MADMAX – Management and analysis database for multiple ~omics experiments

Ke Lin[1,2], Harrie Kools[3], Philip J. de Groot[4], Anand K. Gavai[1,5], Ram K. Basnet[2], Feng Cheng[6], Jian Wu[6], Xiaowu Wang[6], Arjen Lommen[3], Guido J. E. J. Hooiveld[4], Guusje Bonnema[2], Richard G. F. Visser[2], Michael R. Muller[4], Jack A. M. Leunissen[1,5*]

[1] Laboratory of Bioinformatics, P.O. Box 569, 6700 AN Wageningen, the Netherlands

[2] Laboratory of Plant Breeding, P.O. Box 386, 6700 AJ Wageningen, the Netherlands

[3] RIKILT, P.O. Box 230, 6700 AE Wageningen, the Netherlands

[4] Nutrition, Metabolism and Genomics Group, P.O. Box 8129, 6700 AA Wageningen, the Netherlands

[5] Netherlands Bioinformatics Centre (NBIC), P.O. Box 9101, 6500 HB Nijmegen, the Netherlands

[6] Institute of Vegetable and Flowers, Chinese Academy of Agricultural Sciences, 12 Zhongguancun South Street, Beijing, 100081, China

Summary

The rapid increase of ~omics datasets generated by microarray, mass spectrometry and next generation sequencing technologies requires an integrated platform that can combine results from different ~omics datasets to provide novel insights in the understanding of biological systems. MADMAX is designed to provide a solution for storage and analysis of complex ~omics datasets. In addition, analysis results (such as lists of genes) can be merged to reveal candidate genes supported by all datasets. The system constitutes an ISA-Tab compliant LIMS part, which is linked to the different analysis pipelines. A pilot study of different type of ~omics data in *Brassica rapa* demonstrates the possible use of MADMAX. The web-based user interface provides easy access to data and analysis tools on top of the database.

1 Introduction

To better understand how phenotypes emerge, increasingly series of ~omics technologies (genomics, transcriptomics, proteomics, metabolomics) rather than individual measurements are necessarily used within a single study. Such efforts boost the demands of both data storage and data analysis of different high-throughput approaches. However, in the past it was hardly possible to store metadata from different ~omics technologies in the same repository. To accommodate this demand the ISA-Tab [1] format was proposed to build up a common structured representation of the metadata of studies from a combination of technologies. This also triggered attempts to develop data processing tools tailored to the needs of biologists. Unfortunately most of these tools have high demands on hardware requirements, or contain non-intuitive command line-based interfaces.

* To whom correspondence should be addressed: jack.leunissen@wur.nl

doi:10.2390/biecoll-jib-2011-160

Here we present MADMAX, a multi-purpose database for the management and analysis of data from multiple ~omics experiments. It includes an ISA-Tab compliant backend database and a series of analysis pipelines for transcriptomics, metabolomics and genomics datasets; these pipelines are connected to the database through webservices such that other pipelines can be easily integrated into the current system (Figure 1). The currently supported pipelines include gene function annotation for next generation sequencing of genomes, quality control and statistical analysis (such as Gene Set Enrichment Analysis [2], GSEA) using R [3] and Bioconductor [4] for different microarray platforms, and the Metalign software [5] for LCMS, GCMS and GC-GC-MS metabolomics studies. Because the quality of the gene function annotation is the key to reliable transcriptomics and metabolomics analysis in newly sequenced species, MADMAX uses the ProGMap database for orthologous group information [6] to obtain the function annotation for each gene. Besides the original output from microarray and metabolomics analyses, it will further mine the results and generate over- and under-expressed genes from microarray studies and genes responsible for producing enzymes that affect steps in the pathways of metabolites detected in metabolomics studies. The intersection of genes listed in different omics results may lead to a manageable number of candidate genes for experimental validation.

