GRASSIUS: A Platform for Comparative Regulatory Genomics across the Grasses[1][W][OA]

Alper Yilmaz[2], Milton Y. Nishiyama Jr.[2], Bernardo Garcia Fuentes, Glaucia Mendes Souza, Daniel Janies, John Gray, and Erich Grotewold*

Department of Plant Cellular and Molecular Biology and Plant Biotechnology Center (A.Y., B.G.F., E.G.), and Department of Biomedical Informatics (D.J.), The Ohio State University, Columbus, Ohio 43210; Instituto de Química, Departamento de Bioquímica, Universidade de São Paulo, São Paulo, Brazil (M.Y.N., G.M.S.); and Department of Biology, University of Toledo, Toledo, Ohio 43606 (J.G.)

Transcription factors (TFs) are major players in gene regulatory networks and interactions between TFs and their target genes furnish spatiotemporal patterns of gene expression. Establishing the architecture of regulatory networks requires gathering information on TFs, their targets in the genome, and the corresponding binding sites. We have developed GRASSIUS (Grass Regulatory Information Services) as a knowledge-based Web resource that integrates information on TFs and gene promoters across the grasses. In its initial implementation, GRASSIUS consists of two separate, yet linked, databases. GrassTFDB holds information on TFs from maize (*Zea mays*), sorghum (*Sorghum bicolor*), sugarcane (*Saccharum* spp.), and rice (*Oryza sativa*). TFs are classified into families and phylogenetic relationships begin to uncover orthologous relationships among the participating species. This database also provides a centralized clearinghouse for TF synonyms in the grasses. GrassTFDB is linked to the grass TFome collection, which provides clones in recombination-based vectors corresponding to full-length open reading frames for a growing number of grass TFs. GrassPROMDB contains promoter and cis-regulatory element information for those grass species and genes for which enough data are available. The integration of GrassTFDB and GrassPROMDB will be accomplished through GrassRegNet as a first step in representing the architecture of grass regulatory networks. GRASSIUS can be accessed from www.grassius.org.

A large fraction of the genome of any organism is dedicated to specify when, where, and how much of each mRNA needs to be produced. This regulatory information, hardwired into the genomic DNA, is essentially the same in every cell and largely constant over time and generations. Because these regulatory sequences are often in close proximity to the genes they control, we refer to them here as the cis-regulatory apparatus, which is formed by a mosaic arrangement of cis-regulatory elements (CREs). However, depending on the cell type or on the particular environmental circumstance, the same regulatory sequences can be interpreted in very different ways. It is the function of a group of trans-acting proteins, the transcription factors (TFs), to interpret the sequence code hardwired in the cis-regulatory apparatus and execute it in the form of a signal to the basal transcription machinery that will result in RNA production. TFs are organized into hierarchical gene regulatory networks in which one TF, often in cooperation with other proteins, positively or negatively regulates the expression of another TF. This establishes a variety of regulatory motifs, which, when assembled into regulatory modules, provide the free-scale architecture that characterizes gene regulatory networks (Babu et al., 2004; Yu and Gerstein, 2006). A first step in starting to build regulatory networks involves compiling the Parts List, which includes the TFs, promoters, CREs, and interactions between TFs and particular promoters (Schlitt and Brazma, 2007). Providing a comprehensive parts list is the main gap in our knowledge that GRASSIUS (Grass Regulatory Information Services) intends to fill. This is being done within the broader objective of linking regulatory networks and important agronomic traits in the grasses.

Several databases, including AtTFDB (http://arabidopsis.med.ohio-state.edu/AtTFDB; Davuluri et al., 2003), PlnTFDB (plntfdb.bio.uni-potsdam.de/v2.0; Riano-Pachon et al., 2007), PlantTFDB (planttfdb.cbi.pku.edu.cn; Guo et al., 2008), and DBD (dbd.mrc-lmb.cam.ac.uk; Kummerfeld and Teichmann, 2006; Wilson et al., 2008), contain information on plant TFs. In addition, a few databases also provide information on promoters (e.g. AtcisDB [Davuluri et al., 2003], PlantProm [Shahmuradov et al., 2003], and PPDB [Yamamoto and Obokata, 2008]). TF or promoter databases that focus solely on Arabidopsis (*Arabidopsis*

[1] This work was supported by the National Science Foundation (grant no. DBI–0701405 to J.G. and E.G.) and Fundação de Amparo à Pesquisa do Estado de São Paulo (grant to G.M.S.). G.M.S is also a recipient of a CNPq fellowship.

[2] These authors contributed equally to the article.

* Corresponding author; e-mail grotewold.1@osu.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Erich Grotewold (grotewold.1@osu.edu).

[W] The online version of this article contains Web-only data.

[OA] Open Access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.108.128579

thaliana) benefit from a well-annotated genome, a coherent nomenclature for TFs, and a large collection of full-length cDNAs (FLcDNAs), which permit the precise location of transcription start sites (TSSs) and hence promoters. Most of these resources are only now becoming available in the grasses and, while the time is ripe to start building the parts list for establishing regulatory network architecture, significant challenges remain.

Here, we describe the development of a first version of GRASSIUS (GRASSIUS v1 already deployed at www.grassius.org) as a knowledge-based public Web resource that integrates information on TFs (in the GrassTFDB database) and gene promoters (in the GrassPROMDB) for maize (*Zea mays*), rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), and sugarcane (*Saccharum* spp.), yet expected to expand to other grasses as genome information becomes available. In addition to providing the framework for building a comprehensive parts list, GRASSIUS also serves as a centralized clearinghouse for TF synonyms for the grasses. Combined with the discovery of phylogenetic relationships among members of TF families, and as a portal for available TF open reading frames (ORFs) in convenient recombination vectors, GRASSIUS provides a valuable resource for comparative regulatory genomics across the grasses.

RESULTS AND DISCUSSION

GRASSIUS furnishes a user-friendly online database tool developed as a comprehensive resource for retrieving information regarding the components involved in the regulation of gene expression across the grasses, initially focusing on maize, rice, sorghum, and sugarcane. GRASSIUS currently consists of two integrated databases, GrassTFDB and GrassPROMDB. As previously done for AGRIS (Palaniswamy et al., 2006), these databases are linked, providing a first step toward establishing regulatory motifs, the building stones of regulatory networks. Significantly different from AGRIS, however, GRASSIUS integrates data across multiple plant species and serves as a platform for comparative regulatory genomics. As the genomes being used in GRASSIUS continue to be analyzed, the importance of this database for researchers focusing on the grasses will increase. The rice genome is sequenced and annotated, providing plenty of resources regarding this plant. Likewise, the sorghum genome is sequenced, but the extent of the annotation or the availability of resources is not comparable with rice. In contrast, the maize genome is not yet fully sequenced and gene models are largely based on ab initio predictions. Sugarcane has the least information, and EST sequences are the major source in defining genes. Thus, GRASSIUS provides a platform for integrating resources related to regulatory genomics regardless of genome sequence availability and it is designed to grow to meet the future analytical needs of these species. Additionally, GRASSIUS offers information on resources for the experimentalist, including the TFome collection, phylogenetic trees that allow the identification of orthologous pairs and an incipient collection of minimal promoter regions experimentally shown to drive gene expression. Ultimately, GRASSIUS will furnish a venue for linking important agronomic traits to aspects related to the control of gene expression.

