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

Ali Hadizadeh Esfahani[1], Janina Maß[2], Asis Hallab[2], Bernhard M. Schuldt[3], David Nevarez[1], Björn Usadel[2], Mark-Christoph Ott[4], Benjamin Buer[4], Andreas Schuppert[1]*

[1]Joint Research Center for Computational Biomedicine, RWTH Aachen University, Aachen, Germany
[2]IBG-4: Bioinformatics, Forschungszentrum Jülich, 52425 Jülich, Germany
[3]Independent Consultant
[4]Bayer AG, Crop Science Division, Monheim am Rhein, Germany

Abstract

Generalization of transcriptomics results can be achieved by comparison across experiments, which is based on integration of interrelated transcriptomics studies into a compendium. Both characterization of the fate of the organism under study as well as distinguishing between generic and specific responses can be gained in such a broader context. We have built such a compendium for plant stress response, which is based on integrating publicly available data sets for plant stress response to generalize results across studies and extract the most robust and meaningful information possible from them.

There are numerous methods and tools to analyze such data sets, most focusing on gene-wise dimension reduction of data to obtain marker genes and gene sets, e.g. for pathway analysis. Relying only on isolated biological modules might lead to missing of important confounders and relevant context. Therefore, we have chosen a different approach: Our novel tool, which we called Plant PhysioSpace, provides the ability to compute experimental conditions across species and platforms without a priori reducing the reference information to specific gene-sets. It extracts physiologically relevant signatures from a reference data set, a collection of public data sets, by integrating and transforming heterogeneous reference gene expression data into a set of physiology-specific patterns, called PhysioSpace. New experimental data can be mapped to these PhysioSpaces, resulting in similarity scores, providing quantitative similarity of the new experiment to an a priori compendium.

Here we report the implementation of two R packages, one software and one data package, and a shiny web application, which provides plant biologists convenient ways to access the method and a precomputed compendium of more than 900 PhysioSpace basis vectors from 4 different species (*Arabidopsis thaliana*, *Oryza sativa*, *Glycine max*, and *Triticum aestivum*).

The tool reduces the dimensionality of data sample-wise (and not gene-wise), which results in a vector containing all genes. This method is very robust against noise and change of platform while still being sensitive. Plant PhysioSpace can therefore be used as an inter-species or cross-platform similarity measure. We demonstrate that Plant PhysioSpace can successfully translate stress responses between different species and platforms (including single cell technologies).

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

1 Introduction

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

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

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

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

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

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

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

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

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

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

2 Method

2.1 Data Preparation

While setting up a PhysioSpace matrix, our method requires extensive training data for achieving adequate 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 GEO[1] and SRA[2]. 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 **physio**logical 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 Array[3]. 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

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

In LOOCV, by definition, one sample is left out for testing. In our LOOCV scheme though, we left out one GSE (GEO[4] series) data set in each cycle. Due to batch effects, even with proper preprocessing, samples from the same GSE set tend to be similar. With leave-one-GSE-out cross-validation, we make sure that the stress response from different data sets could be successfully matched together.

In each iteration, one GSE data set is chosen as the test set, it is mapped to the rest of the data sets, the training set, and it is counted as a successful match if the analyzed test data and its most similar data set from training set undergone the same stress group. Using the confusionMatrix function from the caret package [21] in R[5], matching accuracy and robustness of the method is evaluated (Fig.2A). With an overall balanced accuracy of 0.43, a Cohen's kappa of 0.385, and an accuracy p-value of 7.35×10^{-42}, PhysioSpace could successfully match the samples going through the same stress group.

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

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

The resulted space, which we call meta-reference space or S_{mr}, and its successive mean space which we denote by $\overline{S_{mr}}$, are the spaces we use as reference throughout the result section.

2.2.2 Physio-mapping

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

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

1. Either

   (a) A new space is extracted from the new data. This means that for each gene in each stress case, a fold change is calculated by modeling the gene behavior under the respective stress given the control. We call this new input space S_i.

   (b) For each stress type γ, genes are sorted from the lowest to highest fold change.