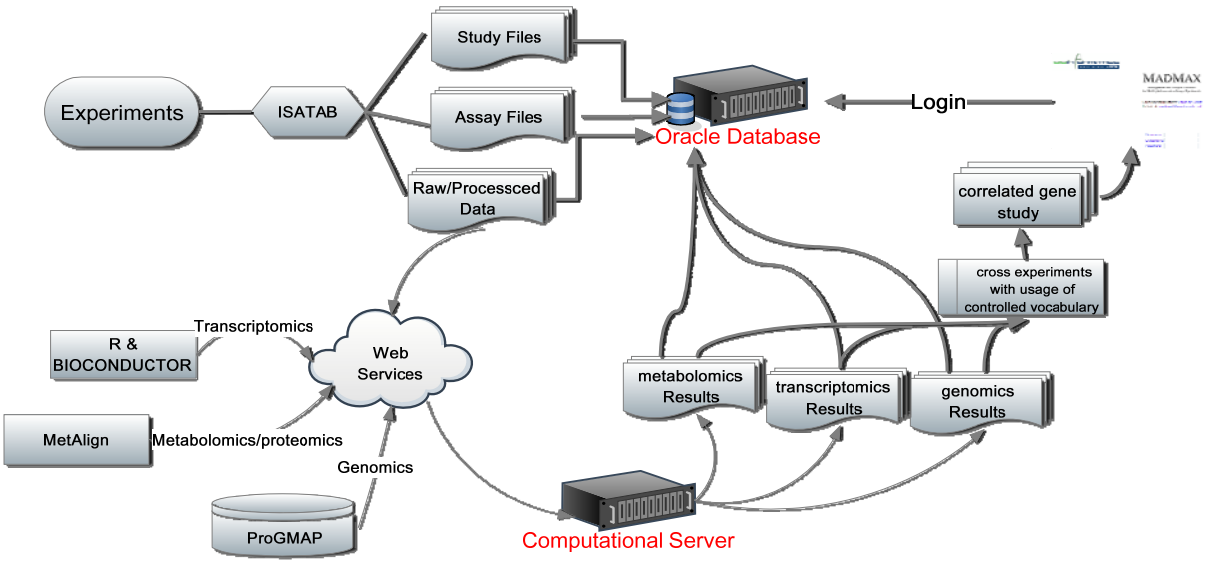

Figure 1: Architecture and pipeline of MADMAX

Through the web interface, the user can store a complete experiment with all fields required in ISA-Tab format, sufficient to allow for subsequent analysis or even repeating the experiment later. Another section on the website is the central access to different analysis pipelines. Both individual analysis results and combined gene lists can be retrieved in the system for download. Centrally stored experiments and analysis results can only be accessed by the creator by default and will be accessible for other users only if the creator desires to share the data. The system is on an automatic backup schedule.

MADMAX can be reached at http://madmax2.bioinformatics.nl/ and is available upon request by sending an email to madmax.request@bioinformatics.nl.

2 Implementation

MADMAX is built upon an Oracle relational database on a Linux server and a computational analysis engine for different ~omics data on a second server. The transcriptomics and

metabolomics analysis pipelines are triggered by web services through a web-based user interface developed using Oracle Application Express. The genomics analysis is done separately outside of the database system, and loaded to the system when it's done. Both metadata and analysis results of an experiment are stored in the database in ISA-Tab compliant format. Different ~omics analysis results can be further combined to yield system-level measurements within one experiment.

2.1 Database infrastructure

The basic unit in MADMAX is the Experiment which contains the overall goals and means used in an experiment as with Investigation in the data model of the ISA-Tab standard. Each experiment can have one or more studies to record the sample preparation information. One study can be used for one or more assays if these assays share the same sample preparation. In an assay, each sample and the corresponding data file generated from it will be connected with additional information about the extracted material; the processed results from the computational server in MADMAX will be attached to the assay where the raw data are used in the analysis.

2.2 Analysis pipeline

Currently MADMAX can handle three types of ~omics datasets: gene models of a partial or complete genome, microarrays from Affymetrix, Illumina, or Agilent, and LC/GC/GCGC mass spectrometry data sets. By default, a list of genes will be generated from the different ~omics data in an experiment, based upon gene models from genomics, over/under expressed genes from microarrays, and genes encoding the enzymes involved in the pathways of the metabolites detected in metabolomics analysis. Not all analysis pipelines used in the system can be accessed by the user interactively. As for genomics analysis, the complete gene models will be loaded in the system and annotated later by the system maintainer. Both microarray and metabolomics analyses can be accessed interactively through the web interface.

2.2.1 Genomics analysis

The *de novo* sequenced genomes using next-generation technology normally contain *in silico* predicted gene models. The gene models in FASTA format are then used as query sequences to BLAST against the ProGMap protein database, which includes protein sequences from Ensembl version 61_8, InParanoid version 7, OrthoMCL version 4, COG/KOG from 2003, eggNOG version 2, HomoloGene version 65, ProtClustDB from Nov 2010, PIRSF from Mar. 9 2010, OMA from Nov 2010 and KEGG ORTHOLOGY database [7-16] from Apr 2011 (at the time of this writing). At present, 21,521,041 proteins sequences are used in the search with default BLASTX [17] settings. When the BLAST search finishes, the matched sequences are filtered through three thresholds: matched sequence length is greater than 100 amino acids, percentage of sequence similarity is greater than 40% and the e-value of the sequence is lower than 10E-20; the results are reformatted into a tab delimited file. Gene models will then be mapped to an orthologous group of databases mentioned above one by one using passed sequences. The criteria of orthologous group selection are mainly based on four features: database source, group evidence level, number of matched sequences in the orthologous group and sequence evidence level. Specifically, orthologous groups from the database with highest priority score in Supplementary Table S1 will be taken into account in the first place. Priority scores assigned to each database are determined on the basis of their curator, the number of protein sequences used, and the frequency of database release. The Ensembl and InParanoid databases get relatively low priority due to their pairwise comparison implementation, which may reduce the reliability for some conserved function group

detection. If more than one orthologous group comes from the same database or databases with the highest priority score, orthologous groups with the best group evidence level will be taken into account. The group evidence level is based upon the status of manual curation and only available in orthologous groups from KEGG, ProtClustDB and PIRSF. In case when several orthologous groups have the same group evidence level, the one with the highest number of matched proteins will be chosen. The annotation of each gene model is inherited from the description of the assigned orthologous group if available. If the orthologous group description is not present, a combination of the descriptions of those proteins with the best sequence evidence level within the assigned orthologous group will be made. The orthologous relationships between the species of gene models and its closest model organism(s) according to the relationships in NCBI taxonomy database will be deduced using both the OrthoMCL and DODO software packages [18, 19].

