# Evolution of a domain conserved in microtubule-associated proteins of eukaryotes

Alex S Rajangam[1]
Hongqian Yang[2]
Tuula T Teeri[1]
Lars Arvestad[2]

[1]KTH Biotechnology, Swedish Center for Biomimetic Fiber Engineering, AlbaNova, Stockholm, Sweden; [2]Stockholm Bioinformatics Center and School of Computer Science and Communication, Royal Institute of Technology, AlbaNova, Stockholm, Sweden

**Abstract:** The microtubule network, the major organelle of the eukaryotic cytoskeleton, is involved in cell division and differentiation but also with many other cellular functions. In plants, microtubules seem to be involved in the ordered deposition of cellulose microfibrils by a so far unknown mechanism. Microtubule-associated proteins (MAP) typically contain various domains targeting or binding proteins with different functions to microtubules. Here we have investigated a proposed microtubule-targeting domain, TPX2, first identified in the Kinesin-like protein 2 in *Xenopus*. A TPX2 containing microtubule binding protein, *Ptt*MAP20, has been recently identified in poplar tissues undergoing xylogenesis. Furthermore, the herbicide 2,6-dichlorobenzonitrile (DCB), which is a known inhibitor of cellulose synthesis, was shown to bind specifically to *Ptt*MAP20. It is thus possible that *Ptt*MAP20 may have a role in coupling cellulose biosynthesis and the microtubular networks in poplar secondary cell walls. In order to get more insight into the occurrence, evolution and potential functions of TPX2-containing proteins we have carried out bioinformatic analysis for all genes so far found to encode TPX2 domains with special reference to poplar *Ptt*MAP20 and its putative orthologs in other plants.

**Keywords:** TPX2 domain, MAP20, evolution, microtubule, cellulose, bioinformatics

## Introduction

Similar to other eukaryotes, the cytoskeleton of plant cells consists of tubulin and actin networks which render multiple morphological functions during various phases of cellular development, growth, and movement (Vassilyev 1996; Hasek et al 2003; Mathur 2004; Wasteneys and Yang 2004). Unlike animal cells, plant cells have prominent cell walls formed by networks of cellulose microfibrils, hemicelluloses, lignin, and structural proteins. In secondary cell walls, the cellulose microfibrils are strictly aligned and form multiple layers designated S1, S2, and S3, which are characterized by different microfibril angles. Several lines of evidence suggest that interphase cortical microtubules somehow influence the ordered deposition of cellulose microfibrils (Goddard et al 1994; Roberts et al 2004; Oda et al 2005). For example, it has been shown that all three secondary cell wall associated CesA proteins colocalize with bands of cortical microtubules in older xylem vessels in *Arabidopsis* (Gardiner et al 2003). It has also been demonstrated recently that the CesA complexes in the *Arabidopsis* plasma membrane move at constant rates in linear tracts that coincide with cortical microtubules (Paradez et al 2006). Furthermore, studies of developing wood cells in both conifers and angiosperms indicate reorientation of microtubules upon changes in microfibril orientation during the formation of the successive cell layers S1–S3 (Chaffey et al 1997, 1999, 2002; Funada et al 2001).

The assembly, bundling and stability of microtubules depend on the activity of various microtubule-associated proteins (MAPs) and their regulatory kinases and phosphatases (Sedbrook 2004; Amos and Schlieper 2005). Domains are conserved sequence units that frequently determine the function of proteins and that often correspond to

Correspondence: Lars Arvestad
CSC and Stockholm Bioinformatics Center, Royal Institute of Technology, AlbaNova, SE–100 44 Stockholm, Sweden
Tel +46 8 5537 8565
Fax +46 8 5537 8214
Email arve@csc.kth.se

structural units (Elofsson and Sonnhammer 1999). Microtubule associated proteins typically contain a variety of conserved domains and motifs (Amos and Schlieper 2005). One of these is TPX2 (Pfam: PF06886), a putative microtubule-targeting domain first identified in Kinesin-like proteins in *Xenopus* (Wittmann et al 2000; Bayliss et al 2003; Brunet et al 2004). The current Pfam database (July 2007) contains about 68 proteins containing significantly conserved TPX2 domains. Among plant MAPs (Sedbrook 2004), WVD2 and WVL1 in Arabidopsis contain a TPX2 domain (Korolev et al 2007; Perrin et al 2007). We have recently identified a new MAP, denoted *Ptt*MAP20, which exhibits particularly high level of expression in the wood forming tissues of hybrid aspen (*Populus tremula* × *P. tremuloides*) (Rajangam et al pers comm). We also showed that this protein binds specifically the herbicide 2,6-dichlorobenzonitrile (DCB), which is a known inhibitor of cellulose synthesis (Sabba and Vaughn 1999). This finding is consistent with the hypothesis that cellulose synthesis is coupled with cortical microtubules (Goddard et al 1994; Roberts et al 2004; Oda et al 2005), and suggests that *Ptt*MAP20 may have a role in mediating such interactions in poplar secondary cell walls.

In order to get more insight into the occurrence, evolution, and potential functions of TPX2-containing proteins we have carried out a bioinformatic analysis for all genes so far found to encode TPX2 domains with special reference to the poplar *Ptt*MAP20 and its putative orthologs in other plant species.

# Material and methods

## Sequence retrieval

Genome sequences and their allied gene predictions were retrieved for *Arabidopsis* (v 5, January 2004, tigr.org), rice (v 4, January 12, 2006, rice.tigr.org) and *Populus trichocarpa* (v 1.1, 2006, jgi.org, DoE Joint Genome Institute and Poplar Genome Consortium). A prerelease of *Medicago truncatula* genome (December 14, 2006), including gene predictions, was downloaded from http://www.medicago.org/. *Zea maize* contigs were analyzed and downloaded from www.plantgdb.org in April 2007. Animal TPX2 proteins (Q2U500, A2APB8, Q6P9S6, Q6DDV8, Q643R0, Q805A9, Q6NUF4, Q5ZIC6, TPX2, Q5RAF2, Q96RR5 and Q8BTJ3) were downloaded from the Pfam website (http://pfam.sanger.ac.uk/), and redundant sequences were ignored.

## Similarity search and alignments

Sequence similarity searches for mRNA and EST support for genes were primarily conducted using NCBI's online Blast using full-length protein sequences. The HMMER package (Eddy 1998) was used for the alignment and identification of TPX2 domains, using the Pfam model PF06886.1. The Pfam website was used for general domain architectural analysis. The assembly of ESTs to putative transcripts was done with CAP3 (Huang and Madan 1999). Multialignments were computed using Kalign (Lassmann and Sonnhammer 2005), Muscle (Edgar 2004), and MAFFT (Katoh et al 2002, 2005). Visualization of the alignments was done using TeXShade (Beitz 2000). The quality of conservation of a sequence alignment of plant TPX2 domains was plotted using the EMBOSS plotcon software (Available from: http://bioweb.pasteur.fr/seqanal/interfaces/plotcon.html), with a standard window size of 4 for both DNA and protein sequences.

## Phylogenetic analysis

Phylogeny studies were done using MrBayes (Ronquist and Huelsenbeck 2003). Each analysis had four MCMC chains running for $4 \times 10^6$ iterations, default thinning, and the first 10% iterations removed as burnin. The mixed amino acid model with gamma-distributed rate-change over sites was chosen. Analysis of possible adaptive evolution was performed with codeml in the PAML package (Yang 1997).

## Gene mapping

Mapping of the transcripts or proteins to genome sequences and the determination of exon/intron-structure was done using Exonerate (Slater and Birney 2005). Promoter analyses were conducted using TSSP (Solovyev and Shahmuradov 2003) and PLACE (Higo et al 1999). Phylogenetic footprinting for motif finding was done with MEME (Bailey and Elkan 1994; Bailey and Gribskov 1998) with settings chosen to look for up to five motifs, present in some but not necessarily all sequences. Local genomic alignments were computed using DBA (Jareborg et al 1999).

## Expression profiling

The expression profiling data for the *Arabidopsis* TPX2 genes were extracted from the Gene Atlas performed with ATH1 (22K full genome Arabidopsis Affymetrix GeneChip) available online http://www.arabidopsis.org/.

# Results and discussion

## Identification of TPX2 proteins

Proteins containing a TPX2 domain were identified in the completely sequenced genomes of poplar (*Populus*

*trichocarpa*) (Tuskan et al 2006), *Arabidopsis thaliana* (Arabidopsis Genome Initiative 2000), and rice (*Oryza sativa*) (Goff et al 2002; Goff 2005), using the Pfam model of TPX2. The fully sequenced animal genomes only contain one gene encoding a single TPX2 domain. In contrast, plant genomes were found to encode rich repertoires of different proteins containing a TPX2 domain, 14 in *Arabidopsis*, 10 in rice, and 19 in *Populus*. Here we have compared the TPX2 domains in all fully sequenced plant and animal genomes as well as other known TPX2 containing proteins from the Pfam database. Sequence alignments show

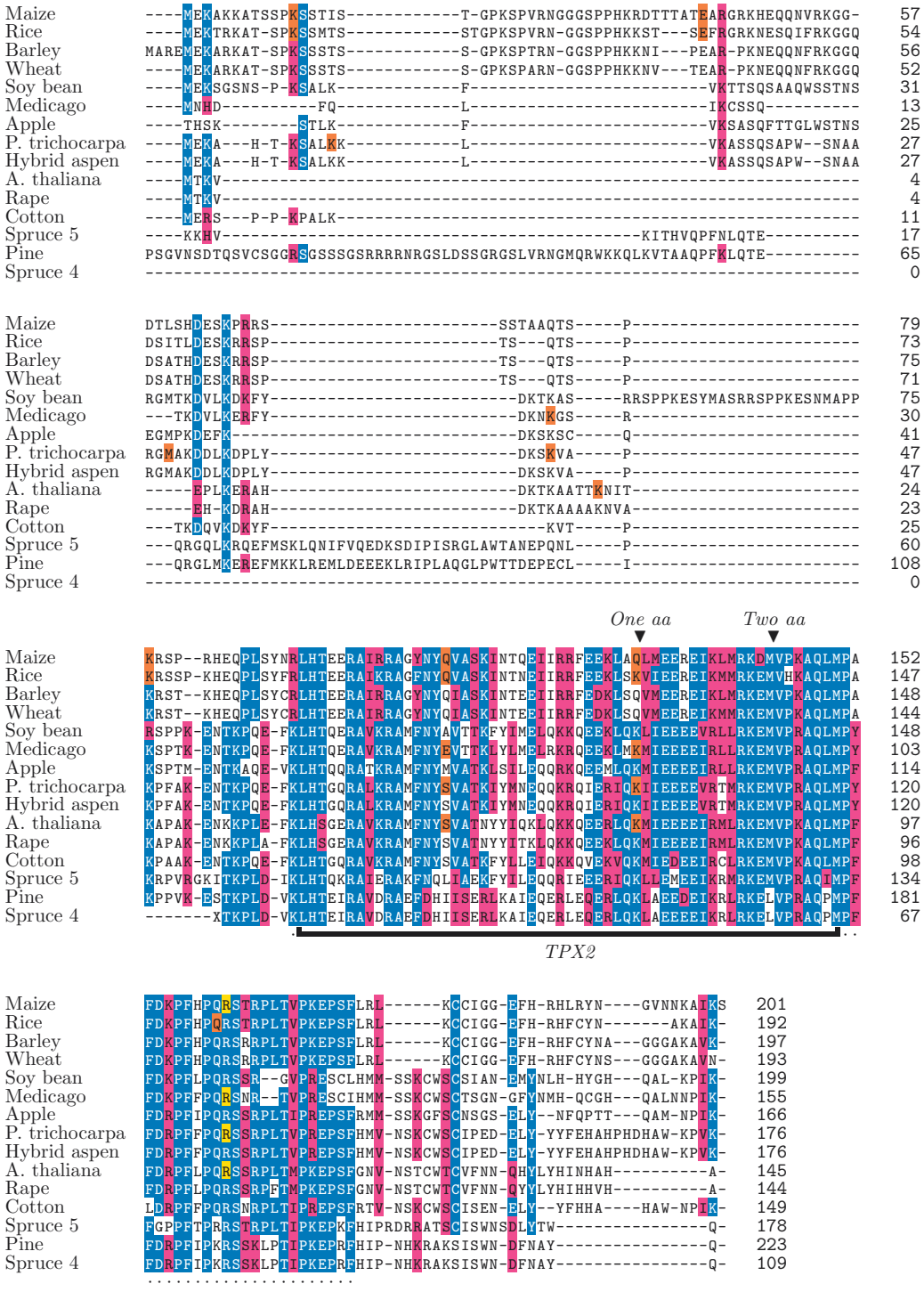



**Figure 1** Sequence alignment of M20L proteins made with the MAFFT alignment tool. The TPX2 domain is marked with a black bar and the extended TPX2 domain by a dotted line under the sequences. Notes above the TPX2 domain indicate sites where residues are lost relative to the Pfam domain model of TPX2. The identity or similarity to a consensus sequence is indicated in blue and red (respectively), the last codon of an exon in orange, and the exon/intron boundary in frame 2 in yellow.

that these proteins exhibit low or no similarity with each other beyond their TPX2 domains (Figure 1). This indicates that the TPX2 domain is a common nominator of many multi-domain proteins that overall do not have common evolutionary origins. However, a common origin is apparent for the TPX2 domains suggesting that these domains share a similar function.

