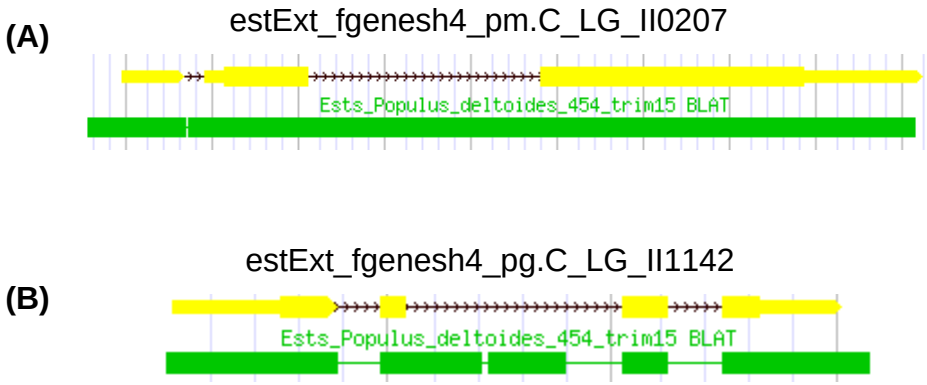
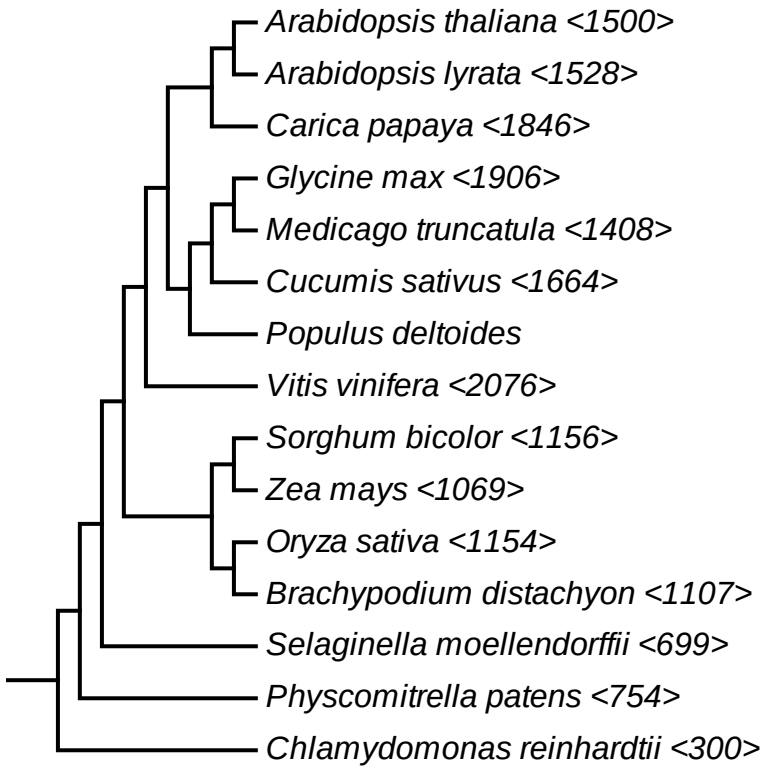
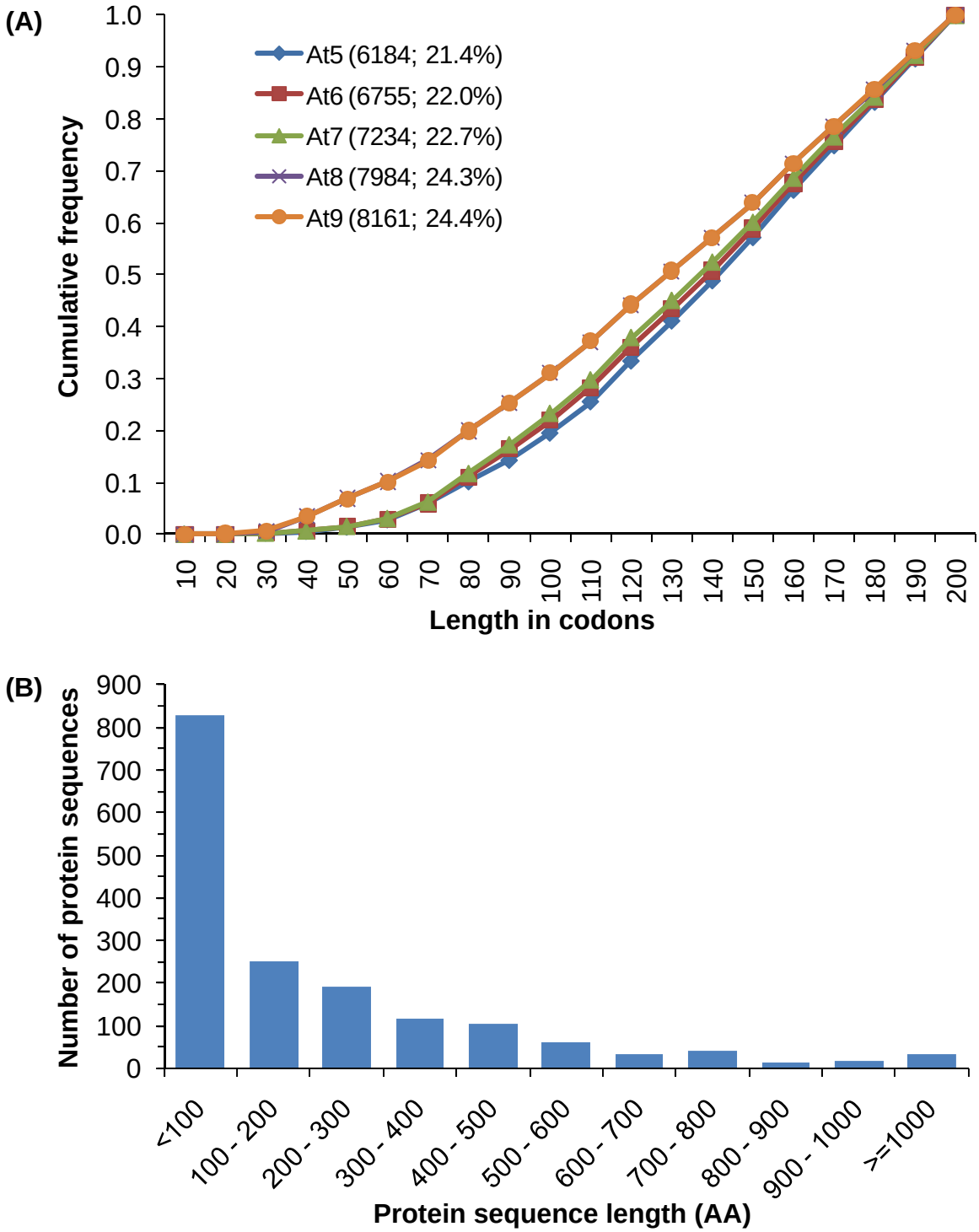