2.2.2 Transcriptomics analysis

MADMAX's microarray analysis support is tightly integrated with the R and Bioconductor projects. R and Bioconductor are updated biannually and so is the MADMAX microarray analysis system. Annotation libraries are up-to-date and - consequently - new innovative scientific algorithms are available and accessible. The MADMAX analysis pipelines are split in 2 parts: microarray quality control and microarray statistical analysis. The quality control pipeline supports the 3 major microarray platforms: Affymetrix (via the affy and xps libraries [20]), Illumina (via the lumi library [21]), and Agilent (via the Limma library[22-26]). Madmax supports the latest microarray chips, such as the Affymetrix GeneTitan plates. The quality control pipeline creates an extensive quality control report in the pdf-format that is easy to browse. It contains various informative and descriptive plots, such as probe intensity plots (before and after normalization), clustering plots (e.g. PCA and correlation plots), NuSE and RLE plots, and so forth. In addition, an Excel file with normalized intensities and an extensive annotation (gene names, descriptions, chromosomal locations, GO descriptions, etc.) is generated and available for direct use. The statistical pipeline utilizes the Limma package and its Bayesian correction to obtain either paired or unpaired t-test results including FDR-corrected p-values. Again, an extensive Excel file is generated with the statistical results including extensive gene annotations. In addition, the Limma output is utilized for conducting GO enrichment analysis using ermineJ and performing Gene Set Enrichment Analysis (GSEA). The major microarray platforms are supported: Illumina and Agilent arrays are, during the normalization step, converted in a so-called expressionSet that serves as input for the other steps: group assignment, filtering, and the final statistics including ermineJ [27] and GSEA. For the Affymetrix platform, MADMAX lets the user decide whether the (outdated) Affymetrix probe annotation or a recent custom probe annotation is being used. The custom probe annotation libraries are provided by the Microarray Lab, Department of Psychiatry / Molecular and Behavioral Neuroscience Institute, University of Michigan and are updated once or twice a year.

2.2.3 Metabolomics analysis

Metabolomics analysis capability has been recently added to the MADMAX LIMS system. Initial capabilities are limited to targeted search capabilities in preprocessed metabolomics data. Metabolomics platforms for which analysis tools are provided are LCMS[28, 29], GCMS and GCGCMS. Preprocessed data can be acquired using the MetAlign software and stored in MADMAX. Targeted searches of the stored data involve matching of sample data to available libraries of annotated compound data. Two types of compound libraries can be stored in MADMAX: firstly, LCMS compatible entries containing compound identifier, compound retention and mass and retention and mass window used in searching sample data.

Secondly, GCMS / GCGCMS compatible NIST formatted spectral fingerprint libraries containing compound identifiers and fragment mass and (relative) intensity information. Using either publicly available libraries or personal, proprietary libraries MADMAX users can annotate their metabolomics samples. LCMS results contain quantitative information on the assigned compound in each sample, where GCMS/GCGCMS adds quality parameters based on fragment library match completeness. The metabolomics analysis tools available to MADMAX are implemented as web services, separating the computationally intensive algorithms from the database. Current on-going development aims to add untargeted data analysis and metabolomics sample alignment to MADMAX.

2.3 Graphical web user interface

MADMAX was developed as a database-driven web application using Oracle Application Express. Owing to the management and analysis purpose of MADMAX, the web interface constitutes four main tabs, after users pass the authentication from the login page. The "Home" tab displays a short description of the system and information about recent changes and/or updates; users are able to change their personal information in the "User Information" tab; metadata information of the experiment can be collected/viewed in "Experiment" tab; execution of different analysis pipelines can be triggered in "Analysis" tab; and finally, both individual analysis results and integrated results are available from the "Results" tab.

Any omics-based experiment contains metadata and raw/derived data generated using different omics techniques. The metadata are categorized into Investigation, Study and Assay in ISA-Tab format and are collected in the "Experiment", "Protocol", "Study" and "Assay" pages under the "Experiment" tab in MADMAX. LIMS data of the experiment will be collected in the system directly or with the reference URI in "LIMS data upload" page and can be retrieved in "LIMS data download" page. A complete ISA-Tab compliant experiment can be imported to MADMAX via the "Import" page as long as investigation, study and assay files are all uploaded successfully.

Microarray analysis listed in "Analysis" tab can be executed for either quality control or statistical purpose. When the analysis is submitted, user will be redirected to a page with frequent updated progress information. The metabolomics analysis pipeline can be reached from the "Metabolomics Analysis" page.

Gene model annotation of newly sequenced species will be shown on the "Genomics Analysis Results" page. On the same page, experiments with well-studied species like model organisms will use the existing gene model (protein) annotation in UniProt. Both microarray (quality control and statistical analysis) and metabolomics analysis will generate an archive file and a separate text file which will be used to nail down a set of genes if the run is chosen as standard for the assay. The text file from the microarray analysis is the normalized fold change of each probe and the one from metabolomics is the normalized intensity of targeted metabolites. Different gene lists can be obtained either from different analysis rounds of the same dataset or from the analysis of different ~omics data. These separate gene lists can subsequently be integrated in the "Integration" page under the "Results" tab.

