Genetics and Molecular Biology, 28, 3 (suppl), 511-519 (2005)

www.sbg.org.br

Research Article

Identification and characterization of homeobox genes in *Eucalyptus*

Graça Celeste Gomes Rocha[#,1], Régis Lopes Corrêa[#,2], Anna Cristina Neves Borges[1], Claudio Bustamante Pereira de Sá[3] and Márcio Alves-Ferreira[1]

[1]*Universidade Federal do Rio de Janeiro, Instituto de Biologia, Departamento de Genética, Laboratório de Genética Molecular Vegetal, Rio de Janeiro, RJ, Brazil.*
[2]*Universidade Federal do Rio de Janeiro, Instituto de Microbiologia Prof. Paulo de Góes, Departamento de Virologia, Laboratório de Virologia Molecular Vegetal, Rio de Janeiro, RJ, Brazil.*
[3]*Instituto Brasileiro de Geografia e Estatística, Diretoria de Pesquisas, Coordenação de Estatísticas Agropecuárias, Rio de Janeiro, RJ, Brazil.*

Abstract

Homeobox genes encode transcriptional factors, usually involved in molecular control of plant developmental patterns. They can be divided into several classes according to conserved sequences within the homeobox region and the presence of specific additional sequences. Based on these conserved sequences, we developed a search procedure to identify possible homeobox genes in the *Eucalyptus Genome Sequencing Project Consortium* (FORESTs) database. We were able to identify 50 *Eucalyptus* sequences (EST-contigs) containing the homeodomain sequence. Phylogenetic analysis was applied to this ESTs-contigs and 44 of them were found to have similarities with one of three well-known homeobox classes: Bell, Knox and HD-Zip, and their sub-classes. However, no EST-contig grouped with the fourth important homeobox class, the *PHD*-finger homeobox. On the other hand, two sequences have showed pronounced similarity to the *Arabidopsis thaliana Wuschel* gene, considered an "atypical" homeobox gene. Hierarchical clustering analysis of the expression pattern of these putative *Eucalyptus* homeobox genes revealed the presence of ten distinct expression groups. Combining phylogenetic analysis and expression patterns for some of the *Eucalyptus* genes revealed interesting aspects about some of the potential homeobox genes, which might lead to a better understanding of the *Eucalyptus* biology and to biotechnological applications.

Key words: plant development, abiotic stress, *Eucalyptus*, FORESTs, phylogenetic analysis.

Received: May 28, 2004; Accepted: March 17, 2005.

Introduction

The homeobox is a semi-conserved sequence motif of about 180 base pairs, which was first found in morphogenesis controlling genes of *Drosophila melanogaster*. Homeobox sequences encode 60 amino acid sequences, collectively referred as the homeodomain, which is highly conserved among animal, fungal, and plant proteins. The homeodomain sequences fold into a characteristic DNA-binding structure, composed of three α-helices separated by a loop and a turn (Gehring, 1987; Laughon, 1991; Scott *et al*., 1989). Homeodomain-containing proteins are basically transcription factors that control many developmental processes and other cellular mechanisms, such as positional information, spatial patterning, cell fate determination and cell differentiation (Lawrence and Morata, 1994).

In plants, homeobox genes are involved in a variety of functions, including developmental programs and response to stress. They are usually divided into four main classes, namely Knox, Bell, PHD-finger and HD-Zip, based on some conserved features inside the homeodomain sequence, as well as on the presence of additional sequences outside the domain (Kerstetter *et al*., 1994). Members of each class show characteristic structural and functional properties and can be further grouped into sub-classes or families (Sessa *et al*., 1998).

Most of the members of the *Knox* (*Knotted1*-like) class are associated with the maintenance and growth of the shoot meristems (Bowman and Eshed, 2000). *Knox* genes harbor a second conserved region just upstream of the homeodomain, the ELK domain, the basis on which they may be classified in two families: *Knox I* and *Knox II* genes (Bharathan *et al*., 1999). *Knox I*, such as *KNOTTED-1*

Send correspondence to Marcio Alves-Ferreira. Universidade Federal do Rio de Janeiro, Ilha do Fundão, Departamento de Genética, Laboratório de Genética Molecular Vegetal, Sala A2 76, 21944-970 Rio de Janeiro, RJ, Brazil. E-mail: alvesfer@biologia.ufrj.br.
#Both these authors contributed equally to the article.

(*kn1*) and *SHOOTMERISTEMLESS* (*STM),* are preferentially expressed in shoot and floral meristems (Reiser *et al*., 2000), while *Knox II* genes seem to have much wider expression domains.

In the Bell class, the homeodomain is located within the C-terminal third of the protein (Reiser *et al*., 1995). They usually show regions rich in serine/threonine and proline outside the homeodomain (Quaedvlieg *et al*., 1995), which may function as transcription-activating domains. The *Bell1* and *ATH1* are examples of Bell 1 class genes. They are required for integument development and signal transduction during photomorphogenesis, respectively (Quaedvlieg *et al*., 1995).

Proteins encoded by both *Knox* and *Bell* genes present a three-amino acid extension in the loop connecting the first and second helices of the homeodomain (Bürglin, 1997). They are thus classified in the TALE (Three Amino acid Loop Extension) super-class (Chan *et al*., 1998; Bellaoui *et al*., 2001; Becker *et al*., 2002).

The *Plant Homeodomain* (*PHD*) finger genes encode proteins containing a Cys4-His-Cys3 motif, which is thought to bind two zinc ions, even though the space between the cysteine/histidine residues may be too variable (Aasland *et al*., 1995). They are frequently associated with chromatin-mediated transcriptional regulation, as in the case of the *Arabidopsis* Pathogenesis-related homeodomain protein (*PRHA*), which regulates elicitor-mediated expression of the parsley pathogenesis-related protein 2 gene (*PR2*) (Korfhage *et al*., 1994).

