


Brief Communication

PsORF: a database of small ORFs in plants

Yanjun Chen¹, Danyang Li¹, Weiliang Fan^{1,2}, Xiaoming Zheng¹, Yifan Zhou¹, Hanzhe Ye¹, Xiaodong Liang¹, Wei Du¹, Yu Zhou^{1,2} and Kun Wang^{1,*} 

¹College of Life Sciences, Wuhan University, Wuhan, China

²State Key Laboratory of Virology, Wuhan University, Wuhan, China

Received 13 December 2019;

revised 14 April 2020;

accepted 18 April 2020.

*Correspondence (Tel +86 27 68754887; fax: +86 27 68754887; email: wangk05@whu.edu.cn)

Keywords: database, small ORF, plant, Ribo-seq, mass spectrum.

Dear Editor,

Small open reading frames (sORFs) which are translated to small peptides (100 amino acids or fewer in length) have been always excluded from genome annotations. In recent years, more and more biologically significant sORFs have been discovered to encode functional peptides or play regulatory roles on mRNA translation. In plants, an evolutionarily ancient micro-peptide, AtLURE1, promotes and maintains reproductive isolation through accelerating conspecific pollen tube penetration (Zhong *et al.*, 2019). The sORFs in the 5' UTR of mRNA, usually named as upstream ORFs (uORFs), were reported to mediate translational regulation of their downstream main ORFs (mORFs) (Xu *et al.*, 2017).

Recent advances in translomics (especially the ribosome profiling, Ribo-seq) and MS-based proteomics have indicated that sORFs were pervasively present in non-coding RNAs, UTR regions of mRNAs, and circleRNAs etc (Wang *et al.*, 2019). In animals, there have been two public databases for sORF collection: SORFS.ORG (Olexiouk *et al.*, 2016) and smProt (<http://bioinfo.ibp.ac.cn/SmProt/>) (Hao *et al.*, 2017). The two databases integrated Ribo-seq and MS-based proteomics data in animals to annotate the sORFs. In plants, a database ARA-PEPs (<http://www.biv.kuleuven.be/CSB/ARA-PEPs>) has been constructed (Hazarika *et al.*, 2017). The ARA-PEPs identified sORFs based on criteria that the peptide sequences of at least 10 amino acids beginning with a canonical start codon and not truncated by a stop codon. It is a repository only for sORFs in *Arabidopsis thaliana*, in which the 13 748 candidate sORFs lack translational evidence, but have only RNA expression evidence (microarray and RNA-seq). Therefore, a database of systematic sORF annotations in plants is still missing, which will not only hinder cross-species studies in plants, but also restrict the possibility of cross-kingdom comparison analysis between animals and plants.

In this study, we collected multi-omic data including genome, transcriptome, Ribo-seq and mass spectrum (MS) from public database, and built a pipeline to identify sORFs in 35 different plant species. Based on the results, we designed a web-accessible database, PsORF (<http://psorf.whu.edu.cn/>).

The PsORF integrates released data from multiple databases to acquire a set of sORFs generated from non-coding region annotated in reference genomes. We collected 35 reference genomes from PLAZA database (<https://bioinformatics.psb.ug>

nt.be/plaza/) with well-annotated UTRs and lncRNAs. The five plant species including two eudicots *Arabidopsis thaliana* and *Gossypium arboreum*, two monocots *Oryza sativa* and *Zea mays*, and an algae *Chlamydomonas reinhardtii* which have available data of Ribo-seq and MS in public database were selected to analyse and get the translational evidence for sORFs. Totally, we collected 103 Ribo-seq for the five major species from NCBI (<https://www.ncbi.nlm.nih.gov/>) and EBI (<https://www.ebi.ac.uk/>), together with 93 mass spectral (MS) projects generated by high sensitivity mass spectrometry instrument (Q Exactive or LTQ Orbitrap Elite) in PRIDE database (<https://www.ebi.ac.uk/pride/archive/>).

To integrate above data, we built a pipeline which is shown in Figure 1a. When defining the candidate sORFs, all three possible reading frames of RNA transcript were examined, and ATG and near-cognate codons (ATG, TTG, GTG, CTG, AAG, AGG, ACG, ATA, ATT, ATC), and TAG, TAA, TGA were considered as start and stop codons, respectively. To determine whether a candidate sORF is translated, the Ribo-seq and MS data were analysed separately using different softwares. The PRICE (v 1.0.2) (Erhard *et al.*, 2018) was used to analyse the 3 nt periodic feature of ribosome footprints from Ribo-seq data. The SearchGUI (v 3.3.13) (Barsnes and Vaudel, 2018) was used to find the peptides matching with the translational reading frame in MS data. Then, the two sets of sORF from Ribo-seq and MS were filtered to retain sORFs with length of 18–300 nt and combined by taking the union set to get the core sORF registry for the five plant species.

