A knowledge graph and topological data analysis framework to disentangle the tomato-multi pathogens complex gene regulatory network

Maxime Multari[1], Mathieu Carrière[2], Xavier Amorós-Gabarrón[1], Alexina Damy[1], Sebastian Lobentanzer[3], Julio Saez-Rodriguez[3,4], Stéphanie Jaubert[1], Aurélien Dugourd[3,4*] and Silvia Bottini[1*#]

1 INRAE, Université Côte d'Azur, CNRS, Institut Sophia Agrobiotech, Sophia-Antipolis, France

2 Data Shape Team, Centre Inria d'Université Côte d'Azur, France

3 Heidelberg University, Faculty of Medicine, and Heidelberg University Hospital, Institute for Computational Biomedicine, Heidelberg, Germany

4 European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Cambridgeshire, United Kingdom

*Those authors contributed equally

#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

Global population is rapidly increasing, representing a major challenge for food supply, exacerbated by climate change and environmental degradation. Despite the pivotal role of agriculture, plant health and survival are threatened by various biotic stressors. Although how plants respond to each of these individual stresses is well studied, little is known about how they respond to a combination of many of these bio-aggressors occurring together.

To tackle this question, first, we built TomTom, a knowledge graph gathering molecular interactions from nine publicly available databases, including transcription factors- or microRNAs- targets, protein-protein interactions, and functional terms. Then, we selected transcriptomics data of tomato subjected to six distinct pathogens and performed an integrative analysis. We found 5561 candidate genes involved in the multi-stress response of tomato. To study how the response is orchestrated, we mapped those genes in TomTom and extracted a comprehensive gene regulatory network (GRN) composed of 71 transcription factors (TF) and 1618 target genes. By estimating the TF activity, we identified 43 TFs responding either specifically to one or multiple bio-aggressors. GRN analyses with a topological data analysis approach allowed to identify 18 clusters of TFs with similar properties, yielding four main configurations localized in specific regions of the GRN. Finally, we found four ERF hubs which cooperatively coordinate the tomato response to multiple pathogens.

Our findings allowed to study the complex molecular reprogramming in tomato upon interaction with different biotic agents, providing tools scalable to other questions involving tomato molecular interactions and beyond.

Introduction

Plants as sessile organisms are exposed to diverse stresses, either abiotic or biotic. These stresses can act singly, in combination, sequentially or in a multifactorial manner [1], and their combination can have dramatic consequences for plant survival even when the effect of each stress applied individually is negligible [2] [3]. Plants developed high phenotypic plasticity including rapid responses to environmental factors and adaptations to changing environments [4]. Plant pathogens have a broad spectrum of characteristics spanning from ectoparasites to endoparasites, from biotrophic to necrotrophic, from avirulent to virulent strains, attacking different tissues from leaves to roots, varying time-life cycles, causing several different possible diseases that can lead to different plant responses. The co-occurrence of multiple stress factors further exacerbates the vulnerability of plants to adverse changing environmental conditions [5]. Although multiple infections affecting a single plant or crop are recognized to be common in plant disease epidemics [6], plant diseases are usually studied as the result of the interaction between a plant host and a single pathogen. Therefore, to be able to reduce crops' vulnerabilities, it is necessary to globally characterize how plants modulate their molecular program to respond to these wide-ranging attacks when they occur in a combined manner.

Single stress studies in model systems have provided invaluable mechanistic insights, but more sophisticated experimental designs are required to replicate the native context of a plant. However, such studies can be expected to be more difficult to perform and interpret.

Advances in systems biology have facilitated the understanding of several complex biological processes including plant immune systems, notably using graph-based techniques. Such techniques rely on an abstract description of molecular interactions as nodes' relationships (edges) [7] where nodes can represent molecules such as genes, RNAs, proteins, metabolites, etc. [8]. Network inference has emerged as a broadly used methodology with the development of the so-called omics techniques and their spectacular progresses in the last decade [9]. Among those, transcriptomics is by far the most produced omics type, offering the possibility to study transcriptional regulation, an important factor in controlling tolerance and resistance to stresses. Therefore, Gene Regulatory Networks (GRNs) composed of nodes which represent transcription factors (TFs) or target genes, and edges representing the regulatory connections between regulators (TFs) and targets (genes), have gained a lot of attention lately [10]. GRN inference is an evergreen unsolved problem in the field of systems biology. Two main approaches are commonly employed, either using measured data, from transcriptomics and/or chromatin accessibility experiments for instance, to perform a data-driven inference [11] or using prior knowledge, such as known molecular interactions from publicly available databases [12], to build a network skeleton. While reconstructing interactions from observational data is a critical need for investigating natural biological networks, the dimensionality is usually very high. On the other side, summarizing knowledge of an organism is challenging, most of the time relevant information is found across multiple databases, often not interoperable among each other.

To tackle these challenges, we focus on tomato (*Solanum lycopersicum*) not only because of its agroeconomic interest, but mainly because it is susceptible to more than 200 diseases caused by several different pathogens [13], [14] [15] [16], offering a biological model to study multifactorial stress

integration. First, we developed TomTom, a knowledge graph for molecular interactions in tomato. Then, we retrieved transcriptomics data of tomato subjected to six different pathogens from publicly available transcriptomic data. We extracted the molecular signatures which are either specifically responding to one stress or to multiple ones by performing an integrative analysis with the HIVE tool [17]. We used TomTom to generate the GRN including only the molecular signatures extracted from our integrative approach. To study this complex network, we set up a novel approach based on the estimation of transcription factor activity [18] and the study of the topology of the network using the topological data analysis framework [19].

Materials and Methods

TomTom: a knowledge graph to study tomato molecular interactions

Several databases were used to construct the knowledge graph to cover multiple molecular interactions. Supplementary Table 1 summarizes all the input databases. The procedures used to include them in the knowledge graph are explained in the following paragraphs.

Genome and microRNA backbone

The genome of *Solanum lycopersicum* was downloaded from Sol Genomics Network [20], using the latest version available: SL4.1. The GFF file, describing genomic features, was used as a backbone for the knowledge graph to construct and link the genomics entities (genes - transcripts - proteins), using unique identification. The identification (ID) chosen was the Organized Locus Name (OLN) of each gene (eg: Solyc01g010203). This ID is constructed thanks to the species name (*SOlanum LYCopersicum*) followed by the chromosome number (01). A letter (g) specifies the type of the entity, here a gene, followed by a unique 6-number identifier (010203). We extracted the genes from the GFF file and followed the central dogma of biology to add transcripts and proteins. Thus, for one gene we have one transcript and one protein, as alternative splicing was absent in the reference genome. We then identify those entities such as “gene:OLN”, “transcript:OLN” and “protein:OLN”. The previous features build the backbone of the knowledge graph by linking genes to transcripts and transcripts to proteins.

Then, we also constructed the microRNA backbone (microRNAs and precursors) using miRbase [21]. We downloaded miRBase and parsed it to retrieve the microRNA and precursor information for *S. lycopersicum* species. We then linked each precursor to their corresponding mature microRNA.

Integration rules for other databases

We had to define specific rules to integrate other databases that cover multiple interactions.

Since all these databases are based on different genome versions, to have consistent interactions when considering different genome versions, a filtering process was mandatory to avoid adding incorrect information to the graph. Thus, we checked every added entity against the backbone to avoid having ones not present in the genome used (SL4.1). All identifiers present in the input databases must match the identifiers in the backbone, and if one of the identifiers is not recognized in the backbone, the corresponding interaction is removed.

The filtering rules have been applied to all further databases.

Transcription Factor - Target Interaction

The transcription factors (TFs) and their associated family, were downloaded from the PlantTFDB [22] web interface and added to the graph. To connect TFs and target genes, we downloaded the gene regulatory network (GRN) of *S. lycopersicum* from the PlantRegMap [23] web interface. The GRN, composed of TF-Target interaction, was added to the graph.

MicroRNA - Target Interaction

Three microRNA - Target interaction (MTI) databases: TarDB [24], DPMIND [25], and PNRD [26] were added by downloading the interactions from their respective web interfaces.

PNRD also included precursors and not only mature microRNA IDs. We manually curated those cases by checking whether they were present in miRBase using their sequence and renamed those microRNAs accordingly or filtered them otherwise.

The three databases are distinctly represented in the graph, giving database-specific information and the ability to perform specific queries that include one or more MTI databases.

Protein - Protein Interaction

Protein-protein interactions (PPI) were retrieved from the STRING database [27]. STRING uses its identifier, derived from UniProt, instead of the OLN identifier. To associate an OLN, we used the UniProt mapping script that maps the protein's IDs to the associated reference genome. However, the only available version of the *Solanum lycopersicum* genome is the SL3.1 from Ensembl Plants. The drawback of this approach is that by mapping on a different genome version than the one used for the KG, there is a loss of proteins. After converting the protein IDs to the OLN format, the PPI network was added to the knowledge graph. We included interactions with a STRING combined score greater than 0.8 to maintain high quality interactions. The combined score can be adjusted by the user if more PPIs are to be added.

Planteome Terms

To gain more insights into the gene functions, we downloaded a custom dataset from the Planteome web page [28], which included terms from Gene Ontology (GO) or Plant Ontology (PO) as well as the term descriptions. Those terms were then linked to genes accordingly, using a custom procedure to match genes to terms by reformatting the original downloaded file.