The *Homeodomain-leucine zipper* (*HD-Zip*) gene class encodes to proteins in which the homeodomain is closely associated with a leucine zipper (Schena and Davis, 1992, 1994). They are separated into four different families (HD-Zip I, II, III and IV) depending on specific differences of the leucine zipper motif (Chan *et al*., 1998; Aso *et al*., 1999; Ingram *et al*., 2000; Sakakibara *et al*., 2001; Ageez *et al*., 2003). This kind of homeodomain organization is present only in plants and it is speculated that *HD-Zip* genes originated in plant lineage by exon exchange between a homeodomain gene and a leucine zipper containing sequence (Schena and Davis, 1992). Proteins belonging to HD-Zip I and II classes are very similar with respect to the architecture of the HD-Zip domain (Sessa *et al*., 1993). On the other hand, *HD-Zip III* proteins present an insertion of four amino acids both between helix 2 and 3 of the homeodomain and between helix 3 and the leucine zipper domain (Baima *et al*., 1995; Sessa *et al*., 1998), changing the spacing between the homeodomain and leucine zipper. The HD-Zip IV family has a leucine zipper motif separated in two sub-domains by a loop of 10 amino acid residues. Many examples of *HD-Zip* class genes are found in the well-known model plant *Arabidopsis thaliana*, performing a wide array of functions. The *ATHB-1* (*HD-Zip I*) gene is associated with leaf development (Aoyama *et al*., 1995); the ATHB-2 (HD-Zip II) protein probably has a role in mediating cell elongation (Carabelli *et al*., 1996); the ATHB-8 (HD-Zip III) protein acts in the differentiation of the vascular system (Baima *et al*., 1995) and the ATHB-10 (HD-zip IV) regulates trichome development and suppresses root hair formation (Di Cristina *et al.,* 1996).

Finally, there are a number of genes that have been recognized as having the homeodomain motif which do not fit in any of these main classes. Examples include the atypical homeobox WUSCHEL protein from *A. thaliana* which is essential for shoot apical meristem formation and maintenance (Mayer *et al*., 1998).

In this work we searched for potential homeobox genes in the *Eucalyptus Genome Sequencing Project Consortium* (FORESTs) database and used phylogenetic methods to classify those among these well-established groups. We were able to identify 50 high quality assemblies of homeobox genes after an extensive search on the FOREST database. To better understand homeobox gene expression at plant system level and to identify differentially expressed and tissue-specific genes, we conducted a digital expression analysis. By clustering genes according to their relative abundance in the various EST libraries, expression patterns of genes across various tissues were generated and genes with similar patterns were grouped. This information can be useful in devising strategies for the biotechnological improvement of *Eucalyptus* trees for industrial use and may assist in the study of homeobox in other plant species.

Material and Methods

Databases and procedures for searching *Eucalyptus* homeobox sequences

The primary data used in this work were the so-called EST-contigs sequences, from the FORESTs project database. In the FORESTs project, they have been assembled from approximately 130,000 ESTs obtained in the sequencing of 19 *Eucalyptus* cDNA libraries, corresponding to different tissues and physiological states (Detailed information on cDNA libraries, sequencing, clustering and other features of the FORESTs project may be found in https://forests.esalq.usp.br/). Besides *Eucalyptus* EST-contigs, *Arabidopsis thaliana* homeobox sequences, obtained from The National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) and from The Institute for Genomic Resources (TIGR, http://www.tigr.org/tdb/e2k1/ath1/), were used for comparison.

In any single FORESTs database search, just one target amino acid sequence was compared to all EST-contigs nucleotide sequences in all six frames. This was done with the tBLASTn algorithm (Altschul *et al.*, 1997) implemented in the FORESTs system. In the first search, the target amino-acid sequence was a consensus homeodomain sequence (qrcqyvvgreREELAKQLNLTERQVK VWFQNRRAKxKKdqsrdlekra) generated by the COBBLER program (Henikoff and Henikoff, 1997). The

searches were conducted using the TBLASTN algorithm (Altschul *et al.*, 1997) with the BLOSUM60 scoring matrix. Sequences with significant similarity to the COBBLER consensus (E-value of 10^{-5}) were retrieved from the database and were screened for the presence of homeodomain using the InterProScan (Apweiler *et al*., 2001) and PRODOM (Servant *et al*., 2002) programs. Sequences that met these criteria were then used as target amino-acid sequences in new searches against the FORESTs database to locate further potential homeobox sequences. In this new round full EST-contigs were the target sequences, and the searches were conducted using the TBLASTN algorithm with the BLOSUM80 scoring matrix. Newly found EST-contigs harboring significant similarity (again E-value of 10^{-5}) to any of the previously found were also investigated using InterProScan and PRODOM programs. This procedure was repeated until no new contigs were found. Finally, sequences from each selected EST-contig were used for BLAST search against Genebank in order to verify its closest sequence deposited in the world data bank.

Phylogenetic and expression analyses

To classify the selected EST-contigs, we first performed a multiple alignment of their homeodomain region using the CLUSTALW program (Higgins *et al*., 1994). Phylogenetic analysis was conducted using the MEGA 2.1 software (Kumar *et al*., 2001). Among the various options available, we chose the Neighbor-joining method (Saitou and Nei, 1987), derived from a p-distance matrix, and the pair-wise deletion option was adopted, excluding amino acid gaps from the sequence alignment.

The expression analysis used the so-called 'digital northern blot' (Carraro *et al*., 2001; Lambais, 2001) to represent the expression profile of the potential homeobox genes and to identify differential expression among the diverse *Eucalyptus* tissues/treatments used in the FORESTs project. In this approach, the frequency of reads for each EST-contig in each selected library is computed and then normalized with respect to the number of reads in the library and to the total number of reads in all libraries. The obtained values form a matrix relating contigs and libraries. We then used the Cluster and Tree View programs (Eisen *et al.*, 1998) to determine similarity in gene expression patterns among EST-contigs and among libraries. In both cases, aggregation was made by hierarchical clustering, based on Spearman Rank correlation matrix, and substituting formed clusters by their average pattern. Digital blot matrix was ordered, according to similarities in the patterns of gene expression and displayed as an array, where the normalized number of reads for each EST-contig in each specific library is represented in gray scale.