## Identification and sequences analysis of PttMAP20 orthologs

Due to the possible involvement of *Ptt*MAP20 in cellulose biosynthesis (Rajangam et al pers comm.), putative orthologs (as defined by Fitch 1970) were searched for corresponding genes in all published plant genomes and in the NCBI dbEST (Table 1). The current annotations of the gene models Os09g13650.1 in rice and At5g37478 in *A. thaliana* had to be adjusted since sequences homologous to the other TPX2 proteins were detected beyond the current gene models (see Materials and Methods). Reannotation of the rice gene was facilitated by an EST sequence, NM_001069361.1, and in both cases the new gene models were more similar to *PttMAP20* than the current gene models. Table 1 summarizes the gene models identified in different plant species. These gene models are hereafter designated as 'Map20-Like' (*M20L*). None of the animal TPX2 proteins were found to show any significant similarity with *Ptt*MAP20 beyond the TPX2 domain (data not shown).

Putative *PttMAP20* orthologs were easy to detect among monocots and dicots by simple analysis of sequence similarity. In most cases, phylogenetic analysis (see below) would corroborate the highest scoring sequences as orthologs of *Ptt*MAP20. In cases when the best database hit was not an ortholog, the similarity score had already indicated a more distant relationship to *Ptt*MAP20.

In gymnosperms, however, thorough phylogenetic analysis was required in order to find putative orthologs of *Ptt*MAP20. In this way, 11 sequences in *Pinus* and 10 sequences in *Picea* (with ESTs from *P. engelmanii*, *P. glauca*, and *P. sitchensis*) were found to encode a TPX2 domain.

Multiple alignments of the MAP20-like proteins were computed using MUSCLE, MAFFT, and KALIGN, but no reliable full-length alignment could be found. In particular, there was little consensus at the N- and C-terminal ends of the proteins (see Figure 1 and Supplemental Figure 1 for MAFFT and MUSCLE alignments, respectively). However, the programs did agree on a strongly conserved 81-residue region containing the TPX2 domain, here called the extended TPX2 domain. It was verified that no other known domain structures are located in the regions flanking the TPX2 domain of the M20L proteins.

## Conservation of extended TPX2 domain of M20L

To understand the conservation of the TPX2 domain in plants, a multiple alignment was made with DNA sequences of the extended TPX2 domain within the M20L sequences. The similarity plots were made with both DNA and protein sequences of the extended M20L TPX2 domain to check the quality of conservation. The similarity score was low for the DNA sequences compared to protein sequences (Figure 2a), mainly due to silent mutations in the 3rd nucleotide of the triplet codon (Figure 2b).

## Phylogenetic analysis of TPX2 domain

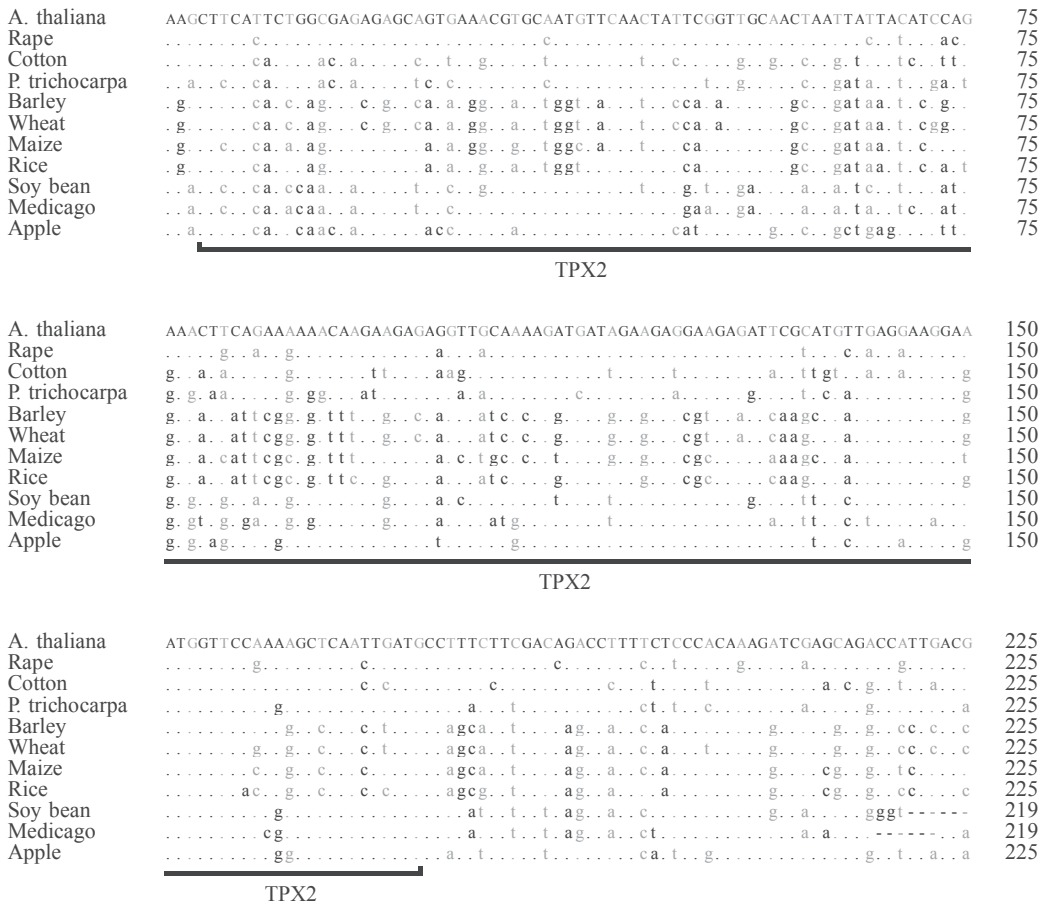
Molecular evolution of the TPX2 domain was studied with phylogenetic analysis of all available TPX2 domain sequences (Figure 3). The phylogenetic tree has two distinct branches,

**Table 1** A list of MAP20-Like gene models (*M20L*) in different plant species

| Species | Accession | Origin | Protein acc |
|---|---|---|---|
| *P. trichocarpa* | eugene3.00440209 (592874) | Gene prediction | *Pt*MAP20 |
| *P. tremula* × *P. tremuloides* | POPLAR.3073.C1 | PopulusDB | *Ptt*MAP20[a] |
| *A. thaliana* | At5g37478 | Gene prediction | *At*M20L, At5g37478_alt[b] |
| *O. sativa* | Os09g13650.1 | Gene prediction | *Os*M20L[b] |
| *Medicago truncatula* | AC147472 | Genome scaffold | *Mt*M20L |
| *Brassica napus* | CN734834 | dbEST | *Bn*M20L |
| *Gossypium hirsutum* | DW520361 | dbEST | *Gh*M20L |
| *Glycine max* | BQ612497 | dbEST | *Gm*M20L |
| *Hordeum vulgare* | CX630106 | dbEST | *Hv*M20L |
| *Malus* × *domestica* | CV083029 | dbEST | *Md*M20L[c] |
| *Triticum aestivum* | BJ280729 | dbEST | *Ta*M20L |
| *Zea maise* | AY110515 | dbEST | *Zm*M20L[d] |

[a]Identical to PtMAP20; [b]New gene model; [c]Single EST support, may be short on 5' side; [d]Ambiguous codons were resolved using the shorter AI795385.

a




b

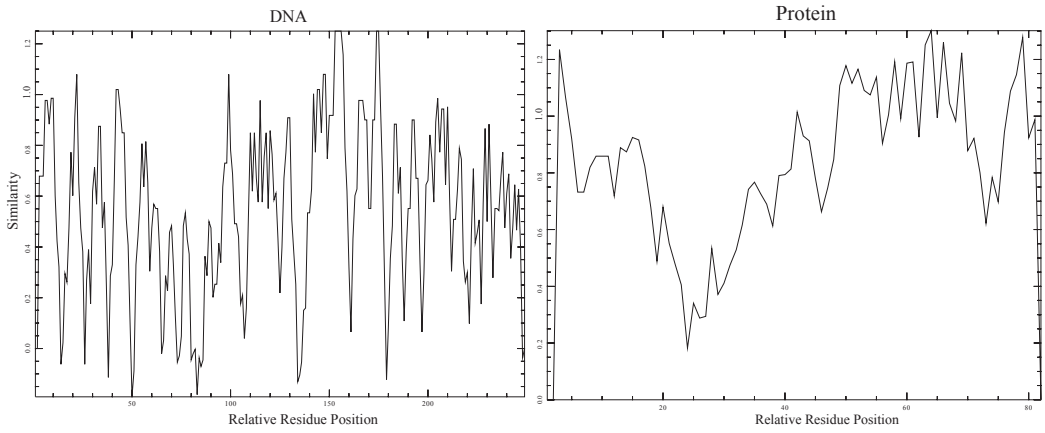



**Figure 2** (**a**) A multiple alignment made with DNA sequences of the extended TPX2 domains of the M20L protein sequences. Differences with respect to *A. thaliana* are noted, and the conserved bases are indicated by dots. The third codon position is printed in grey. (**b**) Similarity plot graph made with the extended TPX2 domain with reference to relative residue position (both DNA and Protein sequences) using EMBOSS plotcon software using a standard window size of 4.

one with and one without animals. The plant-only branch contains a subtype of TPX2 containing the so-called KLEEK motif previously identified in the *Arabidopsis* WVD2 protein (Yuen et al 2003). Proteins containing a KLEEK motif form a monophyletic clade, and have several branches each with its own monocot and dicot sub-branches, as exemplified by WDL5/WDL6 and WDL4 in Figure 3. The animal-containing branch is where our analysis put the M20L genes. Notice that gymnosperm sequences are present in both major branches and that there are two likely orthologs to *Ptt*MAP20 in *Picea*.

