Mining the Conserved Transcriptional Response to Phytophthora Infection for Factors Beyond the Known PR Genes

Claude Simo
University of Douala: Universite de Douala

Kay Hofmann (✉ kay.hofmann@uni-koeln.de)
University of Cologne  https://orcid.org/0000-0002-2289-9083

Keywords: Pathogenesis-related genes, signaling peptides, cytochrome c oxidase, MOCCI, MISTR-A

Posted Date: December 6th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1117924/v1

License: ☎️ This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License
Mining the conserved transcriptional response to phytophthora infection for factors beyond the known PR genes

Claude Simo¹, Kay Hofmann²

¹ Department of Plant Biology, Faculty of Science, University of Douala. PO Box 24157 Douala-Cameroon
² Institute for Genetics, University of Cologne, Zülpicher Straße 47a, D-50674 Cologne, Germany. ORCID: 0000-0002-2289-9083

Short title: Transcriptional Response to Phytophthora

Correspondence should be addressed to:
Kay Hofmann (kay.hofmann@uni-koeln.de)
Tel: +49-(0)221 470-1701
Fax: +49-(0)221 470-3402

Key Message
By analyzing transcriptomics data from plant/phytophthora infection systems, we identified 310 gene clusters with a conserved upregulation. All clusters were bioinformatically annotated and revealed similarity to animal defense responses.
Abstract

Oomycetes of the genus *Phytophthora* are devastating plant pathogens that affect many commercially important plants. Considerable efforts have been made to investigate the transcriptional response of individual plant species to *phytophthora* infection, often showing a concerted upregulation of pathogen-response (PR) gene families, which are also induced upon infection by fungi and other biotic and even non-biotic stressors. By integrating four transcriptomics datasets derived from three different plants (arabidopsis, soybean, cocoa), a core set of upregulated sequence clusters was derived, which represents a conserved multi-species response to *phytophthora* infections. We annotated more than 300 common induced gene clusters and subjected them to bioinformatical analysis. Besides the expected PR genes, several novel gene families without known links to biotic stress were found to be strongly induced in all tested datasets. Among the most prominent response genes are two families of putatively secreted peptides and a family of predicted mitochondrial complex-IV associated proteins. Interestingly, the latter sequences are related to the mammalian NDUFA4 family, which also contains members with constitutive and pathogen-induced expression. This recurrent functional diversification points toward an important role of complex IV regulation within the biotic defense response in multiple kingdoms.

Keywords: Pathogenesis-related genes, signaling peptides, cytochrome c oxidase, MOCCI, MISTR-A.

Introduction

Domestication of flowering plants forms the basis of a productive and sustainable agriculture that is under constant challenge by the attack of emerging pathogens (Scholthof, 2003; McDonald and Stukenbrock, 2016). Most species in the genus *Phytophthora* of oomycetes are devastating plant pathogens that cause severe attacks on many wild and cultivated plants (Grunwald and Flier, 2005; Meng et al., 2014; Marelli et al., 2019; Boevink et al., 2020). In general, *Phytophthora* species have a very flexible life cycle and ability to adapt to changing environmental conditions (Jeger and Pautasso, 2008; Naveed et al., 2020). Today, there are more than 100 species of the genus *Phytophthora* with varying host ranges; some of these species like *P. sojae* and *P. megakarya* are restricted to a single known species, while others like *P. capsici, P. ramorum, P. parasitica* and *P. palmivora* can infect a wide range of plants from several genera (Blair et al., 2008; Meng et al., 2014). A particular threat to cocoa plantations in West and Central Africa is *P. megakarya*, a restrictive but very aggressive species, while infections of cocoa by *P. palmivora* are more widespread around the world, causing significant crop losses (Doungous et al., 2018; Marelli et al., 2019).

*Phytophthora* species share a hemibiotrophic lifestyle. Their asexual life cycle in plants is characterized by the adhesion of mobile zoospores to the host tissue, followed by encystment and germ
tube formation (Boevink et al., 2020; Chepsergon et al., 2020) Entry into the plant is achieved by
surface appressoria and is followed by establishment of an apoplastic hyphal network. At the
biotrophic stage, Phytophthora produce haustoria in plant cells. These structures participate in
the acquisition of nutrients and release virulence proteins known as effectors (Petre and Kamoun,
2014; Dong and Ma, 2021). This is followed by the necrotrophic stage, characterized by tissue
necrosis of the host and the production of numerous sporangia that release zoospores (Judelson
and Blanco, 2005; Hardham, 2007; Boevink et al., 2020).

