Aus dem Institut für
Pflanzenzüchtung, Saatgutforschung und Populationsgenetik
der Universität Hohenheim
Fachgebiet Angewandte Genetik und Pflanzenzüchtung
Prof. Dr. A. E. Melchinger

Development and applications of Plabsoft: A computer program for population genetic data analyses and simulations in plant breeding

Dissertation
zur Erlangung des Grades eines Doktors
der Agrarwissenschaften

vorgelegt

der Fakultät Agrarwissenschaften

von
Diplom-Agrarbiologe
Hans Peter Maurer
aus Neuendettelsau

2008

Die vorliegende Arbeit wurde am 17. August 2007 von der Fakultät Agrarwissenschaften als ,,Dissertation zur Erlangung des Grades eines Doktors der Agrarwissenschaften (Dr. sc. agr.)" angenommen.

Tag der mündlichen Prüfung:	9. Januar 2008
1. Prodekan:	Prof. Dr. W. Bessei
Berichterstatter, 1. Prüfer:	Prof. Dr. A.E. Melchinger
Mitberichterstatter, 2. Prüfer:	Prof. Dr. H.-P. Piepho
3. Prüfer:	Prof. Dr. R. Blaich

Contents

[1]Maurer, H.P., A.E. Melchinger, and M. Frisch. 2007. Theor. Appl. Genet. 115:393–398.
[2]Maurer, H.P., C. Knaak, A.E. Melchinger, M. Ouzunova, and M. Frisch. 2006. Maydica 51:269–279.
[3]Schrag*, T.A., H.P. Maurer*, A.E. Melchinger, H.-P. Piepho, J. Peleman, and M. Frisch. 2007. Theor. Appl. Genet. 114:1345–1355.
[4]Maurer, H.P., A.E. Melchinger, and M. Frisch. 2008. Euphytica 161:133–139.
[5]Prigge*, V., H.P. Maurer*, D.J. Mackill, A.E. Melchinger, and M. Frisch. 2007. Theor. Appl. Genet. 116:739–744.
[6]Falke*, K.C., H.P. Maurer*, A.E. Melchinger, H.-P. Piepho, C. Flachenecker, and M. Frisch. 2007. Theor. Appl. Genet. 115:289-297.
[7]Stich, B., A.E. Melchinger, H.-P. Piepho, S. Hamrit, W. Schipprack, H.P. Maurer, and J.C. Reif. 2007. Theor. Appl. Genet. 115:529–536.
*Both authors contributed equally.

Abbreviations

AFLP	amplified fragment length polymorphism
DNA	deoxyribonucleic acid
HWE	Hardy–Weinberg equilibrium
IRRI	International Rice Research Institute
LD	linkage disequilibrium
M	Morgan
QTL	quantitative trait locus (or loci, depending on the context)
SSR	simple sequence repeat

Chapter 1

General Introduction

The availability of molecular markers and DNA sequences is no longer a limiting factor for genetic studies of economically important crop species (Varshney et al., 2005). Genotyping of many individuals with a large number of markers is routinely conducted in applied maize plant breeding programs (Bernardo and Yu, 2007). Genotyping of individuals is promising to (i) detect genes and alleles underlying important agronomic traits (Mackay and Powell, 2007), (ii) predict hybrid performance based on marker data from parental lines (Vuylsteke et al., 2000), and (iii) select desirable plants in marker-assisted backcrossing programs (Frisch and Melchinger, 2005). Bioinformatic tools for data analyses and simulation of entire plant breeding programs are urgently required to facilitate the integration of the above applications in applied plant breeding programs (Peleman and Rouppe van der Voort, 2003).

The first concepts for stochastic simulation of population genetical problems were developed with the advent of computers (Fraser, 1957) and applied, for example, to simulate the long-term selection response in reciprocal recurrent selection with different selection schemes (Cress, 1967). However, until recently, the available computing resources strongly restricted the complexity of the investigated scenarios. The first simulations of marker applications in plant breeding investigated marker-assisted backcrossing (Hospital et al., 1992) and marker-assisted selection (Gimelfarb and Lande, 1994). These simulations were carried out with software, written especially for the problem

under investigation, because a generic software for carrying out complex simulations was not available. The programs QU-GENE (Podlich and Cooper, 1998) and Plabsim (Frisch et al., 2000) were simulation tools targeting at a more flexible approach to simulate plant breeding programs. They provided an interface for describing the scenarios to be investigated and did not require knowledge of the underlying programming language. Both were employed in several studies (QU-GENE: Wang et al., 2003, 2005; Plabsim: Frisch et al., 1999; Frisch and Melchinger, 2001), but their functionality was restricted to only a few predefined types of breeding schemes, such as the pedigree and bulk method in wheat breeding (QU-GENE) or marker-assisted introgression of one or two target genes (Plabsim).

In conclusion, the optimization of conventional and molecular marker-based plant breeding programs demands a powerful and user-friendly simulation platform that allows to model complex breeding plans and various genetic architectures of the traits under consideration. A tight integration of the simulation platform with data analysis tools is required to guarantee an efficient integration of marker-based selection schemes into applied breeding programs. The development of such a simulation software was the subject of my thesis work.

1.1 Data analysis

1.1.1 Hardy–Weinberg Equilibrium

The assumption of Hardy–Weinberg equilibrium (Hardy, 1908; Weinberg, 1908) is the basis of many concepts in population genetics and quantitative genetics (Crow, 1988). Therefore, tests for Hardy–Weinberg equilibrium are of crucial importance in plant, animal and human genetics as well as evolutionary studies. Tests for Hardy–Weinberg equilibrium are employed to (i) gather information on the mating system and genetic structure of wild

and breeding populations (*e.g.*, Semerikov et al., 2002; Reif et al., 2004), (ii) detect population admixture (*e.g.*, Deng et al., 2001), (iii) reveal marker phenotype associations (*e.g.*, Nielsen et al., 1999), and (iv) identify systematic genotyping errors (*e.g.*, Xu et al., 2002).

Asymptotic goodness-of-fit tests or exact tests based on the probability of occurrence of genotype arrays can be used to test for Hardy–Weinberg law (Weir, 1996). If the contingency table of observed genotype counts has sparse cells or the sample size is small, it is known that asymptotic goodness-of-fit tests have poor statistical properties. Exact tests are computationally demanding, but they are to be preferred over asymptotic goodness-of-fit tests, because they do not require large sample assumptions. Exact p-values can be calculated for small population sizes via computationally demanding enumeration methods (Louis and Dempster, 1987) and approximated for large population sizes via Monte Carlo methods (Guo and Thompson, 1992; Huber et al., 2006).

Aoki (2003) proposed a network algorithm for an incomplete enumeration method to reduce the computational efforts of Hardy–Weinberg tests. However, it is still not possible to carry out exact tests for many molecular marker data sets with Aoki's (2003) algorithm, because the required computing time is still to long. It is of great importance to extend the computational feasibility of exact tests to data sets commonly available in plant breeding. Therefore, faster tests need to be developed and implemented in software.

1.1.2 Linkage disequilibrium

Linkage disequilibrium (LD) is the non-random association of alleles at different loci within a population, *i.e.*, the alleles at two loci are occurring together more often than it is expected under random mating. The amount and distribution of LD across the genome depends on the genealogy of a population sample. Moreover, mutation, random genetic drift, selection within

populations, and migration resulting in admixture between populations can also cause LD, while in random mating populations LD is reduced in each generation through recombination.

LD mapping (often also called association mapping) is an approach to detect genes and alleles of interest in breeding populations (Lynch and Walsh, 1997). The resolution of LD mapping studies depends on the extent and distribution of LD in the population. A prerequisite for LD mapping are chromosome segments in LD, which ideally harbor a molecular marker and a locus responsible for the trait of interest. LD mapping studies were suggested as a strategy for a systematic exploitation of the diversity present in breeding populations (Jannink et al., 2001). To successfully implement LD mapping studies in plant breeding programs, two prerequisites need to be met: (1) a software for determining and comparing the LD present in plant breeding populations with different population sizes, and (2) detailed knowledge about the amount and distribution of LD in the breeding program under consideration.

Available software for carrying out LD analysis such as Arlequin (Excoffier et al., 2005), Powermarker (Liu and Muse, 2005), and other (*c.f.*, Excoffier and Heckel, 2006) lack a significance test for commonly used LD measures such as D', D'_m, r^2 and R (Maurer et al., 2006). However, such a test is of particular importance, because the size of the LD coefficients depends strongly on the allele frequencies in the investigated population. Therefore, the absolute values of LD coefficients are typically less informative than information about their statistical significance. Assessment of the prospects of LD mapping approaches in applied maize breeding programs requires a detailed knowledge about the amount and distribution of LD in modern breeding germplasm and this knowledge is entirely lacking.

1.1.3 Haplotype blocks

Due to linkage, alleles at adjacent loci are mostly inherited together. Therefore, allele frequencies at linked markers are often highly correlated. This can result in an overestimation of QTL effects and in a reduced power of QTL detection. Combining adjacent markers to so-called haplotype blocks can reduce this problem because (1) haplotype blocks correspond directly to the biologically functional unit, (2) population genetics has shown that sequence variation is structured into haplotypes blocks, and (3) haplotype blocks often have the statistical advantage of reducing the dimension of statistical tests involved (Clark, 2004). Alternative approaches have been proposed for finding haplotype blocks (Anderson and Novembre, 2003; Gabriel et al., 2002; Jansen et al., 2003; Patil et al., 2001; Zhang et al., 2002). However, the usability of haplotype block-based approaches for LD mapping or hybrid performance prediction in the context of applied plant breeding programs has not yet been investigated.

A first prerequisite to implement haplotype block-based methods in plant breeding is the availability of suitable algorithms and software to determine the haplotype structure of breeding material with alternative methods. Before starting this thesis work, no such algorithms and software were available. A second prerequisite are investigations on the relative advantages of haplotype-based methods compared with single marker-based methods. An important problem in hybrid breeding is the prediction of the hybrid performance of potential hybrids not tested in field trials. In this context, prediction methods based on haplotype blocks of parental inbred lines are regarded as a promising alternative to single marker-based methods. However, no studies investigating haplotype block-based prediction methods in hybrid breeding were available.

