Biocuration of phenome data

The detailed methodology followed for phenome and transcriptome data mining, analysis, interpretation, and presentation into the SCIPDb is as described below.

The SCIPDb hosts a collection of available literature on various stress combinations known to exist in nature. Based on the existing literature infromation and our knowledge in combined stress we report the presence of 154 (hypothetical number which needs to confirmed by thorough literature analysis) combinations in nature. Presently we have curated 122 stress combinations involving different abiotic and biotic stressors and presented in SCIPDb section.

Deciding the Stress combinations:

Based on the knowledge of the team and the literature in plant science research, we have divided the stress combinations into three categories, namely abiotic-abiotic (consisting of two or more abiotic stress), abiotic-biotic (wherein one stressor is abiotic, and the other is either pathogen or pest) and biotic -biotic (consisting of two or more pathogens/pests). We have not included stress combinations in the present state, including elevated CO_2, herbicides, and insecticides, since these are anthropogenic. We also provide a list of several stress combinations which can exist but are not reported in the literature. The reference of articles that indicate the possible occurrence of such stress combinations is provided. Also, we plan to include many combinations other than the ones presently listed.

The detailed steps followed for phenome data collection are as described below.

a. Literature mining

To retrieve all available articles under combined stress and to have >90% literature coverage, different kind of keywords (refer keywords file) have been used along with the different types of search engines. Bibliography from each article has also been searched to make sure that no articles are left out.

The list of search engines used for literature mining was listed as below,

1. Google- https://www.google.com/
2. NCBI PubMed- https://www.ncbi.nlm.nih.gov/
3. Searchit- http://searchit.libraries.wsu.edu/primo_library/libweb/action/search.do?vid=WSU&dscnt=0&dstmp=1470197360669&fromLogin=true
4. Jstor- http://www.jstor.org/
5. Web of knowledge- http://apps.webofknowledge.com/WOS_GeneralSearch_input.do?product=WOS&search_mode=GeneralSearch&SID=Q25r3nQ9RdXUq7DPaGm&preferencesSaved=
6. Google scholar- https://scholar.google.co.in/
7. Krishikosh (for agriculture-related thesis)- http://krishikosh.egranth.ac.in/
8. Shodhganga (for non-agriculture related thesis)- http://shodhganga.inflibnet.ac.in/
9. Consortium of e-resource in agriculture (CERA)- http://cera.iari.res.in/index.php/en/
10. CABI- https://www.cabi.org/

b. Sorting of articles

After retrieving all the articles related to particular stress combination, articles were sorted as 'main research articles' and 'ancillary articles'. This was done based on the type of articles (e.g., research, review, reports, etc.) and type of data (e.g., morpho-physiological and molecular data) it contains. Main research articles with only morpho-physiological data were considered for data extraction, whereas reports, thesis, book chapters, abstracts, reviews, gene overexpression, and gene-silencing studies were listed under ancillary articles and not considered for data extraction. All these types of articles were integrated into the database under the 'phenomics' tab. Articles with

transcriptome study were considered for integration into the database under the 'transcriptome tab.'

Some intended exclusions:

1. Plant competition, in this database, referred to intraspecific competition and constituted studies related to the effect of plant density involved with other stresses on plants. In most of the cases, inter-specific competitions have not been discussed under this category.

2. The studies involving plant growth promoting bacteria and fungi were excluded as both of them independently do not act as pathogen to plants.

3. In most cases, we have not included tree species. However, in some places where the economically important tree species have been considered, we have included studies wherein the experiments involved young saplings.

4. The studies with gene overexpression or silencing were not considered for the data extraction however they will be included as ancillary articles under respective stress combination.

5. Articles with chemiclal treatment like fungicide, herbicide, fertilizer amendments were also excluded since these are anthropogenic.

c. Listing out parameters and their classification

From the main research articles, parameters studied in each article were listed out and classified into type A, B, and C parameters based on their significance in reflecting the net impact of stress.

- **Type A** includes growth (plant height, biomass, leaf area, leaf number, root length, shoot weight, root weight, etc.) and yield (seed weight, seed number, test weight, etc.), attributing parameters that directly reflect the impact of stress.
- **Type B** includes physiological (photosynthesis, stomatal conductance, transpiration, chlorophyll content, etc.) and pathogenesis (disease index, pathogen load, disease score, etc.) related parameters which indirectly reflect the impact of stress.

- **Type C** includes biochemical parameters such as proline content, MDA content, nutrient content, ROS content, etc., which also explains the impact of stress but to a lesser extent compared to the other two classes of parameters.

For the complete list of parameters hosted in the database, refer “trait ontology” file.

d. Data extraction and depiction

Once parameters were listed out from each article, data values were extracted into the excel file. The values from the table were directly copied into the excel sheet, whereas values from graphs were extracted using the ‘GetData Graph Digitizer’ (http://getdata-graph-digitizer.com/) tool for better accuracy. Since data is heterogeneous and to make it uniform, it was normalized by subjecting to calculation using the formula mentioned below.

$$\text{Change over control (\%)} = \frac{(\text{Control} - \text{stress}) * 100}{\text{Control}}$$

Few parameters, such as electrolyte leakage, pathogenesis-related parameters, etc., were not subjected to calculation. Using both the calculated and un- calculated values (raw values), a table was prepared, reflecting the net impact of stress and interaction between the two stresses at the plant interface. This table was used for preparing the ‘data page’ file for each study on a specific plant and was finally represented in the database in tabular form. For easy understanding, percent change values were shown along with arrows in red and green color. A red-colored downward arrow indicates that the parameter is affected under stress; the higher the positive value greater the damage to the parameter under stress. Green-colored upward arrow indicated parameters are not affected under stress conditions as compared to control.

e. Data analysis and interpretation

Data presented in tabular form was analyzed by comparing the individual and combined stress values of each parameter. Suppose percent change values are greater in combined stress compared to both the individual stressed. In that case, the outcome of combined stress is depicted as ‘negative.’ If percent change values are less in combined stress compared to both the individual stress, then the outcome is depicted as “positive.’ Together with the tabular part, a brief inference was written for

each article which was finally checked for grammar and plagiarism. Each of the data pages starts with a brief introduction of the stress combination, with information about the number of studies available in different crops for that particular stress combination. Mapping of important traits hosted in SCIPDb was done using the Plant Trait Ontology database (http://www.ontobee.org/ontology/TO) wherever possible.

f. Data integration into the SCIPDb

The frontend user interface was implemented using HTML5, CSS and PHP (version: 7.0.12). The dataset was finally integrated and depicted as an HTML page presented to the users based on the three-level-dropdown selection designed in JavaScript, specific to each plant species. The back-end schema was designed using MySQL, an open-source relational database management system, and stored in MySQL tables (Version: 5.7.17). To provide an interactive interface and enhanced user experience, Bootstrap 4 and jQuery were used. For data page format, please visit the 'phenomics' section.

g. Visualizations

Tableau public desktop (Version 2020.4, https://public.tableau.com/en-us/s/) was used to visualize high dimensional phenomics data in the form of an interactive Treemap. For the creation of radial trees, Flourish studio was used (https://app.flourish.studio/). Several other interactive visualizations offered by Flourish studio like Chord diagram, Sankey Diagram were used for representations of the analysed information. An interactive geographical map was generated using Google My Maps. (https://www.google.com/maps/about/mymaps/).