. vinifera* parent to produce n_1 pBC_1 individuals. From this pBC_1 population, one best individual was selected with the selection strategies described below (see

Correspondence to: Dr. M. Frisch, Institute of Agronomy and Plant Breeding II, Justus Liebig University, Heinrich-Buff-Ring 26-32, 35392 Giessen, Germany. E-mail: matthias.frisch@uni-giessen.de

also Fig. 1), and crossed to the next *V. vinifera* parent. This procedure was repeated for $t = 6$ pseudo-backcross generations with constant population sizes $n_{F1} = n_t = 50, 100, 150, 200, 250, 300$.

For all selection strategies, carriers of the donor allele at the target locus were pre-selected in the first selection step. These individuals were then subjected to one of three genome-wide background selection strategies. For Strategy 1, a selection index $i = \sum_m x_m$ was created, where summation is over background markers and x_m is the number of *V. vinifera* alleles at the *m*th marker. One individual with the highest value of *i* was selected in the second selection step and crossed to the next *V. vinifera* parent of generation pBC_{t+1}. The individual with the highest proportion of *V. vinifera* alleles at background marker loci was thus selected as parent for the next generation pBC_{t+1}.

For Strategy 2, a selection index $j = \sum_{c\,xc}$ was created, where summation is over chromosomes and $x_c = 1$ if a chromosome carries *V. vinifera* alleles at all background marker loci. All individuals with the highest value of *j* were selected in the second selection step. For these individuals, the value of *i* was determined as described for Strategy 1 in the third selection step. One individual with the highest value of *i* was selected and crossed to the next *V. vinifera* parent of generation pBC_{t+1}. The best individual with the highest number of chromosomes carrying *V. vinifera* alleles at all background marker loci was thus selected as parent for the next generation pBC_{t+1}.

For Strategy 3, a selection index $k = \sum_{c\,xc}$ was created, where summation is over chromosomes and *xc* = length of chromosome c in M if a chromosome carries *V. vinifera* alleles at all background marker loci. All individuals with the highest value of *k* were selected in the second selection step. For these individuals, the value of *i* was determined as described for Strategy 1 in the third selection step. One individual with the highest value of *i* was selected and crossed to the next *V. vinifera* parent of generation pBC_{t+1}. The best individual with the highest cumulative length of chromosomes carrying *V. vinifera* alleles at all background marker loci was thus selected as parent for the next generation pBC_{t+1}.

To quantify the success of the respective pseudo-backcross programs, the 10th percentile (Q_{10}), the arithmetic mean ($\bar{x}$) and the standard deviation (s_x) of the frequency distribution of *V. vinifera* genome in percentage, the average number of chromosomes carrying *V. vinifera* alleles at all background marker loci, and the average number and length of donor fragments were determined in every backcross generation for the selected individuals. In addition, the required number of chromosome-wise multiplexes for the respective pseudo-backcross programs, and the number of individuals n_i selected for evaluation with selection index *i* were determined in every backcross generation.

Results and Discussion

Up to generation pBC_3, Strategy 2 required a higher resource input than Strategy 1 for recovering equivalent levels of *V. vinifera* genome (Fig. 2). Strategy 3 was always inferior to Strategy 2. A $Q_{10} \geq 98$ % required $n_t = 100$ individuals per generation, 1753 multiplexes and three pseudo-backcross generations with Strategy 1 (Fig. 1). With Strategy 2, a $Q_{10} \geq 98\%$ required $n_t = 200$ individuals and 3162 multiplexes. With Strategy 3, a $Q_{10} \geq 98$ % required $n_t = 250$ individuals and 3947 multiplexes, or an additional pseudo-backcross generation.

In generations pBC_1-pBC_3, pre-selection for chromosomes carrying *V. vinifera* alleles at all background marker loci considerably reduced the number of individuals n_i from which the parent for the next generation was selected (Tab. 1). With Strategy 2 only $n_i = 1.8, 2.4, 10.5$ individuals were evaluated for selection index *i*. For Strategy 1, n_i were all individuals carrying the target gene (Tab. 1). These individuals have on average more, but shorter donor fragments than those selected with Strategy 2, and the probability that an individual with a higher overall proportion of *V. vinifera* genome is selected is higher than for Strategy 2.

Strategy 1	Strategy 2	Strategy 3
Selection of best individual with		
highest proportion of background markers for *V. vinifera* parent	highest number of chromosomes carrying *V. vinifera* alleles at all background markers	highest cumulative length of chromosomes carrying *V. vinifera* alleles at all background markers
Number of plants and multiplexes for selection up to pBC_3 at $Q_{10} \geq 98$ %		
n_t = 100 plants 1753 multiplexes	n_t = 200 plants 3162 multiplexes	n_t = 250 plants 3947 multiplexes

Fig. 1: Strategies and effort required to get $Q_{10} \geq 98$ % of *V. vinifera* genome in generation pBC_3. Number of plants per generation and number of chromosome-wise multiplexes are indicated for each strategy.

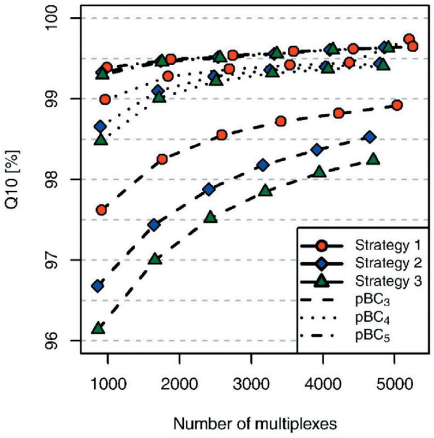

Fig. 2: Tenth percentile Q_{10} values of *V. vinifera* genome recovered in generations pBC_3-pBC_5 for Strategies 1, 2, 3 with constant population sizes n_t = 50, 100, 150, 200, 250, 300 individuals per pseudo-backcross generation plotted against the required number of chromosome-wise multiplexes.

In generations pBC_4-pBC_6, the differences in *V. vinifera* genome between the selection strategies disappeared (Fig. 2 and Tab. 1). For population sizes of $n_t \geq 150$ individuals per pseudo-backcross generation, the differences in Q_{10} values were < 0.1% between Strategy 1 and 2 (Fig. 2). Moreover, Strategy 1 then required more chromosome-wise multiplexes for equivalent levels of *V. vinifera* genome than Strategy 2 (Tab. 2). For n_t = 150-300 individuals per pseudo-backcross generation, 7.24-7.87 % of multiplexes were saved in generations pBC_4-pBC_6 with Strategy 2 compared to Strategy 1 (Tab. 2). In generation pBC_4, all non-carrier chromosomes carried *V. vinifera* alleles at all background marker loci on average, and selection focused on individuals with the shortest donor segment around the target gene for all three strategies (Tab. 1). The number of required multiplexes was then increasing at approximately the same rate for Strategy 1 and Strategy 2. However, Strategy 2 was more efficient than Strategy 1 in selecting for chromosomes which carried *V. vinifera* alleles at all background marker loci in generations pBC_1-pBC_3. While for Strategy 2, 9.3, 15.6, and 18.0 chromosomes on average carried *V. vinifera* alleles at all background marker loci, only 8.1, 14.5, 17.6 chromosomes carried *V. vinifera* alleles at all background marker loci with Strategy 1. These early savings resulted in an overall saving of chromosome-wise multiplexes in advanced pseudo-backcross generations.

Increasing the number of individuals per pseudo-backcross generation from n_t = 150 to n_t = 300 resulted in an additional *V. vinifera* genome recovery of 1.2-0.7 % for Strategy 1 in generations pBC_1-pBC_3 (data for pBC_1 and pBC_2 not shown, for pBC_3 see Fig. 2). In contrast, increasing the number of individuals per pseudo-backcross generation beyond n_t = 150-200 had little effect on *V. vinifera* genome recovery in advanced backcross generations for both Strategy 1 and Strategy 2 (Tab. 2, see also Fig. 2).

Table 1

Recovered level of *V. vinifera* genome (Q_{10}, $\bar{x}$, s_x) in percentage, required number of chromosome-wise multiplexes (CM), average number of chromosomes carrying *V. vinifera* alleles at all background marker loci (CCV), number of individuals evaluated for selection index *i*, number ($\bar{x}$, s_x) and length ($\bar{x}$, s_x) of donor fragments (cM) in generations pBC_1-pBC_6 for Strategies 1, 2, 3 with constant population size n_t = 150 individuals per pseudo-backcross generation

Strategy	pBC_t	*V. vinifera* genome (%)			CM	CCV	n_t	No. donor fragments		Length donor fragments (cM)	
		Q_{10}	$\bar{x}$	s_x				$\bar{x}$	s_x	$\bar{x}$	s_x
1	pBC_1	82.24	84.59	1.92	1425	8.1	75.0	12.9	2.2	33.55	26.22
	pBC_2	94.90	96.24	1.04	2246	14.5	74.9	5.8	1.8	18.14	17.62
	pBC_3	98.55	99.14	0.43	2585	17.6	75.0	2.2	1.1	10.91	9.71
	pBC_4	99.37	99.64	0.20	2692	18.4	75.0	1.4	0.6	7.05	5.08
	pBC_5	99.54	99.75	0.15	2740	18.7	74.9	1.3	0.5	5.51	3.79
	pBC_6	99.64	99.80	0.13	2761	18.9	74.9	1.2	0.4	4.91	3.26
2	pBC_1	79.10	82.75	2.86	1425	9.3	1.8	11.5	1.7	41.96	29.21
	pBC_2	92.89	95.07	1.66	2152	15.6	2.4	4.7	1.5	29.40	26.45
	pBC_3	97.88	98.82	0.69	2407	18.0	10.5	1.8	0.9	18.69	18.39
	pBC_4	99.28	99.60	0.24	2481	18.3	54.9	1.4	0.6	8.25	6.30
	pBC_5	99.52	99.74	0.16	2534	18.7	47.9	1.2	0.5	5.84	4.02
	pBC_6	99.63	99.79	0.13	2558	18.8	61.9	1.2	0.4	5.06	3.34
3	pBC_1	78.29	82.19	3.06	1425	9.2	1.1	11.6	1.7	43.04	28.95
	pBC_2	92.02	94.47	1.87	2163	15.5	1.2	4.8	1.5	32.23	27.28
	pBC_3	97.52	98.63	0.80	2428	18.0	8.9	1.8	0.9	20.95	20.68
	pBC_4	99.22	99.57	0.27	2506	18.3	55.3	1.4	0.6	8.67	6.88
	pBC_5	99.51	99.74	0.16	2560	18.6	48.2	1.2	0.5	5.90	4.03
	pBC_6	99.62	99.79	0.13	2586	18.8	61.4	1.2	0.4	5.10	3.35

Table 2

Recovered level of *V. vinifera* genome (Q_{10}) and percentage of saved chromosome-wise multiplexes (CM (%)) for Strategy 2 compared to Strategy 1 in generations pBC_4-pBC_6 with constant population sizes n_t = 150, 200, 250, 300

Generation	n_t = 150		n_t = 200		n_t = 250		n_t = 300	
	Q_{10}	CM (%)	Q_{10}	CM (%)	Q_{10}	CM (%)	Q_{10}	CM (%)
pBC_4	99.28	7.84	99.36	7.83	99.40	7.63	99.44	7.87
pBC_5	99.52	7.52	99.56	7.50	99.61	7.33	99.64	7.52
pBC_6	99.63	7.35	99.67	7.37	99.68	7.24	99.68	7.47

We conclude that if SSR markers are analyzed as chromosome-wise multiplexes, selecting for individuals with the highest proportion of *V. vinifera* genome at background marker loci is the most efficient selection strategy for short gene introgression programs of up to three pseudo-backcross generations. For such short gene introgression programs, population sizes of $n_t \geq 300$ individuals per pseudo-backcross generation maximize *V. vinifera* genome recovery. For gene introgression programs of four to six pseudo-backcross generations, pre-selecting individuals with the highest number of chromosomes carrying *V. vinifera* alleles at all background marker loci has the potential to considerably reduce the number of required SSR multiplexes. For these longer gene introgression programs, population sizes of n_t = 150-200 individuals are sufficient.

Acknowledgements

We thank N. Hofheinz for proof-reading the manuscript.

References

Blanc, S.; Wiedemann-Merdinoglu, S.; Dumas, V.; Mestre, P.; Merdinoglu, D.; 2012: A reference genetic map of *Muscadinia rotundifolia* and identification of *Ren5*, a new major locus for resistance to grapevine powdery mildew. Theor. Appl. Genet. **125**, 1663-1675.

Di Gaspero, G.; Cattonaro, F.; 2010: Application of genomics to grapevine improvement. Aust. J. Grape Wine Res. **16**, 122-130.

Eibach, R.; Zyprian, E.; Welter, L.; Töpfer, R.; 2007: The use of molecular markers for pyramiding resistance genes in grapevine breeding. Vitis **46**, 120-124.

Merdinoglu, D.; Butterlin, G.; Bevilacqua, L.; Chiquet, V.; Adam-Blondon, A. F.; Decroocq S.; 2005: Development and characterization of a large set of microsatellite markers in grapevine (*Vitis vinifera* L.) suitable for multiplex PCR. Mol. Breed. **15**, 349-366.

Patocchi, A.; Fernández-Fernández, F.; Evans, K.; Gobbin, D.; Rezzonico, F.; Boudichevskaia, A.; Dunemann, F.; Stankiewicz-Kosyl, M.; Mathis-Jeanneteau, F.; Durel, C. E.; Gianfranceschi, L.; Costa, F.; Toller, C.; Cova, V.; Mott, D.; Komjanc, M.; Barbaro, E.; Kodde, L.; Rikkerink, E.; Gessler, C.; van de Weg, W. E.; 2009: Development and test of 21 multiplex PCRs composed of SSRs spanning most of the apple genome. Tree Genet. Genomes **5**, 211-223.

Pauquet, J.; Bouquet, A.; This, P.; Adam-Blondon, A. F.; 2001: Establishment of a local map of AFLP markers around the powdery mildew resistance gene *Run1* in grapevine and assessment of their usefulness for marker assisted selection. Theor. Appl. Genet. **103**, 1201-1210.

Salmaso, M.; Malacarne, G.; Troggio, M.; Faes, G.; Stefanini, M.; Grando, M. S.; Velasco, R.; 2008: A grapevine (*Vitis vinifera* L.) genetic map integrating the position of 139 expressed genes. Theor. .Appl. Genet. **116**, 1129-1143.

Töpfer, R.; Hausmann, L.; Harst, M.; Maul, E.; Zyprian, E.; Eibach, R.; 2011: New horizons for grapevine breeding. In: H. Flachowsky, M. V. Hanke (Eds): Methods in Temperate Fruit Breeding, 79-100. Global Science Books; Fruit, Vegetable and Cereal Science and Biotechnology **5** (Special Issue 1), 79-100.

Vezzulli, S.; Troggio, M.; Coppola, G.; Jermakow, A.; Cartwright, D.; Zharkikh, A.; Stefanini, M.; Grando, M.S.; Viola, R.; Adam-Blondon, A. F.; Thomas, M.; This, P.; Velasco, R.; 2008: A reference integrated map for cultivated grapevine (*Vitis vinifera* L.) from three crosses, based on 283 SSR and 501 SNP-based markers. Theor. Appl. Genet. **117**, 499-511.

Received February 2, 2013

Chapter 5

Selection strategies for the development of maize introgression populations[1]

[1]Herzog, E, Falke, KC, Presterl, T, Scheuermann, D, Ouzunova, M, & Frisch, M. 2014. Selection strategies for the development of maize introgression populations. *PLOS ONE*, **9**(3), e92429.

OPEN ACCESS Freely available online

PLOS ONE

Selection Strategies for the Development of Maize Introgression Populations

Eva Herzog[1], Karen Christin Falke[2], Thomas Presterl[3], Daniela Scheuermann[3], Milena Ouzunova[3], Matthias Frisch[1]*

1 Institute of Agronomy and Plant Breeding II, Justus Liebig University, Giessen, Germany, **2** Institute for Evolution and Biodiversity, University of Münster, Münster, Germany, **3** KWS Saat AG, Einbeck, Germany

Abstract

Introgression libraries are valuable resources for QTL detection and breeding, but their development is costly and time-consuming. Selection strategies for the development of introgression populations with a limited number of individuals and high-throughput (HT) marker assays are required. The objectives of our simulation study were to design and compare selection strategies for the development of maize introgression populations of 100 lines with population sizes of 360–720 individuals per generation for different DH and S_2 crossing schemes. Pre-selection for complete donor chromosomes or donor chromosome halves reduced the number of simultaneous backcross programs. The investigated crossing and selection schemes differed considerably with respect to their suitability to create introgression populations with clearly separated, evenly distributed target donor chromosome segments. DH crossing schemes were superior to S_2 crossing schemes, mainly due to complete homozygosity, which greatly reduced the total number of disjunct genome segments in the introgression populations. The S_2 crossing schemes were more flexible with respect to selection and provided economic alternatives to DH crossing schemes. For the DH crossing schemes, increasing population sizes gradually over backcross generations was advantageous as it reduced the total number of required HT assays compared to constant population sizes. For the S_2 crossing schemes, large population sizes in the final backcross generation facilitated selection for the target segments in the final backcross generation and reduced fixation of large donor chromosome segments. The suggested crossing and selection schemes can help to make the genetic diversity of exotic germplasm available for enhancing the genetic variation of narrow-based breeding populations of crops.

