Population genetic simulation and data analysis with Plabsoft

Hans Peter Maurer · Albrecht E. Melchinger · Matthias Frisch

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.

Keywords Breeding informatics · Computer simulation · Data analysis · Population genetics

H. P. Maurer · A. E. Melchinger (✉) · M. Frisch
Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, Stuttgart 70593, Germany
e-mail: melchinger@uni-hohenheim.de

Introduction

Population genetics provides a theoretical framework to investigate the effects of random mating or planned crossing, selection, drift, and mutation on the genetic variation present in plant breeding populations. Simplifying assumptions are often necessary to obtain tractable mathematical models with analytical solutions. The most important assumptions are absence of selection, infinite population size, and random mating. However, selection in plant breeding is typically carried out in finite populations derived from planned crosses. Mathematical treatment of population genetics problems is therefore often not feasible.

Computer simulations are a widely used tool to obtain approximate solutions for such problems. While a few population genetic simulation programs are available (Tinker and Mather 1993; Podlich and Cooper 1998; Laval and Excoffier 2004), they do not provide the functionality required for many studies. Therefore, simulation software is often written that is capable of performing exactly the task required for a certain study, but is neither available to other researchers nor capable of solving a broad range of population genetics problems. Analysis of simulated data is another challenge, because population genetic software (Excoffier et al. 2005; Rohlf 2002; Liu and Muse 2005) is not designed to analyze repeatedly large simulated datasets and summarize the results.

We developed Plabsoft, a flexible and powerful software for population genetic simulation and data analysis. Our objectives were to provide software for (i) simulation of a broad range of population genetics problems occurring in plant breeding and genetics, (ii) analysis of simulated datasets employing state-of-the-art-methods and algorithms, and (iii) high quality analysis of experimental data.

Methods

Plabsoft is implemented as an add-on package for the statistical software R (Ihaka and Gentleman 1996). Computationally demanding algorithms and data management routines were written in C (Kernighan and Ritchie 1988) using the multiple precision library GMP (GMP 2006) and the scientific computing library GSL (Galassi et al. 2006). Computationally less demanding functions were written in the programming language R, which is not unlike S (Becker and Chambers 1981), and is an integral part of the statistical software R. Plabsoft provides more than 100 functions, which can be (i) called in arbitrary order, (ii) used to build up new, user-defined, high level functions, and (iii) employed in combination with the data analysis routines provided by R and other add-on packages of the R system. Each function is performance optimized, validated, and accompanied by extensive documentation (online as well as hard-copy) and application examples.

Meiosis is simulated with a count-location process (Karlin and Liberman 1978). In a first step (count), the number k of crossovers on a chromosome of length λ M is determined with a realization of Poisson distributed random variable with parameter λ. In a second step (location), the locations of the k crossovers are determined with realizations of a uniformly distributed random variable. This algorithm assumes that (a) the average number of crossovers formed on a chromosome is equal to its length in Morgan units and (b) the locations of crossovers are uniformly and independently distributed on a chromosome. These assumptions imply the absence of interference (Stam 1979) and are mathematically equivalent to those underlying Haldane's (1919) mapping function.

The genotypic value G of an individual for a certain trait is modeled by

$$G = \sum_{S \subseteq N} X_s, \tag{1}$$

where N is the set of all loci on both homologous chromosomes and X_s is an effect assigned to a given combination of alleles at a subset of loci S (Bulmer 1985, p. 46). This model allows for a flexible definition of genetic effects, including additive, dominance, and epistatic effects of any order. An arbitrary number of traits can be simulated, each with its own genetic architecture. Phenotypic values are simulated by adding non-genetic effects to the genotypic values according to arbitrary field designs or error structures.