GrassTFDB

Whereas many other proteins participate in the regulation of gene expression, we limit here our definition of TFs to proteins that contain a characteristic structural motif, the DNA-binding domain, which is involved in recognizing a short (usually 4–8 bp) DNA sequence. Based on the structure of the DNA-binding domain, TFs are classified into a variable number of different families (usually 40–60), and in plants, 5% to 7% of all the protein-encoding genes correspond to TFs meeting these characteristics (Riechmann et al., 2000; Riechmann and Ratcliffe, 2000).

Because many of the grass genomes are not yet completely sequenced or annotated, the total number of TFs that should be expected is hard to predict. As a first step toward estimating the total number of TFs, particularly from maize and sugarcane where genomic information is either incomplete or missing, we performed a correlation between the number of genes in various genomes and the number of identified TFs. For

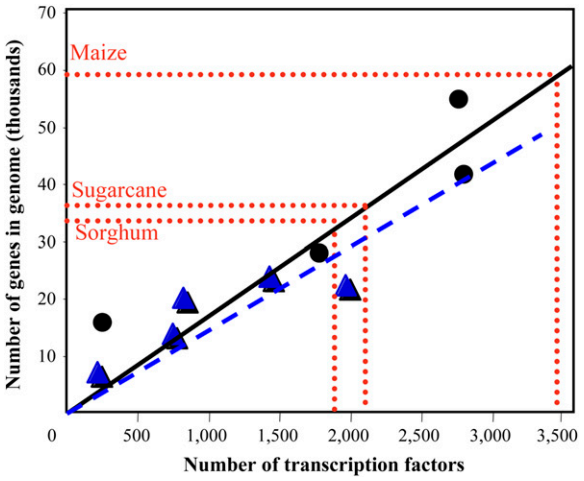


Figure 1. Estimation of TF numbers in grass genomes. Correlation between the number of TFs and the total number of genes in genomes was based on completely annotated plant genomes of Arabidopsis, rice, poplar, and Chlamy (black circles). A best-fit linear regression (r^2 = 0.87) was used to estimate the predicted number of TFs in maize, sorghum, and sugarcane (Table I). A similar analysis was conducted for nonplant organisms (blue triangles), including yeast (*Saccharomyces cerevisiae*), *Caenorhabditis elegans*, fruitfly (*Drosophila melanogaster*), mouse (*Mus musculus*), and human (*Homo sapiens*), and a best-fit linear regression line was drawn (r^2 = 0.74; blue dashed line).

Table I. *Number of grass TFs*

Expected number of TFs was estimated from Figure 1.

Genome Features	Maize	Sorghum	*Saccharum* spp.
Genome size (MB)	2,500	700	900
Total gene no.	59,000	36,338	33,620
Expected TF no.	3,470	2,137	1,997
TFs in GRASSIUS	3,337	2,448	1,647

sugarcane, the gene number was estimated based on the EST data from the SUCEST Project (Vettore et al., 2003). Based on the information available for Arabidopsis, rice, poplar (*Populus trichocarpa*), and Chlamy (*Chlamydomonas reinhardtii*) gene and TF numbers, a linear regression (best-fit line) was estimated (Fig. 1, solid line). To estimate the minimum number of TFs in maize, sorghum, and sugarcane, the size of the respective genomes was overlaid on the regression (Fig. 1, red dotted lines). This analysis suggests that a total of 3,470, 2,137, and 1,977 TFs could minimally be expected for maize, sorghum, and sugarcane, respectively (Table I). Of course, this analysis does not include TFs corresponding to families yet to be identified from the remaining fraction of plant genomes that remain as unknowns.

Interestingly, when a similar analysis was done for several nonplant genomes (fungal and animal), a similar trend was observed (Fig. 1, blue triangles). The fit of the regression ($r^2 = 0.74$, blue dashed line; Fig. 1) was significantly improved by combining the plant and nonplant points ($r^2 = 0.82$; data not shown), suggesting that plants and animals have a similar relationship between total gene numbers and TF numbers.

With these estimates in mind regarding the expected number of TFs that GRASSIUS should contain, we initiated the generation of GrassTFDB. As a first step, publicly available plant TFs from PlnTFDB, PlantTFDB, and DBD were obtained, and a comprehensive and nonredundant list of plant TFs was generated. Then, previously unidentified TFs were searched in the most recent genome sequence releases by scanning for PFAM domains found in plant TFs (Fig. 2). Predicted TFs in each species were sorted into 47 families, following criteria similar to those used for developing AtTFDB and AGRIS (Davuluri et al., 2003; Palaniswamy et al., 2006). GRASSIUS provides various ways to access information on TFs from the various families (Fig. 3).

For rice and sugarcane, the number of TFs currently present in GrassTFDB is very close to the predicted number of TFs (Table I), indicating that the database has good coverage. In the case of sorghum, the number of predicted TFs is higher than the expected number, suggesting that GrassTFDB may contain some duplicates and splice variants that should be collapsed into single TFs. For maize, the number of TFs in GrassTFDB is close to the expected number, in agreement with most of the coding region of the maize genome being already available. The contents of GrassTFDB also compare very favorably to other TF databases in terms of comprehensiveness (Table II).

When all TFs in GrassTFDB are arranged into species and families, interesting differences become evident, according to the online summary table furnished by GRASSIUS (http://grassius.org/summary.html).

