Database, Vol. 2011, Article ID baq038, doi:10.1093/database/baq038

Original article

A genotypic and phenotypic information source for marker-assisted selection of cereals: the CEREALAB database

Justyna Milc[1,*], Antonio Sala[2], Sonia Bergamaschi[2] and Nicola Pecchioni[1]

[1]Department of Agricultural and Food Sciences, University of Modena and Reggio Emilia, via G. Amendola 2, 42122 Reggio Emilia, Italy and [2]Department of Information Engineering, University of Modena and Reggio Emilia, via Vignolese 905, 41125 Modena, Italy

***Corresponding author:** Tel: +39 0522522064; Fax: +39 0522522027; Email: justynaanna.milc@unimore.it

Submitted 20 September 2010; Revised 16 December 2010; Accepted 17 December 2010

The CEREALAB database aims to store genotypic and phenotypic data obtained by the CEREALAB project and to integrate them with already existing data sources in order to create a tool for plant breeders and geneticists. The database can help them in unravelling the genetics of economically important phenotypic traits; in identifying and choosing molecular markers associated to key traits; and in choosing the desired parentals for breeding programs. The database is divided into three sub-schemas corresponding to the species of interest: wheat, barley and rice; each sub-schema is then divided into two sub-ontologies, regarding genotypic and phenotypic data, respectively.
Database URL: http://www.cerealab.unimore.it/jws/cerealab.jnlp

Introduction

The CEREALAB database was created and developed as one of the objectives of the CEREALAB and SITEIA projects, funded by Emilia-Romagna (Italy) regional government, and aims to increase the competitiveness of Regional seed companies through the use of modern selection technologies, i.e. the Marker-Assisted Selection (MAS).

The mission of a seed company is the release, through the process of 'plant breeding', that in synthesis is based on the deployment of genetic variation through selection, of new, improved, plant varieties, to be sold to the farmers. In turn, the release of new cereal varieties is an important target for many seed companies and public institutions in the world, to cope with the increasing food demand. In fact, the most important objective of cereal breeding programs all over the world is to combine in the new varieties high grain yield together with high quality and disease resistance.

In conventional plant breeding, the phenotype of a plant is visually assessed by different measuring (e.g. yield, resistance) and the best-performing plants are chosen. The conventional breeding is thus conducted by direct selection of the traits of interest, within populations of genotypes that contain an appropriate source of genetic variation for those traits, without knowledge of the gene-to-phenotype details. Moreover, many important agricultural traits, including yield, are so-called quantitative traits whose variation is under polygenic control with considerable environmental influence and genotype-by-environment interaction on trait expression. Such traits are the most difficult to breed for, typically requiring large-scale, multi-environment testing in order to obtain sufficient progress from selection. In most cases the successful breeding is thus an outcome from experienced breeder work performed using proven breeding methods.

Although conventional breeding practices have been very successful in producing a large number of improved varieties, the developments in the field of biotechnology and molecular biology can be employed to enhance plant breeding efforts and to speed up the obtention of cultivars by changing the object of the selection from 'the best phenotype' directly to 'the best genotype', independently

© The Author(s) 2011. Published by Oxford University Press.
This is Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.5), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

Database, Vol. 2011, Article ID baq038, doi:10.1093/database/baq038

from the environment, especially to aid selection for quantitative traits. This is the concept at the basis of the MAS, for which the best genotypes to be selected are identified by their specific profiles (alleles) obtained in laboratory thanks to molecular markers. Molecular markers have thus become an important tool with a great impact on plant assisted breeding and with several other areas of application, such as genetic fingerprinting for plant variety protection or for agri-food traceability programs (1).

Because of the growing number of MAS programs applied to various species, the demand for the storage of genotyping data is constantly increasing, due to the speed at which numerous markers, especially those PCR-based, are being developed and applied. Then MAS requires the breeder to have access to different kind of data, which usually either are not organized in databases, or reside in many different sites. Accessing, understanding and composing these data usually require long time and deep domain knowledge as the user has to query different databases and to combine manually the data he/she obtains. Especially for conventional breeders that lack a thorough preparation as far as molecular biology is concerned, handling of such a great quantity of information may be difficult.

The already available online genomic information sources for the crop research community include for instance MaizeGDB database for maize (2), GrainGenes (3) for wheats, barley, rye and oat, and Gramene (4) for grasses. Those databases oriented more towards basic research are however not perfectly suited for the breeders' activity, as either they do not include both genotypic and phenotypic data, or genomic and phenotypic information is managed separately, with no possibility for the user to correlate it. Moreover, often the applicability to breeding of molecular data contained is limited, since they do not show which plant genotypes contain which specific profiles (alleles) of a given molecular marker (i.e. marker genotyping outputs).

