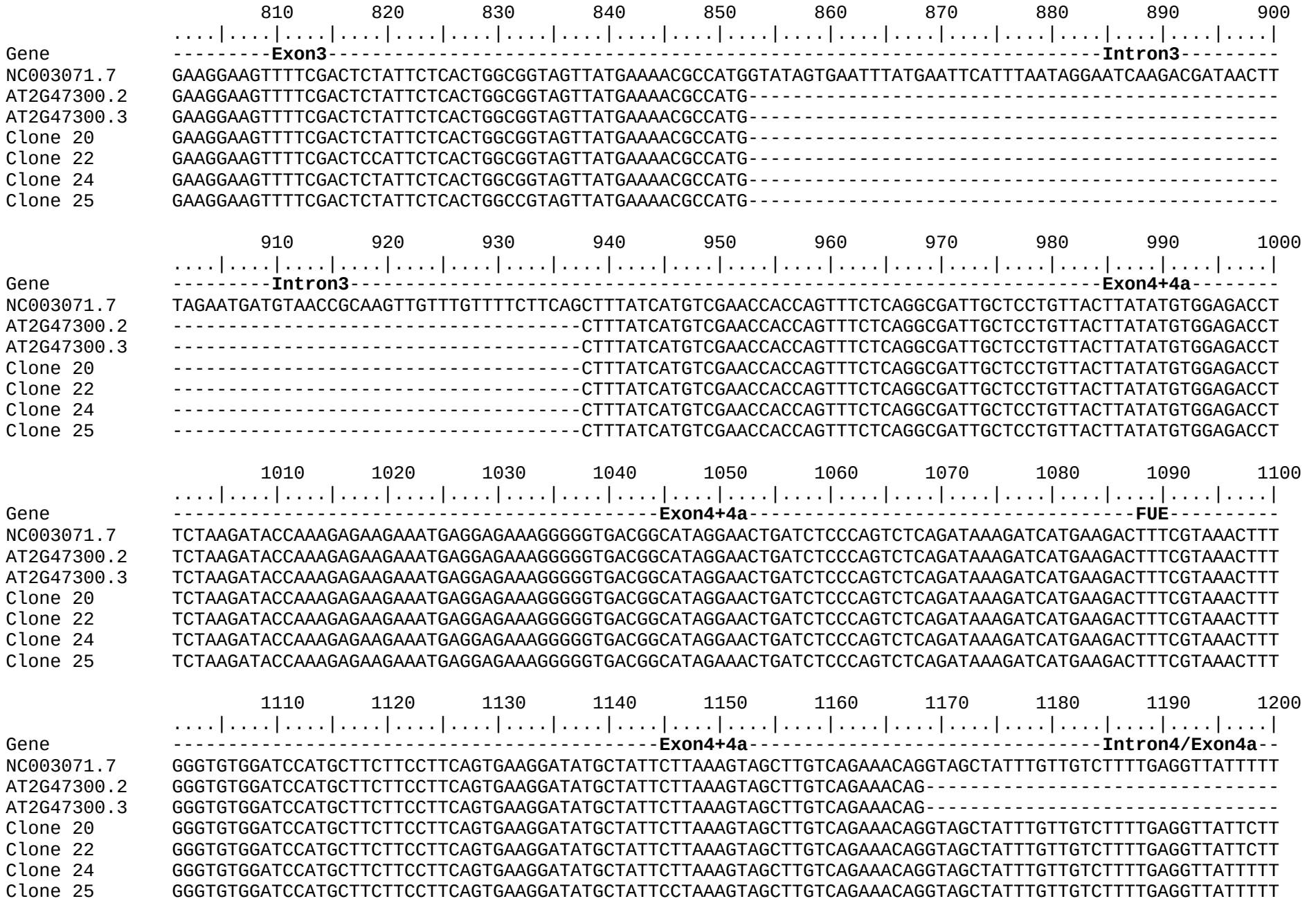
Supplementary sequence information (Krehan et al., RNase MRP RNA ...)

