# A Recommendation for Naming Transcription Factor Proteins in the Grasses

Transcription factors are central for the exquisite temporal and spatial expression patterns of many genes. These proteins are characterized by their ability to be tethered to particular regulatory sequences in the genes that they control. While many other proteins participate in the regulation of gene expression, we limit our definition of transcription factors here to proteins that often 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, transcription factors are classified into 50 to 60 different families, and in plants, 5% to 7% of all the protein-encoding genes are transcription factors, making them, collectively, perhaps the largest functional class of proteins.

The availability of protein sequence information from several plant genomes that have been fully or partially sequenced over the past few years is creating an urgent need to develop a set of criteria for naming and identifying members of large protein families. A common nomenclature would facilitate better communication among scientists, working not just on a particular plant system, but also across different plant species. This need is becoming particularly acute in the grasses, where some members have a rich genetic history (e.g. maize [*Zea mays*], rice [*Oryza sativa*], barley [*Hordeum vulgare*]) and genes have been named for more than 100 years according to the phenotypes of the corresponding mutations, often leading to multiple names for the same gene. In other grasses that do not have such a genetic heritage, such as sugarcane (*Saccharum* ssp.), genes are characterized primarily by EST accession numbers, often with multiple nonoverlapping sequences corresponding to one gene. Gene nomenclature is governed by species-specific communities in many cases (e.g. maize, rice, sorghum [*Sorghum bicolor*]), but such committees do not exist for all species with significant sequence data. In general, names of proteins follow the names of their corresponding genes.

In Arabidopsis (*Arabidopsis thaliana*), shortly after the completion of the genome sequence, criteria were developed to provide unique names to all transcription factors, often in a family-by-family strategy (e.g. Stracke et al., 2001; Jakoby et al., 2002; Bailey et al., 2003). These identifiers are of the form AtXXXyyy, where At provides the species identifier At = Arabidopsis, important when describing transcription factors from multiple plants; XXX corresponds to a two to five or more letter code for the particular transcription factor family; and yyy corresponds to an arbitrary number between one and the total number of members in that particular family. These nomenclature conventions were rapidly embraced by the community, facilitating communication in publications and allowing the development of databases that compile information on Arabidopsis regulatory proteins (Davuluri et al., 2003; Guo et al., 2005).

Here, we propose adoption of similar synonyms for proteins corresponding to transcription factors across the grasses, with a goal of having a uniform naming system as already developed for Arabidopsis. Such names are not meant to replace the often-familiar names for many proteins, but rather to provide synonyms that can link information about all members of the gene family. Briefly, each transcription factor will be identified by a two-letter code corresponding to the species (e.g. Zm for *Z. mays*, Sb for *S. bicolor*, and Os for *O. sativa*), followed by the transcription factor family name and by a number that represents its position within the family. The two-letter species code should suffice for now to unequivocally describe the grasses for which transcription factors have been identified, but clearly in the future it will need to be expanded to three or four letters, as the number of species being included in studies increases. Alternatively, the two-letter code reflecting the species name could be used for organisms whose genomes are currently under study, and if needed, a new two-letter abbreviation, which is not necessarily consistent with the genus and specific epithet will be made up for new ones. Following this criterion, sugarcane will be identified by the Sc letter code, as agreed by the respective community.

Numbers will be assigned arbitrarily, and whenever possible, the numbers should provide a historic perspective of the order in which transcription factors have been first identified. For example, since maize KN1 and C1 correspond to the founding members of their respective families (HD and MYB, respectively), they are assigned the number 1. When a transcription factor has already been numbered, every possible effort should be made to consider that number as part of the new name, e.g. maize Zm38 (Franken et al., 1994) should become ZmMYB38. Since it is realized that many transcription factors are known by their genetic names, this nomenclature will permit the use of various different synonyms. For example, KNOTTED1, which would be ZmHD1, where HD corresponds to the homeodomain family, could also be identified as ZmHD1(KN1) when in need to highlight the genetic locus, *KN1*, for which the protein is also often known. Similarly C1, ZmMYB1, could be also identified as ZmMYB1(C1). Where the same name has already been used for different family members, then an alternative name(s) will be assigned to avoid further confusion in the literature. For example, ZmMYB8 (Fornale et al., 2006), which is supported by a complete cDNA will