1.2 Simulation

1.2.1 Software development

Random mating, infinite population size, and absence of selection are often assumed in the derivation of analytical solutions for population genetical problems. However, in plant breeding, finite populations are derived from planned crosses, and selection is carried out to achieve breeding progress. Therefore, population genetical problems for which no suitable analytical solutions are available oftentimes occur in plant breeding. Examples comprise the expected LD decay under planned crossing schemes such as chain-crossing or the development of the additive genetic variance under recurrent full-sib selection (Falke et al., 2007b). Computer simulations can be employed to solve such problems and to obtain approximate solutions. Several software programs for population genetical simulation have been developed such as GREGOR (Tinker and Mather, 1993), QU-GENE (Podlich and Cooper, 1998), SIMCOAL (Laval and Excoffier, 2004), but they lack functions required for modeling applied plant breeding programs and interfaces to suitable data analysis software.

For a successful integration of molecular marker-based selection schemes in plant breeding, a software is required that allows to simulate entire applied plant breeding programs with and without marker-assisted selection and to determine the cost relevant parameters for alternative experimental settings. Furthermore, integrated routines are required for analyzing the simulated data. Before starting this thesis work such a software was not available.

1.2.2 Validation with experimental data

Simulation studies are used to draw conclusions on the optimum design of marker-assisted plant breeding programs. In particular, they were often em-

ployed to determine the optimum design of marker-assisted backcrossing programs (*c.f.*, Hospital, 2005; Frisch and Melchinger, 2005).

The results of computer simulations depend on the model of meiosis employed in developing the simulation software. In Plabsoft, the absence of interference in crossover formation (Stam, 1979) was assumed, resulting in a model mathematically equivalent to that underlying Haldane's mapping function (Haldane, 1919). Only if the model employed in the software is in good accordance with reality, the simulation results are relevant for optimizing applied breeding programs. However, studies comparing simulated with experimental datasets, and hence, validating whether the models implemented in the software are met to a sufficient degree in reality, are entirely lacking, particularly in the area of marker-assisted backcrossing programs.

1.3 Integrated simulation and data analysis

1.3.1 Recurrent selection

Recurrent selection is a breeding method designed for obtaining long-term selection response by increasing the frequency of favorable alleles in a breeding population while maintaining the genetic variance. A constantly high additive genetic variance is required to obtain long-term selection response, but the extent of the additive variance is reduced by strong LD. Therefore, Johnson (1982) suggested three generations of random intermating before starting with the selection process of a recurrent selection program to reduce the strong initial LD present in F_2 populations. For random intermating, theoretical concepts exist that describe the LD decay due to intermating. However, for mating schemes such as chain-crossing no theoretical results on the LD decay are available.

In a long-term study at the University of Hohenheim, two recurrent selection programs with maize were conducted (Flachenecker et al., 2006a,

b, c; Falke et al., 2007a), in which the F_2 base population was three times intermated with the chain-crossing method before starting selection. To determine the theoretical expectation of the LD decay in the intermating generations of this breeding program, a software was required which would allow to (1) simulate the breeding scheme repeatedly as carried out in practice and (2) analyze the data generated in each simulation replication for LD. Before starting this thesis work such a software was not available.

1.3.2 Linkage disequilibrium in a breeding program

Knowledge about the amount and distribution of LD in a plant breeding population is an important factor to draw conclusions on the possible selection response and on the prospects of LD mapping methods. However, in addition to the information on the amount of LD in an applied breeding program, it is of utmost importance to determine the underlying causes for the observed results. Only if breeders are aware of the factors causing LD in a breeding population, they can influence the extent of LD by adapting their breeding schemes.

Conducting large breeding programs repeatedly for scientific reasons is not possible due to high costs. Therefore, simulation studies are the only possibility to investigate the forces generating and conserving LD in breeding programs. Before starting this thesis work no simulation and data analysis tools were available to start such a simulation study.

1.4 Objectives

The goal of this thesis was to develop an integrated computer program for population genetic simulation and data analyses and apply the software to problems occurring in applied genetics research and plant breeding. In particular, the objectives were to

(1) develop and implement in software an incomplete enumeration algorithm for an exact test of Hardy–Weinberg proportions, which is faster than existing algorithms (Chapter 2),

(2) implement in software the computation of several LD coefficients, develop and implement significance tests for LD coefficients D'_m and R, and employ these by analyzing a commercial hybrid maize breeding program to draw conclusions about the prospects of LD mapping (Chapter 3),

(3) develop and implement in software methods for haplotype block detection and employ these by determining the haplotype block structure in a hybrid maize breeding program to predict hybrid performance (Chapter 4),

(4) develop algorithms for simulation of plant breeding programs and implement these in an integrated software, comprising also the data analysis routines of objectives (1) to (3) (Chapter 5),

(5) compare the agreement between results from computer simulations and those of experimental plant breeding programs by the example of a marker-assisted backcrossing program for transfer of a submergence tolerance QTL in rice (Chapter 6),

(7) determine with simulations the expected LD decay in the intermating generations of two recurrent selection programs in maize (Chapter 7),

(8) investigate with computer simulations the forces generating and maintaining the LD in a hybrid maize breeding program (Chapter 8).

References

Anderson, E.C., and J. Novembre. 2003. Finding haplotype block boundaries by using minimum-description-length principle. Am. J. Hum. Genet. 73:336–354.

Aoki, S. 2003. Network algorithm for the exact test of Hardy–Weinberg proportion for multiple alleles. Biometric. J. 4:471–490.

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

Clark, A.G. 2004. The role of haplotypes in candidate gene studies. Genet. Epidemiol. 27:321–333.

Cress, C.E. 1967. Reciprocal recurrent selection and modifications in simulated populations. Crop Sci. 7:561–567.

Crow, J.F. 1988. Eighty years ago: the beginnings of population genetics. Genetics 119:473–476.

Deng, H.W., W.M. Chen, and R.R. Recker. 2001. Population admixture: Detection by Hardy–Weinberg test and its quantitative effects on linkage-disequilibrium methods for localizing genes underlying complex traits. Genetics 157:885–897.

Excoffier, L., and G. Heckel. 2006. Computer programs for population genetics data analysis: a survival guide. Nat. Rev. Genet. 7:745–758.

Excoffier, L., G. Laval, and S. Schneider. 2005. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47–50.

Falke, K.C., C. Flachenecker, A.E. Melchinger, H.-P. Piepho, H.P. Maurer, and M. Frisch. 2007a. Temporal changes in allele frequencies in two European F_2 flint maize populations under modified recurrent full-sib selection. Theor. Appl. Genet. 114:765–776.

Falke, K.C., H.P. Maurer, A.E. Melchinger, H.-P. Piepho, C. Flachenecker, and M. Frisch. 2007b. Linkage disequilibrium in two European F_2 flint maize populations under modified recurrent full-sib selection. Theor. Appl. Genet. 115:289–297. Erratum: Theor. Appl. Genet. 115:299.

Flachenecker, C., M. Frisch, J. Muminović, K.C. Falke, and A.E. Melchinger. 2006a. Modified full-sib selection and best linear unbiased prediction of progeny performance in a European F_2 maize population. Plant Breeding 125:248–253.

Flachenecker, C., M. Frisch, K.C. Falke, and A.E. Melchinger. 2006b. Trends in population parameters and best linear unbiased prediction of progeny performance in a European F_2 maize population under modified recurrent full-sib selection. Theor. Appl. Genet. 112:483–491.

Flachenecker, C., M. Frisch, K.C. Falke, and A.E. Melchinger. 2006c. Genetic drift and selection effects from recurrent selection programs in two F_2 populations of European flint maize. Theor. Appl. Genet. 113:1113–1120.

Fraser, A.S. 1957. Simulation of genetic systems by automatic digital computers: I. Introduction. Aust. J. Biol. Sci. 10:484–491.

Frisch, M., and A.E. Melchinger. 2001. Marker-assisted backcrossing for simultaneous introgression of two genes. Crop Sci. 41:1716–1725.

Frisch, M., and A.E. Melchinger. 2005. Selection theory for marker-assisted backcrossing. Genetics 170:909–917.

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

Frisch, M., M. Bohn, and A.E. Melchinger. 2000. Plabsim: Software for simulation of marker-assisted backcrossing. J. Hered. 91:86–87.

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229.

Gimelfarb, A., and R. Lande. 1994. Simulation of marker-assisted selection in hybrid populations. Genet. Res. 63:39–47.

Guo, S.W., and E.A. Thompson. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372.

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.

Hardy, G.H. 1908. Mendelian proportions in a mixed population. Science 28:49–50.

Hospital, F. 2005. Selection in backcross programmes. Phil. Trans. R. Soc. B 360:1503–1512.

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

Huber, M., Y. Chen, I. Dinwoodie, A. Dobra, and M. Nicholas. 2006. Monte Carlo algorithms for Hardy–Weinberg proportions. Biometrics 62:49–53.

Jannink, J.L., M.C.A.M. Bink, and R.C. Jansen. 2001. Using complex plant pedigrees to map valuable genes. Trends Plant Sci. 6:337–342.

Jansen, R.C., J.L. Jannink, and W.D. Beavis. 2003. Mapping quantitative trait loci in plant breeding populations: Use of parental haplotype sharing. Crop Sci. 43:829–834.

Johnson, G.R. 1982. Two-locus theory in recurrent selection for general combining ability in maize. Theor. Appl. Genet. 61:279–283.

Laval, G., and L. Excoffier. 2004. SIMCOAL 2.0: a program to simulate genomic diversity over large recombining regions in a subdivided population with a complex history. Bioinformatics 20:2485–2487.

Liu, K., and S.V. Muse. 2005. Powermarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129.

Louis, E.J., and E.R. Dempster. 1987. An exact test for Hardy–Weinberg and multiple alleles. Biometrics 43:805–811.

Lynch, M., and B. Walsh. 1997. Genetics and analysis of quantitative traits. p. 413. Sunderland, MA: Sinauer Assoc.

Mackay, I., and W. Powell. 2007. Methods for linkage disequilibrium mapping in crops. Trends Plant Sci. 12:57–63.

