A new versatile database created for geneticists and breeders to link molecular and phenotypic data in perennial crops: the AppleBreed DataBase

A. Antofie, M. Lateur, R. Oger, Andrea Patocchi, Charles Eric Durel, W.E. van de Weg

BIOINFORMATICS ORIGINAL PAPER

Vol. 23 no. 7 2007, pages 882–891
doi:10.1093/bioinformatics/btm013

Databases and ontologies

A new versatile database created for geneticists and breeders to link molecular and phenotypic data in perennial crops: the *AppleBreed DataBase*

A. Antofie[1], M. Lateur[1], R. Oger[1,*], A. Patocchi[2], C. E. Durel[3] and W. E. Van de Weg[4]

[1]Walloon Agricultural Research Centre (CRA-W), Gembloux, Liroux 9, B-5030, Belgium, [2]Agroscope Changins-Wädenswil (ACW), Phytopathologie, P.O. Box 185, Schloss CH-8820 Wädenswil, Switzerland, [3]Genetics and Horticulture – GenHort, National Institute for Agricultural Research, 15 INRA, BP 60057, F-49071 Beaucouzé Cedex, France and [4]Plant Research International (PRI), P.O. Box 16, 6700 AA Wageningen, The Netherlands

Received on November 20, 2006; revised on January 11, 2007; accepted on January 12, 2007

Advance Access publication January 19, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Objective: *AppleBreed DataBase* (DB) aims to store genotypic and phenotypic data from multiple pedigree verified plant populations (crosses, breeding selections and commercial cultivars) so that they are easily accessible for geneticists and breeders. It will help in elucidating the genetics of economically important traits, in identifying molecular markers associated with agronomic traits, in allele mining and in choosing the best parental cultivars for breeding. It also provides high traceability of data over generations, years and localities. *AppleBreed DB* could serve as a generic database design for other perennial crops with long economic lifespans, long juvenile periods and clonal propagation.

Results: *AppleBreed DB* is organized as a relational database. The core element is the GENOTYPE entity, which has two sub-classes at the physical level: TREE and DNA-SAMPLE. This approach facilitates all links between plant material, phenotypic and molecular data. The entities TREE, DNA-SAMPLE, PHENOTYPE and MOLECULAR DATA allow multi-annual observations to be stored as individual samples of individual trees, even if the nature of these observations differs greatly (e.g. molecular data on parts of the apple genome, physico-chemical measurements of fruit quality traits, and evaluation of disease resistance). *AppleBreed DB* also includes synonyms for cultivars and pedigrees. Finally, it can be loaded and explored through the web, and comes with tools to present basic statistical overviews and with validation procedures for phenotypic and marker data to certify data quality.

AppleBreed DB was developed initially as a tool for scientists involved in apple genetics within the framework of the European project, 'High-quality Disease Resistance in Apples for Sustainable Agriculture' (HiDRAS), but it is also applicable to many other perennial crops.

Contact: oger@cra.wallonie.be

*To whom correspondence should be addressed.

1 INTRODUCTION

Breeding cultivated plants and, in particular, apple trees (*Malus x domestica* Borkh.), has always been an important activity at both the amateur and professional level. Compared with annual crops, breeding perennial crops is complex, long-lasting and time consuming due to their long juvenile phase and long economic lifespan. Cultivars of perennial crops are grown across large geographical areas and consequently new cultivars and breeding selections have to be evaluated over many successive years in various localities. The age of the trees has to be taken into account when phenotypically characterizing material because phenotypic characteristics change with age. Other characteristics give perennial crops some advantages in genetic research, such as their suitability for vegetative propagation. This means that many old cultivars and breeding selections still exist, which allows an identical genotype to be tested in various localities and in various years. In addition, the simultaneous presence of various successive generations within a single experiment is possible. All these specific characteristics affect breeding procedures and genetic experiments and put high demands on the storage of phenotypic data.

The demands for the storage of genotyping data is also increasing tremendously due to the pace at which high numbers of PCR-based molecular markers are being developed. Initially, studies on marker-trait associations were limited in size, usually involving just a single cross. The use of a single cross suffices as long as the genetic basis of a trait is extremely simple (only one locus with one + allele). In all other cases, multiple crosses are needed if sound conclusions are to be reached on the number of loci, alleles and mode of action of genes (intra- and/or inter-locus interactions). Studies on multiple crosses therefore demand high quality and good data management facilities.

In the perennial apple crop, a new concept of gene and QTL identification was initiated called Pedigree Genotyping (Van de Weg *et al.*, 2004). This approach aims to identify marker-gene associations, functional allelic diversity and both intra- and inter-locus interactions by the integrated analysis of multiple plant populations (crosses, breeding selections and commercial

cultivars) that are genetically related by their pedigree. The European project '*High-quality Disease Resistance in Apples for Sustainable Agriculture*' (HiDRAS) (Gianfranceschi and Soglio, 2004), was initiated to test the concept. In this study, more than 2000 genotypes are being extensively phenotyped and genotyped, delivering more than 1 million data points. Each phenotypic data point is associated with its own descriptors for tree, year, sample and locality. Each genetic data point is associated with its own descriptors for DNA sample, tree, genotype, marker and map position. To meet the needs for the storage and accessibility of these data, a database was needed. There are already several databases managing both genomic and phenotypic information for the plant kingdom. MaizeGDB database (Lawrence *et al*., 2004), for instance, is a repository for maize sequence, stock, phenotype, genotypic and karyotypic variation, as well as chromosomal mapping data The GrainGenes database (Matthews *et al*., 2003) focuses on grasses and cereals storing both genetic and phenotypic information. It holds, amongst others, the genealogy and allelic constitution of markers and genes from 69 632 wheat accessions. Other databases have been developed for managing genome molecular information (Rhee *et al*., 2003; Schoof *et al*., 2002) or for storing genes and protein information for *Arabidopsis thaliana* (ABRC, NASC, MATDB).