Citation: Herzog E, Falke KC, Presterl T, Scheuermann D, Ouzunova M, et al. (2014) Selection Strategies for the Development of Maize Introgression Populations. PLoS ONE 9(3): e92429. doi:10.1371/journal.pone.0092429

Editor: Lewis Lukens, University of Guelph, Canada

Received October 18, 2013; **Accepted** February 21, 2014; **Published** March 19, 2014

Funding: Funding from the German Federal Ministry of Education and Research (BMBF Grant 0315951) is gratefully acknowledged. [http://www.bmbf.de/en/index.php] The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have the following interests: Thomas Presterl, Daniela Scheuermann and Milena Ouzunova are employed by KWS Saat AG. There are no patents, products in development or marketed products to declare related to KWS SAAT AG and the subject matter of the publication. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials, as detailed online in the guide for authors.

* E-mail: matthias.frisch@uni-giessen.de

Introduction

Introgression libraries are valuable resources for the identification of alleles of agricultural interest in exotic germplasm. They facilitate the introduction of new genetic variation into elite breeding germplasm by providing favorable chromosome segments from wild or exotic species in an adapted genetic background [1,2]. Ideally, an introgression library consists of a set of homozygous introgression lines (IL*s*) which carry short marker-defined chromosome segments from an exotic donor in a common genetic background. The concept was first described in tomato [3]. In the mean time, introgression libraries have been developed for the model species *Arabidopsis thaliana* [4,5], and in many agriculturally important crops, such as rice [6,7], barley [8,9], wheat [10,11], maize [12,13] and rye [14].

Introgression libraries are usually developed by marker-assisted backcrossing followed by selfing or production of double haploid (DH) lines. The backcross process for their development is costly and labor-intensive if complete coverage of the donor genome by short evenly distributed target chromosome segments is to be achieved. Often additional backcross programs have to be run for the developed IL*s* in order to close gaps in donor genome coverage, or to shorten donor chromosome segments by additional recombination events [3,9]. In spite of the high resource requirements, only incomplete donor genome coverage has been achieved for most of the reported introgression libraries [9,14].

In previous simulation studies on introgression libraries, two generations of selfing were investigated for line development [15,16]. Recent genetic studies in maize were based on IL*s* that underwent two to five generations of selfing [17–19]. The use of DH technology has to our knowledge not yet been investigated in simulation studies on the development of introgression libraries. However, *in vivo* induction of maternal haploids is currently a routine method of DH production in commercial maize breeding programs. The main advantage of the DH technology is that complete homozygosity can be obtained after only two generations. Inspite of this time-saving, the production of DH lines is still considerably more costly than conventional selfing [20]. Moreover, a current drawback of *in vivo* induction of maternal haploids in maize is that on average only one viable DH line can be derived from one backcross individual. It is therefore of economic interest to compare this method with S_2 crossing schemes which require

the same number of generations to evaluate the benefits of DH lines.

A possible approach to tackle the high costs required for the development of ideal introgression libraries would be to resort to introgression populations which are not perfect in appearance, but carry some additional donor segments outside the actual target segments. Such introgression populations could be developed with fewer individuals and marker assays. Complete coverage of the donor genome is desirable in order to capture the whole wealth of alleles of agricultural interest in the exotic donor. It is therefore one component of a minimum standard which introgression populations should meet. A second component are short, evenly distributed target donor chromosome segments in a clean adapted background, as they facilitate the use of the ILs in the following breeding process.

The design of the crossing scheme and the selection strategy are the most important factors that influence the distribution of donor chromosome segments in the introgression population. Falke et al. [16] suggested for the development of ideal introgression libraries that a chromosome-based selection strategy which pre-selects individuals carrying the donor alleles on complete chromosomes in generation BC_1 saves resources. Adapting and advancing this concept to crossing schemes with small population sizes might be an efficient approach to develop introgression populations with a limited number of marker assays.

The objectives of our simulation study were (1) to design selection strategies and crossing schemes for the development of maize introgression populations with limited resources, (2) to compare these selection strategies with respect to the distribution and length of donor chromosome segments and the required investments in terms of time, individuals and marker assays, (3) to give guidelines for the optimal experimental design for constructing introgression populations.

Materials and Methods

Software

All simulations were conducted in R version 3.0.0 [21] with the software package SelectionTools, which is available from http://www.uni-giessen.de/population-genetics/downloads.

Genetic Model

A genetic model of maize with 10 equally sized chromosomes of 200 cM length was used for the simulations. Genetic markers for selection were equally spaced. The distance between two adjacent marker loci was 1 cM. All markers were polymorphic between donor and recipient. It was assumed that markers were analyzed with high-throughput (HT) assays. One HT assay comprised genotyping one individual at all marker loci in the linkage map. Recombination was modelled assuming no interference in crossover formation [22]. Each simulation of an introgression population of 100 ILs was replicated 1,000 times in order to reduce sampling effects and to obtain results with high numerical accuracy and a small standard error.

Crossing Schemes

Four crossing schemes were investigated: BC_2DH, BC_3DH, BC_2S_2, BC_3S_2. Each crossing scheme started with the cross of a homozygous donor and a homozygous recipient to create one F_1 individual. The F_1 individual was backcrossed to the recipient to create a BC_1 population of size n_{BC1}. From the BC_1 population, the best individuals with the highest values of selection indices for the respective selection strategy were selected. Each of the selected BC_1 individuals was backcrossed to the recipient to create BC_2 sub-populations of size n_{BC2}. From these BC_2 sub-populations, the best individuals with the highest values of the respective selection indices were selected. For the DH crossing schemes, *in vivo* induction of maternal haploids was assumed with a success rate of one viable DH line per backcross individual. For the BC_2DH schemes, one DH line was thus created from each of the selected BC_2 individuals. For the BC_2S_2 crossing schemes, the selected BC_2 individuals were selfed to create a fixed number of S_1 individuals. Each of the S_1 individuals was selfed again and one S_2 individual was created. For the BC_3 crossing schemes, each of the selected BC_2 individuals was backcrossed to the recipient to create BC_3 sub-populations of size n_{BC3}. From these BC_3 sub-populations, the best individuals with the highest values of the respective selection indices were selected. The generations S_1, S_2 or DH of the BC_3 crossing schemes were carried out as described for the BC_2 crossing schemes.

Evaluation of Selection Candidates

The final introgression populations should consist of 100 ILs which guarantee an acceptable resolution of QTL detection in maize, and which can be immediately used in further breeding steps. Each IL should ideally carry a 20 cM chromosome segment from the donor to provide a complete and even coverage of the donor genome without overlap. The 20 cM chromosome segments are hereafter simply referred to as "target segments". To determine the selection index for an individual with respect to a given target segment, we denote with t_c the donor genome proportion of the chromosome on which the target segment is located, with t_h the donor genome proportion of the chromosome half on which the target segment is located and with t_s the donor genome proportion of the target segment itself. The values for the genetic background b_c, b_h, b_s correspond to t_c, t_h, t_s and denote the recipient genome proportion outside the respective chromosome region. Depending on the selection strategy, t and b are used to define selection indices.

Selection Strategies

We considered generations $g=\{BC_1,BC_2,BC_3,DH,S_1,S_2\}$ for selection. Generation DH was the generation in which homozygous diploid DH lines were available for selection. In each generation g, the genome was divided into selection regions that could either be 10 complete chromosomes, 20 chromosome halves or 100 target segments. For selection for complete donor chromosomes, a fixed number n_{sel} of best individuals for each of the chromosomes $c=1,2,...,10$ with the highest values for selection index $i=t_c+b_c$ were selected. For selection for donor chromosome halves, a fixed number n_{sel} of best individuals for each of the chromosome halves $h=1,2,...,20$ with the highest values for selection index $i=t_h+b_h$ were selected. For selection for donor target segments, a fixed number n_{sel} of best individuals for each of the target segments $s=1,2,...,100$ with the highest values for selection index $i=t_s+b_s$ were selected.

Selection for complete donor chromosomes, donor chromosome halves and donor target segments were combined to form different selection strategies. Selection for complete donor chromosomes in a backcross generation is denoted by a C in the strategy name, selection for donor chromosome halves is denoted by an H, and selection for donor target segments is denoted by an S. For example, for strategy CH, selection for complete donor chromosomes was conducted in generation BC_1 while selection for donor chromosome halves was conducted in generation BC_2. An overview of the investigated selection strategies is presented in Table 1. The investigated combinations of crossing scheme and selection strategy are listed in the first column of Table 2. For all

Table 1. Definition of the selection index i in generations BC_1, BC_2, BC_3, DH, S_1, S_2 for different selection strategies for developing introgression populations.

	Generation				
Strategy	BC_1	BC_2	BC_3	S_1	**DH/S_2**
C	t_c+b_c	–	–	–	t_s+b_s
H	t_h+b_h	–	–	–	t_s+b_s
CC	t_c+b_c	t_c+b_c	–	–	t_s+b_s
HH	t_h+b_h	t_h+b_h	–	–	t_s+b_s
CH	t_c+b_c	t_h+b_h	–	–	t_s+b_s
CCC	t_c+b_c	t_c+b_c	t_c+b_c	–	t_s+b_s
HHH	t_h+b_h	t_h+b_h	t_h+b_h	–	t_s+b_s
CHH	t_c+b_c	t_h+b_h	t_h+b_h	–	t_s+b_s
HHS	t_h+b_h	t_h+b_h	t_s+b_s	–	t_s+b_s

Selection for complete donor chromosomes (C), selection for donor chromosome halves (H) and selection for donor target segments (S) were combined to form different selection strategies (left column). t_c, t_h and t_s denote the donor genome proportions of the chromosome on which the target segment is located, of the chromosome half on which the target segment is located and of the target segment itself. b_c, b_h and b_s correspond to t_c, t_h, t_s and denote the recipient genome proportion outside the respective chromosome region.

doi:10.1371/journal.pone.0092429.t001

selection strategies, the best 100 ILs for selection index $i=t_s+b_s$ were selected in generation DH or S_2, depending on the crossing scheme.

Population Sizes and Simulation Series

We investigated population sizes of $n_{tot}=360-720$ individuals per backcross generation. This should be within a range which can be realized in practical maize breeding programs. Variations in population size were investigated to determine both the effect on preserving the target segments up to line development as well as on recovering the genotype of the recipient outside the target segments.

In the first series of simulations, basic crossing schemes were investigated. Selection was carried out in generation BC_1 for basic crossing schemes with two backcross generations, and in generations BC_1 and BC_2 for basic crossing schemes with three backcross generations. The total population size per generation was kept constant at $n_{tot}=360$ individuals in every generation g.

In the second series of simulation, crossing schemes with high selection intensity were investigated. Population size was doubled compared to the basic crossing schemes ($n_{tot}=720$) in every generation g, while the number of selected individuals was the same as for the basic crossing schemes. The crossing schemes with high selection intensity are denoted by BC_3-CC', BC_3-HH' and BC_3-CH' (Table 2). In the first and second series of simulations, all backcross individuals generated in the final backcross generation were used for line development for both DH and S_2 crossing schemes. One IL was derived from one backcross individual.

In the third series of simulations, crossing schemes with selection in the final backcross generation were investigated. n_{tot} was doubled to 720 individuals in the final backcross generation for the DH crossing schemes BC_2DH-CC, BC_2DH-HH, BC_2DH-CH, $BC_3DH-CCC$, $BC_3DH-HHH$, $BC_3DH-CHH$. This increase in population size was necessary to enable selection and to keep n_{tot} at 360 individuals in generation DH. For the corresponding S_2 schemes, n_{tot} was kept at 360 individuals also in the final backcross generation.

In the fourth series of simulations, crossing schemes with increasing population sizes were investigated. Selection was conducted in the final backcross generation. The crossing schemes with increasing population sizes are denoted by BC_3-HHH^* and BC_3-HHS^*. The details concerning the total population size n_{tot} and population sizes in the sub-populations n_g for all investigated combinations of crossing scheme and selection strategy are summarized in Table 2. Schematic representations of the crossing schemes $BC_3DH-HHH^*$ and $BC_3S_2-HHS^*$ are given in Figure 1 and Figure 2 for illustration.

Measures

To evaluate and compare introgression populations originating from different crossing and selection schemes, the following measures were determined: (a) the genome coverage of the donor O in percent, which is defined as the proportion of the donor genome which is covered by the introgression population, irrespective of whether by the target segments or other donor segments in the genetic background, (b) the depth of donor genome coverage T, which is defined as the average number of ILs in which each donor allele appears in the introgression population, (c) the number of disjunct genome segments in the introgression population S, (d) the resolution of the introgression population R in cM, which is defined as the total genome length of the genetic model in cM divided by S, (e) the average number of donor segments per IL N, (f) the average length of donor segments per IL L in cM, (g) the average total donor genome proportion of the introgression population D_t in percent, (h) the average donor genome proportion of the chromosomes carrying the respective target segments D_c in percent, (i) the average donor genome proportion of the target segments D_s in percent.

Results

High values for the donor genome coverage O around 99% were observed for all crossing schemes (Table 3). However, the resulting introgression populations differed substantially in the values for the number of disjunct genome segments S, the total donor genome proportion D_t, the donor genome proportion of the carrier chromosomes D_c and the donor genome proportion of the target segments D_s. BC_3 crossing schemes resulted in 2–3% lower values for D_t than BC_2 crossing schemes, even if the number of

Table 2. Subdivision of the total population sizes n_{tot} into sub-population sizes n_g in generations $g = BC_1, BC_2, BC_3, S_1, DH, S_2$ for different crossing and selection schemes for developing introgression populations.

	Generation				
Scheme	BC_1	BC_2	BC_3	S_1	**DH**/S_2
Basic crossing schemes					
BC_2DH-C	$1\times1\times360$	$1\times10\times36$	–	–	$10\times36\times1$
BC_2DH-H	$1\times1\times360$	$1\times20\times18$	–	–	$20\times18\times1$
BC_3DH-CC	$1\times1\times360$	$1\times10\times36$	$10\times1\times36$	–	$10\times36\times1$
BC_3DH-HH	$1\times1\times360$	$1\times20\times18$	$20\times1\times18$	–	$20\times18\times1$
BC_3DH-CH	$1\times1\times360$	$1\times10\times36$	$10\times2\times18$	–	$20\times18\times1$
Crossing schemes with high selection intensity					
BC_3DH-CC'	$1\times1\times720$	$1\times10\times72$	$10\times1\times72$	–	$10\times72\times1$
BC_3DH-HH'	$1\times1\times720$	$1\times20\times36$	$20\times1\times36$	–	$20\times36\times1$
BC_3DH-CH'	$1\times1\times720$	$1\times10\times72$	$10\times2\times36$	–	$20\times36\times1$
Crossing schemes with selection in the final BC generation					
BC_2DH-CC	$1\times1\times360$	$1\times10\times72$	–	–	$10\times36\times1$
BC_2DH-HH	$1\times1\times360$	$1\times20\times36$	–	–	$20\times18\times1$
BC_2DH-CH	$1\times1\times360$	$1\times10\times72$	–	–	$10\times(2\times18)\times1$
$BC_3DH-CCC$	$1\times1\times360$	$1\times10\times36$	$10\times1\times72$	–	$10\times36\times1$
$BC_3DH-HHH$	$1\times1\times360$	$1\times20\times18$	$20\times1\times36$	–	$20\times18\times1$
$BC_3DH-CHH$	$1\times1\times360$	$1\times10\times36$	$10\times2\times36$	–	$20\times18\times1$
Crossing schemes with increasing population sizes					
$BC_3DH-HHH^*$	$1\times1\times180$	$1\times20\times18$	$20\times1\times27$	–	$20\times18\times1$
$BC_3DH-HHS^*$	$1\times1\times180$	$1\times20\times18$	$20\times1\times30$	–	$20\times(5\times3)\times1$
Basic crossing schemes					
BC_2S_2-C	$1\times1\times360$	$1\times10\times36$	–	$10\times36\times1$	$10\times36\times1$
BC_2S_2-H	$1\times1\times360$	$1\times20\times18$	–	$20\times18\times1$	$20\times18\times1$
BC_3S_2-CC	$1\times1\times360$	$1\times10\times36$	$10\times1\times36$	$10\times36\times1$	$10\times36\times1$
BC_3S_2-HH	$1\times1\times360$	$1\times20\times18$	$20\times1\times18$	$20\times18\times1$	$20\times18\times1$
BC_3S_2-CH	$1\times1\times360$	$1\times10\times36$	$10\times2\times18$	$20\times18\times1$	$20\times18\times1$
Crossing schemes with high selection intensity					
BC_3S_2-CC'	$1\times1\times720$	$1\times10\times72$	$10\times1\times72$	$10\times72\times1$	$10\times72\times1$
BC_3S_2-HH'	$1\times1\times720$	$1\times20\times36$	$20\times1\times36$	$20\times36\times1$	$20\times36\times1$
BC_3S_2-CH'	$1\times1\times720$	$1\times10\times72$	$10\times2\times36$	$20\times36\times1$	$20\times36\times1$
Crossing schemes with selection in the final BC generation					
BC_2S_2-CC	$1\times1\times360$	$1\times10\times36$	–	$10\times1\times36$	$10\times36\times1$
BC_2S_2-HH	$1\times1\times360$	$1\times20\times18$	–	$20\times1\times18$	$20\times18\times1$
BC_2S_2-CH	$1\times1\times360$	$1\times10\times36$	–	$10\times2\times18$	$20\times18\times1$
BC_3S_2-CCC	$1\times1\times360$	$1\times10\times36$	$10\times1\times36$	$10\times1\times36$	$10\times36\times1$
BC_3S_2-HHH	$1\times1\times360$	$1\times20\times18$	$20\times1\times18$	$20\times1\times18$	$20\times18\times1$
BC_3S_2-CHH	$1\times1\times360$	$1\times10\times36$	$10\times2\times18$	$20\times1\times18$	$20\times18\times1$
Crossing schemes with increasing population sizes					
$BC_3S_2-HHH^*$	$1\times1\times180$	$1\times20\times18$	$20\times1\times27$	$20\times1\times18$	$20\times18\times1$
$BC_3S_2-HHS^*$	$1\times1\times180$	$1\times20\times18$	$20\times1\times23$	$20\times5\times4$	$100\times1\times4$

