Discovery and annotation of small proteins using genomics, proteomics and computational approaches

Contact: Xiaohan Yang, (865)241-6895, yangx@ornl.gov

Funding Source: DOE Office of Biological and Environmental Research, Genomic Science Program, and DOE BioEnergy Science Center

- *Ab initio* discovery of small proteins (10-200 amino acids in length and) is often overlooked.
- ~2.6 million expressed sequence tag (EST) reads from *Populus deltoides* leaf were obtained by deep RNA sequencing and used to reconstruct full-length transcripts. 12,852 short open reading frames (sORFs) encoding proteins of 10–200 AA in length were identified.
- Three computational approaches were used to enrich for *bona fide* protein-coding sORFs: 1) coding-potential prediction, 2) evolutionary conservation between *P. deltoides* and other plant species, and 3) gene family clustering within *P. deltoides* (Fig. 1A). Sequential filtering led to larger numbers of matches between the sORFs and proteomics measurements (Fig. 1B).
- Deep RNA sequencing, in combination with computational approaches to predict high-likelihood protein-encoding sORFs, indicates that potential sORF candidates remain to be annotated in sequenced genomes, and presents an efficient strategy for discovery of sORFs in species for which no genome annotation is yet available.

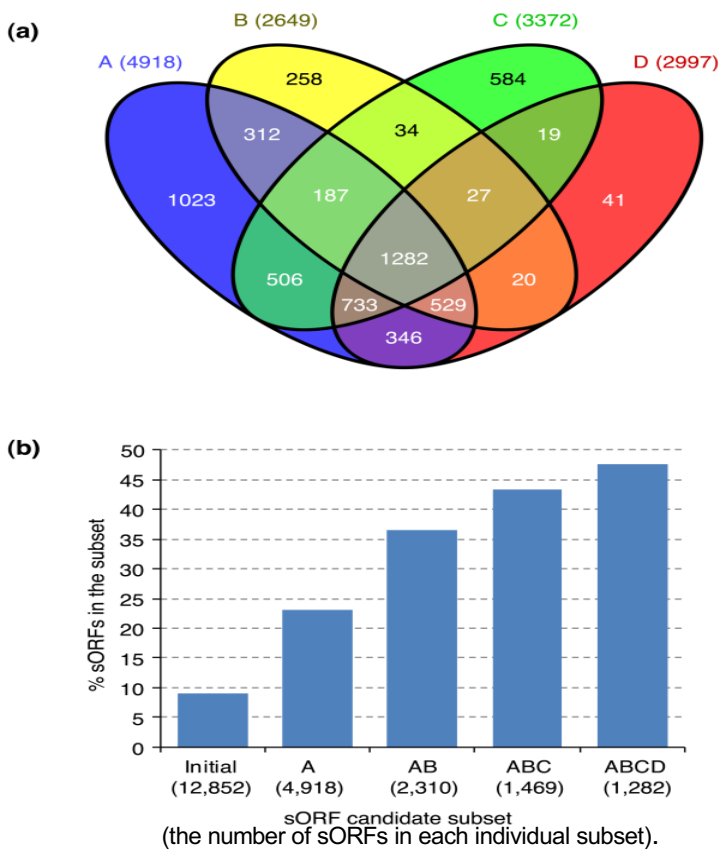

Figure 1. *Populus deltoides* small protein-coding candidate genes enriched from transcription units. (a) The number of sORF candidates in four different subsets and their intersections. (b) Proportion of the sORF subsets having protein mass spectrometry data support. Subset A: sORFs with high protein-coding potential. Subset B: sORFs conserved between *P. deltoides* and at least one other plant species. Subset C: sORFs clustered into families. Subset D: sORFs with known protein domains. “AB”, “ABC”, “ABCD”: the different intersections of the Subsets, respectively.

Yang, X., Tschaplinski, T.J., Hurst, G.B. et al. 2011 Genome Research (doi:10.1101/gr.109280.110)