BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 9 2014, pages 1316–1318
doi:10.1093/bioinformatics/btt739

*Gene expression*

Advance Access publication January 8, 2014

# MAGIC: access portal to a cross-platform gene expression compendium for maize

Qiang Fu[1], Ana Carolina Fierro[1,2,3], Pieter Meysman[1], Aminael Sanchez-Rodriguez[1,2,3], Klaas Vandepoele[3,4], Kathleen Marchal[1,2,3,*] and Kristof Engelen[1,5]

[1]Center of Microbial and Plant Genetics, KU Leuven, Kasteelpark Arenberg 20, B-3001, Leuven, [2]Department of Information Technology, Ghent University - iMinds, 9050 Gent, Belgium, [3]Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, Ghent University, B-9052 Gent, Belgium, [4]Department of Plant Systems Biology, VIB, Technologiepark 927, Ghent University, B-9052 Gent, Belgium and [5]Fondazione Edmund Mach, Research and Innovation Centre, Via E. Mach, 1, 38010 San Michele all'Adige, Trento, Italy

Associate Editor: Martin Bishop

**ABSTRACT**

**Summary:** To facilitate the exploration of publicly available *Zea mays* expression data, we constructed a maize expression compendium, making use of an integration methodology and a consistent probe to gene mapping based on the 5b.60 sequence release of *Z. mays*. The compendium is made available through a web portal MAGIC that hosts a variety of analysis tools to easily browse and analyze the data. Our compendium is different from previous initiatives in combining expression values across different experiments by providing a consistent gene annotation across different platforms.

**Availability and implementation:** http://bioinformatics.intec.ugent.be/magic/

**Contact:** kathleen.marchal@ugent.be

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on June 12, 2013; revised on November 23, 2013; accepted on December 16, 2013

## 1 INTRODUCTION

Owing to its importance as sustainable food and feedstock, maize genomics is of high academic and industrial relevance. As a result, microarrays have been widely applied to interrogate the maize transcriptome, with currently >100 maize gene expression experiments being publicly available in online repositories such as GEO (Barrett *et al.*, 2011) and ArrayExpress (Parkinson *et al.*, 2009). However, cross-platform differences, the lack of consistent platform and measurement descriptions and inconsistent gene annotations in maize complicated the straightforward use of these data.

To integrate the data from different array platforms in a readily usable single compendium, we resolved gene annotation inconsistencies by reannotating probes of previously published *Zea mays* arrays using the published maize genome sequence (Schnable *et al.*, 2009) and made measurements comparable across different platforms/experiments using an adapted version of the data integration method described by Engelen *et al.* (2011). This resulted in a cross-platform expression compendium containing 1749 microarrays covering 24 690 genes. Additional gene information was integrated from various external sources. Experimental annotations were manually curated. A web access portal MAGIC with a specialized set of exploration and analysis functionalities provides public access to this compendium.

*To whom correspondence should be addressed.

## 2 MATERIALS AND METHODS

### 2.1 Compendium creation

The compendium itself corresponds to a matrix of which the rows correspond to *genes* and columns to *sample contrasts*. A sample contrast is defined as the comparison of the gene's expression between two different *biological samples,* one acting as a *test* and the other as a *reference*. Each value in the matrix then represents the expression change of a gene presented as the log-ratio of its expression in the test versus the reference samples.

*2.1.1 Probe reannotation* Probes were reannotated using the latest release of the maize genome. Original probe sequences, if available, were obtained from the respective platform annotation files or otherwise from GenBank based on the corresponding GI numbers. They were used as query against the curated 'Filtered Gene Set' (FGS) of the 5b.60 release of B73 maize genome using MegaBLAST version 2.2.17 (Zhang *et al.*, 2000). Both the gene and the transcript models of FGS are searched to increase the eventual hit rate. Different blast parameters were chosen for oligo and cDNA probes, respectively, as they considerably differ in length and specificity. For probes that mapped to multiple genes, we identified the most specific hit by comparing the hit quality across different targets. Only probes for which a sufficiently unique probe-to-gene mapping (details in Supplementary Methods) could be identified were retained for compendium construction.

*2.1.2 Expression data homogenization* Microarray experiments were retrieved from GEO and ArrayExpress. Each experiment can be composed of multiple arrays. A distinction is made between one- and two-channel arrays, which differ from each other in the number of samples tested per array. Irrespective of the used platform, raw expression intensities were extracted for each channel (sample) separately. To make measurements between one- and two-channel platforms comparable, all expression measurements were converted to log ratios (as contrasts) that compare the expression between a test and a reference sample, respectively, both from the same experiment. For each contrast, proper test and reference samples were assigned based on the corresponding experiment annotation. Raw expression data were subsequently normalized using dedicated procedures (Engelen *et al.*, 2011) and mapped to the corresponding genes based on the probe annotation (see Section 2.1.1).

### 2.2 Compendium annotation

To improve biological interpretation of the compendium, we integrated gene (row) and contrast (column) annotation from publicly available resources and curated all available information.

To facilitate gene selection, we included, next to gene ids from 5b.60 genome release, gene names from MaizeGDB (Lawrence *et al*., 2004) and Xref assignments from www.maizesequence.org to provide mapping from EntrezGene and UniProt ids. As for functional annotations, metabolic pathway information (version 2.0) from Gramene (Youens-Clark *et al*., 2010) and Gene Ontology annotations from www.maizesequence.org were provided.

