Bioenerg. Res. (2009) 2:217–232
DOI 10.1007/s12155-009-9047-9

A Bioinformatic Analysis of *NAC* Genes for Plant Cell Wall Development in Relation to Lignocellulosic Bioenergy Production

Hui Shen · Yanbin Yin · Fang Chen · Ying Xu · Richard A. Dixon

Published online: 16 October 2009

Abstract NAM, ATAF, and CUC2 (NAC) proteins are encoded by one of the largest plant-specific transcription factor gene families. The functions of many NAC proteins relate to different aspects of lignocellulosic biomass production, and a small group of NAC transcription factors has been characterized as master regulators of plant cell wall development. In the present study, a total of 1,232 NAC protein sequences from 11 different organisms were analyzed by sequence phylogeny based on protein DNA-binding domains. We included eight whole genomes (*Arabidopsis*, rice, poplar, grape, sorghum, soybean, moss (*Physcomitrella patens*), and spike moss (*Selaginella moellendorffii*)) and three not yet fully sequenced genomes (maize, switchgrass, and *Medicago truncatula*) in our analyses. Ninety-two potential *PvNAC* genes from switchgrass and 148 *PtNAC* genes from poplar were identified. The 1,232 NAC proteins were phylogenetically classified into eight subfamilies, each of which was further divided into subgroups according to their tree topology. The phylogenetic subgroups were then grouped into different clades each sharing conserved motif patterns in the C-terminal sequences, and those that may function in plant cell wall development were further identified through motif grouping and gene expression pattern analysis using publicly available microarray data. Our results provide a bioinformatic baseline for further functional analyses of candidate *NAC* genes for improving cell wall and environmental tolerance traits in the bioenergy crops switchgrass and poplar.

Keywords NAM/NAC protein · Transcription factor · Secondary cell wall · Stress tolerance · Biomass · Cellulosic ethanol · Phylogeny

Hui Shen and Yanbin Yin contributed equally to this study.

Electronic supplementary material The online version of this article (doi:10.1007/s12155-009-9047-9) contains supplementary material, which is available to authorized users.

H. Shen · F. Chen · R. A. Dixon (✉)
Plant Biology Division, Samuel Roberts Noble Foundation,
2510 Sam Noble Parkway,
Ardmore, OK 73401, USA
e-mail: radixon@noble.org

Y. Yin · Y. Xu (✉)
Computational Systems Biology Lab,
Department of Biochemistry and Molecular Biology,
University of Georgia,
Athens, GA 30602, USA
e-mail: xyn@bmb.uga.edu

Y. Yin · Y. Xu
Institute of Bioinformatics, University of Georgia,
Athens, GA 30602, USA

H. Shen · Y. Yin · F. Chen · Y. Xu · R. A. Dixon
Bioenergy Science Center (BESC),
Oak Ridge, TN, USA

Introduction

The generation of renewable bioenergy from plant biomass has ecological and economic implications of national and global importance. Selection of suitable plant materials for bioenergy production has been conducted in the USA through the Bioenergy Feedstock Development Program at the Oak Ridge National Laboratory since 1978. Fast-growing trees such as poplar (*Populus trichocarpa*) and herbaceous crops including potential forages such as switchgrass (*Panicum virgatum*) have been selected as likely renewable energy sources for the nation's future energy needs [1, 2]. Improving such species for traits such as high biomass yield, abiotic/biotic stress tolerance, high nutrient use efficiency, and reduced cell wall recalcitrance

will be critical for bioenergy production in the future. Recently, cell wall synthetic and regulatory genes, including those involved in cellulose and lignin biosynthesis, have been targeted for transgenic modification to reduce cell wall recalcitrance to enzymatic saccharification. As an example, transgenic alfalfa lines with downregulation of different lignin biosynthetic genes have increased fermentable sugar yields [3]. Such results provide proof of principle for engineering plants for improved bioenergy production.

Plant biomass production is strongly affected by environmental conditions. Because bioenergy crops will have to be planted on marginal lands, they will have to adjust and coordinate their growth, metabolism, and developmental processes under conditions of sometimes severe environmental stress. Regulation of biomass production is under complex and dynamic genetic control, for which transcription factors act as master regulators of gene expression. The protein structure of a typical transcription factor has two domains: the DNA binding domain responsible for binding to specific DNA *cis*-elements in the promoter regions of target genes and the regulatory domain responsible for regulating transcription of the target genes. Many plant-specific transcription factors, such as AP2/EREBP; DNA-binding with one finger protein (Dof); NAM, ATAF, and CUC2 (NAC); WRKY; and SBP proteins have recently been functionally characterized [4]. The NAC transcription factor family, which has more than 100 members in *Arabidopsis* [5, 6], is one of the largest transcription factor families in plants. NAC proteins are involved in diverse processes including orchestration of developmental programs in flowers and roots, response to biotic and abiotic stress [6], control of nutrient remobilization from leaves to developing grains [7], and determination of tiller numbers [8]. Moreover, they also control plant cell wall composition during xylogenesis, fiber development, and wood formation in vascular plants [9–11].

Several *NAC* genes from *Arabidopsis* have been identified as controlling plant cell wall architecture. Overexpression of ANAC104/xylem NAC domain 1 (AtXND1) in *Arabidopsis* suppressed the differentiation of tracheary elements and negatively regulated lignocellulose/secondary cell wall synthesis and programmed cell death in xylem vessel cells [12, 13]. ANAC101/vascular-related NAC-domain 6 and ANAC030/VND7 act as transcriptional switches for trans-differentiation of various cells into metaxylem- and protoxylem-like vessel elements in *Arabidopsis* and poplar [14]. ANAC043/NAC secondary wall-thickening promoting factor 1 (NST1), ANAC066/NST2, and ANAC012/NST3 regulate secondary cell wall formation in cells other than vessels [15]. NST1 and NST2 function redundantly in secondary wall formation in anther endothecium cells in *Arabidopsis*, whereas NST1 and NST3 act redundantly in promoting cell wall thickening of the interfascicular fibers and secondary xylem vessels [15–17]. NST3 was independently identified as secondary wall-associated NAC domain protein 1 [18]. ANAC073/SND2 and ANAC010/SND3 are also involved in secondary cell wall thickening in xylem elements and interfascicular fibers; and dominant repression or loss of function of these genes causes dramatic reduction in the secondary wall thickening of the fibers [11]. Interestingly, SND2 and SND3 seem to work downstream of NST3 and NST1 since the expression level of these genes is downregulated in NST1/NST3 RNAi transgenic lines, and SND2 was identified as a direct target of NST3 [11].

Several groups have attempted phylogenetic analyses of *NAC* genes. The first systematic analysis was of 105 *Arabidopsis* NAC proteins (based on subdomains A to E, Fig. 1) and 75 rice NAC proteins. These 180 NAC protein sequences were classified into two supergroups and 18 subgroups [5]. However, 18 proteins were not placed into any of the subgroups in this study. More recently, a systematic sequence analysis revealed 140 putative *NAC* genes (*ONAC*) in rice. Phylogenetic analyses (based on subdomains A to D, Fig. 1) suggested that the *NAC* family can be divided into five groups (1–5). Sequence analysis, together with the organization of putative motifs, indicated distinct phylogenetic structures and potentially diverse functions for the *NAC* family in rice [19]. Phylogenetic analyses of sequences (based on subdomains A to C, Fig. 1) from tobacco compared with rice, *Arabidopsis*, and poplar revealed a novel *NAC* subfamily termed *TNACS* that appears to be restricted to the Solanaceae since it is absent from currently sequenced plant genomes but present in tomato (*Solanum lycopersicum*), pepper (*Capsicum annuum*), and potato (*Solanum tuberosum*) [20]. The above studies reveal a challenge for phylogenetic analysis of *NAC* genes, as they show that the resulting phylogeny changes when different subsets of the protein domain sequences were included for construction of the phylogenetic tree. We believe that inclusion of the highly diverse C-terminal sequence information and use of additional bioinformatic approaches may lead to a better understanding of the phylogeny of this complex gene family.

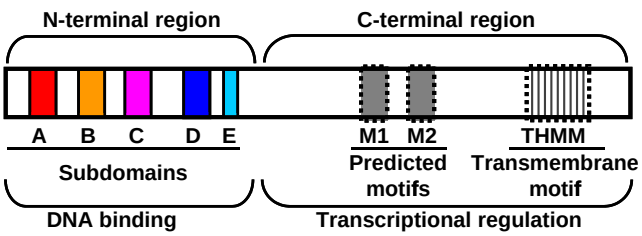


Fig. 1 Structures of NAC protein domains and predicted motifs. The *colored boxes* indicate the protein subdomains *A* to *E*. These comprise the N-terminal NAC domain, the amino acid sequences of which were used to generate the phylogenetic tree. *Dotted boxes* are predicted motifs in the C-terminal regions

In the present study, we employ genome-wide informatic approaches to select NACs relevant to biomass-related traits. A total of 1,232 NAC protein sequences from 11 organisms were analyzed by sequence phylogeny based on the N-terminal protein subdomains A to E. Six whole genomes, namely moss (*Physcomitrella patens*), spike moss (*Selaginella moellendorffii)*, grape (*Vitis vinifera*), poplar, soybean, and sorghum (*Sorghum bicolor*), were added to the genome list for analysis. We classified the NAC proteins into eight subfamilies (NAC-a to NAC-h). Each subfamily was further classified into subgroups according to their tree topology. The subgroups were then grouped into different clades each sharing conserved motif patterns in the C-terminal sequences. Specific protein motifs were found for the groups of NAC proteins involved in cell wall development. In addition, we have made functional predictions for *Arabidopsis NAC* genes based on both sequence homology and coexpression analysis with lignin and secondary cell wall synthetic genes. Together, these approaches provide powerful bioinformatic baseline support for selection of candidate *NAC* genes from poplar and switchgrass with predicted functions in regulating plant cell wall development, stress tolerance, and plant architecture.