KEGG Pathways

KEGG pathways for *S. lycopersicum* were retrieved through KEGG API [29]. Precisely three requests, pathway IDs and names, genes involved in pathways, and mapping from KEGG gene IDs to UniProt IDs. UniProt IDs were then converted to OLN using the same script as before, provided by UniProt. Once converted, links between genes and KEGG pathways were created.

Knowledge graph construction

To construct the knowledge graph, we used the BioCypher (v0.5.40) framework [30]. This allowed us to systematize Python modules for integrating different files or databases into the graph and add complementary information, including node IDs, node properties, edge IDs, and edge properties. The semantics used in the graph were defined in a YAML file that maps graph entities and relationships to concrete concepts from an ontology.

For each database, a corresponding Python module (“adapter”) was created to consider the peculiarities of each database in terms of the information and type of interactions.

The graph was then constructed by running a pipeline orchestration script that controlled each step of the construction. The knowledge graph was finally deployed into a Neo4j graph database instance via a Docker Compose workflow, created from BioCypher's template.

Integrated multi-transcriptomics dataset composition and HIVE analysis

The integrated data comes from five different publicly available studies and is described in Supplementary Table 2. For each project, we selected the corresponding sequencing data from the tomato interaction with pathogens or controls. A summary of all the samples used is described in Supplementary Table 3. In total, 83 samples were included in the dataset, representing seven *S.lycopersicum*-pathogen interactions. Each sample was preprocessed using an internal pipeline using MultiQC [31] then trimmed using fastp [32] with parameters: -5 -3 -M 28 -l α -y. The α is set to be the read length - 2 for each sample. The sequences were then mapped to the *S. lycopersicum* 4.1 genome using STAR [33] with the following parameters: for the indexing: --runMode genomeGenerate – genomeSAindexNbases 10; and the mapping: --alignEndsType EndToEnd –twopassMode Basic. We further counted the reads on the alignment using featureCounts [34] with default parameters. We obtained a count matrix for each dataset. We merged all the count matrices to obtain one multi-transcriptomics dataset.

We applied HIVE [17] to the multi-transcriptomics dataset with default parameters except for the number of bins, which was set to 138 in the "shap_bin_selection_f()" function. HIVE outputted a list of 5561 genes.

Wald statistic calculation

Differential expression analysis was conducted using DESeq2 [35] with the default parameters. Each dataset was analyzed independently, and we compared the controls (uninfected) to the infected samples. The Wald statistics, calculated by dividing the log2FC by its error (lfcSE) for each gene, is used for further analysis.

Gene Regulatory Network inference

To infer the GRN from the HIVE results list, we queried TomTom to have both TF and targeted genes coming from the list. We queried TomTom as follows, where $gene represents the gene list:

MATCH (tf:TranscriptionFactor)-[reg:TFRegulation]->(target:Gene)

WHERE tf.name IN $gene AND target.name IN $gene

RETURN tf.name, target.name, reg.evidence;

We removed the low-confidence "motif" evidence only interactions.

Transcription factor and pathway activity estimation

The transcription factor (TF) activity was computed using the Univariate Linear Model (ULM) from the decoupleR [18] python package on the GRN.

For each sample, ULM fits a linear model for each TF with the stats of targeted genes being the response variable and the TF interaction classes (activating or repressing TF) being the explanatory variable. The t-value and the p-value of the fitted model are computed. The predicted activity of the TF is based on the t-value, representing an active TF (positive value) or an inactive one (negative value), also referred to as TF activity score. In other words, the predicted activity reflects if a TF follows the expression pattern of its targets (active TF) or not (inactive TF).

Since we did not have information on TF classes, all TFs in the GRN are considered activators. Therefore, the activity is dependent on this activating class assumption for all the regulator TFs in the GRN. By applying this model, we retrieved TFs with a p-value inferior to 0.05 in at least one condition.

The pathway activity was computed using the Multivariate Linear Model (MLM) from decoupleR [18]. The network used in this case links genes from the GRN to KEGG pathways, as previously added in the KG. As for the GRN, we missed the interaction classes between genes and pathways; thus, every interaction is considered positive. Like before, the stats of each gene in pathways are used to compute the activities. We then retrieved pathways with a p-value inferior to 0.05 in at least one condition.

Topological data analysis

In order to study large networks of TFs, we use the Mapper algorithm [36] from topological data analysis in combination with the mode-based clustering algorithm ToMATo [37].

Mapper algorithm. This algorithm is a data visualization method that produces a graph representing the dataset, and whose topology (connected components, branches, loops) matches that of the original data. It can be obtained from a continuous function, called filter $f: X \to R^d$, where X is the dataset. Let I be a cover of the image of the filter $im(f)$ with overlapping hypercubes: $im(f) \subseteq \bigcup_{H \in I} H$. This cover can be pulled back to X using f^{-1}; this induces the pullback cover $U = \{f^{-1}(H) \subseteq X \mid H \in I\}$. The pullback cover is then refined by separating every cover element of U into its connected components with a clustering algorithm, this is called the connected pullback cover:

$$V = \{S \subseteq X \mid \exists H \in I \; s.t. \; S \; is \; a \; connected \; component \; of \; f^{-1}(H)\}.$$

Finally, the Mapper graph is defined as the nerve of V, that is, the graph whose nodes are in bijection with the elements of V (each Mapper node thus represents a group of similar data points), and whose edges are drawn between two nodes as soon as the corresponding cover elements in V have at least one data point in common (Supplementary figure 1A).

We applied the Mapper algorithm on the GRN using the first two eigenvectors of the GRN normalized Laplacian as filter (i.e., the two eigenvectors with largest eigenvalues), and a cover of 25 hypercubes with 30% overlap, evenly sampled along the filter components. The clustering algorithm used for producing the connected pullback cover was defined as the connected components of the GRN restricted to the preimages of the hypercubes under the filter. Finally the Mapper nodes were colored

using the average activities of the significant TFs in response to each pathogen, and the Mapper edges were thickened according to the number of target genes shared by the TFs in the corresponding Mapper nodes. We also removed 2 nodes (10 and 12) from the final graph because they were isolated and without any significant TFs. This allowed us to produce compact and geometric representations of the GRN.

ToMATo clustering. In order to group TFs together, we applied the Topological Mode Analysis Tool (ToMATo) clustering algorithm on the Mapper graph. ToMATo is a hierarchical clustering algorithm which uses a user-defined scalar-valued filter $f: X \rightarrow R$ in order to build the cluster hierarchy. Specifically, this algorithm first requires building a graph G on top of the dataset (e.g., a neighborhood graph), and then uses this graph to approximate the gradient $\nabla f(v)$ at every graph node $v \in G$. Finally, the clusters are computed by capturing all local maxima of f using gradient flows (each maximum being seen as a new cluster), and a hierarchy is built by looking at the saddle points of f and connecting all the clusters associated with the saddle to the one with highest prominence. This method allows us to define hierarchies of clusters for any arbitrary function and is known to be robust to noise in the data.

Hence, we applied the ToMATo clustering algorithm using the Mapper graph defined above as the graph G, and the TF activities as well as edge thickness as filters, after some pre-processing has been applied, leading to 8 different clusterings. The pre-processing used was obtained by applying the following mixtures of Gaussians to the TF activities and the edge thickness, showed in Supplementary figure 1B.

Indeed, this pre-processing on the TF activities (resp. edge thickness) allowed us to characterize TFs with low, high, and null activities in average (resp. TFs with no common target genes and many common target genes) as local maxima of the new filter. Finally, the eight clusterings were combined together using Cartesian products, and simplified in the following way: if the TF activities of a Mapper node are significant in response to at least two different pathogens in the Cartesian-product cluster corresponding to the Mapper node, then this node is marked as “multiple response”, otherwise it is marked as “specific response”, and if the edge thickness cluster of the Mapper node is associated to a maxima that is below the saddle point of the Gaussian mixture, it is marked as “exclusive targets”, otherwise it is marked as “shared targets”. We further identify groups of nodes having the same label, allowing us to create four subparts of the original graph.

Hub identification

We performed hub identification in each combined group found by the ToMATo algorithm by computing two metrics: degree and betweenness. We defined hub as a node that have the maximum values for one or both metrics. If multiple nodes have the same value, we considered all of them to be hubs. We retrieved only the hubs with significant TF activity.

Network visualization

We displayed complex networks using Cytoscape [38] (v3.10.2) and yFiles layout algorithms [39] to better visualize it.

Results

TomTom: a comprehensive knowledge graph summarizing molecular interactions in tomato

To build the knowledge graph TomTom, we collected different types of molecular interactions from multiple databases (Supplementary Table 1). The construction of the graph relied on a backbone of 34688 unique genes present in the SL4.1 tomato genome. Interactions were then added following the gene-transcript-protein schema (Figure 1A). For the microRNA backbone, from miRbase [21] 112 precursors and 147 microRNAs were included in the graph. The microRNA-target interactions, were retrieved from DPMIND [25] (159 interactions), TarDB [24] (777 interactions) and PNRD [26] (337 connections). 1717 transcription factors (TFs) were retrieved from PlantTFDB [22] and by using PlantRegMap [23], we included 241 388 TF-gene interactions in the TomTom. Then, 108 274 protein-protein interactions were recovered from STRING [27], corresponding to 6905 proteins. Annotation terms were selected from Planteome [28], (8949 terms) consisting of 794 965 links between genes and terms, and 154 pathways and 14314 links from KEGG [29].

In total, TomTom encompassed 113 415 nodes and 2 864 036 relationships from nine distinct databases (Figure 1B and C). The automatic pipeline can be rerun at any time to provide an updated graph. The graph can be queried using the Neo4j graph database query language, Cypher.