Simulation and data analysis

Plabsoft does not distinguish between experimental and simulated data sets. Linkage map, marker and trait data from experimental studies can be imported from text files, whereas the base populations for simulations can be described by specifying relevant population genetic parameters. After providing Plabsoft with a dataset in one of these two ways, the dataset can be (a) directly analyzed with the provided data analysis routines, or (b) used to conduct simulations, which are subsequently followed by a data analysis (Fig. 1). In the following, we give an overview of the work flow for conducting simulations and data analyses with Plabsoft. For each subsequent step, we give a general description of the available functions, illustrated with examples for specific problems (Table 1).
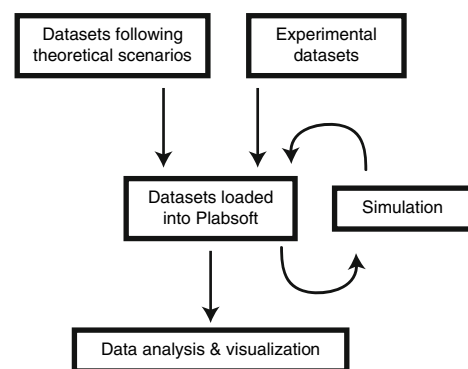
Fig. 1 Work flow of a Plabsoft session

Table 1 Examples for Plabsoft function calls

Command	Description
Genome parameters and genetic architecture	
`genome.parameter.set(10,2,rep(2,10))`	Define the genome parameters for a simulation with a model of the maize genome: 2 homologous chromosome sets, 10 chromosomes of length 2 M)
`linkage.map.load("fsrs.map")`	Load the linkage map for an experimental dataset from a file
`linkage.map.define("*100 ssr mp terminal, *1000 qtl q random")`	Define a linkage map with 100 equally spaced SSR markers and 1,000 randomly distributed QTL
`effect.define("qtl","0, q 1 normal 1/0.05")`	Define additive effects for the QTL. The effect sizes follow a normal distribution with mean 1 and variance 0.05
Initial data sets	
`population.init("P1",homozygote(1))`	Generate a population P1 consisting of one completely homozygous inbred line
`population.matrix.load("bcexperiment.pop")`	Load molecular marker data from a file
Mating schemes	
`cross("BC2","BC1","P2",100)`	Generate a BC2 population of size 100 from a backcross to parent P2
`cross("SYN3","SYN2","SYN2",100,self=0.5)`	Generate a SYN3 population of size 100 assuming a selfing rate of 50%
`ssd.mating("SSD","F2")`	Generate a population of single-seed descent lines from a F2 base population
`dh("DH","F1",100)`	Generate 100 double-haploid lines from a F1 population
Selection strategies	
`select.n.best("RU","C3","resistance",100)`	Select the 100 plants with the highest score for a resistance trait from population C3 and save them in a new population called RU
`select.all.best("BC1s","BC1","target")`	Select all individuals carrying the target gene in a marker-assisted backcrossing program
Data analysis routines	
`evaluate.allele("Pop48","phi056")`	Evaluate the allele frequency distribution at locus phi056 in population Pop48
`evaluate.genotype("Pop48","phi056")`	Evaluate the genotype frequency distribution at locus phi056 in population Pop48
`hwe.monte("Pop48","phi056")`	Test of Hardy-Weinberg equilibrium at locus phi056 in population Pop48
`ld.pw.genotypic.monte("Pop48","Pop48.ld")`	Test of linkage disequilibrium in population Pop48 for all pairs of loci on the Defined linkage map
`ld.matrix.plot("Pop48.ld",measure="DD")`	Plot the result of the linkage disequilibrium analysis in a matrix plot
`ggt.plot("Pop48")`	Plot the graphical genotype of the indivduals of population Pop48
`pcoa.plot("Pop24 Pop22 Pop48",dim=2)`	Plot a two dimensional principal coordinate analysis of all individuals in the three populations Pop24, Pop22, and Pop48

Genome parameters and genetic architecture of traits are defined in the first step of a Plabsoft session. In particular, (a) ploidy level, number and length of chromosomes, (b) linkage maps of markers and loci responsible for qualitative or quantitative traits, and (c) genetic architecture of the traits under consideration, can be specified. Linkage maps can be imported from text files, e.g., from the output of a mapping program or a QTL study. In addition, they can be generated by the software following

assumptions on the number and distribution of loci in the genome. The genetic architecture of a trait can be defined either by import of estimated effects, e.g., from a QTL mapping study, or by specifying assumptions on the mode of inheritance and effect size (Eq. 1).

