Disentangling plant response to biotic and abiotic stress using HIVE, a novel tool to perform unpaired multi-omics integration

Giulia Calia[1, 2,3], Ana Paula Zotta Mota[3], Mame Seynabou-Fall[4], Huyen Trang Nguyen[3,4], Anjana Bhat[3,4], Hannes Schuler[1,5], Carole Gwizdek[6], Ana Cristina Miranda Brasileiro[7], Patricia Messenberg Guimaraes[7] and Silvia Bottini[3*]

[1] Free University of Bolzano, Faculty of Agricultural Environmental and Food Science, Bolzano, Italy

[2] Fondazione Edmund Mach, Research and Innovation Centre, San Michele all'Adige, Italy

[3] INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

[4] Université Côte d'Azur, Center of Modeling, Simulation and Interactions, Nice, France

[5] Free University of Bolzano, Competence Centre for Plant Health, Bolzano, Italy

[6] Université Cote d'Azur, CNRS, IPMC, 06560 Valbonne, France,

[7] Embrapa Genetic Resources and Biotechnology, Brasilia 70770-91, DF, Brazil

*Corresponding Author: silvia.bottini@inrae.fr
To whom correspondence should be addressed:

Silvia Bottini, PhD
Institut Sophia Agrobiotech
UMR 1355 INRAE/7254 CNRS/UniCA
400 route des Chappes – BP 167
06903 Sophia Antipolis Cedex
France
silvia.bottini@inrae.fr

Abstract

Plants live in a constantly changing environment that is often unfavorable or even hostile. Indeed, they developed high phenotypic plasticity that includes rapid responses to aggressive biotic and abiotic factors and adaptations to changing environments. Multiple stresses can occur at the same time, requiring the plants to activate appropriate signaling pathways to respond to both or by prioritising the response to one stress factor. While several studies have been conducted to individual stress factors, only very few studies focus on the simultaneous plant response to multiple stressors. Currently used methods to integrate unpaired experiments consist of performing meta-analysis or finding differentially expressed genes for each condition separately and then selecting the common ones. Although these approaches allowed to find valuable results, they cannot identify non-specific conserved mechanisms that may hold promise for a broader understanding of plant defence response mechanisms.

For this purpose, we developed HIVE (Horizontal Integration analysis using Variational AutoEncoders) to analyse horizontally integrated multi-omics datasets composed of unpaired and/or longitudinal experiments. Briefly, we coupled a variational autoencoder, that captures non-linear relationships and encoded them in the latent space, with a random forest regression and the SHAP explainer to select relevant genes for the studied phenotype.

We illustrate the functionality of HIVE to study the transcriptional changes of two peanut wild relatives' plants during root-knot nematode *Meloidogyne arenaria* infection and/or drought stress from seven unpaired experiments. HIVE performed better than the meta-analysis and the state-of-the-art tool and identified novel promising candidates responsible for triggering effective defense responses to biotic and abiotic stress.

Keywords

Multi-omics integration; unpaired experiments; time-course multi-omics; horizontal integration; variational autoencoders; SHAP explainer; plant responses to stress; root-knot nematode; *Meloidogyne arenaria*; drought stress

Introduction

From omics to multi-omics, we have been assisting to a wide growing of data production with a multitude of experimental designs. Multi-omics can provide a global understanding of the flow of information, by allowing to infer the complete network of interactions from the original biological process to the functional consequences (Großkinsky et al., 2018). In the last decade, multi-omics studies have been employed in several biological fields, including plant immunity (Bolger et al., 2014; Yang et al., 2021). Plants exhibit remarkable cellular plasticity during their growth and development, for instance, they are able to repair tissues after wounding or to regenerate entirely new organisms, as well as to reprogram cells to aid symbiotic interactions with other organisms or to counteract the attack of bio-aggressors (Sugimoto et al., 2018). To understand these mechanisms, omics and multi-omics studies have emerged as a broadly used methodology to characterize how plants respond to the biological environment (Mahmood et al., 2022).

Hence, there is an exponential trend in terms of volumes and complexity of multi-omics data that are currently under-exploited due to the absence of a methodology to successfully integrate data from different experiments. Multi-omics experimental designs have been further extended with a longitudinal dimension to study dynamic relationships between molecules. Despite the impressive effort of the community around this scientific challenge, there is currently a lack of methods to integrate unpaired and longitudinal multi-omics datasets.

Multiple challenges need to be addressed to be able to analyse these data (Mohammadi-Shemirani et al., 2023; Subramanian et al., 2020). Among them, the notorious batch effects caused by external factors may hinder a joint analysis (Goh et al., 2017; Leek et al., 2010). Batch effects are common technical variations in omics and multi-omics data and lead to equivocal results if uncorrected or over-corrected (Han and Li, 2022; Lazar et al., 2013; Molania et al., 2023; Nygaard et al., 2016). Many techniques have been developed to address this challenge; however, they have been developed mainly for large-scale settings (Adamer et al., 2022; Chazarra-Gil et al., 2021; Johnson et al., 2007; Lazar et al., 2013; Rong et al., 2020; Tran et al., 2020). In a recent study, Yu et al. (2023) showed that the best-performing batch correction method is the ratio-based technique. The drawback of this method is the need for reference samples not always available. Several studies applied autoencoder (AE) or its variants to remove the batch effects in single-cell omics and multi-omics data achieving good performances (Lakkis et al., 2021; Li et al., 2020; Wang et al., 2019). The advantage of AEs is their powerfulness to capture non-linear patterns from high-throughput data by transforming them in a lower dimension called latent space. This new compressed representation of the original data has the limitation, prevalent among deep-learning models, of the lack of explainability because the contribution of the originally measured molecules to the latent features is not straightforward. Although some tools have been proposed, such as DeepLIFT (Shrikumar et al., 2019) and DeepSHAP (Lundberg and Lee, 2017) they are usually very computationally intense. Furthermore, they have mainly been used coupled to a supervised approach in which the latent space is used as input for a classifier model to classify samples (i.e. cancer classification).

To circumvent the needs for batch effect correction, the most commonly used approach is a meta-analysis of each experiment separately and the subsequent selection of the most common differentially expressed genes. Among such methodologies, Miodin (Ulfenborg, 2019) and metaRNAseq (Bodein et al., 2022; Rau et al., 2014) are two distinct packages in R language that implement statistical solutions to perform a multi-differential expression analysis and associate a p-value that considers the multiple testings. The major pitfall of this analysis is the small number of common signatures, which decreases with increasing number of datasets. Furthermore, they cannot identify non-specific conserved patterns. An alternative is to use integrative approaches that simultaneously analyse all the experiments. However, very few approaches have been developed for the joint analysis of unpaired and longitudinal experiments. Recently few methods have been proposed specifically to integrate unpaired time course experiments. timeOmics is based on a linear mixed model spline framework to select molecules related to the experimental phenotype (Bodein et al., 2022), while MEFISTO is a flexible and versatile toolbox for modelling high-dimensional data when spatial or temporal dependencies between known samples based on multi-factorial analysis (Velten et al., 2022). However, these methods cannot integrate longitudinal and single time point experiments. In this regard MINT is an approach to analyse unpaired experiments based on partial-least square discriminant analysis (Lê Cao et al., 2011) that corrects for batch effects, classifies samples and identifies key discriminant genes (Rohart et al., 2017a). A possible limitation of this tool is the use of the supervised classification of samples when the experimental design includes the presence of many classes and few replicates that could affect the performance of the tool. To address these limitations, we developed HIVE (Horizontal Integration analysis using Variational AutoEncoders) to analyse at the same time unpaired and/or longitudinal experiments. We show the capabilities of HIVE on a dataset composed of seven distinct transcriptomic experiments of two peanut wild relatives' plants during root-knot nematode *Meloidogyne arenaria* infection and/or drought stress. Altogether HIVE allowed to identify important signatures to improve the understanding of plant-stress response.

Materials and Methods

Datasets collection, preprocessing and integration

We collected RNA-sequencing data from seven different experiments of roots from *Arachis duranensis* or *Arachis stenosperma* before and upon biotic or abiotic stress or both conditions (Table 1). Bioinformatic analyses to extract the gene expression quantification of each experiment were already performed and the gene count matrices were available (for the procedure refer to the original publications listed in Table 1). The integration of all datasets was performed via a custom python script concatenating samples and replicates of each experiment for a total of 36,734 genes and 35 columns. Then, we calculated the variance of each gene along the 35 samples, and we filtered out genes having a value belonging to the first quartile of variance distribution. The final integrated dataset contains 27,557 genes. Finally, a min-max transformation was applied using MinMaxScaler function of Scikit-learn python library (Pedregosa et al., 2011).

HIVE framework overview

HIVE workflow consists of three main blocks described in the next paragraphs and summarized in Figure 1. Briefly, the first block concerns the batch effect correction. Thus, a Variational AutoEncoder (VAE) is used to encode the integrated dataset in a compressed representation called latent space. The two following blocks are independent of each other, but they are consequent to block 1. One block regards the exploration of the latent features. Therefore, the point biserial correlation is applied to calculate the association between the latent features and phenotypic characteristics. The last block is composed of three steps to select relevant genes for the studied phenotype. First, a Random Forest Regression (RFR) is used where each latent feature is considered as the variable to explain, and the original gene expression data as the dependent variable. The problem of inferring the relationship between a set of N latent features and G genes is decomposed into a set of N regression problems. Afterward, the contribution of each gene to each regression was calculated with the SHAP algorithm. Finally, we included a custom method to set a threshold on the SHAP scores to select only the most contributing genes. The third block returns a list of ranked genes that can be used for further analysis.