Results and Discussion

FORESTs database searches

In order to search for homeobox genes in the *Eucalyptus* transcriptome, a consensus homeodomain sequence was generated by the COBBLER program and used to screen the FORESTs dataset. The first round of searches in the FORESTs database, using the COBBLER consensus sequence, selected a total of 2336 putative homeobox EST-contigs, but only forty-three were found to have the homeodomain in the inspection using InterProScan and PRODOM programs.

Each of these forty-three ESTs-contigs were used to perform the second round of searches in the FORESTs database in which we found seven additional putative *Eucalyptus* homeobox sequences. An additional search was done using the conserved sequences outside of the homeodomain of the families HD-Zip, Knox, Bell and PHD-finger but no additional EST-contig was found. A total of 50 EST-contigs were found to contain the homeodomain region in the FORESTs database and, considering the extensive searches and distinct methodologies used, we judge that this should be very close to the actual number of homeobox sequences present in the FORESTs database.

Phylogenetics analysis

To be able to identify the phylogenetic groups based on the homeodomain regions of the 50 sequences from the FORESTs database, we aligned and clustered them together with nine *A. thaliana* homeobox genes, whose genetic, biochemical and taxonomic data have been extensively studied in the literature. Those insertions did not change the structure of the phylogenetic tree, since an identical tree can be obtained from the selected EST-contigs alone (data not shown). The alignment of these sequences showed the existence of the highly conserved amino acid residues, including the absolutely conserved Trp-49 and a conserved secondary structure consisting of a helix-loop-helix-turn-helix motif (Figure 1).

The phylogenetic tree shows eight statistically well-supported groups, which were named as Bell, Knox I, Knox II, HD-Zip I, HD-Zip II, HD-Zip III, HD-Zip IV and Wuschel-like, in agreement with sequence features and the presence of the *A. thaliana* "class marker" genes in each of them (Figure 2). From the 50 EST-contigs encoding to homeobox proteins present in the FOREST databank, 58% of the EST-contigs belong to the HD-Zip class, 16% to the Bell class, 14% to the Knox class and 4% has shown similarity to the unconventional homeobox gene WUSCHEL. Interestingly, our search and analysis did not identify any EST-contig of the PHD-finger family in the FORESTs database. This result is surprising since members of PHD-finger family have been found in *Arabidopsis* and rice ESTs libraries. To further investigate this issue we searched the FOREST databank for EST-contigs contain-