Methods

Sequence Retrieval and Data Sources

We selected 11 plant species for *NAC* gene identification (Table 1), among which six have sequenced and published genomes: *Arabidopsis*, rice, sorghum, poplar, grape, and moss. Two additional genomes, spike moss and soybean (*Glycine max*), have been sequenced but not published in annotated form. The remaining three species, maize (*Zea mays* L.), *Medicago truncatula*, and switchgrass, are important from the biofuel perspective but are not yet fully sequenced. Proteins for these genomes were downloaded from the respective web sites as listed in Table 1. For Medicago, we combined proteins predicted from partial genome release (http://www.medicago.org/ genome/downloads/Mt2/) and peptides predicted from assembled expressed sequence tag (EST) release (ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Medicago_truncatula). Specifically, assembled ESTs of *Medicago* were translated into amino acid sequences in six frames by using the *transeq* command of the EMBOSS package (http://www.ncbi.nlm.nih.gov/pubmed/10827456). The switchgrass peptide sequences were obtained in the same way by translating EST transcripts assembled by plantGDB (http://www.plantgdb.org/). The sequences from tobacco (*Nicotiana tabacum*), cotton (*Gossypium barbadense*), petunia (*Petunia x hybrida*), potato (*S. tuberosum*), pumpkin (*Cucurbita mixta*), soybean, and wheat (*Triticum aestivum*) were retrieved from the National Center for Biotechnology Information nonredundant protein (NCBI-nr) database at ftp://ftp.ncbi.nih.gov/blast/db/FASTA/ as of April 4, 2008.

Homology Search

A Hidden Markov Model (HMM) for Pfam domain PF02365.7 (138 amino acid long) has been developed

Table 1 The 11 plant species analyzed in this study

Index	Abbreviation	Clade	Species	Genome published	Downloaded from
1	pp	Moss	*Physcomitrella patens* ssp. *patens*	[20]	JGI V1.1
2	sm	Spike moss	*Selaginella moellendorffii*	No	JGI V1.0
3	pt	Dicot	*Populus trichocarpa*	[21]	JGI V1.0
4	at	Dicot	*Arabidopsis thaliana*	[22]	ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR7_genome_release
5	vv	Dicot	*Vitis vinifera*	[23]	http://www.genoscope.cns.fr/
6	mt	Dicot	*Medicago truncatula*	No	http://www.medicago.org/genome/downloads/Mt2/ and ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Medicago_truncatula
7	gm	Dicot	*Glycine max*	No	JGI V1.0
8	os	Monocot	*Oryza sativa*	[24, 25]	ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/o_sativa/annotation_dbs/pseudomolecules/version_5.0
9	sb	Monocot	*Sorghum bicolor*	[26]	JGI V1.0
10	zm	Monocot	*Zea mays*	No	http://ftp.maizesequence.org/release-3a.50
11	pv	Monocot	*Panicum virgatum*	No	plantGDB (ftp://ftp.plantgdb.org/download/PUT/Panicum_virgatum/)

corresponding to the A to D subdomains of the N-terminal DNA-binding region of the NAC proteins (Fig. 1) [5]. An additional subdomain E, right after domain D, is also important for DNA binding [6]. We therefore built another HMM including the A to E subdomains based on the multiple-sequence alignment of ten NAC protein sequences from diverse organisms including the moss *Physcomitrella* [6]. This A–E HMM is 155 amino acids long. We also built five HMMs one for each of the five individual subdomains, which are 22 (A), 16 (B), 37 (C), 31 (D), and 17 (E) amino acids long, respectively. The A–E HMM was used to search against the protein databases of the above 11 species (Table 1) and also against the NCBI-nr database. E value cutoff of ≤1.0 was applied to keep significant matches, which were further analyzed by the five individual subdomain HMMs in order to locate boundaries of domains A to E in the full-length proteins. The protein homologs found in the search are listed under their systematic names in Supplemental Table 1 (Table S1). The subfamily groups and the presence or absence of different subdomains in each subfamily are listed in Table S2.

EST Search

EST sequences of the 11 genomes were downloaded from the plantGDB database [27] originally retrieved from the GenBank EST database. The full-length NAC proteins were searched against the ESTs using the E value cutoff of $\leq 1e^{-2}$; ESTs that were more than 98% identical to the query were retained for further analysis.

Phylogenetic Analyses

For phylogenetic reconstruction, multiple protein sequence alignments (MSAs) were performed on the NAC domains (A–E) of all the candidates. MAFFT v6.603 [28] was used in the alignment employing L-INS-I, which is considered to be one of the most accurate MSA methods [29, 30]. Maximum likelihood (ML) trees were built on the MSA datasets by using PhyML v2.4.4 [31]. Specifically, PhyML analyses were conducted with the JTT model, 100 replicates of bootstrap analysis, estimated proportion of invariable sites, four rate categories, estimated gamma distribution parameter, and optimized starting BIONJ tree.

Motif Identification

Protein sequence motifs were identified using the MEME program [32] (http://meme.nbcr.net/meme3/meme.html). The analysis parameters were set as follows: number of repetitions, any; maximum number of motifs, 50; and optimum width of the motif, ≥15. The motif profile for each protein is presented schematically. TMHMM transmembrane motifs were identified with TMHMM Server v. 2.0 (http://www.cbs.dtu.dk/services/TMHMM/) with default setting.

Analysis of Intron–Exon Structure

Gene structure information was parsed from the GFF file downloaded along with the genome data and was used as the input for graphic display at the Gene Structure Display Server of Peking University [33].

Results

Genome-Wide Systematic Phylogeny Analyses of Plant NAC Proteins

Collection of Plant NAC Protein Sequences NAC proteins contain a conserved N-terminal region, which is usually comprised of five subdomains (A–E; Fig. 1). We have built an NAC A–E domain HMM based on the multiple sequence alignment and then searched this HMM model against the proteomes of 11 species (Table 1) using HMMER [34]. Our search identified a total of 1,311 NAC proteins (Tables 2, S1, S2). Proteins of the 11 species were either predicted from the genome sequences (complete or partial) or from assembled ESTs. Our data show that genomes of higher plants generally encode between 100 and 200 NAC proteins except for grape that encodes only 80 NAC proteins. In contrast, genomes of lower plants, namely moss and spike moss, encode 40 or fewer NAC proteins (Tables 2, S1).

Different naming conventions have been used for NAC proteins across different organisms by different annotation teams, making discussions about the whole family of NAC proteins challenging. We therefore renamed most of the NAC proteins following the naming convention of NAC proteins in *Arabidopsis* and *Oryza sativa.* Table S1 provides a mapping table between our assigned names for the NAC proteins and the original IDs.

Compared to the published results on *NAC* gene identification, we identified 11 more *Arabidopsis NAC* genes than those reported in a 2003 paper [5] and nine more rice *NAC* genes than reported in a 2008 paper [19]. Our predictions on the *NAC* genes from the other nine species represent the first such report for any of the nine genomes. In addition to the 1,311 proteins identified from the 11 plant genomes or ESTs, we also included 21 additional proteins (1,332 total) collected from published literature, most of which have been functionally characterized by genetic experiments. These 21 *NAC* genes, together with 29 well-characterized NAC proteins all from *Arabidopsis* (50 in total), were used as references for our protein function prediction (see below).

Table 2 The number of *NAC* genes (in total and specific subgroups) in the 11 plant genomes after removal of the 100 fragmented sequences

Species[a]	All	NAC-a	NAC-b	NAC-c	NAC-d	NAC-e	NAC-f	NAC-g	NAC-h
pp	35	12	4	9	8	1	0	1	0
sm	42	11	2	8	7	10	0	2	2
pt	148	26	32	14	16	19	11	14	16
at	115	17	35	13	17	9	5	15	4
vv	75	22	11	8	9	16	3	6	0
mt	93	26	16	7	18	11	6	8	1
gm	177	41	29	21	31	23	10	22	0
os	144	23	12	10	23	15	10	15	36
sb	113	20	12	10	21	13	3	15	19
zm	177	29	26	24	31	29	12	19	7
pv	92	20	10	6	26	4	19	5	2
Total	1,211	247	189	130	207	150	79	122	87

[a] See legend to Table 1 for species abbreviations

Subfamily Classification of NAC Proteins The A–E domains of the 1,332 NAC proteins were aligned using the multiple-sequence alignment program MAFFT. An ML phylogeny was built using the PhyML program. Eight monophyletic clades were found in the phylogeny, with only five protein sequences not included in any of the eight clades (data not shown). We manually checked the multiple-sequence alignment for the sequences of each clade and found 100 apparent fragmented sequences, mostly having incomplete NAC A–E domains, which were excluded from the multiple-sequence alignment. The resulting 1,232 sequences were realigned, and a new phylogenetic tree was reconstructed using the same procedure (Fig. 2). The five individual ungrouped sequences now all cluster into one of the eight monophyletic clades. Thus, excluding the problematic sequences, the sequence alignment improved and hence the phylogenetic clustering. While we were not able to perform bootstrap analyses as the PhyML program runs very slowly on such large datasets, we did conduct bootstrap analyses for each of the eight clades (see below).