   (c) N percent lowest and highest genes are selected as $L_L(\gamma)$ and $L_H(\gamma)$ for each γ. N is a user-defined parameter. In this paper, it is between 3 to 5 percent.

   or

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

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

$$PS = \begin{bmatrix} ps_{11} & ps_{12} & \dots & ps_{1t} \\ ps_{21} & ps_{22} & \dots & ps_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ ps_{p1} & ps_{p2} & \dots & ps_{pt} \end{bmatrix} \text{ with } ps_{k\gamma} = signed\ log_2(MWW_{p\text{-value}}(s_{L_H(\gamma)k}, s_{L_L(\gamma)k})) \qquad (1)$$

In equation 1, ps_{km} is the physio score between the m^{th} sample and the k^{th} column of reference space $\overline{S_{mr}}$. ps_{km} physio score value shows how similar the m^{th} sample of S_i is to the k^{th} column in $\overline{S_{mr}}$. $MWW_{p\text{-value}}(a, b)$ is a function that calculates the p-value of a Mann–Whitney–Wilcoxon statistical test

[4] https://www.ncbi.nlm.nih.gov/geo/

[5] https://www.r-project.org/

(also known as Mann–Whitney U test or Wilcoxon rank-sum test) between a and b. $s_{L_H(\gamma)k}$ and $s_{L_L(\gamma)k}$ 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* $log_2(x)$ is $\begin{cases} -log_2(x), & \text{if } mean(s_{L_H(\gamma)k}) \geq mean(s_{L_L(\gamma)k}) \\ log_2(x), & \text{otherwise} \end{cases}$

PS is a physio score matrix, containing similarity values of all input samples to all axes (i.e. columns) in the reference space.

For inter-species mapping, for instance in analyzing new data from *Oryza sativa* using a space generated from *A. thaliana*, we resorted to orthologous genes. By using the ideal assumption of orthologs to have identical biological roles in all species, we mapped genes to their orthologs in cases with interspecies translation.

3 Results

3.1 Stress Space Verification by GO Analysis

For substantiating the authenticity of stress information collected using our space generation process, we utilized the gene list analysis section of PANTHER [22, 23]. For each stress group in our generated mean meta-stress space $\overline{S_{mr}}$, we selected the genes with an absolute score value of more than one and a half[6], and tested this gene list by PANTHER overrepresentation Test, against Gene Ontology (GO) biological processes (Fig.3 and supplement files 2 and 3). From 15 different stress groups, 11 were found to have the GO terms corresponding to the same stress enriched, with significant corrected p-values of less than 0.001.

3.2 Inter-Technology Translation

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

Data derived from different platforms are distinctly different. Hence, there are numerous methods to translate measurements from one technology to another [24, 25]. Moreover, with the third generation sequencing right around the corner (PacBio and Nanopore, to name a few [26]), there is a high demand for computational methods capable of transferring useful information between different measurement technologies.

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

3.3 Inter-Species Translation

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

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

[6] For 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.

[7] The 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%).

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

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

3.4 Benchmarking Plant PhysioSpace against Other Methods

We used the results from inter-technology and -species stress response translation to benchmark our method. Plant PhysioSpace is compared to the most common approaches used in bioinformatics for measuring relations between two or more gene expression samples: Gene Set Enrichment Analysis (GSEA) [27], Weighted Connectivity Score (WTCS), which is an advanced version of GSEA used in connectivity map [28], Pearson and Spearman correlations, and Euclidean distance. For each method, fold change values of samples, from different technologies and species, are calculated and used for finding the similarities between samples. Based on our results in inter-species and -platform mapping, Plant PhysioSpace could outperform other methods in all scenarios, except in mapping from wheat to Arabidopsis (Table 1).

3.5 In Depth Investigation of Wheat Stress Response

The poor performance of our method in translating stress response from *Triticum aestivum* to *A. thaliana* may have potentially derived from the microarray used to measure the wheat gene expressions. All microarray samples of *Triticum aestivum* in this study are generated by using Affymetrix Wheat Genome Array[8]. Not only the aged technology could potentially deter the accuracy of transcription measurements, but also as a polyploid, the complex genetics of wheat would make the task of measuring its RNA levels troublesome. Fortunately, advances in NGS gave rise to wheat stress response data sets with higher precision. In this section, we repeated the wheat-to-Arabidopsis translation from inter-species analysis, except with wheat RNA-seq instead of microarray data.

We turned to the Wheat Expression Browser [29, 30] as the source of *Triticum aestivum* RNA-seq data. From this source, we queried all data sets which study stress response, contain more than 30 samples, and include controls. We mapped these data sets into mean meta-reference space $\overline{S_{mr}}$, and plotted Physio scores of three stress groups with the highest values (Fig.4).

In the experiment set ERP013829, wheat response after inoculation with fungal pathogen *Fusarium graminearum* is measured through time [31]. For this experiment, Plant PhysioSpace correctly predicts that wheat is experiencing BioMone (Biotic and Hormone) stress. In addition, the diachronic rise in the response indicates how Physio scores are quantitatively comparable (Fig.4A).