The subtrees in the phylogeny mostly follow the established species history (Figure 4), but there are

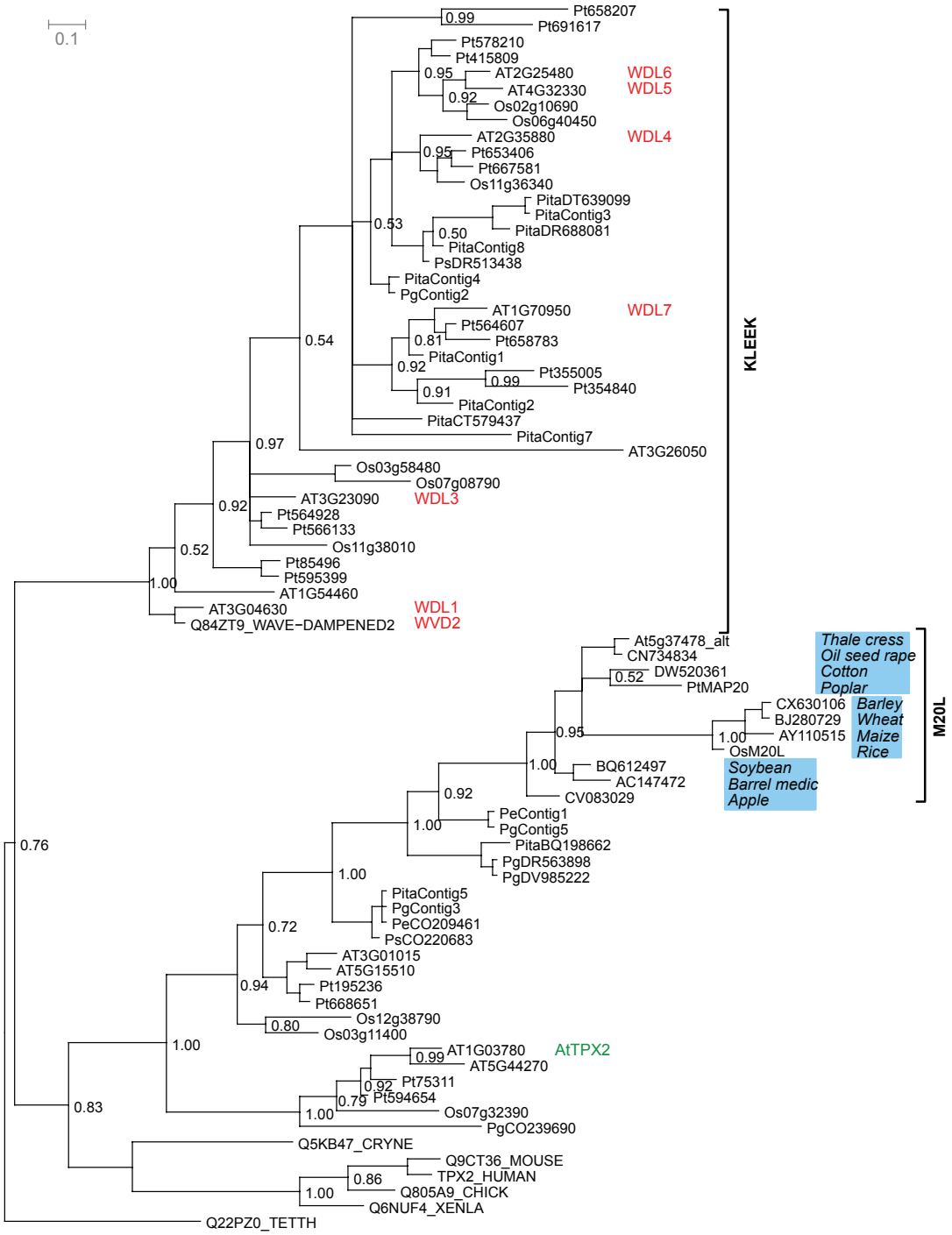



**Figure 3** Phylogenetic tree made with all the available and newly found eukaryotic proteins containing a TPX2 domain. The Bayesian posterior probability is indicated by numbers to the right of the edge in question. The clades with M20L proteins and KLEEK motif are marked. Genes with reported phenotypes in *Arabidopsis* are marked in red (Yuen et al 2003). The gene noted as AtTPX2 by (Perrin et al 2007) is annotated in green.

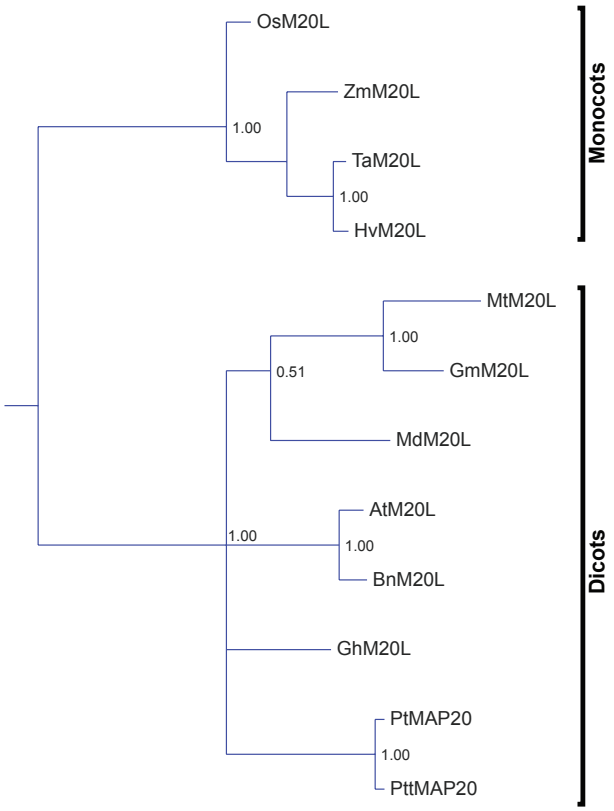



**Figure 4** A phylogenetic tree made with the extended TPX2 domain.

peculiarities to note. For example, the M20L orthologs deviate from the established species phylogeny (see eg, the apple sequence) and the branching is not fully resolved. The lack of phylogenetic resolution is seen also in other parts of the tree, including several branches with low statistical support. However, these problems are to be expected as the phylogeny is estimated from the short TPX2 domain alone, with only 58 informative positions in the alignment. We chose to root the phylogeny using the protozoan *Tetrahymena thermophila* as an outgroup. Although the eukaryotic rooting has long been contested, recent data suggests protozoa as outgroup to animals and plants (Arisue et al 2005; Ciccarelli et al 2006). The displayed structure of the tree, although weakly supported at the crucial edges, allow us to speculate that a first duplication occurred before plants and animals diverged. This would imply that animals lost at least one TPX2 gene, while plants took advantage of the redundancy by further evolutionary diversification. This old duplication could reflect different functions in the two main branches of the TPX2 phylogeny. Regardless of where the root of the phylogeny is correctly placed, the expansion of proteins containing a TPX2 domain among plants is in striking contrast to animal TPX2. There are several duplications implied for plants in both parts of the tree, but animals contain only one TPX2 gene.

All genes in the TPX2 gene family in *P. trichocarpa*, except *MAP20*, seem to come in pairs and they are all more recent than the last species split in this phylogeny. This is consistent with the likely whole-genome duplication event (Tuskan et al 2006).

While there are plenty of duplications in the TPX2 gene family in general, none of the species included in the present study has more than a single copy of *M20L* genes. Such a singleton representation suggest that *M20L* genes have characteristics that render them duplication resistant (Paterson et al 2006) during events of polyploidy. A recent study on gene family evolution dynamics (Wapinski et al 2007) showed that duplication resistant characteristics are typical for genes related to essential growth processes, genes active in organelles and nucleus, and genes essential for viability. It is possible that the proposed, but as yet undefined role of MAP20 in cellulose biosynthesis would fulfill the requirements of such a process in angiosperms.

## Phylogenetic analysis of the extended TPX2 domain in M20L proteins

The M20L clade was revisited by reconstructing a phylogeny based on all Angiosperm-extended TPX2 domains of M20L whereby the branching was significantly improved. Several branches are still not resolved or lack statistical support. Those parts that do have statistical support are in agreement with the accepted species tree. In particular, the M20L phylogeny clearly separates monocots and dicots (Figure 4).

The M20L phylogeny formed the basis for a search for signs of adaptive evolution (Yang 1997; Bielawski and Yang 2003). Only the extended TPX2 domain was studied since we could not derive a reliable multialignment over the full protein sequences (Figure 1 and Supplemental Figure 1). The hypothesis was that key properties in this conserved region could have changed, especially after the monocot/dicot split. However, no signs of positive selection were found, either over branches or on sequence sites. The extended TPX2 domain thus seems to have been under negative selection only.

## Genomic organization of TPX2 genes

The organization of TPX2 genes was mapped in all the three fully sequenced model plants: the universal model *Arabidopsis*, the wood model *Populus* and the grass model *Oryza* (Figure 5a–c). We could not find a general pattern for the TPX2 gene locations in the three genomes. Synteny comparisons suggested that many recent paralogs, eg, WDL1/WVD2

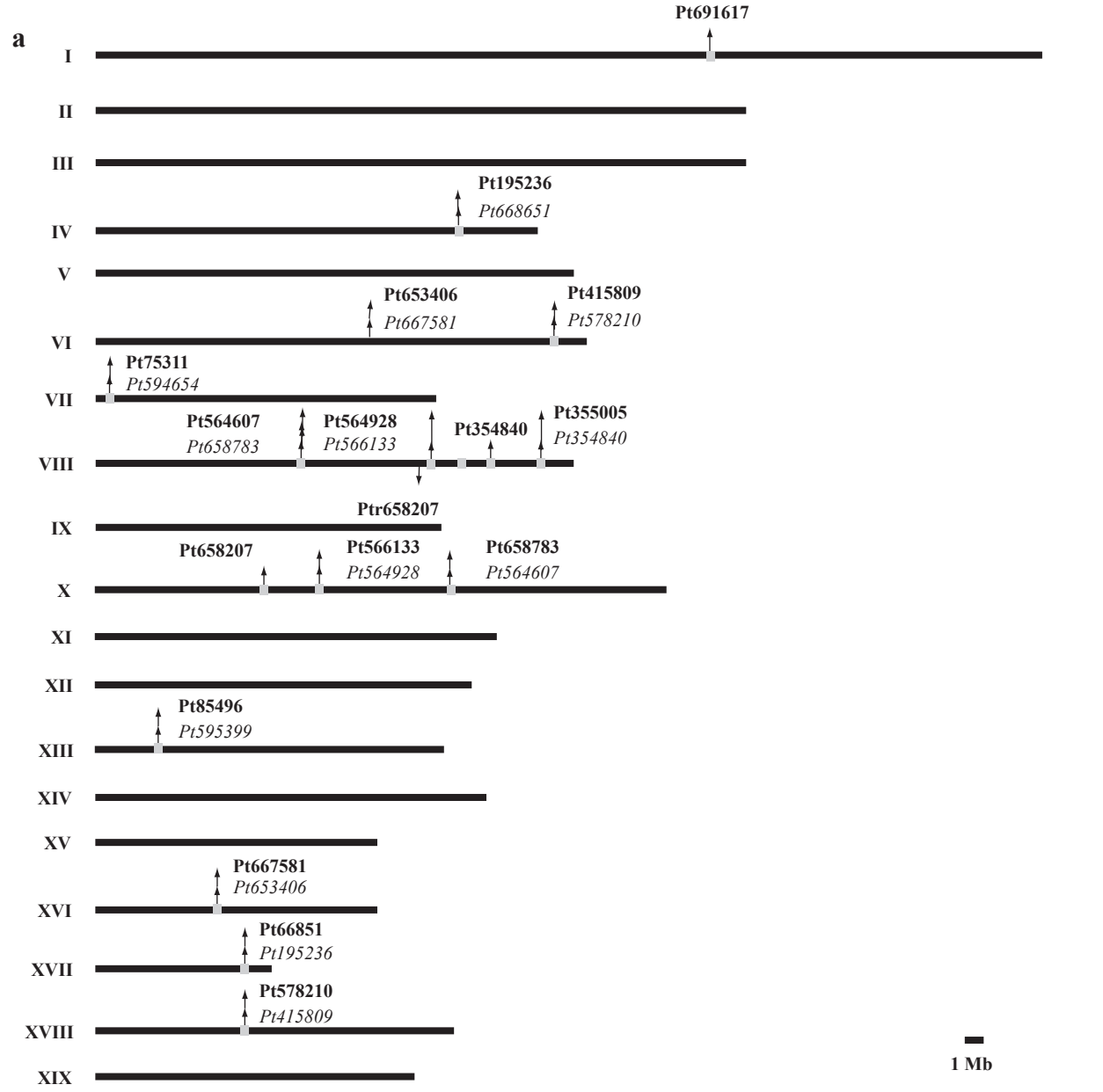



**Figure 5** Genomic organization (to scale) of TPX2 genes in the genomes of (**a**) *Populus trichocarpa*, (**b**) *Oryza sativa*, and (**c**) *Arabidopsis thaliana*.

