D614–D618 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki123

BarleyBase—an expression profiling database for plant genomics

Lishuang Shen[1], Jian Gong[1], Rico A. Caldo[2], Dan Nettleton[3], Dianne Cook[1,3], Roger P. Wise[2,4] and Julie A. Dickerson[1,*]

[1]Virtual Reality Applications Center, [2]Department of Plant Pathology, Center for Plant Responses to Environmental Stresses, [3]Department of Statistics and [4]Corn Insects and Crop Genetics Research, USDA-ARS, Iowa State University, Ames, IA 50011, USA

Received August 12, 2004; Revised and Accepted October 21, 2004

ABSTRACT

BarleyBase (BB) (www.barleybase.org) is an online database for plant microarrays with integrated tools for data visualization and statistical analysis. BB houses raw and normalized expression data from the two publicly available Affymetrix genome arrays, Barley1 and Arabidopsis ATH1 with plans to include the new Affymetrix 61K wheat, maize, soybean and rice arrays, as they become available. BB contains a broad set of query and display options at all data levels, ranging from experiments to individual hybridizations to probe sets down to individual probes. Users can perform cross-experiment queries on probe sets based on observed expression profiles and/or based on known biological information. Probe set queries are integrated with visualization and analysis tools such as the R statistical toolbox, data filters and a large variety of plot types. Controlled vocabularies for gene and plant ontologies, as well as interconnecting links to physical or genetic map and other genomic data in PlantGDB, Gramene and GrainGenes, allow users to perform EST alignments and gene function prediction using Barley1 exemplar sequences, thus, enhancing cross-species comparison.

INTRODUCTION

BarleyBase (BB) is a USDA-funded public database for cereal microarray data. BB was first developed to support the Affymetrix Barley1 GeneChip, and is being expanded to new plant GeneChips and other microarray platforms. The Barley1 GeneChip is a new community-designed, Affymetrix probe array (1), which pioneered the GeneChip design for plants without a fully sequenced genome. Several new GeneChips for wheat, soybean and maize will be released in 2005.

BB includes MIAME-compliant microarray experiment annotations as well as Plant Ontology terms through *BarleyExpress*, its web-based submission tool (2). Links with other sequence and crop databases give BB users the ability to quickly discover all the known facts about any probe set or exemplar sequence on the chip and to compare with other plant species such as rice or wheat. Data queries are integrated with analysis and visualization tools to allow users to explore their experimental data. As of September, 2004, BB hosts 23 completed experiment submissions with a total of 972 hybridizations.

There are many public databases that provide access to microarray data. These include general repositories, such as the Gene Expression Omnibus (GEO) (3), Stanford Microarray Database (4) and ArrayExpress (5) and species-specific resources, such as TAIR (6) and NASCArrays (7). Repositories typically store data for download and later analysis. The general repositories such as GEO and ArrayExpress are intended to act as central data distribution hubs, not to replace gene expression databases that are constructed to facilitate particular analytic methods or comparisons. BB is designed to meet the needs of plant biologists in their analysis of gene expression data and to put the expression data in the context of functional genomics by using controlled gene and plant ontologies to describe experimental conditions. Interconnecting links to plant genomic resources such as PlantGDB (8), Gramene (9) and GrainGenes (10) facilitate access to contig alignments, oligo probe information and a variety of BLAST tools from the NCBI, PlantGDB, TIGR, TAIR or Rice genome databases.

DATABASE DESCRIPTION

BB stores microarray gene expression data in a MIAME-compliant and Plant Ontology enhanced format for plants,

[*]To whom correspondence should be addressed. Tel: +1 515 294 7705; Fax: +1 515 294 8432; Email: julied@iastate.edu

and integrates the data with exploration and analysis tools across experiments. BB stores the following types of information: GeneChip and/or microarray structure data, experimental and labeling protocols, raw and normalized gene expression data and experiment and sample annotations such as summary statistics from R and MAS5.0.

BB uses a hierarchical data model to organize and display microarray gene expression data. The top-level data structure is the experiment, which consists of a set of hybridizations with a treatment structure designed to answer one or more related biological questions. A factorial treatment structure is used to describe BB experiments. Each treatment is associated with a specific level of each of one or more experimental factors. Each treatment has one or more samples as biological replicates; each sample has one or more hybridizations as technical replicates.

