Euphytica (2008) 161:173–179
DOI 10.1007/s10681-007-9478-3

The Plabsoft database: a comprehensive database management system for integrating phenotypic and genomic data in academic and commercial plant breeding programs

Martin Heckenberger · Hans Peter Maurer · Albrecht E. Melchinger · Matthias Frisch

Received: 1 December 2006 / Accepted: 4 June 2007 / Published online: 27 June 2007

Abstract Besides phenotypic data from field trials and molecular data from lab experiments, modern plant breeding programs generate a wide variety of data, for instance pedigree, randomization, geostatistical or climate data. Due to the lack of an integrated database system, breeders generally exploit only part of these data for selection decisions or retrieve only part of the information present in the data. Most approaches in genomics, however, develop their full power only when they are based on analyses of large numbers of genotypes from multiple crosses and current as well as past generations. We have developed a flexible data management and -analyses system for storage and quality control of plant breeding data. It is implemented using the PostgreSQL database management system and linked to the R software environment for integrated statistical analyses of phenotypic and genomic data. The database structure is capable of managing the following types of data observed in breeding programs of all major crops: (a) germplasm data of any species including pedigree data, (b) phenotypic data of any trait and trait complexity, (c) trial management data for any field and trial design, (d) molecular marker data for all common types of markers, as well as (e) project and study management data.

Keywords Breeding informatics · Database · Statistical analyses

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

Introduction

Plant breeding aims at the purposeful manipulation of plant species in order to create desired genotypes and phenotypes for specific requirements. This manipulation involves either controlled pollination, genetic engineering, or both, followed by selection of progenies showing the desired phenotypes. On the one hand, the success of plant breeding efforts depends on the creative intuition of the breeder and the occurrence of random recombination with positive effects, but on the other hand also to a large extent on the focused utilization of various kinds of information on potential parents or progenies of breeding populations. In classical plant breeding, this information consists mainly of the statistical analysis of phenotypic data from field experiments. Modern plant breeding programs, however, are characterized by the extensive use of genomic data (mainly molecular markers, but increasingly metabolite and gene expression data), inclusion of pedigree data (Piepho et al. 2006), information on relatives or spatial data, as well as the application of mobile computing devices throughout all stages of the breeding process.

This increasing complexity of breeding data requires complicated statistical models and high-end computing power for analyses, as well as sophisticated data management and data mining facilities to account for the high degree of integration of the data.

Data management and analysis systems, currently available in plant breeding were developed for analysis of field data with appropriate statistical methods, e.g., "PLABSTAT" (Utz 2001). Few software packages offer tools for administration and biometrical analyses of molecular marker data, but these do not implement links to phenotypic and pedigree data. Software packages written for special tasks such as administration of pedigree records and nursery books generally lack links to genomic data e.g., "Agrobase" (Agrobase 2006). Currently, no database system is available integrating the multitude of functions needed for efficient data management in modern plant breeding programs and providing interfaces to statistical analysis software. Due to this lack of an integrated database system, breeders generally exploit only part of their data for selection decisions (e.g., only data from the current season is used) or retrieve only part of the information present in the data (e.g., information on related genotypes is ignored). Most approaches in genomics (e.g., pedigree-based or haplotype-based QTL mapping), however, develop their full power only when they are based on analyses of large numbers of genotypes from multiple crosses and current as well as past generations.

Because neither of these options are designed for integrated long-term data storage, the following problems often arise: (1) The same data are repeatedly stored in different locations, this may provoke data inconsistency if only one copy of the dataset is changed and requires an unnecessarily large amount of storage capacity; (2) only the experimenter can reproduce the coding and structure of the stored data, which complicates reanalysis of the data; (3) considerable time is required to convert data stored in a certain format into another format that can be input in larger databases or another software; and (4) combining data from several experiments for a joint analysis is difficult.