Overall, TomTom provides a global summary of *Solanum lycopersicum* from molecular interactions to pathway regulation.

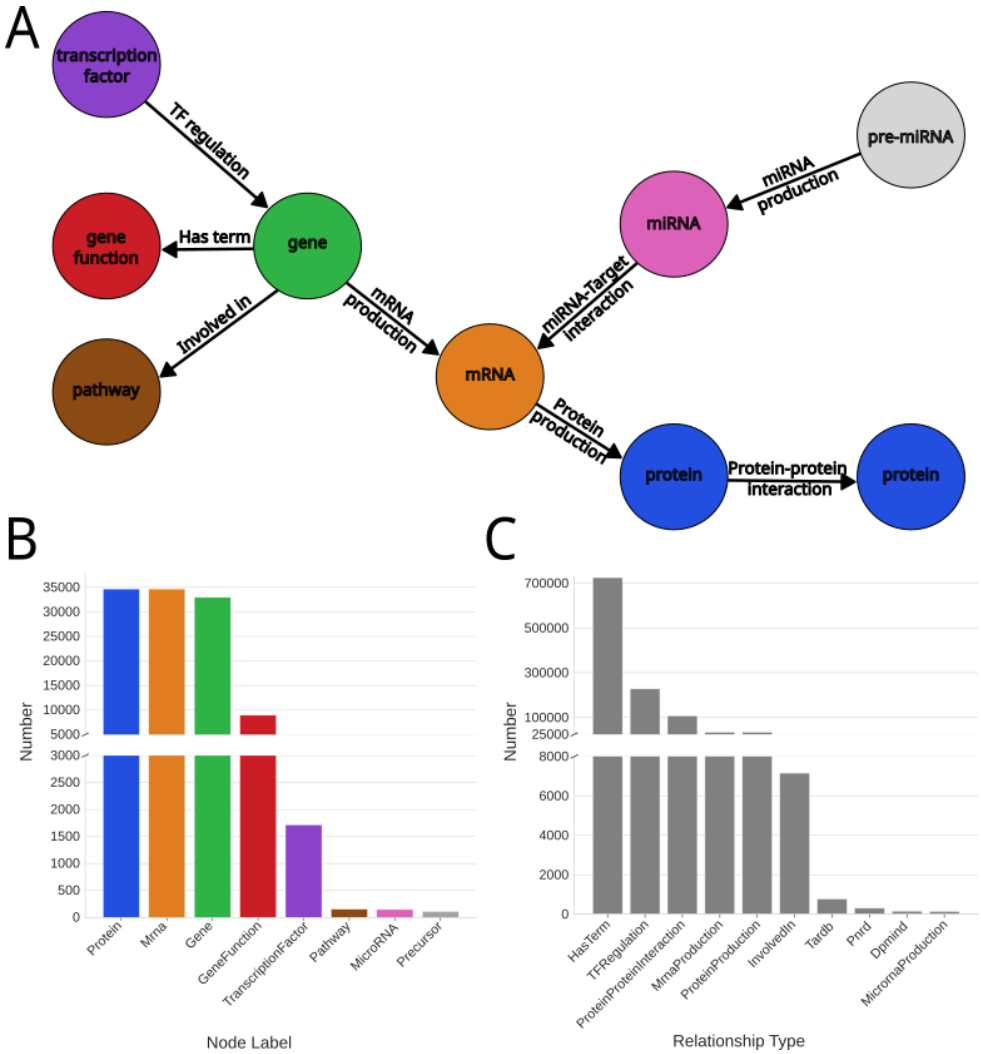

Figure 1. TomTom schema. A) Type of node entities and their relationships in TomTom. Each node represents an entity, and the nodes relate to associated relationships. TF regulation: Transcription Factor-target interaction. B) Number of nodes by molecule type or function. C) Number of edges by relationship type.

The multi-stress transcriptomic data integration and TomTom allowed to infer a global GRN, modeling tomato response to single and multiple attacks.

To study the tomato response to multiple biotic stresses, we collected transcriptomics data from five publicly available studies, each challenging tomato plants with one distinct pathogen (Supplementary Table 2). The final dataset comprises the tomato transcriptome before and upon infection with six pathogens infecting two plant tissues, either leaf or root. The leaf pathogens include the fungi *Botrytis cinerea* and *Cladosporium fulvum*, the oomycete *Phytophthora infestans*. The root pathogens include the root knot nematode *Meloidogyne incognita* (two time points: 7 and 14 days post infection, dpi) and a mild and a severe strain of potato spindle tuber viroid (PSTVd). The choice of those pathogens was made to include the highest variety of species, lifestyle and tissue attacked. The integrated analysis of

those data was performed using HIVE, a novel tool to jointly model single stress omics data, as explained in [17].

To study how regulatory processes are modulated specifically or to respond to multiple pathogens, we built a gene regulatory network (GRN). Briefly, we queried TomTom to extract all relationships corresponding to “TF-target” involving only genes found in the HIVE results, allowing us to study the integrated omics GRN. We excluded the interactions supported only by motif screening analysis. The final GRN consists of 1689 nodes including 71 TFs and 3958 edges representing direct relationships from TFs to targets, although TF can also be targets (Figure 2A, Supplementary Figure 2). To inspect whether the inferred GRN can be used to study tomato response to pathogens infections, we performed a KEGG pathways activity analysis. Briefly, we estimated the activity of KEGG pathways by using a multivariate linear model on the t-values of the genes in each list belonging to the examined pathway; the t-value measures the size of the difference relative to the variation in sample data (see methods). The assumption is that genes belonging to an active pathway should have a coherent pattern of expression [40]. Interestingly, by looking at the overall pattern of pathway activities, we can observe that the more similar are the characteristics of the bio-aggressors, the more similar the tomato respond to pathogen attacks. When focusing only on pathway activities significant in response to at least one pest, although the overall clustering slightly changes, the main patterns are conserved. The response to *M. incognita* is apart and looks overall opposite to the response to the other pathogens, which correspond to the very different mode of infection of this pathogen compared to the others included in this study (Figure 2B, Supplementary Fig 3). Consistently, despite the two viruses’ strains attack the same tissue as *M. incognita*, the roots, we observe that overall, the tomato response is more similar to the response to the leaf pathogens. Finally, the response to the three leaf pathogens and especially to the two fungi is overall very similar. Importantly, among the significantly active pathways, we found: plant-pathogen interaction, MAPK signaling, oxidative phosphorylation, plant hormone signal transduction (Figure 2B). We can also observe some specificities in those pathway activation either in the type of activity or in the specificity to the response to pathogens and none of the studied pathway from KEGG is active in response to all pathogens (Figure 2C). From Figure 2D we observe that leaf pathogens trigger the activation of the highest amount of pathways with a prevalence of negative activity. On the other side, tomato response to both virus strains seems to activate the lowest number of pathways. Finally, the response to *M. incognita* seems to be very different compared to the other pathogens, which is coherent with previous results. To better quantify the number of pathways activated in response to multiple pathogens and their activity levels and significances, we represented those results using a network representation (Figure 2E).