Block 1: batch effect correction

The VAE architecture is constituted of the following elements: (i) three layers of encoder, respectively composed of 5000, 1000, and 500 nodes; (ii) a latent space of 80 nodes; (iii) three layers of decoder mirroring the encoder dimensions and therefore constituted by 500, 1000 and 5000 nodes respectively, with batch size of 32 data points followed by Exponential Linear Unit (ELU) activation function (Supplementary Figure 1). We have evaluated the VAE reconstruction performances by training loss function, KL-divergence, Mean Squared and Absolute Error and perplexity score. The configuration of the VAE can be modified by the user if metrics show bad performances.

The entire VAE is written in python language and uses Tensorflow and Keras libraries (Abadi et al., 2016; Chollet and others, 2018). The output of this first step is the latent space as a matrix having 80 latent features as columns and samples as rows.

Block 2: Association between latent features and phenotypic characteristics

The integrated dataset used as case study for HIVE application has two main phenotypic characteristics, namely the plant species or the type of stress (i.e. biotic or abiotic). To calculate whether the latent space was able to capture these characteristics, we used the point biserial correlation. Point biserial correlation is a correlation coefficient measure, that can be applied when one of the variables is dichotomous (the phenotypic characteristics) and the other is continuous (the latent features) (Kornbrot, 2005). Assuming to have two sets of elements to correlate, i.e. X = {x_1, …, x_n} (dichotomous variable) and Y = {y_1, …, y_n} (continuous variable) with n the total number of observations, n_0 the number of elements in X which are 0 and n_1 the number of elements in X which are 1, the point biserial correlation can be calculated as follows:

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_y} \sqrt{\frac{n_0 n_1}{n(n-1)}}$$

Where $\bar{Y}_0$ and $\bar{Y}_1$ are the means of the corresponding Y elements in X, coded 0 and 1 respectively, and s_y in the standard deviation of Y (Kornbrot, 2005).

We implemented this calculation in a python script and using both Scipy and Scikit-learn libraries (Pedregosa et al., 2011; Virtanen et al., 2020) included in the HIVE pipeline.

Block 3: Gene selection

Random Forest Regression

Before performing the Random Forest Regression (RFR), genes having a raw expression value for each sample lower than 30 to avoid rows with only 0 and latent features scaled sing MinMaxScaler function of Scikit-learn python library (Pedregosa et al., 2011). The RFR (Breiman, 2001) is implemented using RandomForestRegression module from Scikit-learn library, using default parameter and 100 as the number of estimators.

SHAP scores

To extract the most contributing genes to each regression HIVE implements the SHAP explainer (Lundberg and Lee, 2017). SHAP values (SHapley Additive exPlanations) is a method based on cooperative game theory and used to increase transparency and interpretability of machine learning models (Lundberg and Lee, 2017).

The obtained regression models are used as input for the SHAP TreeExplainer from SHAP python library (Lundberg et al., 2020). We implemented this method in HIVE because it showed better performances when compared with the generally used Gini score (see paragraph Selection of the importance score for gene selection).

Threshold calculation

We introduce a custom method to calculate a threshold on the SHAP scores to select only the most contributing genes to each regression. For each latent feature, we calculated the distribution of SHAP

values by grouping the values in 35 ranges, each range is called bin hereafter (supplementary figure 2A). The number of bins is a modifiable parameter in HIVE allowing the user to adjust this value. Since most of the genes have a SHAP score around 0, there is one bin containing a high percentage of the genes (~99% on average on the 80 features for our dataset). We then select the two bins flanking the bin with the highest percentage of genes and we calculate the average SHAP value in each. These values are used as minimal thresholds to retrieve only the genes with SHAP scores in the tails of the distribution, therefore with the most extreme SHAP values, indicating a major contribution to each latent feature. The number of genes selected for each feature and the number of genes selected in more than one latent feature are reported in Supplementary Figure 2B and C.

After repeating the procedure for each latent feature, to provide a unique ranking, we merged in a unique list using the following strategy (Supplementary Figure 2D). The extracted genes for each latent feature are ranked using the absolute value of the SHAP score. To merge the ranking, we proceed by position. For instance, all genes ranked first for each latent feature, occupy the first 80 positions in the final ranking and the order is in agreement with the SHAP value. Then, we extract all genes ranked second for each latent feature, they occupy the following 80 positions (from 81 to 161) in the final list and the order depends on the SHAP score as before. If a gene is selected for more than one feature, then only the highest rank will be maintained in the final list.

Functional analysis

The web-based Gene Ontology toolkit for Agricultural Community, AGRIGOv2 (Tian et al., 2017) was used to retrieve all the GO-terms associated with each gene in the *A. duranensis* genome. GO term enrichment was performed on the same website using the Hypergeometric statistical test and GO terms from Plant GO slim developed by The Arabidopsis Information Resource (TAIR) (Rhee et al., 2003).

To retrieve information such as hormone-related genes and kinases, we performed protein annotation with MERCATOR4 (Schwacke et al., 2019) web-based tool. First, CDS sequences were first downloaded from PeanutBase database (Dash et al., 2016). The obtained sequences accurately parsed to retrieve only the HIVE selected gene sequences are then used as input of MERCATOR4. Here we specified the type of sequence as DNA and we flagged the option to include Swissprot annotations.

The list of NBS-LRR genes was retrieved from two previous publications (Mota et al., 2021, 2018), including the experimentally validated. The list of transcription factors was retrieved from PlantTFDB.

All the output files have been downloaded and then parsed with *ad hoc* python scripts.

Selection of the importance score for gene selection

To select the best method for gene contribution in each regression, we compared the SHAP scores with the Gini importance scores, the most commonly used feature importance measure directly implemented in the RandomForest method of Scikit-learn. This measure uses the Gini impurity index (Ishwaran, 2015) that essentially calculates the goodness of each split in the decision trees in terms of class entropy. Therefore, for each Random Forest Regression model we used the “feature_importance”

function and, for each feature, we selected the same number of genes selected by the SHAP application in HIVE pipeline and then ranked them globally using the same strategy used for SHAP scores (see the paragraph threshold calculation). After gene selection, we divided each gene list into 30 bins and for each bin calculated the agreement of the two methods.

Benchmarking

We used the Multivariate INTegrative method (MINT) framework within the MixOmics R package (Rohart et al., 2017b). MINT is a supervised method performing horizontal integration of unpaired data and feature selection (Rohart et al., 2017a). For the comparison we used MINT sparse Partial Least Squares - Discriminant Analysis (sPLS-DA) model on the same input dataset used for HIVE (Lê Cao et al., 2011).

We followed the steps of the pipeline as indicated in the guidelines, detailed hereafter. We first declared the studies and the label lists. After the first model build-up, we used the “perf()” function to estimate the performance of the model and to determine the best number of components to effectively use, via the Leave One Group Out Cross Validation. Then, we tuned the model with the “tune()” function and with the same criterion as before we calculated the best number of variables to be included in the final model using the Balance Error Rate as error rate measure and the centroids distance. With all those parameters tuned we build-up the final sPLS-DA model.

Hierarchical clustering of gene expression profiles

For each gene, the median value among replicates for each condition was calculated. To visualize the expression profiles as heatmap and perform hierarchical clustering we used the “pheatmap” R package (Kolde, 2019) using Pearson correlation as distance measure for genes clustering.

Gene regulatory network (GRN) inference and analysis

To infer the GRN we used GENIE3 R package (Huynh-Thu et al., 2010), providing as “regulators” the list of transcription factors and as “target” the remaining genes. We retrieved the transcription factors from *Arachis duranensis* in Plant Transcription Factor Data Base (PlantTFDBv5.0) integrated in Plant Transcriptional Regulatory Map (PlantRegMap) (Jin et al., 2017; Tian et al., 2020). We used the Networkx python library to build-up and visualize the resulting GRN (Hagberg et al., 2008).

Specifically, we used the Networkx built-in function, “from_pandas_edgelist()”, to declare the regulatory and target genes from a pandas dataframe obtained by parsing GENIE3 output and the “DiGraph()” module to reconstruct a directed network. The visualization via Networkx is obtained by the built-in function “draw()” and specifying the node position passing to the function “spring_layout()” an indirect version of the previously obtained network.

We used the Leiden community detection method, implemented into the python library leidenalg that is built on the igraph library, to find communities in our network. We used the library on the GRN with the “partition_type” parameter set as leidenalg ModularityVertexPartition. We then explored each

community with *ad hoc* python scripts using methods for functional analysis illustrated in the previous paragraphs.

Data availability

Data, scripts and HIVE source code are available in the github repository: https://github.com/Plant-Net/HIVE.git

Results and Discussion

Case study description

To illustrate the functionalities of HIVE we have selected 16 RNA-seq libraries for a total of 35 samples from Zotta-Mota et al. (2021). The integrated dataset we created consists of seven unpaired transcriptomics datasets of two wild relatives' peanuts plants *Arachis stenosperma*, highly resistant to the root-knot nematode (RKN) *Meloidogyne arenaria* and moderately tolerant to drought stress and *Arachis duranensis*, that shows the opposite behavior. For each plant species, *M. arenaria* infection was studied in a time-course experiment composed of three time points, while drought and desiccation were available in a single time point. The combined stress (drought and *M. arenaria* infection at the same time) was available only for *A. stenosperma* and for a single time point. The complexity of the experimental design combining two plant species, time-course, unpaired experiments, biotic, abiotic and cross-stress makes this dataset the ideal showcase to illustrate the functionalities of HIVE.

The variational autoencoder reduces the batch effects allowing unpaired experiments integrated analysis.

In the HIVE framework, the first step consists of a variational autoencoder (VAE) based method to alleviate the batch effects and to allow the integrated analysis of unpaired experiments. VAE is a deep neural network model that learns a nonlinear mapping function from the original data space to a low-dimensional feature space called latent space (Kingma and Welling, 2022; Pratella et al., 2021). To visualize the presence of batch effects in our dataset, we used the heatmap representation of the correlation matrix among samples (figure 2). First, we calculated the correlation matrix of samples using the original data, we can observe a strong batch effect affecting our dataset as expected (figure 2A). The application of the VAE, allowed to reduce the batch effect while preserving the relationship among samples (figure 2B). We can distinguish two main groups separating samples by biotic/abiotic stress. In the biotic group, we can observe two sub-clusters that separate the samples belonging to the two plant species, as expected because of their different resistance to the two stresses. Interestingly, the samples collected upon cross-stress (drought and *M. arenaria* simultaneously) belong to the abiotic group, suggesting that the plant prioritizes the response to abiotic stress as observed in (Mota et al., 2021).