Recently developed databases, such as "Panzea" (Zhao et al. 2006), or "Germinate" (Lee et al. 2005) were either developed specifically for data management in germplasm collections or lack the necessary flexibility for application in plant breeding programs. To our knowledge, there exists no concept for efficient storage of integrated phenotypic and molecular data, which focuses on applications in a modern plant breeding program. Our main goal was to develop a data structure for storage of molecular marker data in databases, which overcomes the shortcomings of data management in spreadsheets or input files of analysis software. The proposed data structure avoids redundant storage of experimental data and provides a standardized storage format, which facilitates retrieval, reanalysis, and exchange of the data. Our objectives were to (1) develop a flexible database structure capable of integrating phenotypic and genomic data from modern plant breeding programs and genomics research, (2) implement this database structure into a commonly used Database Management System (DBMS), and (3) develop functions for data import, data retrieval and data transfer from and to commonly used statistical analysis software.

Database architecture

Overview

The primary goal of developing the Plabsoft database was the need of a database management system, integrating phenotypic and genomic data available in modern plant breeding programs together with a wide variety of supporting data available from academic or commercial plant breeding processes (e.g., pedigree data, data from lab samples, etc.). Specific needs of commercial and academic project partners in the GABI-BRAIN-Project (Heckenberger et al. 2006) required a database structure, which meets the following criteria: (1) the database structure should be kept as simple as possible for performance reasons, but highly flexible to new developments in genomics or breeding methodology, (2) the database should be easy to maintain and capable of being either used as a single-user database on a desktop computer or within a client server solution for many users, possibly spread across many different places in the world, and (3) the database should be capable of delivering data to commonly used analyses software, such as "R" (R Development Core Team 2004) or Plabsoft (Maurer et al. 2007). The Plabsoft database

is implemented in the PostgreSQL DBMS (PostgreSQL 2006) and is structured in several modules.

Core module

The core module of the Plabsoft database allows the management of projects, project consortia, and project teams (Fig. 1). This enables the database user to specifically assign functions to persons and manage contact data, such as addresses and URLs for persons and institutions participating in projects. The aim of this module is to structure the data according to the projects in which they were developed and to assign personal responsibilities for each data point stored in the database later. In addition, a hierarchy of persons or institutions can be established by defining, which institution is a department of another institution or which person belongs to a certain department.

Trial management

In this module, projects are divided into studies and studies into single experiments to structure the phenotypic and genomic data stored in the database. By default, restrictions in the database model ensure that each experiment is assigned to a study and each study to a project. These restrictions guarantee that each data point stored in the following experiments is uniquely characterized by its affiliation to a project. The designation of an experiment to a study and later to a project is independent of the analysis of the data, as it is

