

1 Plant PhysioSpace: a robust tool to compare stress response 2 across plant species

3 Ali Hadizadeh Esfahani¹, Janina Maß², Asis Hallab², Bernhard M. Schuldt³,
4 David Nevarez¹, Björn Usadel², Mark-Christoph Ott⁴, Benjamin Buer⁴,
5 Andreas Schuppert^{1*}

6 ¹Joint Research Center for Computational Biomedicine, RWTH Aachen University,
7 Aachen, Germany

8 ²IBG-4: Bioinformatics, Forschungszentrum Jülich, 52425 Jülich, Germany

9 ³Independent Consultant

10 ⁴Bayer AG, Crop Science Division, Monheim am Rhein, Germany

11 **Abstract**

12 Generalization of transcriptomics results can be achieved by comparison across experiments,
13 which is based on integration of interrelated transcriptomics studies into a compendium. Both
14 characterization of the fate of the organism as well as distinguishing between generic and specific
15 responses can be gained in such a broader context. There are numerous methods for analyzing
16 such data sets, most focusing on gene-wise dimension reduction to obtain marker genes and
17 gene sets, e.g. for pathway analysis. Relying only on isolated biological modules might lead to
18 missing of important confounders and relevant context.

19 We have developed a novel method called Plant PhysioSpace, which provides the ability
20 to compute experimental conditions across species and platforms without a priori reducing
21 the reference information to specific gene-sets. It extracts physiologically relevant signatures
22 from a reference data set, a collection of public data sets, by integrating and transforming
23 heterogeneous reference gene expression data into a set of physiology-specific patterns. New
24 experimental data can be mapped to these patterns, resulting in similarity scores which provide
25 quantitative likeness of the new experiment to the a priori compendium.

26 Because of its robustness against noise and platform bias, Plant PhysioSpace can be used
27 as an inter-species or cross-platform similarity measure. We have demonstrated its success
28 in translating stress responses between different species and platforms (including single-cell
29 technologies).

30 We also report the implementation of two R packages, one software and one data package,
31 and a shiny web application, which provides plant biologists convenient ways to access our
32 method and precomputed models from four different species.

33
34 **Keywords**— crop science, stress analysis, single cell, computational method, web-tool

35 **1 Introduction**

36 As a consequence of their non-motile nature, plants developed a peculiarly organized yet labyrinthine re-
37 sponse system to external biotic and abiotic stresses. Exploiting this complex system has been playing
38 an important role in achieving sustainable plant protection in agriculture. Instances of tweaking the plant
39 defense system for obtaining better crops are numerous. For instance, priming, i.e. promoting plants to a
40 primed state of defense, has been known, investigated and utilized for decades if not centuries [1, 2]. By
41 exposure to biotic stresses (e.g. microbe-, pathogen-, herbivore-associated molecular patterns) or abiotic
42 stresses (for instance harsh temperatures, drought or damage-associated molecular patterns), plants switch
43 to a primed reinforced defense state. In this primed state, they can display sharper stress response, which

*Correspondence: Schuppert@combine.rwth-aachen.de

B.M.S. and D.N. conceived the original method and research plans; A.S. supervised the method development and calculations; A.H.E. designed the improved method and analyzed the data; D.N., J.M., A.H., M.O., B.U. and B.B. provided technical assistance to A.H.E.; A.H.E. conceived the project and wrote the article with contributions of all the authors; A.S., A.H., B.B., M.O., B.U. and B.M.S supervised and completed the writing. A.S. agrees to serve as the author responsible for contact and ensures communication.

44 in turn results in more robust and resilient organisms. By artificially exposing plants to biotic and abiotic
45 stresses directly, or to some natural or synthetic chemicals which provoke the same defense response, it is
46 possible to engineer tougher plants [3]. Another example of crop engineering is by genetically modifying
47 (GM) plants to attain higher tolerance to stress [4]. Introducing a single gene encoding C-5 sterol desat-
48 urase (FvC5SD) from *Collybia velutipes* to tomato is an instance of GM crop research, and it brings about
49 a drought-tolerant and fungal resistant crop [5, 6]. Obtaining resistance to papaya ringspot virus (PRSV)
50 in transgenic papaya is another famous example. The resistant papaya gains the protection by expressing
51 the PRSV coat protein transgene [7].

52 In research experiments aiming to modify the plant's defense system, such as the examples mentioned
53 above, the stress responses of plants under study are to be thoroughly examined and contrasted to wild
54 types. We argue that a tool, which is capable of quantitatively and dependably measuring the speed and
55 intensity of stress responses in plants, can be of great assistance in this field of research. Hence, we present
56 Plant PhysioSpace, an advanced computational tool based on PhysioSpace [8], for quantitative analysis of
57 stress responses in plants.

58 Sequencing technologies are commonly used for studying the changes in the plants under examination.
59 However, analysis of the results mostly focuses on gene-wise dimension reduction of data to obtain a list of
60 genes, with the rest of the analysis pipeline fixating on the genes in the list. By design, Plant PhysioSpace
61 extracts physiologically relevant information out of intricately convoluted gene expression data without
62 reducing dimensions, providing a direct link from sequencing data to physiological processes. Since it is
63 computationally cheap, the tool is able to train on a vast amount of retrospectively available data, allowing
64 explicit integration of established knowledge and data, eventuating in robust results when testing the method
65 on small data sets generated in specific experiments.

66 Plant PhysioSpace comprises two compartments: the space generation, the algorithm which elicits in-
67 formation from big data, and the physio-mapping, the process with which new data can be analyzed by
68 comparison to the extracted information (Fig.1). Compared to the machine learning nomenclature, space
69 generation is analogous to *training* and physio-mapping to *testing*.

70 In this study, we focused on the application of our novel method in stress response analysis. As one of
71 the fiercest adversaries of plants, biotic and abiotic stresses take a toll on commercial agriculture. Plant
72 PhysioSpace can aid in engineering impervious crops, by quantitatively analyzing the effect of a new mutation
73 or treatment on plant's resistance.

74 Another long-lasting question in the field of stress response research is the potential heterogeneity in
75 response among cell types under the stress. Generally, sequencing is done on thousands to millions of cells,
76 revealing only the average effect on the bulk tissue, thus lacking the direct assessment of cells. The individual
77 and unique role of cell types shape the function of their parent tissue. Hence, by careful examination of
78 stressed tissue cells, the difference between their behavior, and the in-between interplay among them, one
79 can gain new insights into the complex mechanisms shaping the plant stress response.

80 Since 2009, more and more single cell data sets are becoming available publicly. As with other new
81 technologies, the focus is mainly on human and animal tissue sequencing. Lack of data availability is espe-
82 cially true for plant studies on account of processing tissues with cell walls has been a bothersome obstacle
83 for single cell technologies, as they mostly result in low capture rates. But recent leaps in single cell se-
84 quencing technologies, e.g. the 10X platform, increased the resolution of single cell data, eventuating in a
85 few plant single cell experiments [9, 10, 11, 12, 13]. Mostly, scRNA-seq studies follow the same analysis
86 pipeline [14, 15]. In a nutshell, the highest variable genes are selected and gone through principal com-
87 ponent analysis (PCA) and t-distributed stochastic neighborhood embedding (tSNE) or uniform manifold
88 approximation and projection (UMAP) to demonstrate the underlying structure. Subsequently, Clustering
89 or regression algorithms are used to identify biologically relevant groups (e.g. groups of cells with a similar
90 response), or trends (e.g. pseudotemporal axis of cell development), from the underlying data structure [16].
91 Although such technologies as 10X made plant single cell sequencing possible, they are far from perfect. For
92 instance, compared to bulk sequencing technologies, technical noise has a higher interference on single cell
93 reads, which calls for developing sophisticated bioinformatic analysis tools to handle those interferences.

94 This paper has been organized in the following way: It begins with a brief explanation of the Plant Phys-
95 ioSpace algorithm, which includes a review of the already published method, plus the modifications adapting
96 the method to the field of plant stress research. The paper will then go on to the benchmarking section,
97 in which the method's performance is assessed in translating stress response among different experiments,
98 platforms, and species. Benchmarking is followed by two application showcases, in which we demonstrate
99 two Plant PhysioSpace use-cases: investigating time-series data from biotic-stressed wheat, and analyzing a
100 heat-stressed single cell data set. Finally, the discussion gives a brief summary and critique of the findings.

101 2 Materials and Methods

102 2.1 Data Preparation

103 While setting up a PhysioSpace matrix, our method requires extensive training data for achieving adequate
104 robustness. This training data can be retrieved from retrospective data sets. To that end, we curated more

than 4000 plant stress response gene expression samples from GEO¹ and SRA². More specifically, 2480 *A. thaliana* (*Arabidopsis thaliana*) array samples, 967 *A. thaliana* RNA-seq samples, 146 *Oryza sativa* array samples, 172 *Glycine max* array samples, and 104 *Triticum aestivum* array samples were used for space generation. Each sample is annotated with a label from a stress set. In this study, samples are divided into Aluminum, Magnesium, Biotic, Cold, Drought, FarRed, FeDeficiency, Genotoxic, Heat, Herbicide, Hormone, Hypoxia, Light, LowPH, Metabolic, Mutant, Nitrogen, Osmotic, Radiation, Salt, Submergence, UV and Wounding stress groups. For samples which underwent more than one stress, new labels were generated by concatenating existing labels from the stress set. For example, 'Biotic.Drought' designates a sample which sustained both Biotic and Drought stresses.

Samples corresponding to each species are normalized in bulk to remove the batch effect. We used robust multi-array average or RMA [17] for normalizing microarray RAW data files and a pipeline consisting of Fastq-dump, Trimmomatic [18], Star aligner [19] and featureCounts [20] to derive counts from SRA records.

2.2 PhysioSpace Method

PhysioSpace is a supervised dimension reduction method, which aims to extract relevant **physiological** information from big data sets and store it in a mathematical **space** [8]. The method can be divided into two main steps: space generation and physio-mapping (Fig.1).

2.2.1 Space Generation

After preparing the data, we derive the physiologically relevant information from normalized data and store this information in a mathematical space. This step is comparable to *training* in machine learning terminology. Space generation is done in two stages: space extraction and space trimming. The former stage is identical to the method described previously in [8]. However, the latter, space trimming, is a novel addition for adapting the method for studying plant stress response.