To facilitate smooth data exchange across databases, plant ontologies for growth stage and organism parts (11), and other controlled vocabularies are required in the experiment description and sample annotation in BarleyExpress. BB follows the MIAME standards (12) and the implementation used in MIAMExpress (http://www.ebi.ac.uk/miamexpress). BarleyExpress adds plant-specific fields such as links to the Plant Ontology terms on growth stages and tissue types are added in the experiment submission process (2). The use of controlled vocabularies allows cross-experiment comparisons based upon common identifiers, facilitating interoperability between existing plant databases to identify homologous genes. Biological annotation for probe sets and exemplars includes sequence description, BLAST hits from related sequence databases or species, Gene Ontology, and pathway and gene family information.

BB requires raw CEL data files for gene expression data for which EXP and DAT files are recommended. BB processes all submissions in a standardized way which ensures ease of cross-experiment comparison. After the submitter uploads the experiment data, the curator checks the data integrity and computes the normalized expression measures, summary statistics and graphs. Unique accession numbers are assigned to each experiment for data access. Processed data, sequence annotation and pre-computed analyses results are stored for online access and analysis. Finally, BB generates MAGE-ML and text files for batch download and data exchange. The MAGE-ML files can be submitted to ArrayExpress or read by many microarray data analysis programs.

BB uses open-source tools or tools free for academic institutions. The server uses RedHat Linux version 9. The website is powered by an Apache server, PHP and Javascript for dynamic web pages, and a MySQL 4.0 relational database as the back end. The data pre-processing uses R (13), Bioconductor and Perl. R is an open platform for statistical computation and Bioconductor is a project written in R for microarray data analysis. Robust Multichip Average (RMA) (14) in the affy package of Bioconductor and Affymetrix MAS 5.0 (15) are used to compute normalized expression measures from the raw expression values.

Data access policy

BB has secure and flexible account and data access management, which allows data owners to protect their data before publication and yet enables dispersed collaboration. The submitter can specify the accessibility to data of an experiment as 'public', 'private' or 'group accessible'. Public access allows any users to access data; private allows data to be viewed only by the data owner; and group access allows group members to access the data. Registered users can create groups and add selected users to the groups to grant access to data from designated experiments. Reviewers can anonymously access datasets referenced by a manuscript to verify the conclusions using reviewer's login ID. All users are strongly encouraged to make their data public as soon as possible.

DATA ANALYSIS AND VISUALIZATION

Microarray gene expression datasets are large and multivariate in nature and require flexible approaches for analysis. Instructive data visualization and presentation of the data are indispensable for users to efficiently mine the data and derive meaningful biological interpretations. The visualization pages are provided at different levels based on the query hierarchy. Visualizing the expression data will aid users in choosing suitable parameters for gene filtering and analysis. The analysis and visualization tools can be accessed by using a traditional pipeline of experiment analysis or searching for a gene(s) of interest using sequence comparisons then finding genes that behave similarly. The Supplementary Material guides a user through some of the analysis tools available at BB (http://www.barleybase.org/quicktour.php).

Data visualization for experiments

Experiment queries are typically the starting point for data retrieval and analysis flow. Based on the information captured for experiment design in BarleyExpress, the Experiment Query allows users to search and browse the experiments, protocols and array designs.

Quality checking and understanding the experimental data are essential before conducting gene-centric analysis. Users can navigate the expression values by hybridizations and experimental factor, and check sample annotation. The summary statistics and visualizations allow users to quickly assess experiment quality. Box plots and histograms of raw Perfect Match (PM) intensities and normalized expression values are used to check the distribution of the expression data and the quality across hybridizations in an experiment. Histograms of the PM values detect signal saturation, and help to quickly catch problems such as incorrect scanner parameters. Side-by-side boxplots of the normalized expression data are used to assess normalization results. These boxplots are ideally almost identical as shown in Figure 1.