To compensate for the often cryptic and incomplete condition annotations available in public expression repositories, we provided curated annotations incorporating the information from both online repositories and the corresponding publications. Note that in our compendium, expression values are represented as log-ratios of a contrast between two samples. Experimental annotation is provided both at the level of the individual sample and that of the contrast. Annotation at the sample level includes tissue, development stage and genotype specifications (breeding line). The first two are described using Plant Ontology (Avraham *et al*., 2008)-derived ontology terms, whereas genotype specifications are based on the names of cultivars or wild-type. At the contrast level, we associated perturbation annotations specified as a set of relevant properties and corresponding values of change that reflect the stimuli that trigger expression alterations in the test versus the reference.

### 2.3 Compendium exploration

To facilitate the exploration of the compendium, we constructed a web access portal MAGIC providing a set of analysis functionalities. At first, MAGIC allows users to specify their own subcompendium of interest, which only contains contrasts sharing the same characteristics. Users can choose between various predefined subcompendia. Subcompendia focusing on environmental perturbations and on comparisons between lines, between tissues and between development stages are available. Alternatively, users can generate customized subcompendia based on the sample and contrast annotations.

Once a (sub)compendium is selected, the system provides tools to explore and visualize the expression data in a *module*-centralized manner where a module is defined as a subset of the (sub)compendium containing the expression values of a set of genes under a set of contrasts. A module can be created starting from a query set of genes or contrasts to which contrasts or genes are added, respectively, based on their properties (such as the coexpression level, or the expression consistency). An existing module can easily be altered, merged with other modules or split into several modules. Each module can be visualized as an interactive heatmap accompanied with corresponding annotation information.

Because most of the microarray platforms that MAGIC relies on were developed before the genome release of maize, none of them cover the full FGS gene set, and overlap in measured genes can be low for some platform combinations (Supplementary Tables S2 and S3). We developed functionalities that help users to control the number of missing values in their analysis results (Website online help, section 6).

## 3 RESULTS AND DISCUSSION

The compendium available through MAGIC currently contains 24 690 genes and 1310 sample contrasts. It covers 62% of the genes in FGS of the 5b.60 release of B73 maize genome; the remaining genes are not represented on any of the 27 platforms included. The contrasts consist of 69 experiments obtained from GEO and ArrayExpress, amounting to 1749 microarrays. Details on the composition of the compendium in terms of the number of genes and experiments covered per platform can be found in Supplementary Tables S1 and S2. On average, a gene is measured in 9 of the 26 platforms and has been measured in 592 of 1310 contrasts (Supplementary Figs S1 and S2). The compendium can be downloaded through its webportal MAGIC. An elaborate online help, together with two tutorial case studies, illustrates how the various functionalities of MAGIC can be used to infer new biology. The case studies clearly illustrate how information from different platforms contributes to the results obtained by MAGIC, and how to cope with missing values that could become abundant if information from certain sets of platforms is combined.

In contrast to MAGIC, comparable initiatives treat data from different platforms or experiments separately. Genevestigator (Hruz *et al*., 2008) and CORNET (De Bodt *et al*., 2012) construct separate compendia for the Affymetrix Maize Genome Array (GPL4032) (containing 558 and 340 arrays for Genevestigator and CORNET) and the Nimblegen Maize 385 k Array (GPL12620) (containing 180 arrays in both systems). PLEXdb (Dash *et al*., 2011), on the other hand, provides access to the data from 44 Affymetrix and Nimblegen experiments. In this system, data derived from each experiment are treated separately instead of being merged in a larger compendium. Of all these systems, only CORNET provides a restricted meta-analysis tool that allows combining information across the different compendia. Compared with these related initiatives, our approach is unique in directly combining data from different platforms in a single compendium, obviating the need for an additional meta-analysis step (Fierro *et al*., 2008) and enabling the construction of a much larger compendium and the direct data analysis across different platforms and experiments.

## ACKNOWLEDGEMENTS

The authors thank Lucia Pannier and Nicolas Dierckxsens for their contributions.

*Funding*: This work is supported by (i) KU Leuven: [GOA/08/011, PF/10/010 (NATAR)], (ii) Agentschap voor Innovatie door Wetenschap en Technologie (IWT): NEMOA, (iii) Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) G.0428.13N and (iv) Ghent University [Multidisciplinary Research Partnership 'N2N'].

*Conflict of Interest*: none declared.

## REFERENCES

Avraham,S. *et al.* (2008) The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Res*., **36**, D449–D454.

Barrett,T. *et al.* (2011) NCBI GEO: archive for functional genomics data sets—10 years on. *Nucleic Acids Res*., **39**, D1005–D1010.

Dash,S. *et al.* (2011) PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res*., **40**, D1194–D1201.

De Bodt,S. *et al.* (2012) CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations. *New Phytol*., **195**, 707–720.

Engelen,K. *et al.* (2011) COLOMBOS: access port for cross-platform bacterial expression compendia. *PLoS One*, **6**, e20938.

Fierro,A.C. *et al.* (2008) Meta analysis of gene expression data within and across species. *Curr. Genomics*, **9**, 525–534.

Hruz,T. *et al.* (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinformatics*, **2008**, 420747.

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

Parkinson,H. *et al.* (2009) ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Res.*, **37**, D868–D872.

Schnable,P.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Youens-Clark,K. *et al.* (2010) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.*, **39**, D1085–D1094.

Zhang,Z. *et al.* (2000) A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.*, **7**, 203–214.