In the data set ERP013983, responses of two different mutants of wheat are studied to wheat yellow rust pathogen *Puccinia striiformis f.sp. tritici* (PST). The authors focused on the pathogen suppression of basal defense in plants [32]. From their results, they deduced the pathogen overcame the defense by rapidly suppressing the genes involved in chitin perception on day 2 after inoculation. In the susceptible interaction, this provides the possibility of invasion and colonization, while in resistant plants, this suppression is quickly reverted. Plant PhysioSpace results expose this mechanism correctly: Both plant types are inducing a BioMone response on day 1, followed by a suppression of the plant defense response on day 2. Eventually,

[8] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL3802

the quick resurgence of intense BioMone response in resistant wheat helps it in withstanding PST, while the reaction of the susceptible trait might be too slow to deflect the pathogen (Fig.4B).

Plants in ERP009837 went through the infection cycle of the hemibiotrophic fungus *Zymoseptoria tritici*. Similar to ERP013829, plants respond to the pathogen with a dominant BioMone response (Fig.4C). Although unlike ERP013829, in which the experiment spanned a few hours, in ERP009837 plants were studied for a longer period. Physio scores suggest that wheat responds to the presence of the pathogen by increasing its BioMone response. This response starts to degrade from day 14, which is in alignment with the original publication of the data set [33], in which the authors state that on day 21, plant tissue is completely defeated.

In SRP048912, the responses of two different traits of resistant and susceptible wheat to Fusarium crown rot are studied [34]. Only two different time points are included in this experiment: 3 and 5 days after inoculation (dai). Plant PhysioSpace results correctly suggest the most dominant stress response present in plants is BioMone, with the resistant wheat having a stronger response than the susceptible (Fig.4D).

Among the Wheat data sets we analyzed in this section, ERP003465 is arguably the most complex, and consequently most interesting as a testing scenario for our method. ERP003465 examined the behavior of 5 different genotypes under the disease pressure of *Fusarium graminearum* [35]. Two well-validated and highly reproducible QTLs (quantitative trait loci), *Fhb1* and *Qfhs.ifa-5A*, are studied from samples taken 30 and 50 hours after inoculation (hai). Five different genotypes were investigated: CM-82036, a progeny of the resistant Sumai-3, and four near-isogenic lines (NILs) bearing either, both, or none of the resistant alleles *Fhb1* and *Qfhs*. Among the four, NIL1 is a mutant with both QTLs, expected to have the highest resistance after CM-82036, NIL2 and NIL3 are mutants harboring *Fhb1* and *Qfhs* QTLs respectively, with both predicted to behave moderately resistant, and NIL4 missing both QTLs, and is likely to be susceptible.

Data analysis in the original paper was mainly based on differential expression analysis. As a first step, the total number of differentially expressed genes for each genotype at each time point was taken as a surrogate for stress response intensity to *Fusarium graminearum*. In the next steps, the weighted gene co-expression network analysis (WGCNA) was used to detect clusters of genes with similar patterns, and Gene Ontology analysis was utilized to infer the role of each cluster in the stress response.

Being able to quantify the intensity of each stress type at each time point, Plant PhysioSpace can provide much more insight into the characteristics and dynamics of the stress responses that are at play in the ERP003465 experiment (Fig.4E). As this data set encompasses a high number of samples distributed between only two time points, we plotted the results as a bar graph. And because the results cover a wide range of values for this experiment, we used log-scaled PhysioScores in the graph, and replaced values smaller than one by one (i.e. zero in log-scale).

Among the concluding remarks in the original paper, some are in concordance with the results from our method. For example, lines lacking *Qfhs.ifa-5A* are regarded as “slow responders” by the original authors, since they lack resistance against initial infection inferred by *Qfhs.ifa-5A*. This lack of early response can be seen in our results (Fig.4E): lines lacking *Qfhs.ifa-5A*, i.e. NIL2 and NIL4, have no BioMone (Biotic and Hormone) stress response at the early time point, while NIL1 and NIL3 show a considerable BioMone stress response at the same time point. Another remark from the original paper suggested that a lack of timely defensive reaction could result in a higher infection in a later time, and consequently stronger response, and vice versa: a quick response may reduce the intensity and infection at a later time. This can be seen in the contrasting response dynamics of NIL1 versus the other lines (Fig.4E). NIL1 possesses both QTLs: *Qfhs.ifa-5A* ensures an early and fast stress response, evident on 30 hai time point. And a strong follow up, courtesy of *Fhb1*, results in a non-existent BioMone response at 50 hai. NIL3 contains *Qfhs.ifa-5A*, so it benefits from a quick response at 30 hai, but due to the absence of *Fhb1*, it cannot be rid of infection at 50 hai, evident by the high BioMone response at that time point. As mentioned, lines NIL2 and NIL4, which lack *Qfhs.ifa-5A*, do not have an early response and have to play catch up with other lines on the later time point.