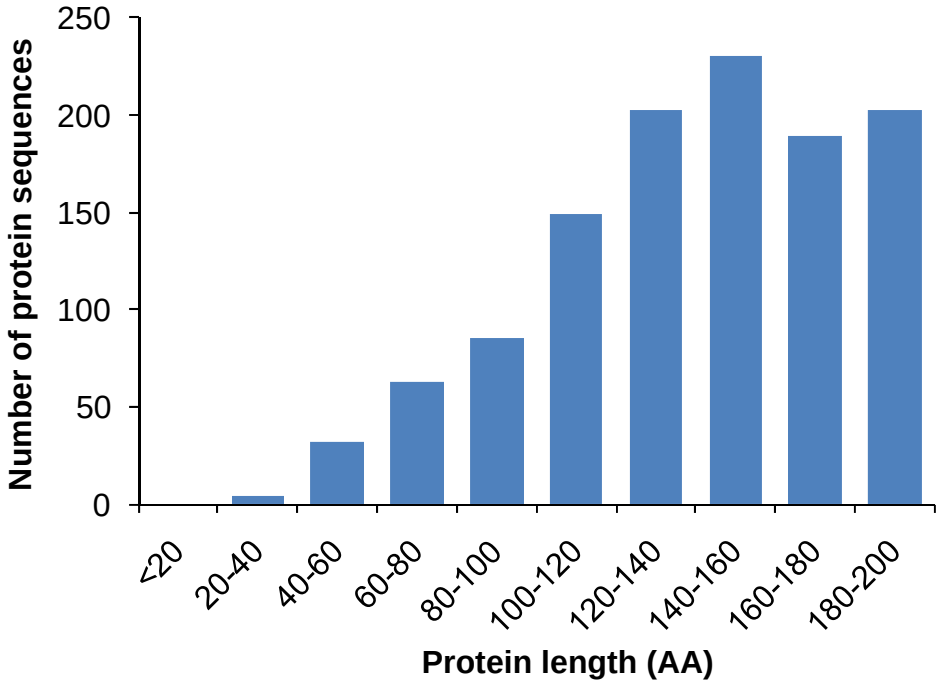
Supplementary Figure 1. Examples showing coverage of full-length *P. trichocarpa* gene models by transcriptome sequencing reads. The yellow represents the annotated gene models estExt_fgenesh4_pm.C_LG_II0207 (**A**) and estExt_fgenesh4_pg.C_LG_II1142 (**B**) in the *P. trichocarpa* genome annotation v1.1 (http://genome.jgi-psf.org/poplar/poplar.home.html). The green represents assembly of ESTs obtained using 454 sequencing technology in this study. The closed boxes indicate exons and the solid lines indicate introns.