Maurer, H.P., C. Knaak, A.E. Melchinger, M. Ouzunova, and M. Frisch. 2006. Linkage disequilibrium between SSR markers in six pools of elite lines of an European breeding program for hybrid maize. Maydica 51:269–279.

Nielsen, D.M., G. Ehm, and B.S. Weir. 1999. Detecting marker-disease association by testing for Hardy–Weinberg disequilibrium at a marker locus. Am. J. Hum. Genet. 63:1531–1540.

Patil, N., A.J. Berno, D.A. Hinds, W.A. Barrett, J.M. Doshi, C.R. Hacker, C.R. Kautzer, D.H. Lee, C. Marjoribanks, D.P. McDonough, B.T. Nguyen, M.C. Norris, J.B. Sheehan, N. Shen, D. Stern, R.P. Stokowski, D.J. Thomas, M.O. Trulson, K.R. Vyas, K.A. Frazer, S.P. Fodor, and D.R. Cox. 2001. Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. Science 294:1719–1723.

Peleman, J., and J. Rouppe van der Voort. 2003. Breeding by design. Trends Plant Sci. 8:330–334.

Podlich, D.W., and M. Cooper. 1998. QU-GENE: a simulation platform for quantitative analysis of genetic models. Bioinformatics 14:632–653.

Reif, J.C., X.C. Xia, A.E. Melchinger, M.L. Warburton, D.A. Hoisington, D. Beck, M. Bohn, and M. Frisch. 2004. Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR markers. Crop Sci. 44:326–334.

Semerikov, V., A. Belyaev, and M. Lascoux. 2002. The origin of Russian cultivars of red clover (*Trifolium pratense* L.) and their genetic relationships to wild populations in the Urals. Theor. Appl. Genet. 106:127–132.

Stam, P. 1979. Interference in genetic crossing over and chromosome mapping. Genetics 92:573–594.

Tinker, N.A., and D.E. Mather. 1993. GREGOR: Software for genetic simulation. J. Hered. 84:237.

Varshney, R.K., A. Graner, and M.E. Sorrells. 2005. Genomics-assisted breeding for crop improvement. Trends Plant Sci. 10:1360–1385.

Vuylsteke, M., M. Kuiper, and P. Stam. 2000. Chromosomal regions involved in hybrid performance and heterosis: their AFLP (R)-based identification and practical use in prediction models. Heredity 85:208–218.

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

Wang, J., H.A. Eagles, R. Trethowan, and M. van Ginkel. 2005. Using computer simulations of the selection process and known gene information to assist in parental selection in wheat quality breeding. Aust. J. Agric. Res. 56:465–473.

Weinberg, W. 1908. Über den Nachweis der Vererbung beim Menschen. Jahreshefte des Württembergischen Vereins für vaterländische Naturkunde 64:369–382.

Weir, B.S. 1996. Genetic data analysis II, 2nd edition. p. 91. Sunderland, Massachusetts: Sinauer Associates.

Xu, J., A. Turner, J. Little, E.R. Bleecker, and D.A. Meyers. 2002. Positive results in association studies are associated with departure from Hardy–Weinberg equilibrium: hint for genotyping error? Hum. Genet. 111:573–574.

Zhang, K., M. Deng, T. Chen, M.S. Waterman, and F. Sun. 2002. A dynamic programming algorithm for haplotype block partitioning. Proc. Natl. Acad. Sci. USA 99:7335–7339.

An incomplete enumeration algorithm for an exact test of Hardy–Weinberg proportions with multiple alleles

H.P. Maurer[1], **A.E. Melchinger**[1], **M. Frisch**[1]

[1] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*

Received: 17 October 2006 **Accepted:** 27 April 2007 **Published online:** 3 July 2007
Communicated by D.A. Hoisington.

Abstract

Testing of Hardy–Weinberg proportions (HWP) with asymptotic goodness-of-fit tests is problematic when the contingency table of observed genotype counts has sparse cells or the sample size is low, and exact procedures are to be preferred. Exact p-values can be (1) calculated via computational demanding enumeration methods or (2) approximated via simulation methods. Our objective was to develop a new algorithm for exact tests of HWP with multiple alleles on the basis of conditional probabilities of genotype arrays, which is faster than existing algorithms. We derived an algorithm for calculating the exact permutation significance value without enumerating all genotype arrays having the same allele counts as the observed one. The algorithm can be used for testing HWP by (1) summation of the conditional probabilities of occurrence of genotype arrays with smaller probability than the observed one, and (2) comparison of the sum with a nominal Type I error rate α. Application to published experimental data from seven maize populations showed that the exact test is computationally feasible and reduces the number of enumerated genotype count matrices about 30% compared with previously published algorithms.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

Linkage disequilibrium between SSR markers in six pools of elite lines of an European breeding program for hybrid maize[1]

H.P. Maurer[2], C. Knaak[3], A.E. Melchinger[2,*], M. Ouzunova[3], M. Frisch[2]

[2] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*
[3] *KWS SAAT AG, Grimsehlstr. 31, 37555 Einbeck, Germany*

Received: 23 May 2005

Abstract

Detailed knowledge about the patterns and distribution of linkage disequilibrium (LD) is of fundamental importance for mapping of genes by whole-genome association studies. In this study, we (i) analyzed the molecular genetic diversity, (ii) investigated the patterns and distribution of LD in elite maize inbreds, and (iii) assessed the prospects of association mapping methods in breeding materials. Six germplasm pools with a total of 497 elite lines of a commercial breeding program in Europe were fingerprinted by 81 SSR markers covering the entire maize genome. The extent of LD among marker loci was estimated by different measures of LD. Across all germplasm pools, the SSR markers were highly polymorphic, and four groups of inbred lines were detected by principal component and Bayesian cluster analysis. We detected genome-wide LD among pairs of loci. LD between linked markers decreased not only with distance but also with a smaller number of alleles and lower gene diversity. In conclusion, association mapping via a genome-wide scan may be applied to detect associations between marker and traits, but LD between unlinked loci will most likely result in high rates of false positives.

Key Words: Molecular genetic diversity; Linkage disequilibrium; Association mapping

[1] This paper is dedicated to Dr. Donald N. Duvick in recognition of his outstanding contributuins to specific progress in maize gentics and breeding and its successful application in commercial maize breeding.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield

T.A. Schrag[1,⋆], H.P. Maurer[1,⋆], A.E. Melchinger[1], H.-P. Piepho[2], J. Peleman[3], M. Frisch[1]

[1] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*
[2] *Bioinformatics Unit of the Institute for Crop Production and Grassland Research, University of Hohenheim, 70599 Stuttgart, Germany*
[3] *Keygene, P.O. Box 216, 6700 AE Wageningen, The Netherlands*

Received: 26 October 2006 **Accepted:** 4 February 2007 **Published online:** 24 February 2007
Communicated by H.C. Becker.

Abstract

Marker-based prediction of hybrid performance facilitates the identification of untested single-cross hybrids with superior yield performance. Our objectives were to (1) determine the haplotype block structure of experimental germplasm from a hybrid maize breeding program, (2) develop models for hybrid performance prediction based on haplotype blocks, and (3) compare hybrid performance prediction based on haplotype blocks with other approaches, based on single AFLP markers or general combining ability (GCA), under a validation scenario relevant for practical breeding. In total, 270 hybrids were evaluated for grain yield in four Dent × Flint factorial mating experiments. Their parental inbred lines were genotyped with 20 AFLP primer enzyme combinations. Adjacent marker loci were combined into haplotype blocks. Hybrid performance was predicted on basis of single marker loci and haplotype blocks. Prediction based on variable haplotype block length resulted in an improved prediction of hybrid performance compared with the use of single AFLP markers. Estimates of prediction efficiency (R^2) ranged from 0.305 to 0.889 for marker-based prediction and from 0.465 to 0.898 for GCA-based prediction. For inter-group hybrids with predominance of general over specific combining ability, the hybrid prediction from GCA effects was efficient in identifying promising hybrids. Considering the advantage of haplotype block approaches over single marker approaches for the prediction of inter-group hybrids, we see a high potential to substantially improve the efficiency of hybrid breeding programs.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

⋆ Both authors contibuted equally to this work.

Population genetic simulation and data analysis with Plabsoft

H.P. Maurer[1], **A.E. Melchinger**[1], **M. Frisch**[1]

[1] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*

Received: 30 November 2006 **Accepted:** 19 June 2007 **Published online:** 26 July 2007

Abstract

Computer simulations are a useful tool to solve problems in population genetics for which no analytical solutions are available. We developed Plabsoft, a powerful and flexible software for population genetic simulation and data analysis. Various mating systems can be simulated, comprising planned crosses, random mating, partial selfing, selfing, single-seed descent, double haploids, topcrosses, and factorials. Selection can be simulated according to selection indices based on phenotypic values and/or molecular marker scores. Data analysis routines are provided to analyze simulated and experimental datasets for allele and genotype frequencies, genotypic and phenotypic values and variances, molecular genetic diversity, linkage disequilibrium, and parameters to optimize marker-assisted backcrossing programs. Plabsoft has already been employed in numerous studies, we chose some of them to illustrate the functionality of the software.

Key Words: Breeding informatics; Computer simulation; Data analysis; Population genetics

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice

V. Prigge[1,⋆], H.P. Maurer[1,⋆], D.J. Mackill[2], A.E. Melchinger[1], M. Frisch[1]

[1] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*
[2] *International Rice Research Institute, DAPO Box 7777 Metro Manila, The Philippines*

Received: 16 May 2007 **Accepted:** 18 December 2007 **Published online:** 31 January 2008
Communicated by D.A. Hoisington.

Abstract

Computer simulations are useful tools to optimize marker-assisted breeding programs. The objective of our study was to investigate the closeness of computer simulations of the recurrent parent genome recovery with experimental data obtained in two marker-assisted backcrossing programs in rice (*Orzya sativa* L.). We simulated the breeding programs as they were practically carried out. In the simulations we estimated the frequency distributions of the recurrent parent genome proportion in the backcross populations. The simulated distributions were in good agreement with those obtained practically. The simulation results were also observed to be robust with respect to the choice of the mapping function and the accuracy of the linkage map. We conclude that computer simulations are a useful tool for pre-experiment estimation of selection response in marker-assisted backcrossing.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

⋆ Both authors contibuted equally to this work.