and WDL6/WDL5 in our phylogeny, could be due to large scale duplications (Table 2). The linkage groups previously found in the three species were used to overlay the location of these gene pairs (Arabidopsis Genome Initiative 2000; Goff et al 2002; Goff 2005; Tuskan et al 2006). The many recent paralogs, especially in Poplar, could make the gene localization ambiguous, but the highest scoring alignment stood out in cases. When aligning a protein sequence to the genome, the second-best match had incomplete sequence coverage, several or many mutations, and significantly lower score compared to the highest scoring match. The TPX2 paralogs in the poplar genome that arose due to the genome duplication are noted in the gene models and their respective linkage groups (Table 2 and Figure 3). Microarray experiments in *Arabidopsis* indicate that WDL1 (At3g04630) and WVD2 (At1g3780) have similar expression patterns (Figure 6) with increased expression in inflorescence tissue, rich in cellulose biosynthesis, consistent with a duplicated regulatory module. Even WDL5 (At4g32330) exhibits somewhat increased expression in inflorescence tissue.

A search for TPX2-containing pseudogenes was conducted in the *P. trichocarpa* genome using translating

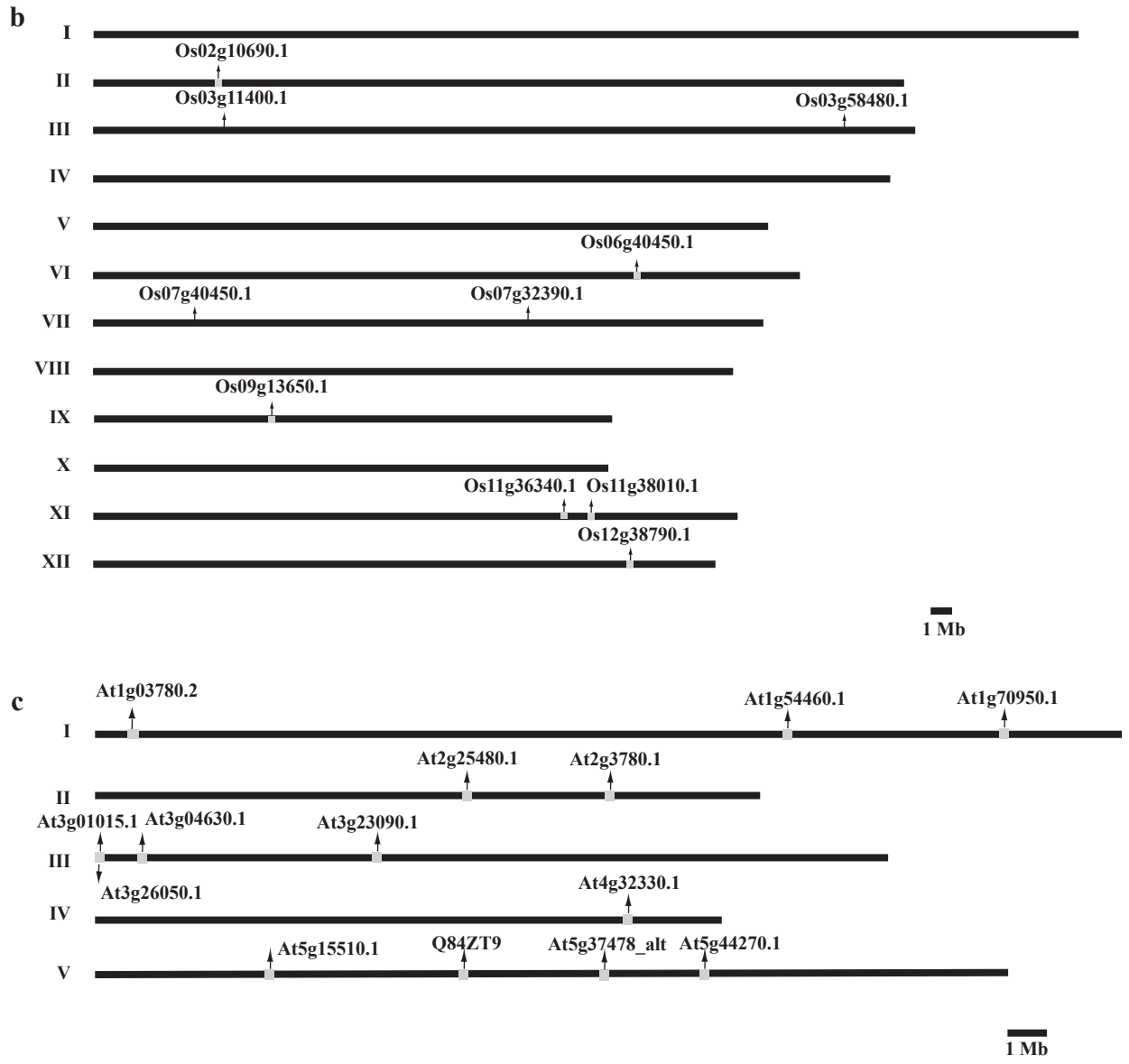



**Figure 5** Continued.

Blast and the known TPX2 domains. Even at a relatively low threshold (E = 0.001), no hits were found outside the known TPX2 genes. Unless some predicted genes are false positives, the *P. trichocarpa* does not have any TPX2 containing pseudogenes.

## Gene structure determination and comparison of M20L genes

A comparison of the exon/intron structures of known *M20L* genes in *A. thaliana*, *P. trichocarpa*, *O. sativa*, *M. truncatula*, and *Z. mays* revealed variation in both the exon numbers and the intron lengths (Figure 7). In particular, monocot members showed longer introns. Interestingly, the TPX2 domain intersects three exons in all studied species and the domain covers 48 bp in the first and 63 bp in the last of the three exons. Consequently, the mid exon is 60 bp in all cases. This is another indication of strong selective pressure on the TPX2 domain.

In order to identify potential gene regulatory elements, 3000 bp of the genome sequence was extracted upstream from the *M20L* start-of-translation point in *O. sativa*, *A. thaliana*, *M. truncatula*, and *P. trichocarpa*. In *M. truncatula*

the M20L gene was found to be located in a contig (CT963077, from chromosome 3) including 3000 bp upstream of the translation start of the gene. The *Zea mays* gene had a match against two contigs (ZmGSStuc11-12-04.70298.1 and ZmGSStuc11-12-04.9118.1). As they were seemingly overlapping, they were reassembled (using CAP3) into one contig from which 1975 bp upstream of start-of-translation was extracted. In an alignment of the five genomic regions, the pairwise identity was lower than 44% in all cases, indicating considerable divergence on regions without selection. However, two motifs were revealed using phylogenetic footprinting (Figures 8 and 9). Motif 1 (39 bp) was significant in all five species ($E = 1.7 \times 10^{-14}$). In dicots, a slightly wider motif of 43 bp was found (Figure 9). Motif 2 consisting of 28 bp was also found in all five species ($E = 2.8 \times 10^{-5}$) but this width was reduced by two bases on the 5' flanking region if applying MEME to dicots only. No further significant motifs could be found in all five species. These motifs were, with small differences, corroborated by pairwise genome alignments using DBA. It was furthermore verified that the motifs are noncoding and not covered by any ESTs, and are hence not likely to be part of the translated regions of these genes. It has been demonstrated (Bejerano et al 2004; Sandelin et al 2004), that extremely conserved regions can be regulators for genes far away from the motifs, but the proximity to the M20L genes make them interesting and ideal candidates for experiments. These motifs are larger than what is common for a transcription factor binding site (TFBS), suggesting that they may represent two regulatory modules. It is interesting to note that although dicots and monocots share little similarity in TPX2 gene structure and protein sequence, their TFBS motifs are well conserved.

**Table 2** Gene pairs and their respective linkage groups present in *Populus trichocarpa, Oryza sativa and Arabidopsis thaliana*

| Species | Duplicated gene pairs | Linkage group |
|---|---|---|
| *Populus trichocarpa* | Pt668651: Pt195236 | IV: XVII |
| | Pt415809: Pt578210 | VI: XVIII |
| | Pt658783: Pt564607 | VIII: X |
| | Pt667581: Pt653406 | VI: XVI |
| | Pt658207:Pt658207 | VIII:X |
| *Oryza sativa* | Os02g10690.1: Os06g40450.1 | II:VI |
| | Os03g58480.1: Os12g38790.1 | III: XII |
| *Arabidopsis thaliana* | AT2g25480.1: At4g32330.1 | II:V |
| | Q84ZT9: At3g04630.1 | III:V |

There were no promoter regions found by homology and known cis-acting regulatory elements had no hits of statistical significance. Putative TATA-boxes upstream of M20L were predicted and the same regions were also used to search dbEST in order to identify likely UTRs. In Poplar

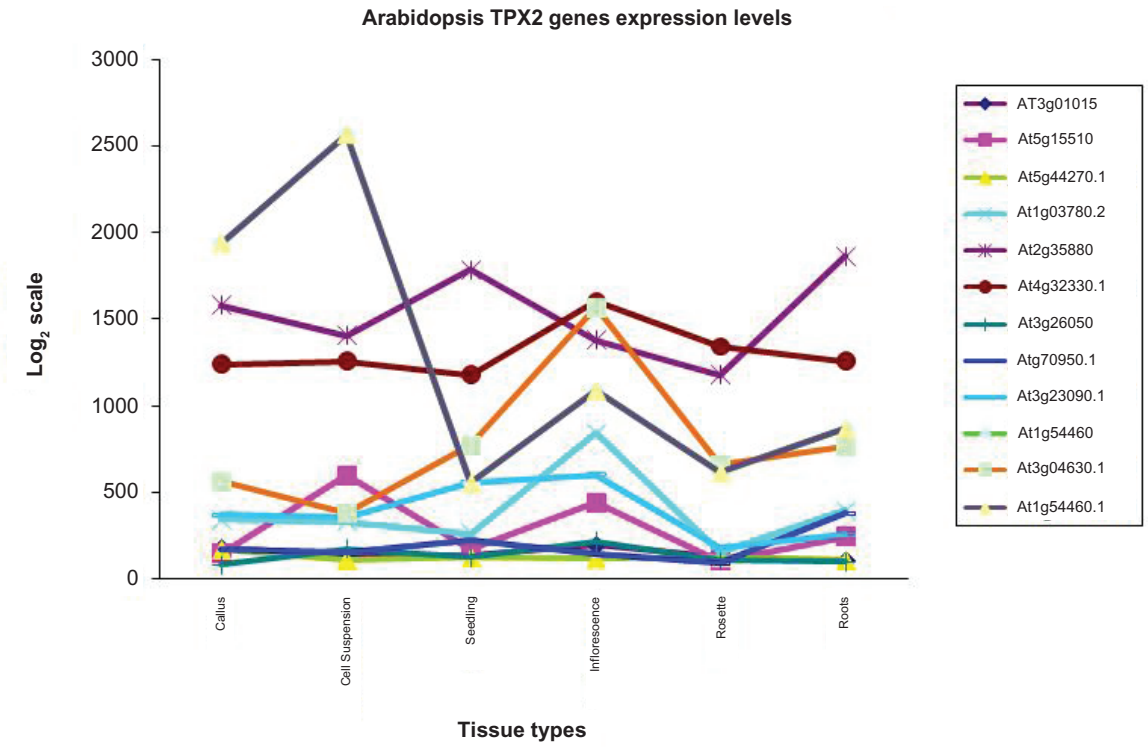



**Figure 6** Relative expression levels of the TPX2 genes in different tissue types of *Arabidopsis thaliana*. Data from http://www.arabidopsis.org/. The expression patterns for WDL1 (At3g04630) and WVD2 (At1g3780) are roughly the same, with elevated levels in inflorescence tissue.

