MEDWIN PUBLISHERS
Committed to Create Value for Researchers

Open Access Journal of Microbiology & Biotechnology
ISSN: 2576-7771

In Silico Prediction of Small Open Reading Frames from Intergenic Regions of *Cucumis sativus* L. Var. *Hardwickii*

Chieng GSW, Tan BC and Teo CH*

Centre for Research in Biotechnology for Agriculture (CEBAR), University of Malaya, Malaysia

***Corresponding author:** Chee How Teo, Centre for Research in Biotechnology for Agriculture (CEBAR), University of Malaya, Kuala Lumpur 50603, Malaysia, Email: cheehow.teo@um.edu.my

Research Article

Volume 7 Issue 4

Received Date: November 07, 2022

Published Date: November 23, 2022

DOI: 10.23880/oajmb-16000242

Abstract

Small open reading frames play important roles in growth and development regulation in plant species. However, their sequences and functions remain poorly understood in many plant species including *Cucumis sativus* which is Asia's fourth most important vegetable. The breeding of climate-resilient cucumbers is of great importance to ensure their sustainability under extreme climate conditions. In this study, we aim to predict the intergenic sORFs from *C. sativus* var. *hardwickii* and determine their expression profiles in transcriptome datasets. We identified a total of 50,191 coding sORFs from var. *hardwickii* genome. In addition, 1,311 transcribed sORFs were detected in RNA-seq datasets of var. *hardwickii* and shared homology to sequences deposited in the cucumber EST database, and among these, 91 transcribed sORFs with translation potential were detected. The findings of this study provide insight into sequence diversity and expression patterns of sORFs in *C. sativus*, which could help in developing climate-resilient cucumbers.

Keywords: *Cucumis sativus* var. *Hardwickii*; Small Open Reading Frame; Transcribed sORF; Cucumber; Coding sORF

Abbreviations: uORF: Upstream ORF; sORF: Small Open Reading Frame; lncORF: Intergenic sORF, Long Noncoding ORF; dORF: Downstream ORF; SEPs: sORF-encoded Proteins; GO: Gene Ontology; KEGG: Kyoto Encyclopedia of Genes and Genomes; BP: Biological Processes; MF: Molecular Function; CC: Cellular Component; KO: KEGG Ontology.

Introduction

Cucurbitaceae used to be documented as a monophyletic family without any close relatives [1], but with the addition of more mitochondrial and chloroplast genome sequences of old and new plant materials, a few of the closest relatives were discovered in this family [2]. Within Cucurbitaceae, there are roughly 66 species in the genus *Cucumis*, and cucumber (*Cucumis sativus*) is the only one having 2n = 2x = 14 chromosomes [3]. Among the many varieties of *Cucumis sativus*, wild cucumber (*C. sativus* var. *hardwickii*), semi-wild Xishuangbanna cucumber (*C. sativus* var. *xishuangbannesis*), the Sikkim cucumber (*C. sativus* var. *sikkimensis*) and the cultivated cucumber (*C. sativus* var. *sativus*) are cross-compatible [4]. Today, China has been ranked as one of the top producers and largest domesticators of cucumbers. Based on the statistics from FAOSTAT [5], the total cucumber production in China in 2018 was 56.24 million tons from 1.044 million hectares. The cucumber production and area utilised for cucumber production in China stand at 52.7% and 74.8% of the corresponding world totals respectively. As for the yield per unit area, China exceeded the world average by 42% with a total of 53.86 kg/ha. Apart from China, India

is also a competitive player in terms of its annual cucumber production. In a study by Sanjeev, et al. [6], the annual production of cucumber was 0.698 million tons from 45,000 ha with a productivity of 15.5 t/ha. For both countries, the most concerning issue is having low productivity and challenging climatic diversity [6,7].