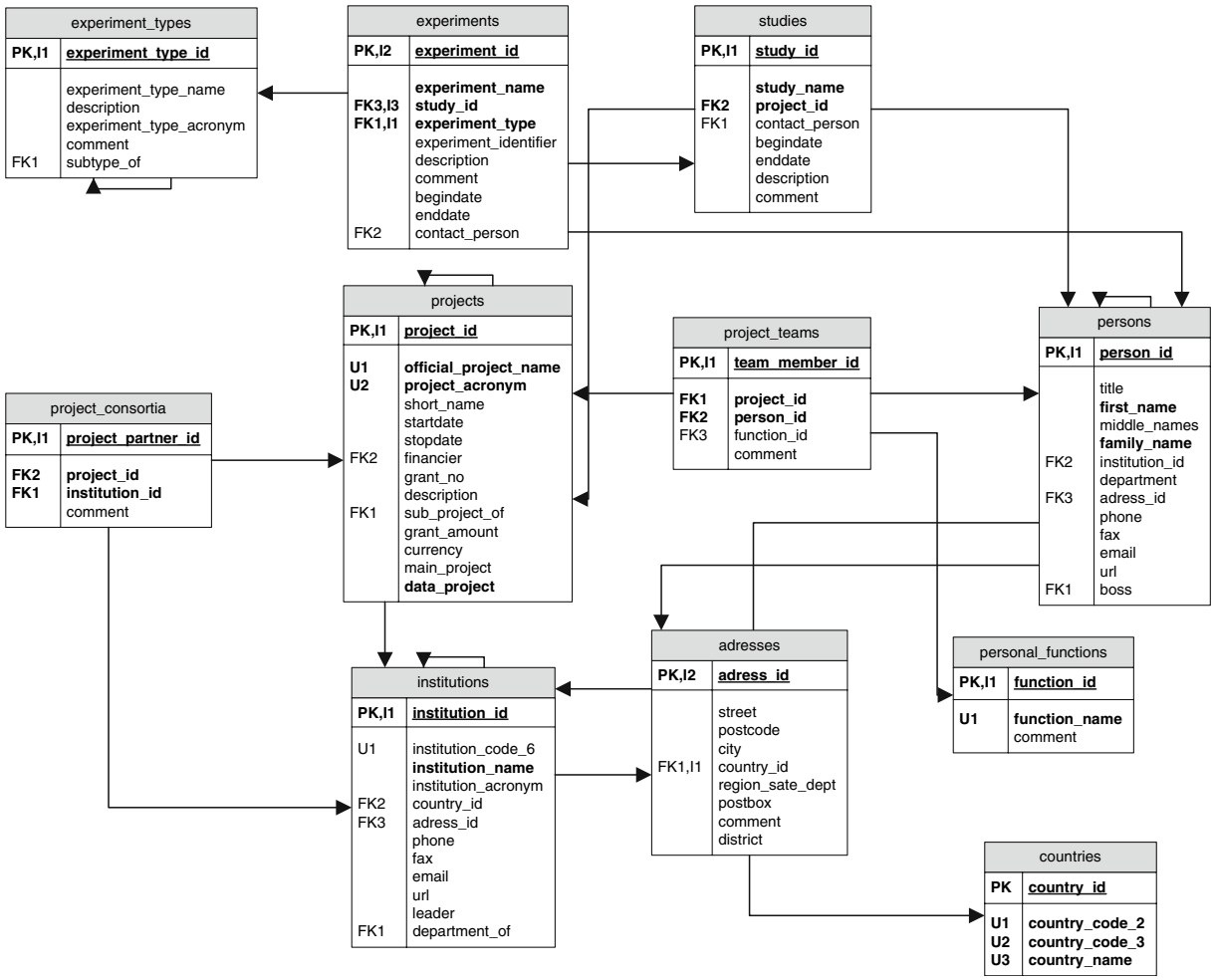

Fig. 1 Entity-Relationship-Diagram of the project and trial management modules of the Plabsoft database. Codes: PK, primary key; U, unique index; I, index; FK, foreign key (the numbers following indicate if multiple columns are a part of the same index)

possible to link data from any project, study, or experiment for joint analyses.

Experiments can be of any size and type, where the types of experiments can be specifically defined by the user in the table "experiment types". Typical experiment types are "field trial", "lab experiment", or "greenhouse trial". A recursive relation enables the user to define general and subtypes of experiments and allows, therefore, to implement a hierarchy of experiment types, which may be useful for data mining approaches and analyses across different experiments. For example, "SSR-study" could be a subtype of "marker experiments" which is itself a subtype of "lab experiments".

Germplasm management

Operational taxonomic units (OTUs) of any type can be handled in the table "otus" (Fig. 2). The possible types of OTUs are defined in the table "otu types", such as "inbred line", "population", or "F2 individual". If a population is only to be treated as such, without consideration of specific individuals within the population, the population is an ordinary OTU. If specific individuals of a population need to be considered separately, they are stored as independent OTUs. The affiliation of OTUs to populations is then managed in table "pop components".

Germplasm attributes of any class, such as "number of grain rows", "quality level", or "germplasm pool" can be defined for each OTU in the table of the same name. Either, the possible values for a given category can be predefined or entered as free text. The crop, which an OTU belongs to, is defined in table "species", where also subtaxa can be defined, e.g., when exotic germplasm of different convarieties or subspecies shall be introduced into the breeding pool.