www.plantphysiol.org/cgi/doi/10.1104/pp.108.128504

remain ZmMYB8, whereas the partial ESTs MYB8 (Jiang et al., 2004) and ZmMYB-IP20 (Rabinowicz et al., 1999), which both correspond to the same protein, will be assigned a new name (e.g. ZmMYB67).

While it would be attractive for this new naming system to provide information about orthologous pairs across species, as for example ZmMYB32 being the closest relative to OsMYB32, this was considered impractical and perhaps misleading for several reasons. First, the genomes are not yet completely sequenced or fully annotated, hence new transcription factors are likely to be identified in the future, significantly affecting the reconstruction of protein phylogenies. Second, different tree-building methods are likely to yield slightly different results, which would create significant confusion. Third, while it has been tempting to assume that high similarity is likely to correspond to the control of similar cellular processes, this has often not been the case, particularly when considering regulators of metabolic pathways (Grotewold, 2008).

We propose that GRASSIUS (www.grassius.org) will serve as an initial centralized clearinghouse for transcription factor synonyms for the grasses, starting with maize, rice, sorghum, and sugarcane, following the criteria outlined above. GRASSIUS will provide a source of cross-reference between the new names, synonyms, National Center for Biotechnology Information accession codes for ESTs, and cDNAs and unique gene identifiers, as they become available. This will be achieved dynamically, ensuring that immediately after a new synonym has been given to a transcription factor, it will be reflected in GRASSIUS. The community is invited to comment on these assignments for a defined period of time (e.g. until the end of 2009), at which time the names will become official. The community will be presented with these guidelines and will be provided with ample opportunity to become aware and discuss these recommendations at conferences and meetings for the corresponding organisms. In addition, we will work with the various model organism and clade-oriented databases (e.g. MaizeGDB, BrachyBase, and Gramene) to ensure that the proper nomenclature is represented in these community resources as well. These databases will also serve as optimal clearinghouses for the respective organisms if GRASSIUS ceases to serve this purpose. Conflicts and issues that may arise with respect to how to name a particular transcription factor (or family of transcription factors) will be opened for discussion to experts in the field. For example, if there is a disagreement on how to name a new family of transcription factors, scientists working with those specific transcription factors will be invited to comment.

## SOME PARTICULAR CASES

### Multiple Proteins from One Gene

Multiple transcripts derived from one gene are frequently present in plants, with more than 21% of the rice and Arabidopsis genes being alternatively spliced (Wang and Brendel, 2006). Several examples of transcription factors displaying alternate splice variants have also been described in maize (e.g. Grotewold et al., 1991; Burr et al., 1996). We recommend that transcription factor proteins derived from alternate spliced mRNAs be named with the .1, .2, .3 suffixes after the number of the protein. For example, the new synonyms corresponding to the two proteins derived from the alternatively spliced variants of maize *PERICARP COLOR1* (*P1*) would be ZmMYB3.1 and ZmMYB3.2. In those instances, as well as in cases when multiple gene models exist for a particular transcription factor gene, the suffixes in the protein will match those in the gene models. For instance, if the rice LOC_Os02g36880 gene shows four different gene models, from .1 to .4, then OsNAC1.1 should match with LOC_Os02g36880.1 and OsNAC1.4 should match LOC_Os02g36880.4.