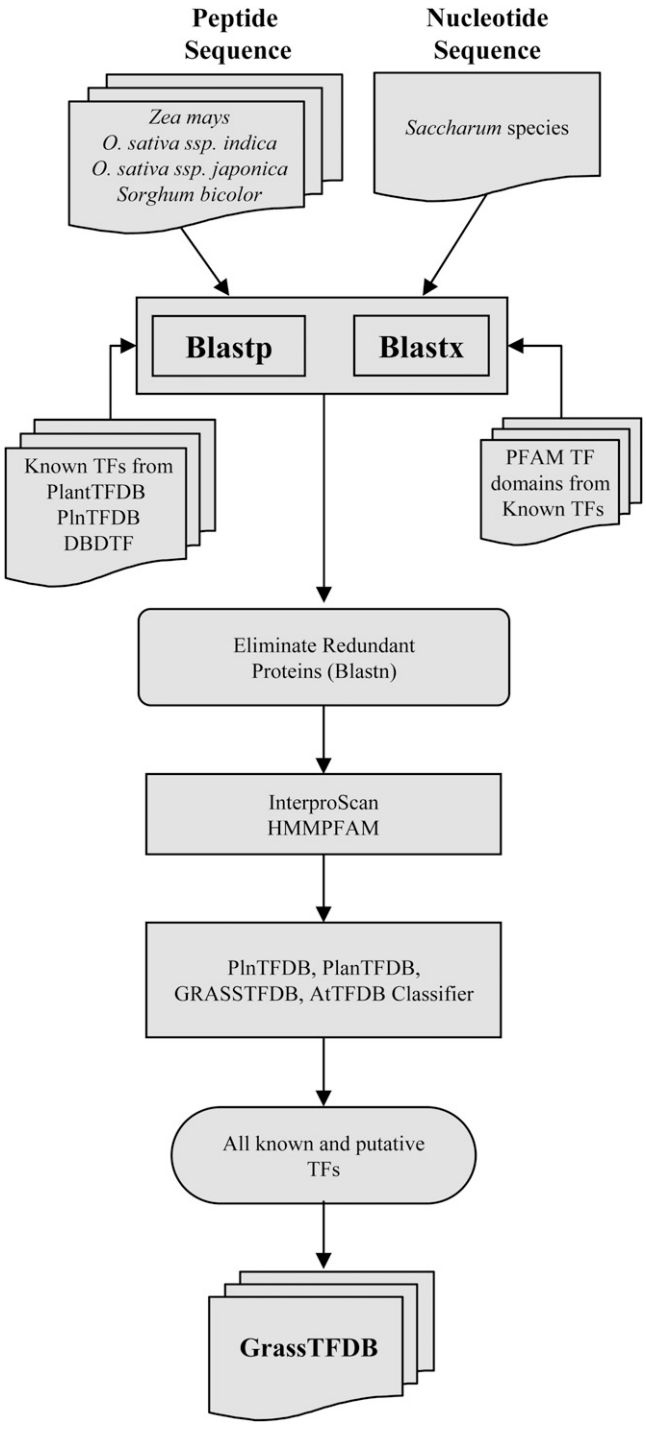


Figure 2. Flow diagram describing the steps involved in the generation of GrassTFDB. Details available in "Materials and Methods."

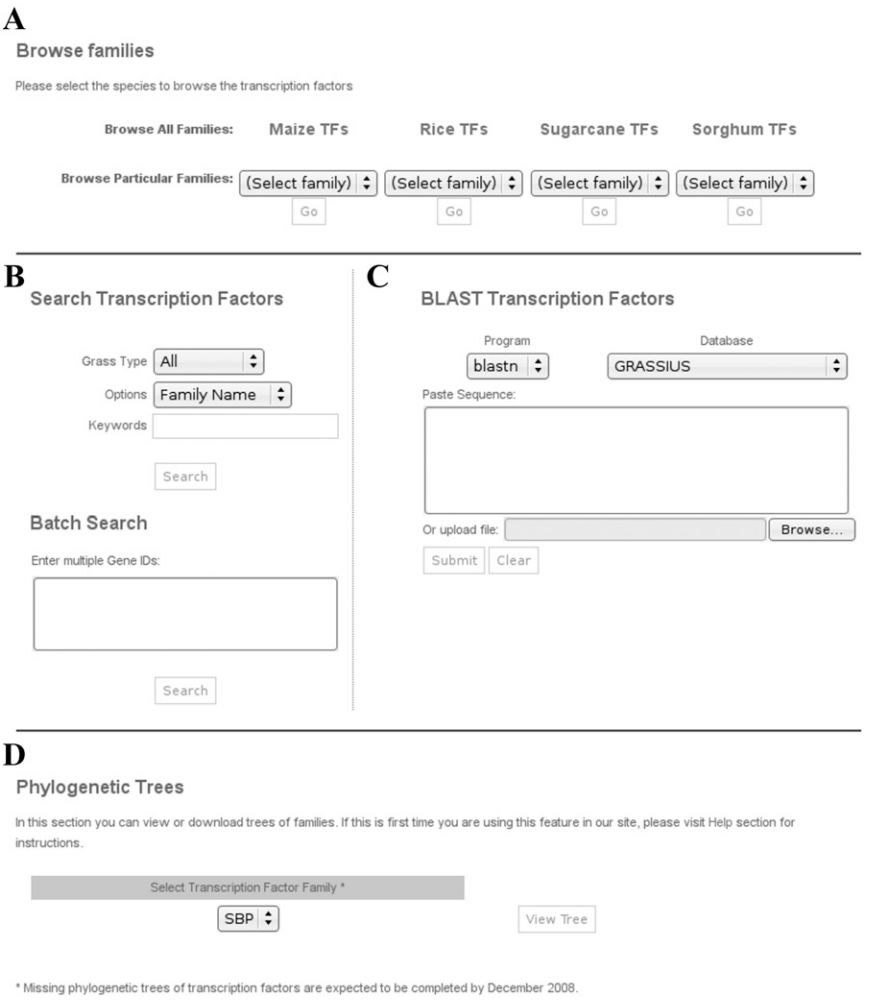


Figure 3. Screen shot showing query possibilities for the GrassTFDB database of GRASSIUS. A, All families in a species or members of a single family can be retrieved by clicking the species name or selecting the family name from the pull-down menu, respectively. B, Specific TFs or families can be retrieved by performing searches by selecting a particular family, or by keywords. Multiple TFs can be searched simultaneously by using the batch search option. C, BLAST application allows TFs to be searched in GrassTFDB by protein (blastp from the pull-down menu) or DNA (blastn from the pull-down menu) sequence. D, Phylogenetic trees of TF families can be retrieved by selecting the families.

For example, although maize has the largest gene count, the number of maize TFs in all families is not the highest. Maize has significantly more TFs only in the ABI3VP1, AP2-EREBP, bZIP, C2C2-YABBY, CPP, E2F-DP, Homeobox, Jumonji, MYB, NAC, SBP, and TUB families. In all other cases, the numbers are about the same as for the other grasses, with the exception of the GRAS family, which shows a significantly ($P < 0.05$; Weisberg t test) lower TF number. Similarly, the C2C2-CO-like family in rice has significantly fewer members than those found in the same family in the other species ($P < 0.05$; Weisberg t test), while the rice Trihelix family is significantly larger ($P < 0.05$; Weisberg t test). In sugarcane, the C3H family has significantly more members than the C3H families of maize, sorghum, or rice ($P < 0.05$; Weisberg t test). These trends may reflect the expansion/contraction of individual families in a particular taxon. The recent amplification of R2R3-MYB regulators during the radiation of the grasses (Dias et al., 2003) provides one possible mechanism for the expansion of particular families. Contraction could be associated with gene loss (Bennetzen, 2007) or domain loss, a phenomenon that has also been reported for MYB domains (Braun and Grotewold, 2001). The significant difference in the size of particular TF families between sugarcane and sorghum is potentially also of interest since the coding regions for the respective genomes have been shown to be 94.5% identical (Jannoo et al., 2007).