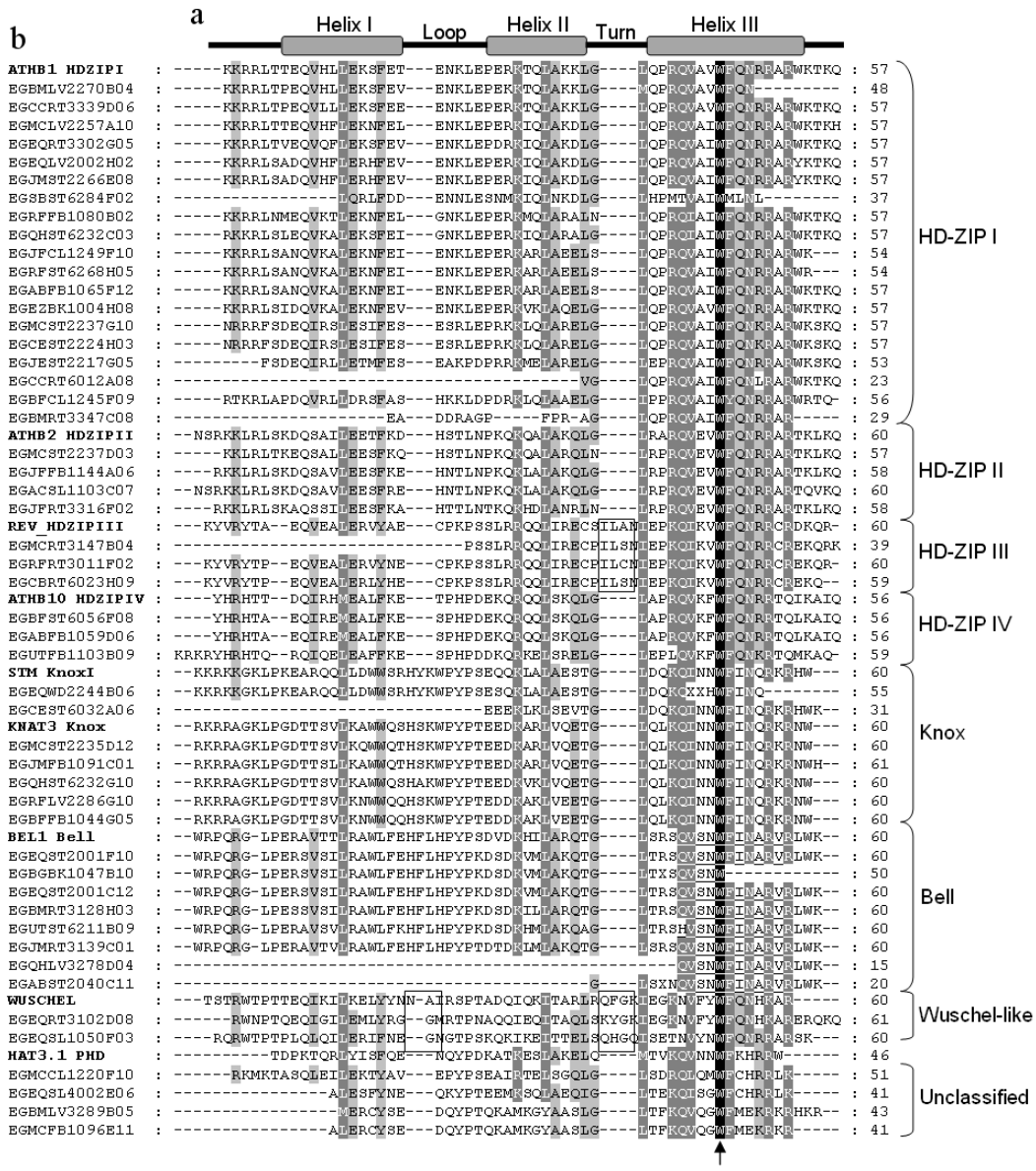

Figure 1 - Multiple sequence alignment of 50 FORESTs and nine *Arabidopsis* homeodomain sequences. (a) Schematic diagram illustrating a typical homeobox protein. (b) Black boxes indicate residues that are present in all sequences, while alignments present in more than 60% but less than 100% of the sequences are shaded in gray. Shading was automatically defined, using the GeneDoc program (http://www.psc.edu/biomed/genedoc/). Residues representing the absolutely conserved Trp-49 are indicated (arrow). The conserved sequence "QVSNWFINARVR" in the C-terminal region of the Bell class homeodomain is underlined. Open boxes indicate extra amino acid residues present in HD-ZIP III and Wuschel-like proteins. The names of *Arabidopsis* homeodomain sequences are in bold letters.

ing a PHD-finger motif. We found 2336 EST-contigs with significant similarity to the PHD-finger COBBLER consensus (YCSVCGKPDDGGELLQCDGCDRWYHQTCL GPPLLIEEPDGKWYCPKCK), but none of these EST-contigs encoded for homeobox proteins. The misrepresentation of *PHD-finger* homeobox genes may be due to their atypical sequence features. These genes usually have extremely long transcripts that might impair reverse transcription and/or subsequent cDNA cloning (Alba *et al*., 2004).

Four of the *Eucalyptus* EST-contigs (EGMCC11220 F10, EGEQSL4002E06, EGMCF81096E11 and EGBM LV3289B05) did not cluster to any of these well-defined groups (Figure 2). Blast search against Genebank revealed that the homeodomain region from ungrouped sequences have similarities with other uncharacterized sequences from *A. thaliana* (CAB77810 is the best hit for EGEQSL 4002E06 and EGBMLV3289B05; NP_199231 for EGMC CL1220F10 and AAF16763 for EGMCFB1096E11). To further typify these sequences, a phylogenetic analysis was performed with 97 homeobox sequences from *Arabidopsis*

and sequences with significant similarities found in genebank blast search. Our results confirmed that contigs EGMCCL1220F10.g, EGEQSL4002E06.g, EGMCFB 1096E11.g and EGBMLV3289B05.g, together with at least 15 sequences from *Arabidopsis*, present a high divergent homeodomain and, thus, were not grouped into one of the traditional plant homeobox families (data not shown).

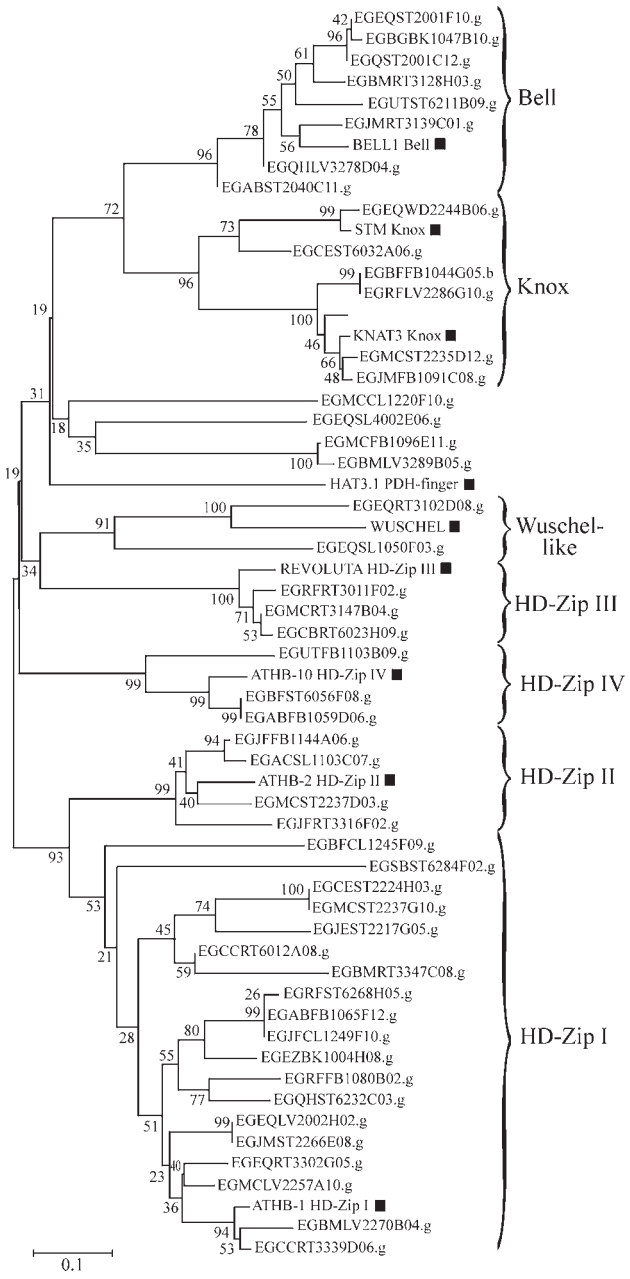

Figure 2 - Phylogenetic unrooted tree for the fifty FORESTs contigs and nine *Arabidopsis* genes, based on the amino acid sequences corresponding to homeodomain region, aligned with ClustalW. In phylogenetic tree construction, we chose the Neighbor-joining, p-distance and pair-wise deletion options, and ran 1000 bootstrap replications. Seven major homeobox groups are identified in the tree. *Arabidopsis* homeobox proteins are indicated by solid black boxes. *Arabidopsis* sequences were obtained from the GenBank database under the following accession numbers: *Arabidopsis thaliana* homeodomain protein (BEL1), AT5G41410; SHOOT MERISTEMLESS (STM), AT1G62360; homeobox protein knotted-1 like 3 (KNAT3), AT5G25220; homeobox protein HAT3.1 (HAT3.1), AT3G19510; WUSCHEL (WUS), AT2G17950; Revoluta (REV), AT5G60690; homeobox-leucine zipper protein 10 (ATHB-10), AT1G79840; homeodomain-leucine zipper protein 2 (ATHB-2), AJ431183; homeobox-leucine zipper protein 1 (ATHB-1), AT3G01470.