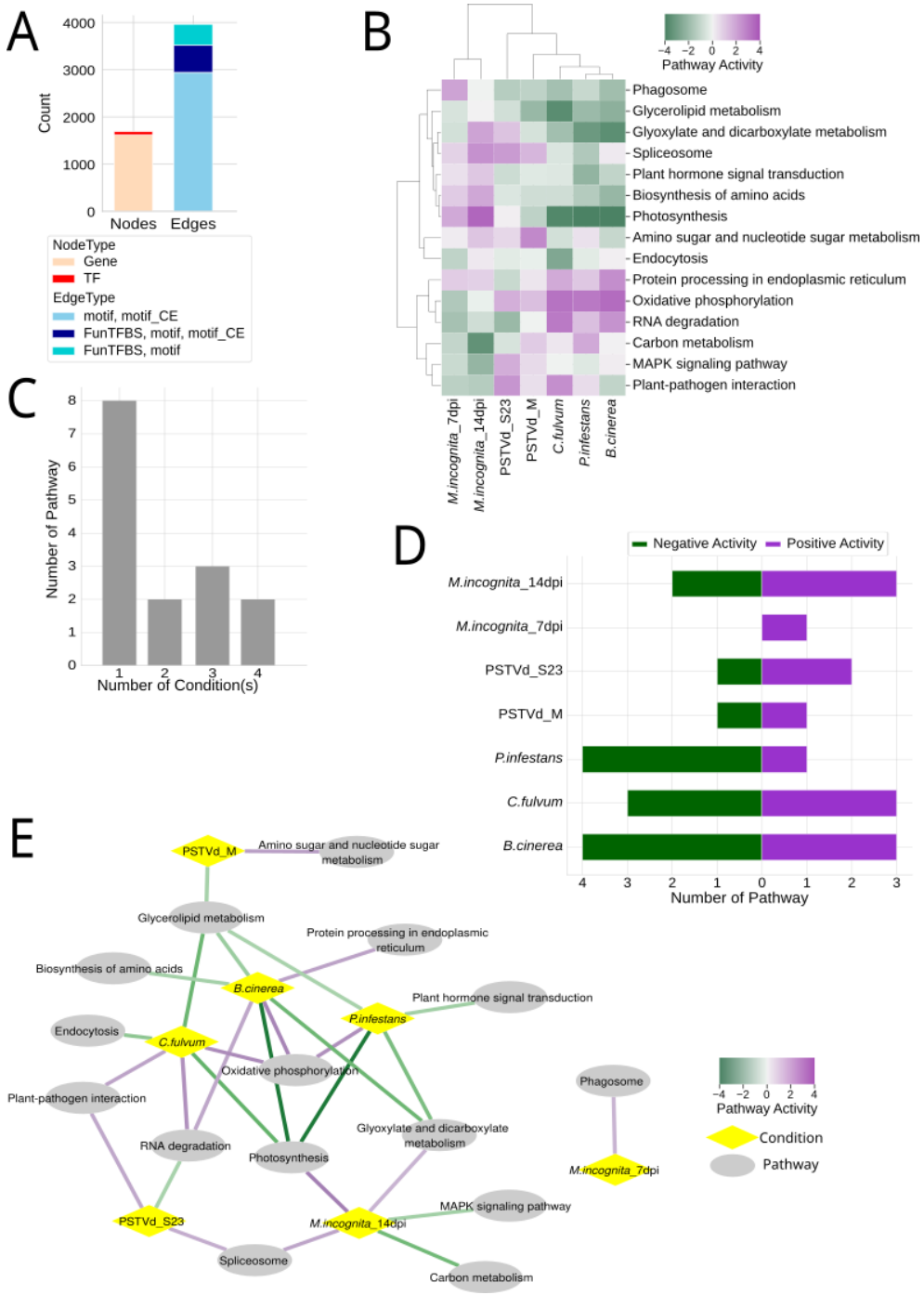

Figure 2. Characteristics of the GRN inferred by TomTom and the transcriptomic integrative analysis of tomato subjected to multi-stress. A) Number of nodes or edges by type in the GRN. B) Heatmap showing the levels of pathway activity in each condition only for significant results in at least one condition. C) Barplot reporting the number of pathways with significant activity in one or more combination of conditions. D) Barplot showing the number of pathways with significant positive (on the right) or negative (on the left) activity according to conditions. E) Network representation of the pathway activity analysis. Nodes are either pathways with significant activity or conditions, edges represent significant activity in the connected condition, the activity value is reported as color scale of the edges.

We observe that the photosynthesis and the glycerolipid metabolism are strongly negatively regulated upon interaction with the leaf pathogens. The oxidative phosphorylation pathway is significantly positively active in response to the three leaf pathogens. The endocytosis is negatively regulated only upon *C. fulvum* attack, the phagosome positively upon *M. incognita* infection and the protein processing in the endoplasmic reticulum is positively active in response to *B. cinerea*. Overall, we observe that the three leaf pathogens trigger the activation of shared pathways in tomato, suggesting that there is a core common response to those pathogens activated in tomato at this tissue level. Interestingly, we found four pathways activated in response to both pathogens attacking the leaves or the roots, although with different sign of activity in three cases. This supports the hypothesis that different pathogens can trigger the same pathways but with different modulation.

These results corroborate the reliability of the GRN built from the signatures found by the integrative analysis and relationship extracted by TomTom, since we observe that the pathway activity reflects the expected response to the different pathogens, although the model is unaware of the similarities among pathogens' characteristics or infected tissue.

Estimating the transcription factor activity identified 43 TFs orchestrating specific and multi-factorial tomato response to multiple biotic stressors.

TFs included in the GRN belong to 23 different families. Among the families with the highest number of TFs in the GRN we found bZIP and ERF (8 TFs each), MYB and NAC (6 TFs each), C2H2 with 5 TFs and WRKY with 4 TFs (Figure 3A). Since the expression of TFs is not always linked to their regulatory activity, we estimate the putative activity of all TFs in each condition based on the expression of their regulated targets. Similarly to the pathway activity estimation explained in the previous paragraph, by using a univariate linear model, we leverage the t-value of the expression level in each condition of all the targets of each TF to estimate whether the TF is significantly active in its target regulation and whether the activity is positive or negative [18]. This analysis yields 43 out of 68 TFs with a significant activity in response to at least one pathogen infection. Of note, three TFs have less than 5 targets therefore we could not estimate the activity. First, we calculated the correlation between the estimated activity and the t-value of expression of those 43 TFs in the different conditions. As shown in Supplementary Figure 4A, the correlation spans from -0.01 to 0.25, therefore validating our hypothesis that the TF activity is not correlated to the t-value. Furthermore, we found that there is no relationship between the number of targets and the activity of TF, therefore there is no detectable size effect in the TF activity estimation (Supplementary Figure 4B).

The TFs with a significant activity belong to 14 families, including all bZIP, WRKY and 7 out of 8 of the ERF, while 9 families present in the GRN do not play a significant regulatory role in response to the studied biotic stresses (Figure 3A).

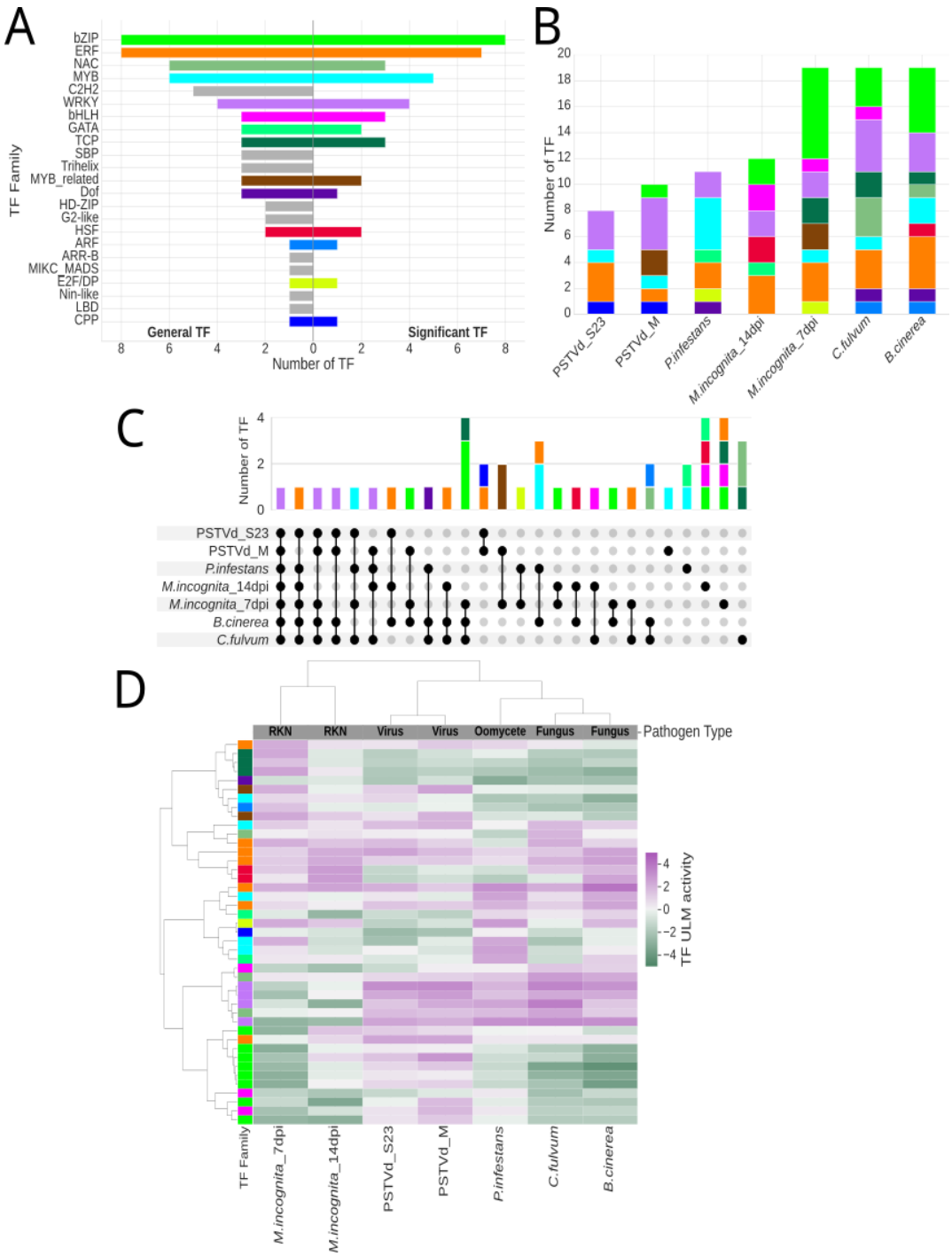

Figure 3. TF activity. A) Number of TF in the GRN per family (on the left) and only for significant activity (on the right). B) Number of TF with significant activity by family, accordingly to A, per condition. C) Upset plot showing the number of conditions in which TFs were reported to have a significant activity. Colors accordingly to TF family, legend in A. D) Heatmap and hierarchical clustering, showing the pattern of TF activity (only for significant in at least one condition) in response to each pathogen infection. The family of the TF is reported on the left, colored accordingly to A.

We then focused on studying the relationship between the family and the condition in which TFs are significantly active (Figure 3B). Infection of either *C. fulvum*, *B. cinerea* and *M. incognita* (7 dpi)

activates the highest number of TFs in tomato (Figure 3B). All 8 TFs of bZIP family are active upon *M. incognita* infection in at least one of the two time points. Four out of seven TFs from ERF family are active upon *B. cinerea* attack and three out of four MYB TFs are active in response to *P. infestans*. The three NAC TFs are specifically active upon fungi infection, all upon *C. fulvum* and one also upon *B. cinerea*.