Space Extraction In the PhysioSpace method, all samples are analyzed contrastively, i.e. using differential expression analysis. "Space" is a matrix which is built upon reference data. In this paper, reference data contains all *Arabidopsis* array samples that are measured by the Affymetrix *Arabidopsis* ATH1 Genome Array³. For each stress group in each data set, gene-wise fold changes are calculated between stressed plants and their corresponding controls. The fold changes fill one column of the space matrix. This generated matrix, which we call reference space (S_r), contains all stress-relevant information represented in the reference data. In addition to S_r , we calculated the mean reference space ($\overline{S_r}$). For constructing $\overline{S_r}$, for each stress group, the gene-wise mean value of fold changes in S_r is calculated and stored in a column in $\overline{S_r}$. More detailed information, as well as a step-by-step guide for creating S_r and $\overline{S_r}$, are provided in supplementary file 1.

Space Trimming The stress grouping in this study is done based on the expert annotation provided alongside public data sets. Therefore, this grouping doesn't necessarily reflect the different classes of biological mechanisms that shape the plant response spectrum. There are groups of biologically-related stresses, which in turn make some stress responses very similar in their full genome signature. Logically, stresses to which plants respond using the same common mechanisms and pathways, have similar gene expression fingerprints. On the other hand, stresses have significantly different gene expression patterns when few to no common genes are involved in their corresponding stress responses.

From the mathematical point of view, the distance between distinct stress responses manifests itself in the collinearity of axes of the extracted space. Collinearity in a mathematical space is a source of redundancy, and in our application, can result in lower accuracy and robustness.

We came up with a new algorithm named space trimming: an unsupervised approach which in combination with space extraction, makes up a hybrid method that can detect new groups of stress responses. We call these new-found groups meta-stress groups.

Space trimming uses a consecutive combination of hierarchical clustering and leave-one-out cross-validation (LOOCV) to remove the aforementioned redundancy from a space. Space trimming consists of three steps:

1. Clustering and cross-validation analyses are done on the space under study, and a dendrogram based on the calculated similarities is constructed.
2. Groups of stresses that are close and have low accuracy are combined to make meta-stress groups. Groups that merge under the 50% of the maximum height in the dendrogram (i.e. groups with the distance of 50% of the maximum distance or lower) are considered close, and groups of stresses that mostly have an accuracy of less than 0.7 are considered low in accuracy.
3. Any newly-generated meta group that has at least the same performance as its subgroups is kept. All other meta groups are reverted back to their former groups.

¹<https://www.ncbi.nlm.nih.gov/geo/>

²<https://www.ncbi.nlm.nih.gov/sra/>

³<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL198>

160 We applied the Space trimming algorithm to the reference space S_r , generated in the last section, from
161 Arabidopsis microarray data (Fig.2).

162 In LOOCV, by definition, one sample is left out for testing. In our LOOCV scheme though, we left out
163 one GSE (GEO⁴ series) data set in each cycle. Due to batch effects, even with proper preprocessing, samples
164 from the same GSE set tend to be similar. With leave-one-GSE-out cross-validation, we make sure that the
165 stress response from different data sets could be successfully matched together.

166 In each iteration, one GSE data set is chosen as the test set, it is mapped to the rest of the data sets,
167 the training set, and it is counted as a successful match if the analyzed test data and its most similar data
168 set from training set undergone the same stress group. Using the confusionMatrix function from the caret
169 package [21] in R⁵, matching accuracy and robustness of the method is evaluated (Fig.2A). With an overall
170 balanced accuracy of 0.43, a Cohen's kappa of 0.385, and an accuracy p -value of 7.35×10^{-42} , PhysioSpace
171 could successfully match the samples going through the same stress group.

172 As expected, clustering analysis exposed the similarities among different stress responses. For instance,
173 responses to Osmotic, Drought and Salt stresses seem to have common underlying activated gene groups. Or
174 regarding Biotic, Hormone, and Biotic.Hormone (double stress), their close proximity points toward a very
175 similar stress response. They also predominantly have lower accuracy comparable to other stress groups
176 (Fig.2A). This led us to the assumption that these groups of stress responses share one or few underlying
177 defensive mechanisms, such as an innate immune response.

178 Merging the similar stress groups and constructing the meta-stress groups result in an improved perfor-
179 mance of the method (Fig.2B). We constructed three new meta-stress groups: BioMone, which comprises
180 of Biotic, Hormone, and Biotic.Hormone stress groups, DrouSaTic, that was built by combining Drought,
181 Salt, and Osmotic stresses, and LighUV, which is made by merging Light and UV stresses (Fig.2B). Redoing
182 the LOOCV on the new grouped space demonstrates the performance gain, with an accuracy of 0.57 and a
183 Cohen's kappa of 0.49, which increased 0.14 and 0.105 respectively in comparison to the classical grouping
184 of stresses. And the accuracy p -value stays significant, as it is equal to 2.91×10^{-39} .

185 The resulted space, which we call meta-reference space or S_{mr} , and its successive mean space which we
186 denote by \bar{S}_{mr} , are the spaces we use as the reference throughout the result section, though it is possible to
187 use S_r and \bar{S}_r as a reference too, for example in cases that individual stresses are needed to be characterized,
188 e.g. if Salt and Drought stresses are needed to be studied separately.

189 2.2.2 Physio-mapping

190 After acquiring a space from known training data, we can map new data from any technology or species
191 back into the space and find similarities between the new unknown information and the known training data.
192 Physio-mapping is a nonlinear, model-free mapping, designed to take advantage of omics data structures and
193 to compensate biases from heterogeneous assessment protocols. Omics are mostly framed in high-feature
194 low-sample arrangements. In most cases, the majority of features in these types of data sets are intrinsically
195 dominated by noise rather than physiologically informative facts about the samples under study. Results
196 commonly acquired by differential expression analysis are great examples of this phenomenon; in most cases
197 of differential expression analyses, only a small proportion of features in an omics data set can be found to
198 be significantly different, i.e. correlated with circumstances that are being studied.

199 Presuming this assumption, the mapping is done by taking the following steps:

200 1. Either

- 201 (a) A new space is extracted from the new data. This means that for each gene in each stress case,
202 a fold change is calculated by modeling the gene behavior under the respective stress given the
203 control. We call this new input space S_i .
- 204 (b) For each stress type γ , genes are sorted from the lowest to highest fold change.
- 205 (c) N percent lowest and highest genes are selected as $L_L(\gamma)$ and $L_H(\gamma)$ for each γ . N is a user-
206 defined parameter. In this paper, it is between 3 to 5 percent.

207 or

208 Differential expression analysis is done on the input data, and for each stress type γ , down- and
209 up-regulated gene sets are calculated, which are called $L_L(\gamma)$ and $L_H(\gamma)$, respectively.

210 2. For each axis on the reference space (i.e. each column in \bar{S}_{mr}), a statistical test is performed between
211 $L_L(\gamma)$ and $L_H(\gamma)$ gene groups to form the PS matrix:

$$PS = \begin{bmatrix} ps_{11} & ps_{12} & \dots & ps_{1t} \\ ps_{21} & ps_{22} & \dots & ps_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ ps_{p1} & ps_{p2} & \dots & ps_{pt} \end{bmatrix} \text{ with } ps_{k\gamma} = \text{signed log}_2(MWW_{p\text{-value}}(s_{L_H(\gamma)k}, s_{L_L(\gamma)k})) \quad (1)$$

⁴<https://www.ncbi.nlm.nih.gov/geo/>

⁵<https://www.r-project.org/>

212 In equation 1, ps_{km} is the physio score between the m^{th} sample and the k^{th} column of reference space
213 $\overline{S_{mr}}$. ps_{km} physio score value shows how similar the m^{th} sample of S_i is to the k^{th} column in $\overline{S_{mr}}$.
214 $MWW_{p\text{-value}}(a, b)$ is a function that calculates the p -value of a Mann–Whitney–Wilcoxon statistical test
215 (also known as Mann–Whitney U test or Wilcoxon rank-sum test) between a and b. $s_{L_H(\gamma)k}$ and $s_{L_L(\gamma)k}$
216 are the sets of values in k^{th} column, and $i \in L_H$ and $i \in L_L$ rows of $\overline{S_{mr}}$, respectively. And *signed log₂(x)*
217 is $\begin{cases} -\log_2(x), & \text{if } \text{mean}(s_{L_H(\gamma)k}) \geq \text{mean}(s_{L_L(\gamma)k}) \\ \log_2(x), & \text{otherwise} \end{cases}$

218 PS is a physio score matrix, containing similarity values of all input samples to all axes (i.e. columns)
219 in the reference space. This physio score matrix is the output of Plant PhysioSpace, the scores quantify the
220 intensity of different stress types in different samples, and they are comparable within the same physio score
221 matrix⁶. To elaborate on how to interpret the physio scores, we analyzed the GSE13739⁷ [22] data set from
222 GEO using Plant PhysioSpace. In GSE13739, the responses of wild type and mutant *Arabidopsis thaliana* to
223 *Golovinomyces orontii* (powdery mildew) are studied till 7 days post infection. The mutant in this study is
224 Salicylic acid biosynthetic mutant *ics1*, which is expected to be more resilient compared to the wild type. We
225 plotted the physio score matrix (calculated using equation1) as a heatmap (Fig.3), with samples in columns
226 and stress groups in rows of the heatmap, and the color showing the magnitude of each stress in each sample.
227 Based on the experiment description, we anticipated 1) the plants to behave as they are under Biotic stress,
228 and 2) the mutant plants to have a milder response compared to the wild type since they are more resilient.
229 The results show similar dynamics in both wild type and mutant groups of samples: BioMone (Biotic or
230 Hormone) is the dominant stress response, and also its corresponding score increases with time, with wild
231 types reaching a stronger response levels compared to the mutants. Our results are aligned with what we
232 expected from the experiment. In addition, they show that physio scores are comparable between samples
233 (row-wise in the heatmap), as well as among stress groups (column-wise in the heatmap).

234 For inter-species mapping, for instance in analyzing new data from *Oryza sativa* using a space generated
235 from *A. thaliana*, we resorted to orthologous genes. By using the ideal assumption of orthologs to have iden-
236 tical biological roles in all species, we mapped genes to their orthologs in cases with interspecies translation.
237 For the mapping, we utilized the Basic Local Alignment Search Tool BLAST [23]. More specifically, we used
238 the BLASTn tool to align sequences from *Oryza sativa*, *Glycine max*, and *Triticum aestivum* against *A.*
239 *thaliana*, with $\alpha = 10^{-2}$ as the cutoff for the "expect value". Expect value, also known as expectation value
240 or e-value, represents the number of random alignments with scores equivalent to or better than the resulted
241 match [24]. In other words, the expect value depicts the statistical significance of the alignment. Therefore,
242 in cases with multiple matching, the gene with the lowest expect value is chosen as the match, to have a
243 one-to-one matching between genes from different species and *A. thaliana*.