We named the eight NAC clades as NAC-a to NAC-h subfamilies (colored clades, Fig. 2). The breakdown of different NAC subfamilies in different species shows that all of the eight subfamilies tend to be found in both monocot and dicot species (Table 2), except that the NAC-h clade

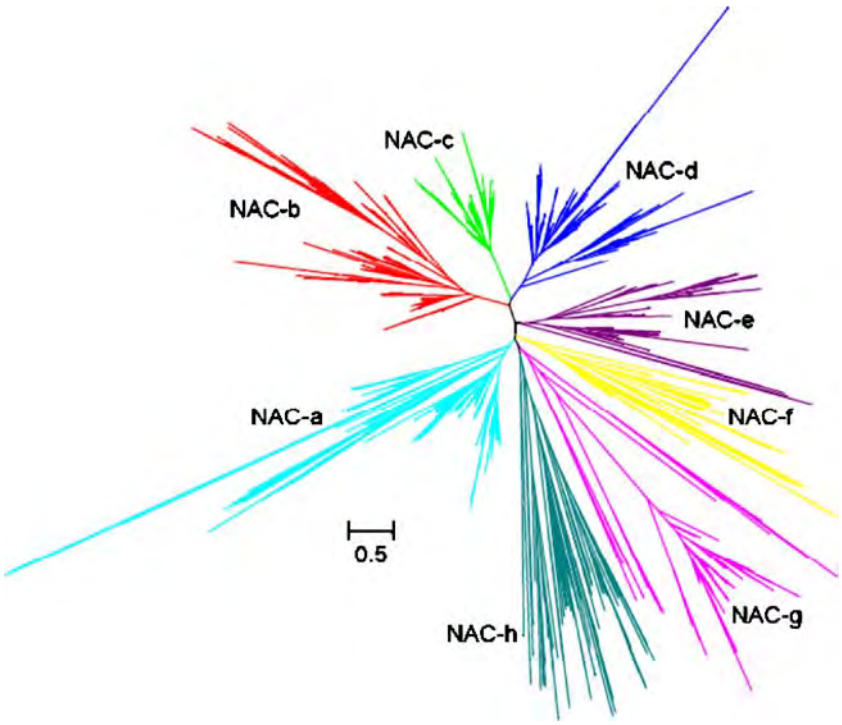


Fig. 2 ML phylogeny of 1,232 NAC proteins. Only the NAC domain regions A–E were used in the reconstruction. The 1,232 proteins include 1,211 NAC proteins from the 11 plant species listed in Table 1 and an additional 21 NAC proteins collected from the literature. One hundred NAC proteins were excluded from the analysis as they appear to be fragmented. *Bar* shows the distance scale for branch length (amino acid substitutions per site)

was not found in the grape and soybean genomes. Interestingly, the NAC-f subfamily was not found in moss (*P. patens,* bryophyte) or spike moss (*S. moellendorffii*, lycophyte). Spike moss is thought to be the earliest evolved vascular land plant. These data suggest an early divergence of the NAC subfamilies before the emergence of vascular plants. It has been suggested that all land plants evolved from the aquatic algae. However, a search of the NAC A–E HMM model against six fully sequenced chlorophyte green algae did not find any NAC homologs (data not shown). Hence, our data did not support this speculation.

We also checked the general grouping patterns of the 50 reference proteins across the eight subfamilies. Interestingly, *NAC* genes with different functions appear to fall into different subfamilies (Table S3). For example, the proteins involved in responses to environmental stress or viral infection grouped into subfamily NAC-a. The membrane-associated NAC transcription factors (MTFs) that mediate either cytokinin signaling during cell division or endoplasmic reticulum stress responses [35] were grouped into the NAC-b subfamily (Table S3). These proteins contain transmembrane α-helices in their C-terminals (Fig. 1) and are predicted to be membrane-associated. We checked the 1,232 NAC proteins for having any transmembrane regions using the TMHMM program, which gave 63 hits (Table S2). Interestingly, 89% (56 out of 63) of the MTF proteins are grouped into the NAC-b subfamily, indicating that this subfamily may be associated with cell division and stress responses (Tables S2, S3). The AtNST and AtVND proteins associated with secondary wall formation during fiber development and xylem differentiation, all grouped into the NAC-c subfamily, while the *NAC* genes that function in organ initiation and differentiation seem to be grouped into the NAC-d subfamily [36] (Table S3). Additional functional grouping patterns may appear as more protein functions are characterized in the NAC-e to NAC-h subfamilies. A comparison of our phylogenetic classification with those of two previous reports on *Arabidopsis* and rice NACs [5, 19] is given in Table S4.

Heat Map for the Presence/Absence of the A–E Subdomains Among the 11 species under study, only *Arabidopsis* and rice are fully sequenced and well annotated. Gene prediction and annotations for the other nine genomes are of relatively low quality. For example, some of the identified NAC proteins without some of the five NAC A–E subdomains may represent misannotations while others may represent correct annotations since some annotated full-length rice NAC proteins do not contain some of the five subdomains [19]. To differentiate these two cases, we have built five HMM models for the five individual subdomains. The presence/absence of the A–E subdomains in the 1,232 NAC proteins and the similarity scores are plotted as a heat map in Fig. 3. The heat map clearly shows eight well-clustered clades consistent with our phylogenetic grouping (Fig. 2). Most of the NAC proteins have the A to E subdomains. The NAC-g and NAC-h subfamilies have weaker NAC domain signals and tend to have no B and E subdomains. Therefore, they may be annotated as NAC-like proteins instead of NAC proteins. Although some subfamilies have weaker E subdomain signals, this subdomain is present in most of the NAC proteins, indicating its likely critical biological function. For this reason, we feel that inclusion of subdomain E in the phylogeny analysis is more appropriate than using subdomains A–C or A–D only.

Further Classification of the Subfamilies Each NAC subfamily contains 79–247 proteins (Table 2). These subfamilies can be further classified phylogenetically into smaller sub-subfamilies, named *subgroups* in our study. The subfamily classification is at a higher phylogeny level than those of the previous reports (Table S4). We have built MSAs and ML phylogenies for the eight NAC subfamilies for further classification. As a result, each of the subfamilies was divided into different subgroups according to its tree topology (Fig. S1). For example, the NAC-a subfamily is divided into nine subgroups (Fig. S1a), NAC-b into ten (Fig. S1b), and NAC-c into five (Fig. S1c). Most of the moss or spike moss proteins were grouped into distinct subgroups, indicating their phylogenetic differences from those of higher plant species. Dicot and monocot NACs appear as pairs in the subgroups, such as in NAC-a-1, NAC-a-2, NAC-a-3, NAC-c-3, NAC-c-5, NAC-d-2, NAC-d-5, NAC-d-7, NAC-d-8, NAC-d-9, NAC-d-10, etc. (Fig. S1). This suggests that the biological functions of the NAC proteins in each subfamily diverged before the monocot and dicot separation.

Intron and Exon Structures We plotted the intron–exon structures for the *NAC* genes from the fully sequenced genomes (Fig. S1a–h). The plots have high error bars, presumably as a result of the less-than-perfect annotation of the gene models in the analyzed genomes. Nevertheless, it appears that neighboring genes on the phylogenetic tree tend to have similar intron–exon structures. All the moss and spike moss *NAC* genes tend to have no introns in the N-terminal NAC domain except for *SmNAC19* and *SmNAC20* in the NAC-h subfamily. NAC-h genes have fewer introns than the genes from other subfamilies, supporting the hypothesis that the NAC-h subfamily may contain *NAC-like* genes (Fig. 4b).

NAC Genes for Plant Cell Wall Development

A group of NAC proteins, including AtNSTs, AtSNDs, and AtXND, function as master transcriptional switches for cell

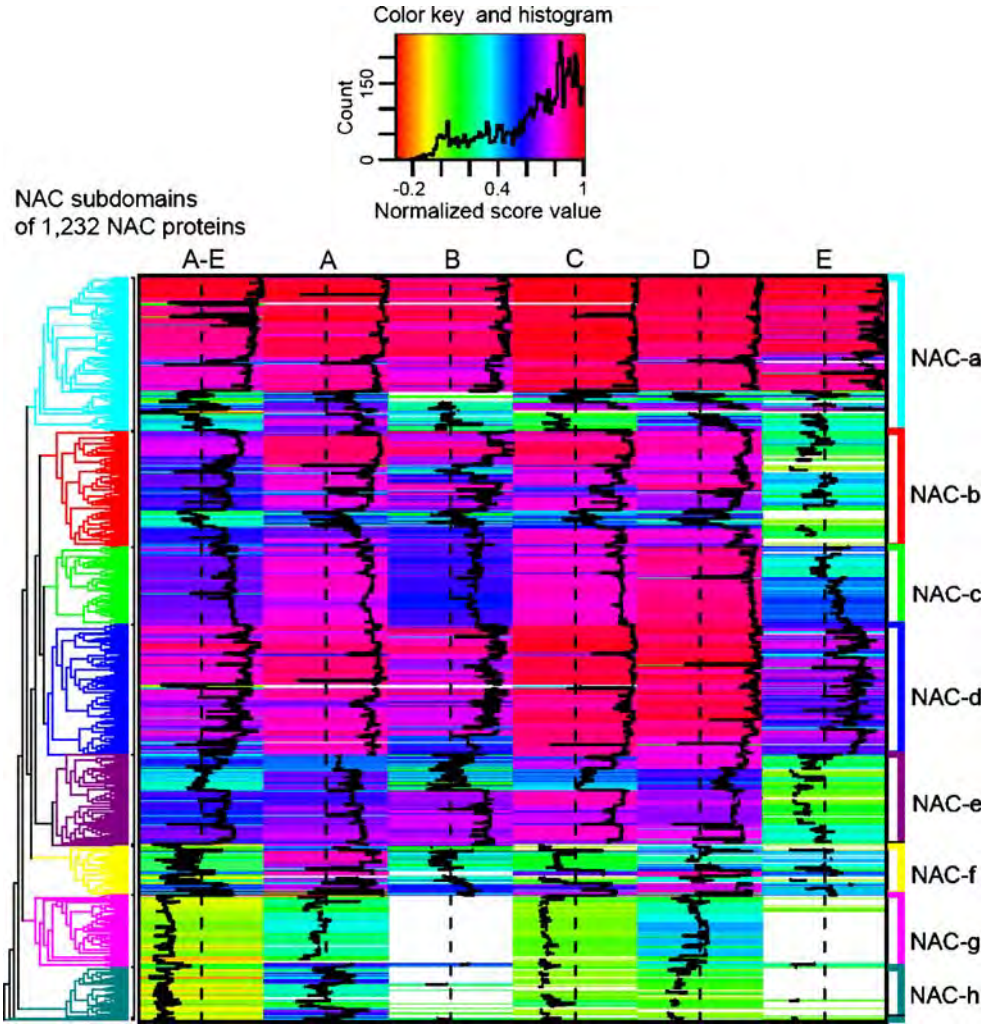


Fig. 3 Heat map showing the presence/absence of the six NAC subdomains. HMMs were built for both the complete NAC A–E region and the five separate A to E subdomains and then searched against the 1,232 NAC proteins (see "Methods"). The HMM search scores were used to measure the strength of the NAC domain and subdomain signals which are shown by the *color key*. Count means the frequency of each of the specific normalized score values across all proteins and all subdomains, shown by the *peaked lines*. *Columns* in the figure correspond to the five A–E subdomains, and *rows* represent the 1,232 proteins ordered according to the phylogeny shown in Fig. 2. If a subdomain is absent in a protein, the color is shown as *white*. If the subdomain is present, the color is defined according to the normalized HMM search score (Table S2). The figure was generated using the gplots package of the R statistics software. (http://cran.r-project.org/web/packages/gplots/index.html)