Overall, the VAE corrects the batch effects thus enabling the integrated analysis of unpaired experiments with complex experimental design.

Latent features are associated with phenotypic characteristics.

We explored the latent features to study whether they capture patterns associated with the different phenotypic characteristics in the dataset. We studied the two main characteristics, namely the type of stress (biotic or abiotic) and the type of plant species (*A. Duranensis* or *A. Stenosperma*). By using the point-biserial correlation, we calculated the most significant latent features correlating either with the type of stress or the species (see methods). Among the 80 latent features, we found 19 features correlated with plant species and 16 with biotic/abiotic stress (figure 3A). Interestingly, three of them were found significantly correlated to both characteristics. Then we plotted the two most significative latent features associated with each phenotypic characteristic and we observed samples grouping accordingly (figure 3B and 3C). As highlighted in the previous paragraph, samples from the experiment with the cross-stress (RKN infection and drought at the same time) cluster together with the other samples of abiotic stress.

Overall, these results show not only that the latent features reduce the batch effect but also that they capture the phenotypic characteristics of the dataset.

The random forest regression coupled with SHAP explainer select genes with insightful biological meaning.

The VAE transformed the original dataset composed in a latent space composed of 80 latent features representing a compressed version of the original transcription profiles for the 35 samples. A major drawback of this novel representation is that the 80 latent features are not trivially connected to the original gene measurements thus hindering a straightforward biological interpretation. Therefore, we used a random forest regression (RFR) model, where each latent feature is considered as the variable to explain, and the original gene expression data as the dependent variable. The problem of inferring the relationship between a set of N latent features and G genes is decomposed into a set of N regression problems. The RFR were evaluated using common metrics such as: mean absolute error, median absolute error, mean squared error and R2 (supplementary table 1).

To find the genes that contribute the most to the latent features, we compared the genes selected by applying the SHAP explainer and the RFR importance score to each regression and retrieve the genes with the highest scores. To compare the two scores, since SHAP scores are both positive and negative, we first needed to find a criterion to select a threshold considering the asymmetry of the distribution. Thus, we divided the SHAP score distribution into an arbitrary number of intervals, and used as thresholds the average value of the SHAP scores in the two bins flanking the bin containing the highest number of genes (genes with SHAP value around 0), refer to methods section for detailed explication. Therefore, we selected 5841 genes based on the SHAP scores. By selecting the first 100 most contributing genes for each latent feature based on the RFR importance score, we achieved a similar number of genes for a total of 5850. The overall agreement on the two scores consists of around 70% of genes (Supplementary Figure 3A). Then we divided the gene lists into 50 bins, and we calculated the gene ranking agreement for each bin. We observe that the concordance of 70% is mainly maintained at the top of the distribution and oscillates around 60% in the following bins (Supplementary figure 3B). By comparing the enrichment in GO terms in the two lists (Supplementary Figure 3C), we

observed very similar behavior although the genes selected by SHAP score present terms related to primary metabolism and cellular protein metabolic process that are absent for the RFR importance score selection. To decide which list to use for further analysis, we performed the GO terms enrichment analysis only selecting the most important gene for each feature with either SHAP or RFR importance score for a total of 80 genes for each score (supplementary Table 2A and 2B). The overlap between the two lists drops to 23% we thus expect different GO terms (Supplementary Figure 3D). Indeed, the genes selected by SHAP are involved in terms related to stress response and metabolic processes that are not present in the genes selected by RFR importance score. Therefore, the SHAP score performs better in selecting genes with biological meaning.

Benchmarking tools for gene selection from integrated datasets.

We compared the results of HIVE with the MINT tool from the MixOmics R package (Rohart et al., 2017b, 2017a). MINT is a supervised approach to performing integrative analysis of unpaired experiments. The application of MINT on our case study dataset leads to the identification of 81 genes. We compared with the set of 5841 genes selected by HIVE, and we observed a small overlap of 22 genes (figure 4A). Thus, we performed the GO term enrichment analysis on the 81 genes selected by MINT to investigate the biological significance associated with these genes. As shown in figure 4B, there is no enriched term related to stress response or specifically correlated to plant response as showed by genes selected by HIVE, but mainly very generic processes like gene expression and cellular processes (refer to supplementary table 2C for the complete list). The possible limitations in the performances of MINT in this case study could be due to the complex experimental design mixing time-course and single time point experiments making difficult to choose the classes to perform the classification and the limited number of replicates of some classes, reinforcing the need to develop a new tool to analyse more complex experimental designs.

These analyses corroborate the reliability of the results found by HIVE and the achievement of better performances compared to the state-of-the-art tool.

HIVE selects stress-responsive genes.

The meta-analysis of the biotic, abiotic and cross stress separately performed in the original study yielded a total of 5728 differentially expressed genes (DEG), including 1224, 2200 and 3345 DEG, respectively (Mota et al., 2021). Importantly, the overlap of the three lists of DEG is of only 86 genes. The small overalp of the three DEG lists obtained from each separated study represents the major limitation of the meta-analysis since the number of common genes usually decrease with the increase of the number of studies considered. The integrated analysis with HIVE retrieved 1416 genes found by the meta-analysis thus yielding a higher overlap than the three separated meta-analyses (figure 5A). The selected genes by HIVE do not only retrieve DEG because the meta-analysis was performed to select commonly regulated genes among all studies instead HIVE can capture more complex patterns including not similarly regulated genes.To further inspect the biological meaning of genes found by HIVE compared with the meta-analysis, we performed GO term enrichment analysis of each group separately (figure 5B and supplementary table 2D). We notice that terms related to metabolic processes

and responses to stimulus and stress are in common between the two analyses. Furthermore, among the terms found enriched in the genes selected by HIVE not in common with the meta-analysis, we found processes commonly impacted when stresses occur such as cell communication, cellular protein metabolic processes and reproduction.
Altogether these results showed the advantages of performing an integrated analysis over the meta-analysis.

HIVE identifies genes involved in the main hormone signalling pathways, kinases and resistance in response to stress.

To further characterize whether HIVE can capture genes involved in plant responses to stresses, we inspected the 5841 genes found by our integromics approach searching for genes involved in the main phytohormones signalling pathways, kinases and resistance genes.
82 genes were identified belonging to seven hormone signalling pathways, three directly connected with abiotic/biotic stress response (abscisic acid, jasmonate and ethylene) and four growth-promoting hormones (gibberellins, auxins, cytokinins and brassinosteroid). Remarkably, 60% of these genes were not previously found in the meta-analysis, reinforcing the validity of our novel approach (figure 6A). Considering the pivotal importance of plant hormones in mediating the regulation of stress responses, we focused on studying the expression profiles of these genes using a hierarchical clustering approach. Overall we observe that these genes are mainly activated under biotic stress and mostly in *A. stenosperma* and repressed in abiotic stress, especially in *A. duranensis* (figure 6B). Nevertheless, genes involved in Jasmonate and Brassinosteroids have a very similar profile of expression in both plant species and are mainly activated in biotic stress with higher expression at 6dpi and 3dpi for *A. stenosperma* and *A. duranensis*, respectively and mostly repressed in abiotic stress for *A. duranensis*. Similarly, genes in the Auxin pathway show mostly repression in abiotic stress for both species and mainly for *A. duranensis* and are activated during biotic stress however with different patterns depending on the plant species. The two growth hormones Gibberellins and Cytokines related genes show different expression profiles: while some genes are activated in both stresses and plant species, others show a more specific pattern. Interestingly, genes in the Ethylene pathway are strongly activated in *A. duranensis* and mildly in *A. stenosperma* during RKN infection and genes in the Abscisic pathway show the opposite pattern of expression during biotic stress and also activation in response to abiotic stress.
Previous analysis of the separated conditions showed that the jasmonic acid pathway was prevalently induced during biotic interaction, while abscisic acid upon abiotic stress (Mota et al., 2021). The integrated analysis allowed to identify other important hormone pathways involved in defence response and their different modulation depending on the plant species.
Plants have evolved a variety of mechanisms to cope with environmental changes, among which kinases play an important role in regulating both plant development and stress responses (Jones and Dangl, 2006; Soltabayeva et al., 2022; Takahashi and Shinozaki, 2019). Plant kinases are originally sensors for a variety of signals like phytohormones and small peptides, facilitating intercellular communication. Among HIVE-selected genes, we have found 208 genes coding for kinases. By

inspecting their transcriptional profiles, we identified 6 clusters (Figure 6C). Three clusters (B, D, F) group kinases induced in a stress and plant species-specific fashion (biotic - *A. duranensis*, biotic - *A. stenosperma*, abiotic - *A. stenosperma*, respectively). Kinases in cluster C are activated in response to biotic stress in both plant species. The last two clusters bear genes showing modulated activation in multiple stresses and plant species at the same time.

Other key players of plant immunity are resistance genes (R-genes) (Jones and Dangl, 2006). They code for the R-proteins, which have a main role in recognizing secreted pathogen-specific proteins (effectors) and activate a powerful, pathogen-specific immune response, which often includes transcriptional re-programming (Tsuda and Katagiri, 2010). Most of the R genes in plants encode nucleotide-binding site leucine-rich repeat (NBS-LRR) proteins (McHale et al., 2006). This large family is encoded by hundreds of diverse genes per genome and can be subdivided into the functionally distinct TIR-domain-containing (TNL) and CC-domain-containing (CNL) subfamilies. Despite being very difficult to find because of their very low expression (von Dahlen et al., 2023), HIVE identified 66 NBS-LRR genes whose expression profiles are reported in Figure 6D. We mainly observe three patterns: either the genes induced upon biotic stress mainly in one of the two plant species, and a small group with very heterogeneous patterns that were not previously found by non-integrated analysis methods (Mota et al., 2018). Of note, nine of them were previously validated by qRT-PCR experiments (Mota et al., 2018), reinforcing the reliability of the results found by HIVE.