Small open reading frame (sORF) as its name suggests, is a shorter version of the canonical ORF. The size of sORF ranges from 30 bp to 300 bp [8]. Their minuscule size has caused them to be excluded from most gene prediction methods [9]. The length cut-off filter used in most gene prediction methods is 300 bp and any sequences below this cut-off will be considered as being non-functional [10]. Another contributing factor causing sORF to be diminished is that short sequences normally have low evolutionary conservation scores, an indicator of the functionality of a gene [11]. To date, there is still no standard classification for sORF and small peptide derived from sORF but researchers have made attempts to classify them into different categories, namely upstream ORF (uORF), intergenic sORF, long noncoding ORF (lncORF), short CDS, short isoform, downstream ORF (dORF), CDS-sORF, interlaced-sORF, miREP, microprotein, hormone-like peptide and defensin [9,12-16].

sORFs have been reported to get translated into small peptides and have functional roles in plants [17-19]. Ong, et al. [16] reviewed the roles of sORF-encoded proteins (SEPs) in several biological processes, including cell signaling, abiotic stress responses, morphogenesis, and growth regulations. In addition, studies also reported that short proteins also function as secreted peptides and hormones. Quio, et al. [19] reported an 18 aa plant polypeptide hormone known as systemin that engages in plant defence mechanisms and secretes phytosulfokine pentapeptides (PSK; 100 aa) that function in regulating plant growth and stress responses.

In this study, we used the reference genome of *C. sativus* var. *hardwickii* PI183967 for *in silico* sORF prediction and characterisation. The main objective of this study was to identify and characterise intergenic sORFs from *C. sativus* var. *hardwickii* PI183967 genome. To achieve this objective, we first identify the coding sORFs from the genome sequences of *C. sativus* var. *hardwickii* PI183967 using sORFfinder. We then classified the coding sORFs into coding sORF with transcription potential (transcribed sORF) and with both transcription and translation potential (translated sORF) based on the outcomes from the transcript (RNA-seq and EST) and protein sequence homology search (SWISS-PROT) analysis. We determined the sORF expression profiles in RNA-seq datasets of *C. sativus* var. *hardwickii* using gene expression tools. Finally, the potential biological functions of the translated sORFs were annotated using Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) pipeline. The findings from this study will set an important foundation for the development of climate-resilient cucumbers and other crop species.

Material and Methods

Data Retrieval of *Cucumis sativus* Reference Genome and Transcriptomes

The reference genome sequences and annotation file of *C. sativus* var. *hardwickii* PI183967 were retrieved from CuGenDBv2 (http://cucurbitgenomics.org/). The transcriptome datasets of *C. sativus* var. *hardwickii* (Project ID: PRJNA624798) were retrieved from NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra).

In silico Prediction of Small Open Reading Frame

The CDS, exon, intron, and intergenic regions of *C. sativus* var. *hardwickii* PI183967 were extracted from its reference genome using gff2sequence [20]. sORFfinder [8] was used to predict the coding sORFs from *C. sativus* var. *hardwickii* PI183967 genome. RepeatMasker (http://www.repeatmasker.org) was used to mask the repeat sequences in the coding sORFs and the masked sequences were removed using an in-house script. Sequence clustering of the coding sORFs was performed using CD-HIT [21] with a clustering threshold of 95% to cluster the redundant sequences into sequence clusters.

Characterisation of Small Open Reading Frame

To identify transcribed sORFs, the coding sORFs were blast searched against Cucumber EST collection version 3 (http://cucurbitgenomics.org/est/cucumber) using the blastn algorithm with homology search parameters "-evalue 1e-5 -per_identity 97 -qcov_hsp_perc 100". The nucleotide sequences of transcribed sORFs were translated to amino acid sequences using gotranseq (https://github.com/feliixx/gotranseq). The amino acid sequences of transcribed sORFs were blast searched locally against the high quality manually annotated and non-redundant protein sequences retrieved from the SWISS-PROT database (https://www.uniprot.org/help/downloads) using the blastp algorithm with homology search parameters "-evalue 1e-5 -per_identity 97 -qcov_hsp_perc 100". The transcribed sORFs that shared high homology to SWISS-PROT protein sequences were designated as transcribed sORFs with translation potential (translated sORFs). The amino acid sequences of translated sORFs were then blast searched against the plant sORF database, PsORF (http://psorf.whu.edu.cn/)[22].