All these databases focus on annual plants and most of them manage genomic or phenotypic information separately. None of them allows the management of pluri-annual data on the same individual plants (Reiser *et al*., 2002; Sakata *et al*., 2000). As none of the existing public databases were able to support extensive studies on marker-trait associations in pedigreed populations of perennial crops, *AppleBreed DB* was developed. In the context of database construction, apples could serve as a model for perennial crops. Apples are a woody perennial and have a 3–7 year juvenile phase, which is a significant handicap in combining high fruit quality and durable disease resistance by classical breeding. Apples are self-incompatible due to a gametophytic incompatibility system, and therefore inbreeding methods are not applied (Lespinasse, 1992). Apples are vegetatively propagated, have an economic lifespan of about 15 years during which they produce 13 crops, are economically important and are highly rated among consumers, being ranked third in a fresh fruit consumption survey after banana and citrus (Pollack, 2001). Currently, there are more than 10 000 apple cultivars (Morgan and Richardson, 2002; Way *et al*., 1991) throughout the world. Nevertheless, world apple production is based on a handful of cultivars that are grown in commercial orchards. The most important commercial cultivars are highly susceptible to the most important apple diseases (scab, powdery mildew and European canker), and most of the resistant cultivars do not yet meet the quality demands of consumers. The most important objective of worldwide apple breeding programmes is therefore to combine high fruit quality with good disease resistance. To achieve this aim, breeders need a better understanding of the genetic basis of fruit quality traits and disease resistance, and to obtain access to molecular markers for the most important genes controlling these traits.

AppleBreed DB supports breeders and geneticists in their genetic studies and in their exploration of germplasm collections. Structured information stored in the database should help them not only to elucidate the genetics of complex traits and to assess marker-trait associations, but also to choose more easily and more quickly the most interesting genitors to cross with (e.g. with good disease resistance, a particular taste, or a skin colour preferred by consumers). In this way, it is expected that breeders will more easily be able to create new cultivars meeting consumer preferences and allowing sustainable production systems. This article describes the database model of *AppleBreed DB*. *AppleBreed DB* is sufficiently generic to allow it to be used as a model database for many other perennial crops.

2 METHODS

AppleBreed DB is intended to be a functional tool for geneticists and breeders. Its data model combines conceptual and logical data models. The aim is to represent the implementation level of information. The conceptual data model (CDM) was the first step in processing the database design, followed by the logical data model (LDM). The CDM includes the main entities/structures and the relationships among them, without specifying attributes or primary keys. Here, the highest levels in the relationships among the different entities are identified. The CDM is divided into super-classes, classes and sub-classes. By definition, a class includes entities characterized by the same profile (e.g. the class *marker* includes the identities *SSR*, *RFLP*, *SCAR* and *AFLP*). A super-class is a generalization of the classes. For example, the *molecular markers*, *alleles* and *map* classes give different information on the genome and can be grouped at a superior level into a super-class called '*Molecular data*'. The same principle is used to define the sub-classes.

The LDM features specify all the entities and links defined in the model, as well as the attributes and the primary and foreign keys (keys identifying the link established between different entities). In line with examples of data models described in the literature and with our objectives, the relational data model was chosen to ensure traceability of the collected data. To facilitate access to the database by users, a web interface had to be developed. Consequently, database management applications were chosen that would be accessible through the web (e.g. MySQL) and preferably based on open source programmes and operating systems (e.g. Linux).

3 RESULTS

3.1 Conceptual data model (CDM)

The CDM includes six main super-classes: GENOTYPE, PHENOTYPE DATA, MOLECULAR DATA, GROWTHSITE, ORGANIZATION and REFERENCE. As shown in Figure 1, all entities are structured around the super-class GENOTYPE, which is the core element of the model. It covers all plant material by individual trees and DNA samples which can come from any kind of material (cultivars, breeding selections, segregating populations and gene bank accessions). GENOTYPE is subdivided into three classes: PLANT MATERIAL, PASSPORT and SYNONYMS. PLANT MATERIAL includes the two main sub-classes TREE and DNA-SAMPLE, PASSPORT includes the PEDIGREE and ACCESSION main sub-classes, and SYNONYMS includes the SYNONYM and PATRONYM sub-classes.

TREE and DNA-SAMPLE hold the identity descriptors for each individual tree, DNA sample and genotype name. TREE also