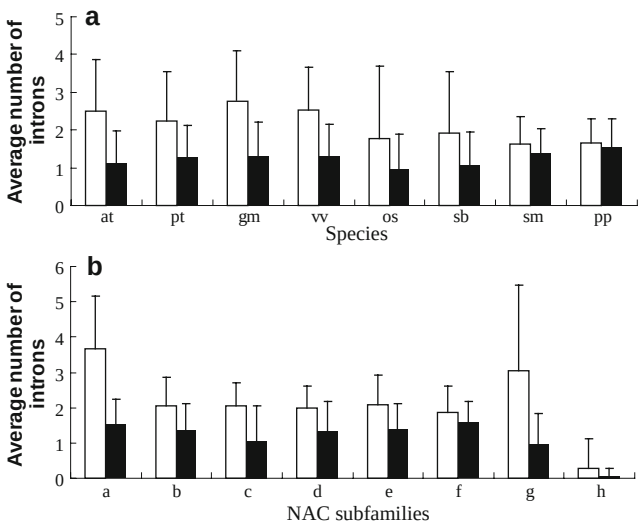


Fig. 4 Intron number statistics for the eight NAC subfamilies. The statistics are based on the gene structure information shown in Fig. S1a–h. *Error bar* shows standard deviation. *Open bar*: average number of introns overall; *black bar*: average number of the introns in the A–E subdomains

wall development, especially in the regulation of secondary cell wall and fiber development and wood formation in vascular plants [9–11]. Genetic modification of these *NAC* genes in *Arabidopsis* has caused dramatic changes in cell wall recalcitrance properties. For this reason, we focused on these subgroups for further analyses.

NST, VND, SND, and XND Are Classified into Different Phylogenetic Subgroups and Motif Clades Our phylogenetic analyses show that ANAC104/AtXND1 is grouped into subgroup a-4 of the NAC-a subfamily, and ANAC073/SND2 and ANAC010/SND3 are grouped into subgroup g-9 of the NAC-g subfamily. AtNST 1, 2, and 3 [11, 14, 15] and AtVND 1–7 [13, 37] are grouped into subgroups c-3, c-4, and c-5 of the NAC-c subfamily. These subfamilies were therefore chosen for further study.

The phylogenetic classification of NAC proteins is based only on sequence similarities among their subdomains A–E (Fig. 1), which did not take the C-terminal sequences into consideration. Considering the C-terminal sequences of the

NAC proteins are highly diverged, we have performed motif analyses across the C-terminals of these proteins, in hope of finding conserved motif patterns across some of the C-terminals. We ran the MEME program to find sequence motifs in these regions and then grouped the subgroups into different clades sharing conserved motif patterns. This resulted in 20 motif clades for the NAC-a subfamily (named as a-sc1 to a-sc20), 17 motif clades for the NAC-c subfamily (c-sc1 to c-sc17), and 11 motif clades for the NAC-g subfamily (g-sc1 to g-sc11; Figs. 5, S3, S4).

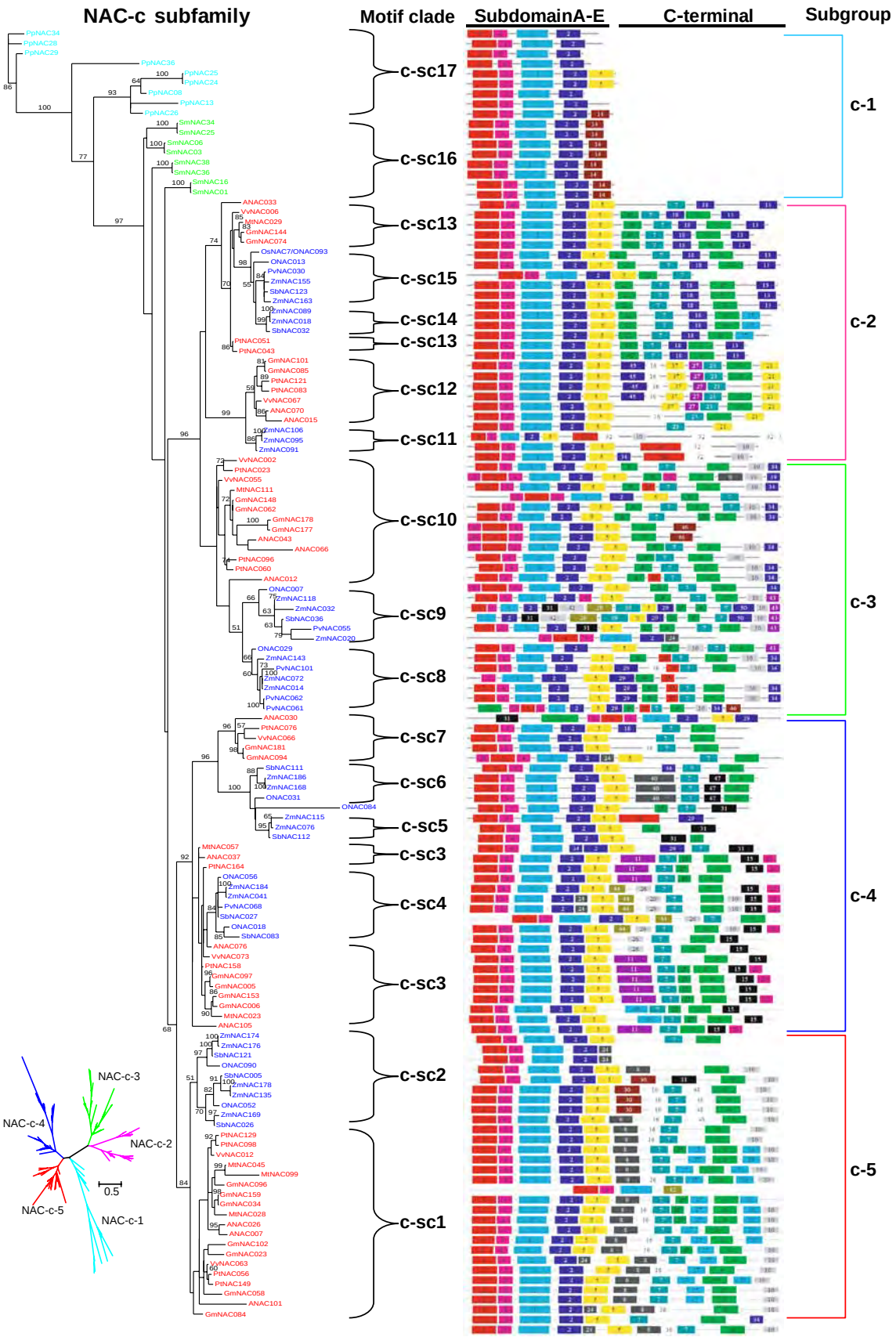
The analysis results clearly show that the C-terminal motifs also form conserved patterns, in addition to the N-terminal subdomains. While the above results indicate that there are conserved motif patterns across subgroups, there are no conserved motif patterns across different subfamilies. For example, the c-3 subgroup contains three motif clades, c-sc8, c-sc9, and c-sc10 (Fig. 5), and the g-9 subgroup has two motif clades, g-sc1 and g-sc2 (Fig. S4). The motif pattern clusters in the same way as the subgroup pattern, suggesting that our phylogenetic clustering result is accurate. The NAC proteins from moss and spike moss can be classified into distinct motif clades. Most of these proteins have no C-terminal motifs, and translation stops after the E subdomain. Only SmNAC26, SmNAC18 (from the a-sc19 clade), PpNAC02, PpNAC22, PpNAC23 (from the a-sc6 clade), and PpNAC027 (from the NAC-d subfamily) have distinguishable motifs downstream of subdomain E. Both monocot and dicot motif patterns are present within each subgroup. For example, the c-5 subgroup contains c-sc1 (dicot) and c-sc2 (monocot) motif clades; the a-9 subgroup contains a-sc1 (monocot) and a-sc2 (dicot) motif clades, and the g-9 subgroup contains g-sc1 (monocot) and g-sc2 (dicot) clades (Figs. 5, S3, S4). The dicot and monocot clades share similar motif patterns but have distinct amino acid variations within the motifs (Figs. 6, S4a-e, S5a-e).

NST, VND, SND, and XND Have Different Sequence Motifs at Their C Termini Similar C-terminal motif patterns were found in different subgroups within one subfamily (Figs. 6, S4, S5). For example, the C.M9, C.M7, C.M6, and C.M10 motifs were found in subgroup c-3 with *AtNST1*, *AtNST2*, and *AtNST3* as representative genes. C.M7, C.M6, and C.M10 motifs were found in subgroup c-1 with *AtVND4*, *AtVND5*, and *AtVND6* as representative genes. C.M7 and C.M6 were also found in other subclades except for sc5, sc11, sc16, and sc17 (Fig. S4a). We searched all the 1,232 NAC proteins for the C.M7 and C.M6 motifs, using HMM models for both motifs; the search did not give any significant hits except for proteins in the NAC-c subfamily although not every NAC protein in the NAC-c subfamily has both the C.M7 and C.M6 motifs. For example, ANAC033, ONAC031, ONAC084, VvNAC066, and MtNAC111 have only the C.M7 motif; ANAC070, PtNAC023, GmNAC177, GmNAC178, GmNAC 023, GmNAC058, ONAC090, and SbNAC121 have only the C.M6 motif, and ANAC015, PtNAC121, ONAC031, and ONAC084 have no hits for either C.M7 or C.M6.