### Allelic Variants

The sequencing of multiple inbred lines/subspecies displaying significant natural variation makes it necessary to incorporate an option to represent from which allele a particular transcription factor protein sequence is derived. We propose that whenever necessary, a superscript is added. This superscript could represent the source of the allele, when known. For example, the P1 protein obtained from the W22 maize inbred could be represented as ZmMYB3.1$^{W22}$, and that from B73 as ZmMYB3.1$^{B73}$. When formatting issues prevent the use of the superscripts, then it would also be acceptable to use ^B73 to represent the allele. In that case, ZmMYB3.1$^{B73}$ and ZmMYB3.1^B73 would be equivalent. Such criterion could also be used to indicate, whenever known, whether a transcription factor protein sequence in rice is derived from the sequenced *japonica* genome (cv Nipponbare) or the sequenced *indica* genome (cv 9311). Thus, the OsNAC6 factor, involved in biotic and abiotic stress response in rice (Ohnishi et al., 2005), could be OsNAC6$^{Nipp}$ (or OsNAC6$^{9311}$) when intending to capture aspects of the protein that relate to variation. If the origin is not precisely known or a name/accession_ID is too cumbersome to be represented as a superscript, then numbers could be used to distinguish alleles (e.g. OsNAC6$^{1}$, OsNAC6$^{2}$, etc.), with cross-references to inbred/accession names maintained within species databases. Of course, within a species, a single transcription factor name (e.g. OsNAC6) will correspond to the products of corresponding gene models.

### Products from Tandem Gene Arrays

In some instances, for example the various alleles of the maize *p1* gene (Chopra et al., 1998), very similar (but not necessarily identical) proteins are encoded by individual members of a multigene array. In those instances, we recommend using letters (a, b, c) to indicate the proteins that come from each copy. For example, if three different copies of the *p1* gene

from B73 were shown to encode slightly different proteins, then those products would be identified as ZmMYB3$^{B73}$a, ZmMYB3$^{B73}$b, and ZmMYB3$^{B73}$c.

## GENE AND PROTEIN NOMENCLATURE

The guidelines described here for naming transcription factors are expected to apply solely to proteins (or predicted open reading frames) and not necessarily to genes. Indeed, as the sequencing of various genomes progresses, nomenclature committees have been established to address the issue of how to name genes and gene products. It is therefore of paramount importance that the guidelines described here are in line with those being developed by the corresponding committees. Toward this objective, the criteria described here have already been discussed and accepted for the maize transcription factors by the Maize Genetics Nomenclature Committee (http://www.maizegdb.org/maize_nomenclature.php), by the Sugarcane Nomenclature Committee, and by the International Brachypodium Initiative (http://www.brachypodium.org/). The corresponding nomenclature committees will make these guidelines available to the respective communities.

**John Gray, Michael Bevan, Thomas Brutnell, C. Robin Buell, Karen Cone, Sarah Hake, David Jackson, Elizabeth Kellogg, Carolyn Lawrence, Susan McCouch, Todd Mockler, Stephen Moose, Andrew Paterson, Thomas Peterson, Daniel Rokshar, Glaucia Mendes Souza, Nathan Springer, Nils Stein, Marja Timmermans, Guo-Liang Wang, and Erich Grotewold***