The total population size in generation g is defined as $n_{tot} = n_{pop} \times n_{sel} \times n_g$. n_{pop}: number of sub-populations in generation $g-1$; n_{sel}: number of individuals selected from the sub-populations in generation $g-1$; n_g: population size per sub-population in generation g.

doi:10.1371/journal.pone.0092429.t002

generations of selection was the same. For example, the basic crossing scheme BC_3DH-CC resulted in a D_t of only 5.0%, while crossing scheme BC_2DH-CC with selection in the final backcross generation resulted in a D_t of 7.8%. An additional generation of selection in BC_2 schemes only resulted in minor improvements of D_t of 0.4–1.4% compared to the basic crossing schemes without selection. For example, scheme BC_2DH-CC improved D_t only by 0.5% compared to scheme BC_2DH-C.

The DH crossing schemes had in most cases better values for T, D_t, D_c, D_s and especially S than the S_2 crossing schemes (Table 3).

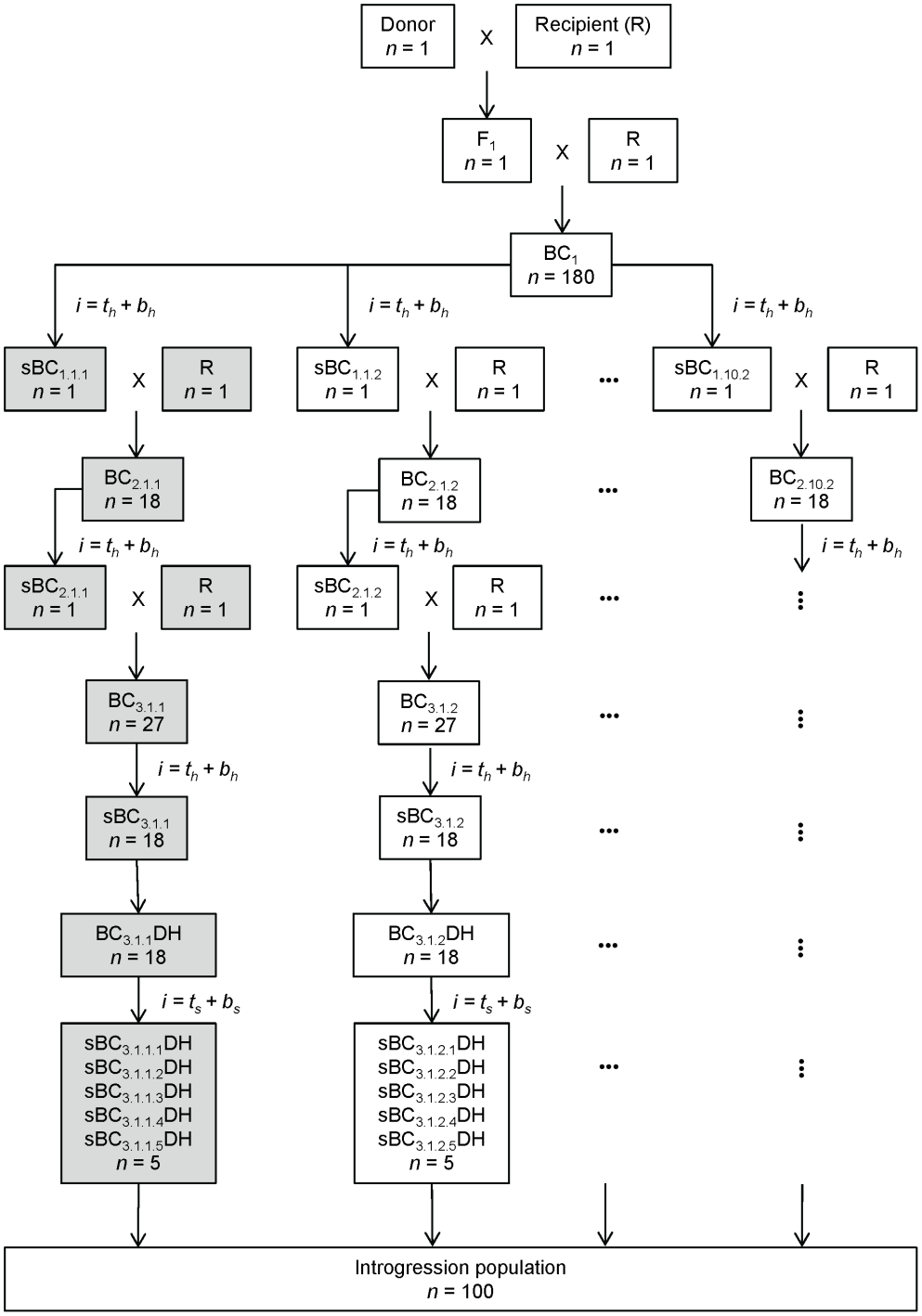
Donor
n = 1
X
Recipient (R)
n = 1
F1
n = 1
X
R
n = 1
BC1
n = 180
i = th + bh
i = th + bh
i = th + bh
sBC1.1.1
n = 1
X
R
n = 1
sBC1.1.2
n = 1
X
R
n = 1
sBC1.10.2
n = 1
X
R
n = 1
BC2.1.1
n = 18
BC2.1.2
n = 18
BC2.10.2
n = 18
i = th + bh
i = th + bh
i = th + bh
sBC2.1.1
n = 1
X
R
n = 1
sBC2.1.2
n = 1
X
R
n = 1
BC3.1.1
n = 27
BC3.1.2
n = 27
i = th + bh
i = th + bh
sBC3.1.1
n = 18
sBC3.1.2
n = 18
BC3.1.1DH
n = 18
BC3.1.2DH
n = 18
i = ts + bs
i = ts + bs
sBC3.1.1.1DH
sBC3.1.1.2DH
sBC3.1.1.3DH
sBC3.1.1.4DH
sBC3.1.1.5DH
n = 5
sBC3.1.2.1DH
sBC3.1.2.2DH
sBC3.1.2.3DH
sBC3.1.2.4DH
sBC3.1.2.5DH
n = 5
Introgression population
n = 100

Figure 1. Schematic representation of crossing scheme $BC_3DH-HHH^*$. Crossing scheme $BC_3DH-HHH^*$ is characterized by increasing population sizes in the backcross generations and selection for donor chromosome halves in the final backcross generation. The parts highlighted in gray represent one branch of the crossing scheme. Sub-populations are indexed by BC_g, $BC_{g.c.h}$ and $BC_{g.c.h.s}$, where g is the respective backcross generation, c is the respective chromosome, h is the respective chromosome half, s is the respective target segment; $sBC_{g.c.h}$ and $sBC_{g.c.h.s}$ denote individuals selected for the respective selection regions.
doi:10.1371/journal.pone.0092429.g001

Very high values of $S>1000$ segments were observed for the basic crossing schemes BC_2S_2-C and BC_2S_2-H. These crossing schemes had on average $N=1$ additional donor segment per IL compared to the corresponding DH crossing schemes. However, they were also characterized by incomplete homozygosity (Figure 3B). The S_2 crossing schemes with selection in the final backcross generation required 360 individuals and HT assays less then the corresponding DH crossing schemes (Tables 2 and 3). Nevertheless, the differences between DH and S_2 crossing schemes then diminished. For example, scheme BC_3S_2-HHH resulted in similar values for most measures as the corresponding scheme $BC_3DH-HHH$ (Table 3).

The differences in the total donor genome proportion D_t between selection for complete donor chromosomes and selection for donor chromosome halves ranged only between 0.1–0.7% for the same number of backcross generations and generations of selection. However, substantial differences were observed for the donor genome proportion of the carrier chromosomes D_c and the donor genome proportion of the target segments D_s. For selection for complete donor chromosomes, high values for D_c of up to 48% were observed. They were clearly visible in the graphical genotypes for schemes BC_2S_2-CC and BC_3DH-CC (Figure 3B and C). For selection for donor chromosome halves, the values for D_c were much lower and did not exceed 42% (Table 3). Without selection in the final backcross generation, selection for donor chromosome halves resulted in substantially reduced values for D_s. For example, the basic crossing schemes BC_3DH-HH and BC_3S_2-HH resulted in values for D_s of only 94% and 90%. Moreover, the ranges for D_s for these crossing schemes were substantially greater (Figure 4 for S_2 crossing schemes, for DH data not shown).

The basic crossing schemes BC_3DH-CH and BC_3S_2-CH which combined selection for complete donor chromosomes and selection for donor chromosome halves resulted in similarly low values for D_s of 93.7% and 89.9% as selection for donor chromosome halves only (Table 3). In addition, the combined strategies CH and CHH resulted in high values for D_c of up to 45.9%. The low values for D_s and the high values for D_c were reflected in the graphical genotype of scheme BC_3DH-CH, e.g. in ILs 47 and 54 (Figure 3A).

Doubling population sizes n_{tot} from 360 to 720 individuals in the crossing schemes with high selection intensity reduced the total donor genome proportion D_t from 5.0–5.1% to 3.6–3.8% compared to the basic DH crossing schemes, and from 5.3–5.7% to 4.3–4.4% compared to the basic S_2 crossing schemes (Table 3). The donor genome proportion of the carrier chromosomes D_c was reduced by about 4.2–7.5% for the DH crossing schemes, and by about 0.9–6.9% for the S_2 crossing schemes. The reduction of the donor genome proportion of the target segments D_s in combination with increased ranges that was observed with selection for donor chromosome halves in the basic crossing schemes was not observed in the crossing schemes with high selection intensity (Table 3 and Figure 4). D_s was increased by 5.2% for crossing scheme BC_3DH-HH' and by 8.6% for crossing scheme BC_3S_2-HH' compared to the basic crossing schemes BC_3DH-HH and BC_3S_2-HH. However, these improvements were only achieved with 2160 HT assays compared to 1080 HT asssays in the basic crossing schemes (Table 3).

The crossing schemes with selection in the final backcross generation resulted in values for D_t that were 1.1–1.2% higher for the DH crossing schemes und 0.6–1.4% higher for the S_2 crossing schemes compared to the crossing schemes with high selection intensity. The ranges of D_s for selection for donor chromosome halves were about the same size as for the crossing schemes with high selection intensity (Figure 4). The average values for D_s were 0.5% lower for scheme $BC_3DH-HHH$ and 0.3% lower for scheme BC_3S_2-HHH (Table 3). The number of required HT assays was reduced by 360 for the DH crossing schemes and by 720 for S_2 crossing schemes compared to the crossing schemes with high selection intensity. For the crossing schemes with selection in the final backcross generation, selection for donor chromosome halves was the most advantageous selection strategy with respect to the genetic background and to the target segments. Most notably, the crossing schemes $BC_3DH-HHH$ and BC_3S_2-HHH resulted in the lowest values for the donor genome proportion of the carrier chromosomes D_c. Compared to the most efficient basic crossing schemes BC_2DH-C and BC_3DH-CC, the crossing schemes BC_2DH-HH and $BC_3DH-HHH$ resulted in small improvements of both the genetic background and D_s. However, in both cases 720 additional HT assays had to be invested. For the S_2 crossing schemes with selection in the final backcross generation, high values of D_c of 38.1–48.3% were observed. Large donor chromosome segments on the carrier chromosomes were also visible in the graphical genotypes for schemes BC_2S_2-CC and BC_3S_2-HHH (Figure 3B and D). The high values for D_c were associated with a considerable reduction of the number of disjunct genome segments S of >200 segments for the BC_2 crossing schemes and of 100–200 segments for the BC_3 crossing schemes compared to the basic S_2 crossing schemes (Table 3).

The crossing schemes with increasing population sizes reduced the number of required HT assays for DH crossing schemes by 360 in comparison to the crossing schemes with selection in the final backcross generation and constant population sizes. The crossing schemes $BC_3DH-HHH^*$ and $BC_3DH-HHS^*$ resulted in similar values for most measures as the crossing scheme $BC_3DH-HHH$. However, D_c and D_s were slightly reduced for crossing scheme $BC_3DH-HHS^*$. Compared to the most efficient basic crossing scheme BC_3DH-CC, crossing scheme $BC_3DH-HHH^*$ required 360 additional HT assays, but reduced D_c by 1.9% and increased D_s by 0.6%. The crossing scheme $BC_3S_2-HHH^*$ resulted with 38.0% in a much higher D_c than the crossing scheme $BC_3S_2-HHS^*$ with 30.4%. For crossing scheme $BC_3S_2-HHS^*$, the average D_s was only 96.2% and the range for D_s was higher than for the crossing schemes $BC_3S_2-HHH^*$ and BC_3S_2-HHH (Figure 4). However, D_t and D_c were the lowest for all investigated crossing schemes, with the exception of the crossing schemes with high selection intensity and $n_{tot}=720$ (Table 3). The clear-cut separation of the target segments is also visible in the graphical genotype (Figure 3F).