Recently some databases designed to store and manage both phenotypic and genotyping data have been reported: Germinate (5), Panzea (6), AppleBreed (7), PlantDB (8) and IWIS (9). Panzea is specific for maize, and allows an advanced Genomic Diversity and Phenotype Connection (GDPC) search. AppleBreed, specific for perennial crops, allows the management of pluri-annual data on the same individual plants, and it supports apple breeders and geneticists in their genetic studies and in their exploration of germplasm collections also for trait/marker associations, being sufficiently generic to allow it to be used for other perennial crops (7). PlantDB is a simple and versatile MS Access-based downloadable database designed to manage plant genetic resource collections, although it is mostly focused on experimental and research purposes (8). Germinate is a generic plant data management system, downloadable under the terms of the GNU public license, and designed to hold passport data and a range of additional data types including molecular markers, for different plant species (5). The International Wheat Information System™ (IWIS™) manages and integrates diverse information for wheats, triticale and barley in a single database (9) that now is migrated to International Crop Information System (ICIS) platform which facilitates distribution and integration of data. Those databases are often designed to store the experimental data and the data available are generally restricted to those implemented by the developers/users with no possibility to take advantage of already available information that resides in other data sources. Moreover, such a kind of database specific for breeding of wheat, barley and rice, fundamental crops for the world agriculture, still does not exist.The experimental data obtained from different research projects are often stored in files generated with different spreadsheet softwares. Those data are repeatedly stored in different locations and a considerable time is needed to convert data stored in a certain format into another format that can be input in larger databases, and combining data from different experiments for a joint analysis is difficult. The general goal of the scientific community should be to favour the exchange of information and to integrate all the data available as much as possible in order to reduce the redundancy. What is needed to solve these problems is the definition of methods for extracting and fusing the information coming from different (and heterogeneous) information sources (e.g. web sites and web databases) and subsequent presentation of the information according to a unique interface.

In order to help the users to better exploit the great amount of data already available, and to provide a unique access to all the information needed by wheat, barley and rice cereal breeders to perform MAS, the CEREALAB database was realized. The objective of the database is to combine data obtained by genotyping activity of the CEREALAB research group with the already available molecular, genotypic and phenotypic data in order to create a tool for breeders. This integration is obtained using the MOMIS system (Mediator envirOnment for Multiple Information Sources), a framework developed by the Database Group of the University of Modena and Reggio Emilia that performs information extraction and integration from both structured and semistructured data sources allowing to query the information in a transparent way for the user regardless of the specific languages of the sources. MOMIS maintains the integrated databases mutually independent, performing an on the fly integration of the data only when necessary, i.e. when a query is posed on the CEREALAB database.

Thanks to the way in which the database is organized, it can help the breeders in unravelling the genetics of

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

economically important phenotypic traits; in identifying and choosing molecular markers associated to key traits; in identifying alleles of such markers associated to trait positive variants; and in choosing the desired parentals for breeding programs.

Methods

Database outline and content

Since the database was developed as one of the outputs of the CEREALAB regional project, its design and content followed strictly the needs expressed by the breeders of the regional seed companies, partners of the project and focused on bread and durum wheat, barley and rice variety release. As one of the activities of CEREALAB consisted in genotyping the collections of varieties grown or used in breeding programs in Italy, it was necessary to create a tool to distribute those results to their end-users, considering especially that only very few data had been available on molecular marker characterization of Italian varieties till now.

The idea of the CEREALAB database developers was to enable the breeder to find directly in the database the variety of his interest to be used in the breeding program he conducts. Starting his query with the trait of interest he should find varieties genotyped for molecular markers associated with genes and QTLs that govern that trait, and choose the variety that harbours the high-scoring allele. In alternative, he should find a molecular marker associated with a gene/QTL underlying the trait of his interest and use it in his breeding program in order to speed up the selection. As a first step the breeders were asked what characters they were most interested in; the characters chosen were then grouped to form six main super-classes following the trait ontology (abiotic stress, biotic stress, growth and development, anatomy and morphology, quality and yield).

Since bread wheat (*Triticum aestivum*) is an allohexaploid species, consisting of three sub-genomes (A, B and D) and shares the former two sub-genomes with the allotetraploid durum wheat (*Triticum turgidum ssp. durum*) (10), it was decided to merge the data regarding those two species and organize the database into three main sections corresponding to the genus *Triticum*, *Hordeum* and *Oryza.*

Data sources

To create the CEREALAB database, first a relational database was created with data obtained from a systematic genotyping work performed by the research group of the CEREALAB laboratory. This database was then integrated with other, already available sources of molecular and phenotypic data that have been chosen among those particularly relevant for MAS of wheats (*T. aestivum* and *T. turgidum ssp. durum*), barley (*Hordeum vulgare*) and rice (*Oryza sativa*). The choice was also driven by the need of having easily access to the data to be integrated: as it will be discussed in the next section, the preference was given to sources in the form of relational databases that provide some sort of public access directly to the data, not only through a web interface. This starting set of sources can easily be extended with other sources as they become available, including other databases that were not taken into consideration initially. As far as molecular data are concerned, the Gramene database (http://www.gramene.org, release 23) was used as a data source for rice, and GrainGenes (http://www.graingenes.org) for wheat and barley.