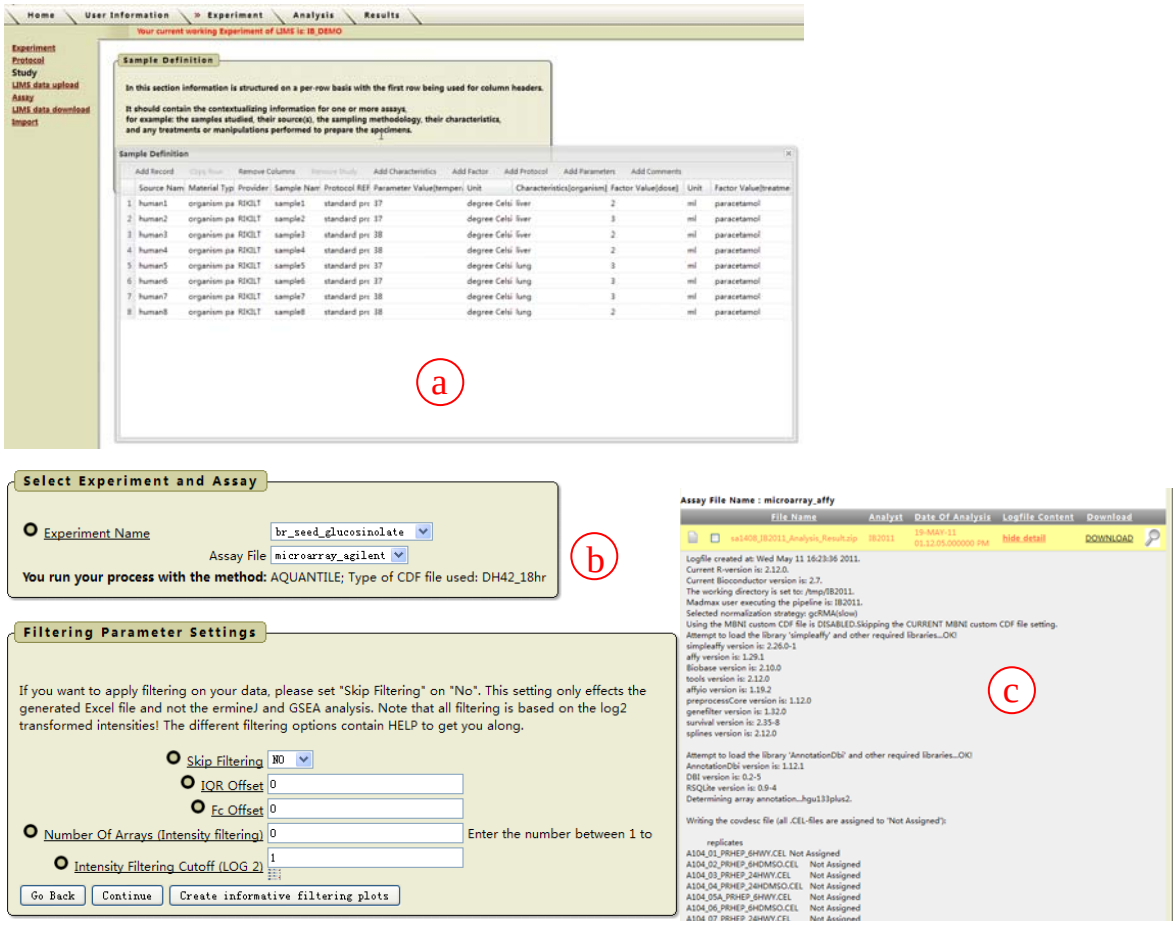

Figure 2: The web user interfaces of MADMAX. a. Form to fill in the sample related information which is content of study file in ISA-Tab specification, eg. sampling methodology, characteristics, factor value of each sample. b. Microarray statistical analysis interface in MADMAX. It need normalization, group assignment and filtering steps before user can run statistical tests. c. Analysis results download page. The individual analysis results of each assay can be downloaded and the detailed process information can be viewed as well. The highlighted results are standard of each assay which can be used for the integration later.

3 Case Study

Brassicaceae, the mustard or cabbage family, includes several well-known species such as the crop species *Brassica rapa* (Chinese cabbage, turnip, etc.), *Brassica oleracea* (cauliflower, broccoli, etc.), *Brassica nigra* (black mustard, etc.), *Brassica napus* (rapeseed, rutabaga, etc.), *Brassica carinata* (Ethiopian mustard, etc.), *Brassica juncea* (Indian mustard, etc.) and the model species *Arabidopsis thaliana*). *Brassica rapa* is together with *B. oleracea* one of the most important economic vegetable crops all over the world. In the wake of discovery on the healthy function of a subset of the secondary metabolites glucosinolates, *Brassicaceae* also become more significant for nutritional studies. In order to determine which genes may take effect on producing seed- or seedling-specific glucosinolates in *Brassica rapa*, we combined a *de novo* sequenced *Brassica rapa* genome using next-generation sequencing technology with an experiment on transcript profiling of developing seeds from two genotypes over 6 time points after flower opening, as well as a targeted metabolomics study on mature seeds.

3.1 Genomics analysis

The reference genome sequence of *B. rapa L. ssp. pekinensis* inbred line “Chiifu” (~70x genome coverage) and the re-sequenced genome sequences of two other *B. rapa* accessions

(one rapid cycling oil-like accession and Japanese turnip accession, 27x genome coverage) has been sequenced and will be released publically soon (submitted to Nature Genetics). In total, 41,132 gene models are predicted from 43,201 scaffolds [30]; after orthologous group assignment, 29,009 gene models can be mapped to existing orthologous groups which have annotations either at the group level or protein member level. Further analysis reveals that 8,635 of 29,009 genes have strong existence evidence on either group level (full or full/description) or protein level (1 or 2). Orthologous relationships between *Arabidopsis thaliana* and *Brassica rapa* were performed using two software tools: DODO and OrthoMCL. The first program found 17,677 pairs, the latter 26,226, while they have 13,135 genes pairs in common. A keyword search of 23 terms used against protein annotations, which contain all passed orthologous groups rather than the best one used in the final assignment, resulted in 514 genes that are potential glucosinolate-related (Supplementary Table S2). Since glucosinolates have been well studied in *Arabidopsis thaliana*, to which *Brassica rapa* is evolutionary relatively close (both belong to the *Brassicaceae*), the orthologous relationships between the two genomes can be used to discover candidate genes resulting in 57 genes. Both the keywords and *Arabidopsis thaliana* genes used for the genomics analysis are coming from the latest version of known glucosinolate related metabolic pathways of *Arabidopsis thaliana*.