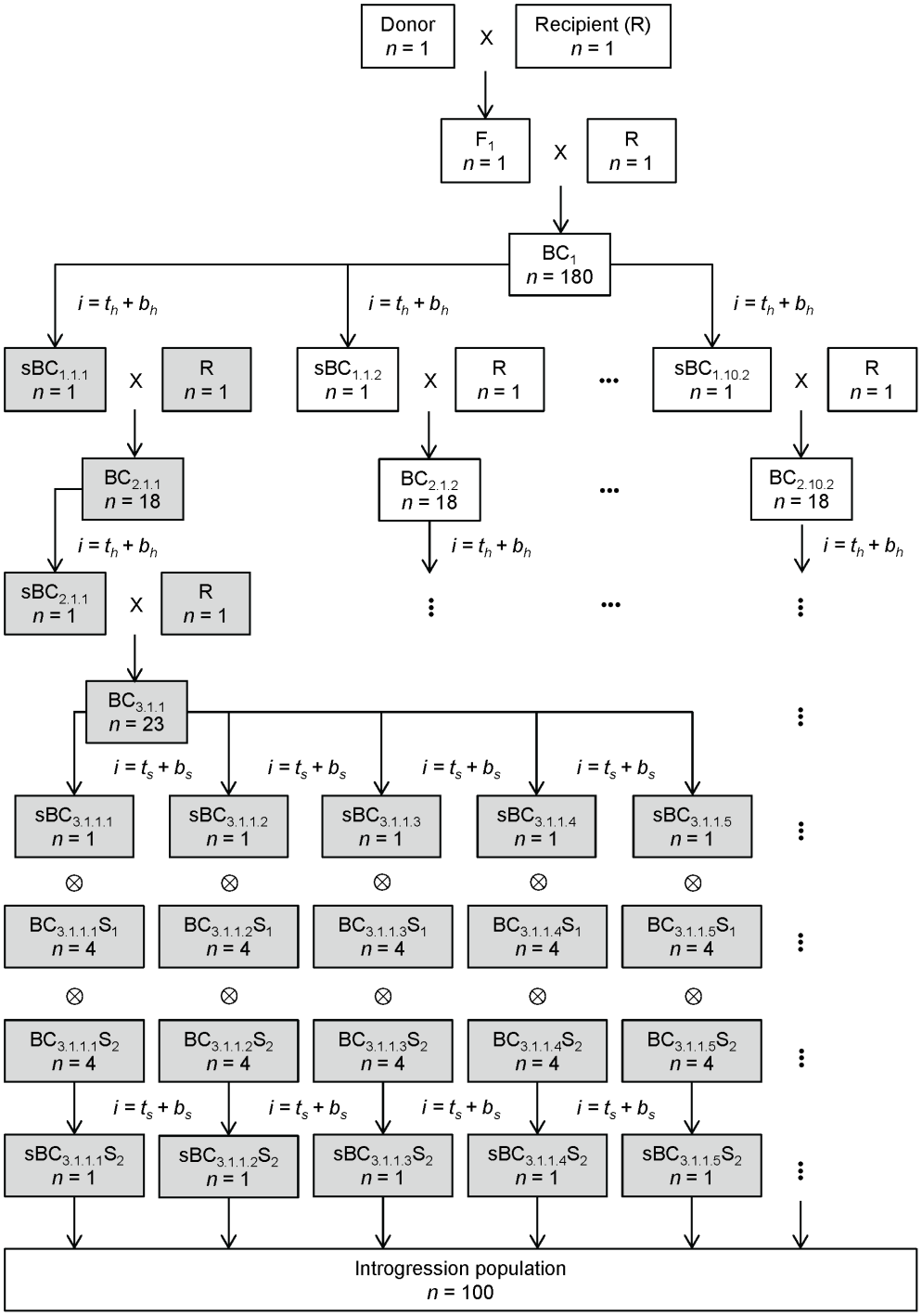
Donor
n = 1
X
Recipient (R)
n = 1
F1
n = 1
X
R
n = 1
BC1
n = 180
i = th + bh
i = th + bh
i = th + bh
sBC1.1.1
n = 1
X
R
n = 1
sBC1.1.2
n = 1
X
R
n = 1
sBC1.10.2
n = 1
X
R
n = 1
BC2.1.1
n = 18
BC2.1.2
n = 18
BC2.10.2
n = 18
i = th + bh
i = th + bh
i = th + bh
sBC2.1.1
n = 1
X
R
n = 1
BC3.1.1
n = 23
i = ts + bs
i = ts + bs
i = ts + bs
i = ts + bs
sBC3.1.1.1
n = 1
sBC3.1.1.2
n = 1
sBC3.1.1.3
n = 1
sBC3.1.1.4
n = 1
sBC3.1.1.5
n = 1
⊗
BC3.1.1.1S1
n = 4
BC3.1.1.2S1
n = 4
BC3.1.1.3S1
n = 4
BC3.1.1.4S1
n = 4
BC3.1.1.5S1
n = 4
⊗
BC3.1.1.1S2
n = 4
BC3.1.1.2S2
n = 4
BC3.1.1.3S2
n = 4
BC3.1.1.4S2
n = 4
BC3.1.1.5S2
n = 4
i = ts + bs
i = ts + bs
i = ts + bs
i = ts + bs
sBC3.1.1.1S2
n = 1
sBC3.1.1.2S2
n = 1
sBC3.1.1.3S2
n = 1
sBC3.1.1.4S2
n = 1
sBC3.1.1.5S2
n = 1
Introgression population
n = 100

Figure 2. Schematic representation of crossing scheme $BC_3S_2-HHS^*$. Crossing scheme $BC_3S_2-HHS^*$ is characterized by increasing population sizes in the backcross generations and selection for target segments in the final backcross generation. The parts highlighted in gray represent one branch of the crossing scheme. Sub-populations are indexed by BC_g, $BC_{g.c.h}$ and $BC_{g.c.h.s}$, where g is the respective backcross generation, c is the respective chromosome, h is the respective chromosome half, s is the respective target segment; $sBC_{g.c.h}$ and $sBC_{g.c.h.s}$ denote individuals selected for the respective selection regions.
doi:10.1371/journal.pone.0092429.g002

Discussion

Measures for Characterizing Introgression Populations

Measures for the description of introgression populations should allow to distinguish between introgression populations of different structure. Complete donor genome coverage O is desirable in order to make the complete genetic variation of the donor available for the breeding process. However, high values for O can also be caused by donor segments outside the target segments which could not be removed from the genetic background. O is therefore only informative if interpreted in relation to measures which reflect the distribution of the donor genome in the introgression population. A distinctive description of introgression populations is possible with the total donor genome proportion D_t, the donor genome proportion on the carrier chromosomes D_c and the donor genome proportion of the actual target segments D_s.

A high total donor genome proportion D_t is often associated with a high number of disjunct genome segments S. S determines the resolution R, which is an important parameter for the accuracy of QTL detection. However, if S is greater than the number of ILs, the problem of overparameterization arises with classical linear model approaches. This issue has only in part been resolved by using statistical methods which pre-select a reduced number of ILs for the linear model [23].

High values for the donor genome proportion on the carrier chromosomes D_c and the depth of donor genome coverage T reflect undesired donor segments attached to the actual target segments. Such large donor segments which overlap between ILs have been reported to increase the risk of false-positive effects in QTL detection and reduce the power of QTL detection [24]. This is mainly a problem if linkage maps with large distances between adjacent markers of 10 cM or more are employed, because QTLs located between the last marker of the target segment and the next marker outside the target segment are incorrectly assigned to the target segments. With dense marker maps which are now available this problem should be overcome. However, large donor segments also increase the risk of linkage drag in the breeding process and often require further steps of separation [24].

Low values for the donor genome proportion of the target segments D_s indicate a loss of target segments and potentially useful alleles. This is a problem that arises with small population sizes as were investigated in the present study [16]. Even if the missing target segments are present in the genetic background of other ILs, this might impair QTL detection and the further use of the ILs for the breeding progress.

We therefore argue that short non-overlapping target segments in a clean recipient background are advantageous also with dense marker maps. For 20 cM target segments and a genomic model of 10 equally sized chromosomes of 200 cM length, this corresponds to $D_t = 1\%$, $D_c = 10\%$ and $D_s = 100\%$ in the ideal case. The effort and time required for developing introgression populations with such characteristics is beyond the scope of most breeding programs. With the limited population sizes and number of HT assays investigated in this study, these ideal values could not be achieved with two or three backcross generations (Table 3). We therefore considered those crossing and selection schemes as efficient which with a given limited resource input resulted in the highest coverage of target segments D_s in combination with low overlap of target segments reflected in D_c and T and a low total donor genome proportion D_t.

With respect to QTL detection, it can be expected that the optimal values for the suggested measures will depend on the statistical method and the genetic architecture of the trait. They could be determined for a given statistical method by including QTLs of different number and effect in future simulation studies. We plan further investigations in this area of research.

Crossing Schemes

BC_3 crossing schemes had 2–3% lower values for the total donor genome proportion D_t than BC_2 crossing schemes (Table 3), even if no selection for the genetic background was conducted in generation BC_3. Selection in generation BC_2, as was investigated with the crossing schemes BC_2-CC, BC_2-HH and BC_2-CH, only resulted in a reduction of D_t of 0.4–1.4% compared to the basic crossing schemes BC_2-C and BC_2-H (Table 3). An explanation for this comparatively small reduction is that the limiting factor for the reduction of D_t is the number of recombinations during meiosis. Hence, even though BC_2 crossing schemes have a time advantage, the effect of a third backcross generation cannot be compensated by investing in additional marker analyses. We therefore conclude that BC_3 crossing schemes result in introgression populations with an improved structure, and that the time investment in the additional backcross generation is worthwhile.

DH crossing schemes were for most measures superior to the corresponding S_2 crossing schemes. The differences were most pronounced in the number of disjunct genome segments S. Even though the S_2 schemes on average had a slightly higher number of donor segments per IL N, it seems that the very high values for S that were observed especially in the BC_2-S_2 crossing schemes mainly had to be attributed to incomplete homozygosity (Figure 3B). It can be expected that introgression populations with $S > 1000$ segments in 100 ILs (Table 3) are not suitable for effective QTL detection. We therefore conclude that the DH method is essential for short crossing schemes with only two backcross generations.

A drawback of the DH method is that with current protocols of *in vivo* DH induction of maternal haploids, only a very limited number of viable DH lines can be derived from one backcross individual. We expect that our assumption of one DH line per backcross individual is a conservative, but realistic estimate. In contrast, with selfing, many progenies can be derived from one selected backcross individual. In the S_2 crossing schemes, it is consequently comparatively cheap and easy to conduct selection in the final backcross generation. For the DH schemes, selection in the final backcross generation could only be conducted if population size in this generation was higher than the desired number of final DH lines. As a result, the S_2 crossing schemes with selection in the final backcross generation required 360 HT assays less than the corresponding DH schemes (Table 3). Moreover, the selected fractions of best backcross individuals were much greater for the DH than for the S_2 crossing schemes (Table 2). This resulted in a lower selection intensity for both the selection region of the final backcross generation and the genetic background in

Table 3. Measures evaluated for introgression populations resulting from different crossing and selection schemes.

Scheme	O	T	S	R	N	L	D_t	D_c	D_s	HT
Basic crossing schemes										
BC_2DH-C	99.9	8.3	691	2.9	6.2	28.7	8.3	38.9	98.7	720
BC_2DH-H	100.0	8.9	751	2.7	6.4	29.2	8.9	37.3	94.1	720
BC_3DH-CC	99.2	5.1	457	4.4	3.8	29.8	5.0	35.2	97.8	1080
BC_3DH-HH	99.8	5.2	487	4.1	3.9	29.8	5.1	33.3	94.2	1080
BC_3DH-CH	99.6	5.2	469	4.3	3.8	30.5	5.1	35.1	93.7	1080
Crossing schemes with high selection intensity										
BC_3DH-CC'	99.3	3.6	389	5.1	3.0	27.8	3.6	27.7	98.8	2160
BC_3DH-HH'	99.9	3.8	406	4.9	3.0	29.7	3.8	29.1	99.4	2160
BC_3DH-CH'	99.6	3.9	399	5.0	3.0	30.0	3.8	29.9	99.0	2160
Crossing schemes with selection in the final BC generation										
BC_2DH-CC	99.9	7.8	676	3.0	5.9	28.7	7.8	41.0	98.9	1440
BC_2DH-HH	100.0	8.1	716	2.8	6.0	29.0	8.1	38.9	99.3	1440
BC_2DH-CH	99.9	8.5	679	2.9	6.0	30.4	8.5	43.7	98.8	1440
$BC_3DH-CCC$	99.1	4.9	457	4.4	3.7	30.2	4.8	35.8	98.0	1800
$BC_3DH-HHH$	99.9	4.8	464	4.3	3.6	31.0	4.7	33.9	98.9	1800
$BC_3DH-CHH$	99.6	4.8	450	4.4	3.5	31.4	4.7	35.0	98.5	1800
Crossing schemes with increasing population sizes										
$BC_3DH-HHH^*$	99.9	5.0	492	4.1	3.8	29.7	5.0	33.3	98.4	1440
$BC_3DH-HHS^*$	99.8	4.9	484	4.1	3.8	29.4	4.8	32.5	97.5	1440
Basic crossing schemes										
BC_2S_2-C	100.0	11.4	1021	2.0	7.3	26.4	9.3	41.4	97.6	720
BC_2S_2-H	100.0	11.4	1073	1.9	7.4	25.9	9.3	36.3	90.4	720
BC_3S_2-CC	99.3	6.9	684	2.9	4.5	27.7	5.7	39.2	97.0	1080
BC_3S_2-HH	99.9	6.3	702	2.8	4.4	26.3	5.3	33.0	90.3	1080
BC_3S_2-CH	99.7	6.4	681	2.9	4.3	26.9	5.3	35.0	89.9	1080
Crossing schemes with high selection intensity										
BC_3S_2-CC'	99.5	5.1	585	3.4	3.6	26.6	4.3	32.3	98.8	2160
BC_3S_2-HH'	99.9	5.1	591	3.4	3.5	27.8	4.4	32.1	98.9	2160
BC_3S_2-CH'	99.7	5.2	581	3.4	3.5	28.2	4.4	33.4	98.4	2160
Crossing schemes with selection in the final BC generation										
BC_2S_2-CC	99.3	10.5	795	2.5	6.5	27.2	8.3	48.3	98.0	1080
BC_2S_2-HH	99.9	9.7	785	2.5	6.2	27.0	7.9	42.8	98.8	1080
BC_2S_2-CH	99.8	9.7	761	2.6	6.1	27.6	7.9	45.9	98.4	1080
BC_3S_2-CCC	98.3	7.1	588	3.4	4.2	30.5	5.7	44.6	97.5	1440
BC_3S_2-HHH	99.7	5.9	510	3.9	3.7	31.0	5.0	38.1	98.6	1440
BC_3S_2-CHH	99.3	6.0	509	3.9	3.7	31.3	5.0	39.5	98.2	1440
Crossing schemes with increasing population sizes										
$BC_3S_2-HHH^*$	99.7	5.8	508	3.9	3.8	30.7	4.9	38.0	98.7	1440
$BC_3S_2-HHS^*$	99.8	5.1	596	3.4	3.7	26.1	4.3	30.4	96.2	1400

O: donor genome coverage in percent; T: depth of donor genome coverage; S: number of disjunct genome segments; R: resolution; N: number of donor segments per IL; L: length of donor segments per IL in cM; D_t: total donor genome proportion in percent; D_c: donor genome proportion of carrier chromosomes in percent; D_s: donor genome proportion of target segments in percent; HT: the required number of HT assays. Measures are arithmetic means over 1,000 replications.

doi:10.1371/journal.pone.0092429.t003

the DH crossing schemes. We therefore suggest that a comparison of DH and S_2 crossing schemes should take the distinctive features of both methods into account. The evaluation of efficiency should also be based on the number of required HT assays. Considering this, S_2 crossing schemes which exploit their selection advantages represent economic and easy-to-handle alternatives to DH crossing schemes.

Selection Strategies for Small and Constant Population Sizes

For a given genetic model and crossing scheme, the selection strategy is the most important factor that influences the structure of the resulting introgression population. In the following paragraphs, different aspects such as the length of the selection

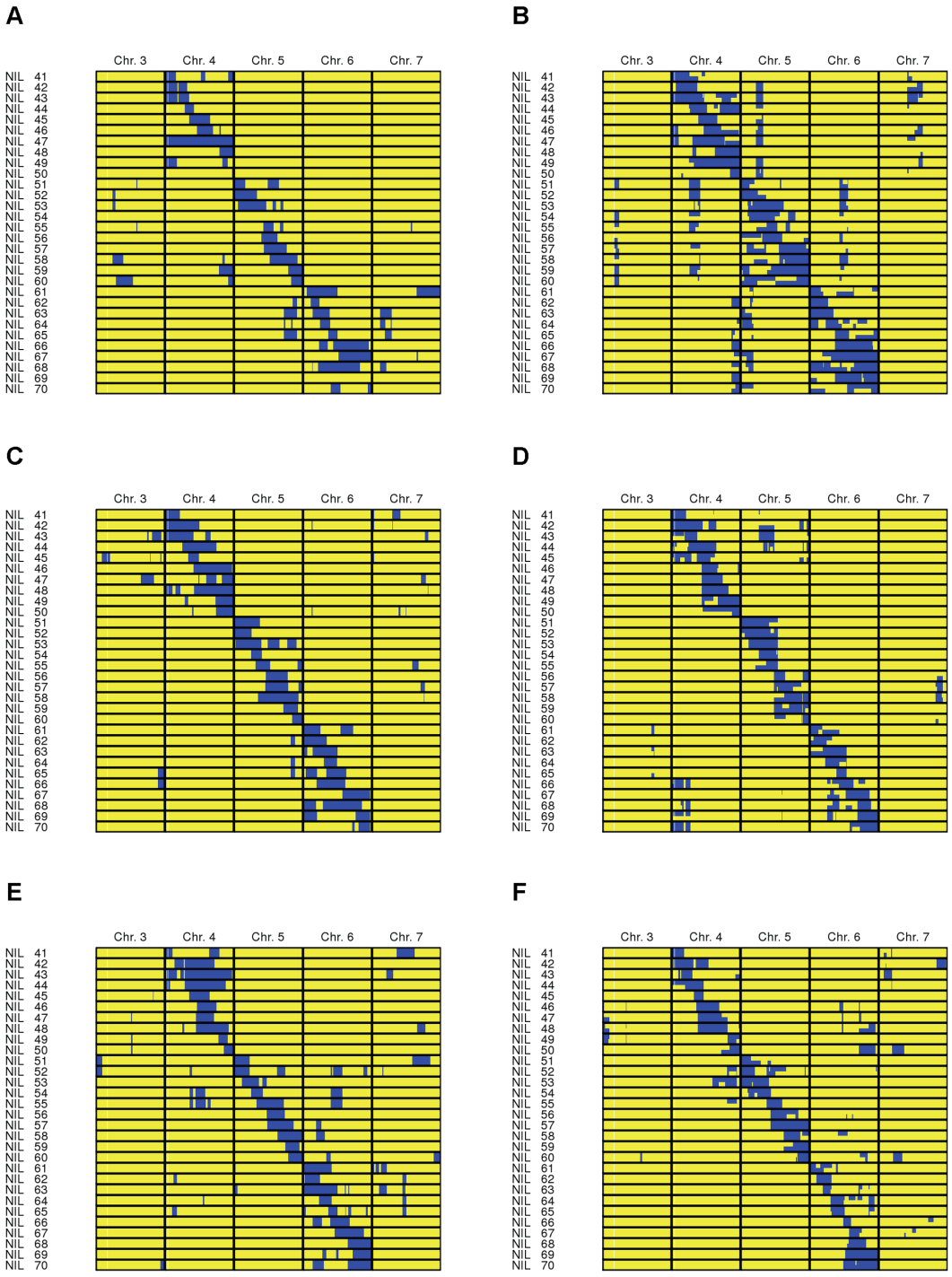
A
B
C
D
E
F
Chr. 3
Chr. 4
Chr. 5
Chr. 6
Chr. 7
NIL 41
NIL 42
NIL 43
NIL 44
NIL 45
NIL 46
NIL 47
NIL 48
NIL 49
NIL 50
NIL 51
NIL 52
NIL 53
NIL 54
NIL 55
NIL 56
NIL 57
NIL 58
NIL 59
NIL 60
NIL 61
NIL 62
NIL 63
NIL 64
NIL 65
NIL 66
NIL 67
NIL 68
NIL 69
NIL 70

Figure 3. Graphical genotypes of introgression populations resulting from six different crossing schemes. A: BC_3DH-CH; B: BC_2S_2-CC; C: BC_3DH-CC; D: BC_3S_2-HHH; E: $BC_3DH-HHH^*$; F: $BC_3S_2-HHS^*$. The graphical genotypes display the chromosomes 3 to 7 of ILs 41–70 and are examples from one simulation run. Chromosome segments which stem from the donor are displayed in blue, whereas chromosome segments which stem from the recipient are displayed in yellow. The graphical genotypes illustrate the differences between the alternative crossing schemes with respect to their suitability to create introgression populations with complete donor genome coverage and clearly separated, evenly distributed target donor chromosome segments.
doi:10.1371/journal.pone.0092429.g003

regions, the number of generations of selection and the required population sizes for effective selection are discussed.