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

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

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

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

3.6 Plant PhysioSpace Application in Single Cell Analysis

Single cell technologies facilitate investigating transcription profiles in single cell resolution, in order to perceive the genetic basis of each cell type and its function. Although relatively new, more and more plant single cell data sets are becoming available to the community [9, 10, 11, 12, 13]. For now, most sequence data sets are focused on Arabidopsis roots. They try to gain an in-depth understanding of transcription patterns of different cells in different developmental stages of wild-type non-stressed plant roots. To our knowledge, the only publication in which stressed single cells were sequenced is the paper by Jean-Baptiste et al. [12]. In this work, 38°C heat stress was applied to 8-day-old seedlings for 45 minutes. Subsequently, roots of the seedlings were harvested, along with the roots of age- and time-matched control seedlings. The authors could capture and sequence 1,009 cells from the stressed group and 1076 from the control group. For processing the sequencing results, they followed the usual single cell analysis pipeline: PCA, UMAP and clustering, followed by differential gene expression analysis on clusters and enrichment tests on genes related to heat-shock. The results show the ”promise and challenges inherent in comparing single cell data across different conditions and treatments”. In this section, we demonstrate how a dedicated method, such as Plant PhysioSpace, can bring forth more benefits than using the methodological norms.

To analyze the single cell data set, we used the gene-wise mean value of all control cells as the reference, calculated fold changes for each single cell, and fed those fold change values into the Plant PhysioSpace pipeline (Fig.5). Regardless of the cell type, heat-stressed single cells had significantly higher heat stress scores, compared to control single cells (Fig.5A). For studying the heat-induced cell type disparity, we overlaid heat stress scores on UMAP and tSNE plots (Fig.5B&5C). In both tSNE and UMAP plots, coordinate values calculated in the original paper of Jean-Baptiste et al. were used. As a result, cells are bundled in cell type clusters in the UMAP plot, while in the tSNE plot, cells are clearly separated into two big clusters of control and stressed. Although, inside these two big clusters, sub-clusters of different cell types are evident (Fig.S1). On the UMAP plot on the other hand, big clusters represent cell types (Fig.S2), and inside each cell type cluster, groups of control and stressed cells may or may not be distinct, depending on the cell type. For example, in Hair and Non-hair clusters, control and heated cells are separated, while the separation is less pronounced in Stele cells (Fig.S2).

To look into the distinct behavior of different cells under stress, we also plotted cell heat scores grouped by the corresponding cell types (Fig.5D&S3). The results show how Hair and Non-hair cells have higher heat scores, which demonstrates how the outer layers of roots are sharper in their response to heat. This finding is in concordance with one of the conclusions in the original paper, in which based on the behavior of the heat-relevant genes, they concluded the three outermost cell layers of the root went through higher levels of changes caused by the heat stress. The authors hypothesized this may be because of more direct exposure of the outer layers to the heat shock, resulting in a quicker and stronger response.

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

1. convenience: unlike the original paper, there was no need for search and curation of heat stress gene clusters, as they are already available in Plant PhysioSpace, as well as clusters for other common stresses.
2. precision: not only the stress type but also the magnitude of the stress response could be quantified by our method, something which is lacking in traditional gene list enrichment approaches. For example, Plant PhysioSpace results suggest a stronger response in Hair response, compared to Non-hair response (Fig.5D). This inference could not be concluded by the results of traditional methods.
3. optimization: in one run, our tool calculated responses of 20 different stresses for 2085 single cells, in less than 3 minutes on a 2-core laptop CPU. This swift performance is accomplished by precalculating the stress space, in combination with an optimized mapping algorithm, all of which is readily available for the community to use.

3.7 Availability

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

In addition, we made a shiny[9] web application of Plant PhysioSpace (Fig.6). We hosted the web app on shinyapp.io (http://physiospace.shinyapps.io/plant/), to be freely available to use (under the terms of GPL-3 license). We also built a docker image of the ready-to-use tool (https://github.com/usadellab/physiospace_shiny).

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 [36, 37].

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.

5 Data Availability Statement

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

References

[1] Kenneth S Chester. The problem of acquired physiological immunity in plants. *The Quarterly Review of Biology*, 8(3):275–324, 1933.

[2] Joseph Kuć. Translocated signals for plant immunization a. *Annals of the New York Academy of Sciences*, 494(1):221–223, 1987.