Almost all *Eucalyptus* EST-contigs in the Bell class showed the conserved sequence "QVSNWFINARVR" in the C-terminal region of the homeodomain (Figure 1). This conserved sequence resulted in good statistical support of this group. Knox sequences also showed a high level of similarity, especially in helix 3 and in the N-terminal end, and were also well supported as a monophyletic group. The two classes of Knox proteins were also easy to distinguish. There is strong support for the Knox II group, even though the evidence for Knox I group is weaker (Figure 2). The difference in the tightness of associations between the two Knox subfamilies was noted before in other plants (Bharathan *et al*., 1999).

Knox and Bell homeobox proteins may be grouped into the TALE superclass. This class is characterized by three extra amino acids between homeodomain helix 1 and helix 2 (Chan *et al*., 1998; Bellaoui *et al*., 2001; Becker *et al*., 2002). However this characteristic three amino acid insertion was not found in two *Eucalyptus* EST-contigs classified as *Bell* (EGQHLV3278D04 and EGABST2040C11) and in one classified as Knox (EGCEST6032A06), probably due to incomplete sequencing (Figure 1). Although these two homeodomain families harbor specific structural similarities, our phylogenetics reconstruction revealed only weak statistical support for the TALE super class (Figure 2). This uncertain evolutionary relationship between TALE members has already been observed in the literature (Bürglin, 1997; Becker *et al*., 2002).

Our analysis identified twenty-nine potential *HD-Zip* genes from the FORESTs EST-contigs, among which nineteen are *HD-Zip I,* four *HD-Zip II*, three *HD-Zip III* and three *HD-Zip IV* (Figure 2). *HD-Zip I* genes were the least supported clade (53%). Analysis of the conserved amino acids among *HD-Zip* genes (Sakakibara *et al*., 2001) showed that the HD-Zip I family had an evolutionary rate higher than the others after the split of the monocot and eudicot plants. It has been postulated that this higher evolutionary rate in *HD-Zip I* genes is due to the absence of strict interactions with cofactors. Because of that, phylogenetic trees often do not assign the HD-Zip I family to a monophyletic group (Sakakibara *et al*., 2001). On the other hand, the HD-Zip II, HD-Zip III and HD-Zip IV families were highly supported as monophyletic groups (99%, 100% and 99%, respectively). The high similarity of *HD-Zip II* genes corroborates with the hypothesis that they form a specialized and recent group (Chan *et al*., 1998). All contig sequences grouped with the HD-Zip III protein REVOLUTA, from *A. thaliana*, showed the characteristic four amino acid insertion between helix 2 and helix 3 (Figure 1).

Two *Eucalyptus* sequences formed a monophyletic group with the *Arabidopsis WUSCHEL* gene (91%), and

were thus called Wuschel-like (Figure 2). WUSCHEL have an "atypical" homeodomain sequence that is present only in plants. These "atypical" homeodomain sequences present a four amino acid insertion in the turn of the homeodomain, like HD-Zip III proteins. However, the Wuschel-like sequences has also an additional insertion of one or two amino acids in the loop region of the homeodomain (Figure 1) (Kamiya *et al*., 2003). The two FORESTs Wuschel-like sequences have structural similarities with this "atypical" group of homeobox genes and may function in analogous biochemical ways (Figure 1).

Expression pattern

It is already recognized that analysis of gene expression can be performed *in silico*, based on the frequency of sequence tags in cDNA libraries, allowing comparisons of the expression profiles of specific genes in plant tissues (Ewing *et al.*, 1999). This digital expression analysis method has advantages over the conventional microarray approaches in that the liability of the later can be reduced due to cross-hybridization of closely related sequences (Kuo *et al*., 2002; Lipshutz *et al*., 1999) and to the development of stable probe secondary structures (Southern *et al*., 1999*).*

The Hierarchical Clustering analysis performed over this obtained expression pattern revealed the presence of ten distinct gene expression groups. For each of these groups we assigned the name of the library for which their components are overrepresented in terms of the normalized number of ESTs (Figure 3). The groups were: dark formed *calli* (group I); dark growing seedlings with three hours of light exposition (group II); roots (group III); leaves damaged by *Thyrinteina* (group IV); stem from drought stress susceptible seedlings (groups V and VI); leaves of Phosphate/Boron deficient plants susceptible to canker and rust (group VII); dark growing seedlings (group VIII); flower buds, flowers and fruits (group IX); and wood (group X). Two of the EST-contigs (EGBGBK1047B10 and EGCCRT6012A08) were not grouped in any of these (Figure 3). Considering the normalized number of reads, we can say that the *Eucalyptus* homeobox genes are preferentially expressed in six libraries: Dark growing seedlings *E. urophylla* (SL6, 7.93% of total number of reads); stem from frost resistant and susceptible plants (ST7, 7.08%); stem from drought stress susceptible seedlings (ST2, 7%); leaves damaged by *Thyrinteina* for seven days (LV3, 6.67%); stem from drought stress susceptible seedlings (ST6, 6.46%); and flower buds, flowers and fruits (FB1 - 6.29%). The result observed for the libraries ST2, ST6 and FB1 is mainly due to the higher number of different EST-contigs, 16, 19 and 19 respectively. On the other hand, the libraries SL6, ST7 and LV3 have few EST-contigs, 3, 6 and 9 respectively, but they present high expression levels. The high number of reads and EST-contigs associated with abiotic stress (drought stress, frost resistance and susceptibility) suggests that *Eucalyptus* homeobox genes may be important in stress adaptation.