Selection strategies which pre-select individuals carrying complete donor chromosomes reduce the number of simultaneous backcross programs to the number of chromosomes [16]. They are therefore suitable for breeding programs with limited resources. However, for long chromosomes of 200 cM length, selection for complete donor chromosomes preserved large donor chromosome segments on the carrier chromosomes up to line development (Figure 3C and B). This was reflected in high values for the proportion of donor genome on the carrier chromosomes D_c of up to 48% (Table 3). The selection regions for selection in the backcross generations were therefore reduced to donor chromosome halves for selection strategies H, HH and HHH. In all four series of simulations, selection for donor chromosome halves resulted in the desired reduction of D_c compared to selection for complete donor chromosomes (Table 3). Other measures for the genetic background were approximately equivalent. We therefore conclude that for crop species with long chromosomes such as maize, wheat or rapeseed, selection for donor chromosome halves reduces the length of the donor segments attached to the actual target segments and the risk of linkage drag.

However, for crossing schemes without selection in the final backcross generation and constant population sizes of $n_{tot}=360$ individuals, selection for donor chromosome halves resulted in a considerable reduction of the donor genome proportion of the target segments D_s of up to 7%. Moreover, the estimated values for D_s were less reliable for these crossing schemes, *e.g.*, in schemes BC_2S_2-H and BC_3S_2-HH (Figure 4). These findings have to be attributed to the small population sizes n_g in the sub-populations and the structure of the selection index i. In generation DH or S_2, population sizes were reduced to $n_g=18$ individuals with selection for donor chromosome halves (Table 2). Without selection in the final backcross generation, around 50% of the ILs developed from the backcross individuals are expected to carry no donor allele at a given locus within the respective target segment. The probability to find five ILs with complete donor target segments for the introgression population was therefore even further reduced. As the selection index $i=t_s+b_s$ weighed the target segments and the genetic background equally, a clean genetic background sometimes outweighed a reduced D_s and led to the observed loss of target segments in these small sub-populations. We therefore conclude that a sufficiently large population size is the crucial

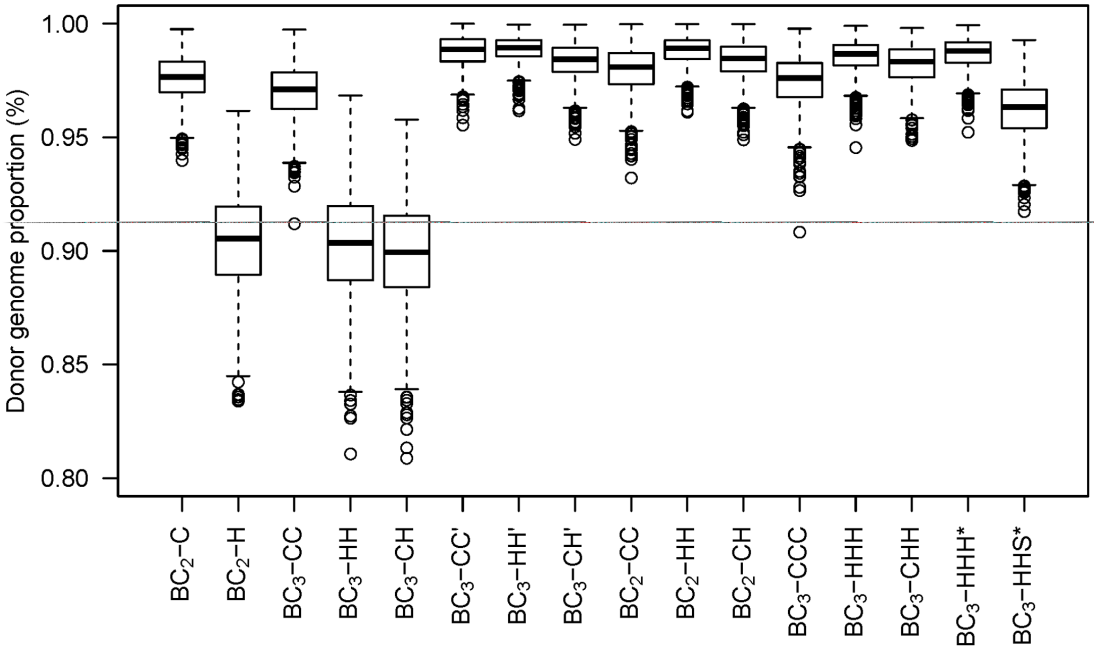

Figure 4. Donor genome proportion of target segments D_s for all investigated S_2 crossing schemes. The boxplots represent the distribution over 1,000 replications of the simulations. The basic crossing schemes BC_2S_2-H, BC_3S_2-HH and BC_3S_2-CH which select for donor chromosome halves are characterized by higher ranges for D_s.
doi:10.1371/journal.pone.0092429.g004

factor for the successful application of selection for donor chromosome halves.

A loss of target segments caused by small population sizes was also observed for the basic combined selection strategy CH which selected for complete donor chromosomes in generation BC_1 and for donor chromosome halves in generation BC_2. In addition, the combined strategies CH and CHH resulted in high values for the donor genome proportion on the carrier chromosomes D_c of up to 45% (Table 3). This can be explained by the efficient selection for complete donor chromosomes from the comparatively large BC_1 population of $n_g = n_{tot} = 360$ individuals (Table 2). The pre-selected complete donor chromosomes are in large part preserved up to line development. The combination of missing target segments with large donor chromosome segments on carrier chromosomes was also reflected in the graphical genotype for scheme BC_3DH-CH, *e.g.*, in IL 47 and 54 (Figure 3A). The selection strategies CH and CHH therefore combine the drawbacks of both selection for complete donor chromosomes and selection for chromosome halves. They are not suitable for crossing schemes with small and constant population sizes, in which the population sizes n_g in the sub-populations are subsequently reduced over the backcross generations. We conclude that for for small breeding programs with a constant population size of $n_{tot} = 360$ and a limited number of HT assays for selection, selection strategies which only select for complete donor chromosomes in the backcross generations should be employed in both DH and S_2 crossing schemes to avoid the loss of target segments.

Finding more Carriers of Donor Target Segments for Line Development

To employ selection for donor chromosome halves effectively for reducing the donor genome proportion of the carrier chromosomes D_c without losing the target segments, it is necessary to increase the frequency of carriers of donor target segments for line development. Using larger population sizes is a straightforward solution for this problem, which in addition can improve the overall structure of introgression populations. The crossing schemes with high selection intensity and double population sizes of $n_{tot} = 720$ individuals resulted in small improvements of the total donor genome proportion D_t of about 1–1.5% compared to the basic crossing schemes (Table 3). The desired increase in the donor proportion of the target segments D_s was achieved. For selection for donor chromosome halves, D_s was increased by 5.2–8.6%. Selection for donor chromosome halves was then even superior to selection for complete donor chromosomes. Moreover, the donor genome proportion on the carrier chromosomes D_c was reduced by up to 7.5%, indicating an improved separation of target segments. The observed improvements were greater for the DH than for the S_2 crossing schemes. Nevertheless, the comparatively small improvements of the introgression populations required 1080 additional HT assays. We therefore conclude that such large population sizes are only suitable for breeding programs with access to DH technology, less stringent resource restrictions and high requirements with respect to the genetic background. If the requirements concerning the structure of the introgression population are not that high, it might be more economic to increase population size only in the final backcross generation and/or to invest in additional HT assays only in this generation.

For crossing schemes with selection in the final backcross generation, the total donor genome proportion D_t was similar to the values of the basic crossing schemes, and about 1% higher than for the crossing schemes with higher selection intensity (Table 3). However, the average values for the donor genome proportion of the target segments D_s were similar to the crossing schemes with higher selection intensity (Table 3) and the ranges were effectively reduced (Figure 4). Moreover, the number of required HT assays was reduced by 360 for the DH crossing schemes and by 720 for S_2 crossing schemes compared to the crossing schemes with higher selection intensity (Table 3). The decision for doubling population sizes requires the same resources as would be required for generating an additional introgression population. This large effort seems not to be justified by the relatively small improvements compared to the basic crossing schemes. We therefore conclude that selection in the final backcross generation is the more efficient solution for both DH and S_2 crossing schemes.

Selection for donor chromosome halves was the best strategy with selection in the final backcross generation for both DH and S_2 crossing schemes (Table 3). However, for the DH schemes, only small improvements for schemes BC_2DH-HH and $BC_3DH-HHH$ were observed compared to the most efficient basic crossing schemes BC_2DH-C and BC_3DH-CC (Table 3). For these small improvements, 720 additional individuals and HT assays had to be invested. For the S_2 schemes, considerable reductions in S of 174 and 236 segments were observed for schemes BC_2S_2-HH and BC_3S_2-HHH with selection in the final backcross generation compared to the basic crossing schemes BC_2S_2-C and BC_2S_2-CC. D_t was only slightly reduced. However, the donor genome proportion on carrier chromosomes D_c was in general very high for the crossing schemes with selection in the final backcross generation with 38–48%. This indicates a fixation of the selection regions of the final backcross generation (Figure 3B and D). In schemes BC_2S_2-CC and BC_3S_2-HH, complete donor chromosomes and donor chromosome halves still appear as blocks around the target segments. These blocks lead to an overlap of donor segments between ILs that reduces the effective resolution of the introgression population for QTL detection. The overlap also hampers the further use of the ILs in the breeding process, as further steps of separation of the target segments by backcrossing are required. We therefore conclude that the crossing schemes with selection in the final backcross generation have the potential to improve the resulting introgression populations at moderate cost. However, for the DH crossing schemes, the number of required HT assays and individuals has to be reduced. For the S_2 crossing schemes, the fixation of large donor chromosome segments has to be avoided. Optimizations of the respective crossing schemes are presented in the following.

Increasing Population Sizes Over Backcross Generations

With constant population sizes of $n_{tot} = 360$ individuals, the population size in generation BC_1 was large in relation to the genetic gains that could be achieved by selecting a comparatively small fraction of 10 or 20 individuals (Table 2). Starting with smaller population sizes in generation BC_1 and gradually increasing population sizes in the following backcross generations was therefore an efficient option to reduce the overall number of required individuals and HT assays for selection in the final backcross generation. Larger population sizes in generation BC_3 also enabled selection for target segments, which was investigated as an option to avoid the fixation of large donor chromosome segments especially for the S_2 crossing schemes.

The schemes $BC_3DH-HHH^*$ and $BC_3DH-HHS^*$ resulted in similar values for all measures (Table 3). However, D_c and D_s were slightly lower for scheme $BC_3DH-HHS^*$. We therefore conclude that selection for target segments already in the final backcross generation is not efficient for DH crossing schemes. In

comparison to the best but also very expensive scheme $BC_3DH-HHH$ with selection in the final backcross generation, scheme $BC_3DH-HHH^*$ can be considered equivalent, but required 360 individuals and HT assays less. In comparison to the more economic basic crossing scheme BC_3DH-CC, scheme $BC_3DH-HHH^*$ improved D_c and D_s and thus the separation of target segments. This is also visible in the graphical genotype (Figure 3). The investment in the additional 360 HT assays seems therefore worthwhile (Table 3).

The scheme $BC_3S_2-HHS^*$ resulted in better values than the schemes BC_3S_2-HHH and $BC_3S_2-HHH^*$. Most notably, it resulted in a much lower D_c of 30% compared to 38%. Scheme $BC_3S_2-HHS^*$ resulted in a D_s that was 2.4% lower compared to schemes BC_3S_2-HHH and $BC_3S_2-HHH^*$ and the ranges for D_s were higher (Figure 4). Nevertheless, it resulted in the lowest values of D_t and D_c and the best separation of target genes of all investigated DH and S_2 crossing schemes with comparable population sizes. The comparatively high value of S of 596 segments in combination with reduced values for D_t can in this case be explained by a greatly improved separation of target segments compared to the other S_2 schemes with selection in the final backcross generation. The improved separation of target segments is also visible in the graphical genotype (Figure 3F). This was achieved with 40 HT assays less (Table 3). We therefore expect that scheme $BC_3S_2-HHS^*$ will result in an improved power of QTL detection, and recommend selection for target segments in the final backcross generation for S_2 crossing schemes.

Compared to the best but expensive comparable DH crossing scheme $BC_3DH-HHH$, the S_2 crossing scheme $BC_3S_2-HHS^*$ resulted in similar values and required 400 HT assays less. Overall, we conclude that increasing population sizes over backcross are advantageous and economic for both DH and S_2 crossing schemes. Moreover, crossing scheme $BC_3S_2-HHS^*$ can provide a cheap alternative to comparable DH crossing schemes.

Conclusions

Our study has shown that introgression populations with complete coverage of the donor genome and reasonably clean recipient background can be developed with a limited number of backcross individuals and HT assays. It has provided further insight on how different crossing and selection schemes influence the structure of the resulting introgression populations. The guidelines which have been derived for maize are transferable to other crop species with similar number and length of chromosomes. For crops with different genome size, some considerations are discussed in the following.

Rapeseed is a crop with a large genome of 19 chromosomes, for which efficient protocols of microspore culture are available for DH production. For the large genome of rapeseed, it can be expected that the values for the total donor genome proportion D_t will be lower than those observed for the smaller genome of maize. With the investigated selection index i, the selection pressure on the carrier chromosomes will be reduced with increasing genome size and number of chromosomes. It might therefore be an interesting option for rapeseed to put more weight on the background markers on the carrier chromosomes to achieve an efficient reduction of D_c. As with microspore culture many DH lines can usually be derived from one backcross individual, the advantages of DH production should be more pronounced than for maize. However, the optimal selection strategies for DH crossing schemes in rapeseed should then be similar to those for selfing in maize.

Sugar beet is a crop with a small genome of 9 chromosomes, for which the guidelines for selfing should be most relevant. In smaller genomes, equivalent values of D_t can usually be reached with smaller population sizes and with fewer backcrosses. However, the average length of the chromosomes in cM is also much shorter than in maize. This implies that fewer crossovers occur per meiosis, and that it might require more individuals and backcross generations to effectively separate the target segments. The combined effects of genome size and chromosome length will also depend on the desired number and length of the target segments.

Simulations can considerably facilitate the planning process for the development of introgression populations in different crop species. The derived guidelines can help breeders and geneticists to enhance the genetic variation of narrow based breeding populations of crops.

Author Contributions

Conceived and designed the experiments: EH KCF TP DS MO MF. Performed the experiments: EH KCF. Analyzed the data: EH. Wrote the paper: EH MF.