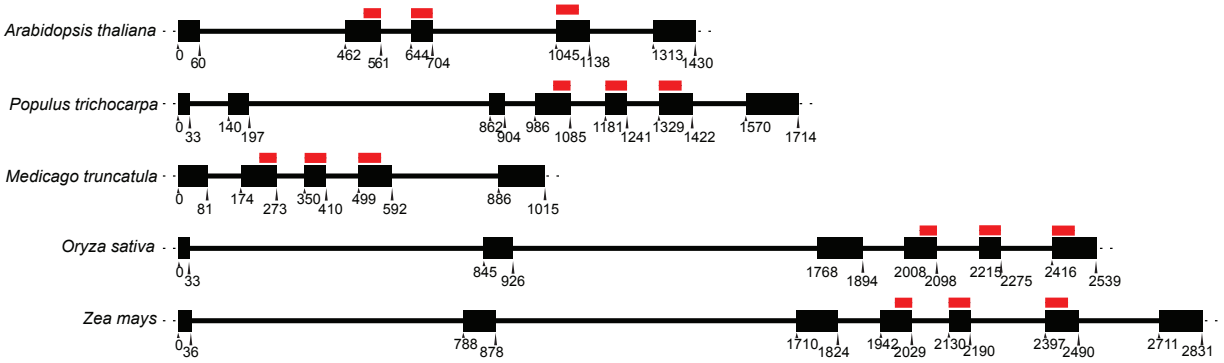



**Figure 7** *M20L* gene structure: the positions (in base pairs from the translation start) of gene CDS are marked, to scale, with dark boxes while the thick lines correspond to introns. The TPX2 regions are marked with dark red lines.

there are 4 putative TATA boxes. Considering EST hits at this region, the TATA box at the position -100 might be important (Figure 8). A similar arrangement is seen in *M. truncatula* although with two TATA boxes downstream of the TFBS motifs. In this case, ESTs suggest a promoter, which is close to the conserved TFBS motifs similar to *Populus*, but this would render both predicted TATA boxes false. The locations of the motifs in *Arabidopsis* are quite different in that they are distant from each other. Short EST hits are reported for the downstream region of the first motif, but offer no real support for any of the predicted TATA boxes. Looking at the monocots, *O. sativa* has the motifs in the same order as the dicots and EST hits are found in the downstream region of motif 2, clearly an extension of the first exon. It is reasonable to assume these ESTs represent an UTR and a predicted TATA box is consistent with this hypothesis. The results are similar for *Z. mays*, with exon-extending and UTR-indicating EST hits, except that the order of the motifs is switched. The long EST hit prior to the motifs in *Oryza*, as well as the long EST hit in *Medicago*, escape our interpretation, but in the light of the results of the ENCODE project (Birney et al 2007) this might be randomly transcribed regions.

## Conclusions

Proteins containing a TPX2 domain are only found in highly evolved members of Eukaryota. Representative members of

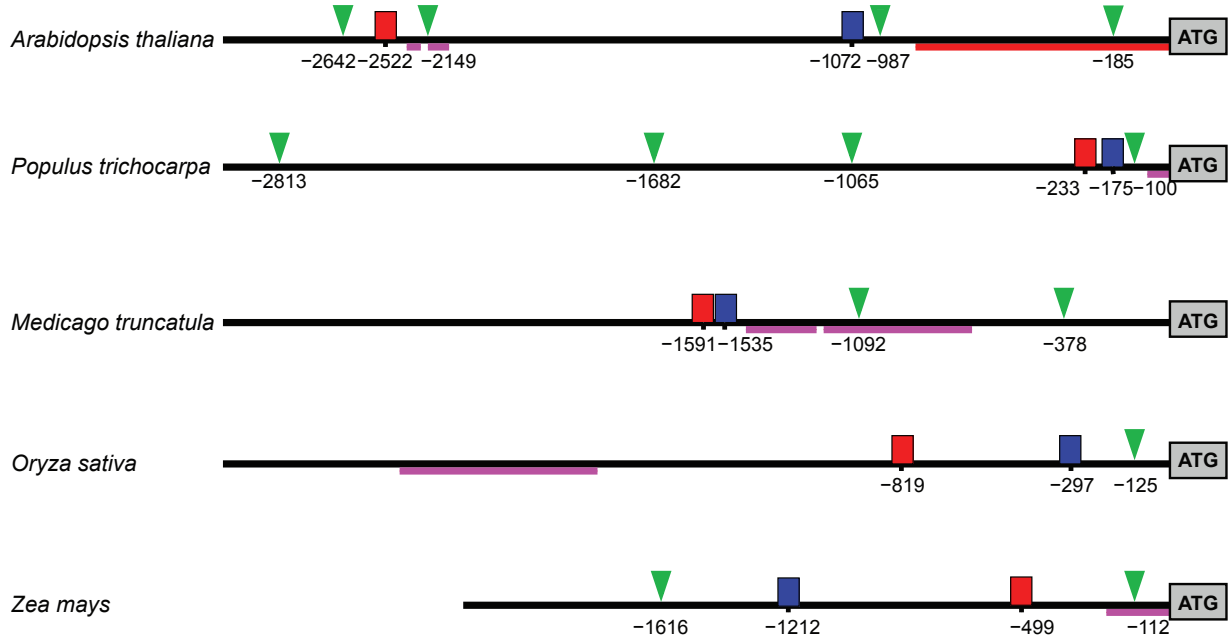



**Figure 8** Possible regulatory elements and promoters of *M20L*. Upstream regions with predicted TATA boxes are indicated by a green triangle, motif 1 by a red box, motif 2 by a blue box, and EST hit regions by dark pink lines. For *A. thaliana*, the core promoter given in AGRIS (Palaniswamy et al 2006) is indicated by a red line. The lines are scaled to actual sequence length and positions are given in base pairs from translation start.

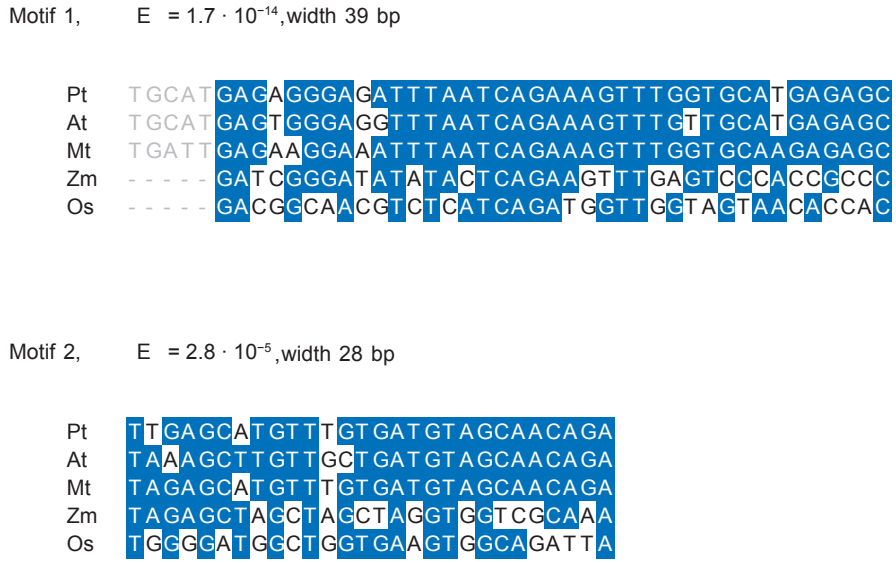



**Figure 9** Sequence alignments for the motifs 1 and 2. Blue color indicates identity to the consensus sequence. The left flank of motif 1 shows the extension of this motif in dicots.

this family of proteins show an interesting species distribution as reported in Pfam (July, 2007). Though this representation might still reflect the prevailing sequence data available across the different species, there are no exceptions so far. Among the Protists, TPX2 is reported in the Ciliate *Tetrahymena thermophila* SB210 where the cilia is made up of assembled tubulin and axonemal protein units (Redeker et al 2005). There are 15 reported TPX2 proteins in animals, representing only the vertebrates, and 2 in fungi (order: Tremellales). As many as 50 TPX2-containing plant proteins have been reported in Pfam and all of them so far in Angiosperms. However, with stringent searches using the EST database, more TPX2 gene sequences were identified also in the Gymnosperms. The TPX2 domain thus seems to be a common nominator of a number of plant and animal proteins.

Analyses of domain structures in different multidomain proteins reveal that domains pairing with other domains generally occur in the same combination (Apic et al 2001; Vogel et al 2004). This suggests that such domains have evolved interdependent functions and that the order of the domains has therefore been maintained during evolution. All known animal TPX2 proteins have an Aurora binding domain as well as RNA binding and recognition domains, and thus probably fulfil the same function in the different species (Weiner and Bornberg-Bauer 2006). In contrast, the TPX2 domain is the only clearly conserved part of these proteins in different plant species, and the overall sizes and sequences of these proteins are quite different. It is therefore likely that, apart from a capacity to interact with microtubules, these proteins have different functions.

Plant TPX2 genes in general are duplicating liberally while *M20L* genes are apparently strictly resistant to duplications. The extended TPX2 domains in the M20L proteins are under strong selective pressure as evidenced by the conservation of residues and nucleotide patterns and the exon/intron structure. Since the extended TPX2 domain is the sole strictly conserved part of these genes, it is arguably the source of the apparent duplication resistance in the M20L clade. We draw the conclusion that M20L is, in some respect, more important than its paralogs.

There is presently too little data to make a good interpretation of the origins of the TPX2 domain. However, there are signs that there was a first duplication before the separation of animals and plants. As more data become available, especially protist TPX2 genes and more gene-order data, this question may be resolvable.

## Acknowledgments

We thank Professor Jens Lagergren for critical reading of the manuscript. This work was supported by the Swedish Research Council Formas (ASR, TTT). The authors report no conflicts of interest in this work.

## References

Amos LA, Schlieper D. 2005. Microtubules and maps. *Adv Protein Chem*, 71:257–98.

Apic G, Gough J, Teichmann SA. 2001. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, 310:311–25.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408:796–815.

Arisue N, Hasegawa M, Hashimoto T. 2005. Root of the Eukaryota tree as inferred from combined maximum likelihood analyses of multiple molecular sequence data. *Mol Biol Evol*, 22:409–20.

Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, AAAI Press, Menlo Park, CA, pp. 28–36.

Bailey TL, Gribskov M. 1998. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14:48–54.

Bayliss R, Sardon T, Vernos I, et al. 2003. Structural basis of Aurora-A activation by TPX2 at the mitotic spindle. *Mol Cell*, 12:851–62.

Beitz E. 2000. TeXshade: shading and labeling of multiple sequence alignments using LaTeX2e. *Bioinformatics*, 16:135–9.

Bejerano G, Pheasant M, Makunin I, et al. 2004. Ultraconserved elements in the human genome. *Science*, 304:1321–5.

Bielawski JP, Yang Z. 2003. Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Func Genomics*, 3:201–12.

Birney E, Stamatoyannopoulos JA, Dutta A, et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816.