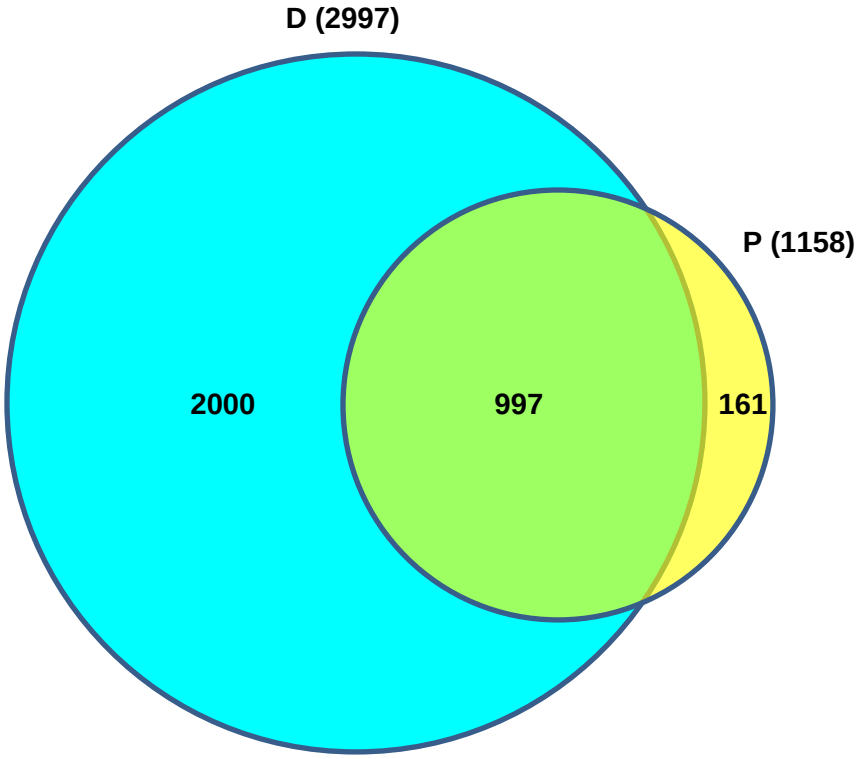
Supplementary Figure 2. The phylogenetic relationship between the plant genomes used to identify conserved genomic sequences containing sORFs. The numbers in brackets indicate the number of sORFs conserved between each species and *P. deltoides*. The phylogenetic relationship was adapted from http://www.phytozome.net/.



Supplementary Figure 3. Length distribution of predicted small protein sequences in *Arabidopsis* genome annotations. (**A**) Small proteins (<200 AA in length) in annotation versions 5–9 (At5–At9); the numbers in parentheses represent the number of small protein-encoding genes and their corresponding percentage of the whole-genome annotation. (**B**) New protein sequences added into *Arabidopsis* genome annotation version 8 (At8).



Supplementary Figure 4. Size distribution of sORF-encoded proteins identified in the *initial sORF candidate set* by mass spectrometry.



Supplementary Figure 5. Venn diagram showing the overlapping between the sORF subset containing known protein domains (**Subset D**) and the sORF subset with proteomics support (**Subset P**) in the *initial sORF candidate set*. The value in parentheses represents the number of sORFs in each individual subset.