During this infection stage, the host plant can sense the pathogen, either directly or by virtue of
its activity, and then mount a more or less targeted defense response. As the first line of defense,
pathogen-associated molecular patterns (PAMPs) are recognized by plant surface receptors,
which then initiate a defense response called PAMP-triggered immunity (PTI) (Naveed et al., 2020;
Ngou et al., 2021). Most pathogens, including Phytophthora, secrete a number of effectors into
the host cell, which interfere with the PTI response. (Petre and Kamoun, 2014; Chepsergon et al.,
2020; Dong and Ma, 2021). These effectors, in turn, are recognized by specific resistance genes
or R genes, which initiate a second branch of the immune system called effector-triggered im-
munity (ETI) (Naveed et al., 2020; Ngou et al., 2021). One consequence of ETI activation is the
hypersensitive response (HR), a form of programmed cell death surrounding the infection site,
which deprives biotrophic pathogens of a food source and prevents their spread (Balint-Kurti,
2019). Another consequence of ETI activation is the transcriptional induction of a wide range of
host genes coding for proteins that either fight the pathogen directly or modify the host cell and
its environment to limit pathogen spread. The genes induced by this pathway are often effective
against multiple classes of pathogens and have been classified into different families of patho-
genesis-related genes (PR-genes) (Stintzi et al., 1993; Sels et al., 2008; Ali et al., 2018; Zribi et
al., 2020).

Besides the dedicated pathogen-response genes, a number of plant signal transduction pathways
are involved in pathogen defense, among them the signaling by the phytohormones salicylic acid
(SA), jasmonic acid (JA), abscisic acid (ABA), gibberellic acid (GA), brassinosteroids (BR), eth-
ylene (ET) and auxins (McDowell and Dangl, 2000). In most cases, the phytohormone-mediated
pathways also control plant development and the abiotic and biotic stress response. The im-
portance of these pathways for different pathogens varies and depends on the biotrophic/necrot-
rophic lifestyle of the pathogen class, but also on the presence of pathway-inhibitory factors en-
coded by specific pathogens or even pathogen strains (Glazebrook, 2005; Pieterse et al., 2012;
Di et al., 2016).

The identification of pathogen-responsive host genes by transcriptomics methods has a long tra-
dition (Wise et al., 2007) and was instrumental for defining the PR-gene classes (Stintzi et al.,
1993). While RNA-sequencing gradually supersedes the traditional microarray-based techniques,
host transcriptomics continues to be a valuable approach to study host-pathogen interactions,
leading to the identification of new response genes (Le Berre et al., 2017) or helping with the
identification of resistance traits (Meng et al., 2021). The aim of the current study was to mine
plant transcriptomics data for novel and unconventional regulatory events that broaden our un-
derstanding of the host response to phytophthora infections. Virtually every dataset measuring
the transcriptional response to pathogens contains a number of upregulated genes that do not belong any of the known PR-classes and don’t have a (predicted) function connected to the defense response – if they have any annotated function at all. Some of these observations might be due to artefacts (Petri et al., 2012), but others might represent unexpected but genuine regulations. In order to reliably identify novel and biologically relevant regulations over a background of transcriptional noise, we have applied a bioinformatical pipeline that searches for conserved regulation events in small gene families from different species, measured by different transcriptomics platforms. The gene families targeted by this approach may either be groups of orthologs, or include closely related paralogs as they are often observed in lineage-specific gene duplications (Sonnhammer and Koonin, 2002). Requiring a conserved regulation mode should deplete spurious false-positive observations and emphasize gene families, whose pathogen-responsive induction is biologically meaningful and has been selected for by evolution.

As the basis of our analysis, we selected four high-quality transcriptomics datasets generated from three plant species (arabidopsis, soybean and cocoa) and four different phytophthora species (P. parasitica, P. sojae, P. megakarya and P. palmivora) (Attard et al., 2014; Lin et al., 2014; Fister et al., 2016; Ali et al., 2017). After defining species-specific sets of strongly induced genes and clustering them for sequence similarity, we created a resource of 310 upregulated sequence clusters (USCs) that were subjected to manual functional annotation, supported by bioinformatical sequence analysis. Among these clusters were several novel and unexpected gene families, including some that code for secreted signaling proteins, and others that hint towards a novel role of mitochondria in plant pathogen response.

**Results**

**Identification of Phytophthora-induced genes**

The first step of the analysis was to define sets of genes that are strongly up-regulated upon infection of the three dicotyledon plants *Arabidopsis thaliana*, *Glycine max* and *Theobroma cacao* with four different phytophthora species. For Arabidopsis, we selected a dataset describing the time course of the response to *Phytophthora parasitica* (Attard et al., 2014). For Glycine, a dataset was selected that describes the response of susceptible and resistant soybean lines to Phytophthora sojae (Lin et al., 2014). For Theobroma, two datasets were analyzed: one using cDNA microarrays to measure the response to *Phytophthora palmivora* (Fister et al., 2016), the other one using RNA-Seq to measure the response of cocoa pods to *Phytophthora palmivora* and *Phytophthora megakarya* infection (Ali et al., 2017).