In conclusion, by integrating experiments on different plant species each affected by one stress, HIVE can capture genes triggered in the different conditions linked to main hormone signalling pathways, kinases and R-genes in response to stress.

The integrated analysis allowed to infer the complex gene regulatory network induced by biotic and/or abiotic stress.

Plant responses to stress are controlled by a complex gene regulatory network (GNR). To reconstruct a comprehensive GNR, we first extracted the transcription factors (TF) 328 among the genes selected by HIVE analysis (supplementary table 3). Among the TF families with more than 10 TF found by our analysis (cross-checked with PlantTFDB), we retrieve families with known pivotal roles in development, plant growth and stress responses (figure 7A) (Baillo et al., 2019; Huynh-Thu et al., 2010). Then we collected TF information from PlantRegMap and we found 57 of the selected TF distributed in 21 families (figure 7B). We inferred two GNRs. The first is by considering the 57 TF found also in PlantRegMap as the “regulators” and the rest of the 5513 genes (without the 271 remaining TFs) as targets. By using the GENIE3 (Huynh-Thu et al., 2010) model and selecting edges with a weight >= 0.09(see methods), we inferred 2916 connections. Then we filtered out the pairs not in agreement with the prediction from PlantRegMap, obtaining 146 interactions (figure 7C). To show the advantage of using HIVE to select the genes to be used to infer the network instead of the meta-analysis approach, we performed the same analysis on the genes selected by the latter method. Among the 379 TFs in the meta-analysis, 76 are also present in PlantRegMap, yielding 5349 genes as putative targets (figure

7D). Applying the same threshold, we selected 2919 pairs. By filtering out the pairs not in agreement with PlantRegMap we obtained 86 interactions, 40% less than the pairs found with HIVE genes. These results suggest that more real direct targets can be predicted from HIVE genes.

We inspected the 146 pairs inferred from the genes identified by HIVE. We have found four TFs regulating genes involved in phytohormone pathways. The Dof TF Aradu.XZ0HK and the Gibbrellin-related gene Aradu.J1UD9 and the Golden2-like TF Aradu.LB6JY with the gene Aradu.IL3D8 is also involved in the Gibberellin pathway, with both pairs showing correlated expression profiles (supplementary figure 4A). The NAC TF and the Aradu.ZM0JI Auxin-related gene, mainly expressed during biotic interaction with mildly anticorrelated patterns (supplementary figure 4B). Finally, the MYB TF Aradu.545W2 and the Abscisic-related gene show opposite activation: while the TF is mainly expressed during biotic interaction in *A. stenosperma*, the predicted target is induced in abiotic stress in the same plant species, suggesting a possible antagonist mechanism (supplementary figure 4C).

To reconstruct a more general regulatome, we inferred a second network using the same strategy, indicating the remaining 271 TFs as “regulators” for which no predicted interaction was reported in PlantRegMap and the 5367 genes as “targets”. By selecting the top 2500 pairs, we obtained a big connected community and several isolated interactions (supplementary figure 5). We focused on the main network for further investigations, and we applied the Leiden algorithm to find communities of genes densely connected (ref). Leiden community detection is a fast and accurate method that allows us to find well-connected and refined partitions of the network (ref). We obtained 17 communities as shown in Figure 7E. Then, we quantified the number of TF belonging to the different families in each community (supplementary table 4). Interestingly, in community 1 (figure 7F), the most represented family is C2H2 zinc-finger domain proteins. These TFs play a pivotal role in response to abiotic stress, mainly drought (Han et al., 2020), and this family is not very present in other communities. Among the predicted targets by GENIE3, there are genes involved in nitrogen compound metabolism (FDR=0.00037) and genes connected to phytohormone pathways connected with plant growth and abiotic stress response (Ethylene, Gibberellin, Abscisic and Cytokines) (supplementary table 5). Nitrogen is an essential nutrient for plant growth and development and this process can be severely impacted during drought stress (Han et al., 2022). This analysis suggests a putative role in nitrogen metabolism mediated by C2H2 TF family, as was recently shown in rice (Han et al., 2022). Similarly, in community 3 (figure 7G) we have found mainly TFs of the Auxin Responsive Factor (ARF) family as long as genes involved in carbohydrate binding (FDR 0.0004) suggesting a role in the regulation of this process during the stress response, as shown in tomato (Yuan et al., 2019). Community 6 (figure 7H) shows a high amount of Ethylene Responsive Factors (ERF) TFs and genes involved in the Jasmonate and Auxin hormones pathway. The crosstalk among these pathways was already suggested in Zota-Motta et al. for the studied biological model as alternative pathways of response to drought stress, RKN infection or both stresses simultaneously. With this analysis, we suggest potential key interactions to further inspect. Finally, Golden2-like (G2-like) TFs are mostly abundant in communities 0 and 7 (figure 7I). Of note these communities are also enriched in genes involved in stress response (community 0: FDR= 0.0069, community 7: FDR=0.0033) and response to stimulus (community 0: FDR= 0.015,

community 7: FDR= 0.03) and they bear 50% of the NBS-LRR genes present in the network (supplementary table 6). Importantly in community 7, there is also the gene ARADU.4196P encoding for an endochitinase, that was previously shown to confer resistance to both stresses *in planta* (Mota et al., 2021). The induction of chitinases by specific agents leads to the simultaneous enhancement of other defense reactions and pathogenesis-related proteins in the same plant tissues. With this analysis, we suggest a transcription factor of the G2-like family as a potential regulator of this key gene and genes mainly involved in stress response and resistance genes.

Overall, the integrated analysis with HIVE allowed the selection of genes with pivotal roles in stress responses and joined to GENIE3 model, to infer a comprehensive GNR to better understand the crosstalk among the different strategies used by plants to respond to diverse stresses.

Conclusions

In this paper, we presented HIVE, a novel method to integrate and analyse multi-unpaired omics data with a complex experimental design. HIVE is an unsupervised model that uses a variational autoencoder to remove the batch effects by creating a compressed representation of the original data called latent space. This is a novel representation that captures the salient patterns to describe the experimental phenotype. With the aim of finding genes with expression profile characteristics of the studied phenotype, we developed a strategy to find genes explaining the latent features by using a random forest regression coupled with the SHAP explainer. This procedure decomposes the problem to infer the relationship between a set of N latent features and G genes into a set of N regression problems. Although variational autoencoders have already been extensively applied to analyse omics and multi-omics data, HIVE is the first model adapted for unpaired omics data. Furthermore, autoencoders and variants have mainly been employed for sample classification in a supervised fashion (Adamer et al., 2022; Chazarra-Gil et al., 2021; Johnson et al., 2007; Lazar et al., 2013; Rong et al., 2020; Tran et al., 2020). Sample classification is not among the tasks of this implementation and our model is unsupervised to capture the data structure without imposing constraints. Indeed, we showed that our unsupervised model performs better than the supervised alternative MINT (Rohart et al., 2017a) in the case of complex experimental design mixing longitudinal and static experiments with few replicates for each condition. The main advantage of the integrated analysis with HIVE is the possibility of finding genes modulated in different conditions without performing several comparisons that usually yield a very poor agreement.

HIVE is particularly appropriate in the field of plant biology. Indeed plants, as sessile organisms, need to continuously adapt to the changing environment and various stresses by reprogramming gene expression profiles. Characterizing the regulation of gene expression is pivotal to understand how plants respond to the biological environment. The selected case study to test HIVE is a dataset composed of seven distinct transcriptomic experiments of two peanut wild relatives' plants during root-knot nematode *Meloidogyne arenaria* infection and/or drought stress. Overall HIVE selected several genes implicated in pivotal molecular processes involved in plant response to stress. Furthermore, by studying the pattern of expression of these genes we showed how these pathways are modulated depending on the different phenotypes. Indeed, HIVE selected genes that are exclusively activated or repressed upon one stress in a particular plant species but also more complex patterns such as genes modulated differently in the studied conditions. These results can help to better understand plant response to different stresses constituting a useful resource of potential candidate genes for engineering a broader stress tolerance. We also showed that by applying the network inference model GENIE3 (Huynh-Thu et al., 2010) on genes selected by HIVE we were able to reconstruct a comprehensive interactome of *Arachis* to respond to *M. arenaria* infection and drought stress. We showed that the candidates provided by HIVE, enabled to reconstruct a network with more reliable connections than by using genes selected by the meta-analysis. The study of the inferred network allowed to better understand the crosstalk in response to biotic and abiotic stress conditions.

Overall, HIVE is a valuable tool to study multi-omics data for plant-stress response. HIVE can be applied on other phytopathosystems and will provide improvements in our understanding of the mechanisms set up by the plant upon biotic interactions and can elucidate important signatures to block the development of plant-pathogen interactions and allow to engineer durable disease resistance.

Competing Interests

The authors declare that they have no competing interests.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: a system for large-scale machine learning, in: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI'16. USENIX Association, USA, pp. 265–283.