Transcriptome Analysis of Small Open Reading Frames

The coding sORF (csORF) sequences of *C. sativus* var. *hardwickii* PI183967 were mapped to its reference genome using blat (https://github.com/djhshih/blat) and the outputs were converted to gene annotation file using blat2gtf.pl (https://github.com/IGBIllinois/HOMER/blob/master/bin/blat2gtf.pl). The sORF annotation file was combined with the reference genome annotation file using agat_sp_merge_annotations.pl from AGAT package (https://github.com/NBISweden/AGAT). The RNA-seq reads were aligned and mapped to *C. sativus* var. *hardwickii* PI183967 genome using HISAT2 (http://daehwankimlab.github.io/hisat2/) and the expression profile of sORFs in the RNA-seq datasets were determined using Stringtie (https://ccb.jhu.edu/software/stringtie/#install) together with the updated gene annotation file. The sORF sequence IDs were retrieved from the Stringtie output file using an in-house script and the sORF sequences were then extracted from the coding sORF file using the subseq function in the seqtk package (https://github.com/lh3/seqtk). The sORF sequences identified by HISAT2/Stringtie pipeline were combined with the sORF sequences that shared homology to cucumber ESTs to form a final set of transcribed sORF.

Functional Annotation of Small Open Reading Frame

To assign a biological function to transcribed sORF, Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) were performed using the clusterProfiler R package (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html). The transcribed sORF amino acid sequences were first blast searched against SWISS-PROT database using the blastp algorithm. The SWISS-PROT protein IDs were extracted from the blast output using an in-house script and converted to Gene ID using the "Retrieve/ID Mapping" function from the UniProt website (https://www.uniprot.org/id-mapping). The GO category analysis was conducted using the group GO function embedded in the clusterProfiler R package. The enrichment analyses of GO and KEGG were performed using the enrich GO and enrich KEGG function embedded in the clusterProfiler R package.

Results and Discussion

Identification of Small Open Reading Frame in *C. sativus* var. *Hardwickii*

The total number of sORFs in an organism varied from species to species. In *Arabidopsis thaliana*, approximately 33,809 sORFs were predicted from the intergenic regions using sORFfinder with 7,159 sORFs are coding sORFs and 2,996 coding sORFs likely expressed in at least one experimental condition of the tilling array data [23]. Using the same sORF detection pipeline, a total of 850,540 sORFs were predicted from the genome of *C. sativus* var. *hardwickii* PI183967 and 50,584 coding sORFs were detected (Table 1).

Description of sORF	Number of sORF
sORF	850,540
Coding sORF	50,584
Unique coding ORF	50,191
Transcribed sORF with homolog in cucumber EST database	693
Transcribed sORF in leaf transcriptome	1,311
Transcribed sORF with homolog in SWISS-PROT	91
Transcribed sORF with homolog in PsORF	82

Table 1: Summary of intergenic small open reading frames found in *C. sativus* var. *hardwickii.*

Repeat Masker and CD-HIT were conducted to remove the csORFs with repeat sequence homology and to cluster redundant csORFs to unique coding sORFs. A final set of 50,191 csORFs was blast searched against the Cucumber EST collection to obtain 693 transcribed sORFs that shared high homology to cucumber EST sequences. Out of the 693 transcribed sORFs, 91 showed high homology to protein sequences deposited in the SWISS-PROT database. Using the leaf transcriptome datasets of *C. sativus* var. *hardwickii*, we identified 1,311 transcribed sORFs (Table 1).