[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.

[4] Asis Datta. Genetic engineering for improving quality and productivity of crops. *Agriculture & Food Security*, 2(1):1–3, 2013.

[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.

[9] http://shiny.rstudio.com

[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.

[7] Dennis Gonsalves. Control of papaya ringspot virus in papaya: a case study. *Annual review of phytopathology*, 36(1):415–437, 1998.

[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.

[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.

[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.

[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.

[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.

[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.

[14] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[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.

[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.

[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.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

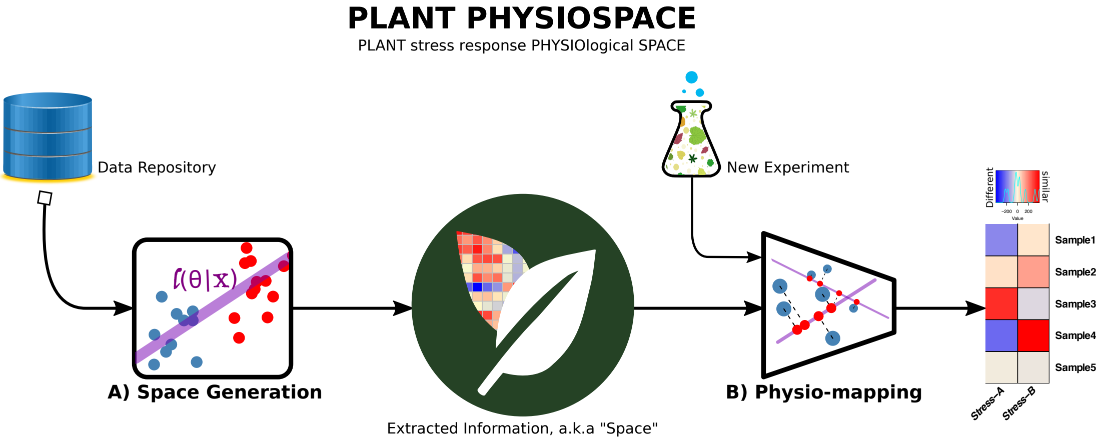


Figure 1: Plant PhysioSpace Overview. The method consists of two main sections: space generation and Physio-mapping. In space generation (A), data from public repositories is processed and its information is extracted. After trimming, the extracted information is stored in matrices called "space", representing 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.