Linkage disequilibrium in two European F_2 flint maize populations under modified recurrent full-sib selection

K.C. Falke[1,⋆], H.P. Maurer[1,⋆], A.E. Melchinger[1], H.-P. Piepho[2], C. Flachenecker[1], M. Frisch[1]

[1] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*
[2] *Institute for Crop Production and Grassland Research, University of Hohenheim, 70593 Stuttgart, Germany*

Received: 18 December 2006 **Accepted:** 27 March 2007 **Published online:** 28 April 2007
Communicated by M. Xu.

Abstract

According to quantitative genetic theory, linkage disequilibrium (LD) can hamper the short- and long-term selection response in recurrent selection (RS) programs. We analyzed LD in two European flint maize populations, KW1265 × D146 (A × B) and D145 × KW1292 (C × D), under modified recurrent full-sib selection. Our objectives were to investigate (1) the decay of initial parental LD present in F_2 populations by three generations of intermating, (2) the generation of new LD in four (A × B) and seven (C × D) selection cycles, and (3) the relationship between LD changes and estimates of the additive genetic variance. We analyzed the F_2 and the intermated populations as well as all selection cycles with 104 (A × B) and 101 (C × D) simple sequence repeat (SSR) markers with a uniform coverage of the entire maize genome. The LD coefficient D and the composite LD measure Δ were estimated and significance tests for LD were performed. LD was reduced by intermating as expected from theory. A directional generation of negative LD between favorable alleles could not be observed during the selection cycles. However, considerable undirectional changes in D were observed, which we attributed to genetic sampling due to the finite population size used for recombination. Consequently, a long-term reduction of the additive genetic variance due to negative LD was not observed. Our experimental results support the hypothesis that in practical RS programs with maize, LD generated by selection is not a limiting factor for obtaining a high selection response.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

⋆ Both authors contibuted equally to this work.

Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations

B. Stich[1], **A.E. Melchinger**[1], **H.-P. Piepho**[2], **S. Hamrit**[3], **W. Schipprack**[1], **H.P. Maurer**[1], **J.C. Reif**[1]

[1] *Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany*
[2] *Institute for Crop Production and Grassland Research, University of Hohenheim, 70593 Stuttgart, Germany*
[3] *Present address: Institute for Molecular Physiology and Biotechnology, University of Rostock, 18051 Rostock, Germany*

Received: 18 December 2006 **Accepted:** 31 May 2007 **Published online:** 28 June 2007
Communicated by G. Wenzel.

Abstract

Knowledge about the forces generating and conserving linkage disequilibrium (LD) is important for drawing conclusions about the prospects and limitations of association mapping. The objectives of our research were to examine the importance of (1) selection, (2) mutation, and (3) genetic drift for generating LD in a typical maize breeding program. We conducted computer simulations based on genotypic data of Central European maize open-pollinated varieties which have played an important role as founders of the European flint heterotic group. The breeding scheme and the dimensioning underlying our simulations reflect essentially the maize breeding program of the University of Hohenheim. Results suggested that in a plant breeding program of the examined dimension and breeding scheme, genetic drift and selection are major forces generating LD. The currently used population-based association mapping tests do not explicitly correct for LD caused by these two forces. Therefore, increased type I error rates are expected if these tests are applied to plant breeding populations. As a consequence, we recommend to use family-based association tests for association mapping approaches in plant breeding populations.

Corresponding author: Albrecht E. Melchinger, Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany. Email: melchinger@uni-hohenheim.de.

Chapter 9

General Discussion

Many population and quantitative genetic problems arising in the context of DNA marker applications in plant breeding programs cannot be answered by classical population genetical theory, which is based on simplifying assumptions to obtain mathematically tractable models. The most important assumptions are random mating, infinite population size, and an infinite number of genes with small effects underlying the phenotypic traits under consideration. However, these assumptions are often not fulfilled. Therefore, efficient use of molecular markers in plant breeding requires a software environment that integrates data analyses routines and simulation modules. In my thesis work, I developed the computer program Plabsoft, which integrates population genetic analyses routines for gene diversity, Hardy-Weinberg equilibrium (HWE), linkage disequilibrium (LD), and haplotype block finding algorithms with simulation tools allowing to model and analyze entire plant breeding programs. Application of Plabsoft in a large number of research projects, of which a considerable number would not have been possible without the software, corroborates the usefulness of this powerful and flexible software environment.

9.1 Hardy–Weinberg Equilibrium

The Hardy–Weinberg law is a cornerstone of diploid population genetics and the basis of many concepts in quantitative genetics (*c.f.*, Falconer and

Table 9.1. Tests for Hardy–Weinberg equilibium implemented in Plabsoft.

Procedure	Description
hwe.chi	Pearson's χ^2 test (Weir, 1996)
hwe.ce.louisdempster	Complete enumeration algorithm for up to four alleles per marker locus based on Fisher's exact test (Louis and Dempster, 1987)
hwe.ce	Complete enumeration algorithm based on Fisher's exact test (Weir, 1996)
hwe.ie	Incomplete enumeration algorithm based on Fisher's exact test by adopting the concept of Pagano and Taylor Halvorsen (1981)
hwe.network	Network algorithm based on Fisher's exact test (Aoki, 2003)
hwe.ie.network	Hybrid algorithm of network and incomplete enumeration algorithm for an arbitrary number of alleles per marker locus based on Fisher's exact test (Maurer et al., 2007)
hwe.monte.guo	Monte Carlo test based on Fisher's exact test (Guo and Thompson, 1992)
hwe.monte.huber	Monte Carlo test based on Fisher's exact test (Huber et al., 2006)

Mackay, 1996). For a correct interpretation of experimental results, it is, therefore, of crucial importance to check whether the assumptions of HWE are met or not. Consequently, I implemented a broad range of test statistics and testing strategies for HWE in Plabsoft (Table 9.1). Furthermore, a novel incomplete enumeration algorithm to test for HWE with multiple alleles was developed (Maurer et al., 2007) by combining the network algorithm (Aoki, 2003) with an adapted incomplete enumeration concept of Pagano and Taylor Halvorsen (1981).

To compare different algorithms for conducting exact HWE tests, the number of enumerated genotype count tables was employed, because this statistic is independent from the used programming techniques, which is not the case, for example, for the required computing time. The tests were applied to an experimental data set of maize (Reif et al., 2003a). For this dataset, the new incomplete enumeration algorithm required less enumerations than all algorithms described previously. Our results indicate that asymptotic goodness-of-fit tests are not necessary with todays computing resources. For an exact HWE test the newly developed algorithm can be carried out directly or its p-value can be estimated with Monte Carlo methods. In conclusion, the newly developed algorithm has considerable computational advantages over algorithms described previously and extends substantially the range of problems that can be tested.

9.2 Linkage Disequilibrium

Knowledge about the patterns and distribution of LD in breeding populations is required to investigate the additive and dominance genetic variances of quantitative traits and for applying association mapping methods (Lynch and Walsh, 1997). Therefore, I implemented in Plabsoft a broad range of LD measures and tests (Table 9.2).

Table 9.2. Routines for analyzing linkage disequilibrium implemented in Plabsoft.

Procedure	Description
ld.pw.gametic.chi	Pearson's χ^2 test for gametic data based on linkage disequilibrium coefficient D (Weir, 1996)
ld.pw.genotypic.chi	Pearson's χ^2 test for genotypic data based on the composite linkage disequilibrium coefficient Δ (Weir, 1996)
ld.pw.gametic.monte	Monte Carlo test for gametic data based on Fisher's exact test (Weir, 1996)
ld.pw.genotypic.exact	Complete enumeration algorithm for genotypic data based on Fisher's exact test
ld.pw.genotypic.monte	Monte Carlo algorithm for genotypic data based on Fisher's exact test with or without assuming Hardy–Weinberg equilibrium (Zaykin et al., 1995). In the case of homozygous inbred lines, the LD measures (D^2, D', r^2, D'_m, and R) and p-values for the LD measures D'_m and R are calculated via Monte Carlo methods
ld.multi.genotypic.monte	Monte Carlo algorithm for multilocus genotypic data based on Fisher's exact test
ld.matrix.plot	Plot the results (Δ^2, D^2, D', r^2, D'_m, R, or p-values of D'_m, R, and Fisher's exact test) of a linkage disequilibrium analysis in a matrix plot
ld.versus.distance	Plot the results (Δ^2, D^2, D', r^2, D'_m, R, or p-values of D'_m, R, and Fisher's exact test) of a linkage disequilibrium analysis as a function of recombination frequency or map distance

The suitability of these alternative measures for assessing LD in plant breeding populations was not yet investigated. In the case of bi-allelic loci, D' measures only recombinational history, while r^2 summarizes both recombinational and mutational history (Flint-Garcia et al., 2003). The analysis of our dataset from a hybrid maize breeding program (Maurer et al., 2006) indicated one main drawback of D'_m. For small sample sizes or rare alleles, the estimates of D'_m can be highly inflated. r^2 can be interpreted as the squared correlation coefficient between allele frequencies at two bi-allelic loci (Devlin and Risch, 1995) and is proportional to Pearson's χ^2 test statistic to detect LD (Balding, 2006). Hence, a low value of r^2 means that a large sample size n is required to detect the LD between the markers. The results of our maize data set based on multi-allelic simple sequence repeat (SSR) markers (Maurer et al., 2007) indicate that the values of D'_m and R strongly depended on the sample size, the number of alleles, and the allele frequency distribution at each locus. In consequence, comparisons between D'_m and R values of the same loci in different populations are problematic.

Due to the difficulty in comparing absolute values of D'_m and R between different populations, it is often more important to know whether the observed LD is statistically significant rather than how large the disequilibrium coefficients are (Weir, 1996). To test for statistical significance of D'_m and R, an algorithm was developed and implemented, which is analogous to the estimation of the p-value with Fisher's exact test via Monte Carlo methods. The algorithm determines with a permutation method the probability p of occurrence of contingency tables with larger D'_m or R values and the same marginal totals as the observed one. In conclusion, Plabsoft is the first software that provides significance tests for the most important LD measures.