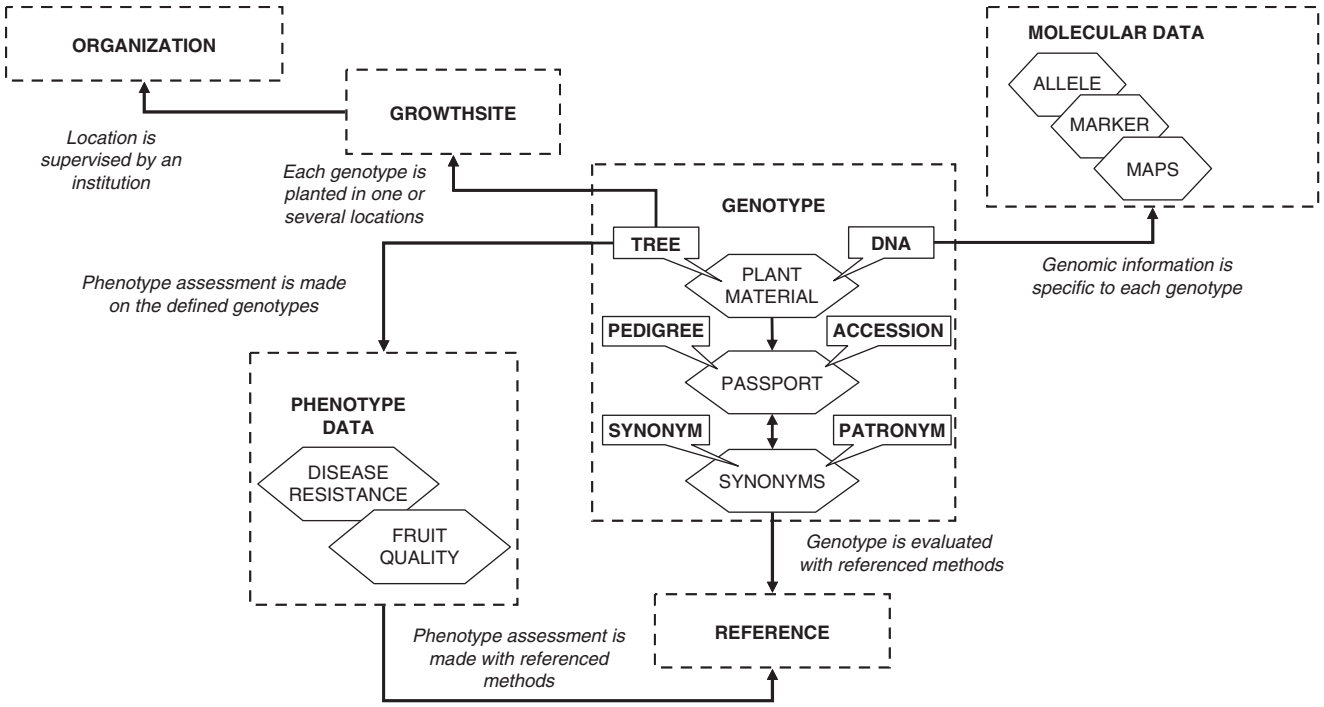

Fig. 1. Conceptual data model of *AppleBreed DB* and existing links between various super-classes.

includes descriptors for the precise location where trees were grown (institute, plot, row and position in row) and their origin (origin of bud wood, year of sowing, planting and grafting, and rootstock). DNA-SAMPLE also includes the origin of each sample (tree from which the sample was derived, date of isolation and position on micro-titre plates of the original sample as their sub-samples etc.).

ACCESSION is used to identify and characterize the plant material (cultivars, breeding selection, segregating population and gene bank accession information). PEDIGREE describes the parentage of each accession up to the founder level and therefore facilitates 'Pedigree Genotyping', a new pedigree-based approach of QTL identification and allele mining in pedigreed populations (Van de Weg *et al*., 2004). The class SYNONYMS holds the known synonyms and patronyms of each genotype, and accounts for the most frequently occurring typing errors.

Figure 1 shows the relationships between GENOTYPE and other elements of the database. Each genotype is localized in one or more specific trial plots (GROWTHSITE) and each institution (ORGANIZATION) supervises its trial plots. Genotypes are evaluated for their fruit quality and disease resistance (PHENOTYPE DATA). The procedures and results of the genotype DNA analyses are stored in MOLECULAR DATA. Each genotype listed in the database is referenced according to the literature references (Silbereisen *et al*., 1996; Smith, 1971) in the REFERENCE super-class. Table 1 summarizes the information included in each super-class and the corresponding main classes.

Most classes are further divided into one to various generations of sub-classes, until the desired level of detail is reached. All these entities have been converted into tables at the LDM level. A class or sub-class may include one or several tables. The most important tables of the database are listed in Table 2.

As stated earlier, GENOTYPE holds information that identifies genotypes (names of cultivars, breeding selections, crosses and gene bank accessions characteristics) and the tangible part of the plant material (trees and DNA samples). Phenotypic information concerns fruit quality and disease resistance. Finally, molecular information relates to molecular markers used to construct genetic linkage maps, information on mining allele, loci and pedigree of the allele. Each genotype listed in the database is considered to be a central key for the traceability of the information stored in the *AppleBreed DB*.

3.2 Logical data model (LDM)

The LDM describes entities defined within each super-class and their relationships with other entities defined above. The database diagrams (see Figs 2–4) give an external view of the *AppleBreed DB* data content. The consistency of data is automatically checked by the database management system itself, at a superior level, according to the rules and the relationships defined when the schema is implemented. The LDM is presented in more detail for the super-classes (i) GENOTYPE, (ii) PHENOTYPE and (iii) MOLECULAR DATA, specifying their primary and secondary keys.

3.2.1 GENOTYPE *super-class* In the GENOTYPE super-class (Fig. 2) the GT_TREE table and GT_DNA_SAMPLE table are the most important because they allow the individual genotype for the phenotype assessment and the molecular data analysis to be set up. Because of the high importance of plant material

Table 1. Super-classes content in *AppleBreed DB*

Super-classes	Acronym	Content	Main classes
GENOTYPE	GT	Information on the material that represents the genotype (tree, DNA sample), passport data of the genotype (accession, pedigree) and synonyms	Plant material, passport, synonyms
PHENOTYPE DATA	PH	Fruit quality results (as external, sensorial, instrumental evaluations and expert panel results) and disease resistance evaluations	Fruit quality, disease resistance
MOLECULAR DATA	MOL	Information on all results related to markers, linkage groups and allelic forms of the markers, and all necessary information for building maps with markers of a specific genotype. Marker information includes sizes of observed bands, PCR protocols, date and laboratory at which the data were raised primer sequences, and the gDNA or EST sequences from which the markers were derived.	Allele Markers Locus Linkage group Maps
GROWTHSITE	GRO	Information on location of trees and orchards	Site Trial plot
ORGANIZATION	ORG	Information about institutions supervising the site and the trial plot	Institution
REFERENCES	REF	Information on literature references used to describe the genotypes and the evaluation procedure	Reference

Table 2. Content and definition of the main tables corresponding to each super-class in *AppleBreed DB*