Phylogenetic Analysis of TF Families

An important function of GRASSIUS will be to provide information that facilitates comparative regulatory genomics studies. Central to this is the identification of orthologous TF pairs between the various grasses. Therefore, GRASSIUS contains an application that permits retrieval of preformed phylogenetic trees for a particular family (Fig. 3D; Supplemental Fig. S1).

Phylogenetic analyses were performed by aligning conserved domains of all members of a particular TF family and trees were constructed using RAXML (see "Materials and Methods"). The branches and nodes of the tree, visualized with A TREE VIEWER (ATV; Supplemental Fig. S1) are hyperlinked to the underlying data within GRASSIUS. A click on a terminal

Table II. *Comparison of GrassTFDB contents with other available plant TF databases*

Table summarizes the number of proteins classified as TFs in particular databases. When several gene models are available for a TF, primary gene models are considered when counting. In PlnTFDB and PlantTFDB, TFs in families that are shared between GrassTFDB are considered for comparison.

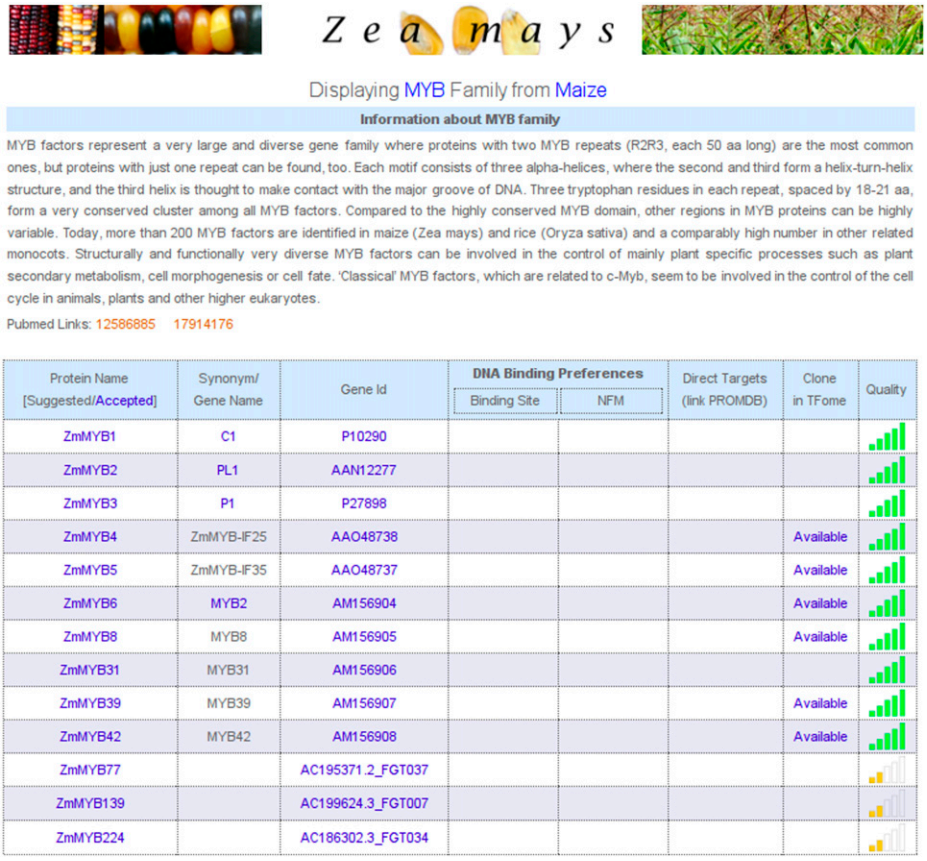
TF Databases	Maize	Rice (subsp. *japonica*)	Rice (subsp. *indica*)	Sorghum	*Saccharum* spp.
GrassTFDB	3,337	1,741	1,836	2,448	1,647
PlnTFDB[a]	N/A	1,875	N/A	N/A	N/A
PlantTFDB[b]	625	1,745	1,831	338	791
DBD[c]	673	1,626	N/A	1,452	N/A

[a]Plant Transcription Factor Database (http://plntfdb.bio.uni-potsdam.de). [b]Plant Transcription Factor Database (http://planttfdb.cbi.pku.edu.cn). [c]Transcription Factor Prediction Database (http://dbd.mrc-lmb.cam.ac.uk).

branch displays a single sequence. A click on an internal node displays all the data for the group of sequences that node subtends.

Establishment of the Grass TF ORFeome (TFome) Collection

The availability of FLcDNA clones and, in particular, the coding sequence prescribed by the ORF, for any given gene greatly advances the potential for further research of the protein encoded therein. Ready access to a cloned ORF accelerates the pace of research by permitting a variety of fusion and overexpression constructs to be engineered. Despite the large number of ESTs available through various projects, researchers are often lacking a FLcDNA for particular genes of interest. Thus, as part of the effort to establish a central resource for grass regulatory genomics, the development of a collection of clones containing ORFs for grass TFs was initiated (Supplemental Fig. S2). The clones in this collection are distinct from FLcDNA in that the coding sequence without 5′- and 3′-untrans-


Zea mays

Displaying MYB Family from Maize

Information about MYB family

MYB factors represent a very large and diverse gene family where proteins with two MYB repeats (R2R3, each 50 aa long) are the most common ones, but proteins with just one repeat can be found, too. Each motif consists of three alpha-helices, where the second and third form a helix-turn-helix structure, and the third helix is thought to make contact with the major groove of DNA. Three tryptophan residues in each repeat, spaced by 18-21 aa, form a very conserved cluster among all MYB factors. Compared to the highly conserved MYB domain, other regions in MYB proteins can be highly variable. Today, more than 200 MYB factors are identified in maize (Zea mays) and rice (Oryza sativa) and a comparably high number in other related monocots. Structurally and functionally very diverse MYB factors can be involved in the control of mainly plant specific processes such as plant secondary metabolism, cell morphogenesis or cell fate. 'Classical' MYB factors, which are related to c-Myb, seem to be involved in the control of the cell cycle in animals, plants and other higher eukaryotes.