9.3 Haplotype Blocks

Strong LD between marker loci was detected in populations from an applied maize breeding program, which results in correlations between alleles

Table 9.3. Routines for analyzing haplotype blocks implemented in Plabsoft.

Procedure	Description
haploblock.fixed	Determines haplotype blocks by a fixed block length of n adjacent and consecutive marker loci (HB1; Schrag et al., 2007)
haploblock.find	Determines an optimum haplotype block solution with the lowest chromosome-wise haplotype allele number using pairwise and multi-allelic LD measures (HB2 and HB3; Schrag et al., 2007)
haploblock.statistic	Returns a summary statistic about the haplotype block partition
haploblock.boundary.add	Adds a haplotype block boundary between two adjacent loci on the same chromosome
haploblock.boundary.remove	Removes a haplotype block boundary between two adjacent loci on the same chromosome
ggt.plot	Plots the graphical genotype of the individuals of a population. Blocks are visualized via vertical lines. Different coloring modes are available: locus / haplotype block

at two loci (Maurer et al., 2006). These correlations can further result in an overestimation of quantitative trait loci (QTL) effects in association studies (Wang et al., 2003), inaccurate prediction of hybrid performance, and in problems to estimate the parameters in multiple linear regression (Schrag et al., 2006). To address the multicolinearity of marker alleles, I developed three algorithms for detecting haplotype blocks in breeding material and implemented these in Plabsoft (Table 9.3).

The algorithm HB1 uses a fixed block size of four markers as suggested by Jansen et al. (2003). The data-driven approaches HB2 and HB3 determine the block boundaries by minimizing the number of haplotype alleles for each chromosome with alternative constraints on the LD within blocks (Schrag et al., 2007) using Dijkstra's (1959) shortest path algorithm.

The haplotype block finding algorithms were applied to predict heterosis and hybrid performance in four factorial crossing designs of a hybrid maize breeding program (Schrag et al., 2007). We found that missing genotype data is a major problem for haplotype block analysis. In future studies, this problem could be tackled by replacing missing genotypes with predicted values that are based on the observed genotypes, which results in a better use of the observed data and simplified analysis (Balding, 2006). The haplotype blocks resulted in an improved prediction of hybrid performance compared with predictions based on single-marker methods, especially when the data-driven algorithms HB2 and HB3 were used (Schrag et al., 2007).

In conclusion, the haplotype finding algorithms were successfully employed in prediction of heterosis in maize hybrid breeding. Further promising applications, where haplotype blocks can replace single-marker models, are association mapping approaches in crops (Mackay and Powell, 2007).

9.4 Simulation of plant breeding programs

The designs and selection decisions in plant breeding programs are complex and can be analyzed only under simplifying assumptions with closed math-

ematical models. Often, computer simulations are the only possibility to model plant breeding programs. Therefore, I developed and implemented a comprehensive simulation module in Plabsoft (Maurer et al., 2008).

The core of the genetic simulation is an algorithm which allows to model the recombination process during meiosis. The algorithm implemented in Plabsoft is based on the count location process (Karlin and Liberman, 1978). From genotypes generated by the simulation of meiosis, genotypic values are obtained using a quantitative genetic model suggested by Bulmer (1985). Phenotypic values can be obtained by modeling masking variances and the error variance for arbitrary experimental designs. In addition to meiosis of diploid species (Maurer et al., 2008), I developed a module for simulation of meiosis for autotetraploid crops such as potato, adopting the meiosis model without double reduction as described by Gallais (2003). An illustration of the principle employed for simulation of meiosis is given in Figure 9.1.

Using the above core functionality, I developed a broad range of simulation functions (Maurer et al., 2008, Table 1). These functions for the simulation of individual breeding steps can be used to model entire breeding programs. This software design provides a high level of flexibility for the advanced user. To make the software also applicable for non-expert users, I designed and implemented an alternative set of high-level interfaces for Plabsoft (Table 9.4). These high-level routines allow to describe entire breeding programs, which follow standardized designs, with only a few lines of program statements without being too much involved with details of the software. The modules are employed in ongoing studies investigating marker-assisted backcrossing programs (Prigge et al., in preparation).

A key question in the application of simulation studies is “How good do the simulations reproduce reality?”. We investigated this question with a dataset of a marker-assisted backcrossing program for introgression of the submergence tolerance QTL *Sub1* (Xu et al., 2006) into the widely grown rice

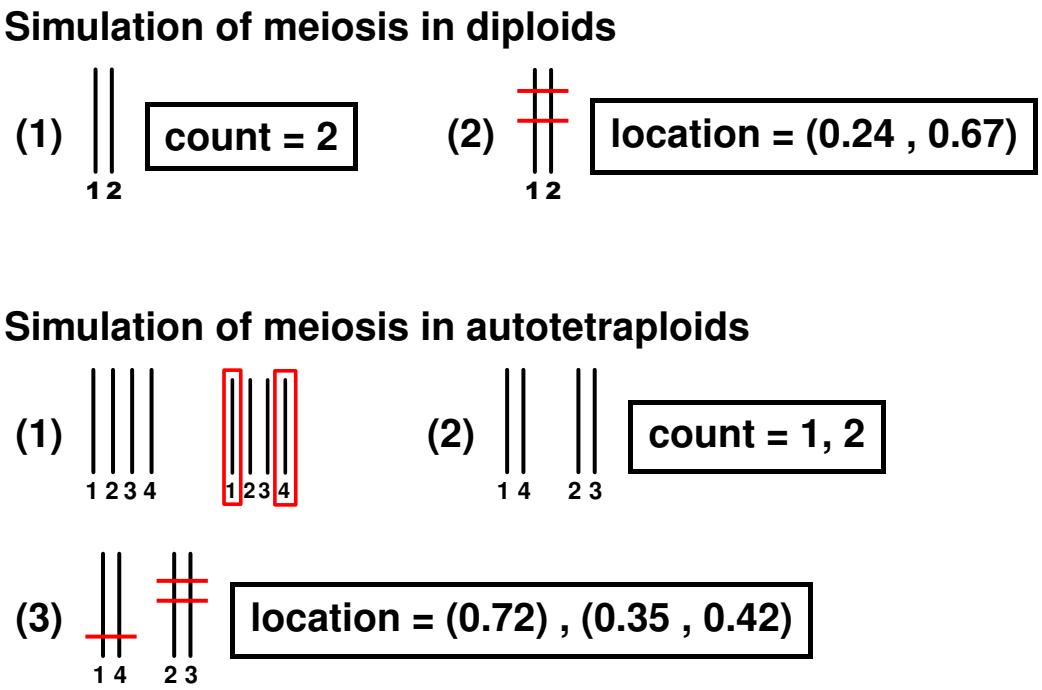

Figure 9.1. Illustration of simulation of recombination during meisosis in dipoid and tetraploid species. In the count location model employed for diploid species, a two-step random process is employed. In the first step, a Poisson distributed random variable determines the count of crossovers on a chromosome. In the second step, the position of each crossover on the chromosome is determined with a uniformly distributed random variable. The figure illustrates this with the example of a chromsome on which two crossovers at 0.24 M and 0.67 M from the telomere occurred. In tetraploid species, a further random process determines which homologous chromosomes are pairing for the recombination process. In the above example, chromosome 1 is paring with chromosome 4 and chromosome 2 with chromosome 3. Subsequently, for each pair of chromosomes the count-location process is modeled.

Table 9.4. High-level interface function to simulate marker-assisted backcrossing programs following standardized designs.

Procedure	Description
genotype.construction.begin	Begin of a simulation run
genotype.construction.end	End of a simulation run
genotype.construction.statistic	Summary statistic about successful and failed simulation runs
population.select	Function for conducting selection experiments with an arbitrary number of stages and selection strategies. A selection strategy is defined by the arguments batch, stage, condition, effect, strategy, x, and MDP.count
watch.batch	Collects information about marker-assisted backcrossing programs with increasing population size up to a point where the simulation run can be successfully continued
watch.gc	Collects information about the genome contribution of a specified ancestor to a population
watch.marker.score	Collects information about the marker score for a specified effect
watch.mdp	Collects information about the required number of marker data points

For each of the four implemented watch functions, further functions are implemented to visualize (watch.plot), compute a summary statistic (watch.statistic), return (watch.get), or save/load from a file the collected information.

lines 'Swarna' and 'Samba Mashuri' (Prigge et al., 2007). A high agreement with simulated and observed proportions of the recurrent parent genome in backcross individuals was observed. Therefore, we concluded that the results obtained with Plabsoft are highly transferable to applied plant breeding programs.

9.5 Application of Plabsoft

Applications of Plabsoft for investigating (a) LD in recurrent selection program in maize (Falke et al., 2007) and (b) the development of LD in the breeding history of maize (Stich et al., 2007b) were used to define the functionality required for modeling applied breeding programs and the requirements of the users with respect to the interface and function calls of the simulation and data analyses routines.

Falke et al. (2007) simulated with Plabsoft three generations of intermating in a recurrent selection program of maize. We carried out an integrated analysis, in which we analyzed genetic distances and LD of simulated and experimental data with Plabsoft. With the simulations, we were able to determine confidence intervals for the expected LD decay under intermating with a chain crossing scheme, a research question for which no theoretical concepts are available. This application demonstrates that Plabsoft can be employed to solve complex population genetical problems for which analytical solutions are not available.

Stich et al. (2007b) simulated the breeding history of the hybrid maize breeding program of the University of Hohenheim. The simulations were based on experimental data of European open pollinated varieties, from which the inbred lines were developed for establishing the hybrid breeding program. Subsequently, eight breeding cycles were simulated, comprising inbred line development, multi-stage selection for general combining ability, and crosses between superior inbred lines. With an integrated analysis,

employing the simulation routines of Plabsoft and the routines for the assessment of LD, we found that random genetic drift and selection are the major forces generating LD. Consequently, association studies in plant breeding programs are best carried out with family-based association tests (Stich et al., 2006b). This application demonstrates that the developed simulation functionality is capable of modeling complex, long-term breeding programs as carried out in practice.