References

1. Zamir D (2001) Improving plant breeding with exotic genetic libraries. Nat Rev Genet 2: 983–989.
2. McCouch S (2004) Diversifying selection in plant breeding. PLoS Biol 2: e347.
3. Eshed Y, Zamir D (1994) A genomic library of *Lycopersicon pennellii* in *L. esculentum*: a tool for fine mapping of genes. Euphytica 79: 175–179.
4. Keurentjes JJ, Bentsink L, Alonso-Blanco C, Hanhart CJ, Blankestijn-De Vries H, et al. (2007) Development of a near-isogenic line population of *Arabidopsis thaliana* and comparison of mapping power with a recombinant inbred line population. Genetics 175: 891–905.
5. Törjék O, Meyer RC, Zehnsdorf M, Teltow M, Strompen G, et al. (2008) Construction and analysis of 2 reciprocal Arabidopsis introgression line populations. J Hered 99: 396–406.
6. Lin S, Sasaki T, Yano M (1998) Mapping quantitative trait loci controlling seed dormancy and heading date in rice, *Oryza sativa* L., using backcross inbred lines. Theor Appl Genet 96: 997–1003.
7. Cheema KK, Bains NS, Mangat GS, Das A, Vikal Y, et al. (2008) Development of high yielding IR64× *Oryza rufipogon* (Griff.) introgression lines and identification of introgressed alien chromosome segments using SSR markers. Euphytica 160: 401–409.
8. Matus I, Corey A, Filichkin T, Hayes P, Vales M, et al. (2003) Development and characterization of recombinant chromosome substitution lines (RCSLs) using *Hordeum vulgare* subsp. *spontaneum* as a source of donor alleles in a *Hordeum vulgare* subsp. *vulgare* background. Genome 46: 1010–1023.
9. Schmalenbach I, Körber N, Pillen K (2008) Selecting a set of wild barley introgression lines and verification of QTL effects for resistance to powdery mildew and leaf rust. Theor Appl Genet 117: 1093–1106.
10. Liu S, Zhou R, Dong Y, Li P, Jia J (2006) Development, utilization of introgression lines using a synthetic wheat as donor. Theor Appl Genet 112: 1360–1373.
11. Pumphrey MO, Bernardo R, Anderson JA (2007) Validating the *Fhb1* QTL for Fusarium head blight resistance in near-isogenic wheat lines developed from breeding populations. Crop Sci 47: 200–206.
12. Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. J Exper Bot 58: 351–360.
13. Szalma S, Hostert B, LeDeaux J, Stuber C, Holland J (2007) QTL mapping with near-isogenic lines in maize. Theor Appl Genet 114: 1211–1228.
14. Falke K, Sušić Z, Hackauf B, Korzun V, Schondelmaier J, et al. (2008) Establishment of introgression libraries in hybrid rye (*Secale cereale* L.) from an Iranian primitive accession as a new tool for rye breeding and genomics. Theor Appl Genet 117: 641–652.
15. Syed N, Pooni H, Mei M, Chen Z, Kearsey M (2004) Optimising the construction of a substitution library in *Arabidopsis thaliana* using computer simulations. Mol Breeding 13: 59–68.
16. Falke KC, Miedaner T, Frisch M (2009) Selection strategies for the development of rye introgression libraries. Theor Appl Genet 119: 595–603.
17. Welcker C, Sadok W, Dignat G, Renault M, Salvi S, et al. (2011) A common genetic determinism for sensitivities to soil water deficit and evaporative

demand: meta-analysis of quantitative trait loci and introgression lines of maize. Plant Physiol 157: 718–729.
18. Belcher AR, Zwonitzer JC, Santa Cruz J, Krakowsky MD, Chung CL, et al. (2012) Analysis of quantitative disease resistance to southern leaf blight and of multiple disease resistance in maize, using near-isogenic lines. Theor Appl Genet 124: 433–445.
19. Mano Y, Omori F (2013) Flooding tolerance in interspecific introgression lines containing chromosome segments from teosinte (*Zea nicaraguensis*) in maize (*Zea mays subsp. mays*). Ann Bot-London 112: 1125–1139.
20. Lübberstedt T, Frei UK (2012) Application of doubled haploids for target gene fixation in backcross programmes of maize. Plant Breeding 131: 449–452.
21. R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
22. Stam P (1979) Interference in genetic crossing over and chromosome mapping. Genetics 92: 573–594.
23. Mahone GS, Borchardt D, Presterl T, Frisch M (2012) A comparison of tests for QTL mapping with introgression libraries containing overlapping and nonoverlapping donor segments. Crop Sci 52: 2198–2205.
24. Falke K, Frisch M (2011) Power and false-positive rate in QTL detection with near-isogenic line libraries. Heredity 106: 576–584.

Chapter 6

General discussion

Relative cost-efficiency of HT and SM assays

Without DNA markers, a backcross program for the introgression of one target gene was usually completed after six backcross generations (Allard, 1960). Simulations have shown that background selection with DNA markers can save up to four backcross generations, but at the cost of thousands of marker data points which could not have been managed with SM assays (Frisch *et al.*, 1999a). For the introgression of one dominant target gene in maize, I estimated that approximately 16,000 marker data points are required to recover 96% of recipient genome in two backcross generations (Herzog & Frisch, 2011). In a CMS conversion program in rapeseed, 33,000 marker data points would have been required to recover 96% of recipient genome in two backcross generations (Herzog & Frisch, 2013). With SNP chips, both applications of MABC would have required about 150 HT assays per backcross generation. It can therefore be expected that with the emergence of HT assays many applications of MABC which previously were not economic have now come into reach. Most notably, it is now possible to exploit the time-saving potential of background selection for fast and complete restoration of the recipient genotype. The focus of the four studies presented in this thesis was therefore on short backcross programs with only two or three backcross generations.

Using HT assays for background selection was cheaper than using SM assays for a wide range of cost ratios of one HT assay compared to one SM assay (HT:SM), both for gene introgression (Herzog & Frisch, 2011) and CMS conversion of seed parent lines (Herzog & Frisch, 2013). The relative costs of using HT assays instead of SM assays were comparable for both applications for genomes of comparable size. HT assays reduced the cost of marker analysis by 10-55% for cost ratios of HT:SM of 50:1-100:1 and marker densities of one marker every 10 cM (Herzog & Frisch, 2011; Herzog & Frisch, 2013). SM assays were only cheaper than HT assays for high cost ratios of HT:SM of 100:1-200:1 in combination with low marker densities, and in crops with small genomes such as rye.

The relative costs of HT assays were lowest in short two-generation backcross programs with high marker densities of one marker every 5 cM, which aimed at the recovery of high levels of recipient genome. For these scenarios, SM assays were always more expensive. For example, using HT assays for recovering 98% of recipient genome in a two-generation CMS conversion program in sunflower reduced the cost of marker analysis by 44-94% for cost ratios of 100:1-10:1 (Herzog & Frisch, 2013). In three-generation backcross programs, the relative cost-efficiency of HT assays decreased compared to two-generation programs (Herzog & Frisch, 2011). This has to be attributed to increasing marker fixation in advanced backcross generations. The effects of marker fixation for the optimum breeding designs in MABC will be discussed in detail in the following chapters.

HT assays have the potential to increase the relative efficiency of background selection for many applications of MABC. It can be expected that this effect will be even more pronounced in the future, as the costs of HT marker analysis are further decreasing. The biggest advantage of HT assays over SM assays is that they allow to run highly intense, short backcross programs with many background markers also in crops with large genomes such as sunflower, rapeseed or wheat.

Optimal breeding designs for HT assays

In older simulation studies based on SM assays, the most efficient selection strategies were those which reached a given level of recipient genome with the fewest marker data points (Frisch *et al.*, 1999a; Frisch & Melchinger, 2001c; Falke *et al.*, 2009; Prigge *et al.*, 2009). With HT assays such as SNP chips, it is possible to genotype all markers on the linkage map in one assay. The efficiency criterion for these HT assays is therefore not the number of required marker data points, but the number of individuals subjected to background selection (Herzog & Frisch, 2011; Herzog & Frisch, 2013). The most important difference between SM and HT assays is that with SM assays only those markers are analyzed in advanced backcross generations which are not yet fixed for the recipient alleles. With HT assays, all markers included in one assay have to be analyzed in every generation as long as some markers are segregating. These different characteristics of SM and HT assays have implications on several aspects of the optimal breeding designs.

Number of backcross generations

For the recovery of recipient genome in MABC, there is a trade-off between the number of backcross generations and the required resource input. A desired level of recipient genome can usually be recovered with fewer markers and considerably smaller population sizes if an additional backcross generation is taken into account (Herzog & Frisch, 2011; Herzog & Frisch, 2013). A central question for MABC therefore is 'Speed at any cost?' (Stam, 2003). For a CMS conversion program in sugarbeet, it was possible to recover 99% recipient genome in three backcross generations with 1,400 marker data points (Herzog & Frisch, 2013). If the same level of recipient genome was to be recovered in two backcross generations, the required number of marker data points was increased ten-fold to 14,000. It therefore seems sensible for SM assays to conduct an additional backcross in order to reduce the financial and

logistic efforts of marker analysis. This principle held also true for HT assays, but the reduction in the number of required assays with an additional backcross was smaller than with SM assays. The increase in the required number of assays would only have been fourfold, and the backcross program could have been completed with 240 HT assays. As in plant breeding the accelerated release of superior cultivars often translates into economic benefits (Morris *et al.*, 2003), it can be expected that the investment in additional HT assays will pay off. A reduction of the number of backcross generations at the expense of more marker analyses is therefore advantageous if HT assays are used.

Variations in population size

Due to increasing marker fixation in advanced backcross generations, the optimum backcross designs for SM assays were characterized by increasing marker densities and population sizes (Frisch *et al.*, 1999a; Frisch & Melchinger, 2001c; Prigge *et al.*, 2009). With the less flexible HT assays, it is difficult or impossible to add markers once an assay has been developed. For constant marker densities, I observed a limit of recipient genome recovery which could not be exceeded by further increasing population sizes (Herzog & Frisch, 2013; Herzog *et al.*, 2013). This limit depended on genome size and the degree of marker fixation. In general, it is reached with smaller population sizes in crops with smaller genomes, and in advanced backcross generations when the number of segregating marker loci and the variance of recipient genome are decreasing. Increasing population size beyond the limit of recipient genome recovery is not economic. Population size should therefore be optimized for every generation of a backcross program. The optimum breeding designs for HT assays are characterized by constant marker densities and decreasing population sizes over backcross generations.

Marker density

With SNP markers, densely covered linkage maps with marker densities of one marker every 5 cM or higher can now be established and genotyped at reasonable costs. However, for most genetic models, marker densities higher than one marker every 10 cM resulted in only marginal improvements of recipient genome recovery (Herzog & Frisch, 2011; Herzog & Frisch, 2013). Only if recipient genome levels of 98-99% were to be achieved in two backcross generations for the large genomes of sunflower and rapeseed, a marker density of one marker every 5 cM was warranted. Beyond this marker density, no further gains in recipient genome recovery were observed for any investigated genetic model. An explanation is that the limiting factor for the recovery of the recipient genome is not the precise estimation of the proportion of recipient genome. Rather, it is the limited number of crossovers per meiosis in short backcross programs with only two to three generations (Frisch *et al.*, 1999a).

However, with marker types that were less abundant in the genome such as AFLPs and SSRs, it was often not possible to evenly cover the genome with polymorphic markers. The effect of the resulting gaps in marker spacing has been studied in simulation studies either by including a random marker distribution, or by using published linkage maps with incomplete marker coverage (Hospital *et al.*, 1992; Frisch *et al.*, 1999a; van Berloo *et al.*, 2001; Herzog & Frisch, 2011). A random marker distribution results in reduced levels of recipient genome even if the number of markers corresponds to an average marker density of one marker every 5 cM (Herzog & Frisch, 2011). The reason is that in uncovered chromosome regions, the content of recipient genome cannot be estimated precisely (Frisch & Melchinger, 2005). The main advantage of SNPs and HT assays for MABC is therefore not primarily that linkage maps with very high marker densities of up to one marker per 1 cM can be generated. Rather, the abundance of SNPs greatly improves the efficiency of MABC by guaranteeing that maps with sufficient density and even marker spacing can be generated.

Selection strategies combining SM and HT assays

Pre-selection on carrier chromosomes

The number of individuals to be genotyped has consistently been identified in simulation studies as the most important factor for the efficiency of MABC (Hospital & Charcosset, 1997; Frisch *et al.*, 1999a; van Berloo *et al.*, 2001; Ribaut *et al.*, 2002). In gene introgression studies based on SM assays, the most efficient breeding designs employed selection strategies which pre-selected a subset of individuals for a few markers located on the chromosomes carrying the target genes (Frisch *et al.*, 1999a; Frisch & Melchinger, 2001c; Prigge *et al.*, 2009). Only this smaller subset of the original backcross population was then subjected to background selection, which resulted in a considerable reduction of the required number of marker data points of up to 75%.

This concept is also valid for HT assays, but only cost-efficient if the pre-selection steps can be conducted with SM assays. In gene introgression programs, preselecting carriers of the target gene in MABC in a two-stage selection strategy will per expectation already reduce the number of HT assays for background selection by 50%. A further reduction in the required number of HT assays can be achieved if an additional selection step at markers flanking the target gene is conducted before background selection.

With such a three-stage selection strategy, the risk of linkage drag is decreased, as tightly linked flanking markers will result in a reduction of the donor chromosome segment attached to the target gene. This donor segment is responsible for the major part of donor genome remaining in the recipient background in gene introgression programs and can still be quite large even in advanced backcross generations (Stam & Zeven, 1981; Young & Tanksley, 1989). While selection at very tightly linked flanking markers will

result in short donor chromosome segments, it can also lead to a reduction of recipient genome recovery, especially in short backcross programs with only two or three generations. The reason is that a high selection intensity on the carrier chromosome can lead to a reduced selection intensity on the non-carrier chromosomes which form the major part of the genome (Hospital *et al.*, 1992; Frisch *et al.*, 1999a; Frisch & Melchinger, 2001a). Three-stage selection therefore entails further design decisions.

In order to reduce selection intensity on the carrier chromosomes, selection at flanking markers can be conducted in two steps (Young & Tanksley, 1989; Hospital & Charcosset, 1997). In the first generation of three-stage selection, all individuals are pre-selected with recombination between the target gene and at least one flanking marker. In the second generation, individuals with recombination between the target genes and both flanking markers are pre-selected. This approach reduces the number of HT assays, but increases the number of required SM assays and requires more logistic effort in the lab than conducting three-stage selection only in one backcross generation (Herzog & Frisch, 2011).

If three-stage selection was only conducted in one backcross generation, selection in generation BC_1 led to a greater reduction in the number of required HT assays than three-stage selection in generation BC_3 (Herzog & Frisch, 2011). This can be explained by the fact that more individuals with recombination between the target gene and both flanking markers were found for background selection in advanced backcross generations. As the recovered levels of recipient genome were approximately equivalent, three-stage selection in generation BC_1 was the more cost-efficient strategy.

The optimal positioning of flanking markers is crucial for efficient control of the donor chromosome segment attached to the target gene (Hospital, 2001; Frisch & Melchinger, 2001a). In general, I considered the smallest distance of flanking markers which had no negative effect on recipient genome recovery as optimal (Herzog & Frisch, 2011). For constant population sizes ranging between 40 and 200 individuals, this was achieved with

flanking marker distances of 20-10 cM. Three-stage selection in generation BC_1 with these optimum distances of flanking markers reduced the overall cost of marker analysis by approximately 20% compared to two-stage selection, irrespective of population size and cost ratio of HT:SM.

A lower marker distance of 5 cM resulted in a reduced recovery of recipient genome compared to two-stage selection. If a reduction of the donor chromosome segment attached to the target gene has high priority, for example because of alleles with negative effect in close proximity, more proximal flanking markers can be chosen in advanced generations (Hospital *et al.*, 1992). The same effect can be achieved by increasing population size in the generation in which three-stage selection is conducted. Larger population sizes increase the probability to find a backcross individual with recombination between both flanking markers and the target gene plus a high proportion of recipient genome (Frisch *et al.*, 1999b). Doubling population size in the generation of three-stage selection shifted the effort in the lab from HT to SM assays and reduced the cost of marker analysis for recovering 99% recipient genome by approximately 20-25% for cost ratios of HT:SM of 200:1-20:1 (Herzog & Frisch, 2011).

It can be concluded that combining SM assays for pre-selection at flanking markers with HT assays for genome-wide background selection is an elegant strategy to efficiently reduce the risk of linkage drag and handle large populations at low cost.

Combining HT and SM assays for background selection on non-carrier chromosmes

For many applications of MABC, a small set of reasonably positioned markers is sufficient for efficient background selection in early backcross generations (Hospital *et al.*, 1992; Visscher *et al.*, 1996; Herzog & Frisch, 2013). Recently, it has been suggested that an efficient option to reduce the required number

of the still costly HT assays would be to pre-select a certain set of individuals with a few SM assays at background marker loci. Only this subset is then subjected to background selection with HT assays (Septiningsih *et al.*, 2013). This option has not yet been investigated in simulation studies, but seems promising for CMS conversion without introgression of target genes.