Super-classes	Main classes	Main tables	Content
GENOTYPE	PLANT MATERIAL	GT_TREE	Trees traced in the model
		GT_DNA_SAMPLE	Information on DNA samples used in the model
	PASSPORT	GT_ACCESSION	Type of material (cultivar, breeding selection, segregating population, gene bank accession)and their names, including synonyms
		GT_PEDIGREE	Parents of each accession, if known
	SYNONYMS	GT_SYNONYM	List of synonyms for each patronym
		GT_PATRONYM	Patronym names with their literature references
PHENOTYPE DATA	FRUIT QUALITY	PH_INSTRUMENTAL_ANALYSIS	Instrumental analysis made during the observation period
		PH_EXTERNAL ANALYSIS	External analysis made duringthe observation period
		PH_SENSORIAL_ANALYSIS	Sensorial analysis made during the observation period
		PH_EXPERT_PANEL	Expert panel evaluation
		PH_SAMPLE	Sampling of the fruit to facilitate the traceability of the information
	DISEASE RESISTANCE	PH_DISEASE	Information on diseases
		PH_DISEASE_ASSESSEMENT	Observations made over several years
		PH_DISEASE_ORGAN	Information on the plant organ evaluated
MOLECULAR DATA	ALLELE	MOL_ALLELE	Allele information
	MARKERS	MOL_MARKERS	Markers used in the molecular analyses
	LOCUS	MOL_LOCUS	Locus names and other information about it
	LINKAGE GROUP	MOL_LG	Linkage group information: link between genotype, loci and allele
	MAPS	MOL_MAPS	Maps description
GROWTHSITE	SITE	GRO_SITE	Sites used to locate the genotype
	TRIAL PLOT	GRO_TRIAL_PLOT	Trial plots used to locate the genotype
		GRO_TP_FERTILITY	Soil fertility classes
		GRO_TP_DRAINAGE	Soil drainage classes
		GRO_TP_ORGANIC_MATTER	Soil organic matter classes
		GRO_TP_TEXTURE	Soil texture classes
ORGANIZATION	INSTITUTION	ORG_INSTITUTIONS	Institutions which supervised a growth
REFERENCES	REFERENCES	REF_REFERENCES	Information on references used to describe the genotypes or the analysis methods

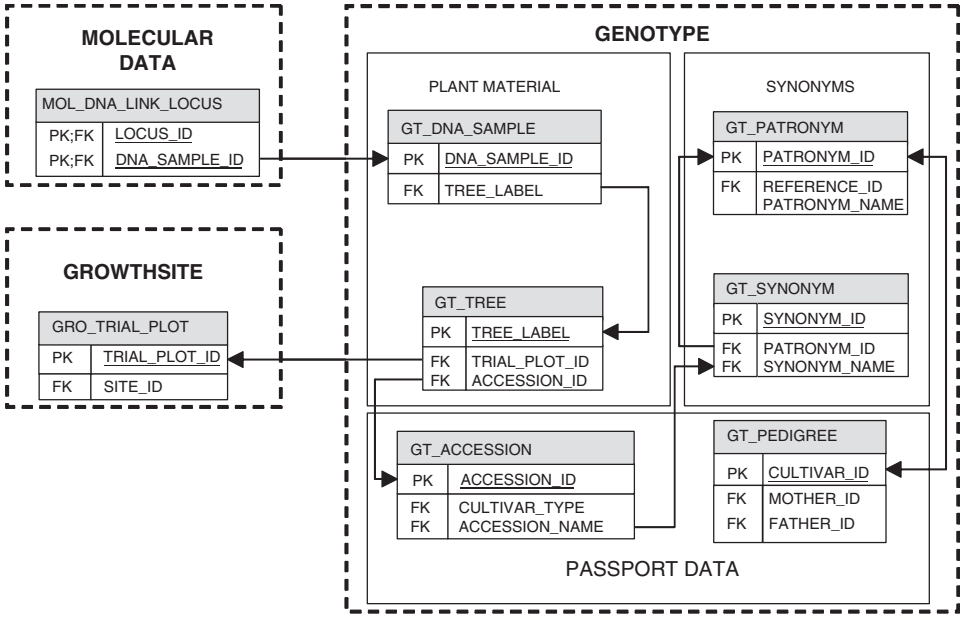

Fig. 2. Detailed structure of the GENOTYPE super-class.