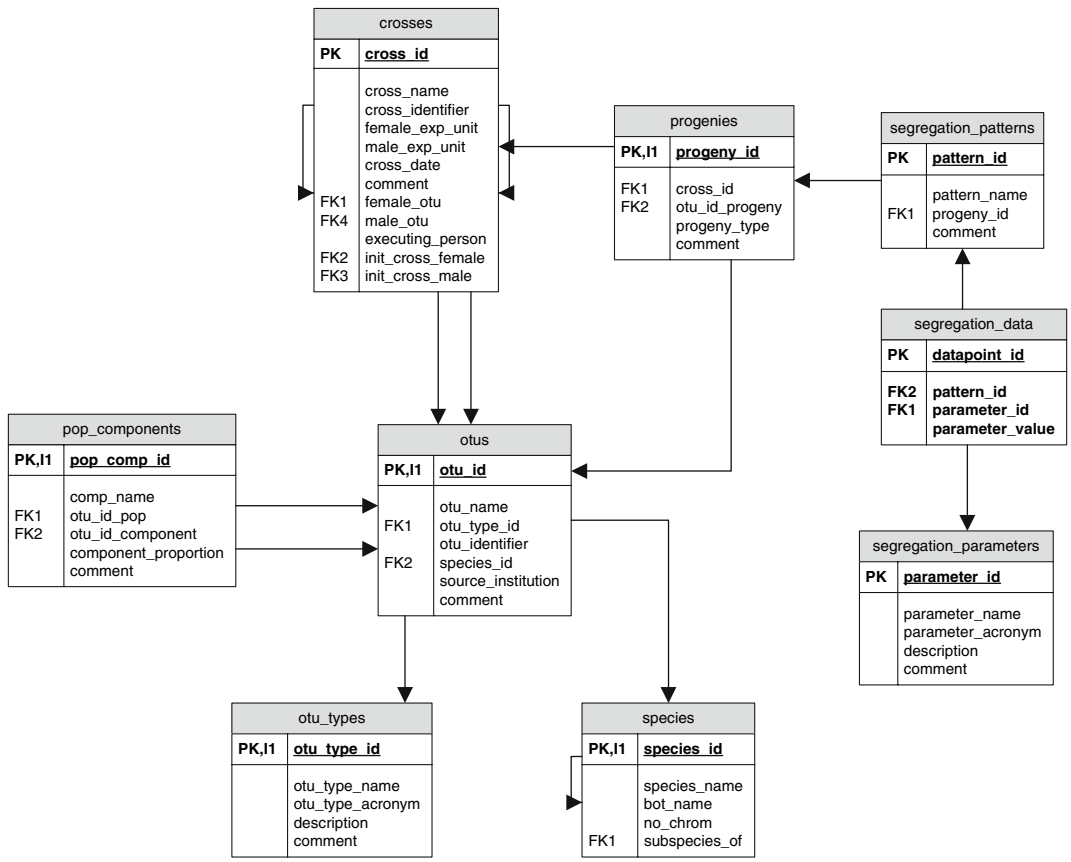

Fig. 2 Entity-Relationship-Diagram of the germplasm and pedigree management modules of the Plabsoft database. Codes: PK, primary key; U, unique index; I, index; FK, foreign key (the numbers following indicate if multiple columns are a part of the same index)

The table "experiment entries" assigns OTUs as entries of a specific experiment. An entry number, valid throughout the experiment is assigned to each OTU, analyzed in the experiment. Once experimental entries are defined, they can be assigned to experimental units. According to the type of the experiment, typical experimental unit types are "field plot", "greenhouse plot" or "lab sample".

Experimental units

Analogous to experiment types, subtypes of the different experimental unit types can be defined (Fig. 3). For example, "yield trial field plot wheat" is a subtype of "field plot wheat", which is itself a subtype of "field plot general". In addition, attributes of any type and class can be defined an assigned to the experimental units. This includes, e.g., plot length and width or the number of sown rows in a plot which is important, for example, when converting plot yield into yield per surface unit of measure.