Brunet S, Sardon T, Zimmerman T, et al. 2004. Characterization of the TPX2 domains involved in microtubule nucleation and spindle assembly in Xenopus egg extracts. *Mol Biol Cell*, 15:5318–28.

Chaffey N, Barlow P, Sundberg B. 2002. Understanding the role of the cytoskeleton in wood formation in angiosperm trees: hybrid aspen (Populus tremula × *P. tremuloides*) as the model species. *Tree Physiol*, 22:239–49.

Chaffey N, Barnett J, Barlow P. 1999. A cytoskeletal basis for wood formation in angiosperm trees: the involvement of cortical microtubules. *Planta*, 208:19–30.

Chaffey NJ, Barnett JR, Barlow PW. 1997. Cortical microtubule involvement in bordered pit formation in secondary xylem vessel elements of Aesculus hippocastanum L (Hippocastanaceae): A correlative study using electron microscopy and indirect immunofluorescence microscopy. *Protoplasma*, 197:64–75.

Ciccarelli FD, Doerks T, von Mering C, et al. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311:1283–7.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics*, 14:755–63.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, 32:1792–7.

Elofsson A, Sonnhammer ELL. 1999. A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, 15:480–500.

Fitch WM. 1970. Distinguishing homologous from analogous proteins. *Syst Zool*, 19:99–113.

Funada R, Miura H, Shibagaki M, et al. 2001. Involvement of localized cortical microtubules in the formation of a modified structure of wood. *J Plant Res*, 114:491–7.

Gardiner JC, Taylor NG, Turner SR. 2003. Control of cellulose synthase complex localization in developing xylem. *Plant Cell*, 15:1740–8.

Goddard RH, Wick SM, Silflow CD, et al. 1994. Microtubule components of the plant-cell cytoskeleton. *Plant Physiol*, 104:1–6.

Goff SA. 2005. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica) (April, pg 92, 2002). *Science*, 309:879.

Goff SA, Ricke D, Lan TH, et al. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp japonica). *Science*, 296:92–100.

Hasek J, Trachtulcova P, Kohlwein SD, et al. 2003. Colocalization of cortical microtubules and F-actin in Dipodascus magnusii using confocal laser scanning microscopy. *Folia Microbiol*, 48:177–82.

Higo K, Ugawa Y, Iwamoto M, et al. 1999. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res*, 27:297–300.

Huang XQ, Madan A. 1999. CAP3: A DNA sequence assembly program. *Genome Res*, 9:868–77.

Jareborg N, Birney E, Durbin R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res*, 9:815–24.

Katoh K, Kuma K, Toh H, et al. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res*, 33:511–8.

Katoh K, Misawa K, Kuma K, et al. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, 30:3059–66.

Korolev AV, Buschmann H, Doonan JH, et al. 2007. AtMAP70-5, a divergent member of the MAP70 family of microtubule-associated proteins, is required for anisotropic cell growth in Arabidopsis. *J Cell Sci*, 120:2241–7.

Lassmann T, Sonnhammer ELL. 2005. Kalign – an accurate and fast multiple sequence alignment algorithm. *BMC Bioinfo*, 6:298.

Mathur J. 2004. Plant cytoskeleton: Reinforcing lines of division in plant cells. *Curr Biol*, 14:R287–9.

Oda Y, Mimura T, Hasezawa S. 2005. Regulation of secondary cell wall development by cortical microtubules during tracheary element differentiation in Arabidopsis cell suspensions. *Plant Physiol*, 137:1027–36.

Palaniswamy SK, James S, Sun H, et al. 2006. AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol*, 140:818–29.

Paradez A, Wright A, Ehrhardt DW. 2006. Microtubule cortical array organization and plant cell morphogenesis. *Curr Opin Plant Biol*, 9:571–8.

Paterson AH, Chapman BA, Kissinger JC, et al. 2006. Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends Genetics*, 22:597–602.

Perrin RM, Wang Y, Yuen CYL, et al. 2007. WVD2 is a novel microtubule-associated protein in Arabidopsis thaliana. *Plant J*, 49:961–71.

Redeker V, Levilliers N, Vinolo E, et al. 2005. Mutations of tubulin glycylation sites reveal cross-talk between the C termini of alpha- and beta-tubulin and affect the ciliary matrix in Tetrahymena. *J Biol Chem*, 280:596–606.

Roberts AW, Frost AO, Roberts EM, et al. 2004. Roles of microtubules and cellulose microfibril assembly in the localization of secondary-cell-wall deposition in developing tracheary elements. *Protoplasma*, 224:217–29.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–4.

Sabba RP, Vaughn KC. 1999. Herbicides that inhibit cellulose biosynthesis. *Weed Sci*, 47:757–63.

Sandelin A, Bailey P, Bruce S, et al. 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5:99.

Sedbrook JC. 2004. MAPs in plant cells: delineating microtubule growth dynamics and organization. *Curr Opin Plant Biol*, 7:632–40.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinfo*, 6:31.

Solovyev VV, Shahmuradov IA. 2003. PromH: promoters identification using orthologous genomic sequences. *Nucleic Acids Res*, 31:3540–5.

Tuskan GA, DiFazio S, Jansson S, et al. 2006. The genome of black cottonwood, Populus trichocarpa (Torr and Gray). *Science*, 313:1596–604.

Vassilyev AE. 1996. The cytoskeleton of animals and higher plants: The comparison of structure and functions. *Zhurnal Obshchei Biol*, 57:293–325.

Vogel C, Bashton M, Kerrison ND, et al. 2004. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*, 14:208–16.

Wapinski I, Pfeffer A, Friedman N, et al. 2007. Natural history and evolutionary principles of gene duplication in fungi. *Nature*, 449:54–61.

Wasteneys GO, Yang YB. 2004. New views on the plant cytoskeleton. *Plant Physiol*, 136:3884–91.

Weiner J, Bornberg-Bauer E. 2006. Evolution of circular permutations in multidomain proteins. *Mol Biol Evol*, 23:734–43.

Wittmann T, Wilm M, Karsenti E, et al. 2000. TPX2, a novel Xenopus MAP involved in spindle pole organization. *J Cell Biol*, 149:1405–18.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*, 13:555–6.

Yuen CYL, Pearlman RS, Silo-Suh L, et al. 2003. WVD2 and WDL1 modulate helical organ growth and anisotropic cell expansion in Arabidopsis. *Plant Physiol*, 131:493–506.

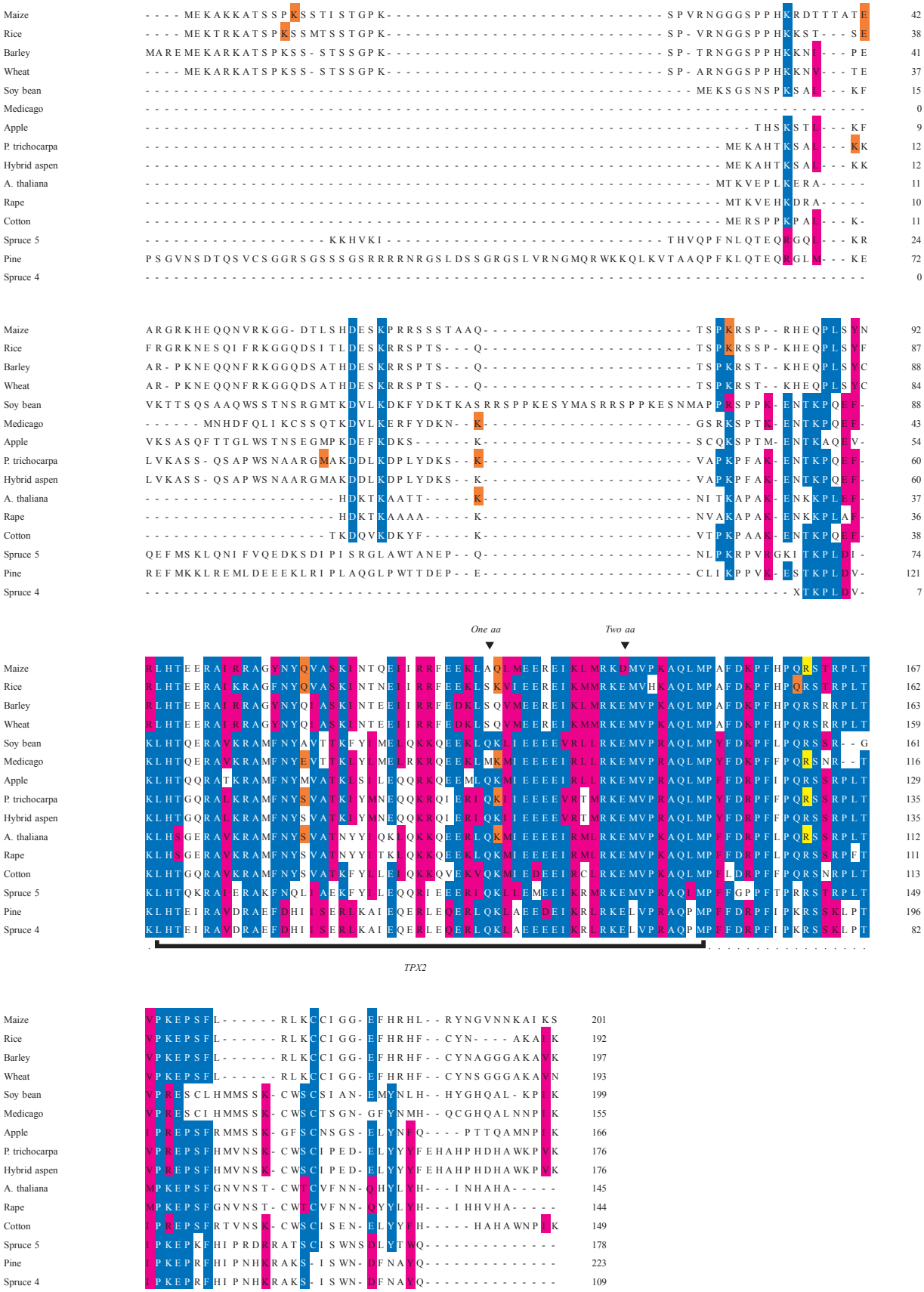
**Supplementary Table 1** Microtubule associated and binding protein reported in pfam by July 2007

| Accession | ID | Description |
|---|---|---|
| PF00225 | Kinesin | Kinesin motor domain |
| PF02991 | MAP1 LC3 | Microtubule associated protein 1A/1B, light chain 3 |
| PF00414 | MAP1B neuraxin | Neuraxin and MAP1B repeat |
| PF00418 | Tubulin-binding | Tau and MAP protein, tubulin-binding repeat |
| PF03271 | EB1 | EB1-like C-terminal motif |
| PF03607 | DCX | Doublecortin |
| PF03999 | MAP65 ASE1 | Microtubule associated protein (MAP65/ASE1 family) |
| PF05672 | MAP7 | MAP7 (E-MAP-115) family |
| PF06886 | TPX2 | Targeting protein for Xklp2 (TPX2) |
| PF08154 | NLE | NLE (NUC135) domain |
| PF00022 | Actin | Actin |
| PF00334 | NDK | Nucleoside diphosphate kinase |
| PF00784 | MyTH4 | MyTH4 domain |
| PF00956 | NAP | Nucleosome assembly protein (NAP) |
| PF00994 | MoCF biosynth | Probable molybdopterin binding domain |
| PF01031 | Dynamin M | Dynamin central region |
| PF01221 | Dynein light | Dynein light chain type 1 |
| PF03028 | Dynein heavy | Dynein heavy chain |
| PF03311 | Cornichon | Cornichon protein |
| PF03378 | CAS CSE1 | CAS/CSE protein, C-terminus |
| PF04402 | DUF541 | Protein of unknown function (DUF541) |
| PF05804 | KAP | Kinesin-associated protein (KAP) |
| PF05937 | EB1 binding | EB-1 Binding Domain |
| PF06098 | Radial spoke 3 | Radial spoke protein 3 |
| PF06705 | SF-assemblin | SF-assemblin/beta giardin |
| PF07202 | Tcp10 C | T-complex protein 10 C-terminus |
| PF07544 | CSE2 | RNA polymerase II transcription mediator |
| PF07781 | Reovirus Mu2 | Reovirus minor core protein Mu-2 |
| PF00307 | CH | Calponin homology (CH) domain |
| PF00091 | Tubulin | Tubulin/FtsZ family, GTPase domain |
| PF03953 | Tubulin C | Tubulin/FtsZ family, C-terminal domain |
| PF00004 | AAA | ATPase family associated with various cellular activities (AAA) |
| PF00018 | SH3 1 | SH3 domain |
| PF00036 | efhand | EF hand |
| PF00041 | fn3 | Fibronectin type III domain |
| PF00069 | Pkinase | Protein kinase domain |
| PF00400 | WD40 | WD domain, G-beta repeat |
| PF00433 | Pkinase C | Protein kinase C terminal domain |
| PF00435 | Spectrin | Spectrin repeat |
| PF00566 | TBC | TBC domain |
| PF00595 | PDZ | PDZ domain (Also known as DHR or GLGF) |
| PF00612 | IQ | IQ calmodulin-binding motif |
| PF00622 | SPRY | SPRY domain |
| PF00627 | UBA | UBA/TS-N domain |
| PF00642 | zf-CCCH | Zinc finger C-x8-C-x5-C-x3-H type (and similar) |
| PF00692 | dUTPase | dUTPase |