3.2 Microarray analysis

Based on the reference genome of *Brassica rapa*, a customized 8 * 60 K Agilent microarray was created for transcription profiling. Two double haploid (DH) lines (DH42 and DH78) from a DH population serived from a cross between Yellow Sarson and Pak Choi as female and male parent respectively are used in a pilot study to identify a stage of seed development when genes related to seed filing process are differentially expressed. The two parents differ in their seed color and flowering time: YS143 is yellow-seeded early flowering and PC175 is black-seeded and intermediate flowering. The two progeny lines DH line 78 is late flowering and DH line 42 is early flowering line. Seeds are the basic unit of plant life and play multiple roles of maintaining life, like establishment of the crop, but seeds can also represent the harvested plant parts, like for oilseed rape. To study the genetics of metabolite biosynthesis processes in seed, a gene expression study was conducted using developing seeds collected after 18, 20, 25, 30, 35 and 40 DAP (days after pollination) on 2-colour 60-mer microarrays. Two parallel experiments were carried out in a double-loop experiment on a pairwise combination of the parental accessions YS143 and PC175, and two progeny DH lines DH 78 and DH 42. Each experiment consisted of 46 samples from two genotypes, six time points and dye-swap between two samples within the same time points. After applying the quality control pipeline, one array (name genotype and time points used) had extremely low average intensities after background correction and hence was marked as bad quality, and omitted for the later analysis. As learned from previous studies on both *Arabidopsis thaliana* and *Brassica napus*, there is a certain time frame in which the gene expression related to seed metabolite regulating genes including glucosinolates occurs. After comparing a number of different group assignment combinations, it was concluded that at time points 25 and 30 days the expression intensity of the genes selected for their relationship with/role in glucosinolate metabolism of two DH lines can generate the largest amount of differentially expressed genes.

3.3 Metabolomics analysis

Different types of GLS (glucosinolate) content were measured from mature seed of the DH population to understand the genetics of GLS metabolite content in seed. Seed metabolites content will be correlated with seed and seedling vigour traits to investigate the role of these metabolites in those traits. For Canola quality B. napus, glucosinolate levels need to be very low. The targeted metabolomics analysis showed 14 glucosinolate-related metabolites

(Supplementary Table S3); 11 of them have found with metabolomics search function in MADMAX in two genotypes while benzyl glucosinolate, 3-methylthiopropyl and hydroxybenzyl show no signal in both DH lines. Kegg compound ids are first retrieved by the names of the 11 metabolites and then 49 enzymes which have a link to the compound IDs are found using kegg webservices [31]. The enzymes EC numbers are used to search against the same protein annotation set used in genomics analysis which results in 369 gene models related.

In total 20 candidate genes which are supposed to be responsible for the different glucosinolate levels between two DH lines are discovered by combining all genomics, transcriptomics and metabolomics experimental results (Supplementary Table S4). There are only enzymes included in the result lists because of the way used for the metabolomics analysis. It is well studied in *Arabidopsis* and *Brassica* that transcription factors like myb gene family can regulate the biosynthesis of glucosinolates [32, 33]. Using word "myb" to search against the assigned gene annotations gives 314 hits. These genes combined with the over-/under-expressed genes provide another 30 candidate genes where Bra002610, Bra006422, Bra006370 and Bra030743 can be mapped to the same scaffold as the highlighted three candidate genes (of 20) in table S4. Paps sulfotransferase, UDP-glucosyltransferase and cytochrome P450 are known to be expressed differentially during seed development from the previous study in Brassica species [34-36].

4 Discussion

The recent exponential increase in sequencing projects also accelerates the discovery rate of how phenotypes emerge for newly sequenced species at other ~omics levels. Managing, sharing, and comparing datasets from different high-throughput techniques thus have become in urgent demand and pose a big challenge for bioinformaticians. There are many freely available analysis systems for a single type of ~omics data but only few of them can also store the metadata, such as Galaxy for genomics, Gene Expression Browser for transcriptomics and metaP-server for metabolomics [37-39]. The MADMAX database was built to store, analyse and integrate the various ~omics datasets together, including their associated metadata. Besides of systems developed at the NCBI and EBI which cannot be installed locally, the only one which claims multiple ~omics analysis integration is Babelomics [40]. It can handle data from transcriptomics, proteomics and genomics experiments individually but not the storage and management of metadata. In contrast to the before-mentioned systems, the most predominant feature of MADMAX is the integration of analysis results from different ~omics level of an experiment. Hence the system can determine the correlation between experiments based on the analysis results of them. Concretely, the intersection of gene lists generated from different analysis pipelines in one experiment will be used to calculate the similarity coefficient with the one in another experiment. It helps biologists to determine which conditions may correspond to certain gene expression levels. Therefore, conditions can still be summarized and categorized manually even though some of the linked metadata do not use the controlled vocabulary or use different terms of a controlled vocabulary.

Due to the reduced price of sequencing, new developed markers are mostly indels and SNPs that can be directly linked to the genome. Hence, normal QTL analysis can be merged into MADMAX for scaling down the candidate gene lists from a genetical point of view. The traits used for QTL analysis can be either a normal phenotype or gene expression/metabolite abundance. As mentioned in the case study, we used a rather unbalanced data set to come to the candidate lists which is quite different then in genetic studies where data of all genotypes are needed. Co-localizing QTLs and candidate genes from current analysis can be the extra filter. If QTLs are mapped to the regions which are distantly away from candidate genes, it

becomes quite interesting to figure out which individual ~omics analysis results can find back the pairs of QTLs and their localized genes. In other cases, QTL map position can also be very important information to validate the function annotation of genes we assigned in the system.