At the hybridization level, pseudo-color images of PM intensities are used for visual detection of spatial abnormalities. Scatter plots and MVA plots show reproducibility and variability among and/or between hybridizations or treatments. These comparative scatter plots can range across experiments, with *x*- and *y*-axes using hybridizations or treatment means from different experiments sharing similarity in experimental material or factors. In the MVA plots, the *M* is the log ratio between two hybridizations and *A* is average of the logged signal intensities. MVA plots can be regarded as a

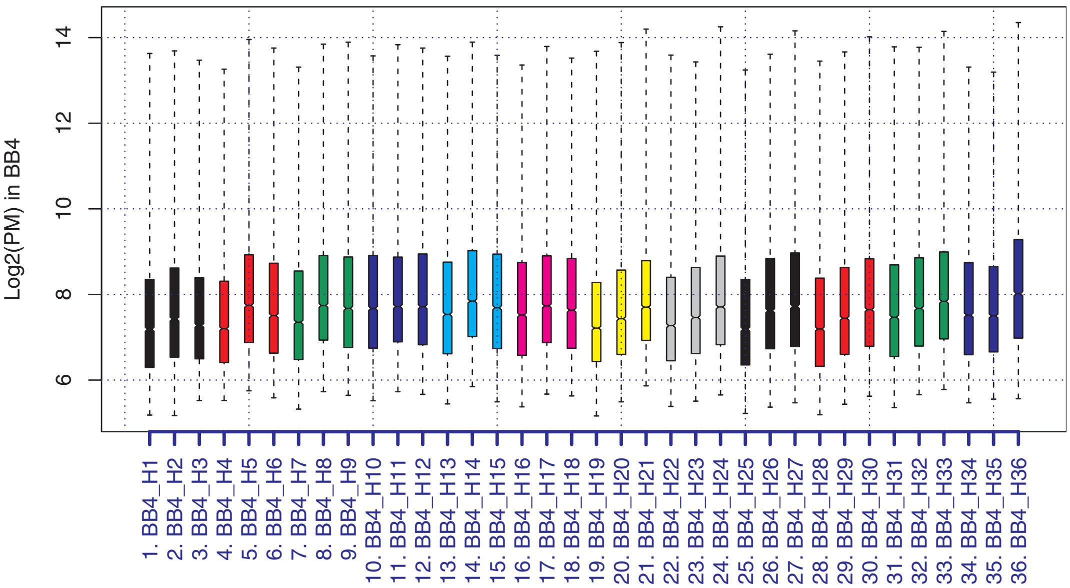


Figure 1. Boxplots can be used to visualize overall experimental quality before expression data normalization. The central box shows the middle 50% of the data. The dotted lines or whiskers, extend to the 10th and 90th percentiles, respectively. The circles represent outliers. Boxes with the same color are technical replicates with the same treatment factors. These data summarize the logarithm of the raw probe set perfect match (PM) expression values in experiment BB4 before normalization.

45° clockwise rotation of scatter plots for easier viewing of differential expression.

Gene-centric expression data analysis tools

Following the initial experiment and hybridization exploration, users can further filter data and create gene lists. Creating gene lists is the first step in most gene-centric analysis for microarray experiments. Saved gene lists can be fed to advanced microarray data analysis and visualization methods. BB provides a full range of gene filters by expression profiles and by biological criteria. Gene-centric expression profiles for single genes or gene lists are displayed as profile-plots (line graphs) and heatmaps. Interactive profile-plots allow the user to gain insight into the way treatments affect expression. An 'Expression view' (heatmap) explores genes with similar expression profiles that may represent co-regulated genes.

Expression profile filters are mainly used to identify differentially expressed genes (probe sets). The filters usually operate on a single experiment, but users may do cross-experiment query for hypothesis generation. The filter can be a single filter or a composite filter that is a combination of several filters linked with various Boolean operators. Filters are based on absolute value range, relative and absolute variation, fold change, MAS5.0 Presence/Absence call or other variation measures. Statistical test filters include most standard two-sample and multiple-sample statistical methods for identifying differentially expressed genes with multiple test corrections. Co-regulated genes can also be identified. For cross-experiment filtering, hybridizations from several experiments are compared with each other. This functions like a virtual experiment in silico using hybridizations from different existing experiments.