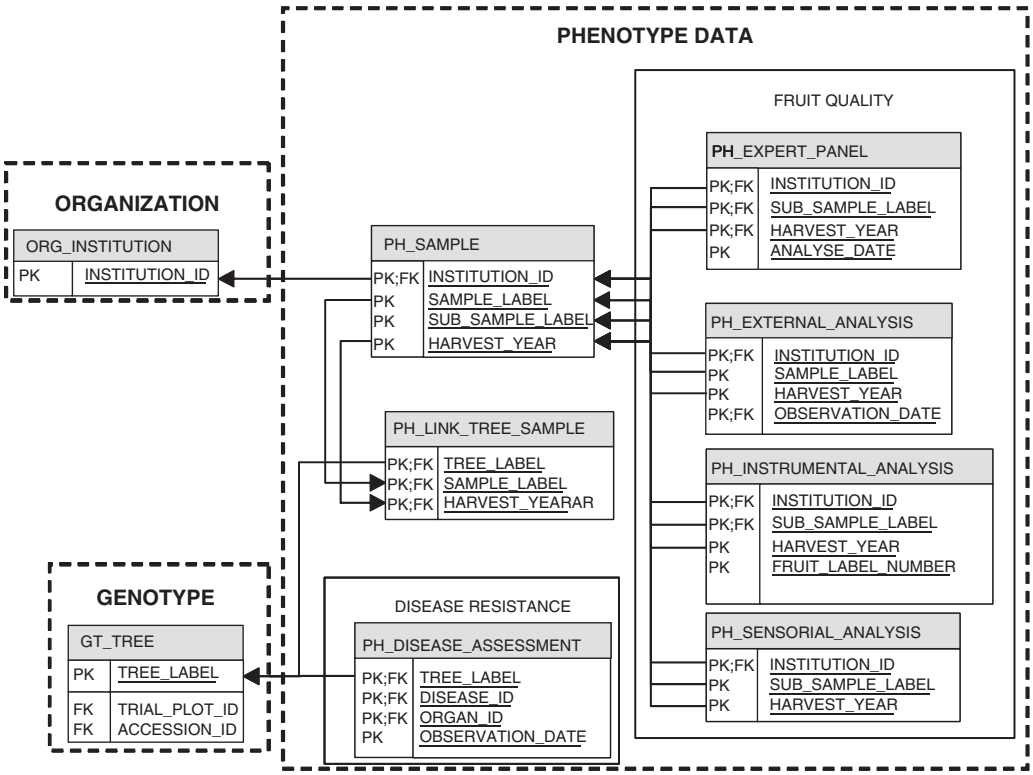

Fig. 3. Detailed structure of the PHENOTYPE DATA super-class.

identification and certification for genetic studies, the emphasis was put on tracking and tracing aspects for the definition of the structure of these tables. Their detailed content is presented in Tables 3 and 4. The link between them is made through the TREE_LABEL primary key.

The GT_ACCESSION table is used to store information that is assigned to an accession when it is entered into a collection. The key element of this table is the accession number, which is unique in the collection. Once assigned, this number can never be reassigned to another accession. The GT_PEDIGREE

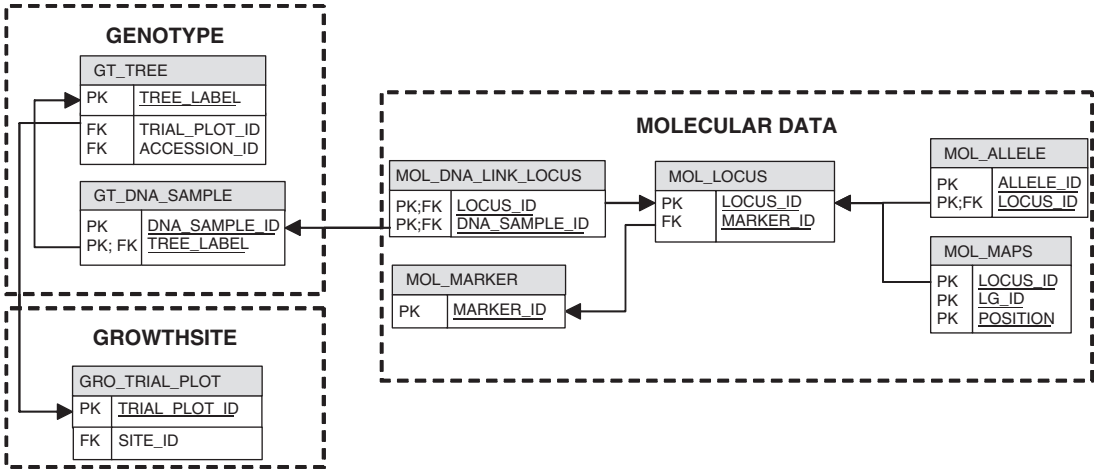

Fig. 4. Detailed structure of the MOLECULAR DATA super-class.

Table 3. Content and definition of the GT_TREE table

Fields	Definition	Remarks
TREE_LABEL	Unique identifier of the tree in the DB	PK
TRIAL_PLOT_ID	Identifier of the trial plot inside the institution	FK to GRO_TRIAL_PLOT
ACCESSION_ID	Accession number in the institution's collection	FK to GT_ACCESSION
ROW_NUMBER	Row number of the tree in the orchard	
POSITION_IN_ROW	Tree position in the row of the orchard	
PLANTING_YEAR	Planting year of the tree	
PLANTING _PERIOD	Planting period in the year	
MULTIPLICATION_TYPE	Type of multiplication used to produce the plant material	
REMARKS	Any further remarks on the plant material or the planting conditions	

PK: primary key; FK: foreign key.

Table 4. Content and definition of the GT_DNA_SAMPLE table

Fields	Definition	Remarks
DNA_SAMPLE_ID	DNA sample identifier	PK
DNA_SAMPLE_NAME	Name of the sample	
TREE_LABEL	Label of the tree from which the DNA sample was collected	FK to GT_TREE
DATE_LEAF_COL	Date of leaves (sample) collection	
COLLECTOR_NAME	Name of the person who collected the sample	
ISOLATION_DATE	Date of the DNA sample isolation	
ISOLATOR_NAME	Name of the person who isolated the DNA sample	
DNA_REF_PROTOCOL	DNA protocol used (file name)	Link to a protocol file
DATA_ENCODE_DATE	Date of data encoding	
DATA_ENCODE_NAME	Person who encoded the data	
ORIG_PLATE_NB	Original plate identifier	
ORIG_PLATE_ROW	Row of the micro-titre plate	
ORIG_PLATE_LINE	Line of the micro-titre plate	
ORIG_INSTITUTION	Institution that provided the sample	FK to ORG_INSTITUTION
NEW_PLATE_NB	New plate identifier	
NEW_PLATE_ROW	New micro-titre plate row	Link to an image file
NEW_PLATE_LINE	New micro-titre plate line	
NEW_INSTITUTION	Institution that conducted the sample analysis	FK to ORG_INSTITUTION
REMARKS	Any further remarks on the DNA sample collection and isolation conditions	

PK: primary key; FK: foreign key.

table allows a user to determine whether a relationship exists between phenotypic characteristics and genomic results from genitors and their progenies.

The high number of synonyms for cultivars is a recurrent problem for breeders, geneticists and managers of gene banks; they impair the efficient management and exploitation of the collections.

This is especially true for old genotypes received or collected in different places and times.