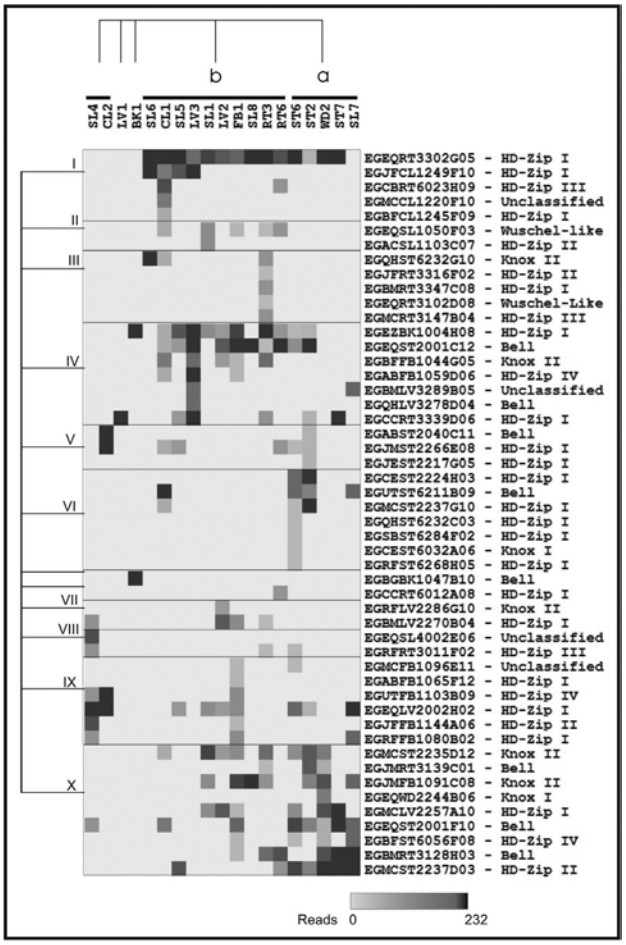

Figure 3 - Digital northern blot representing the expression profile of the potential *Eucalyptus* homeobox genes. The normalized number of reads for each EST-contig in each specific library is represented in gray scale. The different EST-contigs are represented by the rows, while the libraries are represented by the columns. Hierarchical clustering method was used to group homeobox genes with similar expression patterns, resulting in ten expression groups. EST-contig collections were also clustered (bars and the letters **a** and **b** above the draw). The expression groups are indicated by roman numerals to the left of the array: I - *E. grandis* dark formed calli (**CL1**); II - *E. grandis* dark growing seedlings with three hours of light exposition (**SL1**); III- roots from developing plants (**RT3**); IV- leaves damaged by *Thyrinteina* for seven days (**LV3**); V- stem from drought stress susceptible seedlings prepared with 0.6 to 2.0 kb DNA fragments (**ST2**); VI- stem from drought stress susceptible seedlings prepared with 0.8 to 3.0 kb DNA fragments with 0.8 to 3.0 kb(**ST6**); VII- leaves of Phosphate/Boron deficient plants susceptible to canker and rust (**LV2**); VIII- *E. globulus* dark growing seedlings (**SL4**); IX- Flower buds, flowers and fruits (**FB1**); and X- *E. grandis* wood (**WD2**).

The expression profiles from tissue/treatment libraries were also clustered. EST-contig collections derived from the most similar tissues typically clustered together (Figure 3 top). This was observed for one of the cluster (a), where all three stem libraries and the wood library are included, representing wood forming tissues. Interestingly, the library of dark growing seedlings of *E. grandis* (SL7) is

also part of this cluster. The cluster (b) included almost all the rest of the libraries. The libraries (SL4, CL2, LV1 and BK1) do not have enough homeobox ESTs to sustain any correlation. The comparison of these tissue/treatments libraries did not display any consistent trend.

The gene expression groups III, V and VI (roots and plants under drought stress) showed a high number of EST-contigs whose expression are restricted to one or few libraries, indicating that these homeobox genes might be involved in distinctive aspects related to root development and drought stress adaptation. On the other hand, most of the homeobox genes belonging to group X showed a broader distribution over the libraries. This indicates that homeobox genes of the group X might be involved with more general aspects of plant physiology/development like, for instance, vascular development in different tissues.

HD-Zip genes were the type of homeobox genes most commonly found in the FORESTs database. These homeobox genes are involved in a wide range of process in plants and are expressed in different organs and developmental stages. Their expression patterns are modified by processes causing developmental responses like light, hormones, stress and wounding (Schena *et al*., 1993; Aoyama *et al*., 1995; Baima *et al*., 1995; Carabelli *et al*., 1996; Söderman *et al*., 1996; Aso *et al*., 1999; Ingram *et al*., 2000; Sakakibara *et al*., 2001; Ageez *et al*., 2003). *Eucalyptus HD-Zip I* genes were also present in almost all gene expression groups (Figure 3), but the greatest number of EST-contigs (seven out of 19) were observed in expression groups V and VI (stem from drought stress susceptible seedlings). Three *HD-Zip I* genes from *Arabidopsis* (*ATHB-6, ATHB-7 and ATHB-12)* have their expression induced during drought stress (Söderman *et al*., 1996; Lee and Chun, 1998), and this same expression pattern was observed for the potential *Eucalyptus* homologous of *ATHB-7* (EGQHST6232C03) and *ATHB-12* (EGMCST2237G10 and EGCEST2224H03). The association between drought stress and the expression of those genes provides initial insights about their possible function.

The atypical homeobox *WUSCHEL* gene is essential for shoot apical meristem formation and maintenance *in A. thaliana* (Mayer *et al*., 1998). Recently it was shown that a *WUSCHEL*-like gene is also involved in rice root apical meristem formation (Kamiya *et al*., 2003). This contributes to the hypothesis that maintenance of shoot and root apical meristems are regulated by similar processes. The presence of an EST-contig homologous to *WUSCHEL* in the *Eucalyptus* roots indicates that this hypothesis may also be true for tree species (EGEQRT3102D08 - group III, Figure 3). The other *Eucalyptus WUSCHEL*-like gene has broader expression domain including floral buds, roots and callus. Interestingly, this EST-contig has a higher degree of similarity with the *Populus tremula* HB2 protein, a protein involved in wood differentiation. *Populus* HB2 protein is expressed in cambial cells, in xylem and phloem cells undergoing radial enlargement (Hertzberg and Olsson, 1998).