In addition to the publications of this thesis, the development of Plabsoft was further promoted by its use in eight studies, on which I was involved as a co-author. Reif et al. (2005d, e) used the data analyses routines to investigate population genetical parameters in maize. Heckenberger et al. (2005c) simulated breeding schemes to derive concepts for plant variety protection. Stich et al. (2005, 2006a) analyzed LD in European maize to assess the prospects of association mapping in plant breeding. Stich et al. (2006b) conducted simulation studies to evaluate the power of a new family-based association test. Heckenberger et al. (2008) developed a database interfaced for Plabsoft. Stich et al. (2007a) investigated with simulations the power and proportion of false positives in association mapping.

The software was also used in about thirty studies, in which I did not participate, but which underline its broadly usability by applied geneticists and breeders. Plabsoft was used to assess population genetic parameters in maize (Andersen et al., 2005; Reif et al., 2003a, b, 2004, 2005c, 2006; Xia et al., 2004, 2005), triticale (Tams et al., 2005, 2006), celeriac, cornsalad, and radish (Muminović et al., 2004a, b, 2005), and wheat (Dreisigacker et al., 2004, 2005a, b; Reif et al., 2005a; Zhang et al., 2005, 2006). Reif et al. (2005b) conducted simulations to compare genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Heckenberger et al. (2005a, b, 2006) conducted genetic diversity studies in the context of plant variety protection. Frisch et al. (2004) used simulations to investigate the effects of incorrect locus orders caused by duplicate marker loci. Frisch and Melchinger (2006) used

simulations to investigate the extent of systematic prediction errors in the context of marker-based prediction of the parental genome contribution to inbred lines derived from biparental crosses. Frisch and Melchinger (2007) used simulations to verify theoretically derived formulas for the variance of the parental genome contribution to inbred lines derived from biparental crosses.

9.6 Prospects for future developments

Unraveling of the genetic architecture of complex agronomically important traits is very promising through the integration of genomics, metabolomics, and phenomics (Keurentjes et al., 2006). The integration of such information into applied breeding programs requires a paradigm shift from phenotype- to genotype-based selection (Bernardo, 2001; Peleman and Rouppe van der Voort, 2003). To realize this shift, new breeding concepts and strategies are required. These can be investigated and developed with Plabsoft.

In particular, application of Plabsoft is promising in the following areas: (i) investigations of the statistical properties of new QTL mapping methods, (ii) optimizations of breeding programs for an optimal use of the available information about epistatic and genotype × environment interactions (Walsh, 2005), and (iii) fine-tuning of existing breeding strategies for an optimum use of resources in phenotype- and molecular-based selection approaches (Dekkers and Hospital, 2002).

References

Andersen, J.R., T. Schrag, A.E. Melchinger, I. Zein, and T. Lübberstedt. 2005. Validation of *Dwarf8* polymorphisms associated with flowering time in elite European inbred lines of maize (*Zea mays* L.). Theor. Appl. Genet. 111:206-217.

Aoki, S. 2003. Network algorithm for the exact test of Hardy–Weinberg proportion for multiple alleles. Biometric. J. 4:471–490.

Balding, D.J. 2006. A tutorial on statistical methods for population association studies. Nat. Rev. Genet. 7:781–791.

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

Bulmer, M.G. 1985. The mathematical theory of quantitative genetics. p. 46. Oxford: Oxford University Press.

Dekkers, J.C.M., and F. Hospital. 2002. The use of molecular genetics in the improvement of agricultural populations. Nat. Rev. Genet. 3:22–32.

Devlin, B., and N. Risch. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322.

Dijkstra, E.W. 1959. A note on two problems in connexion with graphs. Numer. Math. 1:269–271.

Dreisigacker, S., P. Zhang, M.L. Warburton, M. Van Ginkel, D. Hoisington, M. Bohn, and A.E. Melchinger. 2004. SSR and pedigree analyses of genetic diversity among CIMMYT wheat lines targeted to different megaenvironments. Crop Sci. 44:381–388.

Dreisigacker, S., A.E. Melchinger, P. Zhang, K. Ammar, C. Flachenecker, D. Hoisington, and M.L. Warburton. 2005a. Hybrid performance and heterosis in spring bread wheat, and their relations to SSR-based genetic distances and coefficients of parentage. Euphytica 144:51-59.

Dreisigacker, S., P. Zhang, M.L. Warburton, B. Skovmand, D. Hoisington, and A.E. Melchinger. 2005b. Genetic diversity among and within CIMMYT wheat landrace accessions investigated with SSRs and implications for plant genetic resources management. Crop Sci. 45:653–661.

Falconer, D.S., and T.F.C. Mackay. 1996. Introduction to quantitative genetics. 4th ed. Longman Group Ltd., London.

Falke, K.C., H.P. Maurer, A.E. Melchinger, H.-P. Piepho, C. Flachenecker, and M. Frisch. 2007. Linkage disequilibrium in two European F_2 flint maize populations under modified recurrent full-sib selection. Theor. Appl. Genet. 115:289–297. Erratum: Theor. Appl. Genet. 115:299.

Flint-Garcia, S.A., J.M. Thornsberry, and E.S. Buckler. 2003. Structure of linkage disequilibrium in plants. Ann. Rev. Plant Biol. 54:357–374.

Frisch, M., and A.E. Melchinger. 2006. Marker-based prediction of the parental genome contribution to inbred lines derived from biparental crosses. Genetics 174:795–803.

Frisch, M., and A.E. Melchinger. 2007. Variance of the parental genome contribution to inbred lines derived from biparental crosses. Genetics 176:477–488.

Frisch, M., M. Quint, T. Lübberstedt, and A.E. Melchinger. 2004. Duplicate marker loci can result in incorrect locus orders on linkage maps. Theor. Appl. Genet 109:305–316.

Gallais, A. 2003. Quantitative genetics and breeding methods in autopolyploid plants. p. 38. INRA, Paris.

Guo, S.W., and E.A. Thompson. 1992. Performing the exact test of Hardy–Weinberg proportion for multiple alleles. Biometrics 48:361–372.

Heckenberger, M., M. Bohn, and A.E. Melchinger. 2005a. Identification of essentially derived varieties obtained from biparental crosses of homozygous lines: I. Simple sequence repeat data from maize inbreds. Crop Sci. 45:1120–1131.

Heckenberger, M., M. Bohn, D. Klein, and A.E. Melchinger. 2005b. Identification of essentially derived varieties obtained from biparental crosses of homozygous lines: II. Morphological distances and heterosis in comparison with simple sequence repeat and amplified fragment length polymorphism data in maize. Crop Sci. 45:1132–1140.

Heckenberger, M., M. Bohn, M. Frisch, H.P. Maurer, and A.E. Melchinger. 2005c. Identification of essentially derived varieties with molecular markers: An approach based on statistical test theory and computer simulations. Theor. Appl. Genet. 111:598-608.

Heckenberger, M., J. Muminović, J. Rouppe van der Voort, J. Peleman, M. Bohn, and A.E. Melchinger. 2006. Identification of essentially derived varieties obtained from biparental crosses of homozygous lines. III. AFLP data from maize inbreds and comparison with SSR data. Mol. Breeding 17:111–125.

Heckenberger, M., H.P. Maurer, A.E. Melchinger, and M. Frisch. 2008. The Plabsoft database: a comprehensive database management system for integrating phenotypic and genomic data in academic and commercial plant breeding programs. Euphytica 161:173–179.

Huber, M., Y. Chen, I. Dinwoodie, A. Dobra, and M. Nicholas. 2006. Monte Carlo algorithms for Hardy–Weinberg proportions. Biometrics 62:49–53.

Jansen, R.C., J.L. Jannink, and W.D. Beavis. 2003. Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. Crop Sci. 43:829–834.

Karlin, S., and U. Liberman. 1978. Classifications and comparisons of multilocus recombination distributions. Proc. Natl. Acad. Sci. USA 75:6332–6336.

Keurentjes, J.J.B., J. Fu, C.H.R. de Vos, A. Lommen, R.D. Hall, R.J. Bino, L.H.W. van der Plas, R.C. Jansen, D. Vreugdenhil, and M. Koornneef.

2006. The genetics of plant metabolism. Nature Genet. 38:842–849.

Louis, E.J., and E.R. Dempster. 1987. An exact test for Hardy–Weinberg and multiple alleles. Biometrics 43:805–811.

Lynch, M., and B. Walsh. 1997. Genetics and analysis of Quantitative Traits. p. 413. Sinauer Assoc., Inc., Sunderland, MA.

Mackay, I., and W. Powell. 2007. Methods for linkage disequilibrium mapping in crops. Trends Plant Sci. 12:57–63.

Maurer, H.P., C. Knaak, A.E. Melchinger, M. Ouzunova, and M. Frisch. 2006. Linkage disequilibrium between SSR markers in six pools of elite lines of an European breeding program for hybrid maize. Maydica 51:269–279.

Maurer, H.P., A.E. Melchinger, and M. Frisch. 2007. An incomplete enumeration algorithm for an exact test of Hardy–Weinberg proportions with multiple alleles. Theor. Appl. Genet. 115:393–398.

Maurer, H.P., A.E. Melchinger, and M. Frisch. 2008. Population genetic simulation and data analysis with Plabsoft. Euphytica 161:133–139.

Muminović, J., A.E. Melchinger, and T. Lübberstedt. 2004a. Prospects for celeriac (*Apium graveolens* var. *rapaceum*) improvement by using genetic resources of *Apium*, as determined by AFLP markers and morphological characterization. Plant Genetic Resources: Characterization and Utilization 2:189–198.

Muminović, J., A.E. Melchinger, and T. Lübberstedt. 2004b. Genetic diversity in cornsalad (*Valerianella locusta*) and related species as determined by AFLP markers. Plant Breeding 123:460-466.

Muminović, J., A. Merz, A.E. Melchinger, and T. Lübberstedt. 2005. Genetic structure and diversity among radish varieties as inferred from AFLP and ISSR analyses. J. Am. Soc. Hortic. Sci. 130:79–87.

Pagano, M., and K. Taylor Halvorsen. 1981. An algorithm for finding the exact significance levels of $r \times c$ contingency tables. J. Am. Stat. Assoc. 76:931–934.