GrainGenes, created by the United States Department of Agriculture, Agricultural Research Service (USDA-ARS), is the internationally recognized database for wheats, barley, rye and oats. For these species it constitutes the main and primary repository for information about genetic maps, genes, QTLs, markers, primers and alleles (3). Gramene is a comparative genome mapping database for grasses (*Poaceae*) that uses the rice genome as an anchor, taking advantage of the known genetic colinearity between rice and major crop plant genomes, to provide researches that deal with other species with the benefits of an annotated genome. It combines and interrelates information on the structure and organization of genomes and genes, functions of proteins, various maps (genetic, physical and sequence), mapped markers, quantitative trait loci (QTLs) and literature citations (4).

Gramene is available online and allows downloading the dump of part of or the complete relational database. We exploited this possibility to create its local copy to be integrated in CEREALAB, avoiding possible bottlenecks caused by having a remote data source. Anyway, the data source is maintained completely independent and no modification was made on it. Differently from Gramene, GrainGenes cannot be downloaded entirely, but this was made possible by contacting the curators. Therefore, GrainGenes was restored locally on the server as well, still maintaining it independent.

As far as phenotypic evaluations are regarded, four data sources were used. Germplasm Resources Information Network (GRIN) data for wheat and barley descriptors were used for international locations, while public data of the Agricultural Research Council (CRA) field trials for wheat, barley and rice, public data of the Emilia-Romagna variety field trials for wheat and barley, and finally public data of the Ente Nazionale Risi (www.enterisi.it) field trials for rice were used for Italian locations.

The Germplasm Resources Information Network (GRIN) of the USDA-ARS provides germplasm users with continuous access to databases for the maintenance of passport, characterization, evaluation, inventory and distribution data important for the effective management and

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

Database, Vol. 2011, Article ID baq038, doi:10.1093/database/baq038

utilization of germplasm collections. The trait descriptors list with observed values for each species can be accessed and downloaded from the section 'Research Crops and Descriptor/Evaluation Data Queries' at www.ars-grin.gov. The data of Agricultural Research Council, Ente Nazionale Risi and Emilia-Romagna field trial data were all entered manually using Excel data spreadsheet files.

Integration process details

We exploited the Mediator envirOnment for Multiple Information Sources (MOMIS) system; a mediator-based system developed by the Database Group of the University of Modena and Reggio Emilia, capable of building from scratch a knowledge base that integrates schemas and data extracted from a set of data sources and provides a Global Schema (GS) from the integrated sources. The choice to integrate existing data was driven by the fact that part of the necessary data was already available to the research community. MOMIS performs integration of structured and semi-structured data sources, such as relational databases, XML/XSD or Excel files in a semi-automatic way. It exploits semantic annotations of the elements appearing in the data sources to group semantically related concepts appearing in the different data sources. Mapping tables are automatically created to handle heterogeneity and data conversion functions and resolution functions can be added to solve data conflicts. The result is a GS that is a virtual view of the underlying sources composed of global classes and global attributes. Starting from the annotations of the local sources, each global class and attribute is automatically annotated, thus, each schema element has a well-understood meaning. In this way the GS can be seen as an ontology emerging from the source schemas. Further information about the MOMIS methodology can be found in (11, 12). The MOMIS system allows integrating distributed data sources that can be accessed remotely, since the only requirement is having read access to the data. In fact, its virtual approach assures the independence of the data sources thus not affecting them for the integration purpose.

As already said, the integration process relies on semantic annotations of the schema elements of the data sources. This process in MOMIS exploits WordNet as a lexical resource, but other resources can be used, such as OWL ontologies. In the case of WordNet, the lexical resource can be extended by means of a WordNet Editor implemented in the system that allows adding new terms and relationships between terms. As in our case the annotation is related to a very specific domain, WordNet lacks many of the terms involved. WordNet was thus widely extended to face this new domain adding new lemmas and their relationships with the terms already present in WordNet. This extension process can then be used to annotate other data sources of the same domain, since many terms are repeated in the sources. The advantage is that this extension step has to be performed just the first time a domain is handled: once a set of annotations that are specific to a particular domain has been created, it makes easier extending the integration process to include other data sources that were not taken into consideration initially. In order to ease the integration process, a first level of views on the molecular data sources was created to aggregate information regarding their main concepts like gene, QTL or marker. This first level of views leads to a denormalization of the data sources, but turned out to be particularly helpful to simplify the GS and to make querying the GS easier since all the properties of important concepts are grouped in the same global class.

Results

Conceptual data model: the CEREALAB ontology

The result of the integration process can be seen as ontology of the underlying data sources. This ontology is not defined manually but is automatically obtained exploiting the semantics of the annotation of the elements appearing in the data sources being integrated. The details about this process are further described in ref. (12). The valuable element of this ontology is the combination of existing phenotypic and genotypic data sources with data coming from the CEREALAB project.