244 3 Results

245 3.1 Stress Space Verification by GO Analysis

246 For substantiating the authenticity of stress information collected using our space generation process, we
247 utilized the gene list analysis section of PANTHER [25, 26]. For each stress group in our generated mean
248 meta-stress space $\overline{S_{mr}}$, we selected the genes with an absolute score value of more than one and a half⁸,
249 and tested this gene list by PANTHER overrepresentation Test, against Gene Ontology (GO) biological
250 processes (Fig.4 and supplement files 2 and 3). From 15 different stress groups, 11 were found to have the
251 GO terms corresponding to the same stress enriched, with significant corrected p -values of less than 0.001.

252 3.2 Inter-Technology Translation

253 Next generation sequencing (NGS) has revolutionized the biological sciences. Its speed, cost and data
254 quality outpaced the older DNA-microarray technology, which is why NGS became the standard method to
255 study transcriptomes. Yet, microarrays were used for RNA quantification for decades. The vast microarray
256 backlogs have the potential to grant an invaluable resource for new biological studies. Unfortunately, the
257 measurement technology has an inevitable impact on the transcript measurement levels and the distribution
258 of the resulting data.

259 Data derived from different platforms are distinctly different. Hence, there are numerous methods to
260 translate measurements from one technology to another [27, 28]. Moreover, with the third generation se-
261 quencing right around the corner (PacBio and Nanopore, to name a few [29]), there is a high demand for
262 computational methods capable of transferring useful information between different measurement technolo-
263 gies.

⁶The possible range of the physio scores is dependent on the number of shared features (i.e genes) between the test data and the reference space. Therefore, if in two different mappings, the numbers of shared features are the same, the scores from the two different resulted physio score matrices are comparable as well.

⁷<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13739>

⁸For BioMone and FeDeficiency stresses with cutoff of 1.5, less than 10 genes were selected, which is too small of a set for list enrichment analysis. Hence in these two cases, cutoff is reduced to one.

264 Since PhysioSpace utilizes the differential expression relations of the genes and not absolute values for
265 space generation and mapping, it can translate between each and any technology, as long as there exists a
266 proper method for detecting the differentially expressed genes in the mentioned technology. As proof for
267 this claim, we mapped more than 900 RNA-seq samples into the microarray space S_{mr} (Table 1). Our
268 method can map the same stress type from microarray to RNA-seq data with 78 percent accuracy. We
269 also calculated the probability of acquiring this accuracy by chance, by randomly permuting the sample
270 labels and calculating the random accuracy. The performance of our method is significantly higher than any
271 random accuracy we acquired⁹, with a p -value of less than 10^{-7} .

Method	RNA-seq to DNA-array, both <i>A. thaliana</i>	Rice to <i>A. thaliana</i> , both DNA-array	Soybean to <i>A. thaliana</i> , both DNA-array	Wheat to <i>A. thaliana</i> , both DNA-array
Plant PhysioSpace	Accuracy 0.78 p -value $< 10^{-7}$	0.59 3×10^{-7}	0.57 0.002	0.23 0.015
GSEA	Accuracy 0.68 p -value $< 10^{-7}$	0.30 3×10^{-4}	0.10 0.42	0.06 0.97
WTCS	Accuracy 0.69 p -value $< 10^{-7}$	0.57 $< 10^{-7}$	0.12 0.21	0.13 0.28
Pearson correlation	Accuracy 0.73 p -value $< 10^{-7}$	0.48 $< 10^{-7}$	0.32 0.12	0.14 0.17
Spearman correlation	Accuracy 0.62 p -value $< 10^{-7}$	0.49 $< 10^{-7}$	0.29 0.21	0.27 4.79×10^{-5}
Euclidean distance	Accuracy 0.30 p -value 0.83	0.27 0.03	0.21 0.62	0.39 $< 10^{-7}$

Table 1: Stress translation between platforms and Species. In each column, the best performer is marked in red.

272 3.3 Inter-Species Translation

273 Although not agriculturally relevant, Arabidopsis is arguably the most investigated species in plant sciences.
274 Its availability, compact size, and fast growth made it an ideal model species. Nevertheless, there are
275 significant differences between the Arabidopsis plant model and crop plants, necessitating procedures for
276 converting well-studied physiological knowledge, e.g. regarding plant response to different types of stress,
277 from Arabidopsis to crops. In this section, we show how Plant PhysioSpace can be utilized for this purpose.

278 We chose three of the most commercially relevant crops to study: *Oryza sativa* (rice), *Glycine max*
279 (soybean) and *Triticum aestivum* (wheat). For each crop, more than 100 microarray samples of stress
280 response experiments were curated, normalized, preprocessed and mapped to the Arabidopsis space S_{mr} . For
281 *Oryza sativa* and *Glycine max*, Plant PhysioSpace achieved respective accuracies of 59 and 57 percent, both of
282 which were significantly higher than any accuracy earned by chance. On the other hand, translation of stress
283 response from *Triticum aestivum* DNA array data to *A. thaliana*, with an accuracy of 23 percent and a p -
284 value of 0.015, was not successful (Table 1). In section 3.5 of this paper, we provided a thorough investigation
285 into wheat to Arabidopsis translation, hypothesized and examined the reason behind the translation failure,
286 and provided solutions for fixing it.

287 3.4 Benchmarking Plant PhysioSpace against Other Methods

288 We used the results from inter-technology and -species stress response translation to benchmark our method.
289 Plant PhysioSpace is compared to the most common approaches used in bioinformatics for measuring rela-
290 tions between two or more gene expression samples: Gene Set Enrichment Analysis (GSEA) [30], Weighted
291 Connectivity Score (WTCS), which is an advanced version of GSEA used in connectivity map [31], Pearson
292 and Spearman correlations, and Euclidean distance. For each method, fold change values of samples, from
293 different technologies and species, are calculated and used for finding the similarities between samples. Based

⁹The highest acquired accuracy from 10,000,000 random runs for RNA2DNA translation was 52% (minimum = 12.08%, first quartile = 28.12%, median = 30.87%, third quartile = 33.56% and maximum = 52.35%).

294 on our results in inter-species and -platform mapping, Plant PhysioSpace could outperform other methods
295 in all scenarios, except in mapping from wheat to *Arabidopsis* (Table 1).

296 3.5 In-depth Investigation of Wheat Stress Response

297 The poor performance of our method in translating stress response from *Triticum aestivum* to *A. thaliana*
298 may have potentially derived from the microarray used to measure the wheat gene expressions. All microar-
299 ray samples of *Triticum aestivum* in this study are generated by using Affymetrix Wheat Genome Array¹⁰.
300 Not only the aged technology could potentially deter the accuracy of transcription measurements, but also
301 as a polyploid, the complex genetics of wheat would make the task of measuring its RNA levels troublesome.
302 Moreover, among all translations tested, wheat microarray provides the least amount of mapped orthologs:
303 when translating from RNA-seq to microarray, we could find 12146 common features (i.e. genes). This
304 amount was lower for both translations of rice and soybean to *Arabidopsis*, as the ortholog mapping could
305 find 9455 and 9452 common features for them, respectively. And, in translation from wheat to *Arabidop-*
306 sis, 6202 common features could be found using the ortholog mapping, which is the least among the four
307 translations. These common feature counts are correlated with the translation performance, considering the
308 RNA-seq to DNA-microarray has the highest and wheat to *Arabidopsis* has the lowest accuracy, while rice
309 and soybean mappings are somewhere in between (Table 1). This correlation suggests one possible reason
310 behind the failed wheat translation is the low amount of orthologs. Fortunately, advances in NGS gave rise
311 to wheat stress response data sets with higher precision and a wider range of sequence reads.. In this section,
312 we repeated the wheat-to-*Arabidopsis* translation from inter-species analysis, except with wheat RNA-seq
313 instead of microarray data.

314 We turned to the Wheat Expression Brower [32, 33] as the source of *Triticum aestivum* RNA-seq data.
315 From this source, we queried all data sets which study stress response, contain more than 30 samples, and
316 include controls. Upon the first inspection, 8813 orthologs could be found between these data sets and the
317 genes available in our *Arabidopsis* reference space, which in addition with the improved data quality provided
318 by RNA-seq, increases the prospect of Plant PhysioSpace correctly detecting and quantifying the stress in
319 these data sets. We mapped these data sets into mean meta-reference space \overline{S}_{mr} , and plotted Physio scores
320 of three stress groups with the highest values (Fig.5).

321 In the experiment set ERP013829, wheat response after inoculation with fungal pathogen *Fusarium*
322 *graminearum* is measured through time [34]. For this experiment, Plant PhysioSpace correctly predicts that
323 wheat is experiencing BioMone (Biotic and Hormone) stress. In addition, the diachronic rise in the response
324 indicates how Physio scores are quantitatively comparable (Fig.5A).

325 In the data set ERP013983, responses of two different mutants of wheat are studied to wheat yellow
326 rust pathogen *Puccinia striiformis f.sp. tritici* (PST). The authors focused on the pathogen suppression of
327 basal defense in plants [35]. From their results, they deduced the pathogen overcame the defense by rapidly
328 suppressing the genes involved in chitin perception on day 2 after inoculation. In the susceptible interaction,
329 this provides the possibility of invasion and colonization, while in resistant plants, this suppression is quickly
330 reverted. Plant PhysioSpace results expose this mechanism correctly: Both plant types are inducing a
331 BioMone response on day 1, followed by a suppression of the plant defense response on day 2. Eventually,
332 the quick resurgence of intense BioMone response in resistant wheat helps it in withstanding PST, while the
333 reaction of the susceptible trait might be too slow to deflect the pathogen (Fig.5B).

334 Plants in ERP009837 went through the infection cycle of the hemibiotrophic fungus *Zymoseptoria tritici*.
335 Similar to ERP013829, plants respond to the pathogen with a dominant BioMone response (Fig.5C). Al-
336 though unlike ERP013829, in which the experiment spanned a few hours, in ERP009837 plants were studied
337 for a longer period. Physio scores suggest that wheat responds to the presence of the pathogen by increasing
338 its BioMone response. This response starts to degrade from day 14, which is in alignment with the original
339 publication of the data set [36], in which the authors state that on day 21, plant tissue is completely defeated.