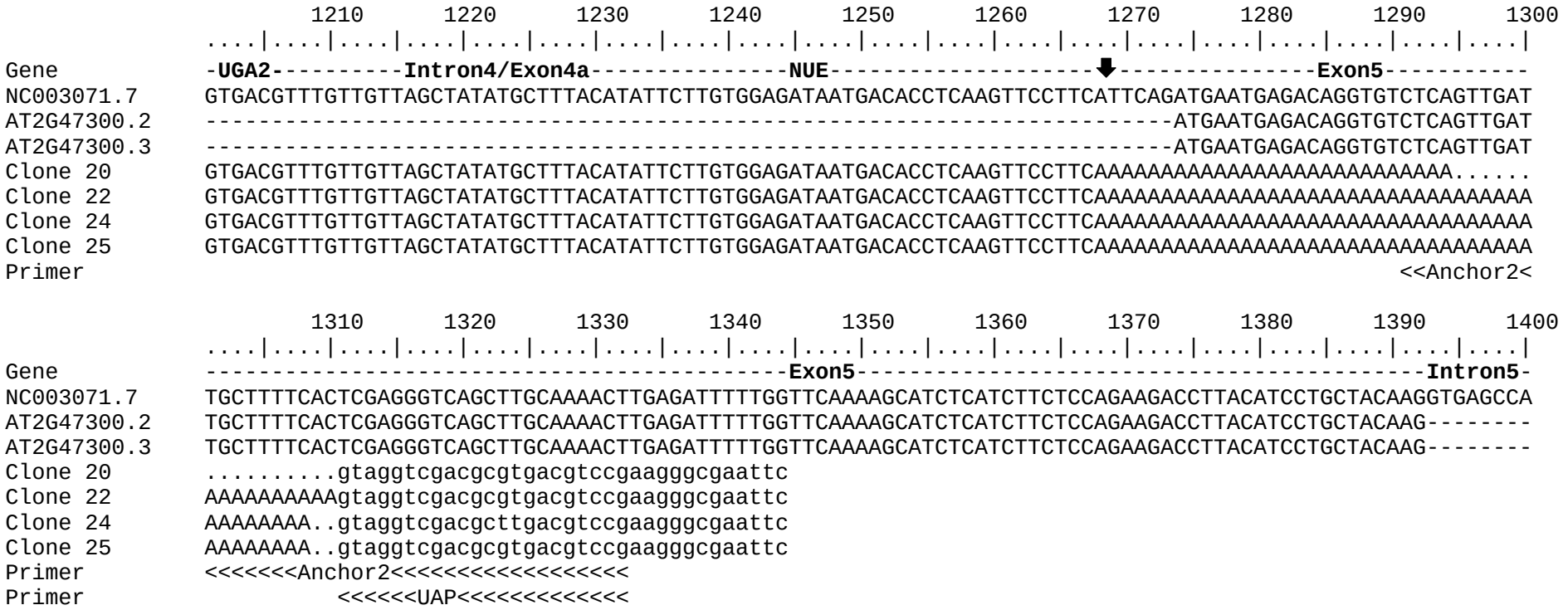
Accessory sequence alignment: Analysis of AtPop1 mRNAs by RACE

```
                 510       520       530       540       550       560       570       580       590       600
               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Gene           ---------Intron1-----------------------------------------------UAG1---------------------Exon2-----------
NC003071.7     GTTAAGAGTCTTTTGAAATCTTAAGGTTGTTTTTGTCATGGTTCTAGAGGAAGGGGATCTAGGGATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
AT2G47300.2    ---------------------------------------------AGGAAGGGGATCTAGGGATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
AT2G47300.3    ----------------------------------------------GAAGGGGATCTAGGGATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
Clone 20       ----------------------------------gaattcgcccttaggaaggggatctagggATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
Clone 22       ----------------------------------gaattcgcccttaggaaggggatctagggATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
Clone 24       ----------------------------------gaattcgcccttaggaaggggatctagggATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
Clone 25       ----------------------------------gaattcgcccttaggaaggggatctagggATGTCTTGAAGCAGTCTCGGCAAGGTGTTCTTGTTC
Primer                                                >>>>>POP1-3'ss>>>

                 610       620       630       640       650       660       670       680       690       700
               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Gene           AUG2-----Exon2-------------------------------------------------UGA1---------------Intron2---------
NC003071.7     ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAGGCAGGCATTCACACTTATTCTTGATATCCTTCTATAAGTTTTTTTCAACTTGGTTG
AT2G47300.2    ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAG-----------------------------------------------------
AT2G47300.3    ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAGGCAG-------------------------------------------------
Clone 20       ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAG-----------------------------------------------------
Clone 22       ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAG-----------------------------------------------------
Clone 24       ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAG-----------------------------------------------------
Clone 25       ATGATGCAAGCTATCACATTGCTGTGCAATTGGAGGGTCCAGAG-----------------------------------------------------

                 710       720       730       740       750       760       770       780       790       800
               ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
Gene           ---------------------------------------------------------------------------------Exon3-----------
NC003071.7     TGTGTTCATATTTTCTTCATGTCTTTTGGCTTTGAATTTCTCAGGGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
AT2G47300.2    ---------------------------------------------GGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
AT2G47300.3    ---------------------------------------------GGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
Clone 20       ---------------------------------------------GGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
Clone 22       ---------------------------------------------GGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
Clone 24       ---------------------------------------------GGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
Clone 25       ---------------------------------------------GGCTCACTCTTATCGATATTAAATATGCTACTAGAGCCTTCTCCGTCATCTCATTC
Primer                                                                          >>>>>>POP1-GSP3F>>>>>
```


