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

Giulia Calia[1,2,3*], Sophia Marguerit[3*], Ana Paula Zotta Mota[3], Manon Vidal[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

* These authors contributed equally to this work

#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-transcriptomics 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 Arachis wild species submitted to 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-transcriptomics integration; unpaired experiments; horizontal integration; variational autoencoders; SHAP explainer; plant responses to stress; Arachis; 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. These studies 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, omics and multi-omics studies have been employed in several biological fields, including the study of 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). Similar to many biological processes, the plant defense response mechanisms are composed of several physiological and molecular changes mediated by complex regulatory interactions. Those processes occur at different biological layers (such as post-transcriptional, post-translational, translational and transcriptional changes), depending on the cells or tissues and the spatio-temporal scales. In the case of parasitism, interactions are based on a molecular dialogue between the pathogen and its host that are influenced by varying spatial and temporal levels, altogether concurring for a successful or unsuccessful infection. During the infection process, pathogen proteins play a crucial role in the re-wiring of multiple biochemical processes occurring in the host, ultimately allowing for the infection's progression. Parasites modulate the immune response, cell signalling, MAPK and hormone signaling, vesicle trafficking systems, ubiquitination systems, metabolism or even plant growth (Asai and Shirasu, 2015). As a counterpart, hosts make use of their protein machinery to trigger defence mechanisms against the pathogen. Furthermore, as sessile organisms, plants must endure abiotic stresses such as drought, high or low temperature, nutrient deficiency, salinity that can occur concomitantly to pathogens attacks. All those conditions affect plant physiology causing several changes in molecular processes that can be non-adaptative or adaptative responses (Zhang et al., 2022). The former are direct responses to the stress, for instance protein structure changes caused by extreme temperatures or disruptions in enzyme kinetics and molecular interactions caused by toxic ions. The second category concerns responses that can improve stress resistance including the repair of stress-induced damage, the rebalancing of cellular homeostasis and the adjustment of growth (Zhang et al., 2020; Zhu, 2016). Those represent potential targets for crop improvement. Therefore, understanding how plants perceive stress signals and adapt to adverse environmental conditions both biotic and abiotic is critical for global food security.

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 omics data that are currently under-exploited due to the absence of a methodology to successfully integrate data from different experiments. 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 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).

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. On the other hand, 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 address these limitations, we developed HIVE (Horizontal Integration analysis using Variational AutoEncoders) to analyse at the same time unpaired and/or longitudinal multi-transcriptomics experiments. We show the capabilities of HIVE on a dataset composed of seven distinct transcriptomic experiments of two Arachis wild species during root-knot nematode *Meloidogyne arenaria* infection and/or drought stress.

Wild species are known to be a good source of resistance genes, due to their high capacity to cope with different stresses. The wild species *Arachis stenosperma* is known by its resistance to the infection of root-knot nematodes, developing a hypersensitive response (Proite et al., 2008). Another wild species from the genus Arachis, A. duranensis was studied for its high tolerance to hydric stresses (Leal-Bertioli et al., 2012). Different studies also assessed the different NLR (nucleotide-binding and leucine-rich repeat LRR domains) gene expression of *A. stenosperma* to different stresses (Mota et al., 2018; Rizwan et al., 2023; Song et al., 2017), and the description of resistance QTL to RKN (Leal-Bertioli et al., 2015).

Plant NLRs are composed of a central nucleotide-binding site (NBS) and a C-terminal LRR domain. Depending on the N-terminal domain, NLRs can be divided into those having a Toll and IL-1 receptor (TIR) domain and those with a coiled-coil (CC) domain (Takken and Goverse, 2012). In addition, some NLRs lack one or more of the canonical domains, commonly termed truncated NLRs, and can also be functional and are found in various plant species (Araújo et al., 2019; Baggs et al., 2017).

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 and a combination of the Kruskal-Wallis test followed by the post-hoc Dunn's test are applied to calculate the association between the latent features and phenotypic characteristics. The last block is composed of four steps to select relevant genes for the studied phenotype. First, a Random Forest Regression (RFR), in a three-fold cross-validation configuration 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 either the SHAP algorithm or the default feature importance method. Finally, we included a custom method to set a threshold on the SHAP scores to select only the most contributing genes. The last step consists in associating the genes to the most varying condition(s) compared to each control. At the end, HIVE returns a list of ranked genes that can be used for further analysis.

Block 1: batch effect correction

A VAE is used to encode the integrated dataset in a compressed representation called latent space. An Autoencoder is a type of deep neural network which is trained in a self-supervised manner (Baldi, 2012; Rumelhart et al., 1986). The symmetric structure is composed of an encoder, a decoder and a bottleneck in the middle which we will refer to as latent space. The network is therefore trained to reconstruct the input data with minimal error by learning to encode the N-dimensional data as a L-dimensional latent space, with L<<N. The VAE originally proposed by Kingma and Welling (Kingma and Welling, 2022), and Rezende et al (Rezende et al., 2014), can be seen as a probabilistic version of an AE, where the resulting model can be also used to generate new data. The encoder is trained to estimate the mean and the variance modeled by a multivariate gaussian mixture. Briefly, the latent variables Z are assumed to have distribution $p\theta$ and the observation X is assumed to be conditional distribution $p\theta(x|z)$. The generative model is trained to estimate the parameters θ to minimize the Kullback-Leiber (KL) divergence between the original data distribution and the model distribution (the marginal likelihood). Being the marginal likelihood analytically intractable because of the nonlinear relationship between the original variables and the latent components, the evidence lower bound (ELBO) approximation is used.

The VAE architecture of HIVE 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 and Mean Squared Error. 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). In case of more than two characteristics, we combine the use of the Kruskal-Wallis test to identify the latent features that are significantly associated with multiple-calls phenotypic characteristics, followed by the post-hoc Dunn's test to identify specifically to which characteristic each latent feature is associated with.

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 are filtered out, to avoid rows with only 0, and latent features are scaled using MinMaxScaler function of Scikit-learn python library (Pedregosa et al., 2011). We implemented a cross-validation strategy to train and test the regression models and identify the best performing one. Cross-validation is very useful in case of limited amount of data, because splitting into training and test may yield a very small test set (Arlot and Celisse, 2010; Efron and Tibshirani, 1997; James et al., 2013; Refaeilzadeh et al., 2009). In case of cross validation, we can build different models to be sure to make prediction on all the data. Here we used a stratified k-fold cross-validation that allow the model to make predictions on class-proportioned unseen data for each fold. Thus, we can use the entire dataset for training and test. The number of folds depends on the minimal number of replicates per condition to assure to have balanced representation of samples in each class during this phase. For the application on our integrated dataset, we performed a three-fold cross-validation. The RFR (Breiman, 2001) is implemented using RandomForestRegressor() method in sklearn.ensemble module from Scikit-learn python library, using default parameter and 100 as the number of estimators.

SHAP and Random Forest Importance scores

For each latent feature, the best performing model from cross-validation is used to calculate the contribution of each gene to each regression with the random forest importance score and SHAP

explainer. Regarding the first method, we used the fitted attribute, “feature_importances_"from RandomForestRegressor() method. This implementation uses the Gini impurity index (Ishwaran, 2015), that essentially calculates the goodness of each split in the decision trees in terms of class entropy. We retrieve the first 100 genes for each model, this is a customizable threshold. On the other hand, SHAP (SHapley Additive exPlanations) is a method based on the Shapley interaction index from game theory and used to increase transparency and interpretability of machine learning models (Lundberg and Lee, 2017). Briefly, the regression model can be seen as a cooperative game and the SHAP explainer assign a score to each feature (i.e. gene) representing its contribution to the model’s output. We used the SHAP TreeExplainer from SHAP python library (Lundberg et al., 2020).

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 is reported in Supplementary Figure 2B.

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 2C). 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 score. 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 by more than one feature, then only the highest rank will be maintained in the final list.

Gene association to condition

To identify in which condition(s) the selected genes have the highest deviation from the respective control sample, we set up a method based on the amplitude of variation. We calculated the absolute value of the log(FC) for each condition using the rlog function in the DESeq2 package (Love et al., 2014). Then for each gene we use the highest value as reference and calculated the distances with each of the other conditions. By selecting as threshold half of the standard deviation of the total distance distribution (Supplementary Figure 2D), we consider each gene to be associated to the condition with the highest abs(log(FC)) and all other condition(s) that eventually have a distance smaller than the selected threshold.