Adamer, M.F., Brüningk, S.C., Tejada-Arranz, A., Estermann, F., Basler, M., Borgwardt, K., 2022. reComBat: batch-effect removal in large-scale multi-source gene-expression data integration. Bioinformatics Advances 2, vbac071. https://doi.org/10.1093/bioadv/vbac071

Baillo, E.H., Kimotho, R.N., Zhang, Z., Xu, P., 2019. Transcription Factors Associated with Abiotic and Biotic Stress Tolerance and Their Potential for Crops Improvement. Genes (Basel) 10, 771. https://doi.org/10.3390/genes10100771

Bodein, A., Scott-Boyer, M.-P., Perin, O., Lê Cao, K.-A., Droit, A., 2022. timeOmics: an R package for longitudinal multi-omics data integration. Bioinformatics 38, 577–579. https://doi.org/10.1093/bioinformatics/btab664

Bolger, M.E., Weisshaar, B., Scholz, U., Stein, N., Usadel, B., Mayer, K.F.X., 2014. Plant genome sequencing - applications for crop improvement. Curr Opin Biotechnol 26, 31–37. https://doi.org/10.1016/j.copbio.2013.08.019

Breiman, L., 2001. Random Forests. Machine Learning 45, 5–32. https://doi.org/10.1023/A:1010933404324

Chazarra-Gil, R., van Dongen, S., Kiselev, V.Y., Hemberg, M., 2021. Flexible comparison of batch correction methods for single-cell RNA-seq using BatchBench. Nucleic Acids Res 49, e42. https://doi.org/10.1093/nar/gkab004

Chollet, F., others, 2018. Keras: The Python Deep Learning library. Astrophysics Source Code Library ascl:1806.022.

Dash, S., Cannon, E.K.S., Kalberer, S.R., Farmer, A.D., Cannon, S.B., 2016. Chapter 8 - PeanutBase and Other Bioinformatic Resources for Peanut, in: Stalker, H.T., F. Wilson, R. (Eds.), Peanuts. AOCS Press, pp. 241–252. https://doi.org/10.1016/B978-1-63067-038-2.00008-3

Goh, W.W.B., Wang, W., Wong, L., 2017. Why Batch Effects Matter in Omics Data, and How to Avoid Them. Trends Biotechnol 35, 498–507. https://doi.org/10.1016/j.tibtech.2017.02.012

Großkinsky, D.K., Syaifullah, S.J., Roitsch, T., 2018. Integration of multi-omics techniques and physiological phenotyping within a holistic phenomics approach to study senescence in model and crop plants. J Exp Bot 69, 825–844. https://doi.org/10.1093/jxb/erx333

Hagberg, A.A., Schult, D.A., Swart, P.J., 2008. Exploring Network Structure, Dynamics, and Function using NetworkX.

Han, G., Lu, C., Guo, J., Qiao, Z., Sui, N., Qiu, N., Wang, B., 2020. C2H2 Zinc Finger Proteins: Master Regulators of Abiotic Stress Responses in Plants. Frontiers in Plant Science 11.

Han, M.-L., Lv, Q.-Y., Zhang, J., Wang, T., Zhang, C.-X., Tan, R.-J., Wang, Y.-L., Zhong, L.-Y., Gao, Y.-Q., Chao, Z.-F., Li, Q.-Q., Chen, G.-Y., Shi, Z., Lin, H.-X., Chao, D.-Y., 2022. Decreasing nitrogen assimilation under drought stress by suppressing DST-mediated activation of Nitrate Reductase 1.2 in rice. Molecular Plant 15, 167–178. https://doi.org/10.1016/j.molp.2021.09.005

Han, W., Li, L., 2022. Evaluating and minimizing batch effects in metabolomics. Mass Spectrom Rev 41, 421–442. https://doi.org/10.1002/mas.21672
Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., Geurts, P., 2010. Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLOS ONE 5, e12776. https://doi.org/10.1371/journal.pone.0012776
Ishwaran, H., 2015. The effect of splitting on random forests. Mach Learn 99, 75–118. https://doi.org/10.1007/s10994-014-5451-2
Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., Gao, G., 2017. PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res 45, D1040–D1045. https://doi.org/10.1093/nar/gkw982
Johnson, W.E., Li, C., Rabinovic, A., 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8, 118–127. https://doi.org/10.1093/biostatistics/kxj037
Jones, J.D.G., Dangl, J.L., 2006. The plant immune system. Nature 444, 323–329. https://doi.org/10.1038/nature05286
Kingma, D.P., Welling, M., 2022. Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114
Kolde, R., 2019. pheatmap: Pretty Heatmaps.
Kornbrot, D., 2005. Point Biserial Correlation, in: Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa485
Lakkis, J., Wang, D., Zhang, Y., Hu, G., Wang, K., Pan, H., Ungar, L., Reilly, M.P., Li, X., Li, M., 2021. A joint deep learning model enables simultaneous batch effect correction, denoising, and clustering in single-cell transcriptomics. Genome Res 31, 1753–1766. https://doi.org/10.1101/gr.271874.120
Lazar, C., Meganck, S., Taminau, J., Steenhoff, D., Coletta, A., Molter, C., Weiss-Solís, D.Y., Duque, R., Bersini, H., Nowé, A., 2013. Batch effect removal methods for microarray gene expression data integration: a survey. Brief Bioinform 14, 469–490. https://doi.org/10.1093/bib/bbs037
Lê Cao, K.-A., Boitard, S., Besse, P., 2011. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics 12, 253. https://doi.org/10.1186/1471-2105-12-253
Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A., 2010. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet 11, 733–739. https://doi.org/10.1038/nrg2825
Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M.P., Hu, G., Li, M., 2020. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun 11, 2338. https://doi.org/10.1038/s41467-020-15851-3
Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.-I., 2020. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9
Lundberg, S.M., Lee, S.-I., 2017. A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Mahmood, U., Li, X., Fan, Y., Chang, W., Niu, Y., Li, J., Qu, C., Lu, K., 2022. Multi-omics revolution to promote plant breeding efficiency. Frontiers in Plant Science 13.

McHale, L., Tan, X., Koehl, P., Michelmore, R.W., 2006. Plant NBS-LRR proteins: adaptable guards. Genome Biology 7, 212. https://doi.org/10.1186/gb-2006-7-4-212

Mohammadi-Shemirani, P., Sood, T., Paré, G., 2023. From 'Omics to Multi-omics Technologies: the Discovery of Novel Causal Mediators. Curr Atheroscler Rep 25, 55–65. https://doi.org/10.1007/s11883-022-01078-8

Molania, R., Foroutan, M., Gagnon-Bartsch, J.A., Gandolfo, L.C., Jain, A., Sinha, A., Olshansky, G., Dobrovic, A., Papenfuss, A.T., Speed, T.P., 2023. Removing unwanted variation from large-scale RNA sequencing data with PRPS. Nat Biotechnol 41, 82–95. https://doi.org/10.1038/s41587-022-01440-w

Mota, A.P.Z., Brasileiro, A.C.M., Vidigal, B., Oliveira, T.N., da Cunha Quintana Martins, A., Saraiva, M.A. de P., de Araújo, A.C.G., Togawa, R.C., Grossi-de-Sá, M.F., Guimaraes, P.M., 2021. Defining the combined stress response in wild Arachis. Sci Rep 11, 11097. https://doi.org/10.1038/s41598-021-90607-7

Mota, A.P.Z., Vidigal, B., Danchin, E.G.J., Togawa, R.C., Leal-Bertioli, S.C.M., Bertioli, D.J., Araujo, A.C.G., Brasileiro, A.C.M., Guimaraes, P.M., 2018. Comparative root transcriptome of wild Arachis reveals NBS-LRR genes related to nematode resistance. BMC Plant Biology 18, 159. https://doi.org/10.1186/s12870-018-1373-7

Nygaard, V., Rødland, E.A., Hovig, E., 2016. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. Biostatistics 17, 29–39. https://doi.org/10.1093/biostatistics/kxv027

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., 2011. Scikit-learn: Machine Learning in Python. MACHINE LEARNING IN PYTHON.

Pratella, D., Ait-El-Mkadem Saadi, S., Bannwarth, S., Paquis-Fluckinger, V., Bottini, S., 2021. A Survey of Autoencoder Algorithms to Pave the Diagnosis of Rare Diseases. Int J Mol Sci 22, 10891. https://doi.org/10.3390/ijms221910891

Rau, A., Marot, G., Jaffrézic, F., 2014. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics 15, 91. https://doi.org/10.1186/1471-2105-15-91

Rhee, S.Y., Beavis, W., Berardini, T.Z., Chen, G., Dixon, D., Doyle, A., Garcia-Hernandez, M., Huala, E., Lander, G., Montoya, M., Miller, N., Mueller, L.A., Mundodi, S., Reiser, L., Tacklind, J., Weems, D.C., Wu, Y., Xu, I., Yoo, D., Yoon, J., Zhang, P., 2003. The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. Nucleic Acids Research 31, 224–228. https://doi.org/10.1093/nar/gkg076

Rohart, F., Eslami, A., Matigian, N., Bougeard, S., Lê Cao, K.-A., 2017a. MINT: a multivariate integrative method to identify reproducible molecular signatures across independent experiments and platforms. BMC Bioinformatics 18, 128. https://doi.org/10.1186/s12859-017-1553-8

Rohart, F., Gautier, B., Singh, A., Cao, K.-A.L., 2017b. mixOmics: An R package for 'omics feature selection and multiple data integration. PLOS Computational Biology 13, e1005752. https://doi.org/10.1371/journal.pcbi.1005752

Rong, Z., Tan, Q., Cao, L., Zhang, L., Deng, K., Huang, Y., Zhu, Z.-J., Li, Z., Li, K., 2020. NormAE: Deep Adversarial Learning Model to Remove Batch Effects in Liquid Chromatography Mass Spectrometry-Based Metabolomics Data. Anal Chem 92, 5082–5090. https://doi.org/10.1021/acs.analchem.9b05460

Schwacke, R., Ponce-Soto, G.Y., Krause, K., Bolger, A.M., Arsova, B., Hallab, A., Gruden, K., Stitt, M., Bolger, M.E., Usadel, B., 2019. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Mol Plant 12, 879–892. https://doi.org/10.1016/j.molp.2019.01.003