340 In SRP048912, the responses of two different traits of resistant and susceptible wheat to *Fusarium* crown
341 rot are studied [37]. Only two different time points are included in this experiment: 3 and 5 days after
342 inoculation (dai). Plant PhysioSpace results correctly suggest the most dominant stress response present in
343 plants is BioMone, with the resistant wheat having a stronger response than the susceptible (Fig.5D).

344 Among the Wheat data sets we analyzed in this section, ERP003465 is arguably the most complex, and
345 consequently most interesting as a testing scenario for our method. ERP003465 examined the behavior of
346 5 different genotypes under the disease pressure of *Fusarium graminearum* [38]. Two well-validated and
347 highly reproducible QTLs (quantitative trait loci), *Fhb1* and *Qfhs.ifa-5A*, are studied from samples taken
348 30 and 50 hours after inoculation (hai). Five different genotypes were investigated: CM-82036, a progeny
349 of the resistant Sumai-3, and four near-isogenic lines (NILs) bearing either, both, or none of the resistant
350 alleles *Fhb1* and *Qfhs*. Among the four, NIL1 is a mutant with both QTLs, expected to have the highest
351 resistance after CM-82036, NIL2 and NIL3 are mutants harboring *Fhb1* and *Qfhs* QTLs respectively, with
352 both predicted to behave moderately resistant, and NIL4 missing both QTLs, and is likely to be susceptible.

353 Data analysis in the original paper was mainly based on differential expression analysis. As a first
354 step, the total number of differentially expressed genes for each genotype at each time point was taken as
355 a surrogate for stress response intensity to *Fusarium graminearum*. In the next steps, the weighted gene

¹⁰<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3802>

356 co-expression network analysis (WGCNA) was used to detect clusters of genes with similar patterns, and
357 Gene Ontology analysis was utilized to infer the role of each cluster in the stress response.

358 Being able to quantify the intensity of each stress type at each time point, Plant PhysioSpace can
359 provide much more insight into the characteristics and dynamics of the stress responses that are at play in
360 the ERP003465 experiment (Fig.5E). As this data set encompasses a high number of samples distributed
361 between only two time points, we plotted the results as a bar graph. And because the results cover a wide
362 range of values for this experiment, we used log-scaled Physio scores in the graph, and replaced values smaller
363 than one by one (i.e. zero in log-scale).

364 Among the concluding remarks in the original paper, some are in concordance with the results from our
365 method. For example, lines lacking *Qfhs.ifa-5A* are regarded as “slow responders” by the original authors,
366 since they lack resistance against initial infection inferred by *Qfhs.ifa-5A*. This lack of early response can be
367 seen in our results (Fig.5E): lines lacking *Qfhs.ifa-5A*, i.e. NIL2 and NIL4, have no BioMone (Biotic and
368 Hormone) stress response at the early time point, while NIL1 and NIL3 show a considerable BioMone stress
369 response at the same time point. Another remark from the original paper suggested that a lack of timely
370 defensive reaction could result in a higher infection in a later time, and consequently stronger response,
371 and vice versa: a quick response may reduce the intensity and infection at a later time. This can be seen
372 in the contrasting response dynamics of NIL1 versus the other lines (Fig.5E). NIL1 possesses both QTLs:
373 *Qfhs.ifa-5A* ensures an early and fast stress response, evident on 30 hai time point. And a strong follow up,
374 courtesy of *Fhb1*, results in a non-existent BioMone response at 50 hai. NIL3 contains *Qfhs.ifa-5A*, so it
375 benefits from a quick response at 30 hai, but due to the absence of *Fhb1*, it cannot be rid of infection at 50
376 hai, evident by the high BioMone response at that time point. As mentioned, lines NIL2 and NIL4, which
377 lack *Qfhs.ifa-5A*, do not have an early response and have to play catch up with other lines on the later time
378 point.

379 Although many conclusions that could be derived from our method are similar to the ones from the
380 original publication, there are some discrepancies between the two groups as well. For instance, in most
381 samples, Wounding stress response is not only present, but it is even stronger than BioMone response in
382 some cases. This is in contrast with the original paper, in which it is mentioned that inoculation was done
383 cautiously without wounding the tissue. Interpretation of CM-82036 defensive behavior is another point of
384 difference between our method and the results from the original paper. Kugler et al. construed the high
385 number of differentially expressed genes (DEG) at 30 hai as a sign of strong early response for CM-82036,
386 even stronger than NIL1 and NIL3. They followed up by studying specific gene families that are relevant to
387 defense mechanisms, such as UGTs and WRKYs, and showed more DEGs from these families can be found
388 at 30 hai in CM-82036 versus other lines. This finding is different from what we can interpret using our
389 method: Although CM-82036 exhibits BioMone response at 30 hai, the magnitude is somewhere between
390 fast responder lines, that is NIL1 and NIL3, and slow responders, i.e. NIL2 and NIL4.

391 We speculate the main reason for the aforementioned inconsistencies is the particular way the prepro-
392 cessing was done in the original paper. In their preprocessing, Kugler et al. mapped the reads to a list of
393 barley high confidence genes and only used the reads with a possible match. This step drastically reduces
394 the number of analyzed transcripts, and also discards wheat-specific genes with no barley homologs. Our
395 method is designed for high-dimensional data, preferably data from the whole genome, therefore the specific
396 preprocessing of this data set might have reduced the performance of Plant PhysioSpace. We should also
397 mention that stress responses are not mutually exclusive; A plant can display multiple different responses
398 at the same time, some of which may even share part of their biological pathways. *Fusarium graminearum*
399 could have damaged the plant tissue at some point, which explains the existence of wounding response
400 alongside BioMone.

401 Albeit the mediocre results of the last experiment, in this section we showed how, in 4 out of 5 data sets,
402 Plant PhysioSpace could:

- 403 1. correctly identify the type of stress plants are going through.
- 404 2. accurately relate the response from RNA-seq test data to DNA-array trained models.
- 405 3. rightly translate *T. aestivum* stress response to *A. thaliana*.

406 3.6 Plant PhysioSpace Application in Single Cell Analysis

407 Single cell technologies facilitate investigating transcription profiles in single cell resolution, in order to
408 perceive the genetic basis of each cell type and its function. Although relatively new, more and more plant
409 single cell data sets are becoming available to the community [9, 10, 11, 12, 13]. For now, most sequence
410 data sets are focused on *Arabidopsis* roots. They try to gain an in-depth understanding of transcription
411 patterns of different cells in different developmental stages of wild-type non-stressed plant roots. To our
412 knowledge, the only publication in which stressed single cells were sequenced is the paper by Jean-Baptiste
413 et al. [12]. In this work, 38°C heat stress was applied to 8-day-old seedlings for 45 minutes. Subsequently,
414 roots of the seedlings were harvested, along with the roots of age- and time-matched control seedlings. The
415 authors could capture and sequence 1,009 cells from the stressed group and 1076 from the control group.
416 For processing the sequencing results, they followed the usual single cell analysis pipeline: PCA, UMAP
417 and clustering, followed by differential gene expression analysis on clusters and enrichment tests on genes
418 related to heat-shock. The results show the “promise and challenges inherent in comparing single cell data

419 across different conditions and treatments". In this section, we demonstrate how a dedicated method, such
420 as Plant PhysioSpace, can bring forth more benefits than using the methodological norms.

421 To analyze the single cell data set, we used the gene-wise mean value of all control cells as the reference,
422 calculated fold changes for each single cell, and fed those fold change values into the Plant PhysioSpace
423 pipeline (Fig.6). Regardless of the cell type, heat-stressed single cells had significantly higher heat stress
424 scores, compared to control single cells (Fig.6A). For studying the heat-induced cell type disparity, we overlaid
425 heat stress scores on UMAP and tSNE plots (Fig.6B&6C). In both tSNE and UMAP plots, coordinate values
426 calculated in the original paper of Jean-Baptiste et al. were used. As a result, cells are bundled in cell type
427 clusters in the UMAP plot, while in the tSNE plot, cells are clearly separated into two big clusters of control
428 and stressed. Although, inside these two big clusters, sub-clusters of different cell types are evident (Fig.S1).
429 On the UMAP plot on the other hand, big clusters represent cell types (Fig.S2), and inside each cell type
430 cluster, groups of control and stressed cells may or may not be distinct, depending on the cell type. For
431 example, in Hair and Non-hair clusters, control and heated cells are separated, while the separation is less
432 pronounced in Stele cells (Fig.S2).

433 To look into the distinct behavior of different cells under stress, we also plotted cell heat scores grouped
434 by the corresponding cell types (Fig.6D&S3). The results show how Hair and Non-hair cells have higher
435 heat scores, which demonstrates how the outer layers of roots are sharper in their response to heat. This
436 finding is in concordance with one of the conclusions in the original paper, in which based on the behavior
437 of the heat-relevant genes, they concluded the three outermost cell layers of the root went through higher
438 levels of changes caused by the heat stress. The authors hypothesized this may be because of more direct
439 exposure of the outer layers to the heat shock, resulting in a quicker and stronger response.

440 Although resulting in generally the same conclusions, in this analysis Plant PhysioSpace provided an
441 advantageous experience for the end-user, through providing:

- 442 1. convenience: unlike the original paper, there was no need for search and curation of heat stress gene
443 clusters, as they are already available in Plant PhysioSpace, as well as clusters for other common
444 stresses.
- 445 2. precision: not only the stress type but also the magnitude of the stress response could be quantified by
446 our method, something which is lacking in traditional gene list enrichment approaches. For example,
447 Plant PhysioSpace results suggest a stronger response in Hair response, compared to Non-hair response
448 (Fig.6D). This inference could not be concluded by the results of traditional methods.
- 449 3. optimization: in one run, our tool calculated responses of 20 different stresses for 2085 single cells, in
450 less than 3 minutes on a 2-core laptop CPU. This swift performance is accomplished by precalculating
451 the stress space, in combination with an optimized mapping algorithm, all of which is readily available
452 for the community to use.

453 3.7 Availability

454 To provide the community with an easy-to-use implementation of our method, we built Plant PhysioSpace
455 into two different R packages: a method package (<https://git.rwth-aachen.de/jrc-combine/PhysioSpaceMethods>)
456 containing functions for generating new spaces and Physio-mapping, and a data package (<https://git.rwth->
457 [aachen.de/jrc-combine/PlantPhysioSpace](https://git.rwth-aachen.de/jrc-combine/PlantPhysioSpace)) comprising plant stress spaces such as \overline{S}_{mr} and \overline{S}_r that were used
458 in this paper.