If an experiment is a field trial, a greenhouse-, or a growth chamber experiment, it might be necessary to randomize the experimental units appropriately. The design factors of the randomization, such as "replication", "incomplete block", "location", or "year" can be freely defined in the table "designfactors". The randomization itself is then managed for each experimental unit in the table "randomization". This flexible model enables to handle any type of randomization for any type of experiment.

Pedigree management

The availability of appropriate pedigree data is crucial for optimized phenotypic analyses (Piepho et al. 2006) as well as for evaluation of the population structure in pedigree-based QTL-Mapping and association mapping approaches. Consequently, the Plabsoft database offers a flexible system for detailed management of pedigree relationships between OTUs. In principle, the pedigree of an OTU is defined by its crossing parents and by its segregation pattern. Crosses can be documented between each OTU, independent of the type of OTU on a cross itself can consist of several steps, e.g., a backcross, a three-way cross or a double-cross. Segregation parameters, such as the identifier of a specific F_2- or DH-plant of a cross can be freely defined and managed separately for each progeny of a cross.

Phenotypic data

Phenotypic data can be stored for each experimental unit and each trait defined in the table "traits" (Fig. 4). By a check constraint, it is assured that only one datapoint per trait and experimental unit is inserted. If time series of traits shall be documented, the different observational times can be either handled as separate traits or the above-mentioned check constraint can be deactivated and the identifier for the observational time can be documented in the "comment"-column of the table.

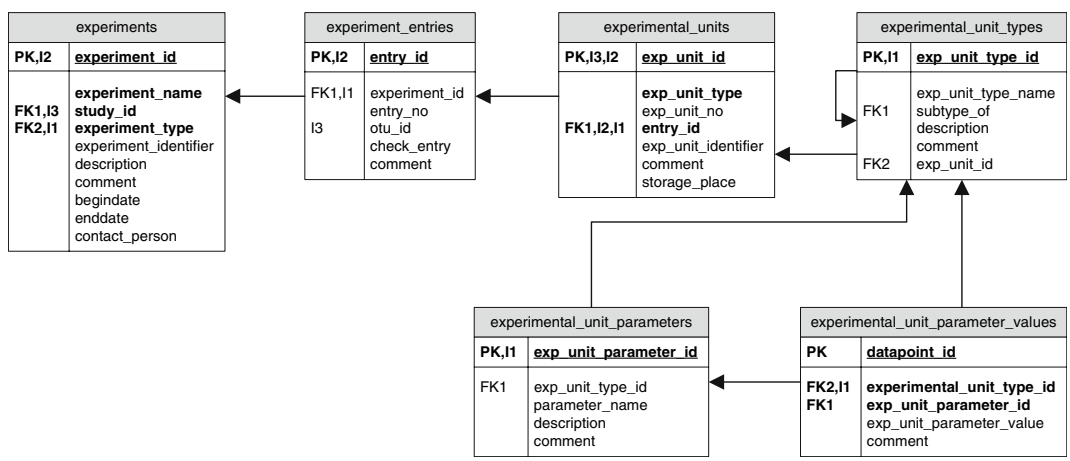

Fig. 3 Entity-Relationship-Diagram of the experimental units module of the Plabsoft Database. Codes: PK, primary key; U, unique index; I, index; FK, foreign key (the numbers following indicate if multiple columns are a part of the same index)