Each sub-schema (*Triticum*, *Hordeum* and *Oryza*) is divided in two parts or sub-ontologies: GENOTYPIC data and PHENOTYPIC data. The main classes of the sub-ontology GENOTYPIC data are Trait, Gene, QTL and Marker. The most appropriate definitions were adopted; a trait is considered as any single feature or quantifiable measurement of an organism. A gene is a hereditary unit consisting of a sequence of DNA that occupies a specific location on a chromosome and determines a particular characteristic in an organism. A QTL is a statistical construct that identifies a particular region of the genome as putatively containing one or more genes associated with a trait. A QTL is represented as an interval in a genetic linkage group, within which the probability of association is plotted for each marker used in the mapping experiment. Molecular markers are specific fragments of DNA that can be identified within the whole genome. They are used to 'flag' the position in the genome of a particular gene, or of a particular anonymous sequence.

The attributes of the classes Trait, Gene, QTL and Marker store information about these entities. For example the information that can be retrieved for the class Gene includes its name, gene class in which it was classified, germplasm in which this gene was identified, chromosome on which it was mapped, its alleles, sequence where available and the bibliographic reference. For the class Marker the main

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

information available includes: marker type (e.g. Microsatellite, RFLP), the PCR protocol with amplification conditions and primer sequences.

Moreover additional information is retrieved from the following relationships:

(i) The information which gene or QTL underlie a certain trait can be found in the classes Trait_affected_by_Gene and Trait_affected_by_Qtl.
(ii) For each gene and QTL it is possible to retrieve the germplasm in which it was identified in the classes Gene_in_Germplasm and QTL_in_Germplasm, respectively. Those classes are reifications of the relationships expressing the identification of a gene or QTL in a Germplasm.
(iii) The relationships among markers and genes and QTLs are reified in the classes Marker_for_Gene and Marker_for_QTL, respectively.
(iv) Moreover, information about which germplasm the markers have been tested on (resulting from the genotyping activity conducted within the CEREALAB project) are available in the class Marker_tested_on_Germplasm.

The PHENOTYPIC data sub-ontology contains field evaluations data for each trait. The data are divided into six super-classes of major interest for the breeders: ABIOTIC STRESS, BIOTIC STRESS, ANATOMY and MOPHOLOGY, GROWTH and DEVELOPMENT, QUALITY and YIELD. Each super-class is divided into several sub-classes, e.g. BIOTIC STRESS has been divided into 14 sub-classes, such as (resistance to) stem rust, leaf rust, powdery mildew, fusarium head blight (FHB) and others. Most classes are further divided according to different distinct phenotypic data sources used (GRIN for international, Emilia-Romagna, CRA and Ente Nazionale Risi evaluation data for Italian germplasm collections). The schema of the CEREALAB integrated ontology is presented in Figure 1.

The database content

The cereal broad-based germplasm collections maintained by the CEREALAB laboratory comprise about 400 rice, 70 barley and 150 bread wheat accessions, containing a significant sample of genetic variation. For instance the bread wheat collection comprises 70 varieties grown and used in breeding programs in Italy and 80 varieties and landraces obtained from Germplasm Resources Information Network (GRIN) of the USDA-ARS. The GRIN varieties were chosen on basis of their origin (Asia, North and South America and Eastern Europe) and reported resistance to *Fusarium spp*. The genotyping and discovery work packages of the CEREALAB project consisted in genotyping (using already known markers and some new protocols) the collection of germplasm with markers associated with resistance to the most common pathogens like *Fusarium*, *Septoria tritici*, *Magnaporthe grisea* and *Blumeria graminis*. All genotyping data produced by the CEREALAB project are available in the class Marker_tested_on_Germplasm. This application can be in the future extended to any genotyping data we are able to retrieve from the scientific literature, collaborating laboratories or from other data sources, in order to provide the breeder with the molecular data of other than CEREALAB germplasm collections.

For the breeders activity of crossing and selection it is necessary and fundamental to know the effective phenotypic performance of the variety in the field and this can be obtained by consulting the evaluation data for such a variety in one or more locations and in one or more years. The CEREALAB source includes thus phenotypic data from different sources, and it should be updated by adding varieties (genotypes), as well as new year and location records. As far as varieties grown in Italy are considered, several data sources available online for Italian field trials were used and the data were entered manually. The data of Agricultural Research Council (CRA) include trials 2004–05 in different regional locations in Italy; Emilia-Romagna data derive from observations made in 20 locations inside the Emilia-Romagna Region and are expressed as average values coming from different year evaluations. As breeders might want to include some new germplasm resources into their breeding programs it was decided to provide them with evaluation data available for Germplasm Resources Information Network (GRIN) germplasm broad collection. These data include observations for several thousand of accessions made in different locations and years. As different phenotypic data sources use different descriptors (due to different national standards), the 'Descriptors' files with a detailed explanation for each data source are available and can be downloaded from www.cerealab.org. Tutorials to get users started can be found in the database homepage as well.