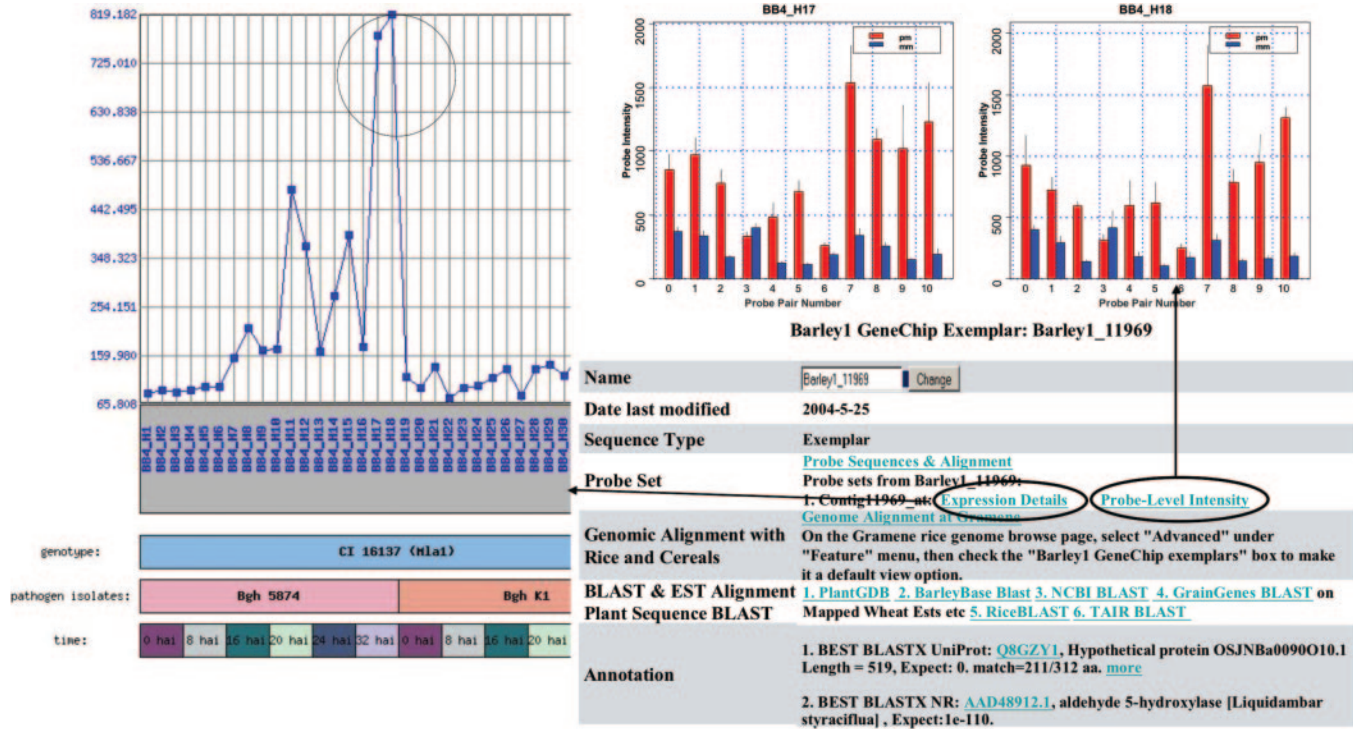
Biologically based filters use annotation keywords and sequence similarity to group genes into a gene list. For the ATH1 GeneChip, gene family and KEGG pathway filters are available to find probe sets corresponding to enzymes from interesting metabolic or regulatory pathways or a given gene family.

Users may import their own list of gene or probe set names. Files of free text containing the gene names can be used directly without tedious editing. Users may export gene lists as tab-delimited text files for names, annotation or expression values. Gene lists can be compared in various combinations: union of two gene lists, intersection of both gene lists and unique genes in either gene lists. This is useful for combining the results of different filters, such as biological and expression-based filters, and for comparing pattern recognition methods. Analysis results are automatically saved, including information about the methods, parameters and gene list analyzed.

Many of the standard supervised and unsupervised pattern recognition methods are implemented for online analysis. Methods include hierarchical and *k*-means clustering, principal component analysis (PCA), self-organizing maps (SOMs) and Sammon's non-linear mapping. For each of the methods, data can be transformed or scaled using logarithm-transformation, mean or median centering, and scaling based on the standard deviation of the probe set in an experiment. The pattern recognition results are visualized using expression profile line graphs, dendrograms and heatmaps for the entire gene list or for each subcluster. Each method also has its specialized visual presentation, such as clustering plots for partitions in *k*-means or partition grids for SOMs.