Pubmed Links: 12586885 17914176

Protein Name [Suggested/Accepted]	Synonym/ Gene Name	Gene Id	DNA Binding Preferences		Direct Targets (link PROMDB)	Clone in TFome	Quality
			Binding Site	NFM			
ZmMYB1	C1	P10290					
ZmMYB2	PL1	AAN12277					
ZmMYB3	P1	P27898					
ZmMYB4	ZmMYB-IF25	AAO48738				Available	
ZmMYB5	ZmMYB-IF35	AAO48737				Available	
ZmMYB6	MYB2	AM156904				Available	
ZmMYB8	MYB8	AM156905				Available	
ZmMYB31	MYB31	AM156906					
ZmMYB39	MYB39	AM156907				Available	
ZmMYB42	MYB42	AM156908				Available	
ZmMYB77		AC195371.2_FGT037					
ZmMYB139		AC199624.3_FGT007					
ZmMYB224		AC186302.3_FGT034					

Figure 4. Screen shot of a part of the maize MYB family query result. A short description of the family is provided, including one or more key references. The first column indicates the name of the TF, following the guidelines provided (see *Letter to the Editor*, this issue [Gray et al., 2009]). Names in blue correspond to those accepted, those in gray correspond to those suggested, waiting comments by the community. The protein names provide clickable links to the general information page for each TF (Fig. 6). The Synonym/ Gene Name column provides alternate names by which the TF (or the gene encoding it) is known. Fields in blue indicate hyperlinks to other databases (such as, for example, MaizeGDB for maize). The Gene Id column provides links to species-specific external databases. If a clone is available for a particular TF in the TFome collection, or if direct targets for a TF are known, the corresponding columns provide links to the corresponding pages.

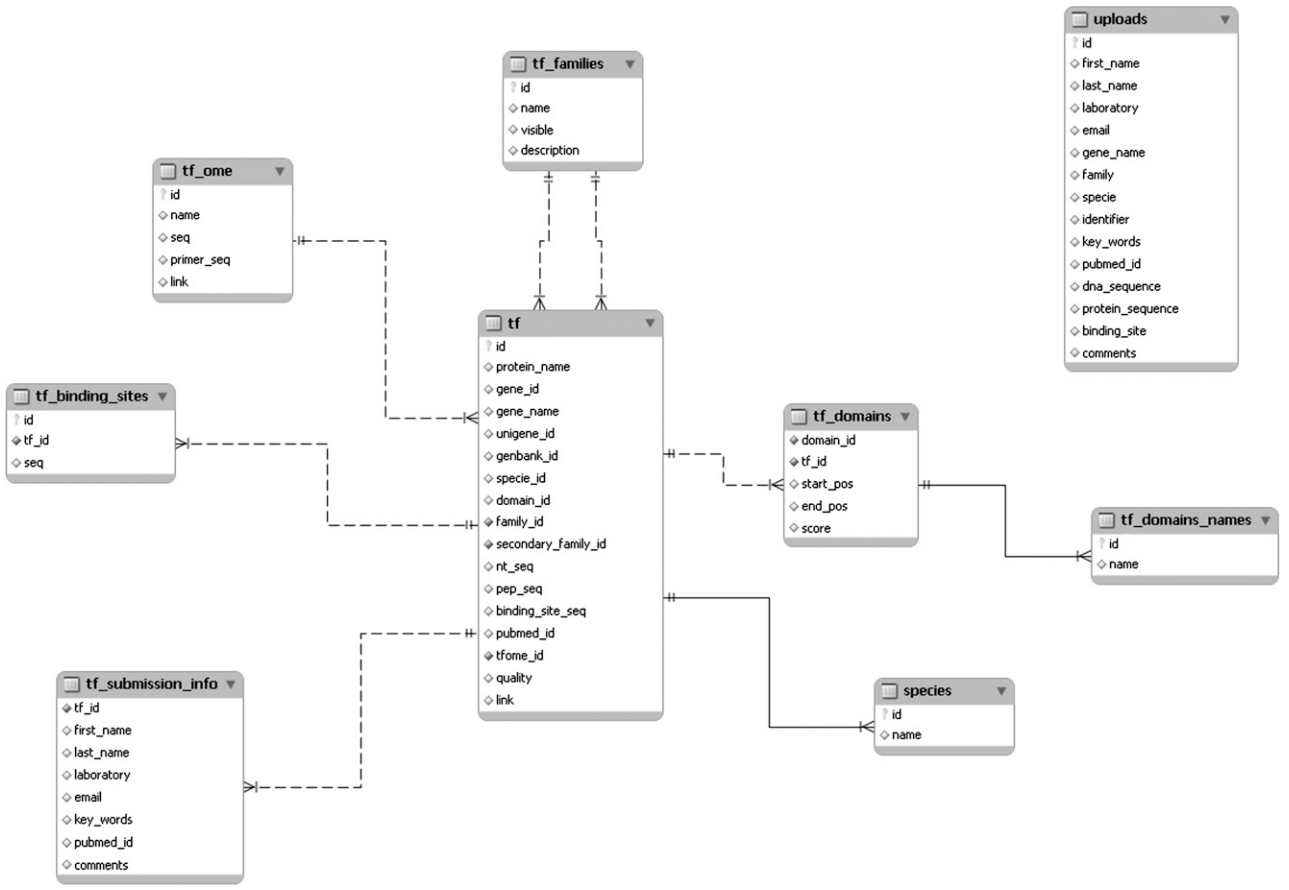


Figure 5. Structure of GrassTFDB. Interconnected MySQL tables contain data for each TF and related TFome clones and binding sites. TF information submitted by the community is stored in the uploads table and integrated into GrassTFDB after review.

lated regions was specifically amplified omitting the stop codon, and then cloned into a Gateway entry vector. Such clones can be easily recombined into a variety of destination vectors (e.g. Karimi et al., 2002; Curtis and Grossniklaus, 2003; Deplancke et al., 2004; Earley et al., 2006) suitable for yeast and bacterial expression, reporter fusion, and overexpression purposes. GRASSIUS provides information on the sequence of these clones, primers, and conditions used for amplification as well as maps (Supplemental Fig. S2). A growing collection of TF ORF clones from maize and rice are currently available to the research community on a distribution cost recovery basis and can be requested at http://grassius.org/tfomecollection.html.

GrassPROMDB

With the ultimate goal of populating GrassPROMDB with all the regulatory sequences in the grasses, this initial release focuses on a set of experimentally verified regulatory sequences as well as predicted rice promoters. Experimentally characterized promoters constitute the gold standard because they furnish information on when and how a particular regulatory sequence is active, often providing information on the CREs responsible for expression and the TFs that they recruit. GrassPROMDB summarizes much of that information for every promoter in a single page (see Supplemental Fig. S3), with links to the corresponding TFs that recognize each CRE, providing the information necessary for building GrassRegNet.