(*Continued*)

**Supplementary Table 1** (*Continued*)

| Accession | ID | Description |
|---|---|---|
| PF00838 | TCTP | Translationally controlled tumour protein |
| PF01302 | CAP GLY | CAP-Gly domain |
| PF01472 | PUA | PUA domain |
| PF01509 | TruB N | TruB family pseudouridylate synthase (N terminal domain) |
| PF01669 | Myelin MBP | Myelin basic protein |
| PF02149 | KA1 | Kinase associated domain 1 |
| PF02187 | GAS2 | Growth-Arrest-Specific Protein 2 Domain |
| PF02971 | FTCD | Formiminotransferase domain |
| PF02985 | HEAT | HEAT repeat |
| PF03451 | HELP | HELP motif |
| PF04961 | FTCD C | Formiminotransferase-cyclodeaminase |
| PF05091 | eIF-3 zeta | Eukaryotic translation initiation factor 3 subunit 7 (eIF-3) |
| PF05217 | STOP | STOP protein |
| PF05622 | HOOK | HOOK protein |
| PF06740 | DUF1213 | Protein of unknown function (DUF1213) |
| PF07058 | Myosin HC-like | Myosin II heavy chain-like |
| PF07145 | PAM2 | Ataxin-2 C-terminal region |
| PF07837 | FTCD N | Formiminotransferase domain, N-terminal subdomain |
| PF08068 | DKCLD | DKCLD (NUC011) domain |
| PF08239 | SH3 3 | Bacterial SH3 domain |
| PF08377 | MAP2 projctn | MAP2/Tau projection domain |
| PF08926 | DUF1908 | Domain of unknown function (DUF1908) |
| PF08953 | DUF1899 | Domain of unknown function (DUF1899) |
| PF08954 | DUF1900 | Domain of unknown function (DUF1900) |
| PF09041 | Aurora-A bind | Aurora-A binding |
| PF09336 | Vps4 C | Vps4 C terminal oligomerisation domain |




**Supplementary Figure 1** Sequence alignment of MAP20 proteins made with the MUSCLE alignment tool. The TPX2 domain is marked with a black bar and the extended TPX2 domain is indicated using dots. Notes above the TPX2 domain indicate where residues are lost relative to the Pfam domain model. Colors indicate similarity to consensus sequence.

**Supplementary Table 2** Protein sequences of gene models

| No | Gene model | Protein sequence |
|---|---|---|
| 1 | Pt195236 gw1.IV.325.1 | LNTKNPKSATPVKDRHGFQSKLSENSNPNLSHLSPCSKPTNSPSTKSQKSASKNPTLNPNPAIFSPRKKI<br>RERRFVVAKKNSKKETLNSNPTTVDCKCKERYGGSVKKCLCLAYETLRASQEEFFKNKNDVEEKDHLMDQ<br>NLDIEDREGSDAQYSCEIEKSGQMGSSKIKRRRNKLLEEARDSAPDNGKVKHLVEAFEKLFTLPNPKESD<br>RTEEEEIKENRKKAMQWALPGFQLPKTNVSSSSFCPSGFFLTSENLGLDTRISISSSWDGSQGSNSSRSS<br>NGGRRSRRNSAESGATTGGRRLKKKQIKITSQKPFKLRTEQRGRLKEEEFTKKIQEIMTEEEKLRIPIAQ<br>GLPWTTDEPECLIKPPVKENTRPIDLKLRSDIRAVERADFDHQVSEKMSLIEQYKMERERQQKLAEEEEV<br>RRLRKELVPKAQPMPYFDRPFIPRRSMKHPTMANEAKLRRHKKIKFCQSWNDVSSYSYDQQ* |
| 2 | Pt75311<br>fgenesh1_pg.C_LG_VII000075 | MAAESDDSSTATTATTTTSLTANATTSLKENTTTLMLVDEMYEFSAPKFYDFVKGESDEDSRNAELWFDVTAAYAPSPFMPRIKTGRSFKVETLCDFSQA<br>DQLHKVAESSDSKPTDSSSDSNSQSEVMPLPEPAASSEMRKEETSSNEENNANLINVISAGDVTCEKEKKVGFACAESNGRCTSSLQIENTDGKSSKNDAY<br>CTPKQPMSSRNRGILTDSKKNQSARHIASLVKNPSSVKPKGQSQSSRVKGIKPSSVKKDPNVKNVAGTANLAQENQAIKKQKLDGGRSRQIVNAKPPQPL<br>MHKSKLGLSSGNSNFCSSVPNKMQKVDRKVYVREQAAPAPFVSMAEMMKKFQSNTRELSLPHDGPASVIQRKPKLTLTRPKEPEFETAQRVRSVKIKST<br>AEIEEEMMAKIPKFKARPLNKKILEAPTLPALPRSTPHRPEFQEFHFVTAARANQNAESASVASTEVSCQSNQWKPHHLTEPKTPVLHTSLRRPAMVKSSL<br>ELEKEELEKIPKFKARPLNRKIFESKGEMGIFCHVKKQVTIPQEFHFATNERIPPASSVVDMFEKHMNKSNTCISSCWSSSQLSLRSEPTNENPIPRNTLPNPF<br>HLHTEERGAEKERKFVMELMQKQMEEERAFHRANPYPYTTDYPVIPPRPEPKPCTKAEPFQLESLVRHEEEMQREMQERERKEKEEAQMRIFRAQPVLK<br>EPIPLPEKVRKPLTQVQQFNLNADHRAVGRAEFDQKVKEKEMLYKRYREESETARMMEEEKALKQLRRTMVPHARPVPNFNRPFFPEKSSKETTNARSP<br>NLRVLQRRERRKMMVNAASSATASGMR* |
| 3. | Pt85496<br>fgenesh1_pg.C_LG_XIII000413 | MEKKPNGLAVKFNGVSHDRVHFAPKLSEGVIKAKEYVEKETAEESEKQDVLGVKSTNFDADVSDEKDEKPEAQKSSDDRNSSSPSLKAGGVGNAHVRQ<br>TVPQPFALATDKRVGRNTFTNSNNAQSPATMKNSQQNSPSTARKPLQPDNKKRHDEEDSWSVASSYPVRFEQHLSLAMFSTAASVRTVKSVTVGTAPTF<br>RSAERAAKRKEYYSKLEEKHRALEKERSQAEERTKEEQEAAIRQLRKNMAYKANPVPNFYYEPPPPKVERKKLPLTRPQSPKLNRRKSCSDAVQTSQEEV<br>GKHCARHRHSIGNHKDSTATSTAKAKVQISSQTANGIRKVTGRSKQERVTAEAVPEKTAEPTNADISVQS* |
| 4. | Pt355005<br>fgenesh1_pg.C_LG_VIII001919 | MGESLVAASSYEDKIGGTVASDPALQASVSFGRFENDSLSWDKWSSFSQNKYLEEVEKCATPGSVAEKRAYFEAHYKKIAARKAELLDQEKQIEHDLSRA<br>NNQNSGDLIVKTSQMDSDFDASNGQTSSEGIRPESKFDNEWDGGHIDKPTEDAAIDAHGQASTNKPYEDTAVDAHGQASSNDPYEDAAFSVHGQASLNE<br>PYEDAAIDVQGQVPLNGRVKEEQDSELDTPVSAKLEEVALMKKEETGSQDMRELPKNLEKEMESILMIKEEKVKLDHRKESPKISPMSKVRDLAMAKKK<br>PEPPITKRPQISSLKFSKPASTSSSLSASQSSIKKVNGSSLPRSKNTPVGGNKKVNPKSLHMSLSMDSPNSETVPLTTTRKSFIMEKMGDKDIVKRAFKTFQN<br>NFSQLKSSAEERSIGAKQMPAKEIGVKVSTSMTPRKENIGSFKSGGVDRRTAKLAPSSSVLKSDERAERRKEFSKKLEEKSKTEAESRRLGTKSKEEREAEI<br>KKPRRSLNFKATPMPGFYRGQKASKSPLDKVFAFGFPFLISVYMLVGCLFMDTKPDKIEKPIK* |
| 5. | Pt354840<br>fgenesh1_pg.C_LG_VIII001754 | MGESIVAASSYEDKIGGTAASDPALQVSVSFGRFENDSLSWEKWSSFSQNKYLEEVEKCASPGSVAEKKAYFEAHYKKIAARKAELFDQEKQMEHESSM<br>ENNHNIGDLTGKNGQTDSSFDVSNGQTSAEGIWHESKLDNERDGGHVDEPYEDAAIDVHGQASLSGLYEDAANDVQSQASSNGRVKEELENKLDSPEST<br>KLEELALIKEEEKGYQDTRELPKNSEKEKESILMIKEEKVKFDHQRGSSKIIPLSKVRDIARAKKKPEPLVTKQPQISTPKVSKRVPTSSSLSASQSSTKKMN<br>GSLLPRSKNPPAGENKKVTSKSLHLSLTMDPSNSEPDPLITTRKSFIREKMGDKDIVKRAFKTFQNNFSQLKSSAEERAIREKQVPAKGTDVKVSTSMTPR<br>TENIVSLKSSRVDRKTAKLAPSSFVLKSDERAKRRKELSMKMEEKSNAKPAESTHLRTKSKEEKEEEIIKQRHSSNFKPTPMPGFYRAQKASKSPLDKVCP<br>VPESCHTLMK* |
| 6. | Pt564607 eugene3.00081202 | MCFKISFYTQNSESGNSRIHPVNKRYSPFLANCKPASSMVSDFKLSIFLCSEKEPRVKKFESPSSRSKRVEPIAHLSTNRTKQNASSINPDTRPSASTFSFKSDER<br>AERRKEFYMKLEEKWHAKEAEMNQIQAKTQEKTEAEIKQFRKSLNFKATPMPSFYHVAVPPASNGNKVQTVQ* |
| 7. | Pt564928 eugene3.00081523 | MGIEVTDICMDKESDSVIVYSNGVSHDQTHETVPHDHGVLESYEPINGVPELHSSEESTEAKEYEVKECTTEVSVEVTELSHAEKSKEGQHVVCSNFEDGLK<br>VKKVKAVNRKSKDIGQQKSSIKRVSKPASAAIARTKHTVPQPFALATEKRASSGIRPSGPEPDITNGVNKSFKANNVLRQNPMKQNQPLSVSRKPLQPNNKK<br>HPDEEDNCSVTSSTTASARPIMSKATAVAAPVFRCTERAEKRKEFYSKLEEKYQALEAEKTQSEARTKEEKEAAIKQLRKSLTFKANPMPSFYHEGPPPKVE<br>LKKLPPTRAKSPKLGRRKSCNNRVNSSQPDKVKRDFNDEKNQSQDSSREDTSNPVSQHSVLKGHAISKFEDETQQAEEINE* |
| 8. | Pt566133 eugene3.00100691 | MGIEVTDVCMDKEPNCVIVYSNGVSHDPTHETVPDDHGVLESYEPINGVPELHSSEESTEAKEYVVKECTTEVSVEVTELSHAEKSKEDQTVVCSNFEDGLK |