Gene function and GeneChip annotation

GeneChips may be searched for a sequence of interest by performing a BLAST search against a particular GeneChip (16). This search gives a list of exemplars and probe sets on a particular GeneChip that match that sequence. This page also allows users to access gene expression data from



Barley1 GeneChip Exemplar: Barley1_11969

Name	Barley1_11969 Change
Date last modified	2004-5-25
Sequence Type	Exemplar
Probe Set	Probe Sequences & Alignment Probe sets from Barley1_11969: 1. Contig11969_at: Expression Details Probe-Level Intensity
Genomic Alignment with Rice and Cereals	Genome Alignment at Gramene On the Gramene rice genome browse page, select "Advanced" under "Feature" menu, then check the "Barley1 GeneChip exemplars" box to make it a default view option.
BLAST & EST Alignment Plant Sequence BLAST	1. PlantGDB 2. BarleyBase Blast 3. NCBI BLAST 4. GrainGenes BLAST on Mapped Wheat Ests etc 5. RiceBLAST 6. TAIR BLAST
Annotation	1. BEST BLASTX UniProt: Q8GZY1, Hypothetical protein OSJNBa0090O10.1 Length = 519, Expect: 0. match=211/312 aa. more 2. BEST BLASTX NR: AAD48912.1, aldehyde 5-hydroxylase [Liquidambar styraciflua] , Expect:1e-110.

Figure 2. Annotation for exemplar Barley1_11969. The exemplar annotation page includes the probe alignment, links to databases such as Gramene, PlantGDB and GrainGenes along with the best BLAST hits. The left-hand side picture shows the normalized expression data for some of the hybridizations in experiment BB4. The blocks at the bottom show the experimental factors laid out in a factorial experimental design. The raw probe intensity data PM (red) and MM (blue) levels for exemplar Barley1_11969 in hybridizations 17 and 18 in experiment BB4 show the different responses across the probe pairs.

BB. This type of search is particularly important for organisms which have not been fully sequenced. Figure 2 shows the results of finding a particular exemplar and its accompanying annotation that links to plant genomic resources such as PlantGDB, Gramene and GrainGenes. Contig alignments from HarvEST:Barley [http://harvest.ucr.edu/Barley1.htm (1)] and oligo probe information from the Barley1 and Arabidopsis GeneChips can be displayed. The sequences can be blasted against the NCBI, PlantGDB, TAIR or Rice genome databases for additional annotation information.

The annotation page also links to expression data related to the probe exemplar as shown in Figure 2. The user can look at how this probe set is expressed in different experiments or search for genes that behave similarly in certain experiments. Probe sets with similar expression profiles as the selected exemplar can be identified using correlation tests. These genes may be used to create gene lists for further analysis on a particular experiment or groups of experiments. This type of analysis is critical in identifying co-regulated genes that may be involved in similar biochemical pathways. The results are displayed using heatmaps or profile plots. For more detail, the raw probe pair PM and MM data can also be displayed to further investigate GeneChip response to a particular hybridization as shown in Figure 2. Barplots with standard deviation are plotted by hybridizations or by probe pair numbers, allowing comparison of intensities across hybridizations for same probe, or across probe pairs for same hybridization. As our data and understanding of the GeneChips accumulate, we plan to exclude probe pairs that are known to be ineffective from the analysis available to BB users.

Comparative genomic analysis

BarleyBase supports comparative genomics capabilities by interconnecting links with established plant databases. Barley1 exemplars are aligned to the sequenced model plant rice genome browser in Gramene, and to other cereal genomes for annotation information integration. Barley1 exemplars can also be queried for Triticeae map positions in GrainGenes. Integrated links with PlantGDB facilitates detailed gene prediction and contig view of the exemplars. ATH1 exemplars, function and pathway information are supported through links with TAIR. A series of BLAST utilities allow users to perform cross-species queries by finding matches for any sequences on plant GeneChips with links to GenBank and other major databases.

Choosing an experiment from different GeneChip platform will automatically initiate cross-platform gene list creation, where the best BLAST hits are used as match from other platforms. Cross-platform gene list creation enhances comparative gene expression analysis to fully utilize microarray data from different plant species.

SUMMARY

Adherence to MIAME standards and controlled plant ontologies facilitates the efficient presentation and organization of the volumes of data from a typical microarray-based investigation. BarleyBase captures and stores all applicable MIAME-compliant information and enforces plant ontology and controlled vocabulary for experiments.

BB explicitly captures factorial experiment design information, enhancing the flow of experiment submission, data analysis and data presentation. It makes data accessible at each data level, from the experiment level to the individual probe level. The online pipeline integrates a broad set of gene query and display options with a full set of analysis and visualization tools. Cross-experiment gene filtering and cross-platform matching provide great flexibility in hypothesis generation.