Fig. 5 Motif clades and subgroups for the NAC-c subfamily. Subgroups are c-1 to c-5; motif clades are c-sc1 to c-sc17. *Red*/*blue-colored* tree ID indicates NAC proteins from dicots and monocots, *light blue* PpNACs, and *green* SmNACs. The bootstrap values of the branches are shown in *black numbers*. The MEME motifs are shown as *different-colored boxes*; the *solid red*, *pink*, *light blue*, *blue*, and *yellow boxes* at the N-terminal indicate the NAC domain region

Some C-terminal motif sequences that were named the same by the MEME program were actually different. The consensus logo and the motif-specific sequences for NST, VND, SND, and XND are shown in Figs. 6, S4, and S5. The sequence similarities of certain motifs are much higher at the clade than at the subgroup level. For example, the sequence similarity between the C.M7 motifs c-sc8, c-sc9, and c-sc10 (in subgroup c-3) is much higher than that between those in sc13, sc14, and sc15 (in subgroup c-2; Fig. S5b). A similar situation was observed with the C.M6 (Fig. S5c), C.M10 (Fig. S4e), A.M8 (Fig. 6), and G.M7 (Fig. S5) motifs. To distinguish the sequence-specific motifs, we call them by clade name, e.g., C.M7 (sc10) and C.M7 (sc1). C.M9 (sc8, sc9, sc10), C.M7 (sc8, sc9, sc10), and C.M6 (sc8, sc9, sc10) are the specific C-terminal motifs for the NST proteins. The C.M9 motif was found only in the c-3 subgroup, where it appears upstream of subdomain E. The L_7, D_8, L_{11}, M_{14}, and G_{15} positions of motif C.M9 are highly conserved (Fig. S4d), although it is not clear what functional role this motif may have in transcriptional regulation (if considered as C-terminal) or in DNA binding (very close to domain E). The C.M7 (sc1-4, sc1-6, and sc1-7) and C.M6 (sc1-4, sc1-6, and sc1-7) motifs are specific to VND proteins (Fig. S4b, c), and the G.M19 and G.M25 motifs appear to be SND protein specific. The G.M7 motif was also found in the g-sc1 to g-sc5 clades (Fig. S5). The G.M47 motif is g-sc1 and monocot specific (Fig. S5b, e). The A.M8 (sc17 and sc18) motif is found in the XND proteins, and the sc18 motif has an aspartic acid (D)-rich region. A.M8 is also found in the a-sc6, a-sc19, and a-sc20 motif clades, but the sequences are different. A.M8 (sc20) has no E_3, S_4, or L_5 but possesses a serine (S)-rich region (Fig. 6). SmNAC26 and SmNAC18 have a specific A.M8 (sc19) motif, and PpNAC22 and PpNAC23 have a specific A.M8 (sc6) motif at their C termini. The potential biological functions of SmNAC26, SmNAC18, PpNAC22, and PpNAC23 in moss and spike moss are intriguing.

The information summarized above provides a highly useful guideline for prioritizing candidate genes from poplar and switchgrass for cell wall modification. For


NAC-c subfamily
Motif clade
SubdomainA-E
C-terminal
Subgroup
c-sc17
c-sc16
c-sc13
c-sc15
c-sc14
c-sc13
c-sc12
c-sc11
c-sc10
c-sc9
c-sc8
c-sc7
c-sc6
c-sc5
c-sc3
c-sc4
c-sc3
c-sc2
c-sc1
c-1
c-2
c-3
c-4
c-5
NAC-c-3
NAC-c-4
NAC-c-2
NAC-c-5
NAC-c-1
0.5

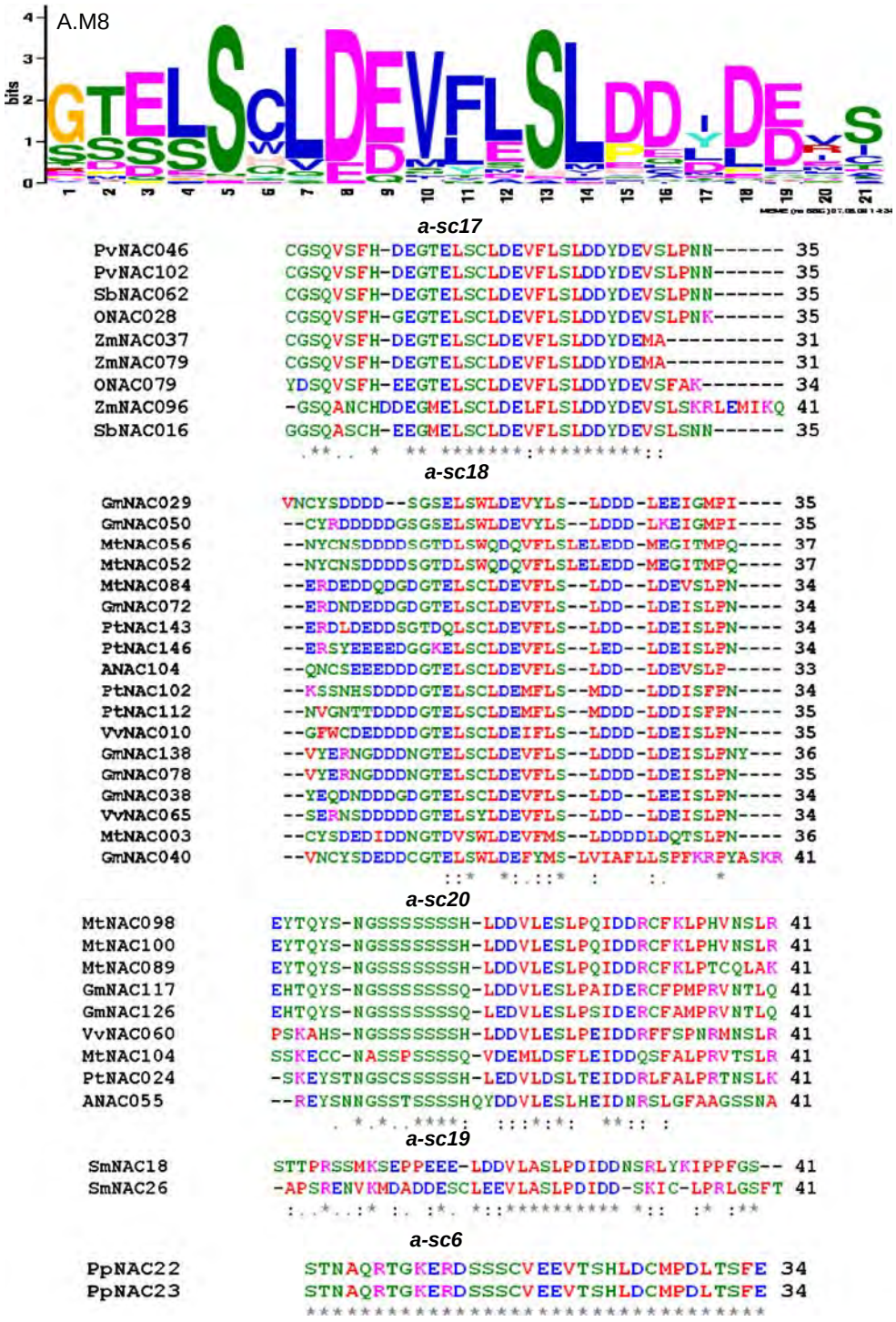


Fig. 6 Sequence-specific MEME motifs for NAC-a (XND) proteins. A.M8 (a-sc17) and A.M8 (a-sc18) are the XND-specific motifs identified in dicot and monocot plants, respectively

example, PvNAC062 (from the c-sc9 clade), PvNAC068 (from the c-sc4 clade), PvNAC046, PvNAC102 (from the a-sc17 clade), PvNAC032, and PvNAC033 (from the g-sc1 clade) seem to be good candidates. PvNAC055 (c-sc9), PvNAC061, and PvNAC101 (c-sc8) are partial ESTs; full-length cDNAs need to be cloned for further analyses of these potential target genes.

Coexpression and Coinduction Pattern Analysis of Plant Cell Wall Development Genes To provide further support for selection of candidate *NAC* genes for cell wall modification study, we performed gene coexpression and induction pattern analysis with all the *Arabidopsis NAC* and secondary cell wall development marker genes using the GENEVESTIGATOR program [39] (see Table S5 for the marker gene list).