Another requirement that may arise soon is the handling of gene expression profiling assays using the RNA-Seq technology. Moreover, special emphasis should be put on improvement of gene function annotation by incorporating statistical models into the current pipeline implementation.

Acknowledgements

The authors would like to thank all colleagues and students who contributed to this study. We are grateful to Harm Nijveen for his help in system maintenance, and to Edouard Severing who assisted in setting up the access to the ProGMap database. This work was supported by the BioRange Programme of the Netherlands BioInformatics Centre (NBIC, to AKG).

References

[1] Rocca-Serra, P., M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, and S.A. Sansone. Isa Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level. Bioinformatics 26(18):2354-6, 2010.

[2] Subramanian, A., P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43):15545-50, 2005.

[3] R Development Core Team. R: A Language and Environment for Statistical Computing. 2011.

[4] Reimers, M. and V.J. Carey. Bioconductor: An Open Source Framework for Bioinformatics and Computational Biology. Methods Enzymol, 411:119-34, 2006.

[5] Lommen, A. Metalign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. Anal Chem, 81(8):3079-86, 2009.

[6] Kuzniar, A., K. Lin, Y. He, H. Nijveen, S. Pongor, and J.A.M. Leunissen. ProGMap: An Integrated Annotation Resource for Protein Orthology. Nucleic Acids Res, 37(Web Server issue):W428-34, 2009.

[7] Kersey, P.J., D. Lawson, E. Birney, P.S. Derwent, M. Haimel, J. Herrero, S. Keenan, A. Kerhornou, G. Koscielny, A. Kahari, R.J. Kinsella, E. Kulesha, U. Maheswari, K. Megy, M. Nuhn, G. Proctor, D. Staines, F. Valentin, A.J. Vilella, and A. Yates. Ensembl Genomes: Extending Ensembl across the Taxonomic Space. Nucleic Acids Research, 38(Database issue):D563-9, 2010.

[8] Ostlund, G., T. Schmitt, K. Forslund, T. Kostler, D.N. Messina, S. Roopra, O. Frings, and E.L. Sonnhammer. Inparanoid 7: New Algorithms and Tools for Eukaryotic Orthology Analysis. Nucleic Acids Research, 38(Database issue):D196-203, 2010.

[9] Chen, F., A.J. Mackey, C.J. Stoeckert, Jr., and D.S. Roos. Orthomcl-Db: Querying a Comprehensive Multi-Species Collection of Ortholog Groups. Nucleic Acids Research, 34(Database issue):D363-8, 2006.

[10] Tatusov, R.L., N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale. The Cog Database: An Updated Version Includes Eukaryotes. BMC Bioinformatics, 4:41, 2003.

[11] Muller, J., D. Szklarczyk, P. Julien, I. Letunic, A. Roth, M. Kuhn, S. Powell, C. von Mering, T. Doerks, L.J. Jensen, and P. Bork. Eggnog V2.0: Extending the Evolutionary Genealogy of Genes with Enhanced Non-Supervised Orthologous Groups, Species and Functional Annotations. Nucleic Acids Research, 38(Database issue):D190-5, 2010.

[12] Sayers, E.W., T. Barrett, D.A. Benson, E. Bolton, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, S. Federhen, M. Feolo, I.M. Fingerman, L.Y. Geer, W. Helmberg, Y. Kapustin, D. Landsman, D.J. Lipman, Z. Lu, T.L. Madden, T. Madej, D.R. Maglott, A. Marchler-Bauer, V. Miller, I. Mizrachi, J. Ostell, A. Panchenko, L. Phan, K.D. Pruitt, G.D. Schuler, E. Sequeira, S.T. Sherry, M. Shumway, K. Sirotkin, D. Slotta, A. Souvorov, G. Starchenko, T.A. Tatusova, L. Wagner, Y. Wang, W.J. Wilbur, E. Yaschenko, and J. Ye. Database Resources of the National Center for Biotechnology Information. Nucleic Acids Research, 39(Database issue):D38-51, 2011.

[13] Klimke, W., R. Agarwala, A. Badretdin, S. Chetvernin, S. Ciufo, B. Fedorov, B. Kiryutin, K. O'Neill, W. Resch, S. Resenchuk, S. Schafer, I. Tolstoy, and T. Tatusova. The National Center for Biotechnology Information's Protein Clusters Database. Nucleic Acids Research, 37(Database issue):D216-23, 2009.

[14] Nikolskaya, A.N., C.N. Arighi, H. Huang, W.C. Barker, and C.H. Wu. Pirsf Family Classification System for Protein Functional and Evolutionary Analysis. Evolutionary Bioinformatics Online, 2:197-209, 2006.

[15] Schneider, A., C. Dessimoz, and G.H. Gonnet. OMA Browser--Exploring Orthologous Relations across 352 Complete Genomes. Bioinformatics, 23(16):2180-2, 2007.

[16] Kanehisa, M., M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. Kegg for Linking Genomes to Life and the Environment. Nucleic Acids Research, 36(Database issue):D480-4, 2008.

[17] Camacho, C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T.L. Madden. BLAST+: Architecture and Applications. BMC Bioinformatics, 10:421, 2009.

[18] Li, L., C.J. Stoeckert, Jr., and D.S. Roos. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. Genome Research, 13(9):2178-89, 2003.

[19] Chen, T.W., T.H. Wu, W.V. Ng, and W.C. Lin. Dodo: An Efficient Orthologous Genes Assignment Tool Based on Domain Architectures. Domain Based Ortholog Detection. BMC Bioinformatics, 11(Suppl_7):S6, 2010.