459 In addition, we made a shiny¹¹ web application of Plant PhysioSpace (Fig.7). We hosted the web app on
460 shinyapp.io (<http://physiospace.shinyapps.io/plant/>), to be freely available to use (under the terms of GPL-3
461 license). We also built a Docker image of the ready-to-use tool (https://github.com/usadellab/physiospace_shiny).

462 3.8 Supplemental Material

463 This paper is accompanied by three supplement files and three supplement figures:

- 464 1. **Supplementary file 1** is a PDF text file that provides a detailed explanation about how "space
465 extraction" is performed.
- 466 2. **Supplementary file 2** is a compendium of all 15 bubble plots related to section 3.1, "Stress Space
467 Verification by GO Analysis".
- 468 3. **Supplementary file 3** is an excel file that contains the results of the PANTHER GO analysis for all
469 the 15 stress groups. In the file, correctly detected GO terms are highlighted, and stress groups that
470 had irrelevant GO terms enriched are labeled.
- 471 4. **Supplementary figures** are related to section 3.6, "Plant PhysioSpace Application in Single Cell
472 Analysis". Supp. figures S1 and S2 are alternative versions of figures 6C and 6B, in which cell types are
473 annotated in place of temperature. And supp. figures S3 is the same as figures 6D, except it contains
474 the control cells physio scores as well.

11 <http://shiny.rstudio.com>

4 Discussion

Gaining proper insight into stress response mechanisms in plants is not only a must for the future of agricultural research, but will prompt advances in the plant research field in general. In this study, we developed an advanced computational method, designed to aid in understanding stress response in plants. The lightweight algorithm allows it to run on either personal computers, or as a web application, making it an ideal tool for experimental quality control, data set annotation, to draw conclusions considering thousands of genes, et cetera.

We built the new method upon a previously published method in humans, called PhysioSpace. We achieved this conformity by curating a multitude of *Arabidopsis* stress response samples to have a rich training data set, adapting the space generation algorithm, i.e. training, to acclimate to the specific characteristics of stress response data in plants, and thoroughly testing against other species and types of data. The results of this study demonstrated that Plant PhysioSpace can be a convenient and practical tool for analyzing new stress response data sets, to apprehend, contrast to state of the art, or to simply quality control.

Notably, our tool could perform adequately even when it was mapping information between different platforms and species. Although, it is crucial to bear in mind these cross translations necessitate for some conditions to be true. In cross-platform translation, it was assumed that with the same experimental setup and samples, there are computational pipelines available which roughly compute the same differentially expressed gene lists regardless of the platform used. And in cross-species mapping, we assumed orthologous genes have the same biological function across all species; evidently impossible to be consistently true for all genes, but a sizable portion of genes have to pass this criterion for the inter-species translation to work.

We demonstrated how Plant Physiospace can provide insights when used for analyzing single cell data sets. Recent advances in single cell technology call for suitable bioinformatic analysis tools, for example for reducing the interfering technical noise [16]. The clear, factual results derived from single cell data analysis in this paper bring a spectrum of applications to mind for the future, especially in the light of approaching plant single cell atlas projects [39, 40].

Plant PhysioSpace may seem paradoxical at first, as it is explained to be a "dimension reduction" method that extracts stress information from gene expression data "without reducing dimensions". Yet, this ostensible contradiction is plausible in actuality, since our method doesn't discard any features from the training data when extracting stress information in the space generation stage (i.e. model training), but only removes the irrelevant features in the physio-mapping (i.e. when applying the method to the new samples). As mentioned before, the majority of genes remain unaltered under stress. We showed this unvarying behavior using the mean meta-reference space $\overline{S_{mr}}$ (Fig.8). Taking the absolute fold change value of 1 as a cutoff, only a small portion of plant genes are changing under stress. To be more specific, from all 22249 genes available in $\overline{S_{mr}}$, only 2905 (~13%) change at least under one stress type. In addition, the number of changed genes is a function of the stress type, ranging from ~0.07% in BioMone (biotic, hormone, or both stresses) to ~5.8% in the Drought.Light (double stress) group (Fig.8A). Moreover, the majority of varying genes are specific to a stress group, as 2175 out of the 2905 genes (~75%) only change in one stress group (Fig.8B). The low dimensionality of the stress responses, along with the highly specific features, may suggest that one can remove the genes with the absolute fold change of less than 1 in the training step, as it seems that keeping all the features till the end is not beneficial, while removing them could result in an even more compressed model. We are against this exclusion, however, on the ground that the removal of genes with an absolute fold change of less than 1 can eliminate relevant information. As proof, we ran the single cell processing of the last section once more, but this time as a reference, we used a "reduced" $\overline{S_{mr}}$, which is a version of $\overline{S_{mr}}$ with all the 2905 varying genes removed (Fig.8C). Although not as high as their counterparts calculated by using the complete $\overline{S_{mr}}$, heat scores calculated using the reduced space are still significantly higher in heated single cells compared to the control cells (Wilcoxon rank-sum test p -value $< 2.2 \times 10^{-16}$), showing that the less varying genes also carry stress-related information. Therefore, we conclude the benefits of keeping all possible features outweigh the advantages that come about by removing them. Having the information from the whole genome at hand is especially helpful when analyzing data sets that don't include a high number of genes, because the overlap between the features present in the model and the new data may become too small. Many new sequencing technologies, the Oxford Nanopore MinION [41] for example, or novel single cell sequencing platforms, provide readings on a limited number of genes, and keeping the whole genome in the trained model can boost the performance of our method on data from the aforementioned technologies. That being said, a proper and clever way of reducing dimensions of the trained spaces, a far more complex way compared to the fold-change cutoff scenario we examined before, has the potential to increase the efficiency and reduce the complexity of our method.

To our knowledge, Plant PhysioSpace is the only computational tool available capable of quantitizing stress response in plant cells. Therefore, it can be used to assess each cell under stress, to grasp an understanding of the complex responses and interplay of cells in plants under stress, and to achieve a comprehensive characterization of plant response to stress as a whole.

536 5 Data Availability Statement

537 All the scripts that generate the results of this paper can be found in <https://git.rwth-aachen.de/jrc-combine/PlantPhysioSpacePaper>.

539 References

- 540 [1] Kenneth S Chester. The problem of acquired physiological immunity in plants. *The Quarterly Review of Biology*, 8(3):275–324, 1933.
- 542 [2] Joseph Kuć. Translocated signals for plant immunization a. *Annals of the New York Academy of Sciences*, 494(1):221–223, 1987.
- 544 [3] Uwe Conrath, Gerold JM Beckers, Caspar JG Langenbach, and Michal R Jaskiewicz. Priming for enhanced defense. *Annual review of phytopathology*, 53:97–119, 2015.
- 546 [4] Asis Datta. Genetic engineering for improving quality and productivity of crops. *Agriculture & Food Security*, 2(1):1–3, 2013.
- 548 [5] Ayushi Kamthan, Mohan Kamthan, Mohammad Azam, Niranjan Chakraborty, Subhra Chakraborty, and Asis Datta. Expression of a fungal sterol desaturase improves tomato drought tolerance, pathogen resistance and nutritional quality. *Scientific reports*, 2:951, 2012.
- 550 [6] Ayushi Kamthan, Mohan Kamthan, Niranjan Chakraborty, Subhra Chakraborty, and Asis Datta. A simple protocol for extraction, derivatization, and analysis of tomato leaf and fruit lipophilic metabolites using gc-ms. *Protocol Exchange*, 10, 2012.
- 552 [7] Dennis Gonsalves. Control of papaya ringspot virus in papaya: a case study. *Annual review of phytopathology*, 36(1):415–437, 1998.
- 554 [8] Michael Lenz, Bernhard M Schuldt, Franz-Josef Müller, and Andreas Schuppert. Physiospace: relating gene expression experiments from heterogeneous sources using shared physiological processes. *PLoS One*, 8(10):e77627, 2013.
- 556 [9] Christine N Shulse, Benjamin J Cole, Doina Ciobanu, Junyan Lin, Yuko Yoshinaga, Mona Gouran, Gina M Turco, Yiwen Zhu, Ronan C O’Malley, Siobhan M Brady, et al. High-throughput single-cell transcriptome profiling of plant cell types. *Cell reports*, 27(7):2241–2247, 2019.
- 558 [10] Tom Denyer, Xiaoli Ma, Simon Klesen, Emanuele Scacchi, Kay Nieselt, and Marja CP Timmermans. Spatiotemporal developmental trajectories in the arabidopsis root revealed using high-throughput single-cell rna sequencing. *Developmental cell*, 48(6):840–852, 2019.
- 560 [11] Kook Hui Ryu, Ling Huang, Hyun Min Kang, and John Schiefelbein. Single-cell rna sequencing resolves molecular relationships among individual plant cells. *Plant physiology*, 179(4):1444–1456, 2019.
- 562 [12] Ken Jean-Baptiste, José L McFaline-Figueroa, Cristina M Alexandre, Michael W Dorrity, Lauren Saunders, Kerry L Bubb, Cole Trapnell, Stanley Fields, Christine Queitsch, and Josh T Cuperus. Dynamics of gene expression in single root cells of arabidopsis thaliana. *The Plant Cell*, 31(5):993–1011, 2019.
- 564 [13] Tian-Qi Zhang, Zhou-Geng Xu, Guan-Dong Shang, and Jia-Wei Wang. A single-cell rna sequencing profiles the developmental landscape of arabidopsis root. *Molecular plant*, 12(5):648–660, 2019.
- 566 [14] Andrew Butler, Paul Hoffman, Peter Smibert, Eifthymia Papalex, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.
- 568 [15] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381, 2014.
- 570 [16] Charlotte Rich-Griffin, Annika Stechemesser, Jessica Finch, Emma Lucas, Sascha Ott, and Patrick Schäfer. Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. *Trends in plant science*, 25(2):186–197, 2020.
- 572 [17] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.

585 [18] Anthony M Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina
586 sequence data. *Bioinformatics*, 30(15):2114–2120, 2014.

587 [19] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe
588 Batut, Mark Chaisson, and Thomas R Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, 2013.

590 [20] Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for
591 assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

592 [21] Max Kuhn. *caret: Classification and Regression Training*, 2020. R package version 6.0-86.