Initial datasets can be provided either by import or by description of population genetic parameters. Datasets can be imported from text files or from SQL queries of a database system such as the Plabsoft database (Heckenberger et al. 2007). In addition, it is possible to generate populations following theoretical scenarios. For example, populations of homozygous inbred lines and populations in Hardy-Weinberg or linkage equilibrium can be generated.

Mating schemes, which are employed in genetic experiments and applied plant breeding programs are available for the simulations. For example, planned crosses, backcrosses, random mating with or without self incompatibility, partial selfing, or recurrent full-sib mating can be simulated. Furthermore, single-seed descent mating, derivation of doubled haploids, as well as more complex mating schemes such as topcrosses, chaincrosses, factorial or diallel crosses are possible.

Selection strategies can be defined by the user based on phenotypic values and/or marker scores. For selection based purely on the phenotype, selection indices combining the performance with respect to an arbitrary number of traits are available. For selection in a marker-assisted backcrossing program, the selection strategies described by Frisch et al. (1999) are available. For marker-assisted selection on quantitative traits, selection indices combining phenotypic value and marker score can be employed. With each of these selection indices, selection can be carried out according to various modes. For example, all individuals above a given threshold or, alternatively, a given number of individuals with a largest selection indices are selected from the base population.

Data analysis routines are available to analyze a broad range of population genetic and quantitative genetic parameters. In the field of general population genetics, estimation of allele and genotype frequencies, approximate and exact tests for Hardy-Weinberg equilibrium (Maurer et al. 2007) and linkage disequilibrium, and calculation of linkage disequilibrium measures like D^2, D', r^2, D'_m, and R (Lewontin 1964; Hill and Robertson 1968; Franklin and Le-

wontin 1970; Maruyama 1982; Hedrick 1987) are implemented. The results of linkage disequilibrium analyses can be visualized, e.g., with matrix plots (cf. Falke et al. 2007a, b). A sophisticated module for plotting graphical genotypes is also available, which allows, for example, interactive sorting of individuals, zooming into chromosome regions and export to common graphics formats. For quantitative genetic studies, the genotypic and phenotypic values of individuals and the genetic variances of populations can be estimated. In marker-assisted backcrossing programs, the genome contribution of a recurrent parent to a population or selected individual, the number and length of chromosome segments originating from the donor parent, and the number of required marker data points can be determined. For genetic diversity studies, various distance measures and similarity coefficients are implemented such as the Roger's distance (Rogers 1972), modified Roger's distance (Wright 1978, p. 78), Nei-Li distance (Nei and Li 1979), Euclidean distance, as well as Dice (1945), Jaccard (1908), and Simple matching coefficients. For visualization, principal coordinate plots and dendrograms are available.

Applications

We give an overview of the functionality of Plabsoft by the example of published studies in different research areas.

Linkage disequilibrium in European maize was analyzed with simple sequence repeat (SSR) and amplified fragment length polymorphism (AFLP) markers to assess the prospects of association mapping in plant breeding (Maurer et al. 2006; Stich et al. 2005, 2006a) and to investigate linkage disequilibrium decay in recurrent full-sib selection programs (Falke et al. 2007a, b).

Examples for the application of Plabsoft in genetic diversity studies with different crops and marker systems are the studies of Dreisigacker et al. (2004, 2005), Muminović et al. (2004a, b, 2005), Reif et al. (2003a, b, 2004, 2005a, b, c, d, 2006a, b), Tams et al. (2005), Xia et al. (2004, 2005), and Zhang et al. (2005).

Plabsoft comprises the functionality of our earlier software Plabsim (Frisch et al. 2000) to optimize marker-assisted backcrossing programs for dominant

target genes (Frisch et al. 1999), recessive target genes (Frisch et al. 2001a), and two target genes (Frisch et al. 2001b). Further routines for pre-test estimation of response to selection in marker-assisted backcrossing programs (Frisch and Melchinger 2005) are implemented.

In the context of plant variety protection, Heckenberger et al. (2005a, b, c) employed simulations to determine thresholds for detection of essentially derived varieties.