FUTURE PLANS

BB is under active development, and several enhancements are planned for the near future. First, BB will expand to support the Affymetrix high-density GeneChips for maize, rice, soybean and wheat that will be available soon, and will evolve into PLEXdb, a comprehensive Plant Expression Data Base. Second, data from spotted cDNA and long oligo microarray platforms will also be added using open-source tools for integrating cDNA microarray data processing and management, such as the TM4 suite from TIGR (17) and BASE (18). Third, plant ontologies for other species beyond barley will be enhanced. These changes will begin to pave the way toward comparative expression data analysis. Gene Ontology and pathway information need to be adapted to BB for exemplar annotation, which will allow functional gene expression analysis with insight on how specific genes are involved in biological processes. Fourth, expression analysis and visualization tool development will add new methods for gene identification and pattern recognition, and enhance BB's web-based interactive visualization capabilities. Overlaying expression data with Gene Ontology, gene network and pathway analysis will be added to aid biological interpretation. Cross-experiment, cross-platform and cross-species data analysis and comparison capabilities will be enhanced for hypothesis generation.

ACKNOWLEDGEMENTS

BarleyBase is hosted at the Iowa State University Virtual Reality Applications Center. Barley1 exemplar sequences and BLASTX NR annotations were provided by HarvEST:Barley (http://harvest.ucr.edu/Barley1.htm). The Nottingham Arabidopsis Stock Centre's microarray database (NASCArrays) shares ATH1 data. The BarleyBase project is funded by the USDA National Research Initiative (NRI) grant no. 02-35300-12619 and USDA-CSREES North American Barley Genome Project.

REFERENCES

1. Close,T., Wanamaker,S., Caldo,R., Turner,S., Ashlock,D., Dickerson,J., Wing,R., Muehlbauer,G., Kleinhofs,A. and Wise,R. (2004) A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol.*, **134**, 960–968.
2. Tang,X.Y., Gong,J., Xin,J.Q., Shen,L., Turner,S.M., Caldo,R.A., Nettleton,D., Wise,R.P. and Dickerson,J.A. (2004) Barleybase—an expression profiling database for cereal genomics. In *Proceedings of the Plant and Animal Genome XII Conference*, Jan 10–14 2004, San Diego, CA, p. 307.
3. Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res.*, **30**, 207–210.
4. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.*, **31**, 94–96.
5. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al*. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
6. Rhee,S.Y., Beavis,W., Berardini,T.Z., Chen,G., Dixon,D., Doyle,A., Garcia-Hernandez,M., Huala,E., Lander,G., Montoya,M. *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.
7. Craigon,D.J., James,N., Okyere,J., Higgins,J., Jotham,J. and May,S. (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res.*, **32**, D575–D577.
8. Dong,Q., Schlueter,S.D. and Brendel,V. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.*, **32**, D354–D359.
9. Ware,D., Jaiswal,P., Ni,J., Pan,X., Chang,K., Clark,K., Teytelman,L., Schmidt,S., Zhao,W., Cartinhour,S.*et al*. (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.*, **30**, 103–105.
10. 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.
11. The Plant Ontology^{TM} Consortium (2002) The Plant Ontology^{TM} Consortium and Plant Ontologies. *Comp. Funct. Genomics*, **3**, 137–142.
12. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet.*, **29**, 365–371.
13. Ihaka,R. and Gentleman,R. (1996) R: a language for data analysis and graphics, *J. Comput. Graphical Stat.*, **5**, 299–314.
14. Gautier,L., Cope,L., Bolstad,B. and Irizarry,R. (2004) Affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.
15. Affymetrix (2001) *Statistical Algorithms Reference Guide*. Affymetrix Inc., Santa Clara, CA.
16. Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
17. Saeed,A., Sharov,V., White,J., Li,J., Liang,W., Bhagabati,N., Braisted,J., Klapa,M., Currier,T., Thiagarajan,M. *et al*. (2003) TM4: a free, open-source system for microarray data management and analysis, *Biotechniques*, **34**, 374–378.
18. Saal,L.H., Troein,C., Vallon-Christersson,J., Gruvberger,S., Borg,Å. and Peterson,C. (2002) BioArray software environment: a platform for comprehensive management and analysis of microarray data. *Genome Biol.*, **3**, software0003.0001–software0003.0006.