HT assays such as SNP chips have the advantage of very high throughput, but are only cost-effective if the major proportion of markers is not yet fixed for the recipient alleles. Simulations have shown that background markers get rapidly fixed at the high selection intensities which are typical for backcross programs, the fixed markers then becoming useless (Hospital *et al.*, 1992). In the studies on gene introgression and CMS conversion, over 90% of the background markers got fixed in the first two backcross generations (Herzog & Frisch, 2011; Herzog & Frisch, 2013). From generation BC_4 on, selection on non-carrier chromosomes in gene introgression programs is no longer efficient, as 99% of background markers outside the carrier chromosome are fixed (Hospital & Charcosset, 1997; Herzog *et al.*, 2013). The few remaining segregating marker loci can be genotyped with SM assays at low cost. It was therefore suggestive to study selection strategies in which HT assays are used for genome-wide background selection in early backcross generations, and SM assays in advanced backcross generations.

A prerequisite for this combination of HT and SM assays is that a flexible marker system exists which can be genotyped with both types of assay. KASP is a flexible assay which allows cost-effective genotyping also of small sets of SNPs (Chen *et al.*, 2010). Furthermore, it can be used in combination with HT assays such as SNP chips, given that a set of versatile SNP markers is available which can be converted from HT assays to KASP. Such a marker set has recently been developed for maize (Mammadov *et al.*, 2012).

Using HT assays for background selection in generation BC_1 and SM assays in the following backcross generations reduced the total cost of marker analysis compared to using only HT assays. The cost reduction ranged

from 1-44% in two-generation backcross programs and from 7-61% in three-generation backcross programs for cost ratios of HT:SM of 50:1-200:1 (Herzog & Frisch, 2011). If HT assays were used in generations BC_1 and BC_2 and SM assays in generation BC_3, the costs were reduced by 18-33%, depending on the genome size of the studied crop species (Herzog & Frisch, 2013).

It can be concluded that combinations of HT and SM assays have the potential to considerably reduce the cost of marker analysis. HT assays are suitable for short, intense backcross programs. SM assays are efficient for pre-selection at a few marker loci, and in advanced backcross generations when the major proportion of marker loci is already fixed for the recipient alleles.

SSR multiplexes: Selection strategies for an intermediate level of throughput

Beside SNP chips, which allow genotyping ten-thousands of markers in one assay, different kinds of HT assays with a lower level of throughput are also available (Appleby *et al.*, 2009). An example is multiplex PCR for SSR markers, which has been suggested to considerably reduce the cost of PCR-related reagents (Merdinoglu *et al.*, 2005). With chromosome-wise SSR multiplexes for background selection, I investigated an example from practical resistance breeding in grapevine in cooperation with the Julius Kühn Institute (JKI), Institute for Grapevine Breeding Geilweilerhof (Herzog *et al.*, 2013). One multiplex comprised all markers located on one chromosome. This resulted in multiplexes of 5 to 11 loci, depending on the chromosome length.

In order to reduce the number of multiplexes, a selection strategy was developed which pre-selected backcross individuals with the highest number of chromosomes completely fixed for the recipient alleles. This selection strategy reduced the number of required multiplexes by about 7%, but only

in backcross programs with four or more backcross generations. For shorter backcross programs, it was more advantageous to select for the highest proportion of recipient alleles at all background marker loci. An explanation is that in early backcross generations the number of individuals with a high number of chromosomes completely fixed for the recipient alleles was very low. Only this very small subset was evaluated at all background markers, which resulted in overall losses of recipient genome.

Frisch *et al.* (Frisch *et al.*, 1999a; Frisch & Melchinger, 2001c) observed in accordance with my results that selection for the highest proportion of recipient alleles at all background marker loci leads to the highest recovery of recipient genome in two-generation backcross programs. In advanced backcross generations, the differences between different selection strategies diminish considerably, as early gains or losses in recipient genome are balanced by the higher carry-over rate of recipient genome recovery in advanced backcross generations.

The general conclusion that can be drawn from investigating marker systems with different levels of throughput is that the optimal selection strategies and breeding designs are determined by the number of markers included in one assay, their assortment with respect to genome location, and the duration of the backcross program.

Selection indices for HT assays combining foreground and background selection

Combinations of marker assays with different levels of throughput in one backcross program require additional effort in the lab and may not be possible in every breeding program. In these cases, HT assays can be used to analyze both foreground and background markers in one assay. This does not only increase the speed and convenience of marker analysis, but allows also more

flexible selection decisions with respect to the donor genome content on the carrier chromosomes in introgression programs.

With pre-selection at flanking markers, the length of the donor chromosome segment attached to the target gene can only be evaluated with respect to the distance of the flanking markers. High marker densities in proximity to the target gene allow more differentiated selection decisions (van Berloo *et al.*, 2001). It is then possible to preselect the individuals with the shortest donor segments attached to the target region. Moreover, the selection pressure against the donor genome on the carrier chromosomes of target genes decreases with increasing genome size. Hence, it might be advantageous for crops with large genomes to put higher weight on the markers on the carrier chromosomes. This approach allows to discriminate between individuals with identical background marker scores in order to select individuals with better carrier chromosomes (Hospital & Charcosset, 1997). The described increase in flexibility can be achieved by combining foreground and background selection with HT assays in one selection index. I investigated this in the context of the development of maize introgression populations (Herzog *et al.*, 2014). The development of introgression populations had previously only been investigated with SM assays and pre-selection strategies in several steps (Falke *et al.*, 2009).

The selection index in my study was defined as the sum of the donor genome proportion in a selection region plus the recipient genome proportion in the rest of the genome outside the selection region. This index was used to develop introgression populations of 100 introgression lines with target regions of 20 cM length in a genome of 2000 cM length with population sizes of 360 individuals per backcross generation (Herzog *et al.*, 2014). After generation BC_1, in which 360 backcross individuals were generated, the population size was divided into sub-populations of equal size, depending on the length of the selection region. Selection regions could be complete donor chromosomes, chromosome halves or target segments. Selection for complete donor chromosomes from a large BC_1 population of 360 individuals resulted

in a high donor genome proportion on the carrier chromosomes which was often preserved in the final introgression lines. Selection of introgression lines carrying the 20 cM target segments from sub-populations of 18 individuals with the selection index, on the other hand, often resulted in a complete loss of the target segments.

An explanation for the efficiency of comparatively large selection regions is that very stringent selection criteria in early backcross generations usually also require high population sizes (van Berloo *et al.*, 2001). Equivalent recipient genome levels can usually be recovered with much smaller population sizes if selection intensity is not too high in early backcross generations. Theoretical solutions for the minimum population sizes required to find with high probability an individual with the desired genotype in the target region are available (Frisch *et al.*, 1999b). However, to combine this with the requirements for a high recipient genome proportion in the rest of the genome is not straightforward with mathematical models. The optimal length of the selection region should therefore be determined with simulations for every generation of the crossing scheme. For the large chromosomes of maize, the best strategy was selection for donor chromosome halves during the backcross generations and selection for the target segments at the stage of introgression lines (Herzog *et al.*, 2014). The most important criterion for the efficiency of a selection index for HT assays was to find a balance between the length of the selection region and population size.

Conclusions

In this thesis, novel strategies for MABC with HT assays have been developed. The results of my simulations suggest that HT assays have the potential to increase the efficiency of MABC both with respect to the costs of marker analysis and selection gain per unit time. HT assays were cheaper than SM assays for genome-wide background selection for a wide range of cost ratios of HT:SM and genetic crop models.

The optimum breeding designs for HT assays differed from those for SM assays with respect to marker density, selection strategy and population size due to the different characteristics of both types of assay. In contrast to SM assays, the number of required marker data points is only of secondary importance for HT assays. Rather, the most important cost factor for HT assays is the number of individuals to be genotyped.

Depending on the level of throughput, the optimum breeding designs for HT assays were determined by the number of markers included in one assay, their assortment with respect to genome location, and the duration of the desired backcross program. HT assays with very high throughput, such as SNP chips, were most efficient in short, highly intense backcross programs with only two or three backcross generations.

Nevertheless, SM assays were more cost-efficient whenever the analysis of only a few marker loci was required. This was the case in foreground selection for target genes, selection at flanking markers for the control of linkage drag, or for background selection in advanced backcross generations when only very few background markers remained segregating. Combining SM and HT assays for different stages of a MABC program consequently further reduced the cost of marker analysis compared to using only HT assays.

Using HT assays for foreground and background selection in one combined index allowed more differentiated selection decisions with respect to the donor genome proportion on carrier chromosomes. This was especially useful for the development of introgression populations with a limited number of backcross individuals and HT assays.

Chapter 7

Summary

Marker-assisted backcrossing (MABC) is the most successful application of DNA markers in plant breeding. While foreground selection for a few loci of interest with single-marker (SM) assays has become a routine application in breeding programs, the large-scale implementation of genome-wide background selection for the recovery of the genotype of the recipient has lagged behind expectations due to the high costs of marker analysis. It has been hypothesized that this problem will be overcome by high-throughput (HT) marker assays which enable genotyping a high number of marker loci at comparatively low cost per individual marker data point. The optimal backcross designs for HT assays have previously not been investigated. The objective of the present study was therefore the development of novel selection strategies for the efficient use of HT assays in different applications of MABC. For this purpose, computer simulations were employed to investigate backcross programs for different crops.

Gene introgression for maize and conversion of seed parent lines to cytoplasmic male sterility (CMS) for rye, sugarbeet, sunflower and rapeseed were simulated with HT and SM assays. Using HT assays for background selection was cheaper than using SM assays for a wide range of cost ratios of one HT assay compared to one SM assay, both for gene introgression and CMS conversion of seed parent lines. The cost-efficiency of HT assays was

greatest in short, highly intense backcross programs, while it decreased with increasing marker fixation in advanced backcross generations.

With SM assays, only those background markers have to be analyzed in advanced backcross generations which have not been fixed for the recipient alleles in previous backcross generations. Due to the increasing degree of marker fixation in advanced backcross generations, the optimal breeding designs for SM assays were characterized by increasing marker densities and population sizes. With HT assays, all markers in the assay have to be analyzed in every analysis step as long as some marker loci remain segregating. Moreover, it is difficult to add additional markers once an assay has been developed. The optimal breeding designs for HT assays in the present study were consequently characterized by few backcross generations, constant marker densities and decreasing population sizes.

A three-stage strategy which employed SM markers for selection for the target gene and at flanking markers, and HT assays for genome-wide background selection reduced the overall cost of marker analysis by about 20%. This strategy also enabled the handling of large population sizes for efficient reduction of the linkage drag by tightly linked flanking markers at low cost. Conducting background selection with HT assays in early backcross generations and with SM assays in advanced backcross generations also reduced the total cost of marker analysis. This was most pronounced when HT assays were the most expensive. Selection strategies which combine SM and HT assays at different stages of a backcross program are therefore an elegant way to further reduce the cost of MABC.

A gene introgression program in grapevine was investigated with HT assays with an intermediate level of throughput. The optimal selection strategies for chromosome-wise SSR multiplexes depended on the duration of the backcross program. Pre-selection of individuals with complete recipient chromosomes reduced the costs of marker analysis by 7% in backcross programs with four or more backcross generations, but not in shorter backcross programs. The optimal selection strategies for a given level of throughput are

consequently determined by the assortment of markers in the assay and the duration of the backcross program.

Combinations of SM and HT assays in one backcross program increase the effort in the laboratory and may not be possible in every breeding program. In these cases, HT assays can be used to analyze both foreground and background markers in one assay. This was investigated with a selection index for the development of introgression populations in maize. The index was defined as the sum of the donor genome proportion in a selection region plus the recipient genome proportion in the rest of the genome outside the selection region. The index allowed more differentiated selection decisions with respect to the ratio of the donor genome proportion on the carrier chromosomes and the recipient genome on non-carrier chromosomes. The most important criterion for the efficiency of this selection index for HT assays was to find a balance between the length of the selection region and population size.

It can be concluded that HT assays have the potential to increase the relative efficiency of background selection for many applications of MABC, as was demonstrated for gene introgression in maize and grapevine, CMS conversion in rye, sugarbeet, sunflower and rapeseed, and the development of introgression populations in maize.

Kapitel 8

Zusammenfassung

Markergestützte Rückkreuzung ist bislang die erfolgreichste Anwendung von DNA-Markern in der Pflanzenzüchtung. Während Vordergrundselektion für einige wenige Zielgene mit Einzelmarkeranalysen mittlerweile routinemäßig in Zuchtprogrammen eingesetzt wird, ist die Anwendung der genomweiten Hintergrundselektion zur Wiederherstellung des Genotyps des Rezipienten lange Zeit hinter den Erwartungen zurückgeblieben. Die Ursachen lagen in den hohen Kosten und dem hohen Aufwand begründet, der für die Vielzahl der benötigten Einzelmarkeranalysen erforderlich ist. Eine Lösung für dieses Problem stellen Hochdurchsatzanalysemethoden wie SNP-Chips dar. Mit diesen Hochdurchsatzmarkeranalysen kann eine hohe Anzahl von Markern zu vergleichsweise geringen Kosten pro Markerdatenpunkt genotypisiert werden. Die optimalen Selektionsstrategien für Hochdurchsatzmarkeranalysen wurden bislang noch nicht untersucht. Das Ziel der vorliegenden Arbeit war es daher, neue Strategien für den effizienten Einsatz von Hochdurchsatzmarkeranalysen in verschiedenen Anwendungen der markergestützten Rückkreuzung zu entwickeln. Zu diesem Zweck wurden Computersimulationen markergestützter Rückkreuzungsprogramme in verschiedenen Kulturarten durchgeführt.

Genintrogression bei Mais sowie die Einlagerung cytoplasmatisch-männlicher Sterilität (CMS) bei Roggen, Zuckerrübe, Sonnenblume und

Raps wurden sowohl mit Hochdurchsatz- als auch mit Einzelmarkeranalysen simuliert. Hochdurchsatzmarkeranalysen reduzierten die Kosten der Hintergrundselektion für eine große Bandbreite an Kostenverhältnissen von Hochdurchsatz- zu Einzelmarkeranalysen. Dies galt sowohl für Genintrogression als auch für die Einlagerung von CMS. Die Kosteneffizienz war am größten in kurzen Rückkreuzungsprogrammen mit dem Ziel hohen Selektionsgewinns in nur zwei Rückkreuzungsgenerationen, nahm jedoch mit zunehmender Markerfixierung in fortgeschrittenen Rückkreuzungsgenerationen ab.

Bei der Verwendung von Einzelmarkeranalysen werden in fortgeschrittenen Rückkreuzungsgenerationen nur die Marker analysiert, die noch nicht für das Rezipientenallel fixiert sind. Effiziente Züchtungsschemata für Einzelmarkeranalysen sind daher durch ansteigende Markerdichten und Populationsgrößen gekennzeichnet. Bei der Verwendung von Hochdurchsatzmarkeranalysen wird der komplette Markersatz in jedem Analyseschritt analysiert, solange noch Marker segregieren. Darüber hinaus ist das Hinzufügen neuer Marker zu einem einmal entwickelten Hochdurchsatzchip nicht einfach umsetzbar. Optimale Züchtungsschemata für Hochdurchsatzmarkeranalysen waren in der vorliegenden Arbeit daher durch wenige Rückkreuzungsgenerationen, konstante Markerdichte und abnehmende Populationsgrößen charakterisiert.

Eine dreistufige Selektionstrategie, die Einzelmarkeranalysen für die Selektion am Ziellocus und an flankierenden Markern nutzte, und Hochdurchsatzmarkeranalysen für die genomweite Hintergrundselektion, senkte die Gesamtkosten für die Genotypisierung um etwa 20%. Diese Selektionsstrategie ermöglichte auch den Einsatz großer Populationen zur Reduktion des Donorchromosomensegments am Zielgen durch eng gekoppelte flankierende Marker zu niedrigen Kosten. Der Einsatz von Hochdurchsatzmarkeranalysen für die Hintergrundselektion in frühen Rückkreuzungsgenerationen und von Einzelmarkeranalysen in fortgeschrittenen Rückkreuzungsgenerationen reduzierte ebenfalls die Kosten für die Genotypisierung. Selektionsstrategien, die den Einsatz von Einzel- und Hochdurchsatzmarkeranalysen in verschiedenen

Phasen eines Rückkreuzungsprogramms kombinieren, sind daher eine elegante Möglichkeit, die Kosten der markergestützten Rückkreuzung weiter zu reduzieren.

Ein Genintrogressionsprogramm bei Reben mit einem Markeranalysesystem mit mittlerem Durchsatz wurde ebenfalls untersucht. Die optimalen Selektionsstrategien für chromosomenweise SSR-Multiplexe wurden durch die Dauer des Rückkreuzungsprogramms bestimmt. Die Vorselektion von Individuen mit kompletten Rezipientenchromosomen reduzierte die Kosten der Markeranalyse um etwa 7% in Rückkreuzungsprogrammen mit vier oder mehr Generationen, nicht aber in kürzeren Rückkreuzungsprogrammen. Die optimalen Selektionsstrategien für ein bestimmtes Durchsatzniveau werden folglich durch die Anordnung der Marker im Assay sowie durch die Dauer des Rückkreuzungsprogramms bestimmt.