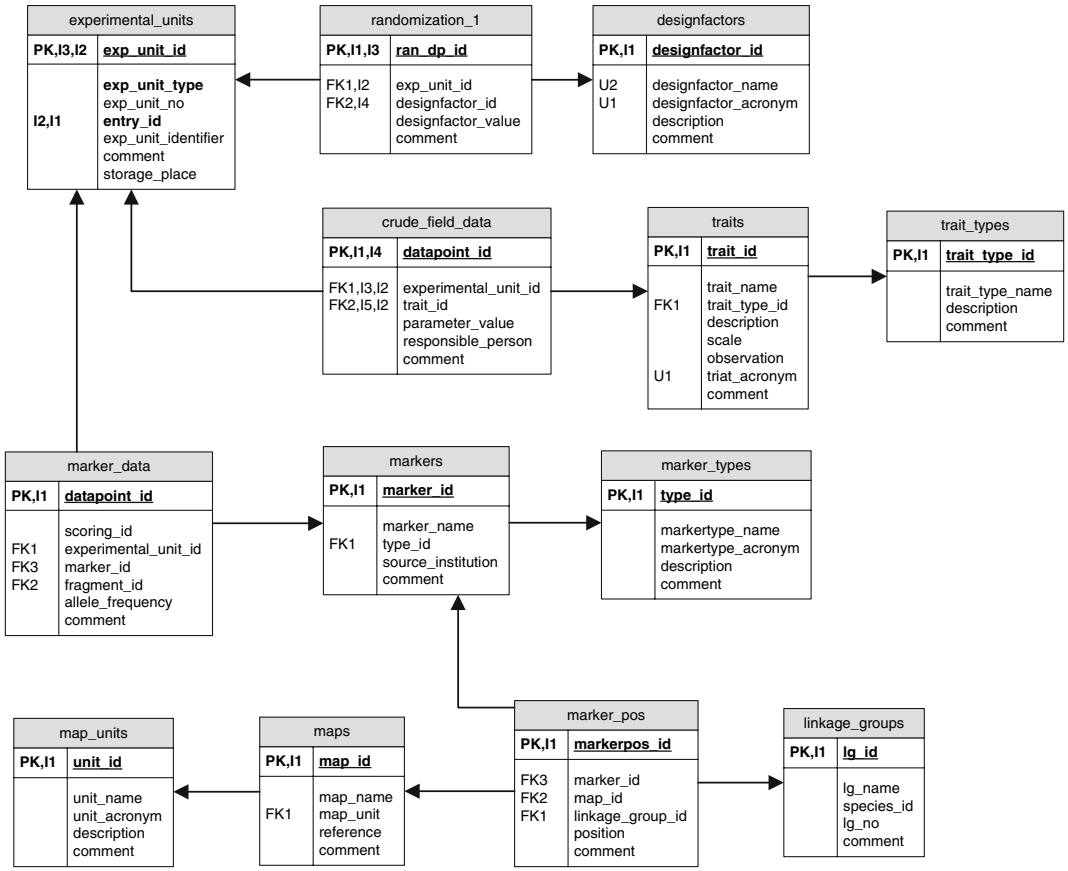

Fig. 4 Entity-Relationship-Diagram of the molecular and phenotypic data management modules of the Plabsoft database. Codes: PK, primary key; U, unique index; I, index; FK, foreign key (the numbers following indicate if multiple columns are a part of the same index)

Molecular data

This module handles molecular marker data of any complexity (Fig. 4). First, properties of molecular markers, such as the type of marker and the institution, where it was developed can be managed in the table "markers". Once markers are defined, marker bands obtained for any experimental unit (given it is any type of DNA-Sample) can be stored in the table "marker data". For polyploids, information on the presumable number of copies of the band can be included. In addition, genetic Maps for molecular markers can be stored in a way that each marker can be assigned to different maps, e.g., for different mapping populations.

Data entry and retrieval

Data entry and retrieval can be done either by using different available graphical interfaces of PostgreSQL, such as "PgAdmin" or the web interface "phpPgAdmin", or by a wide variety of stored procedures and functions which are implemented in R or directly in the database. In addition, the interface to the R software environment together with the Plabsoft (Maurer et al. 2007) add-on package offers the possibility to perform integrated analyses of phenotypic, molecular, and other data, directly retrieved from the database (e.g., association mapping, advanced phenotypic analyses using BLUP, genetic diversity analyses, plant breeding simulations).

Conclusion

The presented database is based on (a) a large body of phenotypic and genomic data from modern applied breeding programs of six crops and (b) information

about the structure and dimensions of these breeding programs. This guarantees practical relevance of the methods and software for all crops.