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 RandomForestRegressor() method of Scikit-learn. Therefore, for each Random Forest Regression model we used the attribute “feature_importances_” and, for each feature, we selected 100 genes.Then we 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 50 bins and for each bin calculated the agreement of the two methods.

Benchmarking

For the differential expression we used the DEseq2 R package (Love et al., 2014), comparing each condition with its own control, while for the meta-analysis the lists were provided in the Supplementary materials of Zotta-Mota et al. (2021) (Mota et al., 2021). 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.

Coefficient of variation

The coefficient of variation was calculated on the log2 of the fold change of each condition compared to the relative control using the function 'variation' of scipy, the upset plots were obtained with the function “upset” of the Python library “upsetplot” with the options subset_size="count" and sort_by='cardinality'.

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 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.

Hierarchical clustering of gene expression profiles

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 using as input the log2 fold-change calculated previously.

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.

Overexpression of NLRs in Arabidopsis

The 1,920 bp coding sequence of AsTIR19 from *A. stenosperma* used here was previously cloned in the pPZP_BAR binary vector, and successfully used to transform tobacco plants against *Sclerotinia sclerotiorum* (Guimaraes et al., 2022).
To identify the complete coding sequence of AsHLR, the Aradu.HLR71 gene model from *A. duranensis* was aligned with the four best BLASTn hits of *A. stenosperma* databases available on NCBI. The consensus sequence (1,296 bp) was synthesized and cloned (Epoch Life Science Inc., TX, United States), under the control of the *Arabidopsis* actin 2 promoter (ACT-2) and the nopaline synthase (NOS) terminator, at the XhoI restriction site of pPZP_BAR (Mota et al., 2019).
Both vectors were used for *Arabidopsis thaliana* ecotype Columbia (Col-0) transformation using the floral dip method (Clough and Bent, 1998). Plants were maintained in a growth chamber at 21°C, 60%

relative humidity, 12-hour photoperiod until seed production. Seeds from the dipped-plants were germinated on plates containing solid basal medium (0.8% w/v agar) Murashige and Skoog (MS) solid (Sigma-Aldrich, St. Louis City, United States) with sucrose (30 g /L) and hygromycin (15 mg/L). T0 plants resistant to hygromycin were transferred to pots containing substrate and maintained in a controlled growth chamber to produce the T1 and T2 generations. The overexpression of the transgenes was confirmed in the OE lines at T2 generation by qRT-PCR analysis.

Assessment of the effect of NLR genes on RKN resistance

For the *M. incognita* bioassay, 4 weeks old plants were challenged with 500 J2 of M. incognita. Ten plants from each of the six Arabidopsis OE lines of each gene construct were used, and after 60 days of infection (60 DAI), the number of eggs per gram of root (N° eggs/gr of root) was estimated in transformed plants and control plants (WT) according to (Júnior et al., 2013)Antonino et al, (2013). The data was statistically evaluated using the Tukey test ($p<0.05$).

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. First, 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. The Kruskal-Wallis followed by the post-hoc Dunn test, was then used to compare pair-wisely the conditions and counting how many latent features are associated to the discrimination of each pair (Figure 3D). *A. duranensis* biotic and *A. stenosperma* drought showed the highest number of latent features (30 latent features) associated to their difference followed by *A. duranensis* biotic and *A. stenosperma* cross and *A. duranensis* biotic and *A. stenosperma* biotic (29 latent features and 28 latent features respectively) as expected therefore confirming the validity of the latent features to resume the multiple phenotypes of the datasets. Interestingly, the similarity of *A. stenosperma* drought and *A. stenosperma* cross was also confirmed by this analysis because only one latent components were found associated with differentiating them.

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 three-fold cross-validation 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 mean absolute error, median absolute error, mean squared error and R2 and the best model selected to minimize the RMSE was retained for each fold and each latent feature (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 6641 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 comparable number of genes for a total of 5764. The overall agreement on the two scores consists of 66% of genes of HIVE selection and 75% of Gini selection (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 a concordance of 50% at the top of the distribution that drops below 10% 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 terms including metabolic processes, response to stress, defence response, oxidation reduction process, although the SHAP selection included more genes than the RFR importance score selection in those categories. Since both selections are very close, we included both in HIVE implementation. Because of the number of genes selected by SHAP score in biologically relevant terms, we decided to continue the analysis with the list of 6641 genes.

Benchmarking tools for gene selection from integrated datasets.

Then we performed the differential expression analysis by comparing each condition with its own control using the popular tool DESeq2. By applying the canonical thresholds of p-value less than 0.05 and absolute value of log2(fold-change) higher than 1, we obtained 12345 genes significantly differentially expressed in at least one condition. 58% of HIVE genes were included in the differential expression selection (Figure 4A). To investigate the difference between those lists, we calculated the coefficient of variation from the top 20 ranked genes to all genes for each selection. From the top 20 to the top 500 genes selected by HIVE show an overall higher coefficient of variation compared to the DE selection, highlighting the efficiency of HIVE to retrieve genes with high variability across conditions compared to DE selection achieved (Figure 4B). Then we inspected whether the two analyses could retrieve specific and/or shared signature i.e. genes deregulated in multiple conditions or in only one specific. While the DE is calculated for each condition separately hence this information is intrinsic with the outcome of the analysis, the genes selected by HIVE were retrieved by using all conditions together. Thus, for HIVE selection, to associate each gene to one or more conditions, we calculated the amplitude of variation

of each gene by calculating the distance of the expression values in each pair of conditions with the last step of the third block of HIVE (see methods, Supplementary Figure 4, Supplementary Table 2). By comparing the results in Figure 4C and Supplementary Table 2 relative to DE and 4D for HIVE, we can observe that 28% (1919 genes) of the genes found by HIVE are deregulated in all conditions while only 1% (128 genes) by DE, as expected HIVE found several common signatures, consistently more than the DE selection thanks to the integrated analysis (Supplementary Figure 4).

The meta-analysis of the biotic, abiotic and cross stress separately performed in the original study yielded a total of 5728 genes, including 1224, 2200 and 3345 genes, respectively (Mota et al., 2021). Importantly, the overlap of the three lists is of only 86 genes. The small overalp of the three 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 1582 genes found by the meta-analysis thus yielding a higher overlap than the three separated meta-analyses (Figure 5A). 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 3A, B). We notice that terms related to metabolic processes and responses to stimulus and stress are in common between the two analyses, altought with a more significant p-value for the HIVE selection. 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 oxidation-reduction processes, cell communication and reproduction.

Finally, we compared the results of HIVE with the MINT tool from the MixOmics R package (Rohart et al., 2017b, 2017a), a classical differential expression analysis for each condition with the tool DESeq2 (Love et al., 2014), and the meta-analysis conducted in the study of Mota et al. on the same dataset (Mota et al., 2021). 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, 43% of them included in the genes selected by HIVE (Figure 5C). Thus, we performed the GO term enrichment analysis to investigate the biological significance associated with these two selections of genes. As shown in Figure 5D, 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 3C 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.

Finally, these results showed the advantages of performing an integrated analysis over the differential expression and meta-analysis by capturing more complex patterns including not similarly regulated genes. Altogether corroborate the reliability of the results found by HIVE and the achievement of better performances compared to the state-of-the-art tools.

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 6641 genes found by our integromics approach searching for genes involved in the main phytohormones signalling pathways, kinases and resistance genes.

93 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). 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 activated or inhibited with very different patters depending on the stress and the plant species (Figure 6A). Nevertheless, genes involved in Jasmonate are mainly activated in *A. stenosperma* biotic as expected due to the known role of this signaling pathway in biotic stress response for plant protection (Huang et al., 2017; Per et al., 2018; Wang et al., 2021). Brassinosteroids are activated/inhibited in a stress and plant species specific manner. The two growth hormones Gibberellins and Cytokines related genes show similar expression profiles including 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. Finally, genes in the Abscisic and Auxin pathways have a very heterogeneous patter of expression with an overall activation in response to drought stress mainly for Abscisic related genes.