593 [22] Divya Chandran, Yu Chuan Tai, Gregory Hather, Julia Dewdney, Carine Denoux, Diane G Burgess,
594 Frederick M Ausubel, Terence P Speed, and Mary C Wildermuth. Temporal global expression data
595 reveal known and novel salicylate-impacted processes and regulators mediating powdery mildew growth
596 and reproduction on arabidopsis. *Plant physiology*, 149(3):1435–1451, 2009.

597 [23] Grzegorz M Boratyn, Christian Camacho, Peter S Cooper, George Coulouris, Amelia Fong, Ning Ma,
598 Thomas L Madden, Wayne T Matten, Scott D McGinnis, Yuri Merezhuk, et al. Blast: a more efficient
599 report with usability improvements. *Nucleic acids research*, 41(W1):W29–W33, 2013.

600 [24] Nicholas H Bergman. *Comparative genomics*. Totowa (NJ): Humana Press, Springer Science & Business
601 Media; 2007, Available from: <https://www.ncbi.nlm.nih.gov/books/NBK1732/>.

602 [25] Huaiyu Mi, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D Thomas. Panther
603 version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools.
604 *Nucleic acids research*, 47(D1):D419–D426, 2019.

605 [26] Paul D Thomas, Anish Kejariwal, Nan Guo, Huaiyu Mi, Michael J Campbell, Anushya Muruganujan,
606 and Betty Lazareva-Ulitsky. Applications for protein sequence–function evolution data: mrna/protein
607 expression analysis and coding snp scoring tools. *Nucleic acids research*, 34(suppl_2):W645–W650, 2006.

608 [27] Matthew E Ritchie, Belinda Phipson, DI Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K
609 Smyth. limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic
610 acids research*, 43(7):e47–e47, 2015.

611 [28] Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion
612 for rna-seq data with deseq2. *Genome biology*, 15(12):550, 2014.

613 [29] Thomas P Niedringhaus, Denitsa Milanova, Matthew B Kerby, Michael P Snyder, and Annelise E
614 Barron. Landscape of next-generation sequencing technologies. *Analytical chemistry*, 83(12):4327–4341,
615 2011.

616 [30] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel,
617 Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity
618 map: using gene-expression signatures to connect small molecules, genes, and disease. *science*,
619 313(5795):1929–1935, 2006.

620 [31] Aravind Subramanian, Rajiv Narayan, Steven M Corsello, David D Peck, Ted E Natoli, Xiaodong Lu,
621 Joshua Gould, John F Davis, Andrew A Tubelli, Jacob K Asiedu, et al. A next generation connectivity
622 map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452, 2017.

623 [32] Philippa Borrill, Ricardo Ramirez-Gonzalez, and Cristobal Uauy. expvip: a customizable rna-seq data
624 analysis and visualization platform. *Plant physiology*, 170(4):2172–2186, 2016.

625 [33] R Ramírez-González, P Borrill, D Lang, SA Harrington, J Brinton, L Venturini, et al. The transcrip-
626 tional landscape of hexaploid wheat across tissues and cultivars. *Science*, 361(6403):eaar6089, 2018.

627 [34] W Schweiger, B Steiner, Sonia Vautrin, Thomas Nussbaumer, Gerald Siegwart, Mina Zamini, F Jun-
628 greithmeier, Verena Gratl, Marc Lemmens, KFX Mayer, et al. Suppressed recombination and unique
629 candidate genes in the divergent haplotype encoding fhb1, a major fusarium head blight resistance locus
630 in wheat. *Theoretical and Applied Genetics*, 129(8):1607–1623, 2016.

631 [35] Albor Dobon, Daniel CE Bunting, Luis Enrique Cabrera-Quio, Cristobal Uauy, and Diane GO Saunders.
632 The host-pathogen interaction between wheat and yellow rust induces temporally coordinated waves of
633 gene expression. *BMC genomics*, 17(1):1–14, 2016.

634 [36] Jason J Rudd, Kostya Kanyuka, Keywan Hassani-Pak, Mark Derbyshire, Ambrose Andongabo, Jean
635 Devonshire, Artem Lysenko, Mansoor Saqi, Nalini M Desai, Stephen J Powers, et al. Transcriptome and
636 metabolite profiling of the infection cycle of zymoseptoria tritici on wheat reveals a biphasic interaction
637 with plant immunity involving differential pathogen chromosomal contributions and a variation on the
638 hemibiotrophic lifestyle definition. *Plant physiology*, 167(3):1158–1185, 2015.

639 [37] Jian Ma, Jiri Stiller, Qiang Zhao, Qi Feng, Colin Cavanagh, Penghao Wang, Donald Gardiner, Frédéric
640 Choulet, Catherine Feuillet, You-Liang Zheng, et al. Transcriptome and allele specificity associated
641 with a 3bl locus for fusarium crown rot resistance in bread wheat. *PLoS One*, 9(11):e113309, 2014.

642 [38] Karl G Kugler, Gerald Siegwart, Thomas Nussbaumer, Christian Ametz, Manuel Spannagl, Barbara
643 Steiner, Marc Lemmens, Klaus FX Mayer, Hermann Buerstmayr, and Wolfgang Schweiger. Quantitative
644 trait loci-dependent analysis of a gene co-expression network associated with fusarium head blight
645 resistance in bread wheat (triticum aestivum). *BMC genomics*, 14(1):728, 2013.

646 [39] Seung Y Rhee, Kenneth D Birnbaum, and David W Ehrhardt. Towards building a plant cell atlas.
647 *Trends in plant science*, 24(4):303–310, 2019.

648 [40] Robert Petryszak, Maria Keays, Y Amy Tang, Nuno A Fonseca, Elisabet Barrera, Tony Burdett,
649 Anja Füllgrabe, Alfonso Munoz-Pomer Fuentes, Simon Jupp, Satu Koskinen, et al. Expression atlas
650 update—an integrated database of gene and protein expression in humans, animals and plants. *Nucleic
651 acids research*, 44(D1):D746–D752, 2016.

652 [41] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery
653 of nanopore sequencing to the genomics community. *Genome biology*, 17(1):1–11, 2016.

654 [42] Wencke Walter, Fátima Sánchez-Cabo, and Mercedes Ricote. Goplot: an r package for visually com-
655 bining expression data with functional analysis. *Bioinformatics*, 31(17):2912–2914, 2015.

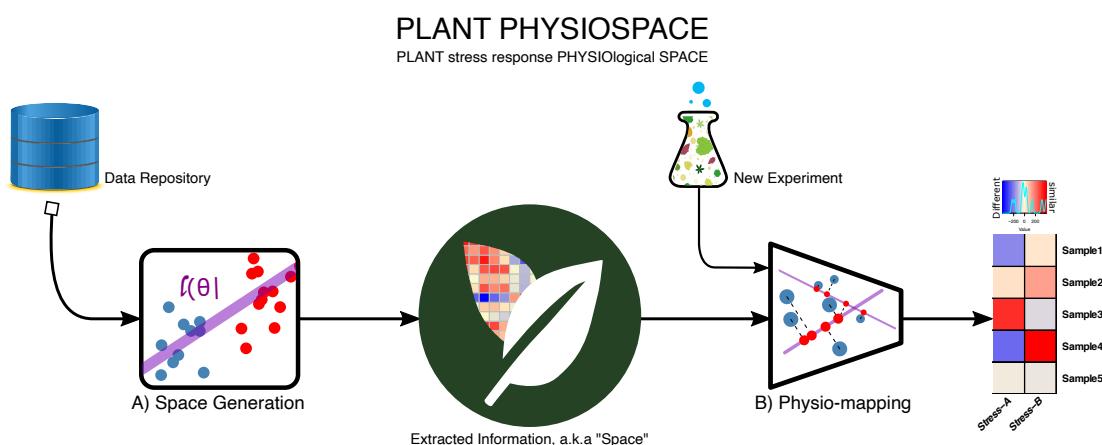


Figure 1: Plant PhysioSpace Overview. The method consists of two main sections: space generation and Physio-mapping. In space generation (A), data from public repositories is processed and its information is extracted. After trimming, the extracted information is stored in matrices called “space”, representing physiology-relevant expression patterns. Physio-mapping (B) uses a space to analyze new, unknown data, for example from a new experiment. The new data is mapped to the generated space, resulting in “similarity” scores that indicate the likeness of the new data to the known physiological processes.

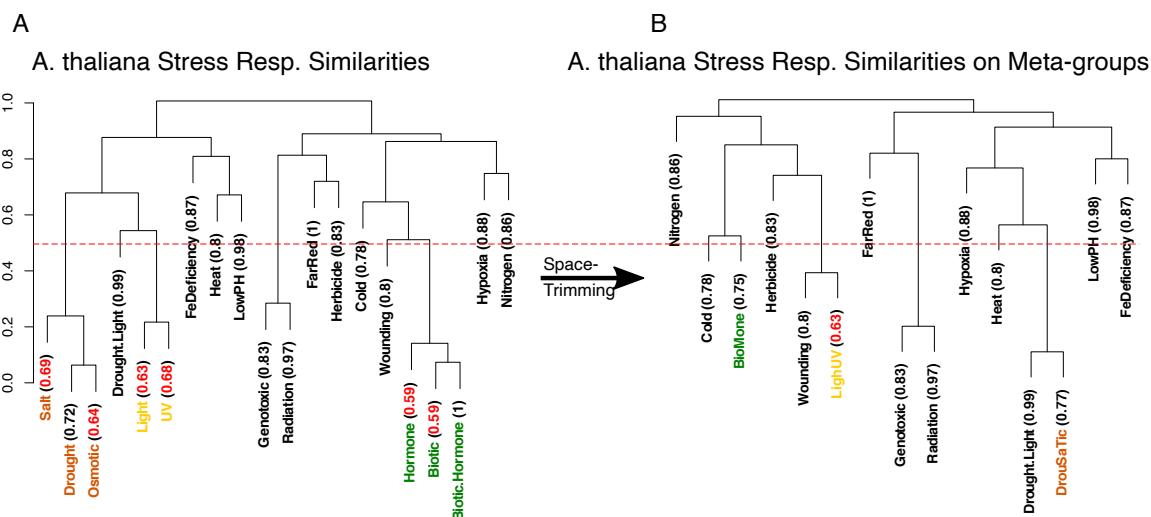


Figure 2: Space Trimming. Stress groups are clustered and for each group, leave-one-out cross-validation accuracy is calculated, written in parenthesis, as shown in panel A. Close groups with low accuracy, written in red, are combined to form new stress groups, called meta-groups, as shown in panel B. Groups are considered close if they merge in the dendrogram in a height lower than 0.5 (50% of maximum height). This cut-off height is shown in the figure with a dashed red line. In this figure, Salt, Drought and Osmotic stress groups, marked with brown color, merge into DrouSaTic meta-group, Hormone, Biotic and Biotic.Hormone groups form BioMone meta-group, written in green, and Light and UV groups combine into LightUV, shown in yellow.

