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

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 \( \lambda M \) is determined with a realization of Poisson distributed random variable with parameter \( \lambda \). 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).

![Diagram of Plabsoft workflow](Fig. 1 Work flow of a Plabsoft session)
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:

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>genome.parameter.set(10,2,rep(2,10))</td>
<td>Define the genome parameters for a simulation with a model of the maize genome: 2 homologous chromosome sets, 10 chromosomes of length 2 M</td>
</tr>
<tr>
<td>linkage.map.load(&quot;fsrs.map&quot;)</td>
<td>Load the linkage map for an experimental dataset from a file</td>
</tr>
<tr>
<td>linkage.map.define(&quot;*100 ssr mp terminal, *1000 qtl q random&quot;)</td>
<td>Define a linkage map with 100 equally spaced SSR markers and 1,000 randomly distributed QTL</td>
</tr>
<tr>
<td>effect.define(&quot;qtl&quot;, &quot;0, q 1 normal 1/0.05&quot;)</td>
<td>Define additive effects for the QTL. The effect sizes follow a normal distribution with mean 1 and variance 0.05</td>
</tr>
</tbody>
</table>

Initial data sets

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>population.init(&quot;P1&quot;, homozygote(1))</td>
<td>Generate a population P1 consisting of one completely homozygous inbred line</td>
</tr>
<tr>
<td>population.matrix.load(&quot;bcexperiment.pop&quot;)</td>
<td>Load molecular marker data from a file</td>
</tr>
</tbody>
</table>

Mating schemes

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>cross(&quot;BC2&quot;, &quot;BC1&quot;, &quot;P2&quot;, 100)</td>
<td>Generate a BC2 population of size 100 from a backcross to parent P2</td>
</tr>
<tr>
<td>cross(&quot;SYN3&quot;, &quot;SYN2&quot;, &quot;SYN2&quot;, 100, self=0.5)</td>
<td>Generate a SYN3 population of size 100 assuming a selfing rate of 50%</td>
</tr>
<tr>
<td>ssd.mating(&quot;SSD&quot;, &quot;F2&quot;)</td>
<td>Generate a population of single-seed descent lines from a F2 base population</td>
</tr>
<tr>
<td>dh(&quot;DH&quot;, &quot;F1&quot;, 100)</td>
<td>Generate 100 double-haploid lines from a F1 population</td>
</tr>
</tbody>
</table>

Selection strategies

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>select.n.best(&quot;RU&quot;, &quot;C3&quot;, &quot;resistance&quot;, 100)</td>
<td>Select the 100 plants with the highest score for a resistance trait from population C3 and save them in a new population called RU</td>
</tr>
<tr>
<td>select.all.best(&quot;BC1s&quot;, &quot;BC1&quot;, &quot;target&quot;)</td>
<td>Select all individuals carrying the target gene in a marker-assisted backcrossing program</td>
</tr>
</tbody>
</table>

Data analysis routines

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>evaluate.allele(&quot;Pop48&quot;, &quot;phi056&quot;)</td>
<td>Evaluate the allele frequency distribution at locus phi056 in population Pop48</td>
</tr>
<tr>
<td>evaluate.genotype(&quot;Pop48&quot;, &quot;phi056&quot;)</td>
<td>Evaluate the genotype frequency distribution at locus phi056 in population Pop48</td>
</tr>
<tr>
<td>hwe.monte(&quot;Pop48&quot;, &quot;phi056&quot;)</td>
<td>Test of Hardy-Weinberg equilibrium at locus phi056 in population Pop48</td>
</tr>
<tr>
<td>ld.pw.genotypic.monte(&quot;Pop48&quot;, &quot;Pop48.ld&quot;)</td>
<td>Test of linkage disequilibrium in population Pop48 for all pairs of loci on the Defined linkage map</td>
</tr>
<tr>
<td>ld.matrix.plot(&quot;Pop48.ld&quot;, measure=&quot;DD&quot;)</td>
<td>Plot the result of the linkage disequilibrium analysis in a matrix plot</td>
</tr>
<tr>
<td>ggt.plot(&quot;Pop48&quot;)</td>
<td>Plot the graphical genotype of the individuals of population Pop48</td>
</tr>
<tr>
<td>pcoa.plot(&quot;Pop24 Pop22 Pop48&quot;, dim=2)</td>
<td>Plot a two dimensional principal coordinate analysis of all individuals in the three populations Pop24, Pop22, and Pop48</td>
</tr>
</tbody>
</table>
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 Lewontin 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.

Acknowledgments

The development of Plabsoft was in part funded by the European Union Grant No. QLK5CT-2000-007221 ‘GENE-MINE’, KWS Saat AG, Einbeck, Germany, and the German Federal Ministry of Education and Research (BMBF) Grant No. (0313126) ‘GABI-BRAIN’. The authors appreciate the editorial work of Dr. J. Muminović, whose suggestions considerably improved the style of the manuscript.

References


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


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


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


Reif JC, Hamrít 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


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