Further examples for simulation with Plabsoft are the studies conducted to evaluate the power of a new family-based association mapping test (Stich et al. 2006b), simulate the breeding history of European maize for investigation of the causes of linkage disequilibrium in breeding programs (Stich et al. 2007), and investigate the prediction error of marker-based predictions of the parental contribution to inbred lines derived from biparental crosses (Frisch and Melchinger 2006).

Availability

Plabsoft runs under the Microsoft Windows and Linux operating systems. Licensing of the Plabsoft modules developed in the research project GABI-BRAIN is planned after 2007.

References

Becker RA, Chambers JM (1981) S: a language and system for data analysis bell laboratories computer information service. Murray Hill, New Jersey

Bulmer MG (1985) The mathematical theory of quantitative genetics. Oxford University Press, Oxford

Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26:197–302

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

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

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

Falke KC, Maurer HP, Melchinger AE, Piepho H-P, Flachenecker C, Frisch M (2007a) Linkage disequilibrium in two European F2 flint maize populations under modified recurrent full-sib selection. Theor Appl Genet 115:289–297

Falke KC, Maurer HP, Melchinger AE, Piepho H-P, Flachenecker C, FrischM (2007b) Linkage disequilibrium in two European F2 flint maize populations under modified recurrent full-sib selection. Theor Appl Genet 115:299

Franklin I, Lewontin RC (1970) Is the gene the unit of selection? Genetics 65:707–734

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 (2000) Plabsim: software for simulation of marker-assisted backcrossing. J Hered 91:86–87

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

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

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

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

Galassi M, Davies J, Theiler J, Gough B, Jungman G, Booth M, Rossi F (2006) GNU scientific library reference manual—revised second edition (v 18). Network Theory Ltd, Bristol UK

GMP (2006) GNU multiple precision arithmetic library. http://www.swox.com/gmp. Cited 27 Nov 2006

Haldane JBS (1919) The combination of linkage values and the calculation of distances between linked factors. J Genet 8:299–309

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

Heckenberger M, Bohn M, Melchinger AE (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, Bohn M, Klein D, Melchinger AE (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, Bohn M, Frisch M, Maurer HP, Melchinger AE (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, Melchinger AE, Maurer HP, Frisch M (2007) The Plabsoft-Database—a comprehensive database management system for integrating phenotypic and genomic data in academic and commercial plant breeding programs. Euphytica doi: 10.1007/s10681-007-9478-3

Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations. Theor Appl Genet 38:226–231

Ihaka R, Gentleman R (1996) R: a language for data analysis and graphics. J Comput Graph Stat 5:299–314

Jaccard P (1908) Nouvelles recherches sur la distribution florale. Bul Soc Vaudoise Sci Nat 44:223–270

Karlin S, Liberman U (1978) Classification of multilocus recombination distributions. Proc Nat Acad Sci USA 75:6332–6336

Kernighan BW, Ritchie DM (1988) The C Programming Language, (2nd edn, ANSI C), Prentice Hall

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

Lewontin RC (1964) The interaction of selection and linkage I General considerations; Heterotic models. Genetics 49:49–67

Liu K, Muse SV (2005) Powermarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21:2128–2129

Maruyama T (1982) Stochastic integrals and their application to population genetics. In: Kimura M (ed) Molecular evolution, protein polymorphism and the neutral theory. Springer-Verlag, Berlin, pp 151–166

Maurer HP, Knaak C, Melchinger AE, Ouzunova M, Frisch M (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 HP, Melchinger AE, Frisch M (2007) An incomplete enumeration algorithm for an exact test of Hardy-Weinberg proportions with multiple alleles. Theor Appl Genet doi: 10.1007/s00122-007-0573-6

Muminović J, Melchinger AE, Lübberstedt T (2004a) Prospects for celeriac (Apium graveolens var rapaceum) improvement by using genetic resources of Apium, as determined by AFLP markers and morphological characterization. Plant Genet Resour 2:189–198

Muminović J, Melchinger AE, Lübberstedt T (2004b) Genetic diversity in cornsalad (Valerianella locusta) and related species as determined by AFLP markers. Plant Breed 123:460–466

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

Nei M, Li WH (1979) Mathematical Models for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci USA 76:5268–5371

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

Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Srinivasan G, Bohn M, Frisch M (2003a) Genetic distance based on simple sequence repeats and heterosis in tropical maize populations. Crop Sci 43:1275–1282