However, given how laborious it is to experimentally dissect promoter function, it is expected that GrassPROMDB will be primarily populated with predicted promoters. Predicted promoters can be of two types, curated promoters and upstream regions. Curated promoters correspond to sequences directly upstream of the TSS. Identifying curated promoters requires the availability of FLcDNAs to precisely determine TSSs. Upstream regions will be used instead of curated promoters when FLcDNAs are not available. Upstream regions consist of sequences 5′ of the translation start codons (ATG), thus including 5′-untranslated regions. Currently (September 2008), GrassPROMDB contains 56,278 rice gene upstream regions corresponding to sequences 5 kb upstream of the ATG, according to the latest release of The Institute for Genomic Research (TIGR) rice genome annotation (release 5). These upstream regions carry the same unique identifier as the genes from which they were extracted.

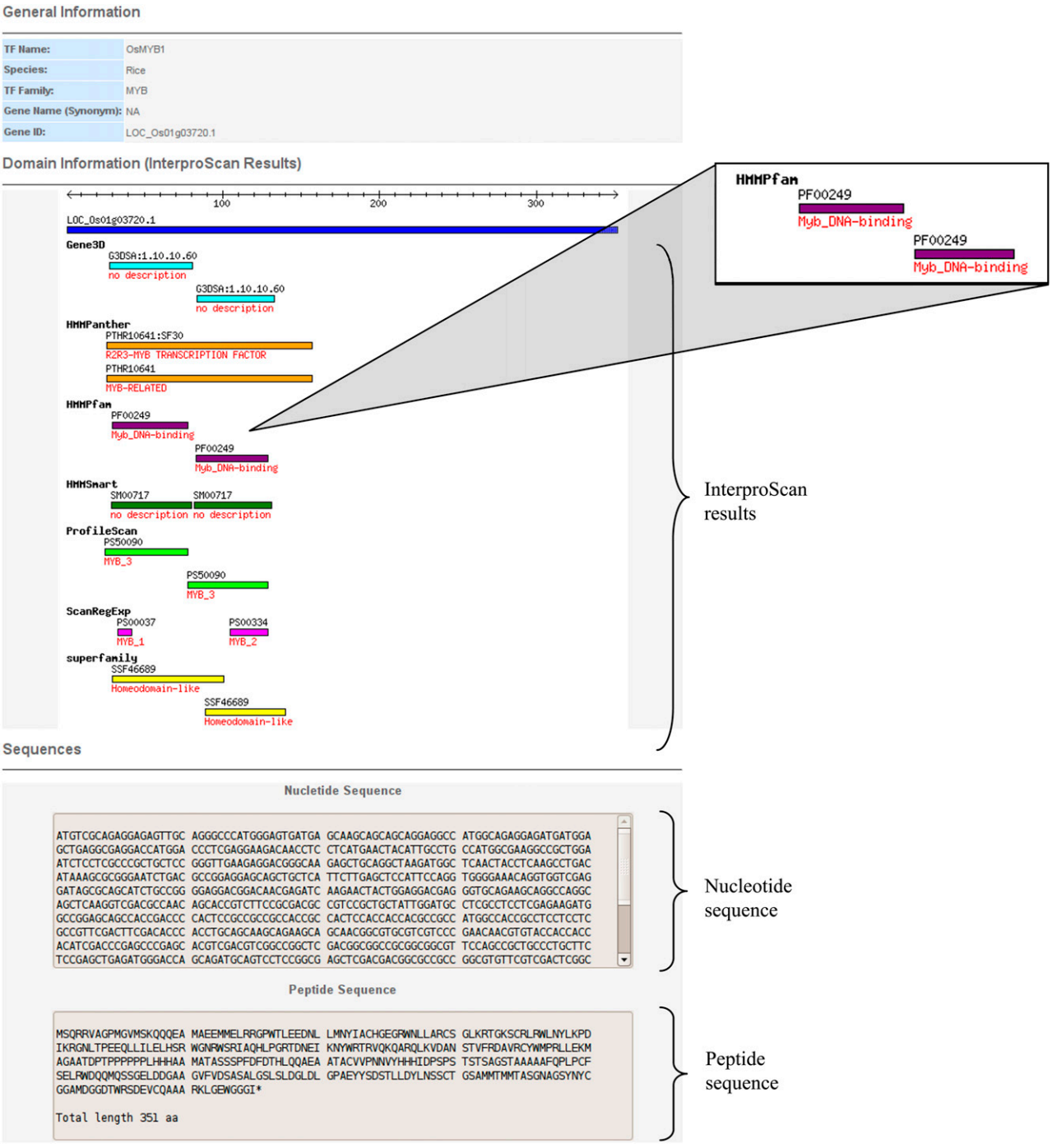


Figure 6. TF general information page. Screen shot of a sample general information page for OsMYB1. The domain structure view is generated using the Bio::Graphics module, which parses InterProScan results. Each box corresponds to an InterProScan hit and provides a database-specific identification number above and a description below. The name of the database from which the information is retrieved is provided on the left side of the image. Both nucleotide and peptide sequences of the TF are provided in scrollable windows at the bottom of the page. This information page will be expanded as more information on TFs becomes available.

CONCLUSION

GRASSIUS provides a first step toward building a comprehensive platform integrating information, tools, and resources for comparative regulatory genomics across the grasses. All the data in GRASSIUS are downloadable and freely available to the community. While initially containing information for maize, sorghum, sugarcane, and rice, as increasing genome

sequence data for other grasses (e.g. wheat [*Triticum aestivum*], Brachypodium) accumulate, GRASSIUS has the potential to include them as well. GRASSIUS also serves as an initial centralized clearinghouse for TF synonyms, and as a source of information regarding orthologous TF pairs between several grasses.

MATERIALS AND METHODS

GrassTFDB

For the development of GrassTFDB, information on TFs was obtained from the respective genome sequence databases and SUCEST, the sugarcane (*Saccharum* spp.) EST database (Vettore et al., 2003), and from existing resources with various levels of information plant TFs, including PlnTFDB (Riano-Pachon et al., 2007), PlantTFDB (Guo et al., 2008), and DBD (Kummerfeld and Teichmann, 2006). PlnTFDB and PlantTFDB classified TFs into families based on specific rules describing which domains are required and which ones are forbidden for assigning a particular protein to a family. As part of the SUCAST project, sugarcane TFs were identified from the SUCEST database and classified based on PFAM domains and BLAST similarity to known TFs (Souza et al., 2001).