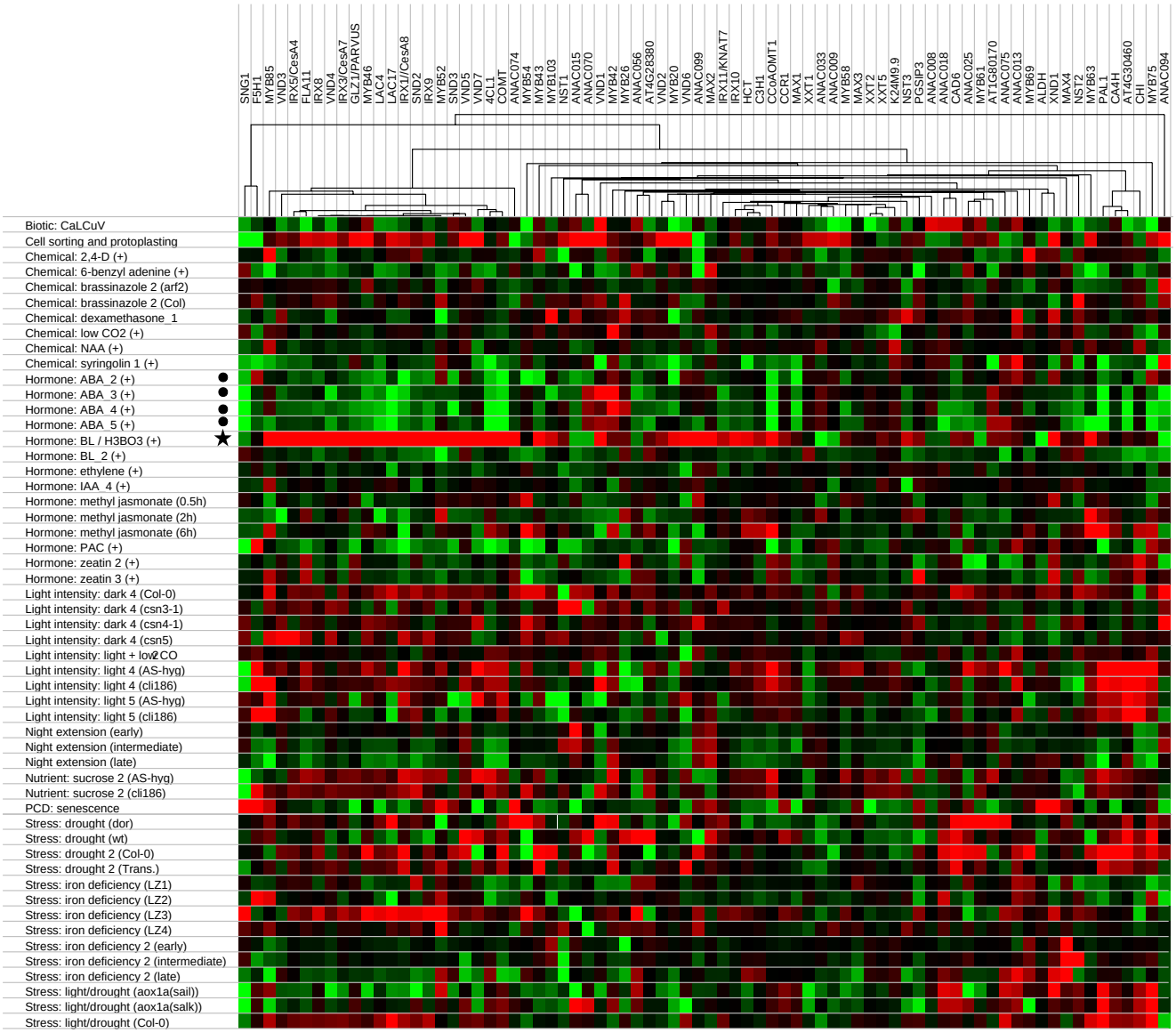


Fig. 7 Induction coexpression analysis of selected *Arabidopsis NAC* genes. The *Arabidopsis NAC* genes from the NAC-a, NAC-c, and NAC-g subfamilies, together with the plant cell wall biosynthesis marker genes (Table S5), were analyzed for coinduction patterns by GENEVESTIGATOR program. The hierarchical clustering was calculated by Pearson correlation. *Star* BL/H_3BO_3 treatment; *dots* ABA treatment

We have observed tissue-specific coexpression between lignin synthetic genes and the secondary cell wall polysaccharide synthetic genes (red and blue boxes, Fig. S6). *AtVND2*, *AtVND3*, *AtVND4*, *AtVND6*, and *AtVND7* from subfamily NAC-c and *SND3* from subfamily NAC-g have similar expression patterns to those of the lignin synthetic genes, and the expression patterns of *AtNST1*, *AtNST2*, and *AtNST3* from the NAC-c subfamily and SND2 from the NAC-g subfamily are similar with those of the secondary cell wall polysaccharide synthetic genes. The *AtNST* and *AtVND* genes have different and tissue-specific expression patterns. AtXND1 is very strongly expressed in hypocotyls and xylem.

The *Arabidopsis* genes from the NAC-a, NAC-c, and NAC-g subfamilies, together with the plant cell wall biosynthesis marker genes, were analyzed for coinduction patterns following different chemical and environmental treatments (Fig. 7). Cell wall developmental processes are controlled by a complex signal matrix, where specific signals trigger unique response patterns for this group of genes. Nevertheless, we did find several interesting coinduction patterns. Expression of most phenylpropanoid and lignin biosynthetic genes and cell wall polysaccharide biosynthesis/maturation genes was highly induced by treatment with brassinolide/boric acid (BL/H_3BO_3), which

induces xylogenesis in *Arabidopsis* suspension cultures [13]; this was accompanied by upregulation of AtNSTs, AtVNDs, AtSNDs, and AtXND1 (Fig. 7). At the same time, expression of the MYB transcription factors MYB54, MYB61, and MYB69 was downregulated, suggesting that these may be negative regulators of these processes. Prolonged treatment with ABA (3~5 h) shows the opposite effect to that of BL/H_3BO_3.

Previous reports place ANAC033, ANAC070, ANAC015, AtVND, and AtNST proteins into a single phylogenetic group: the OsNAC7 group in [5] and the I-1/OsNAC7 group in [19]. Our phylogenetic analysis grouped them with AtNST and AtVND into different subgroups; and our motif analysis suggests that these genes may have different functions than previously suggested. Furthermore, ANAC033 has only the C.M7 motif at the C terminus; ANAC070 has only the C.M6 motif, and ANAC015 has neither C.M7 nor C.M6 motif. The sequences of the C.M7 and C.M6 motifs in ANAC033 and ANAC070 are different from those of the AtVNDs and AtNSTs. The tissue expression pattern from GENEVESTIGATOR [39] shows that *ANAC033*, *ANAC070*, and *ANAC015* are mainly expressed in lateral and primary root cap rather than in secondary cell-wall-enriched tissues (green box, Fig. S6). Indeed, ANAC009, ANAC015, ANAC033, ANAC070, and ANAC094 cluster together in the tissue coexpression analysis (green box, Fig. S6). Moreover, expression of these genes is repressed by BL/H_3BO_3 treatment, whereas the *AtNST* and *AtVND* genes are induced (Fig. 7). Thus, the combination of information from phylogenetic clustering, motif analysis, and coexpression pattern analysis leads to the conclusion that the ANAC009, ANAC033, ANAC070, and ANAC015 proteins are more likely to function in lateral root cap development than in secondary wall synthesis in *Arabidopsis*. Indeed, ANAC033/SMB and ANAC009/FEZ have been shown to control the orientation of the cell division plane in root stem cells (Table S3), but the exact functions of ANAC015 and ANAC070 remain unknown.

Tissue coexpression analysis shows that ANAC104/AtXND1, ANAC056 (from the NAC-a-5 subgroup), and NAC074 (from the NAC-d-6) cluster together (orange box, Fig. S6) and are all expressed in hypocotyls and xylem tissue; ANAC056 is also highly expressed in the silique tissue, and ANAC074 is highly expressed in the petals. This suggests that ANAC056 and NAC074 may have specific functions in xylem cells of the silique and petal tissues. ANAC056 is repressed by BL/H_3BO_3 and ABA treatments, indicating that it might function as a negative regulator of xylem vessel and fiber development during silique maturation.

Tissue coexpression analysis identified ANAC075 as grouping with the lignin synthetic gene pattern (red box, Fig. S6). Phylogenetic analysis indicates that ANAC075 and ANC099 are tightly clustered in the g-sc4 clade of the NAC-g subfamily (Fig. S4). Interestingly, ANAC099 is only expressed in the peripheral endosperm, suggesting a specific role in seed development (Fig. S6). Both ANAC075 and ANC099 are induced by BL/H_3BO_3. Together, these observations suggest that ANAC075 may be a new candidate gene for control of cell wall development in *Arabidopsis*. Clade g-sc5 is the paired monocot group to g-sc4, and PvNAC066 was identified in this clade. PvNAC066, together with PtNAC085 and PtNAC169 from g-sc4, are therefore candidate genes for controlling cell wall development in switchgrass and poplar.

Discussion

Phylogenetic Analysis

We have presented a detailed phylogenetic analysis of NAC proteins, based on the DNA-binding domains. Using this approach, the subfamilies group well with known NAC function classes. For example, the NAC-a, NAC-b, NAC-c, and NAC-d subfamilies represent four well-known functions of the NAC proteins: (1) response to biotic or abiotic stress, (2) cytokinin signaling during cell division or endoplasmic reticulum stress responses, (3) regulation of plant cell wall development, and (4) organ initiation and formation, respectively. NAC proteins with different functions are grouped into different subgroups, as are the NAC proteins from dicots and monocots. Different C-terminal motif patterns fall into different motif clades within the subgroup pattern, and each motif clade has specific C-terminal motif sequences.

Our *NAC* gene classification is largely consistent with previous classifications, although there are some inconsistencies (Table S4). There are several reasons for this. First, some of our families are at a higher level of classification. Second, there are ten times more NAC sequences included in our analysis (1,211 from 11 species) than in the previous studies, limited to *Arabidopsis* and rice (105 and 75 sequences, respectively) [5] and rice alone (140 sequences) [19]. Third, different tree-building algorithms were used in the phylogenetic analyses; we used the ML algorithm and applied more realistic evolutionary models, e.g., by considering and estimating the fraction of amino acids to be invariable during evolution and assigning each site a probability to belong to given evolutionary rate categories. The *Arabidopsis*/rice study used the neighbor-joining (NJ) algorithm (software and parameters not indicated) [5], and the later study on rice used the Bayesian method-based MrBays program (parameters not indicated) [19]. It is well established that the ML and Bayesian algorithms are more accurate in phylogenetic reconstruction

than the NJ algorithm [41, 42]. Finally, the present work and the *Arabidopsis*/rice study used the A–E subdomains for MSAs and subsequent tree building, whereas the rice only study used only the A–D subdomains in the analyses.

Although phylogenetic analysis provides important bioinformatics support for candidate gene selection, we are aware that it alone cannot unequivocally indicate function. For this reason, we combined phylogenetic grouping, specific motif identification, and tissue/induction coexpression analysis to increase the possibilities for finding the best candidates.

Evolution of NAC Proteins

To the best of our knowledge, this is the first systematic analysis of plant-specific transcription factors in the genomes of moss (*P. patens*) and spike moss (*S. moellendorffii*). These two species have special importance for the study of plant evolution. The mosses together with the green algae, liverworts, and hornworts are the four main phyla of nonvascular plants. Except for the green algae, the other three plant groups are traditionally called bryophytes, which are among the earliest land plants [49]. Spike moss is considered to be the first evolved vascular land plant and thus is particularly useful for the study of the development of secondarily thickened plant cell walls, which were necessary to provide support in a nonaqueous environment.

The phylogenetic analyses and the distribution of different NAC subfamilies in different plant species clearly suggest that different NAC protein subfamilies have already diverged in the bryophytic mosses. However, genome-wide transcription factor search in photosynthetic species found that no NAC transcription factors exist in the algae [54]. Currently, all fully sequenced green algae are monocellular chlorophytes, and the existence of *NAC* genes in the charophytic and streptophytic green algae is unknown. More sequencing of lower-order plant genomes will help to resolve the question of the evolutionary origin of the *NAC* genes.