Transcribed sORF with Translational Potential in *Cucumis sativus*

Using a protein sequence homology search approach, 91 transcribed sORFs with translational potential were detected for *C. sativus* var. *hardwickii* (Table 1). Besides SWISS-PROT, we also blast searched the transcribed sORFs against the sORFs with translational potential deposited in the PsORF database. The PsORF database is a collection of plant sORFs from 35 different plant species [22]. The authors collected multi-omics datasets including genome, transcriptome, Ribo-seq, and mass spectrum from public databases and built a bioinformatics pipeline to detect sORFs in these datasets. Results from blast search against the PsORF database showed that 90.11 % of *C. sativus* var. *hardwickii* transcribed sORFs with the homologs in SWISS-PROT also have homologs in the PsORF database (Figure 1). This indicates that these

transcribed sORFs might have translational potential. Using a proteomic approach, Castellana, et al. [17] identified ~5,000 small peptides in *Arabidopsis* and some of these small peptides were novel and/or identified by Hanada, et al. [23].

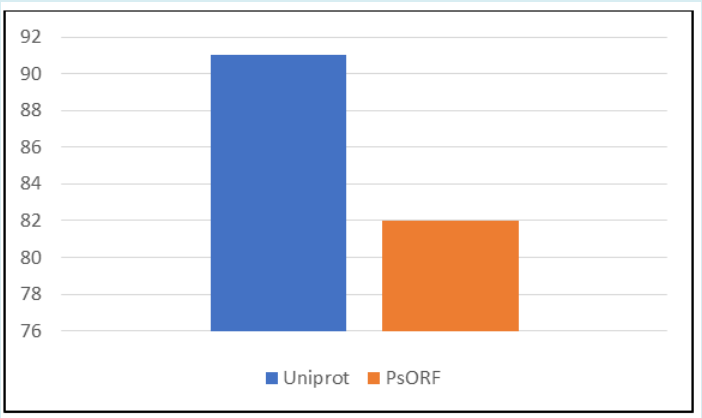

Figure 1: Transcribed sORFs with translational potential in *C. sativus* var. *hardwickki*.

Functional Classification of *Cucumis sativus* var. *hardwickii* sORF

From the 91 transcribed sORFs with the translational potential of *C. sativus* var. *hardwickii*, 2,005 unigenes with Entrez ID were retrieved from the UnitProt database for Gene Ontology (GO) analysis using the clusterProfiler (Figure 2). Among the 3 distinct categories of GO classes, molecular functions were the most represented functional group for sORF functional annotation (Figure 2). The top three GO terms for biological processes (BP) are nitrogen compound metabolic process, reproductive process, and immune response. For molecular function (MF), top three GO terms are protein binding, hydrolase activity, and oxireductase activity whereas for cellular component (CC), top three GO terms are cytoplasm, intracellular organelle, and myosin complex (Figure 2).

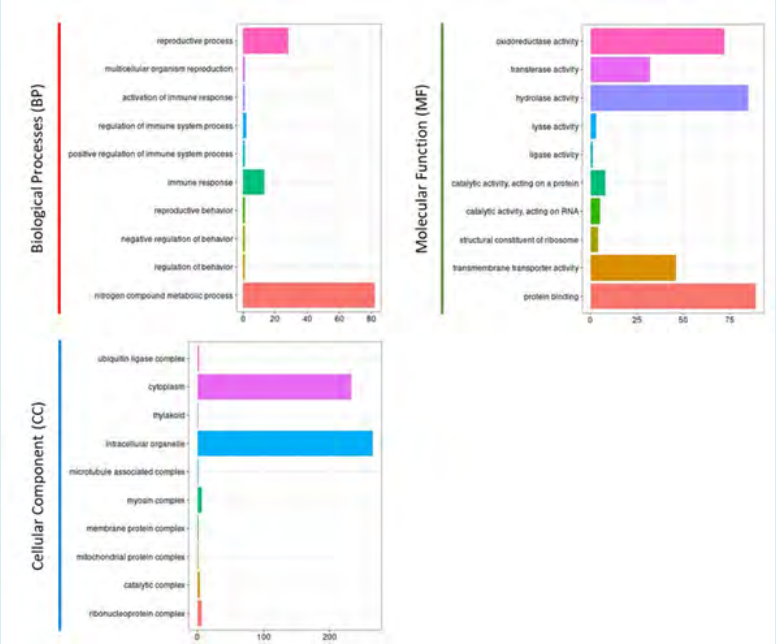

Figure 2: Gene ontology (GO) classification of *C. sativus* var. *hardwickii* transcribed sORFs.