Protein and respective nucleotide sequences for known TFs were integrated and grouped by species, resulting in a nonredundant set. Filtering of redundant nucleotide sequences was performed using the Perl Module Digest::MD5 (available at http://grassius.org/help.html), consisting of a sequence of 32 hexadecimal digits that identifies unequivocally each TF sequence for each species. In a second step, BLAST searches were performed to eliminate redundancy within each species. The proteins were considered duplicated if they were found in the same species, had a query coverage $\geq$90%, had a query identity $\geq$90%, and the query alignment starts less than nine residues from the start codon. If all conditions were satisfied, the longest protein was kept and the eliminated proteins were classified as identical or splicing variants. The proteins identified as TFs without known PFAM identification numbers or containing only SUPERFAMILY domains were classified as Orphan.

Based on the information available in AGRIS and in the other TF databases, we created a comprehensive list of PFAM domains, which was used to generate a database containing all FASTA sequences for each domain cataloged in the PFAM database. Each PFAM domain can be represented by a median of 67 domain sequences from different species, and it was the reference for the approach to identify new TFs from the respective genome databases. This was accomplished by collecting all protein sequences from *indica* (Gramene; http://www.gramene.org) and *japonica* (TIGR5; http://rice.plantbiology.msu.edu) rice (*Oryza sativa*), sorghum (*Sorghum bicolor*; Joint Genome Institute; http://genome.jgi-psf.org/Sorbi1), maize (*Zea mays*; http://maizesequence.org), and nucleotide sequences from *Saccharum* species hybrids (SUCEST; http://sucest-fun.org). For the first three species, we used BLASTP. For sugarcane, BLASTX was used during alignment against the database of TF domain sequences. The first criterion in the BLAST alignment was to retrieve hits with e-value $\leq 10^{-5}$, getting the complete collection of predicted proteins for a given species. To eliminate redundancy, we applied BLASTN alignments in each species and removed all candidates that had coverage $\geq$99% and identity $\geq$99% and the subject had coverage $\geq$99%. The next step involved an InterProScan search against all available PFAM hidden Markov models, keeping only significant hits with e-value $\leq$0.001 and discarding all false positives from the first step. After that, we established the rules for the identification and classification of TFs, according to DNA-binding domain sequences (Luscombe et al., 2000). We utilized a combination of the criteria developed by PlnTFDB (Riano-Pachon et al., 2007) and AtTFDB (Davuluri et al., 2003; Supplemental Table S1).

Construction of Phylogenetic Trees

Phylogenetic analyses were conducted by aligning conserved domains of all members of one TF family. InterProScan of TF sequences revealed the locations of domains, information that was utilized to extract the nucleotide sequence and perform subsequent analyses. A standard workflow that consisted of multiple sequence alignments of nucleotide sequences by ClustalW (Thompson et al., 1994) under default parameters, followed by trimming of ragged ends and tree search by RAXML (Stamatakis et al., 2005) under the gammai model of evolution was used for building the trees. The tree and species data were converted to phyloxml format (http://www.phyloxml.org) suitable for viewing with the ATV tool (Zmasek and Eddy, 2001).

Construction of TFome Collection

The ORFs of selected TFs were amplified from FLcDNA templates available from various sources (mainly the Arizona Genomics Institute) using PCR and directionally cloned into a Gateway entry vector (Invitrogen) according to the manufacturer's protocol. A high-fidelity DNA polymerase (Phusion; New England Biolabs) was employed to minimize errors during amplification. Individual entry clones were picked and plasmids isolated and sequenced to confirm the absence of errors, correct orientation, and remove the stop codon. Clones that passed this quality control were then stored in duplicate at -80°C.

Cloned TFs were named according to nomenclature guidelines developed by the community (see *Letter to the Editor*, this issue [Gray et al., 2009]), and information stored in GRASSIUS. Clones of interest may be conveniently identified using the BLAST tool in GRASSIUS or by browsing through the TF families (Fig. 2). Available clones are highlighted along with a sequence and map of the entry clone generated using the Vector NTI 10.3 software (Supplemental Fig. S2). Initially, clones will be made available by direct request on a distribution cost recovery basis (see instructions at www.grassius.org).

GrassPROMDB

The development of GrassPROMDB was based on gathering published promoter sequence information, along with detailed CRE information extracted from the literature (experimentally verified regulatory sequences). For the predicted promoter sequences, 1-kb regions upstream from the ATG translation start codon of all rice genes were obtained from to the latest TIGR release of the rice genome and extended to 5-kb upstream regions using the available genomic sequence.

Design and Web Implementation of GRASSIUS

The Web interface for GRASSIUS consists of a Perl-embedded HTML and was developed by using HTML::Mason (http://search.cpan.org/dist/HTML-Mason; accessible through http://grassius.org/help.html). Such an approach allowed implementing already available BioPerl (www.bioperl.org; Stajich et al., 2002) modules, such as Bio::Graphics and Bio:SeqIO, easily into the GRASSIUS interface. The databases GrassTFDB and GrassPROMDB were developed using MySQL (http://www.mysql.com). Figure 5 contains a simplified diagram revealing the interaction of the different components behind GrassTFDB, and Supplemental Figure S5 describes the structure of GrassPROMDB.

Data Visualization and User Interface

GrassTFDB

Detailed information on a particular TF can be accessed by either browsing family members (Fig. 3A) or by searching by name (Fig. 3B). TFs can be queried by sequence similarity through BLAST searches (Fig. 3C) or by phylogenetic homology (Fig. 3D). Searching for TF name or browsing a family generates a results table (Fig. 4), where TF name, gene name (synonym), gene accession number, DNA-binding preferences (when available), direct targets, availability of TFome clones, and curation quality levels are shown in separate columns. The TF name column directs users to a page providing details of the particular TF. The Gene Name and Gene Id columns link to external sources depending on the particular species to which the TF belongs. For rice TFs, users are directed to either TIGR (http://www.tigr.org) or RAP-DB (http://rapdb.dna.affrc.go.jp). For maize TFs, users are directed to MaizeGDB (http://www.maizegdb.org) or the Maize Genome Sequencing Project (http://www.maizesequence.org) pages.

The TF information page provides domain information, nucleotide, or peptide sequences, orthologs in the other grasses (when available), and expression information as it becomes available (Fig. 6, sample screen shot for OsMYB1). Domain information is extracted from InterProScan results and gathers information from multiple databases including BlastProDom, FPrintScan, HMMPIR, HMMPfam, HMMSmart, HMMTigr, ProfileScan, ScanRegExp, SuperFamily,

SignalPHMM, TMHMM, HMMPanther, and Gene3D. The positions of the corresponding domains with respect to a schematic representation of the protein (N terminus on the left, C terminus on the right) are represented by boxes generated by the Bio::Graphics module of BioPerl. This provides an identification number specific to the database and descriptions of the particular domain. Each box, when clicked, takes the user to detailed information about the protein domain at corresponding databases.