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 264 genes coding for kinases. By inspecting their transcriptional profiles, we identified 8 clusters (Figure 6B). Three clusters (A, E, G) group kinases induced in a stress and plant species-specific fashion (biotic - *A. duranensis*, biotic - *A. stenosperma*, abiotic - *A. stenosperma*, respectively). Kinases in cluster D are activated in response to biotic stress in both plant species. The remaining four 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 79 NBS-LRR genes whose expression profiles are reported in Figure 6C. Remarkably, the meta-analysis found 55 NBS-LRR, reinforcing the utility to perform integrated analysis using HIVE. We mainly observe similar patterns as for the kinases: either the genes induced upon specific stress-plant species or heterogeneous patterns showing modulations in multiple conditions that were not previously found by non-integrated analysis methods (Mota et al., 2018). Of note, seven 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). Therefore, we focused to study the transcription factors (TF) – targets relationship. Then we collected TF information from PlantRegMap and we found 69 TF among the HIVE selected genes distributed in 21 families (Figure 7A). We inferred the GNRs by considering those 69 TF as the “regulators” and the rest of the 6287 genes that are not 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 2214 connections. Then we filtered out the pairs not in agreement with the prediction from PlantRegMap, obtaining 22 TFs and127 interactions (Figure 7B). Interestingly, by inspected those pairs we have found two TFs regulating genes involved in phytohormone pathways. The bZIP TF Aradu.C0RFP and the Ethylene-related gene Aradu.MA5U7, with uncorrelated expression profiles (Supplementary Figure 5A) and the ERF TF Aradu.K41I0 with the gene Aradu.1C9UI is also involved in the Jasmonic acid pathway, showing correlated expression profiles (Supplementary Figure 5B). 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. Applying the same threshold, we selected 2919 pairs. By filtering out the pairs not in agreement with PlantRegMap we obtained 23 TFs and 86 interactions (Figure 7C), 32% less than the pairs found with HIVE genes. These results suggest that more real direct targets can be inferred from HIVE genes.
To reconstruct a more general regulome, we first extracted all the TFs among the genes selected by HIVE analysis (Supplementary Table 4). Among the families with more than 10 TF found by our analysis we retrieve families with known pivotal roles in development, plant growth and stress responses (Figure 7D) (Baillo et al., 2019; Huynh-Thu et al., 2010). Then we inferred another network using the same strategy, from the 354 TFs in HIVE selection, we removed the 69 used in the previous network for a

total of 285 TFs as “regulators” and the 6287 genes as “targets”. By selecting the top 2500 pairs, we obtained a big connected community and several isolated interactions including 224 TFs and 1410 targets (Supplementary Figure 6). 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 19 communities as shown in Figure 7E. Then, we quantified the number of TF belonging to the different families in each community (Supplementary Table 5). We first inspected community 7 which shows the highest percentage of phytohormones related genes with respect to the total number of genes in the community (Supplementary Table 6), involved prevalently in the Jasmonic and Abscisic acid pathways (Figure 7F). Interestingly this community also present a high percentage of TFs belonging to the Ethylene Responsive Factors (ERF) family. The expression pattern of these genes and TFs are very similar among each other, all showing high activation in *A. stenosperma* under RKN infection but different patterns in the other conditions. Two TFs (Aradu.LD7BF and Aradu.3P75R) highly repressed in *A. duranensis* in the earliest time point of *M. Arenaria* infection, two hormone gene repressed similar to the previous but one (Aradu.EQ5J6) also in drought stress and not Aradu.WB9GS, while Aradu.E2TII (TF) and Aradu.XW8AZ (hormone gene) strongly repressed in *A. duranenesis* in drought condition (Supplementary Figure 7A). The crosstalk among these three hormone pathways was already suggested in the previous publication related to those data. Despite their enrichment we do not observe a direct interaction between the ERF TFs and the hormones genes involved in jasmonic and abscisic acid pathways suggesting that other intermediary TFs and genes intervene to regulate this complex crosstalk therefore providing novel potential key interactions to further inspect. The second most enriched communities in genes related to phytohormones pathways are community 8 and 2. In community 8 there are also TFs of WRKY and bHLH families and stress-response related genes. Because of this last observation, we screened the community for NBS-LRR genes (Supplementary Table 7) and we found two as direct targets of TFs in the two enriched families: Aradu.180Q6 (WRKY TF) – Aradu.NP598 (NBS-LRR) and Aradu.X5F2F (bHLH TF) – Aradu.HLR71 (NBS-LRR) (Figure 7G). By inspecting the expression profiles Aradu.NP598 (NBS-LRR) shows a different patter compared to the other selected genes in the community and strong repression in *A. stenosperma* after 6 and 9 hours of M. Arenaria infection, while Aradu.HLR71 (NBS-LRR) show mild activation in *A. stenosperma* independently by the stress with similar expression profile to the putative TF regulator (Aradu.X5F2F) (Supplementary Figure 7B). In community 2 the most represented family of TF in this community is the far-red-impaired response 1 (FAR1), which is present with the highest percentage also in community 1 and 14. Intriguingly, community 1 bears the highest percentage of NBS-LRR genes. Therefore, we decided to analyse those three communities altogether (Figure 7H). This family of TFs was shown to be implicated into pod development in *A. hypogaea* (Lu et al., 2022). Furthermore, we have observed high percentage of gibberellin related genes, whose modulation was shown to be important for controlling pod size (Wang et al., 2022). In the three communities we observe NBS-LRR genes either as direct targets or as secondary targets of FAR1 TFs, probably suggesting a possible regulatory role for this family in R-gene mediated Arachis response to stress. Among this last kind of interactions, in community 14 we remarked that two NBS-LRR genes (Aradu.RT9LH and Aradu.G29LA) are targeted

by the same TF (Aradu.VL3IZ) of the WRKY family that is found to be a target of a FAR1 TF (Aradu.1I5WA). The expression profiles of the FAR1 TF and the two NBS-LRR is opposite in all condition, with Aradu.G29LA showing higher differences compared with Aradu.RT9LH, suggesting an opposite activity mediated by the intermediary WRKY TF (Supplementary Figure 7C).

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.

Plant functional validation of NBS-LRR genes found by HIVE

Among the genes found in the previous analysis, we selected two genes from the NBS-LRR gene family to be validate *in planta*. One NBS-LRR gene AsTIR19 (Aradu.G29LA) was found for both the highly resistant species *A. stenosperma* and for the moderate resistant *A. duranensis* in response to the infection of the RKN *M. arenaria*. This gene was found in community 14, among the highest activation, and in *A. duranensis* as middle time point. Another NBS-LRR gene AsHLR (Aradu.HLR71) was mildly activated in *A. stenosperma* during the response to both biotic and abiotic stresses in community 8.

To validate these two genes to their response to nematode in plant models, we obtained different overexpressing lines (OE) of *Arabidopsis thaliana* plants for both genes and submitted to the inoculation of *M. incognita*. For the AsHLR gene, although all six OE lines showed a reduction in the infection in relation to Col-0, only one OE line, the HLR-OE 4.1, showed to be significant. An average reduction of 28% in the number of galls/g of root was observed (Figure 8A), suggesting the potential of this gene, both alone and in association with other defense genes, in the production of transgenic plants with a higher level of resistance to *M. incognita*.

We obtained six different OE lines of *A. thaliana* bearing the AsTIR19 gene from *A. stenosperma*. The overexpression of AsTIR19 in the model plant *A. thaliana* was performed to investigate its potential role in defense against root-knot nematode infestation, specifically *Meloidogyne incognita*. The analysis of the overexpression of the AsTIR19 in Arabidopsis showed a significant reduction in the RKN infection in all six OE lines tested (Figure 8B). In average, a 52.8% reduction in the number of galls/g of root was observed. *A. thaliana* is a model plant widely used in the functional analysis of candidate genes for *Meloidogyne* spp resistance therefore, the results obtained here, combining the identification of candidate genes by HIVE followed by *in planta* validation has the potential to be relevant to other plants affected by the pathogen.

Conclusions

In this paper, we presented HIVE, a novel method to integrate and analyse multi-unpaired transcriptomics 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 and to identify two NBS-LRR with a potential role to confer resistance against RKN infection.