Peleman, J. D., and J. Rouppe van der Voort. 2003. Breeding by design. Trends Plant Sci. 8: 330–334.

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.

Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, G. Srinivasan, M. Bohn, and M. Frisch. 2003a. Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci. 43:1275–1282.

Reif, J.C., A.E. Melchinger, X.C. Xia, M.L. Warburton, D.A. Hoisington, S.K. Vasal, D. Beck, M. Bohn, and M. Frisch. 2003b. Use of SSRs for establishing heterotic groups in subtropical maize. Theor. Appl. Genet. 107:947-957.

Reif, J.C., X.C. Xia, A.E. Melchinger, M.L. Warburton, D.A. Hoisington, D. Beck, M. Bohn, and M. Frisch. 2004. Genetic diversity determined within and among CIMMYT maize populations of tropical, subtropical, and temperate germplasm by SSR markers. Crop Sci. 44:326–334.

Reif, J.C., P. Zhang, S. Dreisigacker, M.L. Warburton, M. Van Ginkel, D. Hoisington, M. Bohn, and A.E. Melchinger. 2005a. Wheat genetic diversity trends during domestication and breeding. Theor. Appl. Genet. 110:859–864.

Reif, J.C., A.E. Melchinger, and M. Frisch. 2005b. Genetical and mathematical properties of similarity and dissimilarity coefficients applied in plant breeding and seed bank management. Crop Sci. 45:1–7.

Reif, J.C., A.R. Hallauer, and A.E. Melchinger. 2005c. Heterosis and heterotic pattern in maize. Maydica 50:215–223.

Reif, J.C., S. Hamrit, M. Heckenberger, W. Schipprack, H.P. Maurer, M. Bohn, and A.E. Melchinger. 2005d. Temporal trend of genetic diversity in European maize germplasm. Theor. Appl. Genet. 111:838–845.

Reif, J.C., S. Hamrit, M. Heckenberger, W. Schipprack, H.P. Maurer, M. Bohn, and A.E. Melchinger. 2005e. Genetic structure and diversity of European flint maize populations determined with SSR analyses of individuals and bulks. Theor. Appl. Genet. 111:906–913.

Reif, J.C., S. Hamrit, A.E. Melchinger. 2006. Genetic diversity trends in Central European heterotic groups. Acta Agron. Hung. 54:315-320.

Schrag, T.A., A.E. Melchinger, A.P. Sørensen, and M. Frisch. 2006. Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize using AFLP markers associated with QTL. Theor. Appl. Genet. 113:1037–1047.

Schrag, T.A., H.P. Maurer, A.E. Melchinger, H.-P. Piepho, J. Peleman, and M. Frisch. 2007. Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. Theor. Appl. Genet. 114:1345–1355.

Stich, B., A.E. Melchinger, M. Frisch, H.P. Maurer, M. Heckenberger, and J.C. Reif. 2005. Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor. Appl. Genet. 111:723–730.

Stich, B., H.P. Maurer, A.E. Melchinger, M. Frisch, M. Heckenberger, J. Rouppe van der Voort, J. Peleman, A.P. Sørensen, and J.C. Reif. 2006a. Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol. Breed. 17:217–226.

Stich, B., A.E. Melchinger, H.-P. Piepho, M. Heckenberger, H.P. Maurer, and J.C. Reif. 2006b. A new test for family-based association mapping

with inbred lines from plant breeding programs. Theor. Appl. Genet. 113:1121–1130.

Stich, B., J. Yu, A.E. Melchinger, H.-P. Piepho, H.F. Utz, H.P. Maurer, and E.S. Buckler. 2007a. Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy. Genetics 176:563–570.

Stich, B., A.E. Melchinger, H.-P. Piepho, S. Hamrit, W. Schipprack, H.P. Maurer, and J.C. Reif. 2007b. Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor. Appl. Genet. 115:529–536.

Tams, S.H., A.E. Melchinger, and E. Bauer. 2005. Genetic similarity among European winter triticale elite germplasms assessed with AFLP and comparisons with SSR and pedigree data. Plant Breeding 124:154–160.

Tams, S.H., E. Bauer, G. Oettler, A.E. Melchinger, C.-C. Schön. 2006. Prospects for hybrid breeding in winter triticale: II. Relationship between parental genetic distance and specific combining ability. Plant Breeding 125:331–336.

Walsh, B. 2005. The struggle to exploit non-additive variation. Aust. J. Agric. Res. 56:873–881.

Wang, W.Y.S., H.J. Cordell, and J.A. Todd. 2003. Association mapping of complex diseases in linked regions: Estimation of genetic effects and feasibility of testing rare variants. Genet. Epidemiol. 24:36–43.

Weir, B.S. 1996. Genetic data analysis II, 2nd edition. p. 91. Sunderland, Massachusetts: Sinauer Associates.

Xia, X.C., J.C. Reif, D.A. Hoisington, A.E. Melchinger, M. Frisch, and M.L. Warburton. 2004. Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: I. Lowland tropical maize Crop Sci. 44:2230-2237.

Xia, X.C., J.C. Reif, A.E. Melchinger, M. Frisch, D.A. Hoisington, D. Beck, K. Pixley, M.L. Warburton. 2005. Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: II. Subtropical, tropical midaltitude, and highland maize inbred lines and their relationships with elite U.S. and European maize. Crop Sci. 45:2573–2582.

Xu, K., X. Xu, T. Fukao, P. Canlas, R. Maghirang-Rodriguez, S. Heuer, A.M. Ismail, J. Bailey-Serres, P.C. Ronald, and D.J. Mackill. 2006. *SUB1A* is an ethylene-response-factor-like gene that confers submergence tolerance to rice. Nature 442:705–708.

Zaykin, D., L. Zhivotovsky, and B.S. Weir. 1995. Exact tests for association between alleles at arbitrary numbers of loci. Genetica 96:169–178.

Zhang, P., S. Dreisigacker, A.E. Melchinger, J.C. Reif, A. Mujeeb Kazi, M. Van Ginkel, D. Hoisington, and M.L. Warburton. 2005. Quantifying novel sequence variation and selective advantage in synthetic hexaploid wheats and their backcross-derived lines using SSR markers. Mol. Breeding 15:1–10.

Zhang, P., S. Dreisigacker, A. Buerkert, S. Alkhanjari, A.E. Melchinger, and M.L. Warburton. 2006. Genetic diversity and relationships of wheat landraces from Oman investigated with SSR markers. Genet. Resour. Crop. Ev. 53:1351–1360.

Chapter 10

Summary

Marker-assisted breeding approaches are promising tools for enhancement of the conventional plant breeding process. They have been successfully applied in many areas such as plant variety protection, classification of germplasm, assessment of genetic diversity, mapping of genes underlying important agronomic traits, and using the mapping information for selection decisions. Powerful and flexible bioinformatic tools are urgently required for a better integration of molecular marker applications and classical plant breeding methods. The objective of my thesis work was to develop and apply Plabsoft, a computer program for population genetic data analyses and simulations in plant breeding.

The assumption of Hardy–Weinberg equilibrium is a cornerstone of many concepts in population and quantitative genetics. Therefore, tests for Hardy–Weinberg equilibrium are of crucial importance, but the assumptions underlying asymptotic χ^2 tests are often not met in datasets from plant breeding programs. I developed and implemented in Plabsoft a new algorithm for exact tests of Hardy–Weinberg equilibrium with multiple alleles. The newly derived algorithm has considerable computational advantages over previously described algorithms and extends substantially the range of problems that can be tested.

Knowledge about the amount and distribution of linkage disequilibrium (LD) in breeding populations is of fundamental importance to assess

the prospects for gene mapping with whole-genome association studies. To analyze LD in breeding populations, I implemented various LD measures in Plabsoft and developed a new significance test for the LD measures D'_m and R. The routines were employed to analyze LD in 497 elite maize lines from a commercial hybrid breeding program, which were fingerprinted by 81 simple sequence repeat (SSR) markers covering the entire genome. Strong LD was detected and, therefore, whole-genome association studies were recommended as promising. However, LD between unlinked loci will most likely result in a high rate of false positives.

The prediction of hybrid performance with DNA markers facilitates the identification of superior hybrids. The single marker models used so far do not take into account the correlation between allele frequencies at linked markers. To overcome this problem, the concept of haplotype blocks was proposed. I developed and implemented in Plabsoft three alternative algorithms for haplotype block detection suitable for plant breeding. The algorithms were applied for the haplotype-based prediction of the hybrid performance of 270 hybrids, the parents of which were fingerprinted with 20 amplified fragment length polymorphism (AFLP) primer combinations. Employing haplotypes resulted in an improved prediction of hybrid performance compared with single marker models. Consequently, haplotype-based prediction methods have a high potential to improve substantially the efficiency of hybrid breeding programs.

Computer simulations can be employed to solve population genetic problems in plant breeding, for which the simplifying assumptions underlying the classical population genetic theory do not hold true. However, before the start of my thesis no flexible simulation software was available. I developed algorithms for simulation of single breeding steps and entire plant breeding programs and implemented these in Plabsoft. The routines allow the simulation of plant breeding programs as they are conducted in practice.

The simulation routines of Plabsoft were validated by simulating two marker-assisted backcross programs in rice conducted by the International

Rice Research Institute (IRRI). In the simulations, the frequency distributions of the proportion of recurrent parent genome in the backcross populations were assessed. The simulation results were in good agreement with the experimental data. Therefore, computer simulations are a useful tool for pre-test estimation of selection response in marker-assisted backcrossing.

The application of Plabsoft was exemplified by two studies in maize. In the first study, the expected LD decay in the intermating generations of two recurrent selections programs was determined with simulations. This application demonstrates the use of Plabsoft to solve problems for which analytical results are not available. In the second study, the forces generating and maintaining LD in a hybrid maize breeding program were investigated with computer simulations. This application demonstrates the capability of modeling complex long-term breeding programs as performed in practice.

The studies of my thesis provide an example for the broad range of possible applications of Plabsoft. In addition to the presented studies, Plabsoft has so far been employed in about 40 further studies, which corroborates the usefulness of Plabsoft for integrating new genomic tools in applied plant breeding programs.