810 820 830 840 850 860 870 880 890 900
Gene ---------Exon3----------Intron3---------
NC003071.7 GAAGGAAGTTTTCGACTCTATTCTCACTGGCGGTAGTTATGAAAACGCCATGGTATAGTGAATTTATGAATTCATTTAATAGGAATCAAGACGATAACTT
AT2G47300.2 GAAGGAAGTTTTCGACTCTATTCTCACTGGCGGTAGTTATGAAAACGCCATG----------
AT2G47300.3 GAAGGAAGTTTTCGACTCTATTCTCACTGGCGGTAGTTATGAAAACGCCATG----------
Clone 20 GAAGGAAGTTTTCGACTCTATTCTCACTGGCGGTAGTTATGAAAACGCCATG----------
Clone 22 GAAGGAAGTTTTCGACTCCATTCTCACTGGCGGTAGTTATGAAAACGCCATG----------
Clone 24 GAAGGAAGTTTTCGACTCTATTCTCACTGGCGGTAGTTATGAAAACGCCATG----------
Clone 25 GAAGGAAGTTTTCGACTCTATTCTCACTGGCCGTAGTTATGAAAACGCCATG----------
910 920 930 940 950 960 970 980 990 1000
Gene ---------Intron3----------Exon4+4a--------
NC003071.7 TAGAATGATGTAACCGCAAGTTGTTTGTTTTCTTCAGCTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
AT2G47300.2 ----------CTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
AT2G47300.3 ----------CTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
Clone 20 ----------CTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
Clone 22 ----------CTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
Clone 24 ----------CTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
Clone 25 ----------CTTTATCATGTCGAACCACCAGTTTCTCAGGCGATTGCTCCTGTTACTTATATGTGGAGACCT
1010 1020 1030 1040 1050 1060 1070 1080 1090 1100
Gene ----------Exon4+4a----------FUE----------
NC003071.7 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGGAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
AT2G47300.2 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGGAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
AT2G47300.3 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGGAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
Clone 20 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGGAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
Clone 22 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGGAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
Clone 24 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGGAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
Clone 25 TCTAAGATACCAAAGAGAAGAAATGAGGAGAAAGGGGGTGACGGCATAGAAACTGATCTCCCAGTCTCAGATAAAGATCATGAAGACTTTCGTAAACTTT
1110 1120 1130 1140 1150 1160 1170 1180 1190 1200
Gene ----------Exon4+4a----------Intron4/Exon4a--
NC003071.7 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCTTAAAGTAGCTTGTCAGAAACAGGTAGCTATTTGTTGTCTTTTGAGGTTATTTTT
AT2G47300.2 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCTTAAAGTAGCTTGTCAGAAACAG----------
AT2G47300.3 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCTTAAAGTAGCTTGTCAGAAACAG----------
Clone 20 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCTTAAAGTAGCTTGTCAGAAACAGGTAGCTATTTGTTGTCTTTTGAGGTTATTCTT
Clone 22 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCTTAAAGTAGCTTGTCAGAAACAGGTAGCTATTTGTTGTCTTTTGAGGTTATTCTT
Clone 24 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCTTAAAGTAGCTTGTCAGAAACAGGTAGCTATTTGTTGTCTTTTGAGGTTATTTTT
Clone 25 GGGTGTGGATCCATGCTTCTTCCTTCAGTGAAGGATATGCTATTCCTAAAGTAGCTTGTCAGAAACAGGTAGCTATTTGTTGTCTTTTGAGGTTATTTTT