For other 30 plant species, we used the BLAST to find the homologous sORFs to the core sORF registry. Finally, these sORFs from 30 other plant species and the core sORF registry were combined to get the comprehensive sORF registry of 35 plant species, which was consisted of 112,350 sORF from 51 341 transcripts. Based on their genome location, the sORFs could be divided into five categories: uORF (44,467), uoORF (4788), dORF (53 229), doORF (4403) and sORF (5463) (Figure 1b). Based on their sequence conservation, current version of psORF contains 11 665 homologous sORF family.

In addition, to link the identified sORFs with known knowledge, we collected sROFs in the published literatures by using python-scripted web crawler to discover the key words in the abstract and main text, such as small (coding) ORF/sORF, small protein/peptide, micro-protein/peptide, unannotated translation events, downstream ORF/dORF and upstream ORF/uORF. The known sORFs were made a database which was BLAST against sORFs in the comprehensive sORFs registry by using BLASTp (v 2.6.0+) with parameter setting: cut-offs: e-value ≤ 0.01 , coverage $\geq 30\%$ and identity = 100. The BLAST hits were shown in the gene wiki page of each sORF.

PsORF was deployed on Linux operation system with nginx web server, and all data were stored in MySQL database for query. PsORF offers convenient browse and query services for users (Figure 1a) to get the basic sORF information. In PsORF, users can:

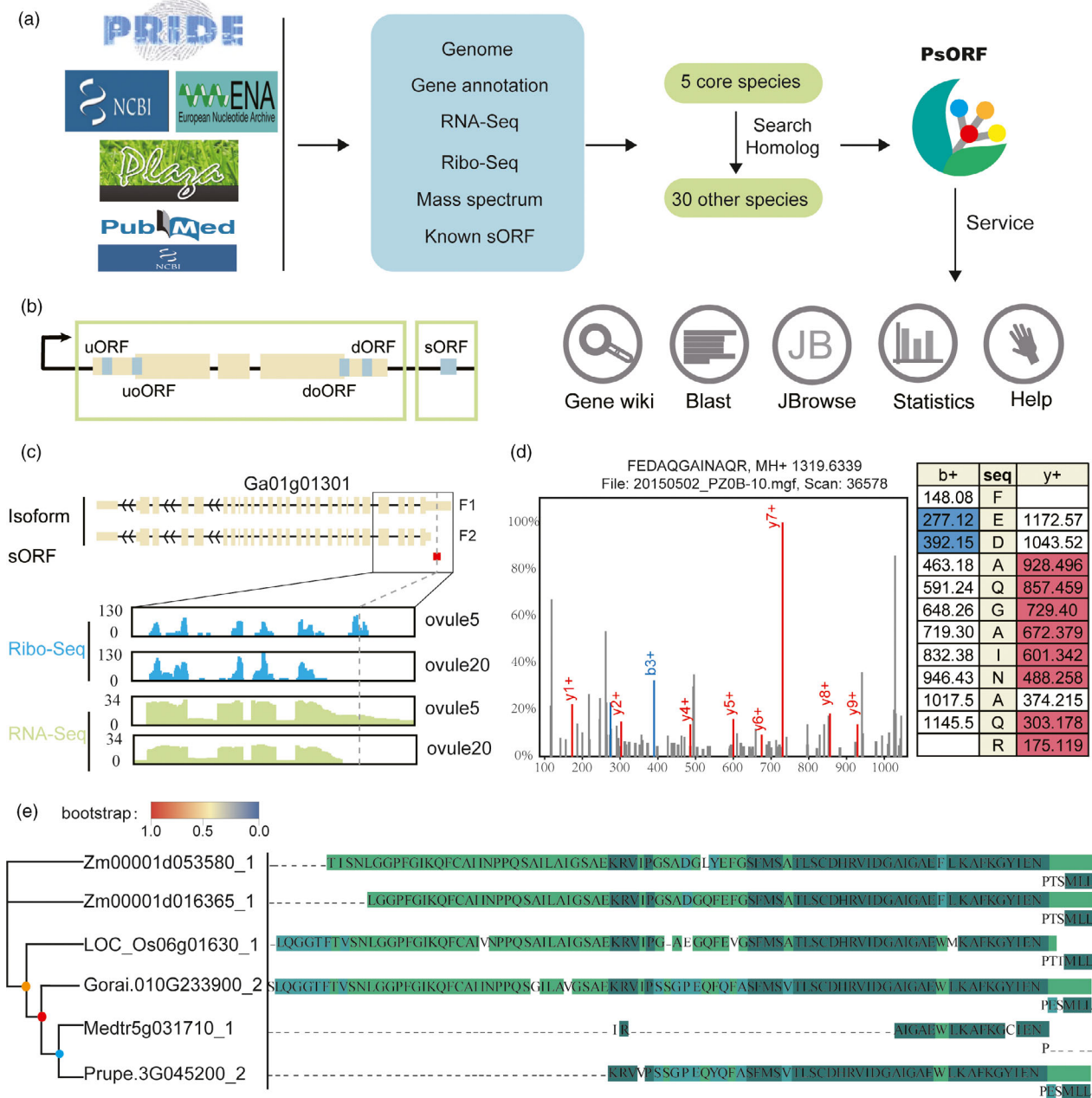


Figure 1 Schematic of PsORF database. (a) Data sources and data processing pipeline of PsORF. (b) The five kinds of sORFs classified by the genome location. uORF, small ORF in the upstream of mORF; uoORF, small ORF across 5'UTR and mORF; dORF, small ORF in the downstream of mORF; doORF, small ORF across mORF and 3'UTR; sORF, other small ORF in the genome. (c) The JBrowse showing a uORF, the associated tracks (Ribo-seq and RNA-seq) of which are showed. (d) The MS spectra of a dORF. The *b* and *y* ion are showed in blue and red colour, respectively. (e) The phylogenetic tree for a conserved sORF and its homologs across five plant species.

(i) browse or search sORFs with ID and sequence; (ii) BLAST the sequence similarity of sORFs across plant species; (iii) browse the Ribo-seq and RNA-seq data and genome location information of sORFs in genome browser JBrowse (Figure 1c) (Buels *et al.*, 2016); (iv) view the MS/MS fragmentation spectra of small peptides (sORFs encoding) in the visual platform (Figure 1d); (v) find the phylogenetic tree of conserved sORFs across different plant species; (Figure 1e); and (vi) check whether the sORFs or their homologs have associated researches in published literature.

To our best knowledge, PsORF (<http://psorf.whu.edu.cn/>) is the unique comprehensive database for plant sORFs. As the

accumulation of translomic data from Ribo-seq and proteomic data from MS, more and more important sORFs and their regulatory roles will be identified. Thus, we will keep on updating PsORF as new data available. We believe that the database will facilitate plant scientists to quickly get the sORF information for further biological discovery.

Acknowledgements

This work was supported by grants: the National Program on Research and Development of Transgenic Plants

(2016ZX08009003-004), the National Natural Science Foundation of China (31770310) and the Fundamental Research Funds for the Central Universities (2042018kf0225) to K.W.; and Innovation Team Program from Wuhan University to Y.Z. and K.W. (2042017kf0233). The Wuhan Gooalgene Technology Co. helped in database construction.

Conflict of interests

The authors declare no competing interests.

Author contributions

K.W. and Y.J. designed the project and wrote the manuscript. Y.J., D.Y., W.F., H.Y., Y.F.Z., Y.Z. and W.D. contributed to data analysis and web design. X.Z. and X.L. contributed to Ribo-seq and MS assays.

References

Barsnes, H. and Vaudel, M. (2018) SearchGUI: a highly adaptable common interface for proteomics search and de novo engines. *J. Proteome Res.* **17**, 2552–2555.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D.M. et al. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol.* **17**, 1–12.

Erhard, F., Halenius, A., Zimmermann, C., L'Hernault, A., Kowalewski, D.J., Weekes, M.P., Stevanovic, S. et al. (2018) Improved Ribo-seq enables identification of cryptic translation events. *Nat. Methods*, **15**, 363–366.

Hao, Y., Zhang, L., Niu, Y., Cai, T., Luo, J., He, S., Zhang, B. et al. (2017) SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci. *Brief. Bioinform.* **19**, bbx005.

Hazarika, R.R., De Coninck, B., Yamamoto, L.R., Martin, L.R., Cammue, B.P.A. and Van Noort, V. (2017) ARA-PEPs: A repository of putative SORF-encoded peptides in *Arabidopsis thaliana*. *BMC Bioinformatics*, **18**, 1–9.

Olexiouk, V., Crappé, J., Verbruggen, S., Verhegen, K., Martens, L. and Menschaert, G. (2016) sORFs.org: a repository of small ORFs identified by ribosome profiling. *Nucleic Acids Res.* **44**, D324–D329.

Wang, K., Wang, D., Zheng, X., Qin, A., Zhou, J., Guo, B., Chen, Y. et al. (2019) Multi-strategic RNA-seq analysis reveals a high-resolution transcriptional landscape in cotton. *Nat. Commun.* **10**, 4714.

Xu, G., Yuan, M., Ai, C., Liu, L., Zhuang, E., Karapetyan, S., Wang, S. et al. (2017) uORF-mediated translation allows engineered plant disease resistance without fitness costs. *Nature*, **545**, 491–494.

Zhong, S., Liu, M., Wang, Z., Huang, Q., Hou, S., Xu, Y.-C., Ge, Z. et al. (2019) Cysteine-rich peptides promote interspecific genetic isolation in *Arabidopsis*. *Science (80-)*, **364**, eaau9564.