Although different NAC subfamilies have diverged in lower plants, our intron–exon gene structure analyses revealed that most of the moss and spike moss *NAC* genes lack the C-terminal exon, only containing the N-terminal exons encoding the NAC A–E subdomains. Some *P. patens* genes even lack the E subdomain. Since the C-terminal domain of NAC transcription factors is the activation domain, the functions of *NAC* genes in moss may be different from those of their homologs in seed plants. For example, the C.M7 and C.M6 motifs in the ANAC012 were called LP and WQ motifs [55], and it was suggested that they may be specific to vascular plants [55]. Eight SmNAC proteins and nine PpNAC proteins are grouped into the NAC-c subfamily. Surprisingly, however, no C-terminal motifs were identified in either PpNAC or SmNAC proteins in the NAC-c subfamily; their translation stops immediately after the E subdomain. Thus, the prediction that the C.M7/LP and C.M6/WQ motifs might be specific for vascular plants appears to be correct. Functional analyses of the *ANAC012* gene in *Arabidopsis* show that both the C.M7/LP and C.M6/WQ motifs are required for transcription activation in a yeast two-hybrid system [55]. Deletion of the C.M6/WQ motif at the C terminus abolished the transcriptional activation of ANAC012 [55]. However, the biological functions of these motifs still need to be investigated in planta. It is not clear whether NST proteins function as key regulators of the vascular development in spike moss, and the evolutionary initiation of the gene-specific motifs at the C-terminal regions of the NAC proteins remains a mystery.

NAC Genes for Biomass Modification

The highly lignified plant cell walls in fibers and xylem vessels contribute to the recalcitrance of biomass. Modifying plant cell wall recalcitrance properties combined with increasing biomass production will be critical for improving the economics of lignocellulosic bioenergy production. In this study, the candidate *NAC* genes that may regulate cell wall development and lignification in bioenergy crops have been identified. However, no studies to date have shown that genetic modification of *NAC* gene expression directly impacts recalcitrance, although this would be predicted from the effects on cell wall properties. Based on the phylogenetic and coexpression studies described above, PtNACs 085 and 169 and PvNACs 032, 033, 046, 055, 061, 062, 068, 101, and 102 are all potential targets for modifying cell wall recalcitrance.

Second-generation bioenergy crops such as switchgrass will be planted on marginal lands where natural resources are limited and environmental conditions are often severe. Therefore, abiotic/biotic stresses, such as drought, high light-induced photo damage, salinity, extreme temperature, chemical toxicity, oxidative stress, nutrient (nitrogen and phosphate) limitations, and virus/insect/fungal infections will all be potential threats to biomass production.

Many NAC proteins from the NAC-a subfamily have been directly linked to abiotic and biotic stress responses (Table S3). Phylogenetic analyses of rice NAC proteins identified a stress-responsive SNAC group. Interestingly, all these ONAC proteins group into different subgroups of the NAC-a subfamily [19] (Table S4). Overexpression of rice *OsNAC6* results in significant cold-stress tolerance when the plants were grown at 4–8°C for 5 days [68]. OsNAC6, ANAC002, ANAC081, CaNAC1 (defense response), and StNAC (wounding response) were clustered in subgroup a-1 of subfamily NAC-a. Overexpression of the stress-responsive gene stress-responsive NAC 1 in rice significantly enhances drought tolerance, yielding a 22–34% higher seed setting rate than controls during the reproduc-

tive stage under severe drought stress conditions in the field. The transgenic rice also shows significantly improved drought and salt tolerance at the vegetative stage [70]. Overexpression of ANAC019, ANAC055, or ANAC072 in *Arabidopsis* resulted in significantly increased drought tolerance [71]. Two *Arabidopsis* NAC transcription factors, ANAC002/ATAF1 [72] and ANAC081/ATAF2 [73], are induced by wounding, and the corresponding *ataf1-1* and *ataf1-2* mutants displayed a recovery rate about seven times higher than wild-type plants in drought response tests [72], and overexpression of ANAC081/ATAF2 results in higher susceptibility to the soil-borne fungal pathogen *Fusarium oxysporum* but increased biomass yield [73]. PvNAC026, PvNAC027, PvNAC028, PvNAC070, and PvNAC071 are clustered with OsNAC6/ONAC048 in clade a-sc1; these are good candidate genes for abiotic and biotic stress engineering.

Delayed senescence also helps to increase biomass production. Several *NAC* genes, including *NtSENU5*, *ANAC029/AtNAP*, and wheat *NAM-B1*, are involved in this process [7, 74, 75]. Leaf senescence in two T-DNA insertion lines of the *AtNAP* gene is significantly delayed compared to wild type [74]. Loss of function of *AtNAC2* underlies the *ore1* "long-living" mutation in *Arabidopsis*, and a trifurcate feed-forward pathway involving ORE1, ethylene reception, and a micro-RNA links aging to cell death in leaves [76]. Down-regulation of *NAM-B1* in wheat delays senescence by more than 21 days [7] and reduces wheat grain protein, zinc, and iron content by more than 30%, possibly via defects in nutrient remobilization from leaves to grains. This type of control is of particular importance for lignocellulosic bioenergy grasses, where vegetative growth for biomass yield of the leaf and stem is more important than grain yield. At the same time, mineral nutrient mobilization from the aboveground tissues to underground root tissue at the mature stage immediately prior to harvesting will benefit sustainability and land conservation. Unfortunately, no PvNACs have yet been found in the a-sc15 subclade of the NAC-a-5 subgroup. From our phylogenetic tree, PtNAC127, PtNAC155, and PtNAC115 from poplar and MtNAC101 from *Medicago* are tightly clustered with NAM-B1 in subgroup NAC-a-5. PvNAC053, PvNAC054, PtNAC049, PtNAC074, and MtNAC009, MtNAC038, and MtNAC046 are very close to ANAC029 in the NAC-a-4 subgroup. Therefore, these are all candidate NAC proteins for promotion of leaf senescence in poplar and *Medicago*. Engineering such targets genes may benefit biomass yield in the future.

Biomass production correlates directly with the tiller number in grasses. Recent studies on switchgrass have shown that tiller density and mass per phytomer consistently show large positive correlations with biomass yield in field tests [77, 78]. Only one report to date indicates that *NAC* genes can control tiller number; OsNAC2/OsTIL1 promotes shoot branching in rice. The activation-tagged gain-of-function *Ostil1* mutant plant showed increased tillers, enlarged tiller angle, and a semi-dwarf phenotype. Overexpression of OsNAC2 promotes tiller bud outgrowth. These data suggest that overexpression of OsNAC2/OsTIL1 may improve plant structure for better light-use efficiency and higher biomass yield potential [8]. OsTIL1 belongs to the d-1 subgroup of the NAC-d subfamily. PvNAC074 clusters with OsTIL1 in this small clade.

Through our phylogenetic analysis, we have predicted many candidate genes that may function in controlling stress responses, senescence, and tiller number, in addition to cell wall properties. Engineering of some of these *NAC* genes to generate transgenic potentially stress-tolerant biofuel crops with improved bioprocessing efficiency is in progress.

Acknowledgments This work was supported by grants to RAD and YX from the US Department of Energy Bioenergy Research Centers, through the Office of Biological and Environmental Research in the DOE Office of Science.

References

1. Ferrell JE, Wright LL, Tuskan GA (1995) Research to develop improved production methods for woody and herbaceous biomass crops. Paper presented at the conference: 2. Meeting on biomass of the Americas, Portland, Oregon, 21–24 Aug 1995
2. McLaughlin S, Bouto J, Bransby D, Conger B, Ocumpaugh W, Parrish D et al (1999) Developing switchgrass as a bioenergy crop. ASHS, Alexandria, pp 282–299
3. Chen F, Dixon RA (2007) Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol 25:759–761
4. Riechmann JL, Heard J, Martin G, Reuber L, Jiang C, Keddie J et al (2000) *Arabidopsis* transcription factors: genome-wide comparative analysis among eukaryotes. Science 290:2105–2110
5. Ooka H, Satoh K, Doi K, Nagata T, Otomo Y, Murakami K et al (2003) Comprehensive analysis of NAC family genes in *Oryza sativa* and *Arabidopsis thaliana*. DNA Res 10:239–247
6. Olsen AN, Ernst HA, Leggio LL, Skriver K (2005) NAC transcription factors: structurally distinct, functionally diverse. Trends Plant Sci 10:79–87
7. Uauy C, Distelfeld A, Fahima T, Blechl A, Dubcovsky J (2006) A NAC gene regulating senescence improves grain protein, zinc, and iron content in wheat. Science 314:1298–1301
8. Mao CZ, Ding WN, Wu YR, Yu J, He XW, Shou HX et al (2007) Overexpression of a NAC-domain protein promotes shoot branching in rice. New Phytol 176:288–298
9. Zhong R, Ye Z-H (2007) Regulation of cell wall biosynthesis. Curr Opin Plant Biol 10:564–572
10. Demura T, Fukuda H (2007) Transcriptional regulation in wood formation. Trends Plant Sci 12:64–70
11. Zhong R, Lee C, Zhou J, McCarthy RL, Ye Z-H (2008) A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in *Arabidopsis*. Plant Cell 20:2763–2782
12. Zhao CS, Craig JC, Petzold HE, Dickerman AW, Beers EP (2005) The xylem and phloem transcriptomes from secondary tissues of the *Arabidopsis* root-hypocotyl. Plant Physiol 138:803–818
13. Zhao C, Avci U, Grant EH, Haigler CH, Beers EP (2008) XND1, a member of the NAC domain family in *Arabidopsis thaliana*,