To provide the users with all further information about molecular markers retrieved, the genes, loci and quantitative trait loci (QTLs) they are associated with, the CEREALAB database provide a direct accession to the several thousand of molecular data of GrainGenes and Gramene databases, to integrate the molecular marker data acquired in the genotyping work package of CEREALAB. Integration of already existing data sources permits to reuse pre-existing data with a considerable time-saving. New molecular markers and new protocols, not available in GrainGenes and Gramene, were entered manually following a specified format.

Querying CEREALAB: graphical interface

The CEREALAB database can be queried by means of a Graphical User Interface (GUI) available as a Java Web

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

Database, Vol. 2011, Article ID baq038, doi:10.1093/database/baq038

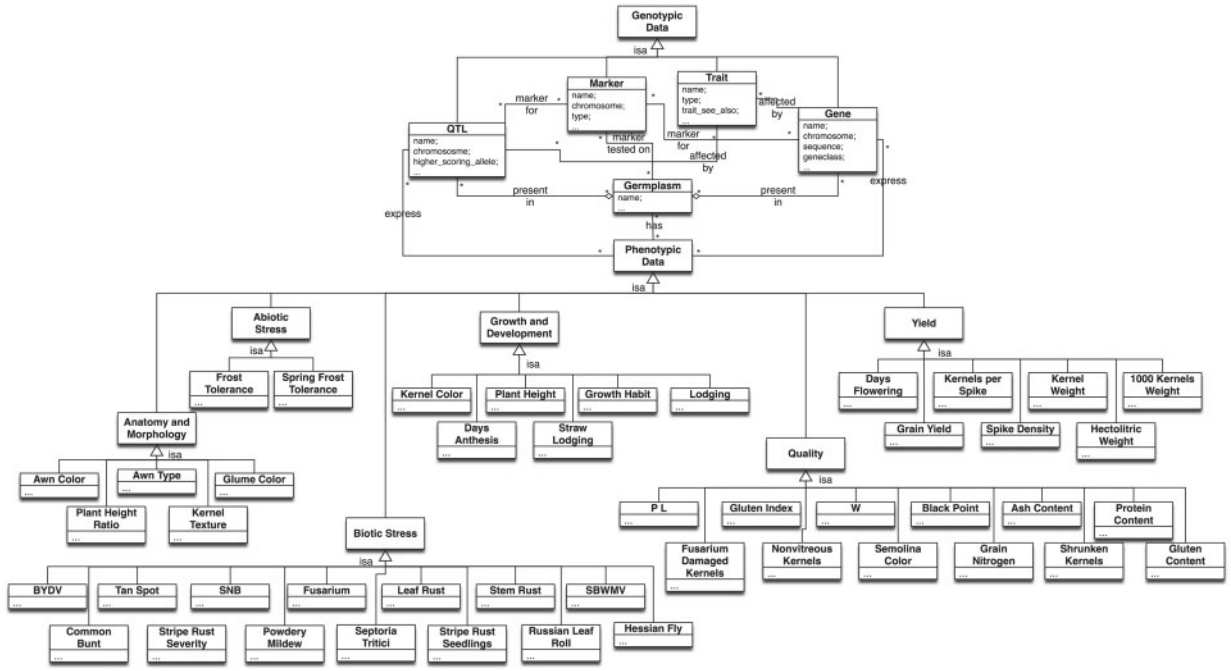

Figure 1. The schema of the CEREALAB database integrated ontology.

Start application at www.cerealab.org. The user has to access the database from the database section of the web page. Choosing 'enter the database' opens the Java application. To start querying the CEREALAB database it is necessary to choose from the 'Menu' the schema we want to query among the three schemas available: *Triticum, Hordeum* or *Oryza*. Once the schema has been loaded, the GS is presented in a tree representation on the left panel, showing ISA (Is-a) relationships among the classes. The user can select the global class to be queried and its attributes are shown in the 'global class attributes' panel with usual relational predicates and logic operators. From this panel, the attributes of interest can be selected, specifying, if necessary, a condition in the 'condition' panel with usual relational predicates and logic operators. More than one global class can be joined just choosing one of the 'referenced classes' of the currently selected class with no need to specify any join condition between the classes as it is automatically inserted. The graphical query, including selections and conditions specified by the user, is then automatically translated into an SQL query and sent to the MOMIS query manager which rewrites it as an equivalent set of queries to be executed on local data sources, and merges their results in a unified answer to be presented to the user [see ref. (13) for a technical description].