For example, Cox's Orange Pippin has more than 40 synonyms. In addition, very modern cultivars often have both a cultivar and a trademark name. Finally, for widely grown cultivars there are often many mutants, each with its own name. For the old cultivars, there are many sources of synonyms. One is translation or transliteration of original names into local languages. There are also spelling errors due to the 'appropriation', over time, of introduced foreign genotypes in local traditions, resulting in new local names adapted to the language or dialect (Oger and Lateur, 2004). This problem can lead to major disappointments. For example, geneticists might believe they are working on different genotypes, but after obtaining their results they realize they are working on the same genotype with different names. The database model addresses this problem by using the SYNONYMS main class. The first appellation found in the literature has, in most cases, to be considered as the patronymic name. This name is filled out in the identifier field PATRONYM_NAME in the GT_PATRONYM table, as displayed in Figure 2, and a link between the patronymic name of a genotype and its synonyms is assured through the PATRONYM_ID.

3.2.2 PHENOTYPE DATA super-class The PHENOTYPE DATA super-class (Fig. 3) includes two main classes: FRUIT QUALITY and DISEASE RESISTANCE. Each genotype is studied for several traits (Gianfranceschi and Soglio, 2004), such as: fruit external characteristics (shape, ground colour, overall colour, fruit size, etc.), fruit internal quality (sugar content, starch index, acidity, etc.), the sensorial evaluations of expert panels to determine the quality of the fruits (sourness, juiciness, firmness, etc.) and the disease levels under natural conditions in the orchard as well as in specially designed greenhouse tests. Data are encoded for each individual assessment, which can be made for a series of individual apples (e.g. firmness data) for different dates (e.g. 0, 2 and 4 months after harvest), localities and years.

Figure 3 also displays the relationships among the main tables of this super-class as well as the relationships with other tables included in other entities, such as GENOTYPE and GROWTHSITE. With regard to the sensorial, instrumental, external and panel expert observations, a composite primary key identifies each observation. This key includes an identifier for the sample, an identifier for harvest times (a date), an identifier for the applied method of assessment and an identifier for the institution making the observations.

This kind of primary key structure gives each institution the possibility of marking its own samples (there is a unique sample code number for each institution).

Each genotype is linked to fruit quality assessment tables (PH_INSTRUMENTAL_ANALYSIS, PH_SENSORIAL_ANALYSIS, PH_EXPERT_PANEL, PH_SENSORIAL ANALYSIS) by the successive tables GT_TREE PK_TREE_LABEL and the table PH_SAMPLE. The PH_SAMPLE_ID field links all the information from instrumental, sensorial and disease observations to trees, and thereby to genotypes.

Each tree is assessed individually, making it possible to connect phenotypic observations with molecular marker data by means of the genotype. This structure allows users to select, for example, a genotype with fruits that have the same level of sugar content and the same starch index, or are similar or dissimilar for other important characteristics. The primary key for PH_DISEASE_ASSESSMENT (the table is included in the DISEASE RESISTANCE class) is also a composite key. This key includes the identifiers of each individual tree, observation date, disease identifier and observed organ plant, as well as an identifier for the applied method of assessment.

3.2.3 MOLECULAR DATA *super-class* The objective of the HiDRAS project is to molecularly characterize all the individuals belonging to a selected pedigree using highly informative markers. Families and their connected progenies are chosen for being representative of apple breeding material and differentiated for fruit quality and disease resistance.

One aspect of the project concerns the development of new highly informative molecular markers to fill the gaps in the available apple linkage maps. The origin of all alleles of each marker/genotype combination is assessed in terms of the alleles of the founding cultivars (identity by descent) by analysing marker data.

The MOLECULAR DATA super-class (Fig. 4) is one of the most important components of the model. Its data describes the genetic constitution of each genotype (allelic composition of molecular markers and major genes) and must allow alleles to be traced over generations. Starting from the genotype, all information is linked in the database as a chain. The molecular information is linked to the genotype by the GT_DNA_SAMPLE table and the DNA_SAMPLE_ID. In the MOLECULAR DATA super-class, the MOL_DNA_LINK_LOCUS and the LOCUS_ID make the link with MOL_LOCUS, MOL_ALLELE, MOL_MAPS and MOL_MARKER tables.

The content of the MOL_LOCUS and MOL_MARKER tables is described in Table 5.

Due to the links between all the tables, the *AppleBreed DB* can easily provide input data for QTL software to search for combinations of certain molecular markers and fruit quality traits (e.g. skin colour, shape or global taste).

3.3 Database implementation

AppleBreed DB was implemented within a MySQL database system and a Linux environment. A web interface was developed in PhP language. Figure 5 illustrates the data management system adopted for the submission and validation of data. Users send their data to the database administrator via specific, standardized templates (Excel files). Data quality control involves three steps: (1) the structure of the encoding templates (templates created to collect data include control concerning the allowed numeric values or class evaluations), (2) the quality check by the database manager and (3) the constraints existing in the database structure itself (a journal

Table 5. Content and definition of the MOL_LOCUS and MOL_MARKERS tables

Fields	Definition	Remarks
MOL_LOCUS table		
LOCUS_ID	Locus identifier	PK
LOCUS_NAME	Locus name	
MARKER_ID	Marker identifier	FK to MOL_MARKERS
PUB_REF	Published reference	FK to REF_REFERENCE
ORIG_SET_REF_CV	Original set reference cultivar	
ADD_INFO_LG	Additional information on the linkage group	
NEW_ALLELE_VAL	New allele value	
DATA_ENCODE_DATE	Date of data input	
DATA_ENCODE_NAME	Person who encoded the data	
INSTITUTION_ID	Institution identifier	FK to ORG_INSTITUTION
MOL_MARKERS table		
MARKER_ID	Marker identifier	PK
MARKER_NAME	Marker name	
MARKER_TYPE	Marker type	
FORWARD_PRIMER	Forward primer-sequence	
REVERSE_PRIMER	Reverse primer-sequence	
GEL_IMG_REF	Reference to a gel image	Link to image file
REQ_TEMP	Temperature used	
LOCI_NUMBER	Number of loci found for the marker	
UPDATE_SHEET	Date of update	
LOCUS_STATUS	Locus status	
RESEARCHER	Researcher name	
PUB_REF	Published reference	FK to REF_REFERENCE
DATA_ENCODE_DATE	Date of data input	
DATA_ENCODE_NAME	Person who encoded the data	
ORIGIN_SEQ	Original sequence	
ORIGIN_FORWARD_SEQ	Original forward primer-sequence	
ORIGIN_REVERSE_SEQ	Original reverse primer-sequence	
PCR_PROTOCOL	PCR protocol used	Link to protocol file