Overall, HIVE is a valuable tool to study multi-transcriptomics 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

Araújo, A.C. de, Fonseca, F.C.D.A., Cotta, M.G., Alves, G.S.C., Miller, R.N.G., 2019. Plant NLR receptor proteins and their potential in the development of durable genetic resistance to biotic stresses. Biotechnology Research and Innovation, Strategies for crop improvement under biotic and abiotic stresses 3, 80–94. https://doi.org/10.1016/j.biori.2020.01.002

Arlot, S., Celisse, A., 2010. A survey of cross-validation procedures for model selection. Statist. Surv. 4. https://doi.org/10.1214/09-SS054

Asai, S., Shirasu, K., 2015. Plant cells under siege: plant immune system versus pathogen effectors. Curr Opin Plant Biol 28, 1–8. https://doi.org/10.1016/j.pbi.2015.08.008

Baggs, E., Dagdas, G., Krasileva, K., 2017. NLR diversity, helpers and integrated domains: making sense of the NLR IDentity. Current Opinion in Plant Biology, 38 Biotic interactions 2017 38, 59–67. https://doi.org/10.1016/j.pbi.2017.04.012

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

Baldi, P., 2012. Autoencoders, Unsupervised Learning, and Deep Architectures, in: Proceedings of ICML Workshop on Unsupervised and Transfer Learning. Presented at the Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings, pp. 37–49.

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.

Clough, S.J., Bent, A.F., 1998. Floral dip: a simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. Plant J 16, 735–743. https://doi.org/10.1046/j.1365-313x.1998.00343.x

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

Efron, B., Tibshirani, R., 1997. Improvements on Cross-Validation: The .632+ Bootstrap Method. Journal of the American Statistical Association 92, 548–560. https://doi.org/10.2307/2965703

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

Guimaraes, P.M., Guimaraes, L.A., Morgante, C.V., Silva, O.B., Araujo, A.C.G., Martins, A.C.Q., Saraiva, M.A.P., Oliveira, T.N., Togawa, R.C., Leal-Bertioli, S.C.M., Bertioli, D.J., Brasileiro, A.C.M., 2015. Root Transcriptome Analysis of Wild Peanut Reveals Candidate Genes for Nematode Resistance. PLoS One 10, e0140937. https://doi.org/10.1371/journal.pone.0140937

Guimaraes, P.M., Quintana, A.C., Mota, A.P.Z., Berbert, P.S., Ferreira, D. da S., de Aguiar, M.N., Pereira, B.M., de Araújo, A.C.G., Brasileiro, A.C.M., 2022. Engineering Resistance against Sclerotinia sclerotiorum Using a Truncated NLR (TNx) and a Defense-Priming Gene. Plants (Basel) 11, 3483. https://doi.org/10.3390/plants11243483

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

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

Huang, H., Liu, B., Liu, L., Song, S., 2017. Jasmonate action in plant growth and development. J Exp Bot 68, 1349–1359. https://doi.org/10.1093/jxb/erw495

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

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. Statistical Learning, in: James, G., Witten, D., Hastie, T., Tibshirani, R. (Eds.), An Introduction to Statistical Learning: With Applications in R, Springer Texts in Statistics. Springer, New York, NY, pp. 15–57. https://doi.org/10.1007/978-1-4614-7138-7_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

Júnior, J.D.A. de S., Coelho, R.R., Lourenço, I.T., Fragoso, R. da R., Viana, A.A.B., Macedo, L.L.P. de, Silva, M.C.M. da, Carneiro, R.M.G., Engler, G., Almeida-Engler, J. de, Grossi-de-Sa, M.F., 2013. Knocking-Down Meloidogyne incognita Proteases by Plant-Delivered dsRNA Has Negative Pleiotropic Effect on Nematode Vigor. PLOS ONE 8, e85364. https://doi.org/10.1371/journal.pone.0085364

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

Leal-Bertioli, S.C.M., Bertioli, D.J., Guimaraes, P.M., Pereira, T.D., Galhardo, I., Silva, J.P., Brasileiro, A.C.M., Oliveira, R.S., Silva, P.I.T., Vadez, V., Araujo, A.C.G., 2012. The effect of tetraploidization of wild Arachis on leaf morphology and other drought-related traits. Environmental and Experimental Botany 84, 17–24.

Leal-Bertioli, S.C.M., Moretzsohn, M.C., Roberts, P.A., Ballén-Taborda, C., Borba, T.C.O., Valdisser, P.A., Vianello, R.P., Araújo, A.C.G., Guimarães, P.M., Bertioli, D.J., 2015. Genetic Mapping of Resistance to Meloidogyne arenaria in Arachis stenosperma: A New Source of Nematode Resistance for Peanut. G3 (Bethesda) 6, 377–390. https://doi.org/10.1534/g3.115.023044

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

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, Q., Liu, Hao, Hong, Y., Liang, X., Li, S., Liu, Haiyan, Li, H., Wang, R., Deng, Q., Jiang, H., Varshney, R.K., Pandey, M.K., Chen, X., 2022. Genome-Wide Identification and Expression of FAR1 Gene Family Provide Insight Into Pod Development in Peanut (Arachis hypogaea). Front. Plant Sci. 13. https://doi.org/10.3389/fpls.2022.893278

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., Oliveira, T.N., Vinson, C.C., Williams, T.C.R., Costa, M.M. do C., Araujo, A.C.G., Danchin, E.G.J., Grossi-de-Sá, M.F., Guimaraes, P.M., Brasileiro, A.C.M., 2019. Contrasting Effects of Wild Arachis Dehydrin Under Abiotic and Biotic Stresses. Front. Plant Sci. 10. https://doi.org/10.3389/fpls.2019.00497
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.
Per, T.S., Khan, M.I.R., Anjum, N.A., Masood, A., Hussain, S.J., Khan, N.A., 2018. Jasmonates in plants under abiotic stresses: Crosstalk with other phytohormones matters. Environmental and Experimental Botany 145, 104–120. https://doi.org/10.1016/j.envexpbot.2017.11.004
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
Proite, K., Carneiro, R., Falcão, R., Gomes, A., Leal-Bertioli, S., Guimarães, P., Bertioli, D., 2008. Post-infection development and histopathology of Meloidogyne arenaria race 1 on Arachis spp. Plant Pathology 57, 974–980. https://doi.org/10.1111/j.1365-3059.2008.01861.x
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

Refaeilzadeh, P., Tang, L., Liu, H., 2009. Cross-Validation, in: LIU, L., ÖZSU, M.T. (Eds.), Encyclopedia of Database Systems. Springer US, Boston, MA, pp. 532–538. https://doi.org/10.1007/978-0-387-39940-9_565

Rezende, D.J., Mohamed, S., Wierstra, D., 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. https://doi.org/10.48550/arXiv.1401.4082

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

Rizwan, M., Haider, S.Z., Bakar, A., Rani, S., Danial, M., Sharma, V., Mubin, M., Serfraz, A., Shahnawaz-ur-Rehman, M., Shakoor, S., Alkahtani, S., Saleem, F., Mamoon-ur-Rehman, H., Serfraz, S., 2023. Evolution of NLR genes in genus Arachis reveals asymmetric expansion of NLRome in wild and domesticated tetraploid species. Sci Rep 13, 9305. https://doi.org/10.1038/s41598-023-36302-1

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

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning Internal Representations by Error Propagation. https://doi.org/10.7551/mitpress/5236.003.0012

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

Song, H., Wang, P., Li, C., Han, S., Zhao, C., Xia, H., Bi, Y., Guo, B., Zhang, X., Wang, X., 2017. Comparative analysis of NBS-LRR genes and their response to Aspergillus flavus in Arachis. PLoS One 12, e0171181. https://doi.org/10.1371/journal.pone.0171181

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

Takken, F.L.W., Goverse, A., 2012. How to build a pathogen detector: structural basis of NB-LRR function. Curr Opin Plant Biol 15, 375–384. https://doi.org/10.1016/j.pbi.2012.05.001

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

Vinson, C.C., Mota, A.P.Z., Oliveira, T.N., Guimaraes, L.A., Leal-Bertioli, S.C.M., Williams, T.C.R., Nepomuceno, A.L., Saraiva, M.A.P., Araujo, A.C.G., Guimaraes, P.M., Brasileiro, A.C.M., 2018. Early responses to dehydration in contrasting wild Arachis species. PLOS ONE 13, e0198191. https://doi.org/10.1371/journal.pone.0198191

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

Wang, Y., Mostafa, S., Zeng, W., Jin, B., 2021. Function and Mechanism of Jasmonic Acid in Plant Responses to Abiotic and Biotic Stresses. Int J Mol Sci 22, 8568. https://doi.org/10.3390/ijms22168568

Wang, Y., Zhang, M., Du, P., Liu, H., Zhang, Z., Xu, J., Qin, L., Huang, B., Zheng, Z., Dong, W., Zhang, X., Han, S., 2022. Transcriptome analysis of pod mutant reveals plant hormones are important regulators in controlling pod size in peanut (Arachis hypogaea L.). PeerJ 10, e12965. https://doi.org/10.7717/peerj.12965

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.