[20] Gautier, L., L. Cope, B.M. Bolstad, and R.A. Irizarry. Affy-Analysis of Affymetrix Genechip Data at the Probe Level. Bioinformatics, 20(3):307-15, 2004.

[21] Du, P., W.A. Kibbe, and S.M. Lin. Lumi: A Pipeline for Processing Illumina Microarray. Bioinformatics, 24(13):1547-8, 2008.

[22] Ritchie, M.E., J. Silver, A. Oshlack, M. Holmes, D. Diyagama, A. Holloway, and G.K. Smyth. A Comparison of Background Correction Methods for Two-Colour Microarrays. Bioinformatics, 23(20):2700-7, 2007.

[23] Ritchie, M.E., D. Diyagama, J. Neilson, R. van Laar, A. Dobrovic, A. Holloway, and G.K. Smyth. Empirical Array Quality Weights in the Analysis of Microarray Data. BMC Bioinformatics, 7:261, 2006.

[24] Smyth, G.K. Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments. Statistical Applications in Genetics and Molecular Biology: 3(1):3, 2004.

[25] Smyth, G.K. and T. Speed. Normalization of cDNA Microarray Data. Methods, 31(4):265-73, 2003.

[26] Smyth, G.K., J. Michaud, and H.S. Scott. Use of within-Array Replicate Spots for Assessing Differential Expression in Microarray Experiments. Bioinformatics, 21(9):2067-75, 2005.

[27] Lee, H.K., W. Braynen, K. Keshav, and P. Pavlidis. Erminej: Tool for Functional Analysis of Gene Expression Data Sets. BMC Bioinformatics, 6:269, 2005.

[28] Gerssen, A., P.P. Mulder, and J. de Boer. Screening of Lipophilic Marine Toxins in Shellfish and Algae: Development of a Library Using Liquid Chromatography Coupled to Orbitrap Mass Spectrometry. Analytica Chimica Acta, 685(2):176-85, 2011.

[29] Lommen, A., A. Gerssen, J.E. Oosterink, H.J. Kools, A. Ruiz-Aracama, R.J. Peters, and H.G. Mol. Ultra-Fast Searching Assists in Evaluating Sub-Ppm Mass Accuracy Enhancement in U-Hplc/Orbitrap Ms Data. Metabolomics : Official Journal of the Metabolomic Society, 7(1):15-24, 2011.

[30] The Brassica rapa Genome Sequencing Project Consortium. The Genome of the Mesohexaploid Crop Species *Brassica Rapa*. Nature Genetics, 2011. (accepted)

[31] Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The Kegg Resource for Deciphering the Genome. Nucleic Acids Research, 32(Database issue):D277-80, 2004.

[32] Hirai, M.Y., K. Sugiyama, Y. Sawada, T. Tohge, T. Obayashi, A. Suzuki, R. Araki, N. Sakurai, H. Suzuki, K. Aoki, H. Goda, O.I. Nishizawa, D. Shibata, and K. Saito. Omics-Based Identification of Arabidopsis Myb Transcription Factors Regulating Aliphatic Glucosinolate Biosynthesis. Proceedings of the National Academy of Sciences of the United States of America, 104(15):6478-83, 2007.

[33] Yuan, Y., L.W. Chiu, and L. Li. Transcriptional Regulation of Anthocyanin Biosynthesis in Red Cabbage. Planta, 230(6):1141-53, 2009.

[34] Hu, Y., G. Wu, Y. Cao, Y. Wu, L. Xiao, X. Li, and C. Lu. Breeding Response of Transcript Profiling in Developing Seeds of Brassica Napus. BMC Molecular Biology, 10:49, 2009.

[35] Klein, M. and J. Papenbrock. The Multi-Protein Family of Arabidopsis Sulphotransferases and Their Relatives in Other Plant Species. Journal of Experimental Botany, 55(404):1809-20, 2004.

[36] Mittasch, J., S. Mikolajewski, F. Breuer, D. Strack, and C. Milkowski. Genomic Microstructure and Differential Expression of the Genes Encoding Udp-Glucose:Sinapate Glucosyltransferase (Ugt84a9) in Oilseed Rape (Brassica Napus). TAG. Theoretical and Applied Genetics. 120(8):1485-500, 2010.

[37] Giardine, B., C. Riemer, R.C. Hardison, R. Burhans, L. Elnitski, P. Shah, Y. Zhang, D. Blankenberg, I. Albert, J. Taylor, W. Miller, W.J. Kent, and A. Nekrutenko. Galaxy: A Platform for Interactive Large-Scale Genome Analysis. Genome Research, 15(10):1451-5, 2005.

[38] Zhang, M., Y. Zhang, L. Liu, L. Yu, S. Tsang, J. Tan, W. Yao, M.S. Kang, Y. An, and X. Fan. Gene Expression Browser: Large-Scale and Cross-Experiment Microarray Data Integration, Management, Search & Visualization. BMC Bioinformatics, 11:433, 2010.

[39] Kastenmüller, G., W. Römisch-Margl, B. Wägele, E. Altmaier, and K. Suhre. metaP-Server: A Web-Based Metabolomics Data Analysis Tool. Journal of Biomedicine & Biotechnology, vol. 2011, Article ID 839862, 2011.

[40] Medina, I., J. Carbonell, L. Pulido, S.C. Madeira, S. Goetz, A. Conesa, J. Tarraga, A. Pascual-Montano, R. Nogales-Cadenas, J. Santoyo, F. Garcia, M. Marba, D. Montaner, and J. Dopazo. Babelomics: An Integrative Platform for the Analysis of Transcriptomics, Proteomics and Genomic Data with Advanced Functional Profiling. Nucleic Acids Research, 38(Web Server issue):W210-3, 2010.