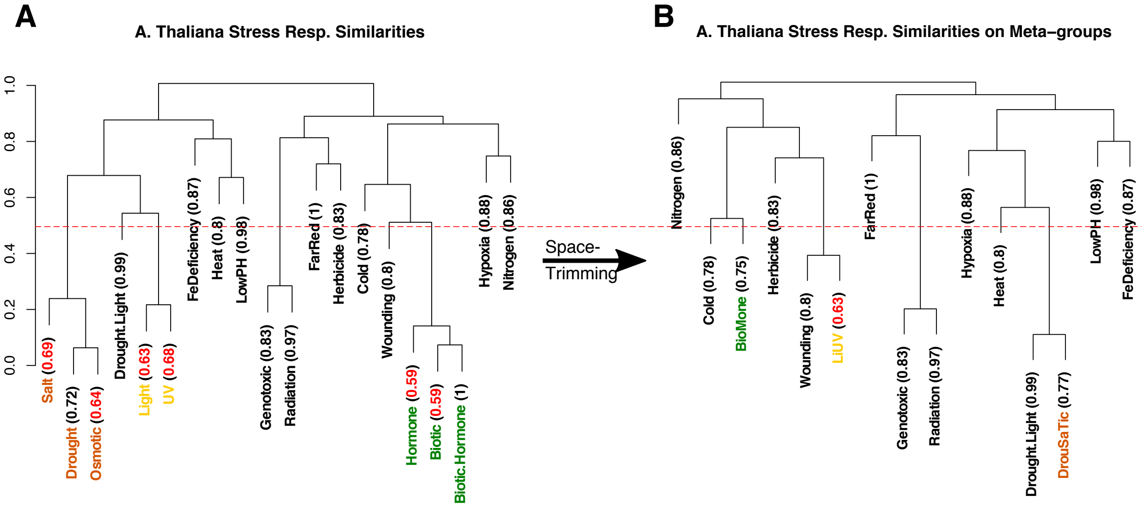
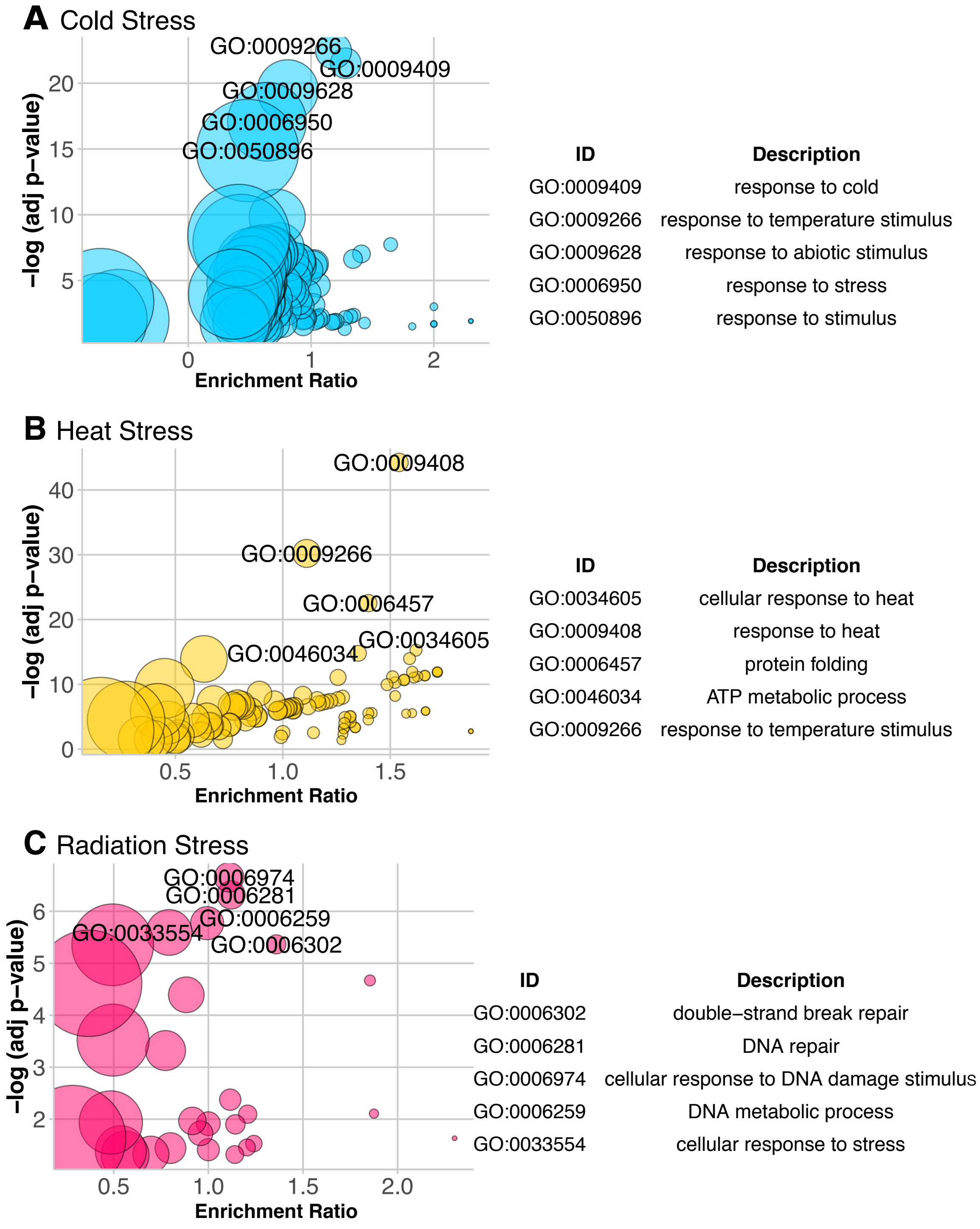


Figure 2: Space Trimming. Stress groups are clustered and for each group, leave-one-out cross-validation accuracy is calculated, written in parenthesis, as shown in panel A. Close groups with low accuracy, written in red, are combined to form new stress groups, called meta-groups, as shown in panel B. 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 LiUV, shown in yellow.



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

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

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 3: 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 GOplot package in R [38].