Zhang, H., Zhao, Y., Zhu, J.-K., 2020. Thriving under Stress: How Plants Balance Growth and the Stress Response. Dev Cell 55, 529–543. https://doi.org/10.1016/j.devcel.2020.10.012

Zhang, H., Zhu, J., Gong, Z., Zhu, J.-K., 2022. Abiotic stress responses in plants. Nat Rev Genet 23, 104–119. https://doi.org/10.1038/s41576-021-00413-0

Zhu, J.-K., 2016. Abiotic Stress Signaling and Responses in Plants. Cell 167, 313–324. https://doi.org/10.1016/j.cell.2016.08.029

Figures and Tables

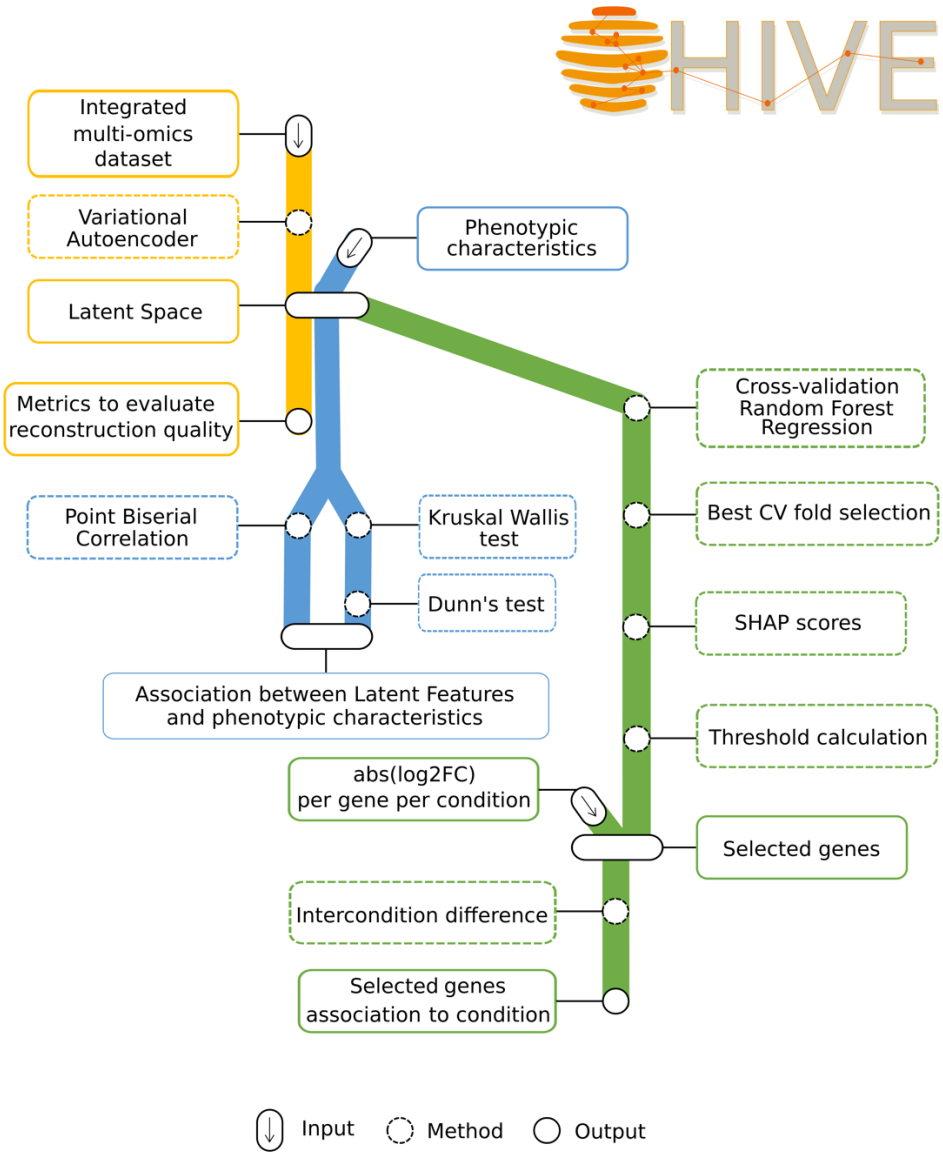

Figure 1. HIVE framework. The pipeline is represented as three underground lines to symbolize the three main blocks composing HIVE. The first block (yellow line) concerns the batch effect correction using a Variational AutoEncoder (VAE) to produce the latent space, that is the common "station" of all 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 four 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, in a cross-validation framework. Afterward, the contribution of each gene to each regression was calculated with the SHAP algorithm. W we included a custom method to set a threshold on the SHAP scores to select only the most contributing genes. The third step returns a list of ranked genes that can be used for further analysis. Lastly, by implementing a novel method based on the amplitude of difference in expression value, each gene is associated to one or more conditions.

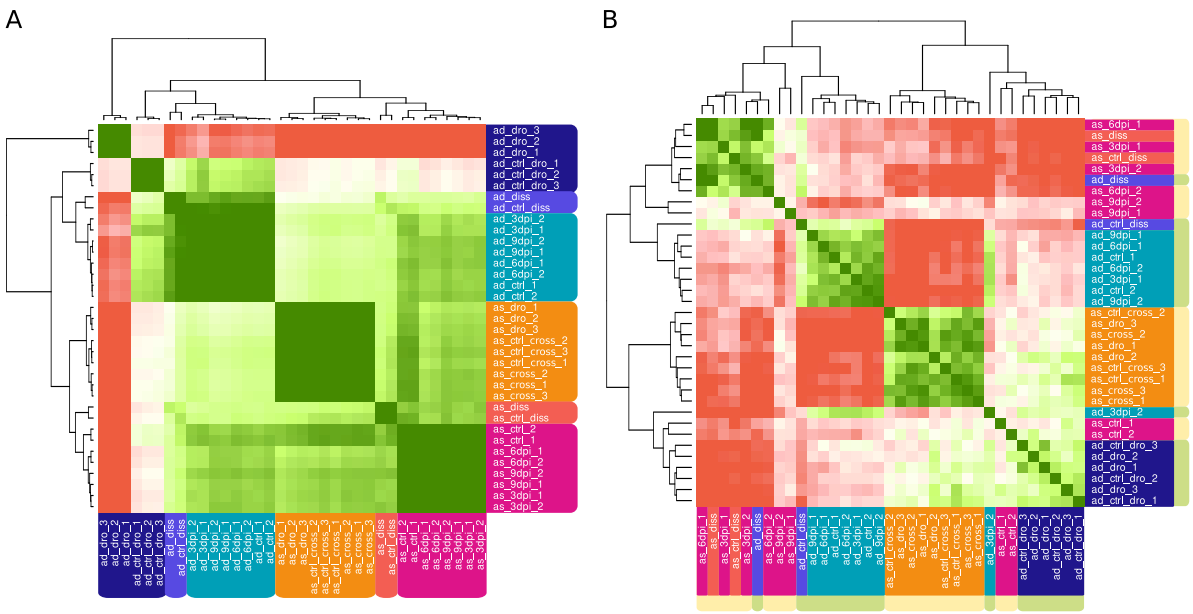

Figure 2. HIVE reduces the batch effects allowing unpaired experiments integrated analysis. Spearman correlation matrix of samples in the original data (A), and after VAE application (B). Hierarchical clustering is applied to show samples groups according to correlation values. Colors of samples according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