By studying the distribution in conditions where the TF are significantly active, we found that 14 TFs were reported active to respond specifically to only one stressor and six are active in at least four conditions, while 29 TFs have a significant activity in response to at least two pathogens including one triggered by all stressors (Figure 3C). Regarding the 14 responding exclusively to one condition, the TFs belong to 8 families, 8 are exclusive for *M. incognita* response (7 or 14 dpi), 3 to *C. fulvum*, 2 to *P. infestans* and one to the virus (mild strain). Intriguingly, the four WRKY TFs are active in at least 4 conditions including one which is active in response to all biotic stressors, suggesting that TFs belonging to this family might have a potential broad role in regulating transcription activity in response to multiple pathogens (Figure 3C). By inspecting the activity profiles of TFs in the different conditions, we can observe conserved patterns in response to the two fungi included in the dataset *C. fulvum* and *B. cinerea*, which are opposite to the activity in response to *M. incognita* (Figure 3D, Supplementary Figure 5).

In conclusion, the GRN inferred by combining TomTom knowledge graph and HIVE selection allowed to estimate TF activity in response to specific or multiple biotic stressors.

Comparing the tomato GRNs in response to multiple pathogens.

After exploring the global GRN signatures, we decided to focus on the analysis of the GRN for each pathogen response. To extract those patterns from the complex GRN, we first needed to find a simplified representation. From the complete GRN, we constructed a network by considering only TFs as nodes and edges were added when TF pairs shared targets using the mapper algorithm [36] from topological data analysis (TDA) [19]. Briefly, the Mapper algorithm is a data visualization method that produces a graph representing the dataset, and its corresponding topology (connected components, branches, loops), refer to the methods section for more detail. Therefore, we obtained a simplified and compressed version of the original GRN, composed of 18 nodes, each node represents a cluster of TFs (Figure 4A). Then, we mapped the average activity of TFs for each cluster and the number of TF with significant activity by stressor response (Figure 4B).

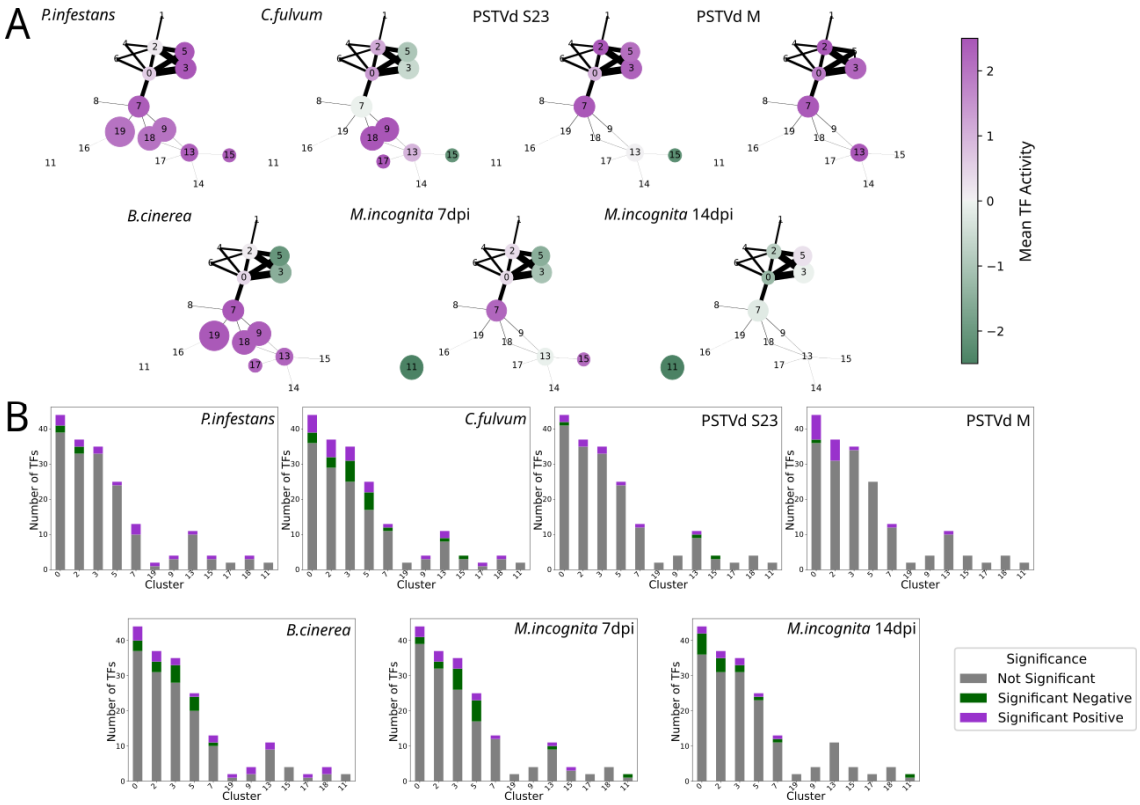

Figure 4. The Mapper algorithm from TDA yielded a simplified and compressed version of the GRN. A) The novel GRN representation composed of 18 nodes, each node represents a cluster of TFs, where the node size represents the number of targets in each cluster and edges represent shared targets among clusters and the edge size is proportional to the number of shared targets, nodes are colored by average TFs activity in clusters per pathogen infection. B) Details on the TFs activity in each cluster by condition.

Infections by the three leaf pathogens lead to the activation of several clusters of TFs compared to the other root pathogens studied and with an overall positive activity. Interestingly, clusters responding to those three pathogens are mainly active in the central and the bottom right region, including clusters 7, 19, 18, 9, 13, and 15. Despite cluster 7 being active also in response to the other pathogens, the others are specific in response to those pests. In those clusters we found TFs from the GATA, E2F, ERF and MYB families. Infections by *C. fulvum* or *B. cinerea* yield similar TFs responses with a predominantly negative activity in the upper region clusters involving twelve and ten TFs respectively, and positive in the bottom right involving TFs from NAC, WRKY and MYB families. The viruses seem to trigger an opposite response compared with the three leaf pathogens. Indeed, responses to viruses are mainly in the upper part of the network including clusters 2, 3, 5, and 0 showing a positive activation opposite to the milder negative activation showed in those clusters in response to leaf pathogens. Clusters with TFs involved in *M. incognita* response have mainly a negative activity, especially in the upper region where are the majority of TFs involved in this response. Interestingly, this pathogen is the only one among the studied pests to trigger a specific response of a TF in the periphery of the network (bottom left region). This is a TF from the bZIP family, the other seven bZIP

TFs with a significant activity are in the upper region, all responding with a negative activity to *M. incognita* attack.

Overall, we can see that the central and the upper regions are active in response to different infections, with the central one showing always positive activity and the upper either positive or negative depending on the stressor. On the other hand, the two bottom regions show a more specialized response either exclusive to one pathogen attack for the left one or to few stressors for the right part.

The topological data analysis of the GRN yielded four configurations linking the network structure and the TFs function.

The novel representation of the GRN obtained with the Mapper algorithm highlights that there are two different structures: clusters with a high number of targets but few connections with close clusters and clusters with similar number of intra/extra connections. We envisage four different possible scenarios: “exclusive targets, specific response”, one or a group of TFs with specific targets, responding to a specific attack; “exclusive targets, multiple response”, one or a group of TFs with specific targets, responding to different attacks; “shared targets, specific response”, a group of TFs with shared targets, each responding exclusively to a specific attack; “shared targets, multiple response” a group of TFs with shared targets, responding to different attacks (Figure 5A).

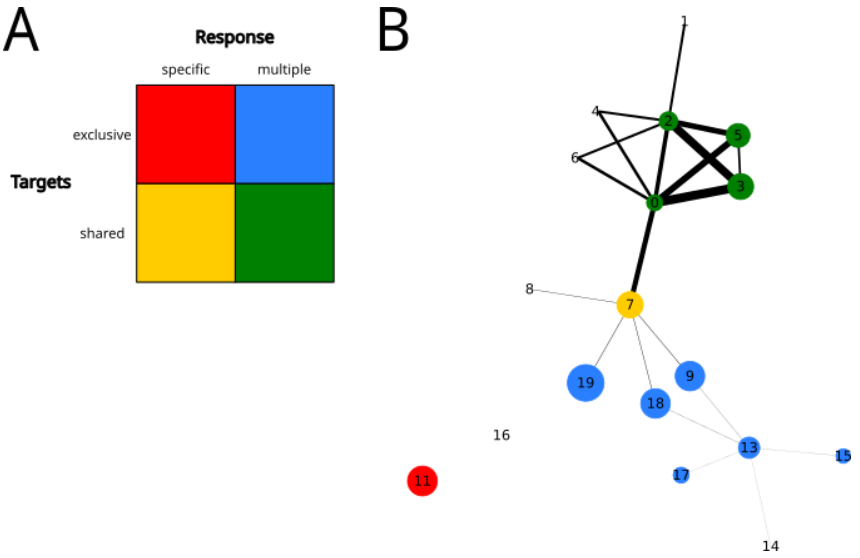

Figure 5. Network structure vs TF activity. A) The four configurations representing different combinations of network structure and TFs activity. B) The four groups of clusters found by the Mapper and ToMATo algorithms from TDA without *a priori* which correspond to the four configurations in A.