Legend to Accessory sequence alignment: 3' RACE analysis of AtPop1 mRNAs

Gene: gene structure (mRNA processing pattern and translational start/stop codons). Numbering starts at AUG1 (see supplementary figure 1), but only the region analysed by RACE is shown here. **Primer:** position of primers used for 3' RACE and sequence analysis; primer length and orientation is indicated by carets. For graphical clarity, the prefix "At" is omitted from primer names.
Symbol characters: ⬇ cleavage and polyadenylation site from this study. FUE, far upstream element; NUE, near upstream element.
RNAs are shown as cDNA sequences and intron positions as dashed lines. Variations in poly(A) length are compensated by dots.
Sequences: NC003071.7: *A. thaliana* Chr.2, genomic region encoding Pop1p;
AT2G47300.2 (TAIR): 2908 nt mRNA variant 2a (synonymous to NM_001084603.1 / DQ069804.1; 826 aa ORF);
AT2G47300.3: 2910 nt mRNA variant 2b (synonymous to NM_001084604.1 / DQ069805.1; 659 aa ORF).
Clone 20/22/24/25: individual clones obtained from the 3' RACE experiment, leading to the novel mRNA variant 3 identified in this study (FN673552; 307 aa ORF starting at AUG1). Sequences derived from amplification primers and vector termini are shown in lower case.

Methods:
The 3' termini of AtPop1 mRNAs *in vivo* were identified by RACE, using nested PCR of cDNAs primed with $dT_{16}V$ as described in the main text. Four independent clones were sequenced completely from both strands, using the universal M13 forward and reverse primers and the internal GSP3F primer. No attempts were made to experimentally verify the annotated long mRNA variants 2a and 2b, because they are supported by plenty of EST data (82 independent entries in TAIR).

Results:
All clones representing the novel AtPop1 mRNA variant 3 extend into the annotated intron 4, thus creating exon 4a. The contiguous reading frame is closed by an in-frame stop codon (UGA2) at position 1202 within this region. The polyadenylation signals for this mRNA are located at positions 1088-1217 (FUE), 1245-1250 (NUE) and 1268 (CS) and conform to the consensus suggested by Loke et al. (2005). An oligo T-stretch which may facilitate transcription termination (52) is present at the distal end of intron 5. The complete sequence of mRNA variant 3 (FN673552) was assembled from several 3' and 5' RACE clones as described in the main text and is shown in supplementary figure 1, together with variants 1, 2a and 2b (corresponding to AT2G47300.1, 2 and 3).

Discussion:
The novel mRNA 3 is created by retention of intron 4, which provides an in-frame stop codon for the ORF starting at AUG1. The RNA terminates close to the annotated 3' splice site and shows all signals required for 3' end processing: the CS at position 1268 is embedded in a cleavage element and preceded by a FUE and NUE. These signals show good agreement to the polyadenylation signal consensus, which is not strictly defined in plant mRNAs (50). All clones are polyadenylated at the same site, indicating that this mRNA is correctly processed *in vivo*.
The 3' termini of the published long mRNA variants 2a and 2b have been defined at a single cleavage/polyadenylation site (see supplementary figure 1, position 3510). However, examination of published ESTs revealed that three additional upstream polyadenylation sites are used at approximately equal distribution (2 or 3 clones/site), as is frequently found in plant mRNAs. Transcription termination may be facilitated by the T-rich regions starting at positions 3351 and 3468.
Of the over 80 independent ESTs analysed, 9 extend to the 3' end within exon 10. The exclusive presence of the long mRNAs in EST databases (TAIR) may be due to these multiple polyadenylation sites; in addition, the genomic sequence immediately downstream of the most distal (and only annotated) poly(A) site contains several A-stretches, which may have led to amplification of partially processed mRNAs and thus to a bias in EST synthesis. Taken together, these points may explain the lack of EST data for the novel mRNA 3.

References:

50. Loke,J.C., Stahlberg,E.A., Strenski,D.G., Haas,B.J., Wood,P.C. and Li,Q.Q. (2005) Compilation of mRNA Polyadenylation Signals in Arabidopsis Revealed a New Signal Element and Potential Secondary Structures. Plant Physiol. **138**, 1457–1468.
52. Hasegawa,K., Yukawa,Y. and Sugiura,M. (2003) *In vitro* analysis of transcription initiation and termination from the Lhcb1 gene family in *Nicotiana sylvestris*: detection of transcription termination sites. *Plant J*, **33**,1063-1072.