



I. Bio curation of transcriptomics dataset:

a. Data mining:

The relevant transcriptome datasets for combined stress in plants were compiled and curated using two major public databanks for microarray data, including Gene Expression Omnibus (GEO) (<https://www.ncbi.nlm.nih.gov/geo/>) and Array Express (<http://www.ebi.ac.uk/arrayexpress/>) was used. The NCBI GEO and ArrayExpress functional genomics repository were queried using keywords, “combined stress” AND “Plants” [organism]. For the compilation of RNA-seq transcriptomics data NCBI, Sequence Read Archive (SRA) (<https://www.ncbi.nlm.nih.gov/sra>) database was used.

b. Data curation

The transcriptomic datasets mined from public databases were manually curated, to determine whether they were from actual combined stress studies. Also duplicates studies were filtered out. For the final analysis studies that showed availability of completely raw and processed datasets and also at least 4 samples (2 combined stress treatments and 2 respective controls, comparable case and control samples) were included.

c. Analysis of Transcriptomics dataset:

Microarray data mined from relevant studies were analyzed in the R environment using the *GEOquery* (<http://www.bioconductor.org/packages/2.8/bioc/html/GEOquery.html>) and *limma* (Linear Models for Microarray Analysis) (<http://www.bioconductor.org/packages/release/bioc/html/limma.html>) R packages from the Bioconductor project. The *GEOquery* R package performs the initial parsing of the GEO data into R data structures. This data is used by the *limma* R package for identifying differentially expressed genes (DEGs) in the input dataset. Following this probe ID conversion was done for DEGs obtained from individual microarray datasets. Based on the platform used in the individual studies probe ids match was done to



convert the DEGs into a common id format. Affymetrix (https://www.affymetrix.com/analysis/netaffx/xmlquery_ex.affx?netaffx=wtgene_transcript) platform was used to do the probe ID conversion to fetch their corresponding transcript ids. For Agilent and other platforms, bioDBnet (biological DataBase network) platform was used (<https://biodbnet-abcc.ncifcrf.gov/db/db2db.php>). In cases where multiple probes matched the same locus, the probe ID with the highest fold change was considered.

For RNA-seq transcriptome datasets analysis, SRA (Sequence Read Archive) toolkit (Version 2.10.8) was used to download and split the SRA raw read data from NCBI (National Center for Biotechnology Information) into FASTQ files (<https://www.ncbi.nlm.nih.gov/>). Raw sequence reads in the format FASTQ files, thus downloaded were subjected to quality check by FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) followed by subsequent trimming of the reads using Trimmomatic (<http://www.usadellab.org/cms/?page=trimmomatic>). This was followed by mapping the reads to the reference genome using HISAT2 (Hierarchical Indexing for Spliced Alignment of Transcripts 2, version 2.0.1) (<http://www.ccb.jhu.edu/software/hisat/index.shtml>) which is a fast and sensitive spliced alignment program for mapping RNA-seq reads. SAM (Sequence Alignment/Map format) files to BAM (Binary Alignment/Map format) files conversion was done using SAMtools (<http://samtools.sourceforge.net>, version 0.1.19). Subsequent to this counting of reads mapped to individual genes or transcripts was done using bedtools (<https://bedtools.readthedocs.io/en/latest/content/installation.html>). In the final step differentially expressed genes between different combined stress and individual stress treatments were identified using DESeq2 in R (<https://bioconductor.org/packages/release/bioc/html/DESeq2.html>). For transcriptome studies where analyzed data was already provided by the author's, DEGs were directly downloaded and used for the further analysis pipelines.

d. Biological interpretation of transcriptome data:

The DEGs identified under each stress combination were associated with several



meta-data, like, gene names and mapped to 'KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway'. In order to aid users to navigate through several other databases and cross-references outside KEGG and fetch a deluge of meta-data corresponding to a particular entry, KEGG genes links were provided in cases where pathway information was not available. KEGG API (Application Programming Interface) (<https://www.kegg.jp/kegg/rest/keggapi.html>) was used to fetch both of these links for respective DEGs. In order to further gain insights into the biological functions of DEGs identified by transcriptome analysis, functional and pathway enrichment analysis was performed using the gProfileR package (<https://cran.r-project.org/web/packages/gProfileR/>). Benjamini– Hochberg adjustment for multiple hypothesis testing was used to correct the P-values. The over-represented GO (Gene Ontology) and KO (KEGG Ontology) terms measured by the adjusted P-values depicted gene function and biological pathway associations.

e. Data integration into the SCIPDb

The frontend user interface was implemented using HTML5, CSS, JQuery and PHP (version: 7.0.12). HTML pages were designed to accept queries based on a three-level dropdown-based selection using JavaScript, specific to each plant species. The back-end schema was designed using MySQL, an open-source relational database management system, and data was stored in mysql tables (Version: 5.7.17). To provide an interactive interface and enhanced user experience, we used Bootstrap 4, to present final results.

f. Visualizations

To visualize the high-dimensional transcriptomics data, several in-house scripts were generated mainly in using shell scripts and R. The heat map was made interactive using the "Heatmaply" (<https://github.com/talgalili/heatmaply>) package from R. "VennDiagram" (<https://cran.r-project.org/web/packages/VennDiagram/>) package was used and scripts were customized to generate color-coded Venn diagrams representing specific categories. Gene ontology enrichment analysis and interactive visualization of the results in the form of Manhattan plots were done using "gprofilerR" package (<https://cran.r-project.org/web/packages/gProfileR/index.html>). Correlates data for the top twenty genes in each category have been collected from ATTED-II



(<http://atted.jp/>) which is a plant co-expression database. Functional annotation and pathway mapping for the correlates was done using KEGG and TAIR. Final visualization of the network together with the other meta-data was done using Cytoscape (<https://cytoscape.org/>), an open-source software platform, widely used for visualizing complex networks. Transcription factor enrichment analysis was done using EatupTF tool (<http://chromatindynamics.snu.ac.kr:8080/EatupTF>).

g. Arabidopsis combined stress transcriptome

Upset plot was generated using UpSetR package, while circos plot was generated using Metascape, a gene annotation & analysis resource. (<https://metascape.org/gp/index.html#/main/step1>). Pathway enrichment analysis was done using major pathway databases like KEGG (<https://www.kegg.jp/kegg/rest/keggapi.html>), Aracyc (<https://plantcyc.org/typeofpublication/aracyc>) and Wikipathways (<https://www.wikipathways.org/index.php/WikiPathways>). Final visualization and network analysis was done using Cytoscape, v3.8.2 (<https://cytoscape.org/>).