A Supplementary Material

Table S1: Some statistics of 11 orthologous group databases used in the pipeline as part of ProGMap database. Priority scores assigned to each database are determined by their curate manner, number of protein sequences used and the frequency of database release. Ensembl and InParanoid database get relatively low priority due to their pairwise comparison implementation which may reduce the reliability for some conserved function group detection.

Database	Release Date	Number proteins	Number clusters	Number species	Manual Curate	Cluster description	Priority
KEGG	Apr 2011	6,375,611	14,396	1,446	Y	Y	5
ProtClustDB	Nov 2010	4,607,655	627,757	unknown	Y	Y	5
PIRSF	Mar 2011	1,952,041	33,741	unknown	Y	Y	5
OMA	Nov 2010	4,586,478	483,374	1,000	N	Y	4
eggNOG	Sep 2009	2,483,276	224,848	630	N	Y	4
OrthoMCL	Feb 2010	1,270,853	116,536	138	N	N	3
HomoloGene	Feb 2011	241,946	43,726	20	N	Y	3
Ensembl	Feb 2011	2,240,884	45,050,607	unknown	N	N	2
nParanoid	Jun 2009	1,940,193	21,378,256	100	N	N	2
KOG	May 2003	60,735	4,852	7	N	Y	1
COG	May 2003	62,139	5,665	66	N	Y	1

Table S2: Keyword list used in the genomics analysis and the number of matched gene models of each keyword in sequenced B*rassica rapa*. Terms used here are curated from the pathways glucosinolates are involved in in *Arabidopsis thaliana*. A complete list of the matched orthologous groups was used instead of the best one assigned to each gene model and the keywords were searched against matched protein member annotations in each group.

Keyword list	Number of hits	Keyword list	Number of hits
S-oxygenase	12	deacetoxyvindoline 4-hydroxylase	28
desulfoglucosinolate sulfotransferase	24	2-hydroxylase	35
UGT74B1	7	sulfotransferase	58
Alkylthiohydroximate,carbon-sulfur lyase	18	epithiospecifier	19
2-oxoglutarate-dependent dioxygenase	62	hydro-lyase	18
3-isopropylmalate dehydrogenase	13	phenylalanine N-hydroxylase	10
nitrile-specifier	16	methylthioalkylmalate synthase	11
amino acid aminotransferase	14	thiohydroximate	28
CYP79B	10	gamma-glutamyl peptidase	12
CYP79F	9	myrosinase	64
CYP81F	31	branched-chain-amino-acid transaminase	12
CYP83B	3	Total	514

Table S3: 14 targeted metabolites with their intensities in two DH lines.

Name	pv-Br088042	pv-Br088078	Name	pv-Br088042	pv-Br088078
2 propenyl or allyl glucosinolate	0.32	0.42	4-methylthiobutyl	0.25	0.04
3-Butenyl	124.56	138.61	phenethyl	0.04	1.98
4-pentenyl	0.41	34.34	5-methylthiopentyl	0.06	2.33
2-hydroxy-3-butenyl	0.04	3.49	hydroxybenzyl	0	0.02
2-hydroxy-4-pentenyl	0.04	0.17	3-indolylmethyl	0.01	0.07
benzyl glucosinolate	0	0	4-hydroxy-3-indolylmethyl	0.64	1.64
3-methylthiopropyl	0	0	underderivitized hoi	0.02	0.29

Table S4: 20 candidate genes which are supposed to be responsible for the different glucosinolate levels between two DH lines. Highlighted genes are located in the same scaffold with four hypothetical myb genes which are differentially expressed in our time course study.

GENE_ID	GROUP_ID	DESCRIPTION	GROUP_EVID
Bra015311	PIRSF000322	1-aminocyclopropane-1-carboxylate oxidase	preliminary
Bra021671	PIRSF000322	1-aminocyclopropane-1-carboxylate oxidase	preliminary
Bra032515	PIRSF000322	1-aminocyclopropane-1-carboxylate oxidase	preliminary
Bra021670	PIRSF000322	1-aminocyclopropane-1-carboxylate oxidase	preliminary
Bra001891	K13811	3-phosphoadenosine 5-phosphosulfate synthase [EC:2.7.7.4 2.7.1.25]	full
Bra033696	K13811	3-phosphoadenosine 5-phosphosulfate synthase [EC:2.7.7.4 2.7.1.25]	full
Bra027963	PIRSF000856	Paps sulfotransferase	full
Bra005921	PIRSF000856	Paps sulfotransferase	full
Bra027118	PIRSF000856	Paps sulfotransferase	full
Bra027666	PIRSF000856	Paps sulfotransferase	full
Bra008132	PIRSF000856	Paps sulfotransferase	full
Bra027117	PIRSF000856	Paps sulfotransferase	full
Bra030620	PIRSF000473	UDP-glucosyltransferase	full
Bra021743	PIRSF000473	UDP-glucosyltransferase	full
Bra033056	PIRSF000045	cytochrome P450 CYP2D6	preliminary
Bra025351	PIRSF000045	cytochrome P450 CYP2D6	preliminary
Bra017819	PIRSF000045	cytochrome P450 CYP2D6	preliminary
Bra017818	PIRSF000045	cytochrome P450 CYP2D6	preliminary
Bra011758	PIRSF000045	cytochrome P450 CYP2D6	preliminary
Bra002747	PIRSF000045	cytochrome P450 CYP2D6	preliminary