GrassPROMDB

Users can query for promoters by either browsing curated promoters in one species, or by searching by gene id, pathway/process name, targeting TF, sequence motif, and BLAST. Not all the links are active at this time because much of the information is in the process of being generated as the genome annotations for maize and sorghum progress. The promoter information page provides a graphic view of CREs as well as the sequence of the promoter (Supplemental Fig. S3). The CREs represented by boxes are located relative to the TSS for curated promoters, ATG for noncurated promoters with their sequence, as well as the name of the TF that binds the respective CRE (when available) being shown, their sequence as well as the name of the TF that binds the respective CRE (when available), is displayed underneath. Upon clicking on a box, a page describing properties and experimental information regarding that CRE is displayed. The TF names link to the corresponding records in GrassTFDB.

Community Contribution

Users are encouraged to contribute to contents of GrassTFDB and GrassPROMDB. After submitting a form describing the details about a particular TF (Supplemental Fig. S4) or promoter sequence, the submitted information is reviewed by GRASSIUS curators and then integrated into GRASSIUS.

Downloads, Help Pages

All the data in GRASSIUS can be freely downloaded from http://www.grassius.org/downloads.html after a swift user registration. Sequences, the alignments of members of TF families and promoter sequences, along with CRE information, are available for download in tab-delimited text format. Any other data are available upon request. Users are guided by comprehensive help pages on how to use all features available in GRASSIUS.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Interactive visualization of phylogenetic trees for TFs from the grasses.

Supplemental Figure S2. Example of the information available in GRASSIUS for a TF ORF clone.

Supplemental Figure S3. Detailed information of genic upstream regions in GrassPROMDB.

Supplemental Figure S4. TF submission form in a community contribution page.

Supplemental Figure S5. GrassPROMDB database structure.

Supplemental Table S1. Family names in four different plant TF databases.

Supplemental Materials and Methods S1. Additional methods description and supplemental figure legends.

ACKNOWLEDGMENTS

We thank Saranyan Palaniswamy, Ramana Davuluri, and Eric Easley with assistance at various stages of this project.

Received August 29, 2008; accepted October 29, 2008; published November 5, 2008.

LITERATURE CITED

Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol **14:** 283–291

Bennetzen JL (2007) Patterns in grass genome evolution. Curr Opin Plant Biol **10:** 176–181

Braun EL, Grotewold E (2001) Fungal Zuotin proteins evolved from MIDA1-like factors by lineage-specific loss of MYB domains. Mol Biol Evol **18:** 1401–1412

Curtis MD, Grossniklaus U (2003) A gateway cloning vector set for high-throughput functional analysis of genes in planta. Plant Physiol **133:** 462–469

Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics **4:** 25

Deplancke B, Dupuy D, Vidal M, Walhout AJ (2004) A gateway-compatible yeast one-hybrid system. Genome Res **14:** 2093–2101

Dias AP, Braun EL, McMullen MD, Grotewold E (2003) Recently duplicated maize *R2R3 Myb* genes provide evidence for distinct mechanisms of evolutionary divergence after duplication. Plant Physiol **131:** 610–620

Earley KW, Haag JR, Pontes O, Opper K, Juehne T, Song K, Pikaard CS (2006) Gateway-compatible vectors for plant functional genomics and proteomics. Plant J **45:** 616–629

Gray J, Bevan M, Brutnell T, Buell CR, Cone K, Hake S, Jackson D, Kellogg E, Lawrence C, McCouch S, et al (2009) A recommendation for naming transcription factor proteins in the grasses. Plant Physiol **149:** 4–6

Guo AY, Chen X, Gao G, Zhang H, Zhu QH, Liu XC, Zhong YF, Gu X, He K, Luo J (2008) PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res **36:** D966–969

Jannoo N, Grivet L, Chantret N, Garsmeur O, Glaszmann JC, Arruda P, D'Hont A (2007) Orthologous comparison in a gene-rich region among grasses reveals stability in the sugarcane polyploid genome. Plant J **50:** 574–585

Karimi M, Inze D, Depicker A (2002) GATEWAY vectors for Agrobacterium-mediated plant transformation. Trends Plant Sci **7:** 193–195

Kummerfeld SK, Teichmann SA (2006) DBD: a transcription factor prediction database. Nucleic Acids Res **34:** D74–81

Luscombe NM, Austin SE, Berman HM, Thornton JM (2000) An overview of the structures of protein-DNA complexes. Genome Biol **1:** REVIEWS001

Palaniswamy K, James S, Sun H, Lamb R, Davuluri RV, Grotewold E (2006) AGRIS and AtRegNet: A platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol **140:** 818–829

Riano-Pachon DM, Ruzicic S, Dreyer I, Mueller-Roeber B (2007) PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics **8:** 42

Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, et al (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science **290:** 2105–2110

Riechmann JL, Ratcliffe OJ (2000) A genomic perspective on plant transcription factors. Curr Opin Plant Biol **3:** 423–434

Schlitt T, Brazma A (2007) Current approaches to gene regulatory network modelling. BMC Bioinformatics (Suppl 6) **8:** S9

Shahmuradov IA, Gammerman AJ, Hancock JM, Bramley PM, Solovyev VV (2003) PlantProm: a database of plant promoter sequences. Nucleic Acids Res **31:** 114–117

Souza GM, Simoes ACQ, Oliveira KC, Garay HM, Fiorini LC, Gomes FS, Nishiyama-Junior MY, da Silva AM (2001) The sugarcane signal transduction (SUCAST) catalogue: prospecting signal transduction in sugarcane. Genet Mol Biol **24:** 25–34

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, et al (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res **12:** 1611–1618

Stamatakis A, Ludwig T, Meier H (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics **21:** 456–463

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res **22:** 4673–4680

Vettore AL, da Silva FR, Kemper EL, Souza GM, da Silva AM, Ferro MI, Henrique-Silva F, Giglioti EA, Lemos MV, Coutinho LL, et al (2003) Analysis and functional annotation of an expressed sequence tag collection for tropical crop sugarcane. Genome Res **13:** 2725–2735

Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA (2008) DBD—taxonomically broad transcription factor predictions: new content and functionality. Nucleic Acids Res **36:** D88–92

Yamamoto YY, Obokata J (2008) ppdb: a plant promoter database. Nucleic Acids Res **36:** D977–981

Yu H, Gerstein M (2006) Genomic analysis of the hierarchical structure of regulatory networks. Proc Natl Acad Sci USA **103:** 14724–14731

Zmasek CM, Eddy SR (2001) ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics **17:** 383–384