negatively regulates lignocellulose synthesis and programmed cell death in xylem. Plant J 53:425–436
14. Kubo M, Udagawa M, Nishikubo N, Horiguchi G, Yamaguchi M, Ito J et al (2005) Transcription switches for protoxylem and metaxylem vessel formation. Genes Dev 19:1855–1860
15. Mitsuda N, Seki M, Shinozaki K, Ohme-Takagi M (2005) The NAC transcription factors NST1 and NST2 of Arabidopsis regulate secondary wall thickenings and are required for anther dehiscence. Plant Cell 17:2993–3006
16. Mitsuda N, Iwase A, Yamamoto H, Yoshida M, Seki M, Shinozaki K et al (2007) NAC transcription factors, NST1 and NST3, are key regulators of the formation of secondary walls in woody tissues of *Arabidopsis*. Plant Cell 19:270–280
17. Zhong RQ, Richardson EA, Ye Z-H (2007) Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of *Arabidopsis*. Planta 225:1603–1611
18. Zhong RQ, Demura T, Ye Z-H (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of *Arabidopsis*. Plant Cell 18:3158–3170
19. Fang Y, You J, Xie K, Xie W, Xiong L (2008) Systematic sequence analysis and identification of tissue-specific or stress-responsive genes of NAC transcription factor family in rice. Mol Gen Genet 280:547–563
20. Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H et al (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69
21. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U et al (2006) The genome of black cottonwood*, Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604
22. The Arabidopsis Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815
23. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467
24. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296:92–100
25. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296:79–92
26. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556
27. Dong QF, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C et al (2005) Comparative plant genomics resources at PlantGDB. Plant Physiol 139:610–618
28. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518
29. Ahola V, Aittokallio T, Vihinen M, Uusipaikka E (2006) A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics 7:484
30. Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471
31. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704
32. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:W369–W373
33. Guo AY, Zhu QH, Chen X, Luo JC (2007) GSDS: a gene structure display server. Yichuan 29:1023–1026
34. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763
35. Kim SY, Kim SG, Kim YS, Seo PJ, Bae M, Yoon HK et al (2007) Exploring membrane-associated NAC transcription factors in *Arabidopsis*: implications for membrane biology in genome regulation. Nucleic Acids Res 35:203–213
36. Xie Q, Frugis G, Colgan D, Chua NH (2000) *Arabidopsis* NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. Genes Dev 14:3024–3036
37. Yamaguchi M, Kubo M, Fukuda H, Demura T (2008) Vascular-related NAC-DOMAIN7 is involved in the differentiation of all types of xylem vessels in *Arabidopsis* roots and shoots. Plant J 55:652–664
38. Kunieda T, Mitsuda N, Ohme-Takagi M, Takeda S, Aida M, Tasaka M et al (2008) NAC family proteins NARS1/NAC2 and NARS2/NAM in the outer integument regulate embryogenesis in *Arabidopsis*. Plant Cell 20:2631–2642
39. Hruz T, Laule O, Szabo G, Wessendorp F, Bleuler S, Oertle L et al (2008) Genevestigator V3: a reference expression database for the meta-analysis of transcriptomes. Adv Bioinformatics 2008:1–5
40. Bu Q, Jiang H, Li CB, Zhai Q, Zhang J, Wu X et al (2008) Role of the *Arabidopsis thaliana* NAC transcription factors ANAC019 and ANAC055 in regulating jasmonic acid-signaled defense responses. Cell Res 18:756–767
41. Hall BG (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol Biol Evol 22:792–802
42. Ogden TH, Rosenberg MS (2006) Multiple sequence alignment accuracy and phylogenetic inference. Syst Biol 55:314–328
43. Rabbani MA, Maruyama K, Abe H, Khan MA, Katsura K, Ito Y et al (2003) Monitoring expression profiles of rice genes under cold, drought, and high-salinity stresses and abscisic acid application using cDNA microarray and RNA gel-blot analyses. Plant Physiol 133:1755–1767
44. Lin R, Zhao W, Meng X, Wang M, Peng Y (2007) Rice gene OsNAC19 encodes a novel NAC-domain transcription factor and responds to infection by *Magnaporthe grisea*. Plant Sci 172:120–130
45. Kikuchi K, Ueguchi-Tanaka M, Yoshida KT, Nagato Y, Matsusoka M, Hirano HY (2000) Molecular analysis of the *NAC* gene family in rice. Mol Gen Genet 262:1047–1051
46. Collinge M, Boller T (2001) Differential induction of two potato genes, *Stprx2* and *StNAC*, in response to infection by *Phytophthora infestans* and to wounding. Plant Mol Biol 46:521–529
47. Liu Z, Shao FX, Tang GY, Shan L, Bi YP (2009) Cloning and characterization of a transcription factor ZmNAC1 in maize (*Zea mays*). Yichuan 31:199–205
48. Oh SK, Lee S, Yu SH, Choi D (2005) Expression of a novel NAC domain-containing transcription factor (CaNAC1) is preferentially associated with incompatible interactions between chili pepper and pathogens. Planta 222:876–887
49. Bowman JL, Floyd SK, Sakakibara K (2007) Green genes—comparative genomics of the green branch of life. Cell 129:229–234
50. Kim Y-S, Kim S-G, Park J-E, Park H-Y, Lim M-H, Chua N-H et al (2006) A membrane-bound NAC transcription factor regulates cell division in *Arabidopsis*. Plant Cell 18:3132–3144
51. Yoon H, Kim S, Park C (2008) Regulation of leaf senescence by NTL9-mediated osmotic stress signaling in *Arabidopsis*. Mol Cells 25:438–445
52. Yoshii M, Shimizu T, Yamazaki M, Higashi T, Miyao A, Hirochika H et al (2009) Disruption of a novel gene for a NAC-domain protein in rice confers resistance to rice dwarf virus. Plant J 57:615–625
53. Mitsuda N, Ohme-Takagi M (2008) NAC transcription factors NST1 and NST3 regulate pod shattering in a partially redundant manner by promoting secondary wall formation after the establishment of tissue identity. Plant J 56:768–778
54. Riano-Pachon DM, Ruzicic S, Dreyer I, Mueller-Roeber B (2007) PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics 8:42

55. Ko JH, Yang SH, Park AH, Lerouxel O, Han KH (2007) ANAC012, a member of the plant-specific NAC transcription factor family, negatively regulates xylary fiber development in *Arabidopsis thaliana*. Plant J 50:1035–1048
56. Hay A, Tsiantis M (2005) From genes to plants via meristems. Development 132:2679–2684
57. Nikovics K, Blein T, Peaucelle A, Ishida T, Morin H, Aida M et al (2006) The balance between the *MIR164A* and *CUC2* genes controls leaf margin serration in *Arabidopsis*. Plant Cell 18:2929–2945
58. Takada S, Hibara K, Ishida T, Tasaka M (2001) The CUP-SHAPED COTYLEDON1 gene of *Arabidopsis* regulates shoot apical meristem formation. Development 7:1127–1135
59. He XJ, Mu RL, Cao WH, Zhang ZG, Zhang JS, Chen SY (2005) AtNAC2, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development. Plant J 44:903–916
60. Kusano H, Asano T, Shimada H, Kadowaki K-i (2005) Molecular characterization of ONAC300, a novel NAC gene specifically expressed at early stages in various developing tissues of rice. Mol Gen Genomics 272:616–626
61. Xie Q, Sanz-Burgos A, Guo H, García J, Gutiérrez C (1999) GRAB proteins, novel members of the NAC domain family, isolated by their interaction with a geminivirus protein. Plant Mol Biol 39:647–656
62. Souer E, Van Houwelingen A, Kloos D, Mol J, Koes R (1996) The *no apical meristem* gene of *Petunia* is required for pattern formation in embryos and flowers and is expressed at meristem and primordia boundaries. Cell 85:159–170
63. Ruiz-Medrano R, Xoconostle-Cazares B, Lucas WJ (1999) Phloem long-distance transport of CmNACP mRNA: implications for supracellular regulation in plants. Development 126:4405–4419
64. Zimmermann R, Werr W (2005) Pattern formation in the monocot embryo as revealed by NAM and CUC3 orthologues from *Zea mays* L. Plant Mol Biol 58:669–685
65. Willemsen V, Bauch M, Bennett T, Campilho A, Wolkenfelt H, Xu J et al (2008) The NAC domain transcription factors FEZ and SOMBRERO control the orientation of cell division plane in *Arabidopsis* root stem cells. Dev Cell 15:913–922
66. Yoo SY, Kim Y, Kim SY, Lee JS, Ahn JH (2007) Control of flowering time and cold response by a NAC-domain protein in *Arabidopsis*. PLoS ONE 2:e642
67. Yokotani N, Ichikawa T, Kondou Y, Matsui M, Hirochika H, Iwabuchi M et al (2009) Tolerance to various environmental stresses conferred by the salt-responsive rice gene ONAC063 in transgenic *Arabidopsis*. Planta 229:1065–1075
68. Ohnishi T, Sugahara S, Yamada T, Kikuchi K, Yoshiba Y, Hirano HY et al (2005) OsNAC6, a member of the NAC gene family, is induced by various stresses in rice. Genes Genet Syst 80:135–139
69. Ogo Y, Kobayashi T, Itai RN, Nakanishi H, Kakei Y, Takahashi M et al (2008) A novel NAC transcription factor, IDEF2, that recognizes the iron deficiency-responsive element 2 regulates the genes involved in iron homeostasis in plants. J Biol Chem 283:13407–13417
70. Hu H, Dai M, Yao J, Xiao B, Li X, Zhang Q et al (2006) Overexpressing a NAM, ATAF, and CUC (NAC) transcription factor enhances drought resistance and salt tolerance in rice. Proc Natl Acad Sci USA 103:12987–12992
71. Tran LSP, Nakashima K, Sakuma Y, Simpson SD, Fujita Y, Maruyama K et al (2004) Isolation and functional analysis of *Arabidopsis* stress-inducible NAC transcription factors that bind to a drought-responsive *cis*-element in the early responsive to dehydration stress 1 promoter. Plant Cell 16:2481–2498
72. Lu PL, Chen NZ, An R, Su Z, Qi BS, Ren F et al (2007) A novel drought-inducible gene, *ATAF1*, encodes a NAC family protein that negatively regulates the expression of stress-responsive genes in *Arabidopsis*. Plant Mol Biol 63:289–305
73. Delessert C, Kazan K, Wilson IW, Van Der Straeten D, Manners J, Dennis ES et al (2005) The transcription factor ATAF2 represses the expression of pathogenesis-related genes in *Arabidopsis*. Plant J 43:745–757
74. Guo Y, Gan S (2006) AtNAP, a NAC family transcription factor, has an important role in leaf senescence. Plant J 46:601–612
75. John I, Hackett R, Cooper W, Drake R, Farrell A, Grierson D (1997) Cloning and characterization of tomato leaf senescence-related cDNAs. Plant Mol Biol 33:641–651
76. Kim JH, Woo HR, Kim J, Lim PO, Lee IC, Choi SH et al (2009) Trifurcate feed-forward regulation of age-dependent cell death involving *miR164* in *Arabidopsis*. Science 323:1053–1057
77. Sharma N, Piscioneri I, Pignatelli V (2003) An evaluation of biomass yield stability of switchgrass (*Panicum virgatum* L.) cultivars. Energy Convers Manag 44:2953–2958
78. Das MK, Fuentes RG, Taliaferro CM (2004) Genetic variability and trait relationships in switchgrass. Crop Sci 44:443–448