We also performed the GO enrichment analysis of transcribed sORFs of *C. sativus* var. *hardwickii* (Figure 3). The transcribed sORFs were enriched in the GO terms of BP and CC. No enrichment was detected for MF. In *C. sativus*

var. *hardwickii,* most of the transcribed sORFs showed significant CC enrichment in the endosome, cytoskeleton, and actin cytoskeleton and to a lesser extent in the Golgi membrane and trans-Golgi network membrane. The cytoskeleton mainly functions as a structure for cell shape and internal organization and is made up of 3 elements, namely, microtubules, intermediate filaments, and actin [24]. As shown in the GO enrichment plot (Figure 3), the transcribed sORFs with enriched GO terms related to actin were detected in *C. sativus* var. *hardwickii*. Actin mainly functions in cellular physiological processes in plants including cell growth, cytokinesis, cell division, and several intracellular trafficking events [25]. This indicates that the transcribed sORFs of *C. sativus* might play a role in growth and development regulations. Hanada, et al. [18] reported that overexpression of sORFs showed varying morphological changes in transgenic *A. thaliana* and was associated with a higher growth rate.

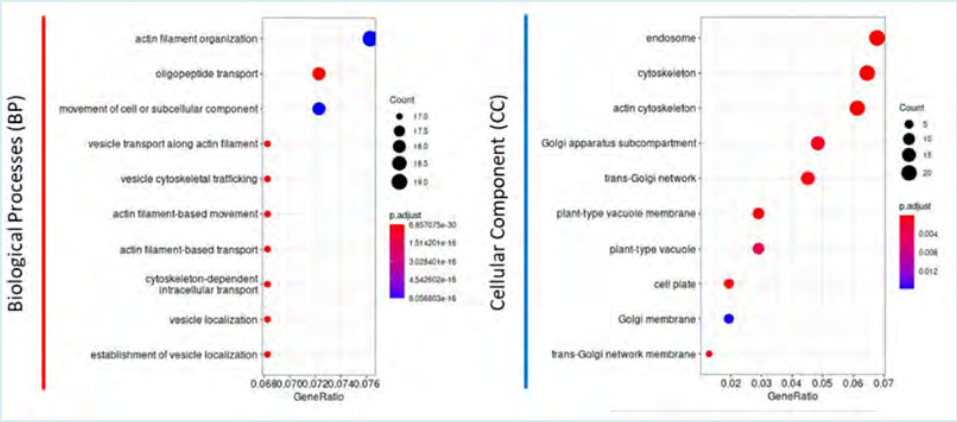

Figure 3: Bubble plot showing the enriched GO terms. X-axis in the bar plot stood for gene ratio, while the y-axis indicates different BP and CC. The size of the circles in each plot is positively correlated with the number of genes involved in each subgroups while the colour of the circles indicate their significance level.

For the enriched KEGG pathways, the transcribed sORFs in *C. sativus* var. *hardwickii* was enriched in KEGG Ontology (KO) terms of plant hormone signal transduction (Figure 4). Plant sORFs have been demonstrated to play important roles in cell signalling, abiotic stress response, morphogenesis, and growth regulation [16,18,26-34]. Apart from that, some of the transcribed sORFs identified in *C. sativus* var. *hardwickii* are significantly associated with KO terms of various metabolisms. This indicates that *C. sativus* transcribed sORFs play a significant role in plant growth and development, and environmental stress responses.