PK: primary key; FK: foreign key.

with the error values is generated). Once these checks are achieved, the results regarding suspicious data are sent back to users for validation. After re-submission, the administrator carries out the transfer and integration of data into the database structure. Finally, users can visualize and upload both the raw and interpreted results by accessing specific web pages. Simple SQL queries allow on-line access to the database through the Internet. Various real-time query tools have been developed, including specific multiple-choice questionnaires for different views of the requested information. Data output formats can be generated 'à la carte', making output directly compatible for a wide range of software packages, including packages for QTL mapping.

4 DISCUSSION

AppleBreed DB is, as far as the authors know, the first database to store both genetic and phenotypic data up to the level of individual observations. This combination of data makes *AppleBreed DB* a powerful tool for extensive genetic studies directed at the assessment of marker-trait associations, for candidate gene validation and for allele mining. *AppleBreed DB* takes into account the particularities of perennials such as: (1) vegetatively propagated, allowing the same genotype to be present at various localities, (2) long juvenile phase, (3) multi-annual crop, (4) long economic lifespan and (5) simultaneous availability of successive generations in the same plot of breeding programs, experimental stations and gene banks. These aims and particularities determined the general structure of the database, and have resulted in a framework quite distinct from models in use for annual crops, such as the ZmDB database (Dong *et al*., 2002; Du *et al*., 2003; Gai *et al*., 2000) or the MaizeGDB database (Lawrence *et al*., 2004).

AppleBreed DB is built on a relational model. The structure of its conceptual model allows for the flexible addition of new entities. In other words, the *AppleBreed DB* structure allows data with new characteristics to be easily and quickly integrated into the database, at least as long as the database integrity rules are respected. The ability to encode new data into the database is checked by the database structure itself.

Due to the relational structure of the database, users' queries are easily handled through SQL requests. Other potential real-time query tools can be easily added, such as specific multiple-choice questionnaires for different views of the

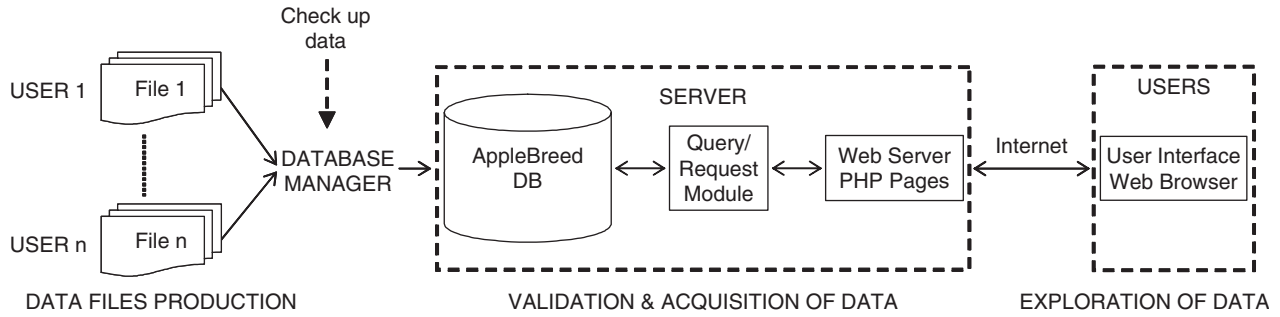

Fig. 5. Data flow setup within the framework of the model.

requested information. Modules to export data in 'à la carte' output formats are also under development, making data directly compatible for a wide range of software packages, including packages for QTL mapping. An interesting point for geneticists and breeders is that it is possible to manage traceability of plant material, a genotype or a family and to follow the parents and their descendants. In addition, the flexibility of the data model makes it possible to adapt this system for other multi-annual botanical species. Unfortunately, one characteristic of relational databases might represent an inconvenience. Direct encoding of results is not allowed, for example, for new genotypes or markers. It is always necessary to insert new data in a particular and logical order and according to a specific and defined format.

AppleBreed DB can store phenotypic data at the level on which they were originally assessed, including at the level of individual samples. In addition, the position of trees in the orchard and the genetic relationships among genotypes are documented. Together, this allows in-depth analysis of the data because experimental design, position effects, genetic relationships and experimental variation can be taken into account.

This not only allows in-depth classical analysis of the phenotypic data itself, such as heritability estimates and the effect of different cultivation practices and environments, but also ensures a high-power detection of marker-trait associations. As it stands *AppleBreed DB* will be a powerful tool for resolving the genetic base of horticulturally important traits. In addition, it has the potential to support valorization of EST and genome sequencing projects, since its phenotypic and genetic data can be helpful in the identification of the candidate genes validated by geneticists.