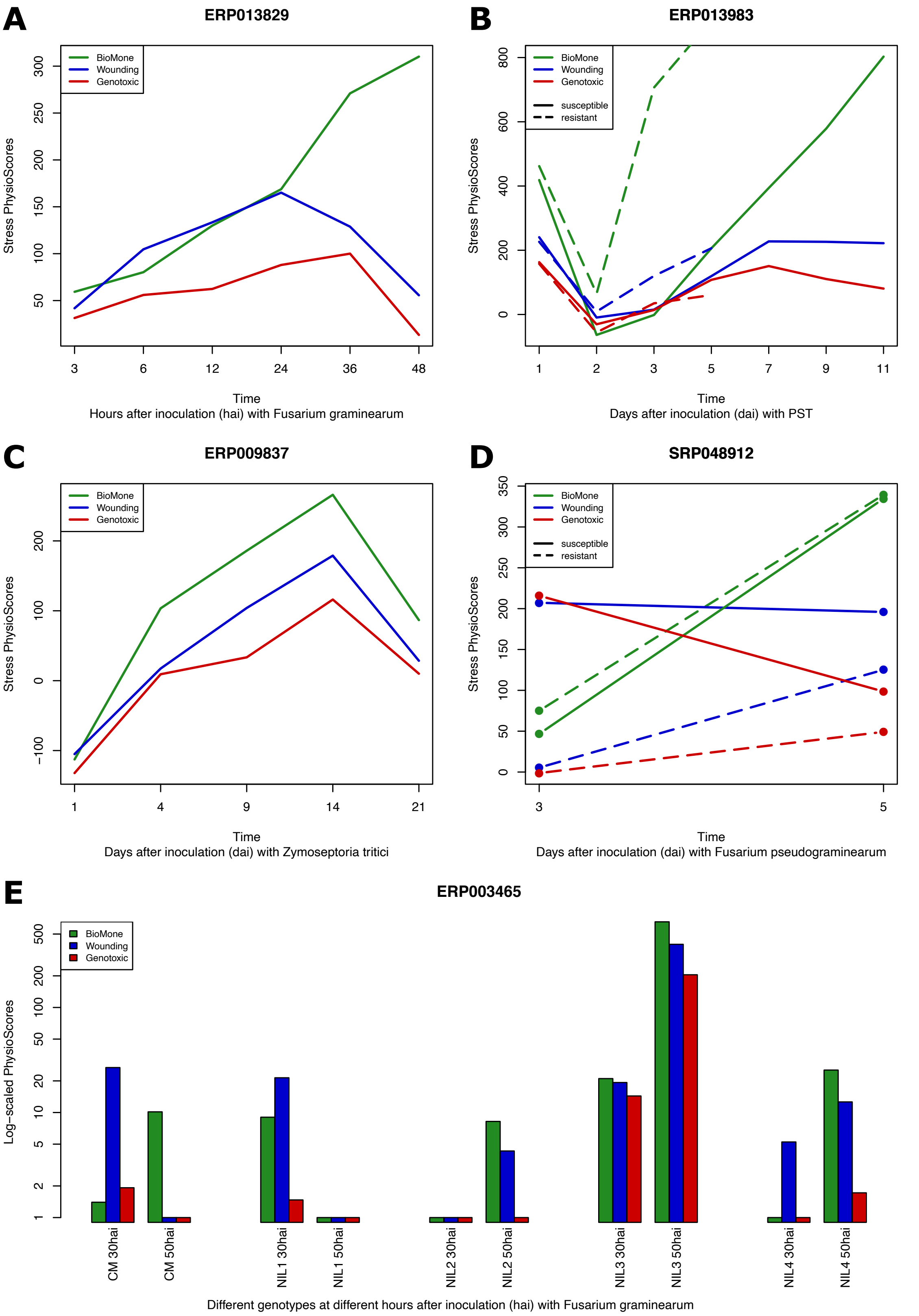


Figure 4: Time Series Analysis of Biotic Stress Response of Wheat RNA-seq data. 5 different biotic-stressed data sets from Wheat Expression Browser are mapped to the Arabidopsis space $\overline{S_{mr}}$, and the three groups with highest stress values are plotted for each data set. In 4 out of 5 cases, 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).