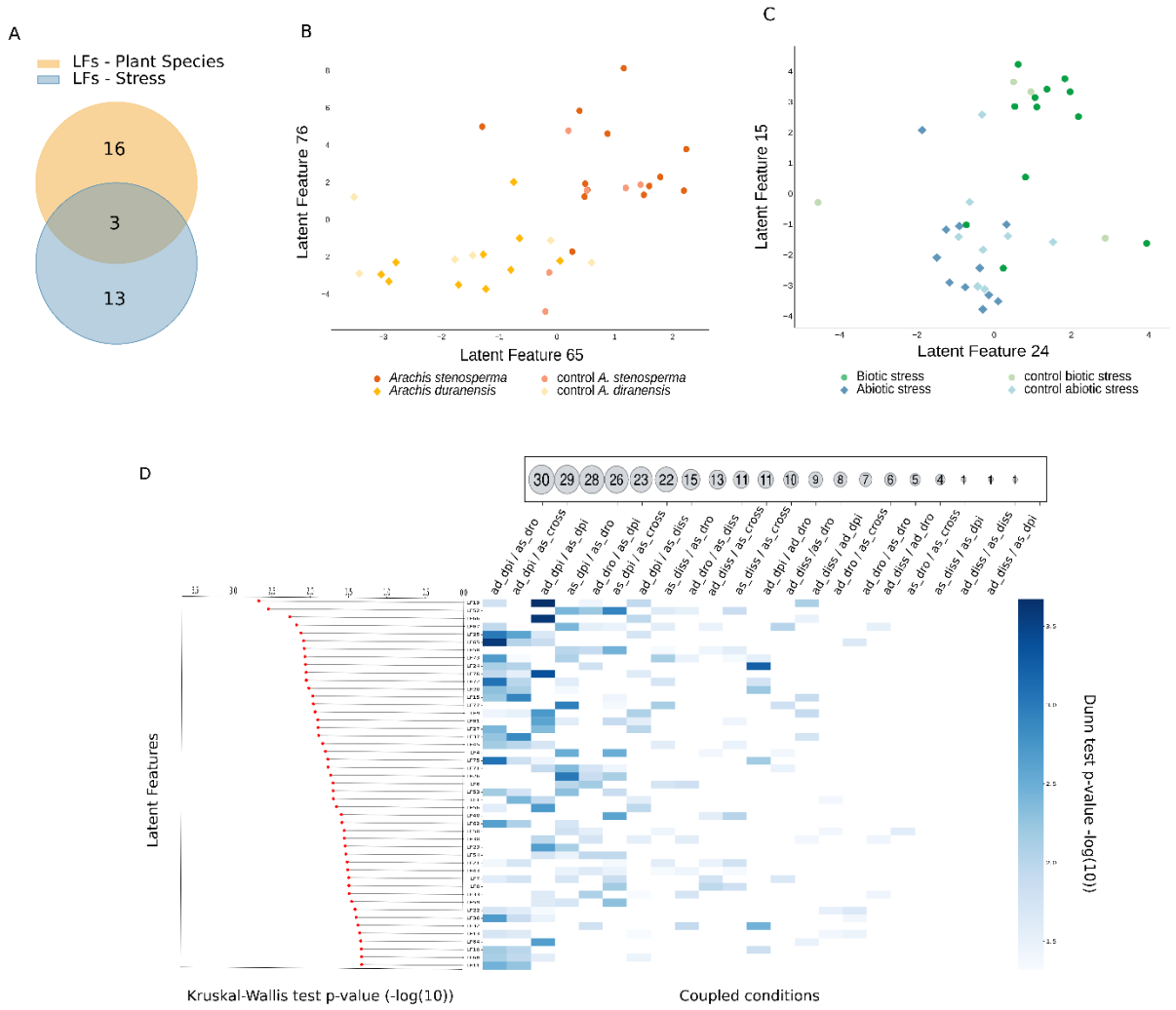

Figure 3. Latent features are associated with phenotypic characteristics.
A) Intersection of latent features selected by the point biserial correlation (p-value < 0.01) as the most correlated either with the type of stress (yellow) or plant species (light blue). Visualization of top two correlated latent features with plant species (B) or with type of stress (C). Points represent samples, the condition is indicated in the correspondent legend. D) Kruskal-Wallis test (on the left) and post-hoc Dunn test (on the right), application to pairs of conditions (couple) for each selected latent feature. On top-right is represented the number of differing latent features accordingly to the Dunn test results.

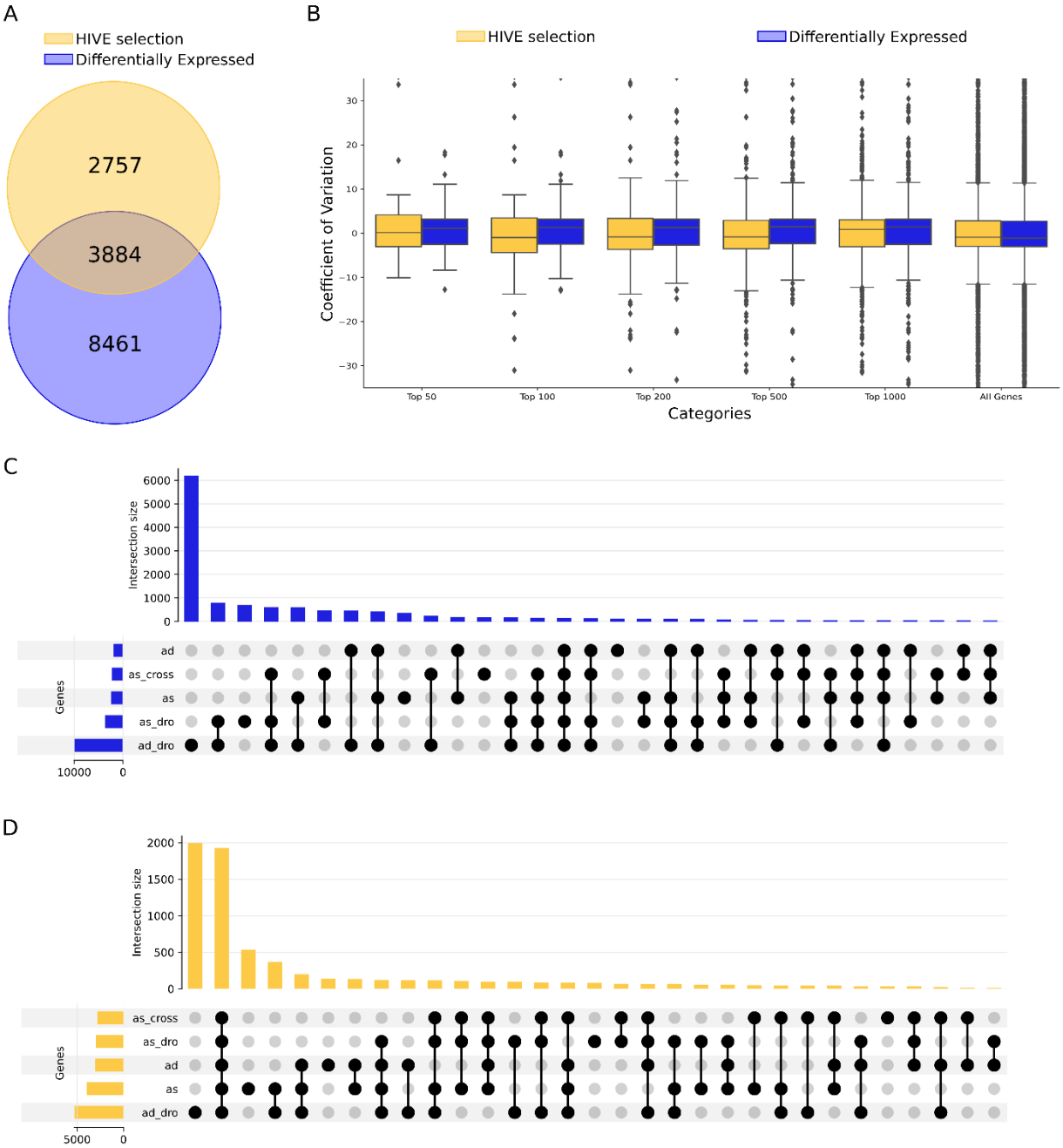

Figure 4: Differential Expression analysis and HIVE selected genes associated to conditions. A) Intersection of HIVE selected genes (yellow) and genes found differentially expressed in at least one condition (blue). B) Coefficient of variation comparing incremental number of genes either selected only by HIVE (yellow) or only via Differential Expression (blue). C) Upset plot showing occurrence and co-occurrence of conditions and the associated number of genes in Differential Expression analysis. D) Shows the same information but for HIVE gene selection. For both upset plots the bar plot on the left represents the number of genes associated to that condition while the top bar plot represents the number of genes found in the occurrence or co-occurrence of the below conditions (single dots and connected dots).