To study if we can identify the four configurations, we used the ToMATo clustering algorithm [37] on the Mapper graph. ToMATo is a hierarchical clustering algorithm which builds the cluster hierarchy

using a user-defined scalar-valued filter. We defined two filters, one for the TF activities and one for the edge thickness, each following different mixtures of Gaussians. The process yields eight clusterings, one for each condition for the TF activity and one for the edges. The consensus clustering was obtained using the Cartesian product (see methods). We obtained four main regions, representing clusters densely connected among each other but poorly with the remaining clusters (Figure 5B). We inspected the characteristics of the TFs belonging to each region to see whether those correspond to the four configurations (Supplementary Table 4). Cluster 11 was found isolated, which corresponds to the TF with significant activity in response only to *M. incognita* infection from bZIP family, described also in the previous paragraph. Being at the periphery of the graph in a poorly connected region, this TF has exclusive targets. This corresponds to the configuration "exclusive targets, specific response". Similarly, cluster 7 was found isolated, which contains seven TFs from 5 families (3 ERF, 2 GATA, 1 E2F, 1 ARF). The overall activity of this region is positive in response to all pathogens, milder only for fungal pathogens, however, for each stress or pair of stressors a different TF is involved despite the shared targets, therefore they belong to the configuration "shared target-specific response". In the bottom-right region, there are in total four TFs with significant activity, from four families NAC, ERF, MYB and WRKY. Since they are spread in different clusters, they have limited shared clusters, however, each TF is involved in response to multiple conditions from two to four, mainly activated in response *P. infestans*, *B. cinerea* and *C. fulvum*. Therefore, this region corresponds to the "exclusive target-multiple response" configuration. Finally, the upper region contains 31 TFs with significant activity, they share multiple targets and most of the TFs are active in response to multiple stressors positively or negatively. This region corresponds to the "shared targets-multiple response". We observe a tendency for negative activity in response to *M. incognita* or *B. cinerea*, and positive for the viruses or *P. infestans*. Intriguingly, 12 of the 14 TFs responding exclusively to one condition are in this region suggesting that they share targets with non-specifically responding TFs.

Overall, we observe a tendency of TFs belonging to the same family to share targets, activity and response to pathogens. Coupling the network structure and the activity estimation, our approach based on the Mapper and ToMATo algorithms from TDA yielded four configurations of how the tomato orchestrates the complex regulatory network to respond to different pests identifying specific and common mechanisms.

Four ERF TFs coordinate the tomato response to multiple pathogens.

To identify key regulators, we partitioned the global regulome accordingly to the four regions identified by the previous analysis. The red region is composed by only one TF with significant activity, therefore was not included in the analysis. Hence, we calculated degree and betweenness and the significance of the activity of TF to retrieve the hubs of each region (see methods). The analysis yielded four hubs, two in the green region and one in each of the other two (Supplementary Table 5). Intriguingly, those TFs found as hubs, all belong to the ERF family and have a significant activity in

response to two or three different pathogens. Furthermore, the activity of those TFs is always positive despite the pest attack. We also observe that each of these TFs were found to be active in response to pathogens with very different characteristics and/or infecting different tissues. The Solyc07g054220 and Solyc06g051840 in the green region actively respond to the virus, *B. cinerea* and *M. incognita* (14 dpi) and *C. fulvum*, *M. incognita* (7dpi), respectively. Solyc03g093550 and Solyc12g009240 in yellow and blue region, actively respond to *C. fulvum*, *B. cinerea* and *M. incognita* (14 dpi) and *B. cinerea* and *P. infestans*, respectively. These findings may suggest that those four key regulators of plant defense are involved in a multi-pathogens response regardless pathogen characteristics or infected tissue. A rising body of evidence shows that ERFs are closely involved in defense responses against various pathogens in plants [41] [42] [43]. Here we found that the same TF from ERF is involved in regulating the response to multiple attacks.

To study if those hubs share targets, we extracted the first neighbors of those four TFs from the global GRN and we obtained a connected sub-network (Figure 6A). Targets are shared by two or more hubs, and we observe a core in the sub-network composed of nodes shared by at least three hubs. We calculated the number of targets shared by all possible pairs or triplets of TFs (Figure 6B). Intriguingly, the three TFs which share the highest number of targets are the Solyc06g051840, Solyc07g054220 and Solyc03g093550, all responding to *M. incognita* infection in addition to either *C. fulvum* or *B. cinerea* or PSTVd_S. The pair with the highest number of shared targets is Solyc06g051840 and Solyc03g093550, both responding to *C. fulvum* and either *M. incognita* or *B. cinerea.* Of note, the number of targets shared by the four hubs is more elevated than the number of targets shared by three or two TF.

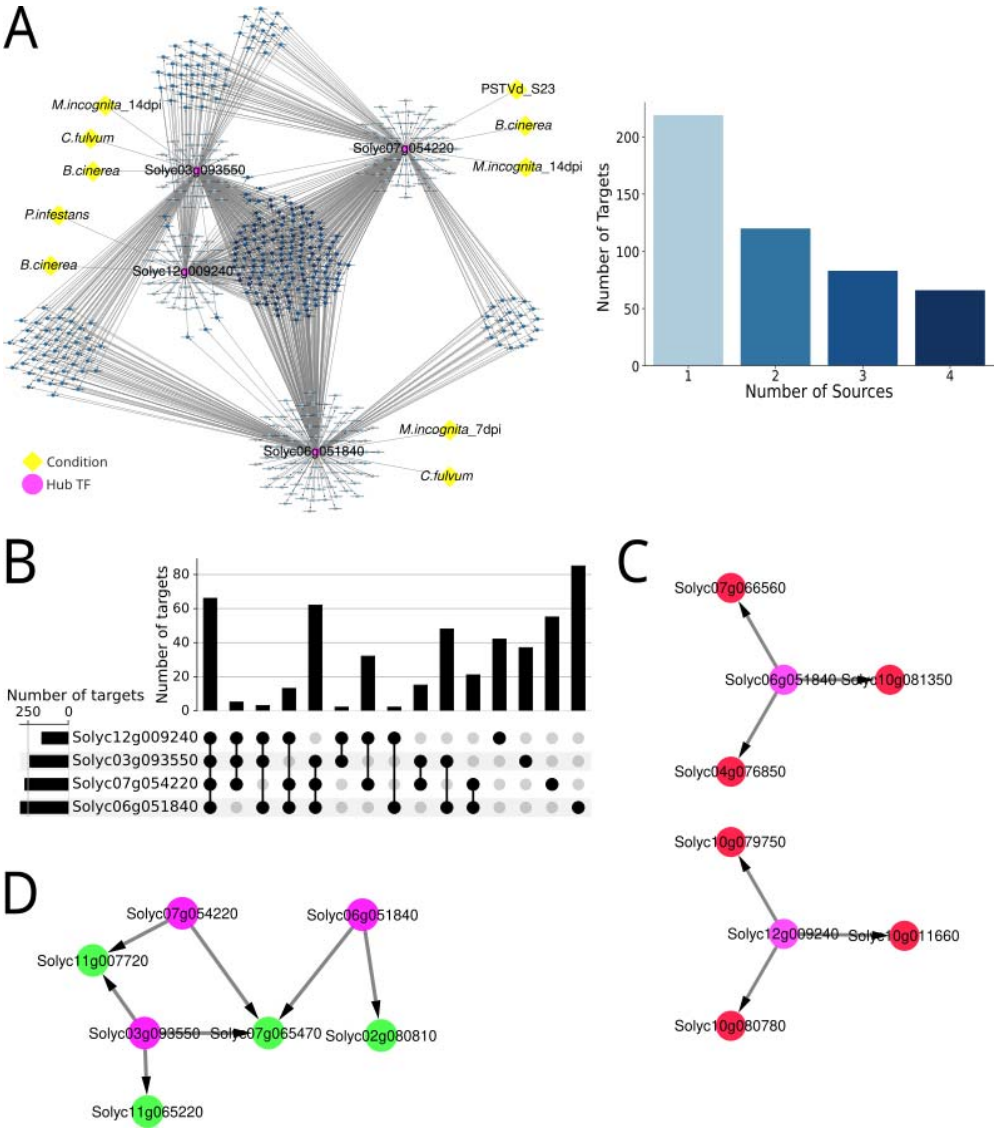

Figure 6. Hubs' analysis on the four groups of clusters found by TDA. A) The sub network of the global GRN composed of the four hubs and their first targets. The conditions in which each TF was found with significant activity are represented as links to the corresponding conditions. The barplot represents the number of targets which are either specific to one TF or shared between multiple hubs from two to four. Colors of nodes in the network accordingly to the barplot. B) Upset plot showing the number of targets exclusive of each hub or shared among multiple hubs. On the left, barplot showing the total number of targets by hub. C) Sub-network of the TFs/targets associated to the pathway "hormone signal transduction" with significant negative activity in response to *P. infestans.* D) Sub-network of the TFs/targets associated to the pathway "carbon metabolism" with significant negative activity in response to *M. incognita.*

Since these TFs share several targets, we hypothesized that they may are involved in common biological functions. Therefore, we performed the pathway activity analysis on the sub-network including only the hubs and their targets, as we did for the entire GRN. We found two pathways with a significant activity; both were found as significant also from the analysis on the entire GRN (Supplementary Figure 6). Although the level of significance was close to the threshold, the

characteristics of those pathways captured our attention. The plant hormone signal transduction pathway, which was particularly found negatively active upon *P. infestans* infection, included two hubs, one of which also has significant activity in response to the same pathogen (Figure 6C), and six targets. Although the two TFs share common targets, the six involved in the significant pathway are not among those. This observation may suggest that the two TFs are involved in two different signaling within the same pathway and therefore targeting different genes to activate different cascades. The role of ERFs in reprogramming expression of defense-related genes involved in hormonal signaling pathways was already found in other phytopathosystems [44] [45] [46]. Our study adds evidence of the role of this family in mediating the tomato response to multiple pathogens.