Concluding Remarks

The FORESTs database was searched for homeobox genes and 50 *Eucalyptus* EST-contigs were identified as sharing significant sequence similarity with the homeobox domain. Sequence alignment and phylogenetic studies provided sound basis for classifying 44 of them (88%) into one of three important homeobox groups: Bell, Knox or HD-Zip. Two sequences were found to have similarity with atypical homeobox genes, like the *A. thaliana WUSCHEL* gene. Based on a digital expression analysis we found ten subsets from the 50 EST-contigs with similar expression patterns. Sequence comparison revealed some *Eucalyptus* genes closely related to *A. thaliana* genes involved in the control of developmental programs, as well as in stress response. Our digital expression analysis facilitated these cross-species comparisons. For instance, we found possible homologous to *ATHB-7* and *ATHB-12* that were preferentially expressed in stems of plants under drought stress. Thus, such genes may act as mediators of the plant growth response to limiting water conditions in leaf, stem and other plant organs, like their potential homologous genes do in *A. thaliana*. We also identified numerous homeobox genes associated with biotic stress, wood development, and pathogen responses, which may be valuable in future efforts to elucidate the genetic mechanisms underlying these processes in trees.

Application of biotechnology for tree improvement offers great potential, and transcription factors, such as those encoded by homeobox genes, may play important roles in such improvements. Homeobox genes control multiple biochemical pathways and cellular processes, and several researchers have recently reported progress in identifying homeobox genes in various plant species, including trees (Hertzberg and Olsson, 1998). For instance, wood is made of secondary xylem and many of its features are determined by xylem growth and cell wall biosynthesis in its cells. Identification of genes involved in xylem formation and/or cell wall biosynthesis would lead to studies on their function that can provide mechanisms to manipulate wood features of interest. The same holds for genes involved in stress resistance, growth, reproductive development and other traits important for tree breeding. Thus, the data obtained in the FORESTs project can be very helpful in developing strategies for the biotechnological improvement of *Eucalyptus* trees. Understanding the molecular mechanisms underlying *Eucalyptus* growth, development and stress reactions can provide important insights into tree development and reveal the means by which tree characteristics could be modified for the improvement of their industrial properties.

Acknowledgments

G.C.G. Rocha is the recipient of a doctoral fellowship from the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). M. Alves-Ferreira is the recipient of a research fellowship from CNPq (307219/2004-6). This work was supported by grants from CNPq CNPq/Centro Brasileiro-Argentino de Biotecnologia (400767/2004-0), CNPq (475666/2004-6) and FORESTs consortium.

References

Aasland R, Gibson TJ and Stewart AF (1995) The PHD-finger: Implications for chromatin-mediated transcriptional regulation. Trends Biotechnol 20:56-59.

Ageez A, Matsunaga S, Uchida W, Sugiyama R, Kazama Y and Kawano S (2003) Isolation and characterization of two homeodomain leucine zipper genes from the dioecious plant *Silene latifolia.* Genes Genet Syst 78:353-3561.

Alba R, Fei Z, Payton P, Liu Y, Moore SL, Debbie P, Cohn J, D'Ascenzo M, Gordon JS, Rose JK, Martin G, Tanksley SD, Bouzayen M, Jahn MM and Giovannoni J (2004) ESTs, cDNA microarrays, and gene expression profiling: Tools for dissecting plant physiology and development. Plant J 39:697-714.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Aoyama T, Dong CH, Wu Y, Carabelli M, Sessa G, Ruberti I, Morelli G and Chua NH (1995) Ectopic expression of the *Arabidopsis* transcriptional activator *Athb-1* alters leaf cell fate in tobacco. Plant Cell 7:1773-1785.

Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA and Zdobnov EM (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic Acids Res 29:37-40.

Aso K, Kato M, Banks JA and Hasebe M (1999) Characterization of homeodomain-leucine zipper genes in the fern Ceratopteris richardii and the evolution of the homeodomain-leucine zipper gene family in vascular plants. Mol Biol Evol 16:544-552

Baima S, Nobili F, Sessa G, Lucchetti S, Ruberti I and Morelli G (1995) The expression of the *ATHB-8* homeobox gene is restricted to provascular cells in *Arabidopsis thaliana*. Development 121:4171-4182

Becker A, Bey M, Bürglin TR, Saedler H and Theissen G (2002) Ancestry and diversity of *BEL1*-like homeobox genes revealed by gymnosperm (Gnetum gnemon) homologs. Dev Genes Evol 212:452-457.

Bellaoui M, Pidkowich MS, Samach A, Kushalappa K, Kohalmi SE, Modrusan Z, Crosby WL and Haughn GW (2001) The *Arabidopsis* BELL1 and KNOX TALE homeodomain proteins interact through a domain conserved between plants and animals. Plant Cell 13:2455-2470.

Bharathan G, Janssen B, Kellogg E and Sinha N (1999) Phylogenetic relationships and evolution of the KNOTTED class of plant homeodomain proteins. Mol Biol Evol 16:553-563.

Bowman JL and Eshed Y (2000) Formation and maintenance of the shoot apical meristem. Trends Plant Sci 5:110-115.

Bürglin TR (1997) Analysis of TALE superclass homeobox genes (*MEIS, PBC, KNOX, IROQUOIS, TGIF*) reveals a novel domain conserved between plants and animals. Nucl Acids Res 25:4173-4180.

Carabelli M, Morelli G, Whitelam G and Ruberti I (1996) Twilight-zone and canopy shade induction of the ATHB-2 homeobox gene in green plants. Proc Natl Acad Sci USA 93:3530-3535.

Carraro DM, Lambais MR and Carrer H (2001) *In silico* characterization and expression analyses of sugarcane putative sucrose non-fermenting-1 (SNF1) related kinases. Genet Mol Biol 24:35-41

Chan RL, Gago GM, Palena CM and Gonzalez DH (1998) Homeoboxes in plant development. Biochim Biophys Acta 1442:1-19.