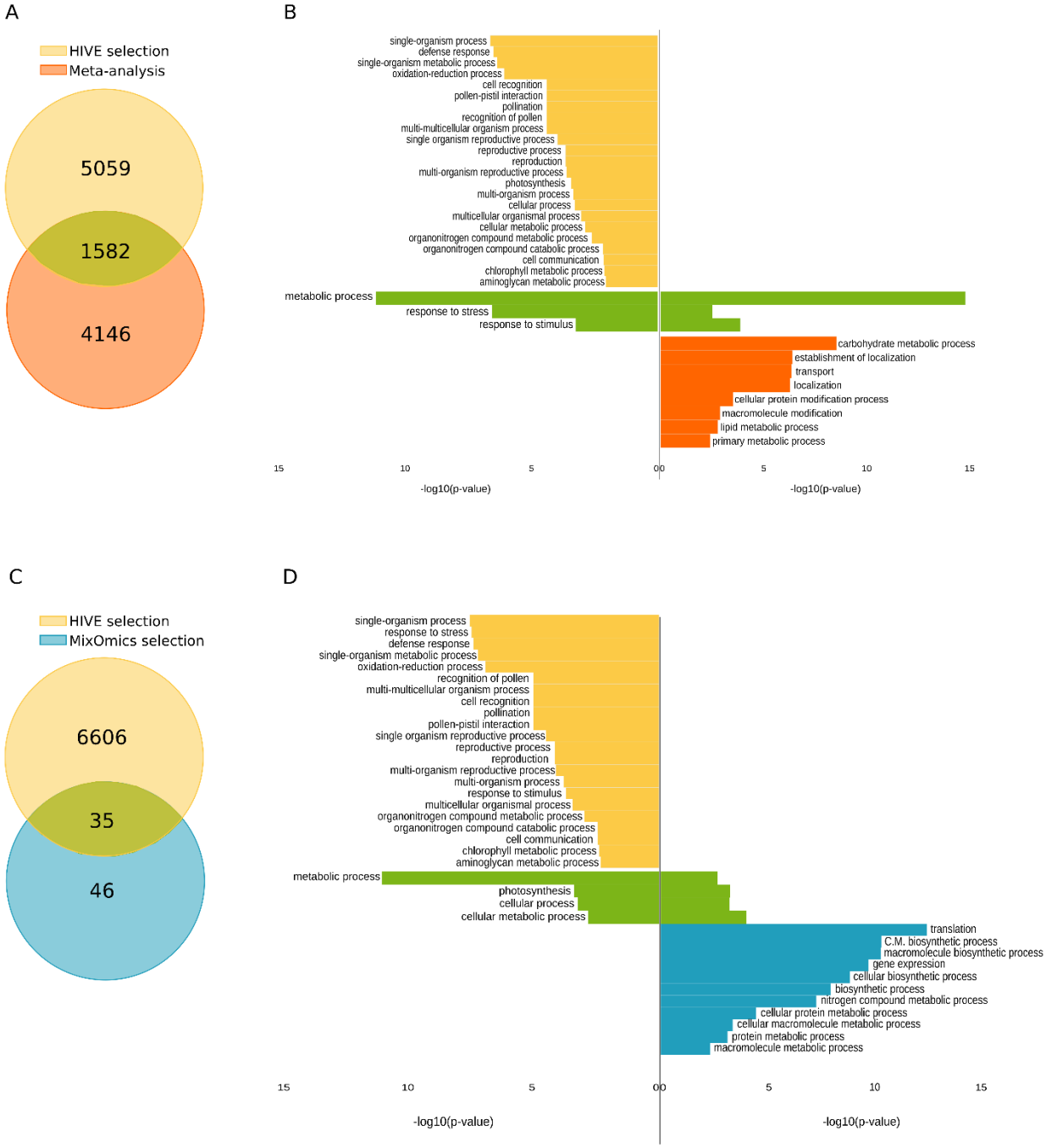

Figure 5: HIVE benchmarking with tools for gene selection from integrated datasets.
A) Intersection of genes selected by HIVE and by DEGs meta-analysis from Mota A., et al, 2021 on the integrated dataset. B) GO-enrichment analysis (p-value <= 0.01), in yellow the GO-terms found enriched only in HIVE, in green GO-terms found enriched in the two selections, and in dark orange the GO-terms found enriched only in the meta-analysis. C) Intersection of gene selected by HIVE and by MixOmics-MINT application on the integrated dataset. D) 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.

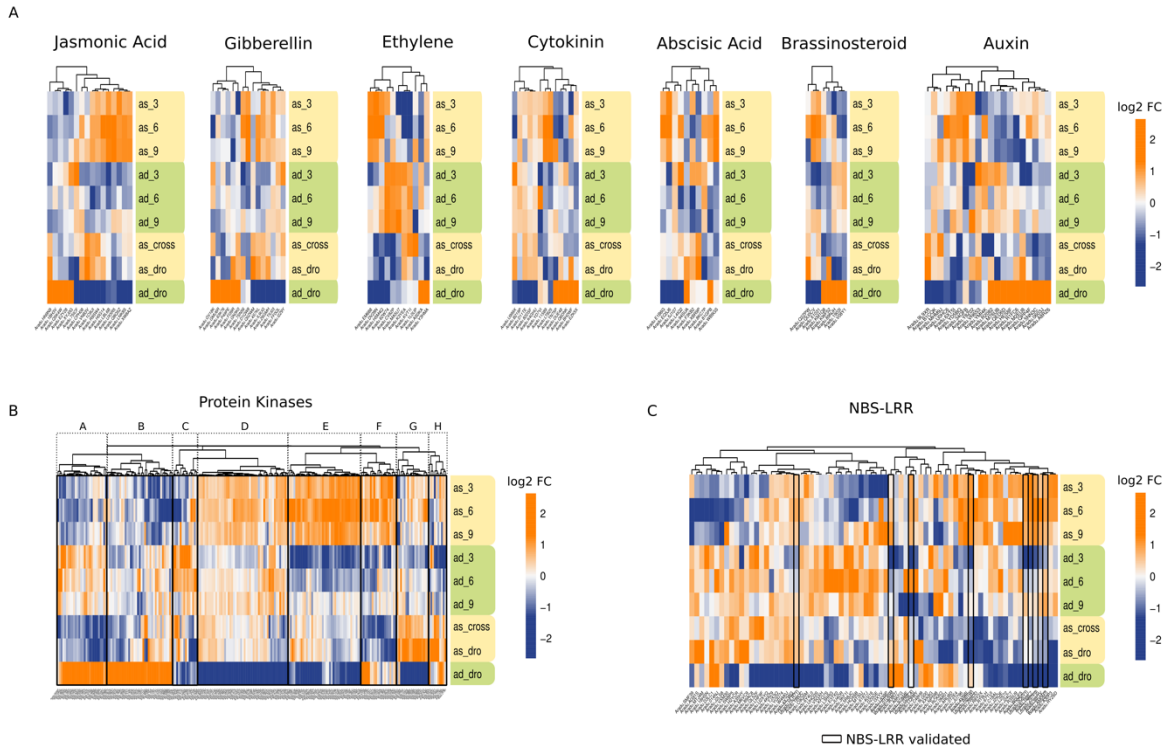

Figure 6. HIVE identify genes involved in the main hormone signalling pathways, kinases and resistance genes in response to stress.