Shrikumar, A., Greenside, P., Kundaje, A., 2019. Learning Important Features Through Propagating Activation Differences. https://doi.org/10.48550/arXiv.1704.02685

Soltabayeva, A., Dauletova, N., Serik, S., Sandybek, M., Omondi, J.O., Kurmanbayeva, A., Srivastava, S., 2022. Receptor-like Kinases (LRR-RLKs) in Response of Plants to Biotic and Abiotic Stresses. Plants (Basel) 11, 2660. https://doi.org/10.3390/plants11192660

Subramanian, I., Verma, S., Kumar, S., Jere, A., Anamika, K., 2020. Multi-omics Data Integration, Interpretation, and Its Application. Bioinform Biol Insights 14, 1177932219899051. https://doi.org/10.1177/1177932219899051

Sugimoto, K., Xu, L., Paszkowski, U., Hayashi, M., 2018. Multifaceted Cellular Reprogramming at the Crossroads Between Plant Development and Biotic Interactions. Plant Cell Physiol 59, 651–655. https://doi.org/10.1093/pcp/pcy066

Takahashi, F., Shinozaki, K., 2019. Long-distance signaling in plant stress response. Current Opinion in Plant Biology, Growth and development 47, 106–111. https://doi.org/10.1016/j.pbi.2018.10.006

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J., Gao, G., 2020. PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res 48, D1104–D1113. https://doi.org/10.1093/nar/gkz1020

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., Su, Z., 2017. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res 45, W122–W129. https://doi.org/10.1093/nar/gkx382

Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., Chen, J., 2020. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biology 21, 12. https://doi.org/10.1186/s13059-019-1850-9

Tsuda, K., Katagiri, F., 2010. Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. Curr Opin Plant Biol 13, 459–465. https://doi.org/10.1016/j.pbi.2010.04.006

Ulfenborg, B., 2019. Vertical and horizontal integration of multi-omics data with miodin. BMC Bioinformatics 20, 649. https://doi.org/10.1186/s12859-019-3224-4

Velten, B., Braunger, J.M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., Stegle, O., 2022. Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. Nat Methods 19, 179–186. https://doi.org/10.1038/s41592-021-01343-9

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W.,

VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors, 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272. https://doi.org/10.1038/s41592-019-0686-2

von Dahlen, J.K., Schulz, K., Nicolai, J., Rose, L.E., 2023. Global expression patterns of R-genes in tomato and potato. Frontiers in Plant Science 14.

Wang, T., Johnson, T.S., Shao, W., Lu, Z., Helm, B.R., Zhang, J., Huang, K., 2019. BERMUDA: a novel deep transfer learning method for single-cell RNA sequencing batch correction reveals hidden high-resolution cellular subtypes. Genome Biol 20, 165. https://doi.org/10.1186/s13059-019-1764-6

Yang, Y., Saand, M.A., Huang, L., Abdelaal, W.B., Zhang, J., Wu, Y., Li, J., Sirohi, M.H., Wang, F., 2021. Applications of Multi-Omics Technologies for Crop Improvement. Frontiers in Plant Science 12.

Yu, Y., Zhang, N., Mai, Y., Ren, L., Chen, Qiaochu, Cao, Z., Chen, Qingwang, Liu, Y., Hou, W., Yang, J., Hong, H., Xu, J., Tong, W., Dong, L., Shi, L., Fang, X., Zheng, Y., 2023. Correcting batch effects in large-scale multiomics studies using a reference-material-based ratio method. Genome Biology 24, 201. https://doi.org/10.1186/s13059-023-03047-z

Yuan, Y., Xu, X., Gong, Z., Tang, Y., Wu, M., Yan, F., Zhang, X., Zhang, Q., Yang, F., Hu, X., Yang, Q., Luo, Y., Mei, L., Zhang, W., Jiang, C.-Z., Lu, W., Li, Z., Deng, W., 2019. Auxin response factor 6A regulates photosynthesis, sugar accumulation, and fruit development in tomato. Hortic Res 6, 1–16. https://doi.org/10.1038/s41438-019-0167-x

Figures and Tables

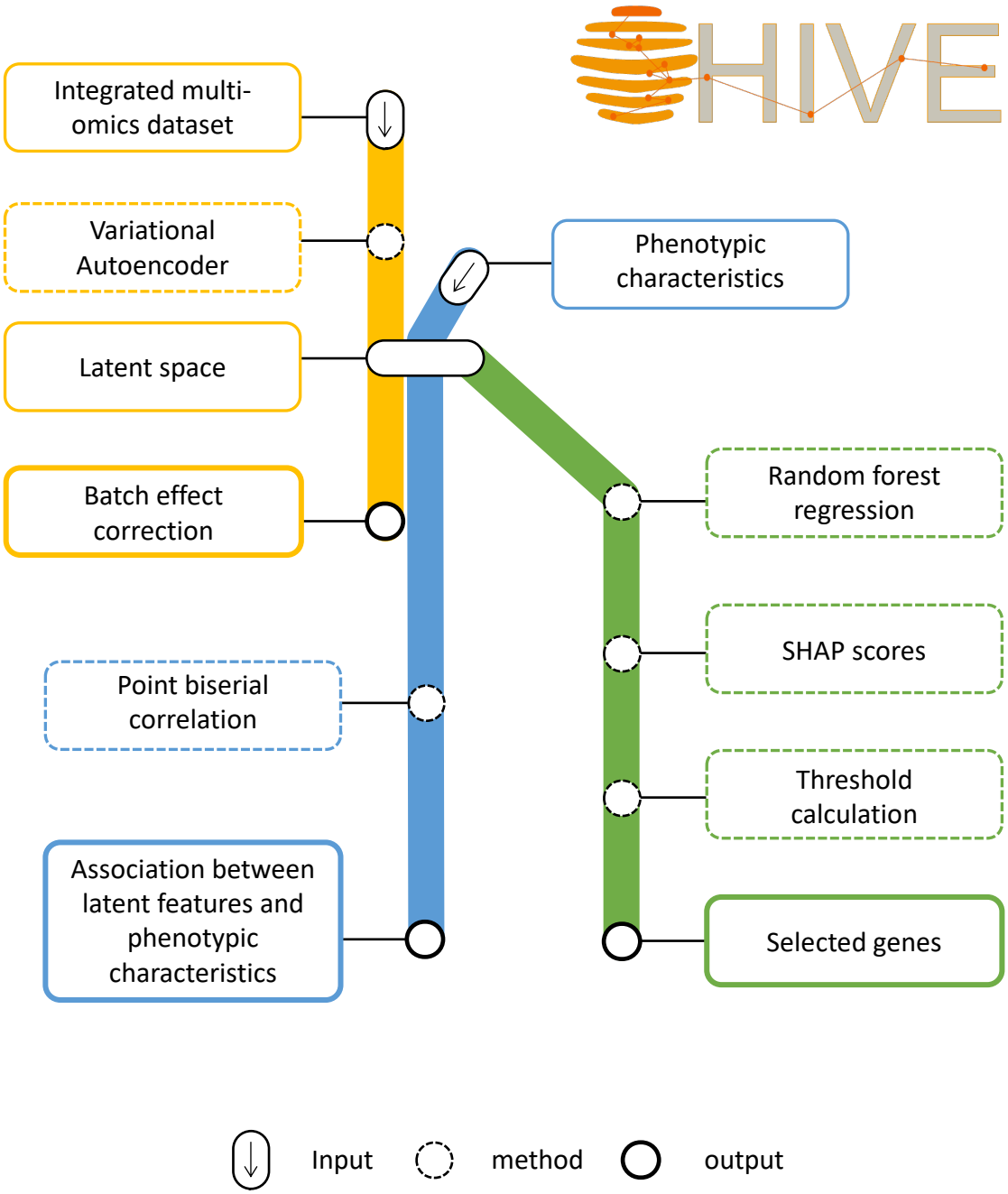

Figure 1. HIVE framework. The pipeline is represented as three underground lines to symbolize the three main blocks composing HIVE. The first block (yellow line) concerns the batch effect correction using a Variational AutoEncoder (VAE) to produce the latent space, that is the common "station" of all the lines. The two following lines are independent of each other, but they are consequent to block 1 because they use the latent space. The blue line regards the calculation of the association between the latent features and phenotypic characteristics using the point biserial correlation. The green line is composed of three steps to select relevant genes for the studied phenotype. First, a Random Forest Regression (RFR) is used where each latent feature is considered as the variable to explain, and the original gene expression data as the dependent variable. Afterward, the contribution of each gene to each regression was calculated with the SHAP algorithm. Finally, we included a custom method to set a threshold on the SHAP scores to select only the most contributing genes. The third block returns a list of ranked genes that can be used for further analysis.