The other significantly active pathway concerns the carbon metabolism, which is a pivotal process in plant response against biotic stress. We found three hubs which share four targets involved in this pathway. Importantly, not only was the pathway found active in response to *M. incognita* aggression, but also each of the three TFs: Solyc06g051840 at 7 dpi, Solyc07g054220 and Solyc03g093550 at 14 dpi (Figure 6D) and the four targets are all annotated as transferase. Carbon metabolism and secondary metabolism are known to play a fundamental role in plants and tomato response to Meloidogyne species [47]. Furthermore, ERF TFs have been shown to be extensively associated with regulation of the secondary metabolites synthesis in response to biotic stresses [44] [48] [49]. Here we found that ERF TFs may have a cooperative role in the modulation of this process during *M. incognita* infection.

In conclusion, our analysis suggests that the four ERF hubs have a cooperative activity by sharing targets involved in pivotal pathways for plant defense and coordinate the tomato response to multiple bio-aggressors.

Conclusions

Here we studied the response of the tomato to six distinct pathogens with a broad range of different characteristics from lifestyle to the infected tissue. The difficulty to realize multi-stress experiments in laboratory, which increases with the number and the type of stress to simultaneously study, has hindered the possibility to investigate plants' molecular mechanisms in a more realistic scenario reproducing the field environment. To address this need, we developed novel *in silico* tools, to integratively analyze single-stress experiments for the identification of multi-stress responding mechanisms. We set up an innovative framework to construct and analyze the complex GRN regulating plant molecular mechanisms to face pathogen infections.

Our framework allows to analyze complex biological questions, such as the plant response to multiple stressors without requiring performing complex experiments. By jointly analyzing several transcriptomic data of tomato subjected to one pest at the time, we found regulatory mechanisms used to respond to either specific or multiple attacks. Although we observed that the general tomato response stratifies accordingly to infected tissues and subsequently with pathogen characteristics, by changing the scale of observation we could highlight different mechanisms. Often the differential expression value of a TF is used as a proxy of its regulatory activity. However, the change in expression of the TF itself is not informative of its role on the targets. We reasoned that the regulatory activity of a TF can be better estimated by inspecting the modulation of the expression of its targets. Thus, we used a model which was developed to infer the TF activity by fitting a univariate linear model on the t-statistics of TF's targets. We showed that TFs activity is very complex, spanning from one TF with a significant activity triggered by all the studied stresses to a total of 29 (out of 43) responding to at least two distinct attacks, regardless of the pathogen's features or infected tissue. By creating a simplified representation of the original network, TDA algorithms allowed to identify 18 clusters of TFs sharing targets either responding specifically to specific pests or having a cooperative or compensating activity by sharing targets and/or activity in response to multiple pathogens, or having an antagonist activity depending on the biotic stressor. By crossing the network structure and TF activity, TDA algorithms pinpointed four main regions in the GRN. Those regions permitted to shed a light on how the tomato orchestrates the complex regulatory network to respond to different pests. The investigation of the properties of these four regions leads to the identification of four hubs, which belong to the same family of TF, the ERF family. Our findings suggest that those four ERFs play a key role in the coordination of regulatory mechanisms to face bio-aggressors. Overall, we demonstrated that TDA can provide innovative and robust instruments to analyze complex GRN.

Beyond the novel analysis methods proposed, we also developed another important resource for the plant scientists' community: TomTom. With the wealth of data generated these days, it is extremely laborious to manually perform a thorough literature survey. The current graph encompasses different molecular interactions, from transcriptional and post-transcriptional regulations to protein-protein interaction, from nine databases. TomTom yields a cartography of known interactions, all grouped

together, which can be used to extract molecular interactions occurring beyond response to biotic stress. Overall, the interrogation of TomTom enables interactive exploration of current knowledge of molecular interactions in tomato and will help to provide mechanistic hypotheses and new insight into experimental observation. Furthermore, the workflow that we set up to build the knowledge graph is easily exportable for other plants; similarly, other interactions can be added for tomato in TomTom to expand the set of regulatory processes to be studied.

Our reproducible framework from knowledge graph to TDA can therefore be used on other biological models to study more generally host-pathogens interaction also in other fields of life sciences. In conclusion, our integrative and comparative approach enabled a more global understanding of tomato-pathogen interactions and will facilitate the development of multi-pathogen control strategies**.**

Declaration of Interests

ADu reports fees from Tempus and MONTAI therapeutics.

Acknowledgments

This work was supported by the French government, through the UCA JEDI Investments in the Future project managed by the National Research Agency (ANR) under reference number ANR-15-IDEX-01.

MC: This work was also partially supported by ANR grant “TopModel”, ANR-23-CE23-0014, and supported by the French government, through the 3IA Cote d’Azur Investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001. This project was supported by INRAE in the framework of the “Métaprogramme DIGIT-BIO” and Plant Health and Environment department.

We are grateful to the bioinformatics and genomics platform, BIG, Sophia Antipolis (ISC PlantBIOs, https://doi.org/10.15454/qyey-ar89) for computing and storage resources, and we would like to thank Martine Da-Rocha for her meaningful guidance regarding the RNA-seq pre-processing.

Authors' contributions

MM: methodology; software; validation; formal analysis; resources; data curation; writing - original draft; visualization. MC: formal analysis. XAG: formal analysis; resources. ADa: software. SL: methodology; software. JSR: review & editing; supervision. SJP: conceptualization; writing - review & editing. ADu: methodology; writing - review & editing; supervision. SB: conceptualization; methodology; writing - original draft; supervision; project administrator; funding acquisition.

References

[1] P. Pandey, V. Irulappan, M. V. Bagavathiannan, and M. Senthil-Kumar, “Impact of Combined Abiotic and Biotic Stresses on Plant Growth and Avenues for Crop Improvement by Exploiting Physio-morphological Traits.,” *Front. Plant Sci.*, vol. 8, p. 537, Apr. 2017, doi: 10.3389/fpls.2017.00537.

[2] M. C. Rillig *et al.*, “The role of multiple global change factors in driving soil functions and microbial biodiversity.,” *Science*, vol. 366, no. 6467, pp. 886–890, Nov. 2019, doi: 10.1126/science.aay2832.

[3] S. I. Zandalinas, S. Sengupta, F. B. Fritschi, R. K. Azad, R. Nechushtai, and R. Mittler, “The impact of multifactorial stress combination on plant growth and survival.,” *New Phytol.*, vol. 230, no. 3, pp. 1034–1048, May 2021, doi: 10.1111/nph.17232.

[4] Y. Wang, R. N. Pruitt, T. Nürnberger, and Y. Wang, “Evasion of plant immunity by microbial pathogens.,” *Nat. Rev. Microbiol.*, vol. 20, no. 8, pp. 449–464, Aug. 2022, doi: 10.1038/s41579-022-00710-3.

[5] S. I. Zandalinas and R. Mittler, “Plant responses to multifactorial stress combination.,” *New Phytol.*, vol. 234, no. 4, pp. 1161–1167, May 2022, doi: 10.1111/nph.18087.

[6] A. Dutt, D. Andrivon, and C. Le May, “Multi□infections, competitive interactions, and pathogen coexistence,” *Plant Pathology*, vol. 71, no. 1, pp. 5–22, Jan. 2022, doi: 10.1111/ppa.13469.

[7] A.-L. Barabási and Z. N. Oltvai, “Network biology: understanding the cell’s functional organization.,” *Nat. Rev. Genet.*, vol. 5, no. 2, pp. 101–113, Feb. 2004, doi: 10.1038/nrg1272.

[8] H. Jeong, S. P. Mason, A. L. Barabási, and Z. N. Oltvai, “Lethality and centrality in protein networks,” *Nature*, vol. 411, no. 6833, pp. 41–42, May 2001, doi: 10.1038/35075138.

[9] R. Peyraud, U. Dubiella, A. Barbacci, S. Genin, S. Raffaele, and D. Roby, “Advances on plant-pathogen interactions from molecular toward systems biology perspectives.,” *Plant J.*, vol. 90, no. 4, pp. 720–737, May 2017, doi: 10.1111/tpj.13429.

[10] C. E. Hernando, A. Romanowski, and M. J. Yanovsky, “Transcriptional and post-transcriptional control of the plant circadian gene regulatory network.,” *Biochim. Biophys. Acta Gene Regul. Mech.*, vol. 1860, no. 1, pp. 84–94, Jan. 2017, doi: 10.1016/j.bbagrm.2016.07.001.

[11] S. Aibar *et al.*, “SCENIC: single-cell regulatory network inference and clustering.,” *Nat. Methods*, vol. 14, no. 11, pp. 1083–1086, Nov. 2017, doi: 10.1038/nmeth.4463.

[12] S. Müller-Dott *et al.*, “Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities.,” *Nucleic Acids Res.*, vol. 51, no. 20, pp. 10934–10949, Nov. 2023, doi: 10.1093/nar/gkad841.

[13] V. K. Singh, A. K. Singh, and A. Kumar, “Disease management of tomato through PGPB: current trends and future perspective.,” *3 Biotech*, vol. 7, no. 4, p. 255, Aug. 2017, doi: 10.1007/s13205-017-0896-1.