The Plabsoft database together with Plabsoft within the R software environment overcomes the shortcomings of currently available data management and analyses systems (e.g., only data from the current season is used or information on related genotypes is ignored). It provides the basis for plant breeders and researchers to perform integrated analyses of phenotypic and molecular data across seasons and experiments, which is documented by the following studies, where the presented database was successfully applied:

- A molecular study on essential derivation in maize (Heckenberger et al. 2006)
- Two studies on the extent of linkage disequilibrium in maize (Stich et al. 2005; Stich et al. 2006, b)
- A study on the development of a new pedigree-based QTL-mapping method (Stich et al. 2006, b)
- Several studies on the optimization of phenotypic analyses in field trials (Piepho et al. 2006, 2007; Piepho and Möhring 2006)

Availability

The Plabsoft database is running under the Linux operating system. Database access from applications under Microsoft Windows can be established using the open-database-connectivity (ODBC) interface. Licensing of the Plabsoft database is planned after the project end of the GABI-BRAIN project.

Acknowledgements The development of the Plabsoft database was funded by the German Federal Ministry of Education and research, Grant No. (0313126) "GABI-BRAIN" within the German genome research initiative "GABI".

References

Agrobase a database management and analysis system for agronomists, plant breeders, and plant researchers (http://www.agronomix.mb.ca/) Cited, 30.11.2006

Heckenberger M, Muminović J, Büchse A, Frisch M, Maurer HP, Melchinger AE, Möhring J, Piepho H-P, Reif JC, Stich B, Utz HF, Lichert F, Braun A, Breun J, Dreyer F, Ebmeyer E, Knopf E, Lübeck J, Schechert A, Stelling D, Streng S, Zacharias A, Wortmann H (2006) GABI-BRAIN: development and implementation of innovative statistical concepts and computational tools for integrating genomics research and applied plant breeding programs. Poster presented at the status seminar of the GABI-Community, 15/16th Feb 2006. Postdam, Germany

Heckenberger M, Muminović J, Van Der Voort JR, Peleman J, Bohn M, Melchinger AE (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 Breed 17:111–125

Lee JM, Davenport GF, Marshall D, Ellis THN, Ambrose MJ, Dicks J, Van Hintum TJL, Flavell AJ (2005) GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. Plant Physiol 139:619–631

Maurer HP, Melchinger AE, Frisch M (2007) Population genetical simulation and data analysis with Plabsoft. Euphytica (in press). doi:10.1007/s10681-007-9493-7

Piepho HP, Büchse A, Truberg B (2006) On the use of multiple lattice designs and α-designs in plant breeding trials. Plant Breed 125:523–528

Piepho HP, Möhring J, Melchinger AE, Büchse A (2007) BLUP for phenotypic selection in plant breeding and variety testing. Euphytica (in press). doi:10.1007/s10681-007-9449-8

Piepho HP, Möhring J (2006) On weighting in two-stage analysis of series of experiments. Biuletyn Oceny Odmian (awaiting approval by collaborating breeders of GABI-BRAIN)

PostgreSQL – (http://www.postgresql.org/) Cited 30. Nov (2006)

R Development Core Team (2004) R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria, ISBN 3–900051-07-0, URL http://www.R-project.org

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, Van Der Voort JR, Peleman J, Sorensen AP, Reif JC (2006) 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 H, Reif J (2006) A new test for family-based association mapping with inbred lines from plant breeding programs. Theor Appl Genet 113:1121–1130

Utz HF (2001) Plabstat, ein Computerprogramm zur statistischen Analyse von pflanzenzüchterischen Experimenten, Version 2P vom 14. Juli 2001 (in German). Universität Hohenheim, Germany

Zhao W, Canaran P, Jurkuta R, Fulton T, Glaubitz J, Buckler E, Doebley J, Gaut B, Goodman M, Holland J, Kresovich S, McMullen M, Stein L, Ware D (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic acids res 34:752–757