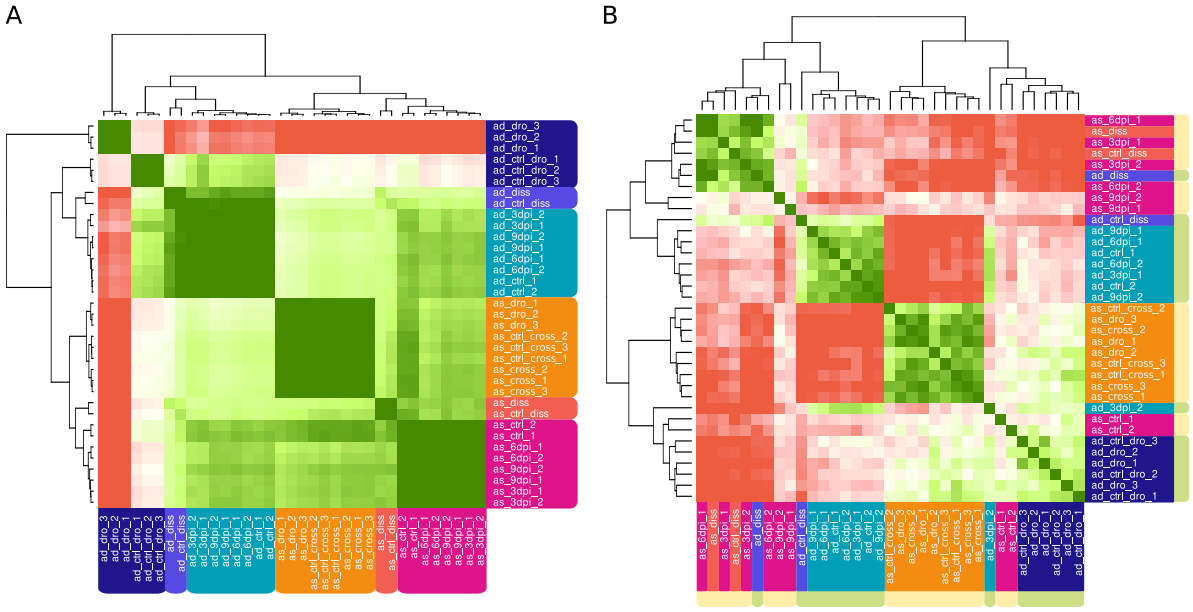

Figure 2. HIVE reduces the batch effects allowing unpaired experiments integrated analysis. Spearman correlation matrix of samples in the original data (A), and after VAE application (B). Hierarchical clustering is applied to show samples groups according to correlation values. Colors of samples according with the experiment batch.

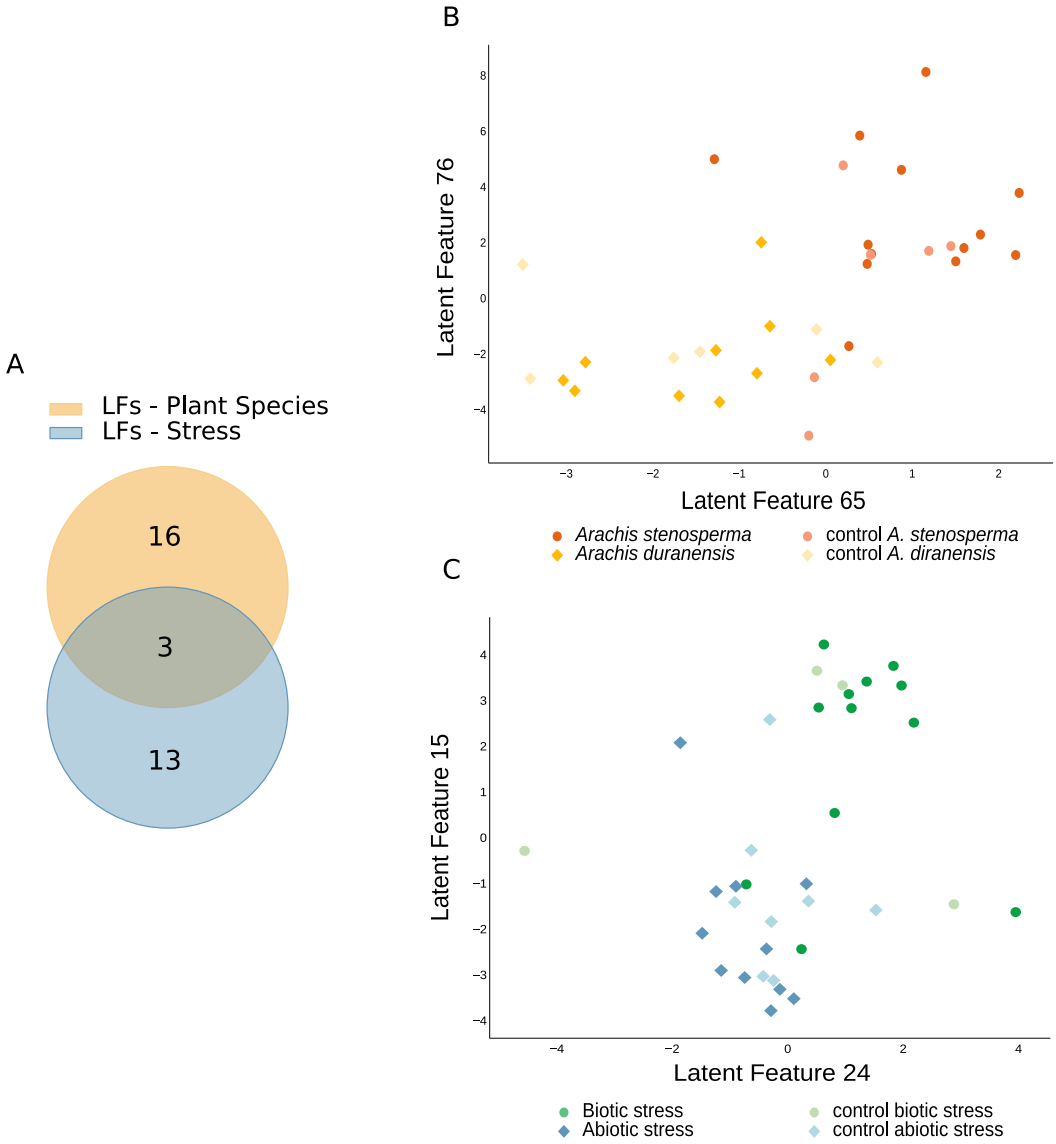

Figure 3. Latent features are associated with phenotypic characteristics. A) Agreement of latent features selected by the point biserial correlation (p-value < 0.01) as the most correlated either with the type of stress (yellow) or plant species (light blue). Visualization of top two correlated latent features with plant species (B) or with type of stress (C). Points represent samples, the condition is indicated in the correspondent legend.

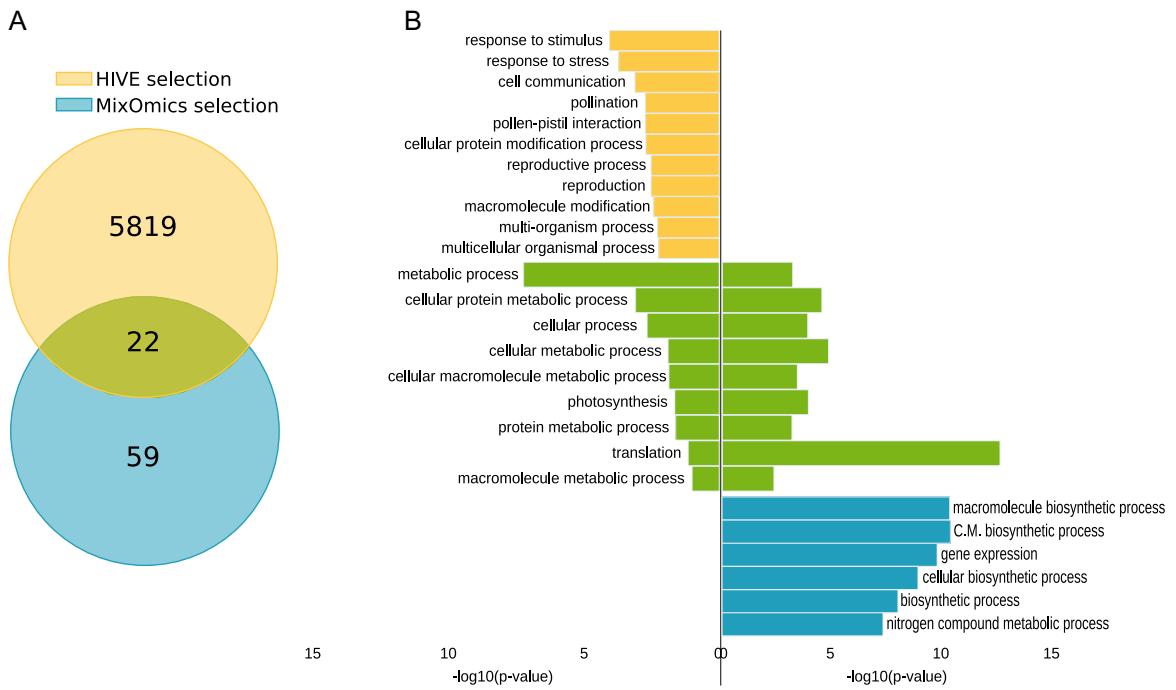

Figure 4. Comparing HIVE and MINT performances. A) Agreement of gene selected by HIVE and by MixOmics-MINT application on the integrated dataset. B) GO-enrichment analysis (p-value <= 0.01), in yellow the GO-terms found enriched only in HIVE, in green GO-terms found enriched in the two selections, and in blue the GO-terms found enriched only by MixOmics-MINT application.