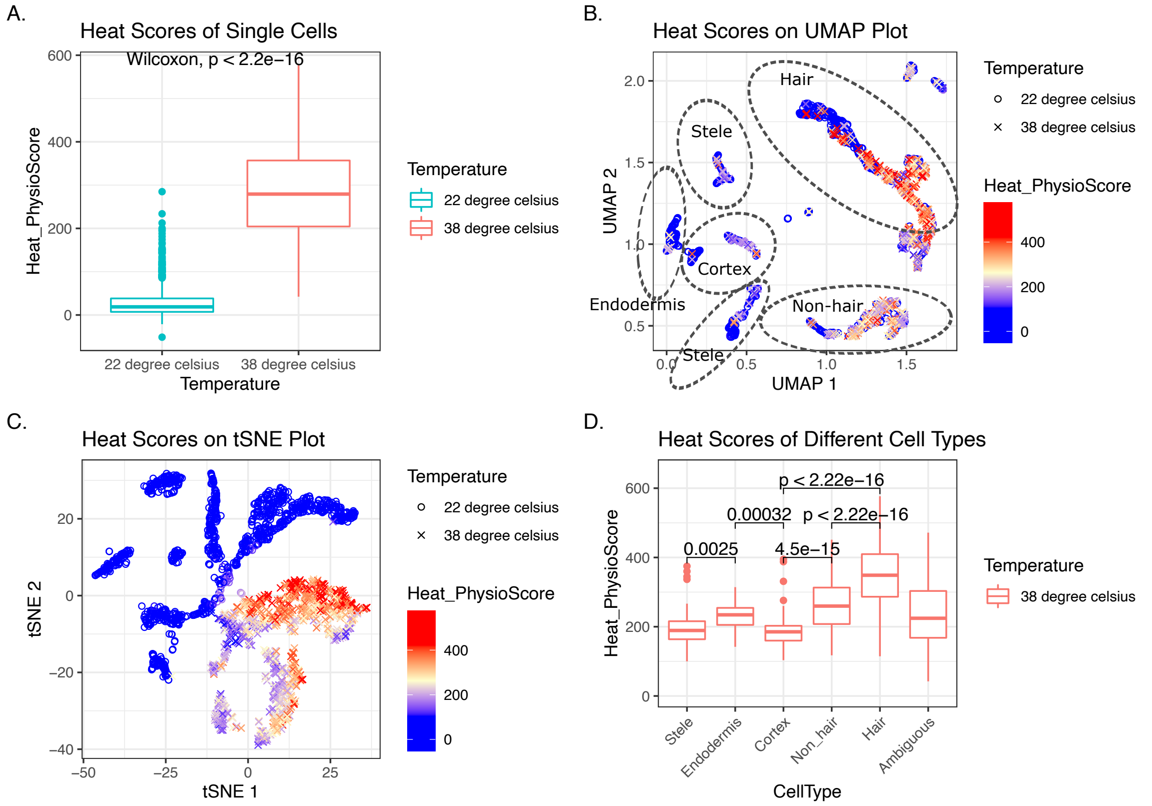
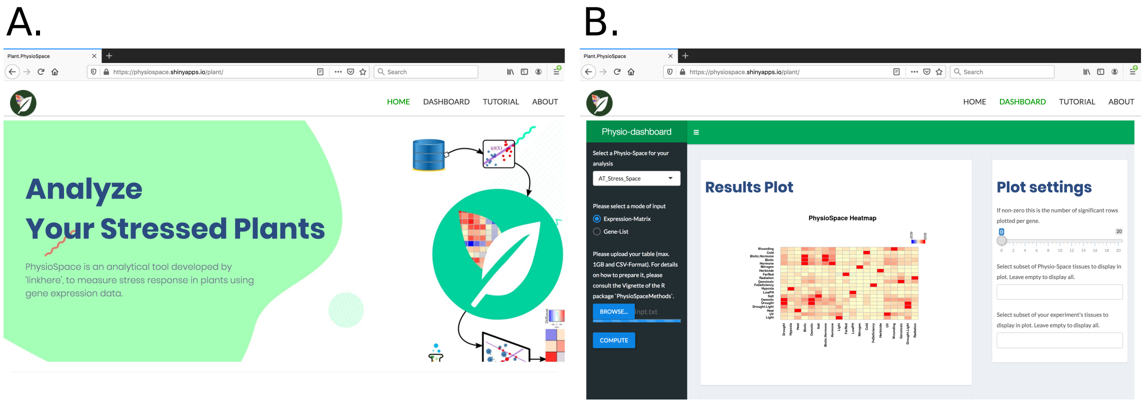


Figure 5: Single Cell Analysis Results of Plant PhysioSpace. Stress scores were calculated for each cell. For demonstrating the outcome, we plotted the heat score of the two big groups of control and stressed, shown in panel A. This box plot proves how Plant PhysioSpace could correctly detect and quantify 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).



Analyze
Your Stressed Plants

PhysioSpace is an analytical tool developed by 'linkhere', to measure stress response in plants using gene expression data.

Select a Physio-Space for your analysis

AT_Stress_Space

Please select a mode of input

● Expression-Matrix
○ Gene-List

Please upload your table (max. 1GB and CSV-Format). For details on how to prepare it, please consult the Vignette of the R package `PhysioSpaceMethods`.

Inpt.txt

Results Plot

Plot settings

If non-zero this is the number of significant rows plotted per gene.

0 — 20

Select subset of Physio-Space tissues to display in plot. Leave empty to display all.

Select subset of your experiment's tissues to display in plot. Leave empty to display all.

Figure 6: Plant PhysioSpace Web-application