The following sample query shows how a breeder interested in yield-related traits can find in the *Triticum* schema the varieties characterized by a 1000 kernel weight (weight in grams of 1000 well-developed whole grains) value of his interest. Once he obtains a list of varieties the user can check which of them have been genotyped for instance with molecular markers associated with QTLs/genes that determine resistance to certain pathogens, in order to choose those that harbour the high-scoring allele. To do this the user has to select the sub-ontology PHENOTYPIC DATA, then YIELD and 1000_Kernel_Weight from the tree representing the GS. All the attributes of 1000_Kernel_Weight are shown in the 'global class attributes' tree. Then, if the user adds to the selection the 'Referenced Class' Marker_Tested_on_Germplasm all the attributes of this class are then automatically added to the 'Global Class Attributes' panel and the user may select attributes from this global class. The user can choose to retrieve the phenotypic evaluation data from either all or only one data source (only selected attributes are added to the query). Then, if he is interested only in varieties that have 1000 kernel weight value higher than 20 g he just has to add in the 'condition' panel the condition '>20' for this attribute and click 'add condition'. Clicking the button 'execute query', all the phenotypic data available for cultivars that meet the criteria are retrieved including field evaluations, repository and pedigree data when available. The molecular data retrieved comprise: QTL or gene, associated marker and genotyping data from CEREALAB project work. In this way the user can choose from the varieties with satisfactory 1000 kernel weight those that have been already genotyped for molecular markers associated with resistance and that harbour the high-scoring allele.

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

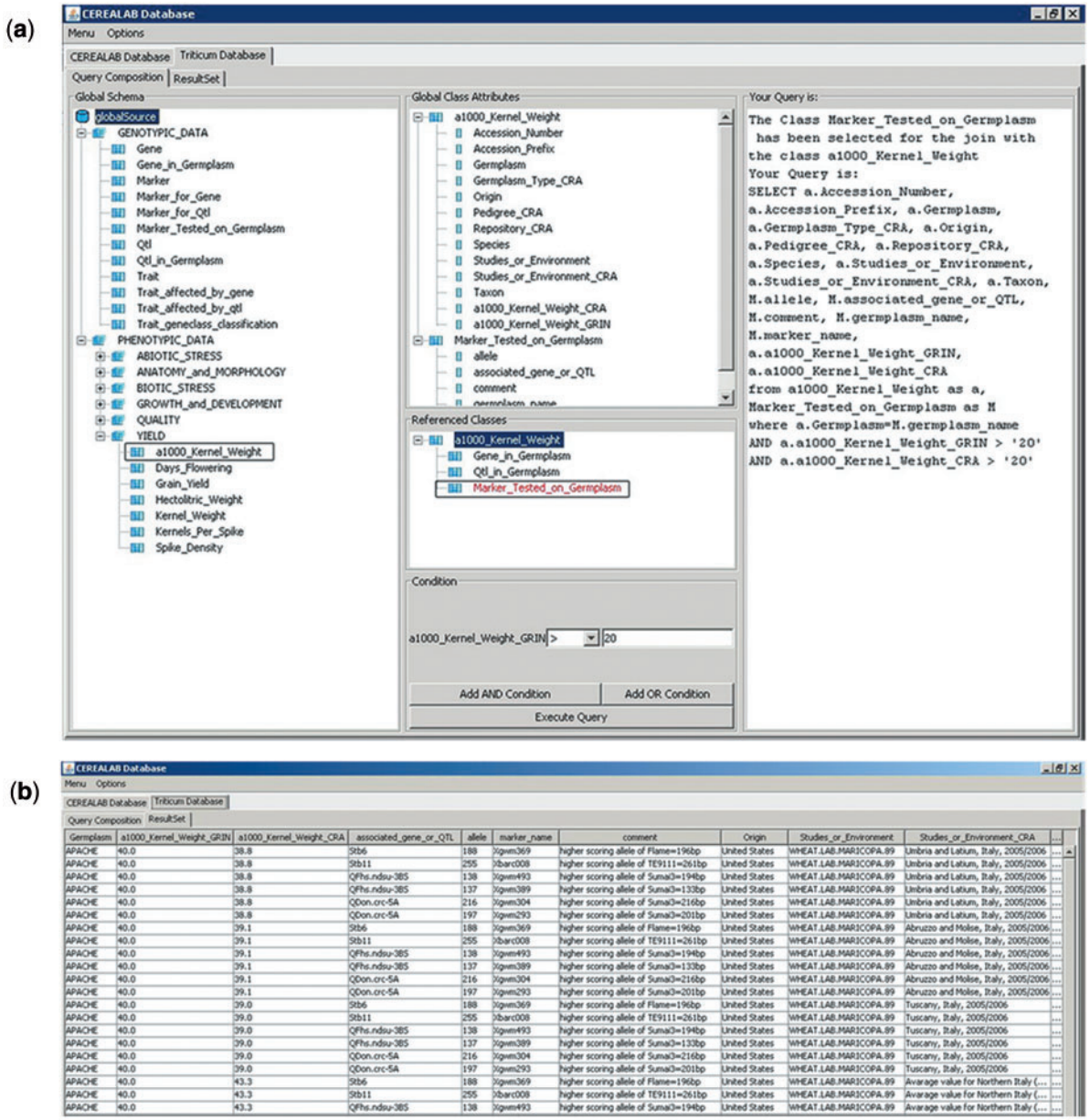

Figure 2. Screenshots of the query example (**a**) and of the results (**b**) returning information about wheat varieties characterized by 1000 kernel weight value >20 g and genotyped with molecular markers associated with genes and QTLs for resistance to major pathogens.