GSE13739 Analysis Using Plant PhysioSpace

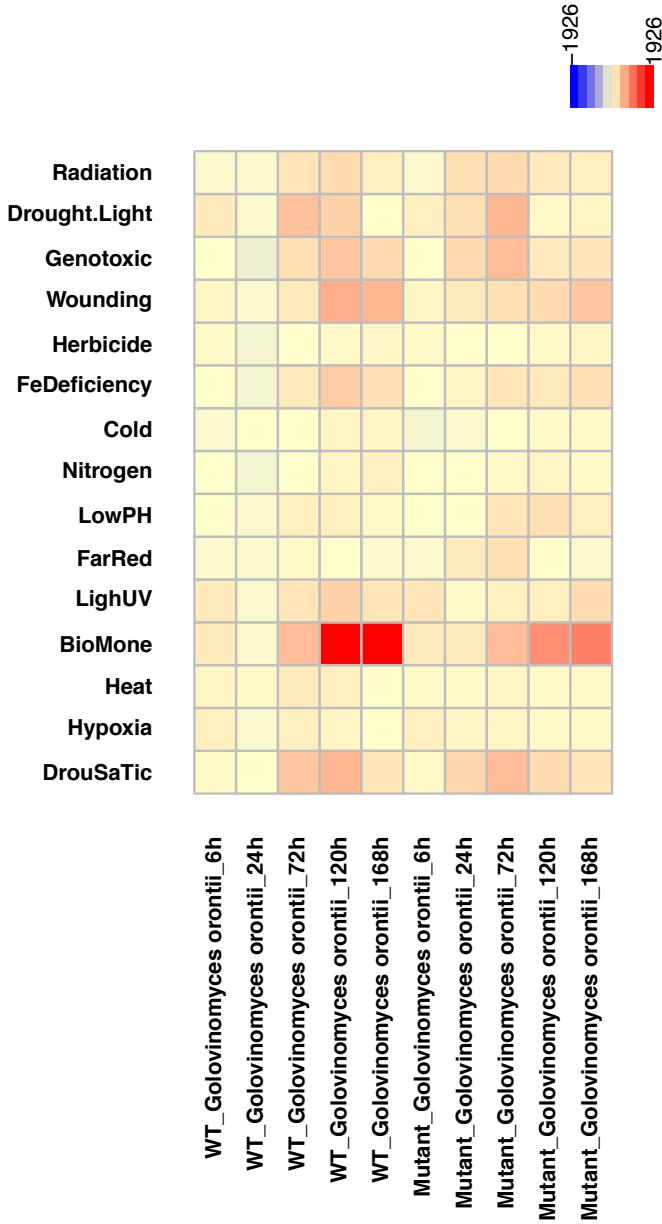
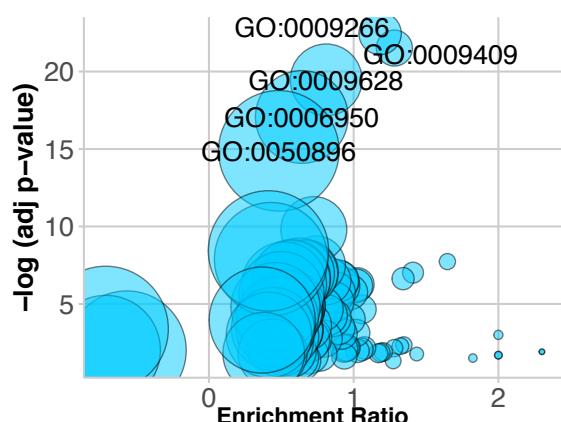


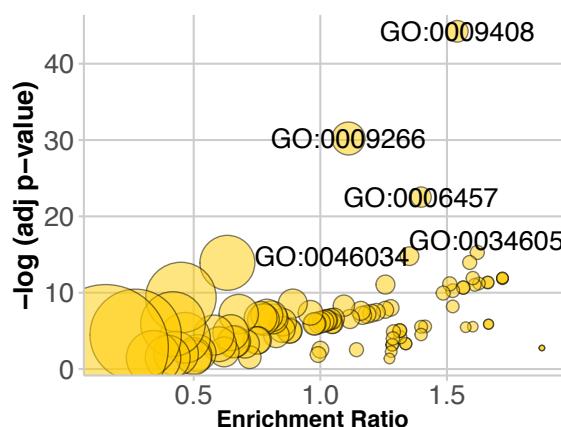
Figure 3: Analysis results of GSE13739. The GEO data set GSE13739 is mapped against the $\overline{S_{mr}}$ reference space using Plant PhysioSpace. The data set provides samples of Wild type (WT) and Mutant (MT) *Arabidopsis thaliana* plants that are infected with *Golovinomyces orontii* (Biotic stress). The mutant plants are expected to be more resilient.

A Cold Stress



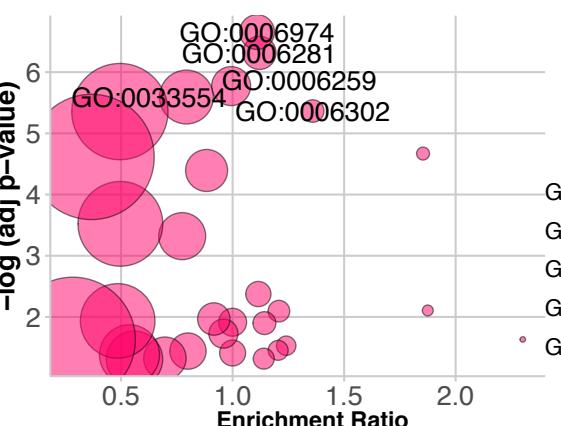
ID	Description
GO:0009409	response to cold
GO:0009266	response to temperature stimulus
GO:0009628	response to abiotic stimulus
GO:0006950	response to stress
GO:0050896	response to stimulus

B Heat Stress



ID	Description
GO:0034605	cellular response to heat
GO:0009408	response to heat
GO:0006457	protein folding
GO:0046034	ATP metabolic process
GO:0009266	response to temperature stimulus

C Radiation Stress



ID	Description
GO:0006302	double-strand break repair
GO:0006281	DNA repair
GO:0006974	cellular response to DNA damage stimulus
GO:0006259	DNA metabolic process
GO:0033554	cellular response to stress

Figure 4: GO Analysis of Mean Stress Space. Results of GO analysis on three stress groups are demonstrated using bubble plots. In the plots, each enriched GO term is represented by a circle, with adjusted p -values as y-axis and enrichment ratio as x-axis. The size of the circle shows the size of the gene list of the corresponding GO term. And enrichment ratio here means the ratio between the actual number of differentially expressed genes and the expected, in each GO group. For each plot, 5 most significant GO terms are labeled on the plot and listed in a table beside each plot. Complete set of bubble plot and set of significant GO terms for all 15 stress groups are provided in supplement files 2 and 3. Plots were generated using the GOpot package in R [42].

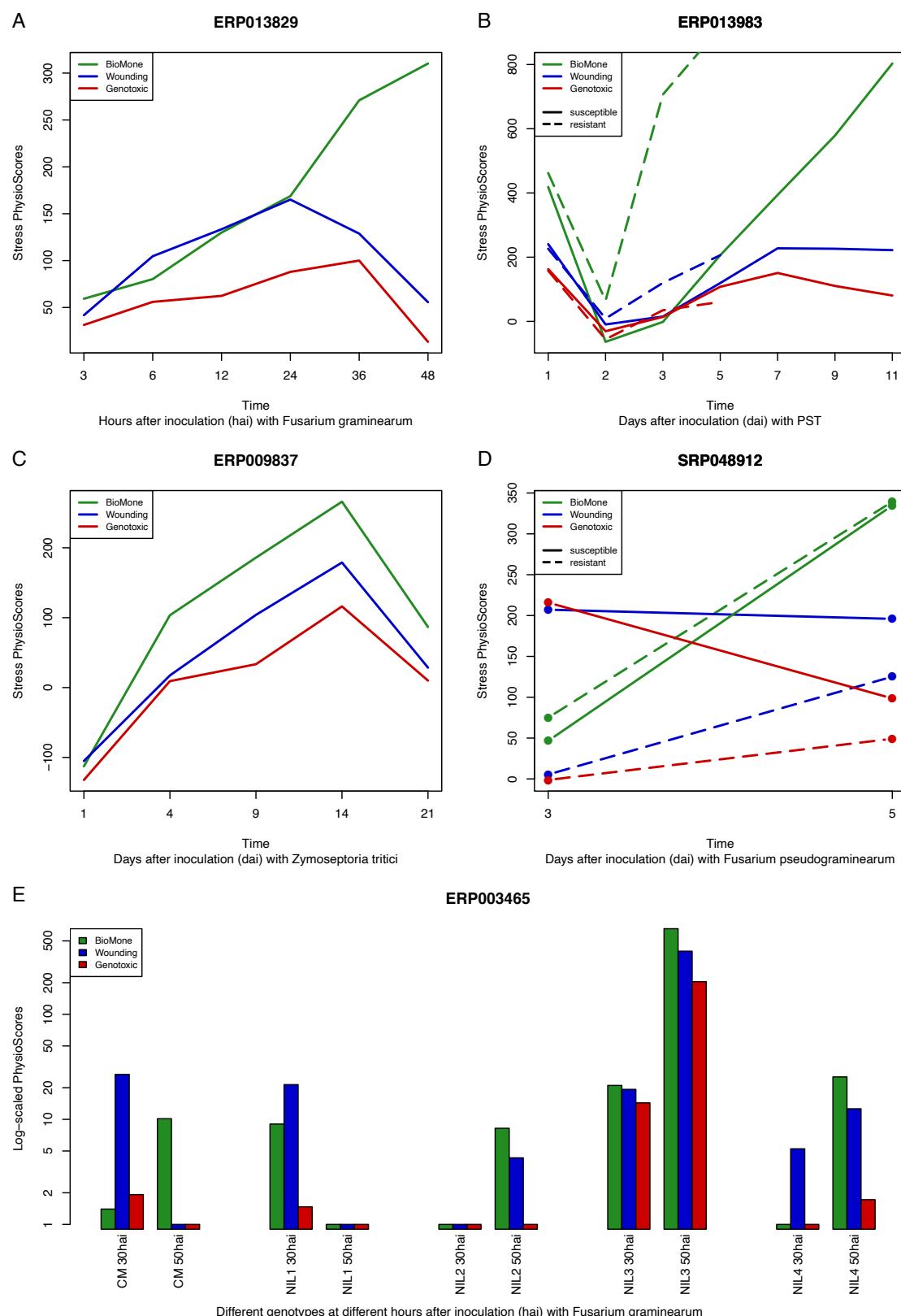


Figure 5: Time Series Analysis of Biotic Stress Response of Wheat RNA-seq data. 5 different biotic-stressed data sets from Wheat Expression Browser are mapped to the *Arabidopsis* space S_{mr} , and the three groups with highest stress values are plotted for each data set. In 4 out of 5 cases, BioMone (Biotic and Hormone) stress group has the highest similarity value, with resistant mutants having higher responses than the susceptible ones (panels A, B, C, and D).