Currently, there are various public databases for perennial crops that are related to different aspects of genetics and breeding. The USDA-ARS Germplasm Resources Information Network (GRIN http://www.ars-grin.gov/npgs/) is a database which stores information about clonal germplasm in the USDA system, including various tree species as apples, pears stone fruits, grapes, etc. The Genome Database for Rosaceae (GDR, http://www.mainlab.clemson.edu/gdr/) is a curated and integrated web-based relational database. GDR contains data on physical and linkage maps, annotated EST sequences and all publicly available Rosaceae sequences. Although this database started as a database for *Prunus*, it is now extending to other families of the Rosaceae. Various databases for the management of genetic resources were created by the European Cooperative Programme for Plant Genetic Resources Networks (ECP/GR). These databases are crop specific and include Apple (http://www.ecpgr.cgiar.org/databases/Crops/Malus.htm [Maggioni *et al*., 2002]), Pear (*http://pyrus.cra.wallonie.be/*) and various stone fruits (*http://www.bordeaux.inra.fr/urefv/base/*). The HiDRAS SSRdb (http://www.hidras.unimi.it/) contains detailed information on more than 300 SSR markers that have been mapped in apple. The *AppleBreed DB* is currently uploading the HiDRAS data, most of which are likely to become public. All these databases are relational, curated and web based. They are continuously extending in content and functionality. Much synergism could be obtained by tuning into their policies, content and formats, and much added value could be obtained if private databases such as the HortResearch Apple EST Database (Crowhurst *et al*., 2005) became part of the network.

5 CONCLUDING REMARKS

The *AppleBreed DB* model provides a unique tool specifically adapted for geneticists and breeders working on perennial crops with a long economic lifespan, especially when the aim is to combine phenotypic and molecular marker data. It supports pedigree-based analysis of the data, including 'Pedigree Genotyping' (Van de Weg *et al*., 2004). This database could be useful in intercontinental collaboration on marker-trait associations, validation of candidate genes and functional allelic diversity. It can be directly applied to apple, and its structure forms a firm foundation on which other users can build their own applications. It can be easily extended to include various crops, thus forming a base for a *RosaceaeBreed DB*. Links to other databases such as GRIN, NCBI (National Center for Biotechnology Information), SINGER (System-wide Information Network for Genetic Resources) and EMBL (Nucleotide Sequence Database), can also be investigated.

ACKNOWLEDGEMENTS

This research was carried out with financial support from the Commission of the European Communities (Contract No. QLK5-CT-2002-01492), Directorate-General Research—Quality of Life and Management of Living Resources Program.

This manuscript does not necessarily reflect the Commission's views and in no way anticipates its future policy in this area. Its content is the sole responsibility of the authors. The authors are deeply indebted to all participants of the HiDRAS project for their involvement, collaboration and support in the development of the conceptual data model of *AppleBreed DB*. Funding to pay the Open Access publication charges was provided by Agricultural Walloon Research Centre of Gembloux (Belgium).

Conflict of Interest: none declared.

REFERENCES

Crowhurst,R.N. *et al*. (2005) The HortResearch apple EST database – a resource for apple genetics and functional genomics. In *Proceedings of Plant & Animal Genomes XIII Conference*, http://www.intl-pag.org/13/abstracts/PAG13_P499.html.

Dong,Q. *et al*. (2002) ZmDB, an integrated database for maize genome research. *Nucleic Acids Res*., **31**, 244–247.

Du,C. *et al*. (2003) Development of a maize molecular evolutionary genomic database. *Comp. Funct. Genomics*, **4**, 246–249.

Gai,X. *et al*. (2000) Gene discovery using the maize genome database ZmDB. *Nucleic Acids Res*., **28**, 94–96.

Gianfranceschi,L. and Soglio,V. (2004) The European 589 project HiDRAS: innovative multidisciplinary approaches to breeding high quality disease resistant apples. *Acta Hortic*., **663**, 327–330.

Lawrence,C.J. *et al*. (2004) MaizeGDB, the community database for maize genetics and genomics. *Nucleic Acids Res*., **32**, D393–D397.

Lespinasse,Y. (1992) Le pommier. In Gallais,A. and Bannerot,H. (eds.) *Amélioration des espèces végétales cultivées*. INRA Editions, Paris, pp. 579–594.

Maggioni,L. *et al*. (2002) Report of a working group on *Malus*/*Pyrus* compilers. Second Meeting 2–4 May 2002, Dresden-Pillnitz, Germany. IPGRI – ECP/GR, Rome.

Matthews,D.E. *et al*. (2003) GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res*., **31**, 183–186.

Morgan,J. and Richardson,A. (2002) *The New Book of Apples*. Ebury Press, London.

Oger,R. and Lateur,M. (2004) Development of a specific software for the management of the recurrent synonymous problem of cultivars inside plant genetic resources databases: the case of the European EUROPEAN ECP/GR Pyrus database. *Acta Hortic*., **663**, 593–596.

Pollack,S. (2001) Consumer demand for fruit and vegetables: the U.S. example. In Regmi,A. (ed.) *USDA Economic Research Service Agriculture and Trade Report WRS-01-1,* USDA, Washington, DC, USA.

Reiser,L. *et al*. (2002) Surviving in a sea of data: a survey of plant genome data resources and issues in building data management systems. *Plant Mol. Biol*., **48**, 59–74.

Rhee,S.Y. *et al*. (2003) The *Arabidopsis* information resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res*., **31**, 224–228.

Sakata,K. *et al*. (2000) INE: a rice genome database with an integrated map view. *Nucleic Acids Res*., **28**, 97–101.

Schoof,H. *et al*. (2002) MIPS *Arabidopsis thaliana* database (MatDB): an integrated biological knowledge resource based on the first complete plant genome. *Nucleic Acids Res*., **30**, 91–93.

Silbereisen,R. *et al*. (1996) *Obstsorten-Atlas*. Eugen Ulmer GmbH & Co, Stuttgard.

Smith,M.W.G. (1971) *National Apple Register of the United Kingdom*. Ministry of Agriculture, Ficheries and Food, London.

Van de Weg,W.E. *et al*. (2004) Pedigree genotyping: a new pedigree-based approach of QTL identification and allele mining. *Acta Hortic*., **663**, 45–50.

Way,R.D. *et al*. (1991) Apple (*Malus*). *Acta Hortic*., **290**, 3–46.