Otherwise if the breeder is interested only in a given variety (for instance '*Eureka*' or '*Bolero*') he can restrict the query only to those two cultivars by adding in the 'condition' panel the condition: 'germplasm' like '*Eureka*' ('add condition') and then '*Bolero*' ('add AND condition'). The query example and its results are presented in Figure 2a and b, respectively.

Data loading in the CEREALAB database

The virtual approach offered by MOMIS facilitates the updating process of the CEREALAB database. In fact, queries are executed directly on the data sources that are integrated in the CEREALAB database, thus providing always up-to-date information. In our approach, the two sources Gramene and GrainGenes have been locally stored on the CEREALAB server, but their independence has been maintained. It is thus possible to update these two sources just downloading and restoring their new releases. The data obtained by the CEREALAB research group are loaded in the CEREALAB data source that is integrated with the other databases, by means of Excel spreadsheets that follow a specified format.

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

Database, Vol. 2011, Article ID baq038, doi:10.1093/database/baq038

A benefit of integrating other data sources with the CEREALAB data is represented by the possibility of exploiting existing information when loading the data: only the additional information has to be inserted, since all the existing data are obtained by the other sources. As an example, let us suppose that a marker already exists in Gramene, but is being tested on a new germplasm in the context of the CEREALAB project: it is not necessary to load all the data about the marker (such as genomic position, primers, PCR protocol, etc.) but only the additional information about the germplasm the marker was tested on. This possibility eases the loading process and is time-saving.

Implementation details

The MOMIS system is entirely implemented in Java, and is available as a Java Web Start application, that guarantees compatibility with all the main operating systems. In particular, the GUI of the CEREALAB database is a standalone Java Web Start application, available at: http://www.cerealab.unimore.it/jws/cerealab.jnlp.

The MOMIS system is currently open source and is available under the terms of GNU General Public License, version 2 at: http://www.datariver.it. This means that the MOMIS system can either be used in other projects for the same purpose of integrating different data sources, or it can be used by other research groups that are interested in contributing to the CEREALAB database by simply sharing with them the set up files (the GS) that can be requested for scientific purposes.

Discussion

CEREALAB database was initially created as a tool to introduce the modern selection technologies of MAS to regional seed companies. The objective of the CEREALAB database developers was thus to combine data obtained by genotyping activity of the CEREALAB research group with the already available molecular and phenotyping data in order to create a unique site and tool for wheat, barley and rice breeders. The database as it is realized, and in English language, can have a broader platform of users, like public researchers and breeders, not only private, and regional or national. Its maintainance and further improvement is ensured at least until 2015 by a follow-up project funded by the same agency, and it is intention of the authors to keep it updated and usable in the long term.

Although wheats and barley are genetically and genomically related, belonging to the tribe of *Triticeae* (10), and in fact they are kept together in the GrainGenes database, it was decided to keep the database divided into three sections corresponding to the three crops for the intrinsic applied nature and aim of the CEREALAB database. In fact breeders use information and make crosses and selection, also marker-assisted, within barley, within rice and within wheats, without ever mixing these three breeding processes.

To date, some other databases such as AppleBreed, Panzea, Germinate and PlantDB have been designed to provide a way to store and correlate genotypic to phenotypic information on plant genetic collections, including genetic stocks and commercial varieties. While AppleBreed is specifically designed for perennial (tree) species, and PlantDB helps managing research and experimental activity in plant science laboratories, Panzea and Germinate aim at more applied utilization for herbaceous crops like cereals.

As far as the existing sources for molecular data as GrainGenes and Gramene are concerned, they are good and extremely comprehensive tools, oriented more towards basic research that are however more suitable for geneticists or for those breeders that are more used to deal with molecular data.

Respect to the existing databases, the unique way in which the CEREALAB database is designed enables an intensive cross-talk between all the integrated data sources. In fact, the MOMIS system was exploited to provide an integrated GS of data sources already existing, to join the data produced by the CEREALAB laboratory, to facilitate breeders in better approaching the large resources available for molecular breeding. It represents a first attempt of simplification of different sources with the mainly applied purpose of giving a tool for marker-assisted breeding. The user can retrieve from one site all essential data. Starting his query with the trait of interest, the user of CEREALAB can find varieties genotyped for molecular markers associated with genes and QTLs that govern that trait, and choose the variety that harbours the positive (high-scoring) allele.

The CEREALAB database is, as far as the authors know, the first database that enables querying both molecular and phenotypic data specific for all three major crop species; wheat, barley and rice. It supports mainly breeders (but it may be used by geneticists as well) in their exploration of germplasm available in various collections. As far as phenotypic data are concerned, we included for example four different data sources to enable the breeder to compare the performance of a variety in different world-wide environments. In this view, it is also the first resource that is releasing publically and in a systematic way an access to the phenotypic data of the Italian germplasms.

The information stored in the database should help the breeders not only to understand the genetics of economically important traits and to identify marker-trait associations, but what's most important, enable them to choose the most interesting genotypes that should be crossed to obtain a desired progeny.