(*Continued*)

**Supplementary Table 2** (*continued*)

| No | Gene model | Protein sequence |
|---|---|---|
| | | VEKVKALNRKSKDIGQKKSSTKHASKPAPAGLARTKHTVPQPFALATEKRASLGMRPSGEPDITNGLNKSFKANNALRPNPIKQNQPLSVSRKPLQPNNKKH<br>PDEEDNCSVTSSYPVSSILTSARPAKSKPAAVAAPVFRCNERAEKRKEFYSKLEEKHLALEAEKTQSEARTKEEKEAAIKQLRKSLMFKASPMPSFYHEGPPP<br>KVELKKLPPTRAKSPKLGRRKSCSNGVNSSQPDRVKGACGDGNNQSQGIFREDTSNPVSQHSIPKGHVICKFEDETREMEGIDELIPLEVSGQSFAGIGLQS* |
| 9. | Pt578210 eugene3.00180434 | MDSDYHLFPDDGLETVHQNGVHEQSAAAREDGVVSNNLSGSMGNTFEVDDCTNDNLSTREVEGELKEGEAKVKDADNSEKARSQKGSGKGGNAKP-<br>SNPK<br>NVSATQVKGKDGRDAVARTAVSNGSVAVNSQLKQSLKSNSFNERQGQASKQSGKSDAVLSAGLVEKAKPLKKGPVVKAEGETESTSSPTAEDAKSRKFGT<br>LPNYGFSFKCDERAEKRKEFYTKLEEKIHAKEVEKSTLQAKSKETQEAEIKLFRKSLAFKATPMPSFYQEPAPLKVELKK*IPTT*RAKSPKLGRKKSPSPADSEG<br>NNSQSNRSGRLSLDEKISSKIPIRGLSPAHPKKPQRKSLPKLPSEKINLYANDEKGKLPKASNEENTTLSDQTNEGVSANQEQEAVSKNEASEFLPPKEEVVVQ<br>EEAATLMKGPIALAV* |
| 10. | estExt_fgenesh4_pg.C_440200 | MEKAHTKSALKKLVKASSQSAPWSNAARGMAKDDLKDPLYDKSKVAPKPFAKENTKPQEFKLHTGQRALKRAMFNYSVATKIYMNEQQKRQIERIQKIIE<br>EEEVRTMRKEMVPRAQLMPYFDRPFFPQRSSRPLTVPREPSFHMVNSKCWSCIPEDELYYYFEHAHPHDHAWKPVK* |
| 11. | Pt594654 eugene3.00640242 | MAAESEDSSTATTMTNATTLMMFVDEAYEFSAPKFYDFVKGESDEDSRNAELWFDVTASYAPSPFMPRIKTGRSFKVETLCDFSQADQFHKVAESSDSKAS<br>DSSSQSEVMPPPAEAAAPIGTGKEEKTSDEDNKENNANLVNVISAGEVTCEKEKKVGFACAEGNERSTSSLQTENADGKESSKNEAYCTPKPPMSSRNRGPL<br>TDSKKNHSARHIASLVRNPSLLKPKSQSQSSQVKGIKPASVKKDRNVKNVAGTTNLAQENQAIKKQKLEGGRSRQILNAKPPQPLTHKSKLGLSSGSSNLCS<br>SVANKMKKEERKVYVREQAAPGPFVSTAEMMNKFQSNTRGLSMPRFNNSISHDGPASVIQRKPKLTLTRPKEPEFETAQRVRSVKIKSSAEIEEEMMAKIPK<br>FKARPLNKKILEAATLPALPRSTPQPPEFLEFHLETAARANQNAESTSVASTEVSHQSNLWKPHHLTEPKTPVLHTSLRARPARVKSSLELEKEEIEKFPKFKA<br>RPLNKKIFESKGAMGIFCHAKKQVTVPQEFHFATNERIPPQAAVADMFDKLSLRSEPILDNPIPRNTKPNPFHLHTEERGAEKERKFWMELVQKQMEEERAR<br>VPRANPYPYTTDYPVVPPRPEPKPCTKPEPFQLESLVRHEEEMQREMEERERKEKEEAQMRIFRAQPVLKEDPIPVPEKARKPLTQVQQFNLHADQRAVERA<br>EFDHKVKEKEMLYKRYREESETAKMMEEEKALKQLRRTMVPHARPVPNFNHPFCPQKSSKEATKAKSPNLRVLQRRERRKMMMVNAASSAAASGMR* |
| 12. | Pt595399 eugene3.00700040 | MGRELQRADMEKKPNGLAAKFHGVSHDKVHISPKLSKAVIEAKEYVEKETAEKSEKQDVLGVKSTNFDADPSDGKDEKPGAQKLSDDKNSSSPSQKIGVN<br>GNKHAHRAHQTVPQQFALATDKHVGGNSSTNSNKTQSPVSMKNSQQNSPSTARKPLHPDNKKHHDEEDSWSVTSSTSASVRTVKSVTVGTAPTFRSSERA<br>AKRKEYYSKLEEKHRALEKERSQAEARTKEEQEAAIKQLRKSMLYKANPVPSFYHEPPPPQVELKKLPLTRPQSPKLNRRKSCSDAVRTSQEEVGKHCARH<br>RHSIGSHKDSTGANTAKAKVQISSQTANGIRKVKDRSKQDHVATKADPEKIAGPTNADISVQS* |
| 13. | Pt653406 grail3.0030003201 | SRANRLAPTGANSKESNINGSKTLTKQTSSTSKSSSQQAASVKSSSLTEAAKCPPPQVSESAADQNSKPETTTFSSKEEDDTHSTTSSATLSGRRSSGSGFSFR<br>LEERAEKRKEFFSKLEEKIHAKEIEQTNLQAKSKESQEAEIKKLRKSLTFKAAPMPCFYKEPPPKVELKK*IPTT*RAKSPKLGRRKSSTTSMNNSLEDVGSSFSP<br>RASHSPHLNQESSNPTKGAQRNGNVDNGASKTPIRKSQPKHQSRQITANGMEGKTVKSKAKLPGAESQTQKANVEKVEVNENNSMKVPVCENGIETMPE<br>NNTPQNNGPVLSSSNPEIMLPHVTVGG* |
| 14. | Pt415809 gw1.VI.182.1 | KVKDADNSENAKSQKGPGKRGTAKPSHLKNASATQVKKGKDGRDAEVQLTVSNGSVAVNSQLKQHLKSKS<br>FNERQGQASKQSGTSDAGPPEGIVEKMKLKPLKKGPVDKAEADTDSTSSPTVEDAKPRKVGALPNYGFSF<br>KCDERAEKRKEFYSKLEEKIHAKEVEKTTLQAKSKETHEAEIKMLRKSLGFKATPMPSFYQEPTPPKVEL<br>KKIPTTRAKSPKLGRRKSSSPADTEGNNSQSYRPGRLSLDEKVSSNIPIKGLSPAHPKKPQRKSLPKLPS<br>EKTKL* |
| 15. | Pt658207 grail3.0036016201 | MPEEKMPNKVNNHPNQEAAEMENVALPNNKRQMSSLSNSLSQSRASKLPKSSAKLSSSTRLNATPNSKKSAGELVGEKRATSKSIHVSINFASRLQDTNK<br>SYVRVSKDRSATPENPTRGSVHGVSKLLPLIFRHSQDRRSKSELNKSVSGKITPGEISQTLSSDCSKSSSAKGSKSRPPLISSPFSFRSEERVAKRKEFFQKLGE<br>KNNAKEDTEKKHLHARPKEKAEHDLKKLRQSAVFRGKPSDDLHRGLHSPYNSMKKIPLTRPQSPKLGRKSTPNAVREASLQLHQRPSVNAETSKPFIQKS<br>NHSSTCPVNLLPKKKALENASPNILW* |
| 16. | Pt658783 grail3.0006023901 | MRSPINGSQFQKILNNVSKTTAKTQNRGEGETPQRAKSEKQSSRATTPTRRTLHRAKNEENSESGNLRLHPVNRSERASRVNKFESPPSRSKKVEPMSHLR<br>ANRNKQIVNSIKPDTMPCAAAFSFKSDERAERRKEFYMKLEEKLHAKEAEMNQIQAKTQEQKKAEIKKFRERLNFKAAPMPSFYRVAVSPGSDGNK |

| | | |
|---|---|---|
| 17. | Pt667581 grail3.0004025702 | RNRCSFPHISVTSSPRLNQANRRVPTGVNSKESNINCSKTLTRQSSSAGKSCSQQATSVKSSSLNEAAKGHPPQASESAAHQNSKPETTTLSSKEDDDTHSTT SSATPSGRRSSGSGFSFRLEERAEKRKEFFSKIEEKIHAKEIEQTNLQEKSKENQEAEIKQLRKSLTFKATPMPSFYKEPPPKAELKKI*PTT*RAISPKLGRRKSST TLTNNSLEDSGSSFSPRASHSPRLNQESSNPTKGIQRNGNKDNGASKTPIRKSQPKLQSHQIMANGLEGKTVKSKAKPPGAENQTQKAGVGKVEENENNSK KIPLCDNGIQTMPENNTPQNDGLVLSSSNPEFMLPQVTVGG* |
| 18. | Pt668651 grail3.0059011801 | MGLRRLKKKKQLKVTSQKPFKLRTEQRGRQKEEEFTKKIQEIMMEEERLRIPVAQGLPWTTDEPECLIKPPVKENTKPVDLKLHSDIRAVERADFDHQVSE KMSLIEQYKMERERQRKLAEEEEIRRLRKELVPKAQPMPYFDRPFIPRRSIKHPTVPREPRFHMPQHKKIKCCLSWSNV |
| 19. | Pt691617 estExt_fgenesh1_pg_v1.C_LG_ I2234 | MGDTTCVMQPFSYAAGISNDAKEGNPIHALGQSISFGRFMSDSLSWEKWSSFSHNRYVEEAEKFSRPGSVAQKKAFFEAHYRNLAARKAAALLEQANAE ANNVQEPENEGGIHDKTTQDSLTVATNSQEAGDREEVHVQQVNCEASFVADDNTRTSNVDMERFESSNVEEVEPSAENEILVENCVKNETLNQIVKVDN KEEVKEMELSVSKQMEKPLLKDFMSCKDDAASMSKKKPAVSSSKSSIYDKASKLPSTPAKPAPSVRAKKENTATPISKKSALESVERRKPTPKSTHKSMN FTPAREFNRITSSIIRKIDNSRVGSHSKSSKDCPTPSRTPMMMVSIAESKHPLATPQSEKRRAKTPLHPSTSGSKTVRSKWHFLPKDCSMFMTSSRNRSQSPS ASIPFSFRTEERAARRKEKLEEKFNAYQAQKVQLQVTLKEKAETELKRLRQSLCFKARPLPDFYKQRVAPNNQMEKVPLTHSESPEPGRKMTPSKIRSASQ LPQWSSLKNSGSKDAMQKKSDNPRSLASRLKASPHENTSPNIQHE* |