A) Heatmaps of gene expression in each set of annotated phytohormones. B) Heatmap of gene expression for annotated protein kinases found in HIVE selection, black squares represent clusters of gene expression (A-H). C) 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.

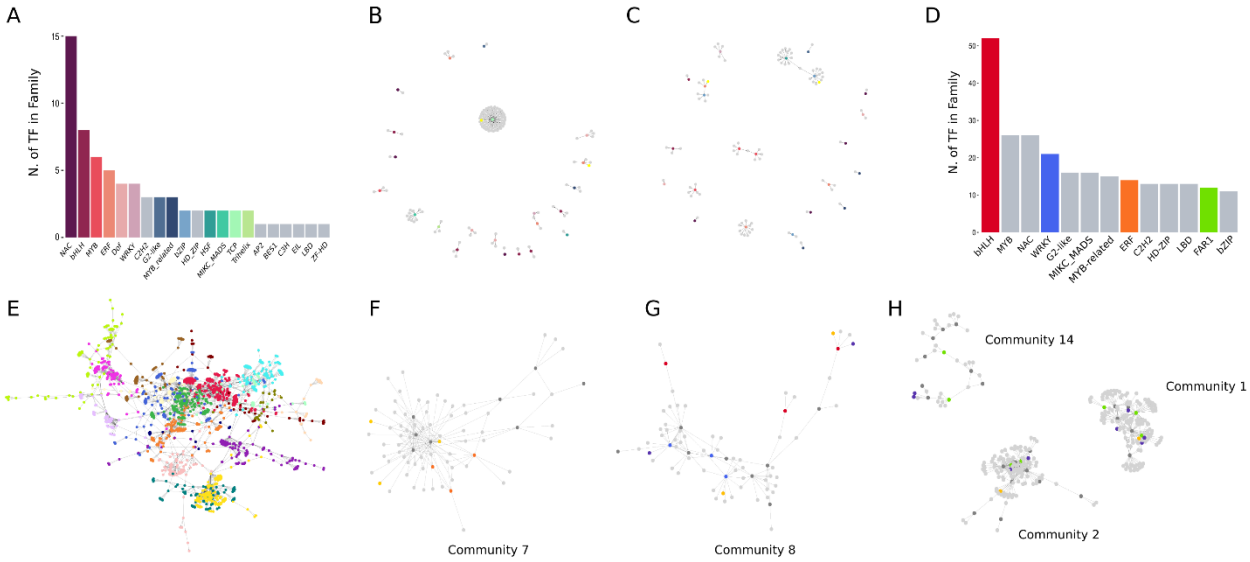

Figure 7. The integrated analysis allowed to infer the complex gene regulatory network induced upon biotic and/or abiotic stress.

A) Number of transcription factors from HIVE selection with predicted binding motif in PlantRegMap. B) GRN inferred by GENIE3 imposing as “regulators”, the TF found by HIVE and PlantRegMap, and as “targets”, genes that are not TFs (after having removed all pairs not in PlantRegMap). Nodes are coloured accordingly to A. C) same as B but with gene selection from the meta-analysis. D) Number of transcription factors in families with at least 10 representatives in HIVE selection. 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 TFs. Colours correspond to the 19 communities found by Leiden algorithm. F) Zoom-in into community n. 7, orange nodes correspond to transcription factors of family ERF. G) Zoom-in into community n. 8, bright blue nodes correspond to transcription factors of family WRKY, and red nodes to transcription factors of family bHLH. H) Zoom-in into communities n. 14, 1, and 2 bright green nodes correspond to transcription factors of family FAR1. In all network figures dark grey nodes represent all the other transcription factors in the network not belonging to the family/families of interest, dark yellow nodes correspond to genes associated with phytohormones, violet nodes indicate NBS-LRR genes and all the light grey nodes are other categories of regulated genes.

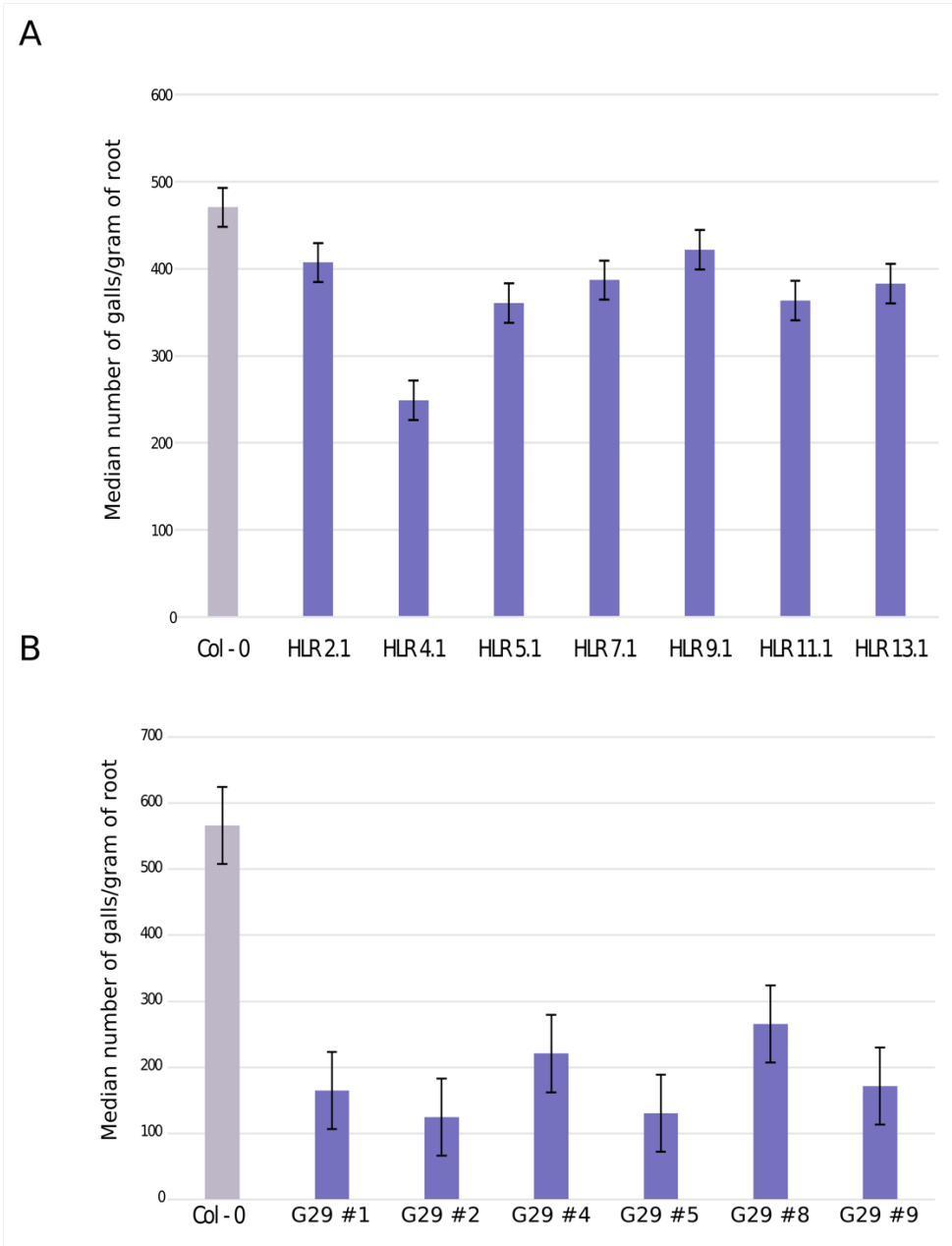

Figure 8. Plant functional validation of NBS-LRR genes found by HIVE.

Number of galls/g of root in *A. thaliana* upon infection with *M. Incognita* overexpressing A) Aradu.HLR71 (AsHLR) or B) Aradu.G29LA (AsTIR19), NBS-LRR genes found by HIVE. Col-0: control; HLR2.1,4.1,5.1,7.1,9.1,11.1,13.1: seven OE overexpressing AsHLR; G29#1,2,4,5,8,9: six OE over-expressing AsTIR19.

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.