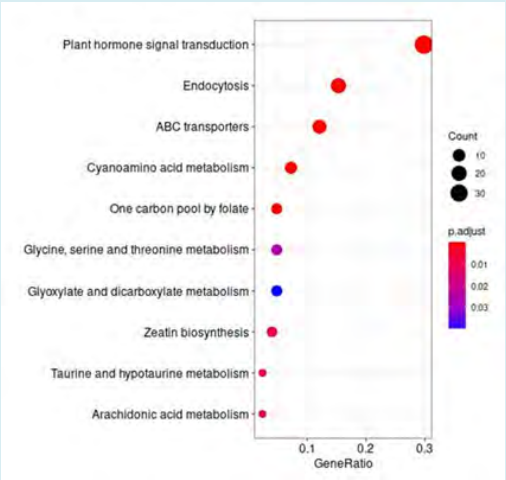

Figure 4: Bubble plot of enriched KEGG pathway. The x-axis indicates the gene ratio and the y-axis stood for the pathway enriched.

Conclusion

In this study, we have established a bioinformatics pipeline for the identification of small open reading frames (sORFs) in *Cucumis sativus*. Using the pipeline, different types of sORFs were identified from the genome and transcriptome datasets of *Cucumis sativus* var. *hardwickii*. GO and KEGG terms that enriched in growth and development, and stress response were predicted for the transcribed sORFs with translational potential. Further classification of sORFs is needed to minimise conflicting sORF annotations and ease categorisations of sORFs. Having a complete database of transcribed sORFs and translated sORFs in *Cucumis sativus* would help us understand the roles of plant SEPs, especially in the biotic and abiotic stress responses. With that being said, cucumber producers will be able to improve crop viability, especially in harsh weather conditions, and produce cucumbers with higher market value.

References

1. Naegele RP, Wehner TC (2017) Genetic Resources of Cucumber. In: Grumet R, et al. (Eds.), Genetics and Genomics of Cucurbitaceae. Springer, Cham 20: 61-86.

2. Chomicki G, Schaefer H, Renner SS (2020) Origin and domestication of Cucurbitaceae crops: Insights from phylogenies, genomics and archaeology. The New phytologist 226(5): 1240-1255.

3. Renner SS, Schaefer H, Kocyan A (2007) Phylogenetics of cucumis (Cucurbitaceae) Cucumber (*C. sativus*) belongs in an Asian/australian clade far from melon (C. melo). BMC Evol Biol 7: 58.

4. Weng Y (2020) *Cucumis sativus* chromosome evolution, domestication, and genetic diversity. In: Goldman I (Ed.), Plant Breeding Reviews. Wiley Online Library, pp: 44

5. FAOSTA (2022) Food and Agriculture Data.

6. Sanjeev K, Patel NB, Saravaiya SN, Desai KD (2015) Economic viability of cucumber cultivation under NVPH. African Journal of Agricultural Research 10(8): 742-747.

7. Liu H, Yin C, Gao Z, Hou L (2021) Evaluation of cucumber yield, economic benefit and water productivity under different soil matric potentials in solar greenhouses in North China. Agricultural Water Management 243: 106442.

8. Hanada K, Akiyama K, Sakurai T, Toyoda T, Shinozaki K, et al. (2010) Sorf Finder: A program package to identify small open reading frames with high coding potential. Bioinformatics 26(3): 399-400.

9. Couso JP, Patraquim P (2017) Classification and function of small open reading frames. Nat Rev Mol Cell Biol 18(9): 575-589.

10. Kute PM, Soukarieh O, Tjeldnes H, Trégouët DA, Valen E (2022) Small open reading frames, how to find them and determine their function. Front Gene 12: 796060.

11. Leong AZX, Lee PY, Mohtar MA, Syafruddin SE, Pung YF, et al. (2022) Short open reading frames (sorfs) and microproteins: An update on their identification and validation measures. J Biomed Sci 29(1): 19.