Reif JC, Melchinger AE, Xia XC, Warburton ML, Hoisington DA, Vasal SK, Beck D, Bohn M, Frisch M (2003b) Use of SSRs for establishing heterotic groups in subtropical maize. Theor Appl Genet 107:947–957

Reif JC, Xia XC, Melchinger AE, Warburton ML, Hoisington DA, Beck D, Bohn M, Frisch M (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 JC, Zhang P, Dreisigacker S, Warburton ML, Van Ginkel M, Hoisington D, Bohn M, Melchinger AE (2005a) Wheat genetic diversity trends during domestication and breeding. Theor Appl Genet 110:859–864

Reif JC, Melchinger AE, Frisch M (2005b) Genetical and mathematical properties of similarity coefficients applied in plant breeding and seed bank management. Crop Sci 45:1–7

Reif JC, Hamrit S, Heckenberger M, Schipprack W, Maurer HP, Bohn M, Melchinger AE (2005c) Trends in genetic diversity among European maize cultivars and their parental components during the past 50 years. Theor Appl Genet 111:838–845

Reif JC, Hamrit S, Heckenberger M, Schipprack W, Maurer HP, Bohn M, Melchinger AE (2005d) Genetic structure and diversity of European flint maize populations determined with SSR analyses of individuals and bulks. Theor Appl Genet 111:906–913

Reif JC, Warburton ML, Xia XC, Hoisington DA, Crossa J, Taba S, Muminović J, Bohn M, Frisch M, Melchinger AE (2006a) Classification of Mexican races of maize revisited with SSR markers. Theor Appl Genet 113:177–185

Reif JC, Hamrit S, Melchinger AE (2006b) Genetic diversity trends in Central European heterotic groups. Acta Agronomica Hungarica 54:315–320

Rogers JS (1972) Measures of genetic similarity and genetic distance VII. Univ Tex Publ 7213:145–153

Rohlf FJ (2002) NTSYSpc: Numerical Taxonomy System, ver 2.1. Exeter Publishing Ltd, Setauket NY

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

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Appl Genet 111:723–730

Stich B, Maurer HP, Melchinger AE, Frisch M, Heckenberger M, Rouppe van der Voort J, Peleman J, Srensen AP, Reif JC (2006a) Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. Mol Breed 17:217–226

Stich B, Melchinger AE, Piepho H-P, Heckenberger M, Maurer HP, Reif JC (2006b) A new test for family-based association mapping with inbred lines from plant breeding programs. Theor Appl Genet 113:1121–1130

Stich B, Melchinger AE, Piepho H-P, Hamrit S, Schipprack W, Maurer HP, Reif JC (2007) Potential causes of linkage disequilibrium in a European maize breeding program investigated with computer simulations. Theor Appl Genet doi: 10.1007/s00122-007-0586-1

Tams SH, Melchinger AE, Bauer E (2005) Genetic similarity among European winter triticale elite germplasms assessed with AFLP and comparisons with SSR and pedigree data. Plant Breed 124:154–160

Tinker NA, Mather DE (1993) Gregor: software for genetic simulation. J Hered 84:237

Wright S (1978) Evolution and the genetics of populations, vol 4: variability within and among natural populations. University of Chicago Press, Chicago, Illinois

Xia XC, Reif JC, Hoisington DA, Melchinger AE, Frisch M, Warburton ML (2004) Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: I Lowland tropical maize. Crop Sci 44:2230–2237

Xia XC, Reif JC, Melchinger AE, Frisch M, Hoisington DA, Warburton ML (2005) Genetic diversity among CIMMYT maize inbred lines investigated with SSR markers: II Subtropical, tropical mid-altitude and highland maize inbred lines and their relationships with elite US and European maize. Crop Sci 45:2573–2582

Zhang P, Dreisigacker S, Melchinger AE, Reif JC, Mujeeb-Kazi A, Van Ginkel M, Hoisington D, Warburton ML (2005) Quantifying novel sequence variation and selective advantage in synthetic hexaploid wheats and their backcross-derived lines using SSR markers. Mol Breed 15:1–10