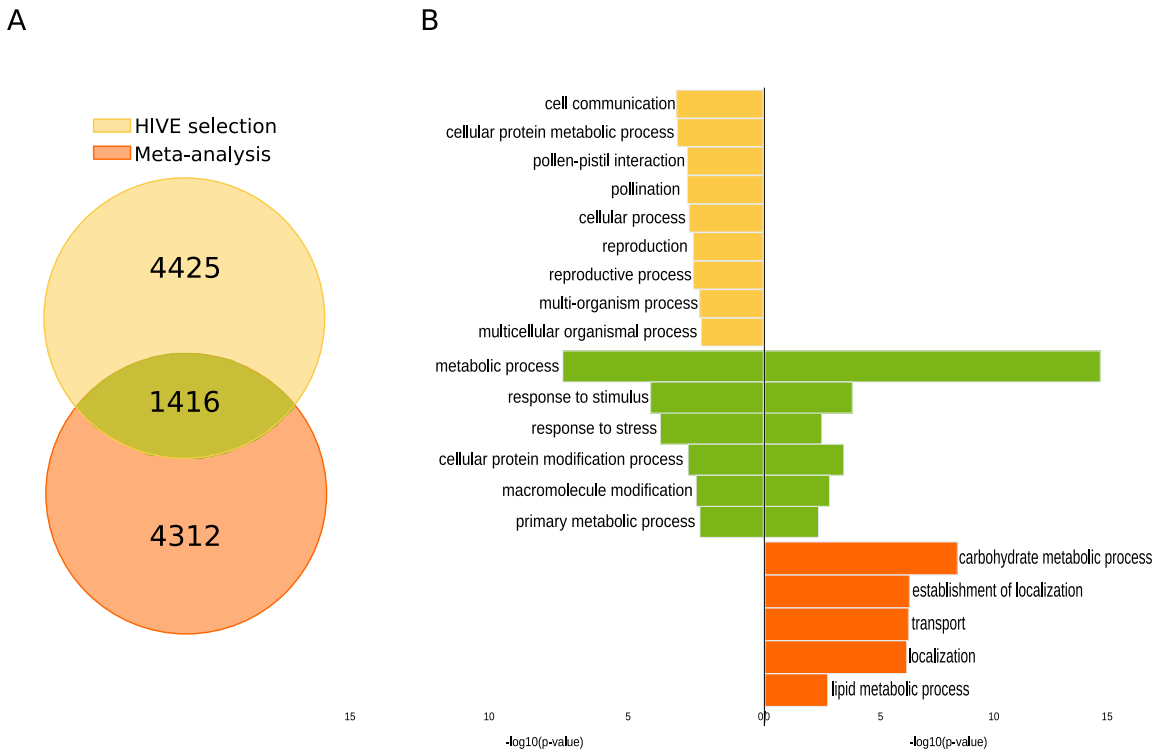

Figure 5. Comparing HIVE and meta-analysis. A) Agreement of gene selected by HIVE and by DEGs meta-analysis from Mota A., et al, 2021 on the integrated dataset. B) GO-enrichment analysis (p-value <= 0.01), in yellow the GO-terms found enriched only in HIVE, in green GO-terms found enriched in the two selections, and in dark orange the GO-terms found enriched only the meta-analysis.

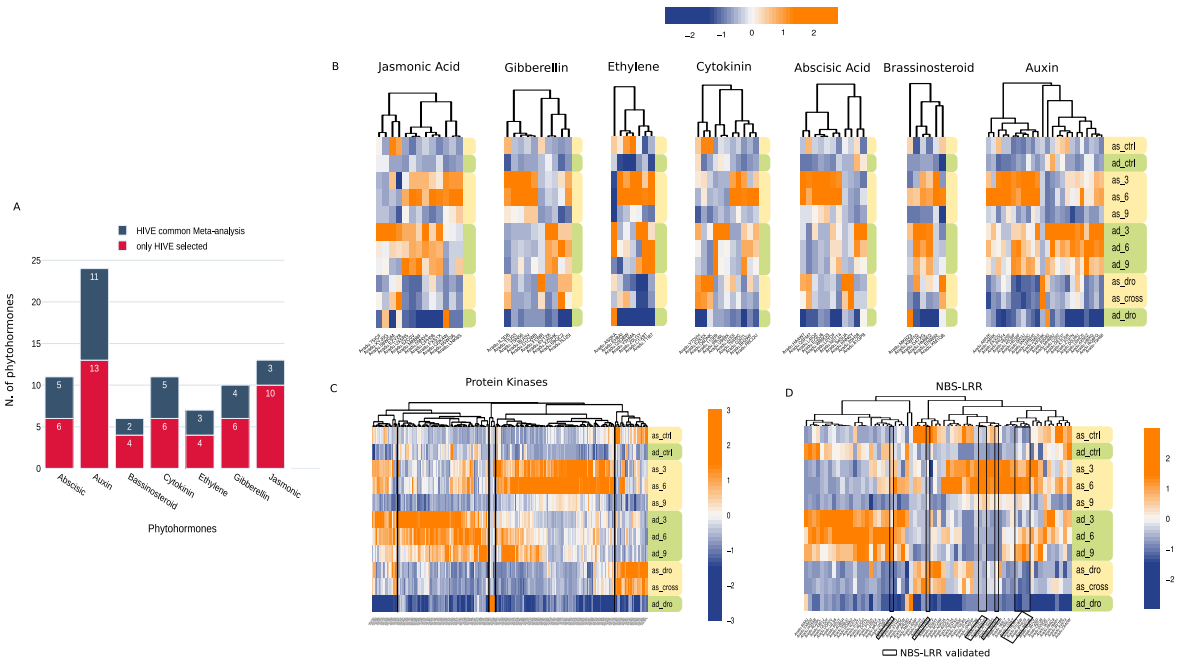

Figure 6. HIVE identify genes involved in the main hormone signaling pathways, kinases and resistance genes in response to stress. A) Phytohormone related genes abundance in HIVE selection, either in common with the meta-analysis from Mota A., et al, 2021 (dark blue) or found only by HIVE (red). B) Heatmaps of gene expression in each set of annotated phytohormones. C) Heatmap of gene expression for annotated protein kinases found in HIVE selection. D) Heatmap of gene expression for NBS-LRR genes found in HIVE selection; the black squares represent the experimentally validated NBS-LRR by Mota A., et al, 2018.

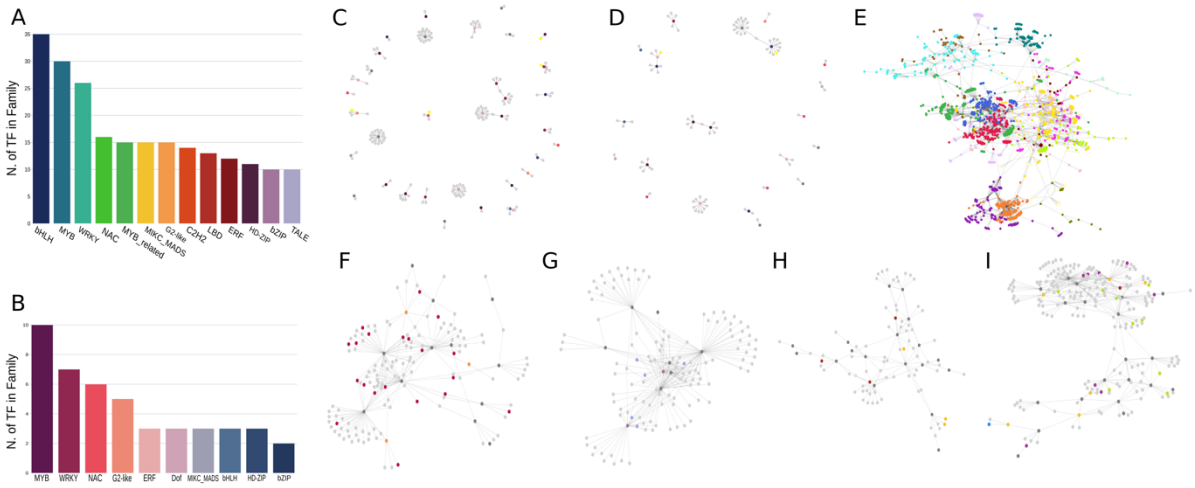

Figure 7. The integrated analysis allowed to infer the complex gene regulatory network induced upon biotic and/or abiotic stress.

A) Number of transcription factors in families with at least 10 representatives in HIVE selection. B) Number of transcription factors from HIVE selection with predicted binding motif in PlantRegMap; top ten most abundant families. C) GRN inferred by GENIE3 imposing as “regulators” the TF found by HIVE and PlantRegMap and as “targets” genes that are not TF, after removing all pairs not in PlantRegMap. Yellow nodes represent phytohormones related genes while the others are colored accordingly D) same as C but with gene selection from the meta-analysis. E) The largest connected network from the GRN inferred by GENIE3 imposing as “regulators” the TF found by HIVE and not in PlantRegMap and as “targets” genes that are not TF. Colors according to the 17 communities found by Leiden algorithm. F) Zoom-in into community n. 1 (green nodes in E)), Bordeaux nodes correspond to transcription factors of family C2H2, dark yellow nodes correspond to genes associated with GO-term GO:0006807 (nitrogen compound metabolism). G) Zoom-in into community n. 3 (electric blue nodes in E)), light purple nodes correspond to transcription factors of family ARF, lilac nodes correspond to genes associated with GO-term GO:0030246 (carbohydrate binding).H) Zoom-in into community n. 6 (electric light blue nodes in E)), bordeaux nodes correspond to transcription factors of family ERF, dark yellow nodes correspond to genes with Jasmonic acid annotation. I) Zoom-in into community n. 0 (red nodes in E)) on top and community 7 (fuchsia nodes in E)), dark yellow nodes correspond to transcription factors of family Golden2-like, bordeaux node corresponds to transcription factor of ERF family, light green nodes correspond to genes associated with GO-term GO:0006950 (response to stress) and GO:0050896 (response to stimulus), violet nodes correspond to NBS-LRR genes, blue node corresponds to the endochitinasi Aradu.4916P validate in planta from Mota A., et al 2021. In all network figures dark grey nodes represent all the other transcription factors in the network not belonging to the family/families of interest and all the light grey nodes are other regulated genes.

Plant Species	CTRL	3 dpi	6 dpi	9 dpi	Cross Stress	drought (7days)	desiccation	references
A.Steno.	2	2	2	2	-	-	-	Guimaraes et al., 2015
A.Steno.	3	-	-	-	3	3	-	Mota et al., 2021
A.Steno.	1	-	-	-	-	-	1	Vinson et al., 2018
total	6	2	2	2	3	3	1	
A.Dura.	2	2	2	2	-	-	-	Mota et al., 2018
A.Dura.	3	-	-	-	-	3	-	Mota et al., 2021
A.Dura.	1	-	-	-	-	-	1	Vinson et al., 2018
total	6	2	2	2	-	3	1	
global	12	4	4	4	3	6	2	

Table 1. Summary of experiments used to construct the integrated dataset. For each condition, the plant species, the number of replicates and the original reference are reported.