12. Fesenko I, Kirov I, Kniazev A, Khazigaleeva R, Lazarev V, et al. (2019) Distinct types of short open reading frames are translated in plant cells. Genome Res 29(9): 1464-1477.

13. Khitun A, Ness TJ, Slavoff SA (2019) Small open reading frames and cellular stress responses. Mol Omics 15(2): 108-116.

14. Takahashi F, Hanada K, Kondo T, Shinozaki K (2019) Hormone-like peptides and small coding genes in plant stress signaling and development. Curr Opin Plant Biol 51: 88-95.

15. Orr MW, Mao YH, Storz G, Qian SB (2020) Alternative orfs and small orfs: Shedding light on the dark proteome. Nucleic Acids Res 48(3): 1029-1042.

16. Ong SN, Tan BC, Idrus AA, Teo CH (2022) Small open reading frames in plantresearch: From prediction to functional characterization. 3 Biotech 12(3): 76.

17. Castellana NE, Payne SH, Shen Z, Stanke M, Bafna V, et al. (2008) Discovery and revision of *Arabidopsis* genes by proteogenomics. Proc Natl Acad Sci USA 105(52): 21034-21038.

18. Hanada K, Takeuchi MH, Okamoto M, Yoshizumi T, Shimizu M, et al. (2013) Small open reading frames associated with morphogenesis are hidden in plant genomes. Proc Natl Acad Sci USA 110(6): 2395-2400.

19. Quio LEC, Herberg S, Pauli A (2016) Decoding sORF translation – from small proteins to gene regulation. RNA Biol 13(11): 1051-1059.

20. Camiolo S, Porceddu A (2013) gff2sequence, a new user friendly tool for the generation of genomic sequences. BioData Min 6(1): 15.

21. Li W, Godzik A (2006) CD-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22(13): 1658-1659.

22. Chen Y, Li D, Fan W, Zheng X, Zhou Y, et al. (2020) PSORF: A database of small ORFs in plants. Plant Biotechnol J 18(11): 2158-2160.

23. Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH (2007) A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res 17(5): 632-640.

24. Scitable (2014) Microtubules and Filaments. Nature Publishing Group.

25. Diao M, Huang S (2021) An update on the role of the actin cytoskeleton in plasmodesmata: A focus on Formins. Front plant Sci 12: 647123.

26. Bashir K, Hanada K, Shimizu M, Seki M, Nakanishi H, et al. (2014) Transcriptomic analysis of rice in response to iron deficiency and excess. Rice 7(1): 18.

27. Doubrava N, Blake JH, Keinath AP, Williamson J (2021) Cucumber, squash, Melon & Other Cucurbit Diseases. Home & Garden Information Center, Clemson University, South Carolina, USA.

28. Kim D, Langmead B, Salzberg SL (2015) HISAT: A fast spliced aligner with low memory requirements. Nat Methods 12(4): 357-360.

29. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL (2019) Graph-based genome alignment and genotyping with Hisat2 and HISAT-genotype. Nat Biotechnol 37(8): 907-915.

30. Ma J, Ward CC, Jungreis I, Slavoff SA, Schwaid AG, et al. (2014) Discovery of human sorf-encoded polypeptides (seps) in cell lines and tissue. J Proteome Res 13(3): 1757-1765.

31. Mochizuki T, Ohki ST (2012) Cucumber mosaic virus: Viral genes as virulence determinants. Mol Plant Pathol 13(3): 217-225.

32. Peeters MKR, Menschaert G (2020) The Hunt for sorfs: A multidisciplinary strategy. Exp Cell Res 391(1): 111923.

33. Sterck L (2022) Draw Venn Diagram. Bioinformatics and Evolutionary Genomics.

34. Wang M, He X, Peng Q, Liang Z, Peng Q, et al. (2020) Understanding the heat resistance of cucumber through leaf transcriptomics. Funct Plant Biol 47(8): 704-715.