Kombinationen von Einzel- und Hochdurchsatzmarkeranalysen in einem Rückkreuzungsprogramm erhöhen den logistischen Aufwand im Labor und sind nicht immer umsetzbar. In diesen Fällen können sowohl Vordergrund- als auch Hintergrundselektion mit einem Hochdurchsatzchip durchgeführt werden. Dies wurde anhand eines Selektionsindexes für die Entwicklung von Introgressionspopulationen bei Mais untersucht. Der Index war als die Summe des Donorgenomanteils innerhalb einer bestimmten Selektionsregion und des Rezipientengenomanteils im Rest des Genoms definiert. Der Index ermöglichte differenziertere Selektionsentscheidungen in Hinblick auf das Verhältnis von Donorgenomanteil auf den Trägerchromosomen der Zielsegmente und Rezipientengenomanteil im Rest des Genoms. Das wichtigste Kriterium für die Effizienz des Selektionsindexes war das Finden einer Balance zwischen der Länge der Selektionsregion und der Populationsgröße.

Wie an den Beispielen der Genintrogression bei Mais und Reben, der CMS-Einlagerung bei Roggen, Zuckerrübe, Sonnenblume und Raps sowie der Entwicklung von Introgressionspopulationen bei Mais gezeigt wurde, können Hochdurchsatzmarkeranalysen die Effizienz vieler Anwendungen der markergestützten Rückkreuzung im Vergleich zu Einzelmarkeranalysen erhöhen.

References

Allard, RW. 1960. *Principles of plant breeding.* John Wiley & Sons, New York.

Appleby, N, Edwards, D, & Batley, J. 2009. New technologies for ultra-high throughput genotyping in plants. *Pages 19–39 of:* Somers, DJ, Langridge, P, & Gustafson, JP (eds), *Plant Genomics. Methods and Protocols.* Humana Press, New York.

Bartlett, MS, & Haldane, JBS. 1935. The theory of inbreeding with forced heterozygosis. *J Genet*, **31**(3), 327–340.

Becker, H. 2011. *Pflanzenzüchtung.* Ulmer, Stuttgart.

Beckmann, JS, & Soller, M. 1986. Restriction fragment length polymorphisms and genetic improvement of agricultural species. *Euphytica*, **35**(1), 111–124.

Botstein, D, White, RL, Skolnick, M, & Davis, RW. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*, **32**(3), 314.

Chen, W, Mingus, J, Mammadov, J, Backlund, JE, Greene, T, Thompson, S, & Kumpatla, S. 2010. KASPar: a simple and cost-effective system for SNP genotyping. *Pages 9–13 of: Final program, abstract and exhibit guide of the XVIII international conference on the status of plant and animal genome research, San Diego, CA.*

Collard, BCY, & Mackill, DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos T Roy Soc B*, **363**(1491), 557–572.

Collard, BCY, Jahufer, MZZ, Brouwer, JB, & Pang, ECK. 2005. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. *Euphytica*, **142**(1-2), 169–196.

Datta, K, Baisakh, N, Thet, K Maung, Tu, J, & Datta, S. 2002. Pyramiding transgenes for multiple resistance in rice against bacterial blight, yellow stem borer and sheath blight. *Theor Appl Genet*, **106**(1), 1–8.

Eathington, SR, Crosbie, TM, Edwards, MD, Reiter, RS, & Bull, JK. 2007. Molecular markers in a commercial breeding program. *Crop Sci*, **47**(Supplement 3), S154–S163.

Falke, KC, Miedaner, T, & Frisch, M. 2009. Selection strategies for the development of rye introgression libraries. *Theor Appl Genet*, **119**(4), 595–603.

Frisch, M, & Melchinger, AE. 2001a. The length of the intact donor chromosome segment around a target gene in marker-assisted backcrossing. *Genetics*, **157**(3), 1343–1356.

Frisch, M, & Melchinger, AE. 2001b. Marker-assisted backcrossing for introgression of a recessive gene. *Crop Sci*, **41**(5), 1485–1494.

Frisch, M, & Melchinger, AE. 2001c. Marker-assisted backcrossing for simultaneous introgression of two genes. *Crop Sci*, **41**(6), 1716–1725.

Frisch, M, & Melchinger, AE. 2005. Selection theory for marker-assisted backcrossing. *Genetics*, **170**(2), 909–917.

Frisch, M, Bohn, M, & Melchinger, AE. 1999a. Comparison of selection strategies for marker-assisted backcrossing of a gene. *Crop Sci*, **39**(5), 1295–1301.

Frisch, M, Bohn, M, & Melchinger, AE. 1999b. Minimum sample size and optimal positioning of flanking markers in marker-assisted backcrossing for transfer of a target gene. *Crop Sci*, **39**(4), 967–975.

Gupta, PK, Roy, JK, & Prasad, M. 2001. Single nucleotide polymorphisms: a new paradigm in molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. *Curr Sci India*, **80**(4), 524–535.

Hanson, WD. 1959. Early generation analysis of lengths of heterozygous chromosome segments around a locus held heterozygous with backcrossing or selfing. *Genetics*, **44**(5), 833.

Herzog, E, & Frisch, M. 2011. Selection strategies for marker-assisted backcrossing with high-throughput marker systems. *Theor Appl Genet*, **123**(2), 251–260.

Herzog, E, & Frisch, M. 2013. Efficient marker-assisted backcross conversion of seed-parent lines to cytoplasmic male sterility. *Plant Breeding*, **132**(1), 33–41.

Herzog, E, Töpfer, R, Hausmann, L, Eibach, R, & Frisch, M. 2013. Selection strategies for marker-assisted background selection with chromosome-wise SSR multiplexes in pseudo-backcross programs for grapevine breeding. *Vitis*, **52**(4), 193–196.

Herzog, E, Falke, KC, Presterl, T, Scheuermann, D, Ouzunova, M, & Frisch, M. 2014. Selection strategies for the development of maize introgression populations. *PLOS ONE*, **9**(3), e92429.

Hill, WG. 1993. Variation in genetic composition in backcrossing programs. *J Hered*, **84**(3), 212–213.

Hillel, J, Schaap, T, Haberfeld, A, Jeffreys, AJ, Plotzky, Y, Cahaner, A, & Lavi, U. 1990. DNA fingerprints applied to gene introgression in breeding programs. *Genetics*, **124**(3), 783–789.

Hospital, F. 2001. Size of donor chromosome segments around introgressed loci and reduction of linkage drag in marker-assisted backcross programs. *Genetics*, **158**(3), 1363–1379.

Hospital, F, & Charcosset, A. 1997. Marker-assisted introgression of quantitative trait loci. *Genetics*, **147**(3), 1469–1485.

Hospital, F, Chevalet, C, & Mulsant, P. 1992. Using markers in gene introgression breeding programs. *Genetics*, **132**(4), 1199–1210.

Kuchel, H, Fox, R, Reinheimer, J, Mosionek, L, Willey, N, Bariana, H, & Jefferies, S. 2007. The successful application of a marker-assisted wheat breeding strategy. *Mol Breeding*, **20**(4), 295–308.

Liu, J, Liu, D, Tao, W, Li, W, Wang, S, Chen, P, Cheng, S, & Gao, D. 2000. Molecular marker-facilitated pyramiding of different genes for powdery mildew resistance in wheat. *Plant Breeding*, **119**(1), 21–24.

Mammadov, J, Chen, W, Mingus, J, Thompson, S, & Kumpatla, S. 2012. Development of versatile gene-based SNP assays in maize (*Zea mays* L.). *Mol Breeding*, **29**(3), 779–790.

Markel, P, Shu, P, Ebeling, C, Carlson, GA, Nagle, DL, Smutko, JS, & Moore, KJ. 1997. Theoretical and empirical issues for marker-assisted breeding of congenic mouse strains. *Nat Genet*, **17**(3), 280–284.

Merdinoglu, D, Butterlin, G, Bevilacqua, L, Chiquet, V, Adam-Blondon, A-F, & Decroocq, S. 2005. Development and characterization of a large set of microsatellite markers in grapevine (*Vitis vinifera* L.) suitable for multiplex PCR. *Mol Breeding*, **15**(4), 349–366.

Moreau, L, Charcosset, A, Hospital, F, & Gallais, A. 1998. Marker-assisted selection efficiency in populations of finite size. *Genetics*, **148**(3), 1353–1365.

Morris, M, Dreher, K, Ribaut, J-M, & Khairallah, M. 2003. Money matters (II): costs of maize inbred line conversion schemes at CIMMYT using

conventional and marker-assisted selection. *Mol Breeding*, **11**(3), 235–247.

National Agricultural Statistics Service. 2013. *Acreage*. United States Department of Agriculture.

Neeraja, CN, Maghirang-Rodriguez, R, Pamplona, A, Heuer, S, Collard, BCY, Septiningsih, EM, Vergara, G, Sanchez, D, Xu, K, Ismail, AM, & Mackill, DJ. 2007. A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. *Theor Appl Genet*, **115**(6), 767–776.

Peng, T, Sun, X, & Mumm, RH. 2014a. Optimized breeding strategies for multiple trait integration: I. Minimizing linkage drag in single event introgression. *Mol Breeding*, **33**(1), 89–104.

Peng, T, Sun, X, & Mumm, RH. 2014b. Optimized breeding strategies for multiple trait integration: II. Process efficiency in event pyramiding and trait fixation. *Mol Breeding*, **33**(1), 105–115.

Prigge, V, Maurer, HP, Mackill, DJ, Melchinger, AE, & Frisch, M. 2008. Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. *Theor Appl Genet*, **116**(5), 739–744.

Prigge, V, Melchinger, AE, Dhillon, BS, & Frisch, M. 2009. Efficiency gain of marker-assisted backcrossing by sequentially increasing marker densities over generations. *Theor Appl Genet*, **119**(1), 23–32.

Ragot, M, & Lee, M. 2007. Marker-assisted selection in maize: current status, potential, limitations and perspectives from the private and public sectors. *Pages 117–150 of:* Guimarães, EP, Ruane, J, Scherf, BD, Sonnino, A, & Dargie, JD (eds), *Marker-assisted selection. Current status and future perspectives in crops, livestock, forestry and fish.* FAO, Rome.

Randhawa, HS, Mutti, JS, Kidwell, K, Morris, CF, Chen, X, & Gill, KS. 2009. Rapid and targeted introgression of genes into popular wheat

cultivars using marker-assisted background selection. *PLOS ONE*, **4**(6), e5752.

Ribaut, J-M, Jiang, C, & Hoisington, D. 2002. Simulation experiments on efficiencies of gene introgression by backcrossing. *Crop Sci*, **42**(2), 557–565.

Schön, CC, Melchinger, AE, Boppenmaier, J, Brunklaus-Jung, E, Herrmann, RG, & Seitzer, JF. 1994. RFLP mapping in maize: quantitative trait loci affecting testcross performance of elite European flint lines. *Crop Sci*, **34**(2), 378–389.

Semagn, K, Bjørnstad, Å, & Ndjiondjop, MN. 2006. Progress and prospects of marker assisted backcrossing as a tool in crop breeding programs. *Afr J Biotechnol*, **5**(25), 2588–2603.

Septiningsih, EM, Collard, BCY, Heuer, S, Bailey-Serres, J, Ismail, AM, & Mackill, DJ. 2013. Applying genomics tools for breeding submergence tolerance in rice. *Pages 9–30 of:* Varshney, RK, & Tuberosa, R (eds), *Translational Genomics for Crop Breeding. Volume II: Abiotic Stress, Yield and Quality.* John Wiley & Sons, Ames, Iowa.

Servin, B, Martin, OC, Mézard, M, & Hospital, F. 2004. Toward a theory of marker-assisted gene pyramiding. *Genetics*, **168**(1), 513–523.

Stam, P. 2003. Marker-assisted introgression: speed at any cost? *Pages 19–21 of: Proceedings of the Eucarpia meeting on leafy vegetables genetics and breeding.*

Stam, P, & Zeven, AC. 1981. The theoretical proportion of the donor genome in near-isogenic lines of self-fertilizers bred by backcrossing. *Euphytica*, **30**(2), 227–238.

Syvänen, A-C. 2005. Toward genome-wide SNP genotyping. *Nat Genet*, **37**, S5–S10.

Tanksley, SD, Young, ND, Paterson, AH, & Bonierbale, MW. 1989. RFLP mapping in plant breeding: new tools for an old science. *Nat Biotechnol*, **7**(3), 257–264.

Tesfaye, LM, Bink, MCAM, van der Lans, IA, Gremmen, B, & van Trijp, HCM. 2013. Bringing the voice of consumers into plant breeding with Bayesian modelling. *Euphytica*, **189**(3), 365–378.

Timonova, EM, Leonova, IN, Röder, MS, & Salina, EA. 2013. Marker-assisted development and characterization of a set of *Triticum aestivum* lines carrying different introgressions from the *T. timopheevii* genome. *Mol Breeding*, **31**(1), 123–136.

Utomo, HS, Wenefrida, I, & Linscombe, SD. 2012. Progression of DNA Marker and the Next Generation of Crop Development. *Pages 1–28 of:* Goyal, A (ed), *Crop Plant*. InTech.

van Berloo, R, Aalbers, H, Werkman, A, & Niks, RE. 2001. Resistance QTL confirmed through development of QTL-NILs for barley leaf rust resistance. *Mol Breeding*, **8**(3), 187–195.

Visscher, PM. 1996. Proportion of the variation in genetic composition in backcrossing programs explained by genetic markers. *J Hered*, **87**(2), 136–138.

Visscher, PM. 1999. Speed congenics: accelerated genome recovery using genetic markers. *Genet Res*, **74**(1), 81–85.

Visscher, PM, Haley, CS, & Thompson, R. 1996. Marker-assisted introgression in backcross breeding programs. *Genetics*, **144**(4), 1923–1932.

Vos, P, Hogers, R, Bleeker, M, Reijans, M, van De Lee, T, Hornes, M, Friters, A, Pot, J, Paleman, J, Kuiper, M, & Zabeau, M. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res*, **23**(21), 4407–4414.

Wilde, F, Schön, CC, Korzun, V, Ebmeyer, E, Schmolke, M, Hartl, L, & Miedaner, T. 2008. Marker-based introduction of three quantitative-trait loci conferring resistance to Fusarium head blight into an independent elite winter wheat breeding population. *Theor Appl Genet*, **117**(1), 29–35.

Willcox, MC, Khairallah, MM, Bergvinson, D, Crossa, J, Deutsch, JA, Edmeades, GO, González-de León, D, Jiang, C, Jewell, DC, Mihm, JA, Williams, WP, & Hoisington, D. 2002. Selection for resistance to southwestern corn borer using marker-assisted and conventional backcrossing. *Crop Sci*, **42**(5), 1516–1528.

Xu, Y, & Crouch, JH. 2008. Marker-assisted selection in plant breeding: from publications to practice. *Crop Sci*, **48**(2), 391–407.

Young, ND, & Tanksley, SD. 1989. RFLP analysis of the size of chromosomal segments retained around the *Tm-2* locus of tomato during backcross breeding. *Theor Appl Genet*, **77**(3), 353–359.

Zietkiewicz, E, Rafalski, A, & Labuda, D. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics*, **20**(2), 176–183.

Acknowledgments

I am very grateful to my academic supervisor Prof. Dr. Matthias Frisch for his advise, suggestions and support during this thesis work.

Thanks to Prof. Dr. Dr. h.c. Wolfgang Friedt for serving on my graduate commitee.

Sincere thanks for sharing information on ongoing research and for good collaboration in our joint projects to Prof. Dr. Reinhard Töpfer, Dr. Ludger Hausmann and Dr. Rudolf Eibach from the Julius Kühn Institute, and the Amaizing team: Priv.-Doz. Dr. Christin Falke, Dr. Thomas Presterl, Dr. Daniela Scheuermann and Dr. Milena Ouzunova.

Many thanks to my office mate Nina Hofheinz for being excellent company on travels around the world and suffering and celebrating with me through the ages.

Many thanks to my colleague Carola Zenke-Philippi for proof-reading.

Thanks to Dr. Gabriel Schachtel for awakening my interest in biometry.

Thanks to Mrs. Renate Schmidt for being of great help in organisational matters.

Last but not least, I would like to thank all my colleagues, family and friends for their encouragement and support.

Eidesstattliche Erklärung

Ich erkläre:

Ich habe die vorgelegte Dissertation selbständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt, die ich in der Dissertation angegeben habe.

Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben, die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht.

Bei den von mir durchgeführten und in der Dissertation erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der „Satzung der Justus-Liebig-Universität Gießen zur Sicherung guter wissenschaftlicher Praxis“ niedergelegt sind, eingehalten.

Eva Herzog

Gießen, 28. Februar 2014