Di Cristina M, Sessa G, Dolan L, Linstead P, Baima S, Rubert I and Morelli G (1996) The *Arabidopsis* ATHB-10 (GLABRA2) is an HD-Zip protein required for regulation of root hair development. Plant J 10:393-402.

Eisen MB, Spellman PT, Brown PO and Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 95:14863-8.

Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S and Claverie J-M (1999) Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 9:950-959.

Gehring WJ (1987) Homeoboxes in the study of development. Science 236:1245-1252.

Henikoff S and Henikoff JG (1997) Embedding strategies for effective use of information from multiple sequence alignments. Protein Sci 6:698-705.

Hertzberg M and Olsson O (1998) Molecular characterization of a novel plant homeobox gene expressed in the maturing xylem zone of *Populus tremula x tremuloides.* Plant J 16:285-295.

Higgins D, Thompson J, Gibson T, Thompson JD, Higgins DG and 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.

Ingram GC, Boisnard-Lorig C, Dumas C and Rogowsky PM (2000) Expression patterns of genes encoding HD-Zip IV homeo domain proteins define specific domains in maize embryos and meristems. Plant J 22:401-414.

Kamiya N, Nagasaki H, Morikami A, Sato Y and Matsuoka M (2003) Isolation and characterization of rice *WUSCHEL*-type homeobox gene that is specifically expressed in the central cells of a quiescent center in the root apical meristem. Plant J 35:429-441.

Kerstetter R, Vollbrecht E, Lowe B, Veit B, Yamaguchi J and Hake S (1994) Sequence analysis and expression patterns divide the maize *knotted1*-like homeobox genes into two classes. Plant Cell 6:1877-1887.

Korfhage U, Trezzini GF, Meier I, Hahlbrock K and Somssich IE (1994) Plant homeodomain protein involved in trans-

criptional regulation of a pathogen defense-related gene. Plant Cell 6:695-708.

Kumar S, Tamura K, Jakobsen IB and Nei M (2001) MEGA2: Molecular evolutionary genetics analysis software. Bioinformatics 17:1244-5.

Kuo WP, Jenssen TK, Butte AJ, Ohno-Machado L and Kohane IS (2002) Analysis of matched mRNA measurements from two different microarray technologies. Bioinformatics 18:405-12.

Lambais MR (2001) *In silico* differential display of defense-related expressed sequence tags from sugarcane tissues infected with diazotrophic endophytes. Genet Mol Biol 24:103-111.

Laughon A (1991) DNA binding specificity of homeodomains. Biochemistry 30:11357-11367.

Lawrence PA and Morata G (1994) Homeobox genes: Their function in *Drosophila* segmentation and pattern formation. Cell 78:181-189.

Lee YH and Chun JY (1998) A new homeodomain-leucine zipper gene from *Arabidopsis thaliana* induced by water stress and abscisic acid treatment. Plant Mol Biol 37:377-384.

Lipshutz RJ, Fodor SP, Gingeras TR and Lockhart DJ (1999) High density synthetic oligonucleotide arrays. Nat Genet 21:20-4.

Mayer KF, Schoof H, Haecker A, Lenhard M, Jürgens G and Laux T (1998) Role of WUSCHEL in regulating stem cell fate in the *Arabidopsis* shoot meristem. Cell 95:805-815.

Quaedvlieg N, Dockx J, Rook F, Weisbeek P and Smeekens S (1995) The homeobox gene *ATH1* of *Arabidopsis* is derepressed in the photomorphogenic mutants *cop1* and *Det1*. Plant Cell 7:117-129.

Reiser L, Sanchez-Baracaldo P and Hake S (2000) Knots in the family tree: Evolutionary relationships and functions of *knox* homeobox genes. Plant Mol Biol 42:151-166.

Reiser L, Modrusan Z, Margossian L, Samach A, Ohad N, Haughn GW and Fischer RL (1995) The *BELL1* gene encodes a homeodomain protein involved in pattern formation in the *Arabidopsis* ovule primordium. Cell 83:735-742.

Saitou N and Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425.

Sakakibara K, Nishiyama T, Kato M and Hasebe M (2001) Isolation of homeodomain-leucine zipper genes from the moss *Physcomitrella patens* and the evolution of homeodomain-leucine zipper genes in land plants. Mol Biol Evol 18:491-502.

Schena M and Davis RW (1994) Structure of homeobox-leucine zipper genes suggest a model for the evolution of gene families. Proc Natl Acad Sci USA 91:8393-8397.

Schena M, Lloyd AM and Davis RW (1993) The *HAT4* gene of *Arabidopsis* encodes a developmental regulator. Genes Dev 7:367-79.

Schena M and Davis RW (1992) HD-Zip proteins: Members of an *Arabidopsis* homeodomain protein superfamily. Proc Natl Acad Sci USA 89:3894-3898.

Scott MP, Tamkun JW and Hartzell GW (1989) The structure and function of the homeodomain. Biochim Biophys Acta 989:25-48.

Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, Peyruc D and Kahn D (2002) ProDom: Automated clustering of homologous domains. Brief Bioinform 3:246-251.

Sessa G, Steindler C, Morelli G and Ruberti I (1998) The *Arabidopsis ATHB-8*, *-9* and *-14* genes are members of a small gene family coding for highly related HD-ZIP proteins. Plant Mol Biol 38:609-622.

Sessa G, Morelli G and Ruberti I (1993) The ATHB-1 and -2 HD-Zip domains homodimerize forming complexes of different DNA binding specificities. EMBO J 12:3507-3517.

Söderman E, Mattsson J and Engström P (1996) The *Arabidopsis* homeobox gene *ATHB-7* is induced by water deficit and by abscisic acid. Plant J 10:375-381

Southern E, Mir K and Shchepinov M (1999) Molecular interactions on microarrays. Nat Genet 21:5-9.

Associate Editor: Carlos F. M. Menck