**Department of Biological Sciences, University of Toledo, Toledo, Ohio 43606 (J.G.); Department of Cell and Molecular Biology, John Innes Center, Norwich Research Park, Norwich NR4 7UH, United Kingdom (M.B.); Boyce Thompson Institute, Cornell University, Ithaca, New York 14853 (T.B.); Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824 (C.R.B.); Division of Biological Sciences (K.C.) and Department of Biology (E.K.), University of Missouri, Columbia, Missouri 65211; Plant Gene Expression Center, Albany, California 94710 (S.H.); Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 (D.J., M.T.); Corn Insects and Crop Genetics Research Unit, Ames, Iowa 50201 (C.L.); Cornell University, Ithaca, New York 14853 (S.M.); Department of Botany and Plant Pathology and Center for Genome Research and Biocomputing, Oregon State University, Corvallis, Oregon 97331 (T.M.); Department of Crop Sciences and Energy Biosciences Institute, University of Illinois, Urbana-Champaign, Illinois 61801 (S.M.); Plant Genome Mapping Laboratory, University of Georgia, Athens, Georgia 30602 (A.P.); Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa 50011–3260 (T.P.); Department of Energy Joint Genome Institute, Walnut Creek, California 94598 (D.R.); Instituto de Química, Departamento de Bioquímica, Universidade de São Paulo, São Paulo, Brazil (G.M.S.); Center for Plant and Microbial Genomics, Department of Plant Biology, University of Minnesota, Saint Paul, Minnesota 55108 (N. Springer); Leibniz-Institute of Plant Genetics and Crop Plant Research, Genebank Department, 06466 Gatersleben, Germany (N. Stein); and Department of Plant Pathology (G.-L.W.) and Department of Plant Cellular and Molecular Biology and Plant Biotechnology Center (E.G.), The Ohio State University, Columbus, Ohio 43210**

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

## LITERATURE CITED

**Bailey PC, Martin C, Toledo-Ortiz G, Quail PH, Huq E, Heim MA, Jakoby M, Werber M, Weisshaar B** (2003) Update on the basic helix-loop-helix transcription factor gene family in *Arabidopsis thaliana*. Plant Cell **15:** 2497–2502

**Burr FA, Burr B, Scheffler BE, Blewitt M, Wienand U, Matz EC** (1996) The maize repressor-like gene *intensifier1* shares homology with the *r1/b1* multigene family of transcription factors and exhibits missplicing. Plant Cell **8:** 1249–1259

**Chopra S, Athma P, Li XG, Peterson T** (1998) A maize Myb homolog is encoded by a multicopy gene complex. Mol Gen Genet **260:** 372–380

**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

**Fornale S, Sonbol FM, Maes T, Capellades M, Puigdomenech P, Rigau J, Caparros-Ruiz D** (2006) Down-regulation of the maize and Arabidopsis thaliana caffeic acid O-methyl-transferase genes by two new maize R2R3-MYB transcription factors. Plant Mol Biol **62:** 809–823

**Franken P, Schrell S, Peterson PA, Saedler H, Wienand U** (1994) Molecular analysis of protein domain function encoded by the *myb*-homologous maize genes *C1, Zm 1* and *Zm 38*. Plant J **6:** 21–30

**Grotewold E** (2008) Transcription factors for predictive plant metabolic engineering: Are we there yet? Curr Opin Biotechnol **19:** 138–144

**Grotewold E, Athma P, Peterson T** (1991) Alternatively spliced products of the maize P gene encode proteins with homology to the DNA-binding domain of myb-like transcription factors. Proc Natl Acad Sci USA **88:** 4587–4591

**Guo A, He K, Liu D, Bai S, Gu X, Wei L, Luo J** (2005) DATF: a database of Arabidopsis transcription factors. Bioinformatics **21:** 2568–2569

**Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F** (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci **7:** 106–111

**Jiang C, Gu J, Chopra S, Gu X, Peterson T** (2004) Ordered origin of the typical two- and three-repeat Myb genes. Gene **326:** 13–22

**Ohnishi T, Sugahara S, Yamada T, Kikuchi K, Yoshiba Y, Hirano HY, Tsutsumi N** (2005) OsNAC6, a member of the NAC gene family, is induced by various stresses in rice. Genes Genet Syst **80:** 135–139

**Rabinowicz PD, Braun EL, Wolfe AD, Bowen B, Grotewold E** (1999) Maize *R2R3 Myb* genes: Sequence analysis reveals amplification in higher plants. Genetics **153:** 427–444

**Stracke R, Werber M, Weisshaar B** (2001) The R2R3 MYB gene family in *Arabidopsis thaliana*. Curr Opin Plant Biol **4:** 447–456

**Wang BB, Brendel V** (2006) Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci USA **103:** 7175–7180