[14] T. Arie, H. Takahashi, M. Kodama, and T. Teraoka, “Tomato as a model plant for plant-pathogen interactions,” *Plant Biotechnology*, vol. 24, no. 1, pp. 135–147, 2007, doi: 10.5511/plantbiotechnology.24.135.

[15] M. M, “Tomato: A Model Plant to Study Plant-Pathogen Interactions,” *FSNT*, vol. 4, no. 1, pp. 1–6, Jan. 2019, doi: 10.23880/fsnt-16000171.

[16] M. R. Ercolano, W. Sanseverino, P. Carli, F. Ferriello, and L. Frusciante, “Genetic and genomic approaches for R-gene mediated disease resistance in tomato: retrospects and prospects.,” *Plant Cell Rep.*, vol. 31, no. 6, pp. 973–985, Jun. 2012, doi: 10.1007/s00299-012-1234-z.

[17] G. Calia *et al.*, “Disentangling plant response to biotic and abiotic stress using HIVE, a novel tool to perform unpaired multi-transcriptomics integration,” *BioRxiv*, Mar. 2024, doi: 10.1101/2024.03.04.583290.

[18] P. Badia-I-Mompel *et al.*, “decoupleR: ensemble of computational methods to infer biological activities from omics data.,” *Bioinformatics Advances*, vol. 2, no. 1, p. vbac016, Mar. 2022, doi: 10.1093/bioadv/vbac016.

[19] L. Wasserman, “Topological Data Analysis,” *Annu. Rev. Stat. Appl.*, vol. 5, no. 1, pp. 501–532, Mar. 2018, doi: 10.1146/annurev-statistics-031017-100045.

[20] N. Fernandez-Pozo *et al.*, “The Sol Genomics Network (SGN)--from genotype to phenotype to breeding.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D1036-41, Jan. 2015, doi: 10.1093/nar/gku1195.

[21] A. Kozomara, M. Birgaoanu, and S. Griffiths-Jones, “miRBase: from microRNA sequences to function.,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D155–D162, Jan. 2019, doi: 10.1093/nar/gky1141.

[22] J. Jin *et al.*, “PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants.,” *Nucleic Acids Res.*, vol. 45, no. D1, pp. D1040–D1045, Jan. 2017, doi: 10.1093/nar/gkw982.

[23] F. Tian, D.-C. Yang, Y.-Q. Meng, J. Jin, and G. Gao, “PlantRegMap: charting functional regulatory maps in plants.,” *Nucleic Acids Res.*, vol. 48, no. D1, pp. D1104–D1113, Jan. 2020, doi: 10.1093/nar/gkz1020.

[24] J. Liu, X. Liu, S. Zhang, S. Liang, W. Luan, and X. Ma, “TarDB: an online database for plant miRNA targets and miRNA-triggered phased siRNAs.,” *BMC Genomics*, vol. 22, no. 1, p. 348, May 2021, doi: 10.1186/s12864-021-07680-5.

[25] Y. Fei, R. Wang, H. Li, S. Liu, H. Zhang, and J. Huang, “DPMIND: degradome-based plant miRNA-target interaction and network database.,” *Bioinformatics*, vol. 34, no. 9, pp. 1618–1620, May 2018, doi: 10.1093/bioinformatics/btx824.

[26] X. Yi, Z. Zhang, Y. Ling, W. Xu, and Z. Su, “PNRD: a plant non-coding RNA database.,” *Nucleic Acids Res.*, vol. 43, no. Database issue, pp. D982-9, Jan. 2015, doi: 10.1093/nar/gku1162.

[27] D. Szklarczyk *et al.*, “The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest.,” *Nucleic Acids Res.*, vol. 51, no. D1, pp. D638–D646, Jan. 2023, doi: 10.1093/nar/gkac1000.

[28] L. Cooper, J. Elser, M.-A. Laporte, E. Arnaud, and P. Jaiswal, “Planteome 2024 update: reference ontologies and knowledgebase for plant biology.,” *Nucleic Acids Res.*, vol. 52, no. D1, pp. D1548–D1555, Jan. 2024, doi: 10.1093/nar/gkad1028.

[29] M. Kanehisa, M. Furumichi, Y. Sato, Y. Matsuura, and M. Ishiguro-Watanabe, “KEGG: biological systems database as a model of the real world.,” *Nucleic Acids Res.*, vol. 53, no. D1, pp. D672–D677, Jan. 2025, doi: 10.1093/nar/gkae909.

[30] S. Lobentanzer *et al.*, “Democratizing knowledge representation with BioCypher.,” *Nat. Biotechnol.*, vol. 41, no. 8, pp. 1056–1059, Aug. 2023, doi: 10.1038/s41587-023-01848-y.

[31] P. Ewels, M. Magnusson, S. Lundin, and M. Käller, “MultiQC: summarize analysis results for multiple tools and samples in a single report.,” *Bioinformatics*, vol. 32, no. 19, pp. 3047–3048, Oct. 2016, doi: 10.1093/bioinformatics/btw354.

[32] S. Chen, “Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp.,” *iMeta*, vol. 2, no. 2, p. e107, May 2023, doi: 10.1002/imt2.107.

[33] A. Dobin *et al.*, “STAR: ultrafast universal RNA-seq aligner.,” *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Jan. 2013, doi: 10.1093/bioinformatics/bts635.

[34] Y. Liao, G. K. Smyth, and W. Shi, “featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.,” *Bioinformatics*, vol. 30, no. 7, pp. 923–930, Apr. 2014, doi: 10.1093/bioinformatics/btt656.

[35] M. I. Love, W. Huber, and S. Anders, “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.,” *Genome Biol.*, vol. 15, no. 12, p. 550, 2014, doi: 10.1186/s13059-014-0550-8.

[36] G. Singh, F. Memoli, and G. Carlsson, “Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition,” *The Eurographics Association*, 2007, doi: 10.2312/spbg/spbg07/091-100.

[37] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba, “Persistence-Based Clustering in Riemannian Manifolds,” *J. ACM*, vol. 60, no. 6, pp. 1–38, Nov. 2013, doi: 10.1145/2535927.

[38] P. Shannon *et al.*, “Cytoscape: a software environment for integrated models of biomolecular interaction networks.,” *Genome Res.*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[39] R. Wiese, M. Eiglsperger, and M. Kaufmann, “yFiles: Visualization and Automatic Layout of Graphs,” in *Graph Drawing*, vol. 2265, P. Mutzel, M. Jünger, and S. Leipert, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 453–454.

[40] A. Dugourd and J. Saez-Rodriguez, “Footprint-based functional analysis of multiomic data.,” *Current Opinion in Systems Biology*, vol. 15, pp. 82–90, Jun. 2019, doi: 10.1016/j.coisb.2019.04.002.

[41] G. Reboledo, A. Agorio, L. Vignale, A. Alvarez, and I. Ponce De León, “The moss-specific transcription factor PpERF24 positively modulates immunity against fungal pathogens in Physcomitrium patens.,” *Front. Plant Sci.*, vol. 13, p. 908682, Sep. 2022, doi: 10.3389/fpls.2022.908682.

[42] Z. Zang *et al.*, “Maize Ethylene Response Factor ZmERF061 Is Required for Resistance to Exserohilum turcicum.,” *Front. Plant Sci.*, vol. 12, p. 630413, Mar. 2021, doi: 10.3389/fpls.2021.630413.

[43] N. Ma *et al.*, “Plant disease resistance outputs regulated by AP2/ERF transcription factor family.,” *Stress Biol.*, vol. 4, no. 1, p. 2, Jan. 2024, doi: 10.1007/s44154-023-00140-y.

[44] C. S. Moffat, R. A. Ingle, D. L. Wathugala, N. J. Saunders, H. Knight, and M. R. Knight, “ERF5 and ERF6 play redundant roles as positive regulators of JA/Et-mediated defense against Botrytis cinerea in Arabidopsis.,” *PLoS ONE*, vol. 7, no. 4, p. e35995, Apr. 2012, doi: 10.1371/journal.pone.0035995.

[45] Y. Maruyama *et al.*, “The Arabidopsis transcriptional repressor ERF9 participates in resistance against necrotrophic fungi.,” *Plant Sci.*, vol. 213, pp. 79–87, Dec. 2013, doi: 10.1016/j.plantsci.2013.08.008.

[46] H. Yang *et al.*, “Genome-wide identification and functional analysis of the ERF2 gene family in response to disease resistance against Stemphylium lycopersici in tomato.,” *BMC Plant Biol.*, vol. 21, no. 1, p. 72, Feb. 2021, doi: 10.1186/s12870-021-02848-3.

[47] R. Gupta, M. F. B. Mfarrej, B. Xhemali, A. Khan, H. Nadeem, and F. Ahmad, “Metabolic responses of plants to Meloidogyne species parasitism: A review on molecular events and functions,” *Journal of King Saud University - Science*, vol. 36, no. 2, p. 103083, Feb. 2024, doi: 10.1016/j.jksus.2023.103083.

[48] S. Imano *et al.*, “AP2/ERF Transcription Factor NbERF-IX-33 Is Involved in the Regulation of Phytoalexin Production for the Resistance of Nicotiana benthamiana to Phytophthora infestans.,” *Front. Plant Sci.*, vol. 12, p. 821574, 2021, doi: 10.3389/fpls.2021.821574.

[49] M. Nakayasu *et al.*, “JRE4 is a master transcriptional regulator of defense-related steroidal glycoalkaloids in tomato.,” *Plant J.*, vol. 94, no. 6, pp. 975–990, Jun. 2018, doi: 10.1111/tpj.13911.