A benefit of integrating different data sources with the CEREALAB data is also represented by the possibility of

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011

exploiting existing information when loading the data; only the additional information has to be inserted, since all the existing data are obtained by the other resources.

Our model approach has been tested on a selected set of data sources to create the CEREALAB database, but it can potentially be applied to every data source related to the domain under consideration to enrich the knowledge available, as well as to other economically important species where genomic and phenotypic databases exist. It means that in future, the information regarding species already included in the database can be enriched as soon as new data are available and the number of species can be extended in a flexible way, to follow the needs of the end-users.

Perspectives

One of the main goals for the future of CEREALAB database is the development of an user-friendly interface to the database, together with identification and construction of some additional tools to maximize and optimize the accessibility of the available information (such as shortcuts to most important protocols) as soon as needs of the end-user are further defined.

The experience acquired creating the CEREALAB GS for wheats, barley and rice can be also exploited with other species and by other research groups. In fact the MOMIS approach can easily be applied to similar schemas concerning other plant and animal species.

Acknowledgements

We thank GrainGenes (http://www.graingenes.org) for the permission to use their datasets, Gramene, (http://www.gramene.org) for a subset of data included in Release 23 and USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network (GRIN) [online Database]. National Germplasm Resources Laboratory, Beltsville, Maryland (http://www.ars-grin.gov) for selected data sets of phenotypic data. Authors wish to thank all partners of the CEREALAB laboratory for their suggestions about the database structure and queries and for the critical discussion of the obtained results. Authors acknowledge that part of the phenotypic data were derived from publicly available online sources of Emilia- Romagna government, CRA and Ente Nazionale Risi.

Funding

Projects P.R.R.I.I.T.T. Misura 3.4—Azione A 'CEREALAB'; 'SITEIA', of the Regione Emilia-Romagna, Assessorato alle Attività Produttive, Italy. The improvement and curation of the database will proceed within BIOGEST-SITEIA Tecnolopo of the High Technology Network of the Regione Emilia-Romagna, Italy. Funding for open access charge: Project P.R.R.I.I.T.T. Misura 3.4-Azione A 'SITEIA' of the Regione Emilia-Romagna, Assessorato alle Attivitá Produttive, Italy.

Conflict of interest. None declared.

References

1. Cooper,M. and Hammer,G.L. (2005) Preface to special issue: complex traits and plant breeding - can we understand the complexities of gene-to-phenotype relationships and use such knowledge to enhance plant breeding outcomes? *Aust. J. Agric. Res.*, **56**, 869–872.
2. Lawrence,C.J., Dong,Q., Polacco,M.L. *et al*. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res.*, **32** (Database issue), D393–D397.
3. Matthews,D.E., Carollo,V.L., Lazo,G.R. and Anderson,O.D. (2003) GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.*, **31**, 183–186.
4. Jaiswal,P., Ni,J., Yap,I. *et al*. (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.*, **34** (Database issue), D717–D723.
5. Lee,J.M., Davenport,G.F., Marshall,D. *et al*. (2005) GERMINATE. A generic database for integrating genotypic and phenotypic information for plant genetic resource collections. *Plant Physiol.*, **139**, 619–631.
6. Zhao,W., Canaran,P., Jurkuta,R. *et al*. (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.*, **34 (**Database issue), D752–D757.
7. Antofie,A., Lateur,M., Oger,R. *et al*. (2007) A new versatile database created for geneticists and breeders to link molecular and phenotypic data in perennial crops: the AppleBreed DataBase. *Bioinformatics*, **23**, 882–891.
8. Exner,V., Hirsch-Hoffmann,M., Gruissem,W. and Hennig,L. (2008) PlantDB - a versatile database for managing plant research. *Plant Methods*, **4**, 1.
9. Payne,T.S., Skovmand,B., Lopez,C.G. *et al*. (eds). (2002) *The International Wheat Information System (IWIS^{TM}), Version 4, 2001*, On compact disk. CIMMYT, Mexico, D.F.
10. Stein,N. (2007) Triticeae genomics: advances in sequence analysis of large genome cereal crops. *Chromosome Res.*, **15**, 21–31.
11. Beneventano,D., Bergamaschi,S., Guerra,F. and Vincini,M. (2003) Synthesizing an integrated ontology. *IEEE Internet Comput.*, **7**, 42–51.
12. Sala,A. and Bergamaschi,S. (2009) A mediator based approach to ontology generation and querying of molecular and phenotypic cereals data. *Int. J. Metadata, Semantics and Ontologies*, **4**, 85–92.
13. Beneventano,D. and Bergamaschi,S. (2007) Semantic search engines based on data integration systems. In: Cardoso,J. (ed). *Semantic Web Services: Theory, Tools and Applications*. Idea Group Publishing, Hershey, PA, pp. 317–342.

Downloaded from database.oxfordjournals.org at Universitàdegli Studi di Modenae Reggio Emilia on January 20, 2011