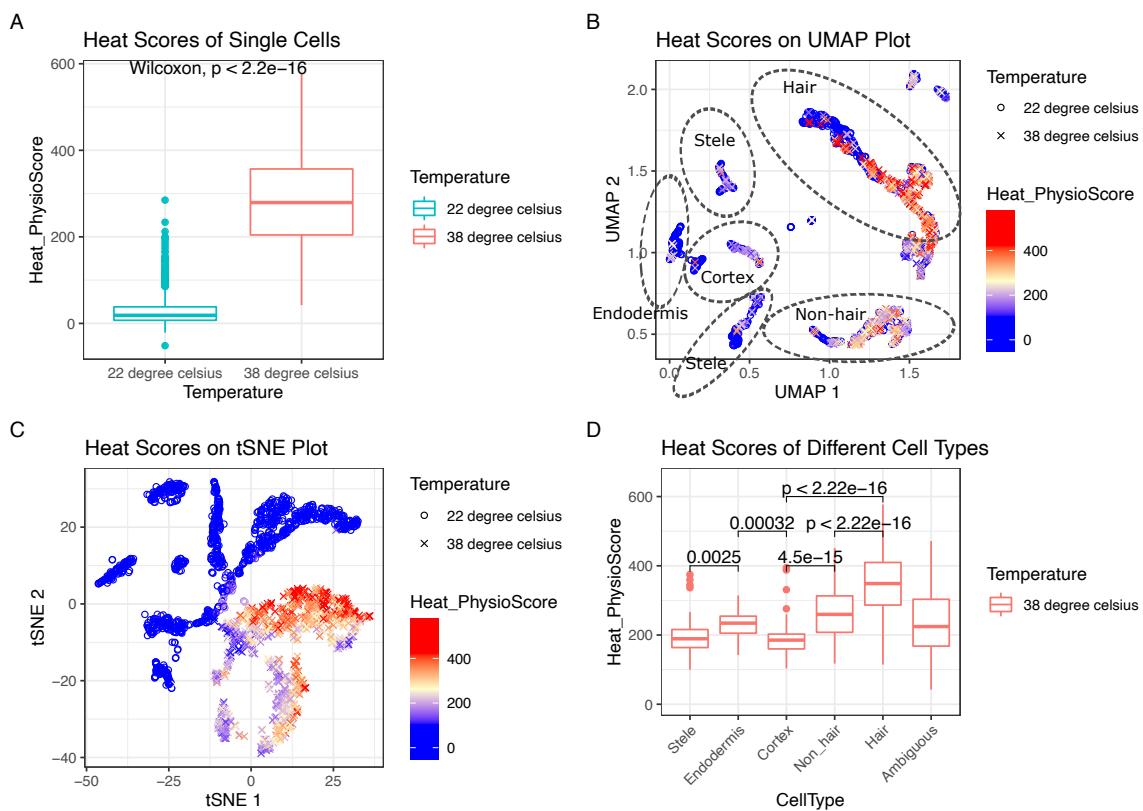


Figure 6: Single Cell Analysis Results of Plant PhysioSpace. Stress scores were calculated for each cell. For demonstrating the outcome, we plotted the heat score of the two big groups of control and stressed, shown in panel A. This box plot proves how Plant PhysioSpace could correctly detect and quantify stress response in single cell data. On panels B and C, we overlaid the heat scores on UMAP and tSNE plots, respectively. On panel D, boxplot of heat scores, on y-axis, were plotted against different cell types, on x-axis. Cell types on the x-axis are ordered based on the morphological anatomy, starting from inner cell types to outermost cell layers (excluding Ambiguous cells, which come at the end).

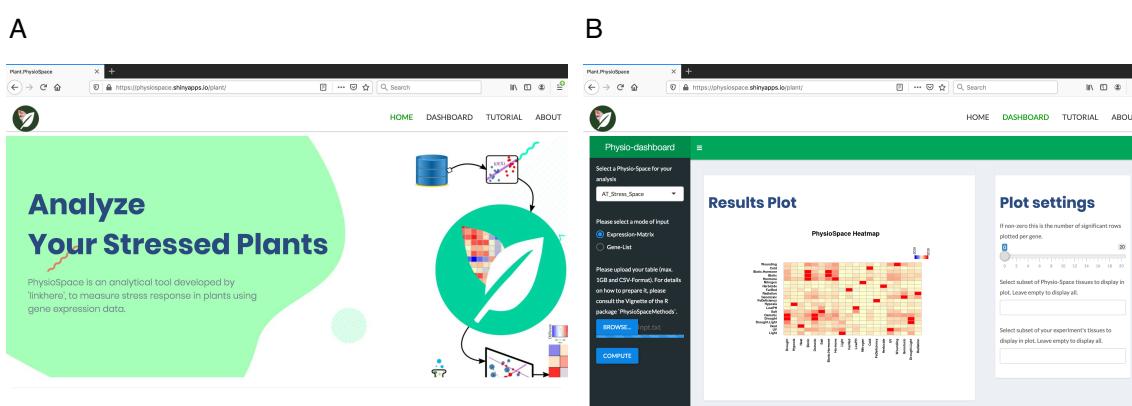


Figure 7: Plant PhysioSpace Web-application

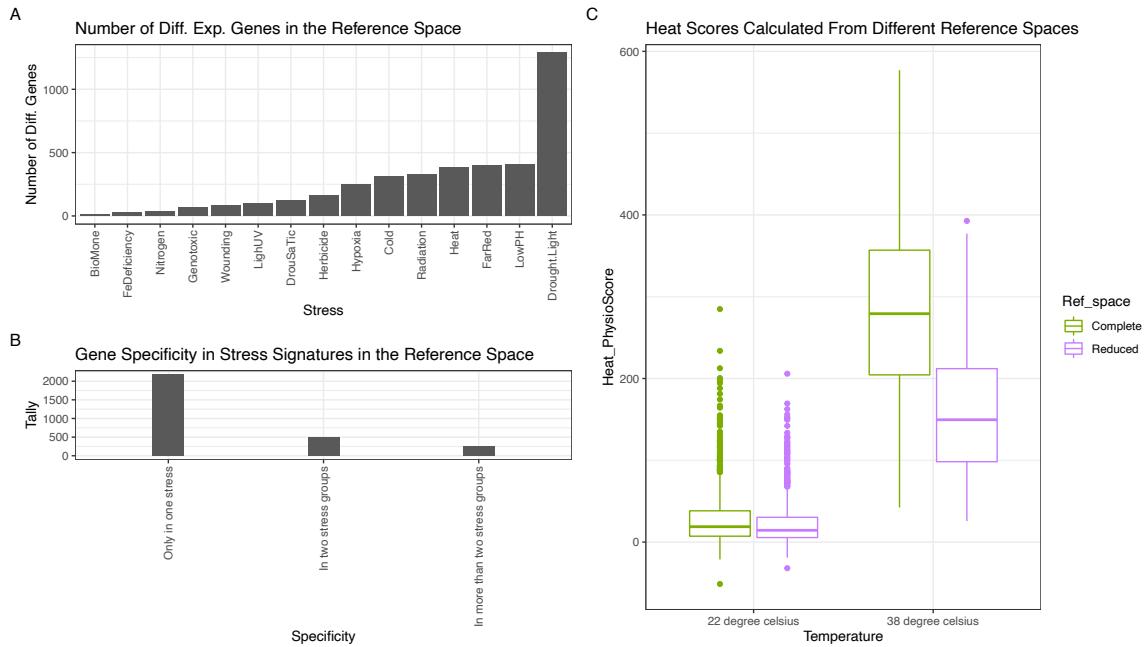


Figure 8: Differentially Expressed Genes in the Reference Space. Behavior of genes with an absolute fold change value of more than 1 in the reference space $\overline{S_{mr}}$, and their effects on the performance of $\overline{S_{mr}}$ as a reference space is examined. In panel A, the number of genes with an absolute fold change value of more than 1 in each stress group is demonstrated. Since there are 22249 genes available in the space $\overline{S_{mr}}$, the ratio of differentially expressed genes to all genes among different stress groups spans from around 0.07% in BioMone (biotic, hormone, or both stresses) to around 5.8% in the Drought.Light (double stress) group. In panel B, we explored the specificity of differentially expressed genes to stress groups. Among all 22249 genes in $\overline{S_{mr}}$, only 2905 (~13%) are differentially expressed in one or more stress groups. From these 2905 genes, 2175 (~10%) are specifically expressed in only one stress group, 488 (~2%) are expressed in two stress groups, and 242 (~1%) in more than two stress groups. Hence, we conclude that the majority of expressed genes in the reference space $\overline{S_{mr}}$ are specific to one stress group. In panel C, we used the heated single cells to study the effect of the 2905 differentially expressed genes on applicability of $\overline{S_{mr}}$ as a reference space. We compared the performance of $\overline{S_{mr}}$ against $\overline{S_{mr}}$ without the 2905 differentially expressed genes, which we called the "reduced" $\overline{S_{mr}}$. As evident from the boxplot, the heat scores are still significantly different between control and heated cells, even when the reduced space is used, although the magnitude of the heat scores is decreased compared to when the complete space is used.