Chapter 11

Zusammenfassung

DNA Marker werden in der Pflanzenzüchtung zum Erkennen von Sortenplagiaten, zur Gruppierung von Zuchtmaterial, zur Überwachung der genetischen Diversität, zur Kartierung von Genen, die für die Ausprägung wichtiger agronomischer Merkmale verantwortlich sind, sowie zur marker-gestützten Selektion eingesetzt. Um die Markertechnologie in die Methodik der klassischen Pflanzenzüchtung zu integrieren, werden dringend flexible und leistungsfähige bioinformatische Konzepte und darauf basierende Computerprogramme benötigt. Das Ziel dieser Arbeit war es, Plabsoft, ein Computerprogramm zur populationsgenetischen Datenanalyse und Simulation von Pflanzenzüchtungsprogrammen, zu entwickeln und anzuwenden.

Die Annahme, dass sich eine Population im Hardy–Weinberg Gleichgewicht befindet, liegt vielen Konzepten der Populationsgenetik und der quantitativen Genetik zugrunde. Deswegen sind statistische Tests auf Hardy–Weinberg Gleichgewicht von großer Bedeutung. In Datensätzen aus Pflanzenzüchtungsprogrammen treffen die statistischen Annahmen, welche den oft verwendeten χ^2-Tests zugrunde liegen, häufig nicht zu. Aus diesem Grund wurde in dieser Arbeit ein neuer Algorithmus für einen exakten Test auf Hardy–Weinberg Gleichgewicht mit multiplen Allelen entwickelt und in Plabsoft umgesetzt. Der neu implementierte Algorithmus ist deutlich schneller als alle vorher beschriebenen Algorithmen und erlaubt somit eine bedeutende Erweiterung für den Anwendungsbereich exakter Hardy–Weinberg Tests.

Die genaue Kenntnis der Höhe und Verteilung von Gametenphasenungleichgewicht (linkage disequilibrium, LD) in pflanzenzüchterischen Populationen ist von großer Bedeutung, um die Erfolgsaussichten genomweiter Assoziationsstudien abschätzen zu können. Zu diesem Zweck wurde die Berechnung der wichtigsten LD Maße in Plabsoft implementiert und ein neuer Signifikanztest für die Maße D'_m und R entwickelt. Die neu entwickelten Routinen wurden zur Analyse des LD in einem kommerziellen Hybridmaiszüchtungsprogramm verwendet. Hierzu wurden 497 Inzuchtlinien mit 81 SSR (simple sequence repeat, Mikrosatelliten) Markern genotypisiert und ein hohes Ausmaß an LD detektiert, so daß genomweite Assoziationskartierungsansätze vielversprechend erscheinen. Jedoch ist zu erwarten, dass aufgrund des hohen Ausmaßes an LD zwischen ungekoppelten Markerloci viele falsch positive Assoziationen beobachtet werden.

Eine markergestützte Vorhersage der Hybridleistung vereinfacht die Identifizierung überlegener Kreuzungskombinationen. Bisher wurden hierfür nur Vorhersagemodelle verwendet, die auf einzelnen Markerloci basieren und die Korrelationsstruktur zwischen Allelen an benachbarten Markerloci nicht berücksichtigen. In der Humangenetik wurde vorgeschlagen, benachbarte Markerloci zu sogenannten Haploblöcken zusammenzufassen, um das Problem der Multikolinearität zu lösen. Im Rahmen dieser Arbeit wurden drei unterschiedliche Algorithmen zur Detektion von Haploblöcken im Zuchtmaterial erarbeitet und in Plabsoft umgesetzt. Die Routinen wurden für eine haplotyp-basierte Vorhersage der Leistung von 270 Hybriden verwendet, deren Eltern mit 20 AFLP (amplified fragment length polymorphism) Primerkombinationen untersucht wurden. Die Vorhersage der Hybridleistung konnte durch die Verwendung von Haploblöcken verbessert werden. Folglich haben haplotyp-basierte Vorhersagemethoden ein großes Potential, die Effizienz von Hybridzuchtprogrammen zu steigern.

Computersimulationen können in der Pflanzenzüchtung zur Lösung populationsgenetischer Fragestellungen auch dann angewendet werden, wenn die Annahmen, welche der klassischen populationsgenetischen Theorie zugrunde liegen, nicht erfüllt sind. Vor Beginn dieser Arbeit stand je-

doch keine Software zur Verfügung, welche auf flexible Art und Weise Simulationen pflanzenzüchterischer Fragestellungen ermöglicht hätte. Aus diesem Grund wurden Algorithmen entwickelt, die die Simulation einzelner Züchtungsschritte sowie kompletter Pflanzenzüchtungsprogramme ermöglichen. Die entwickelten Algorithmen wurden im Computerprogramm Plabsoft umgesetzt, so dass es jetzt möglich ist, komplexe Pflanzenzüchtungsprogramme praxisnah zu simulieren.

Die Simulationsroutinen von Plabsoft wurden an einem experimentellen Datensatz zur markergestützten Introgression eines Überflutungstoleranzgens in Reis validiert. Hierzu wurde das gesamte Zuchtprogramm, wie es in der Praxis durchgeführt wurde, simuliert. In den Simulationen wurde die Häufigkeitsverteilung des rekurrenten Eltergenomanteils in den Rückkreuzungspopulationen erfasst. Die Simulationsergebnisse stimmten nahezu vollständig mit den experimentell beobachteten Daten überein. Dies belegt, dass Computersimulationen ein äußerst effektives Hilfsmittel sind, um den Selektionserfolg bei der markergestützten Rückkreuzung abzuschätzen.

Die Anwendung der Simulations- und Analysesoftware Plabsoft wurde exemplarisch an zwei Studien dargestellt. In der ersten Studie wurde mit Hilfe von Simulationen der zu erwartende Abfall an LD in den Durchkreuzungsgenerationen bei zwei rekurrenten Selektionsprogrammen in Mais bestimmt. Diese Studie demonstriert die Anwendung von Plabsoft zur Lösung von Fragestellungen, für welche keine analytische Lösung zur Verfügung stehen. In der zweiten Studie wurden mit Hilfe von Computersimulationen die Ursachen untersucht, welche in einem Hybridmaiszuchtprogramm LD generieren und aufrecht erhalten. Hiermit wurde gezeigt, dass mit Plabsoft komplexe praktische Zuchtprogramme modelliert werden können.

Die Studien dieser Arbeit geben einen Überblick über das breite Anwendungsspektrum der entwickelten Simulations- und Analysesoftware Plabsoft. Darüber hinaus wurde Plabsoft bis jetzt in vierzig weiteren Studien verwendet, womit die Nützlichkeit von Plabsoft für die Integration neuer genomischer Werkzeuge in die angewandte Züchtungsforschung zweifelsfrei belegt wird.

Acknowledgements

I am very grateful to my academic supervisor Prof. Dr. A.E. Melchinger for his advise, suggestions and continuous support during this thesis work. Thanks to Prof. Dr. H.-P. Piepho and Prof. Dr. R. Blaich for serving on my graduate committee. Sincere thanks to Prof. Dr. Matthias Frisch for many discussions, excellent supervision, keeping the overview when things tended to turn too much into detail, and his never ending patience in proofreading.

The financial support from a fellowship by KWS Kleinwanzlebener Saatzucht AG, Einbeck, Germany, is greatfully acknowledged. In connection with that fellowship, I acknowledge the excellent collaboration with Dr. Milena Ouzunova and Dr. Carsten Knaak and the hospitality during my visits to Einbeck.

Many thanks to the patient beta-testers of Plabsoft: Jürgen Engler, Dr. Carsten Knaak, PD Dr. Jochen Reif, Tobias Schrag, Dr. Benjamin Stich, and Dr. Zoran Sušić. Furthermore, I would like to thank Mrs. H. Beck, Mrs. B. Boesig, Beate Devezi-Savula, and Mrs. S. Meyer for being of great help in organizational matters.

Many thanks to Sankalp Bhosale, Christine Beuter, Christof Bolduan, Dr. Susanne Dreisigacker, Dr. K. Christin Falke, Sandra Fischer, Dr. Christian Flachenecker, Nicole Friedl, Dr. Martin Heckenberger, Christiane Knopf, Elisabeth Kokai-Kota, Vera Kühr, Dr. Barbara Kusterer, Dr. Friedrich Longin, Franz Mauch, Dr. Vilson Mirdita, Dr. Zeljko Micic, Dr. Manuel Montes, Dr. Jasmina Muminović, Vanessa Prigge, Sina Strube, Dr. Anne-Celine Thuillet, Prof. Dr. H.F. Utz, Hans Henning Voß, Thilo Wegenast, Dr. Katinka Wilde, Dr. Pingzhi Zhang, and all unmentioned members within our institute for creating a pleasant work environment and for their support in scientific and non-scientific issues.

Last, but not least, I want to thank my family, my girlfriend, and my friends for their continuous support over the years.

Curriculum vitae

Name	Hans Peter Maurer
Date and Place of Birth	June 18, 1976 in Neuendettelsau
School Education	1982–1986, elementary school (Grundschule Windsbach) 1986–1995, high school (Johann-Sebastian Bach Gymnasium, Windsbach) Abitur June 1995
Civil Service	10/95–10/96, CVJM youth hostel "Burg Wernfels", Spalt
University Education	10/96–03/02, ,,Agrarbiologie", University of Hohenheim, Stuttgart Diplom-Agrarbiologe March 2002 since 10/98, ,,Informatik", University of Stuttgart, Stuttgart. Pre-diploma in July 2001 since 04/02, Doctorate candidate in Applied Genetics and Plant Breeding (Prof. Dr. A.E. Melchinger) at the University of Hohenheim, Stuttgart
Professional Experiences	02/97–06/98, Computing Center, University of Hohenheim, Stuttgart 07/98–10/98, on the farm of Christina Johanna Paulsen-Schlüter, Alte Dorfstr. 42, Tolk 08/00–10/00, Hölle & Hüttner AG, Tübingen 01/01–04/01, "Programming in Java", SIMT, Stuttgart
Employment Record	since 04/02, research associate at the Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Stuttgart