Not only do the selected datasets differ in the host species, they also use a diverse set of infection conditions, time points post inoculation, and measurement modalities. To reach a common frame of reference for inter-species comparisons, a unified ‘combined logarithmic induction factor’ (CLI-factor) was established for each host species. The idea of this combined factor was to select all
genes that are either strongly induced in one of the experimental conditions, or moderately induced in multiple conditions. The details of the CLI calculation are outlined below. In brief, the log-induction values of replicate experiments were averaged, and the remaining N conditions were considered orthogonal coordinates in an N-dimensional ‘regulation space’. The species-specific CLI factor is the length of each gene’s regulation vector. In general, genes surpassing a CLI factor of 4.0 were considered as strongly induced and were included in the multi-species clustering analysis. This condition was met by 1766 Arabidopsis thaliana genes, 1019 genes from Glycine max, and 1451 genes from Theobroma cacao.

Definition of a conserved induction signature

The next step of the analysis was to cluster the strongly Phytophthora-induced genes from three plant species by their sequence relationship. One aim of this clustering was to establish connections between orthologous regulated genes in different species; a second aim was to define families of commonly upregulated genes including those with multiple members in one species. Upon visual inspection of the induced gene lists, it became clear that many of them belong to huge sequence families (e.g. kinases, oxidases, leucine-rich repeat proteins etc.) that also contain many non-regulated members. As a consequence, a finer granularity in sequence clustering is required to identify regulated groups of bona fide orthologs, possibly including their closely related inparalogs (Sonnhammer and Koonin, 2002).

To address this issue, protein sequences encoded by the strongly induced genes were compared to each other using BLAST (Altschul et al., 1990). In addition, all sequences were compared to a database of outgroup sequences, representing various taxa outside the flowering plants. Significant BLAST hits between pairs of induced proteins were accepted only if these hits were better (lower composition-adjusted E-value) than any match between one of the candidates and the outgroup database. In total, 13612 connections fulfilled these criteria and were rendered into a network diagram using the Cytoscape software (Shannon et al., 2003). As can be seen in the overview figure 1, the outgroup filtering prevented the generation of huge non-informative superfamilies and segmented the network into manageable clusters, most of them with less than twenty member-proteins from multiple species. The full network is provided as Supplementary file S1.

Since plant genome databases contain large multi-domain proteins and occasionally also fusions between otherwise unrelated genes, it is not guaranteed that all proteins connected by unbroken network paths are truly related to each other. Fusions or multi-domain proteins can bridge two or more unrelated sub-clusters. Thus, before subjecting the families to manual sequence analysis and annotation, densely-connected clusters with a minimum size of three members were identified using Cytoscape/MCODE (Bader and Hogue, 2003; Shannon et al., 2003). The resulting list of 310 upregulated sequence clusters (USCs) is provided in Supplementary file S2.

The majority of identified clusters encompass sequences from multiple species, including 125 clusters containing putative orthologs from all three plants and 141 clusters with sequences from two plants. The remaining 44 mono-species clusters, including 18 from Arabidopsis, 3 from Glycine and 23 from Theobroma, are either derived from transposons or represent closely related inparalogs resulting from lineage-specific gene duplication events (Sonnhammer and Koonin,
The upregulated gene families code for proteins with known or predicted localizations in all major subcellular compartments. Particularly prominent are families localized in the cytoplasm (85 clusters), extracellular space (54 clusters), plasma membrane (43 clusters), nucleus (39 clusters), ER/Golgi (21 clusters) and mitochondria (18 clusters).

As expected, most of the established pathogenesis-related (PR) gene groups (Stintzi et al., 1993; Sels et al., 2008; Ali et al., 2018; Zribi et al., 2020) were found prominently in the clustering analysis (Table 1, Supplementary file S3). While the more homogeneous PR-families are represented by a single sequence cluster, more diverse families such as β-1,3-glucanases (PR-2), chitinases (PR-3) or peroxidases (PR-9) are subdivided into multiple USCs. The observed segmentation of protein families into multiple USCs usually indicates the presence of upregulated subfamilies within a larger superfamily, which may also contain non-induced genes. One striking example is the peroxidase superfamily, which is very heterogeneous and contains enzymes localized in different compartments and acting on different substrate classes (Pandey et al., 2017). Only selected peroxidase subfamilies are active against pathogenesis-related substrates and are induced under infection conditions. Interestingly, members of all five peroxidase USCs found in the present study are predicted to be localized in the extracellular space (Supplementary file S2).

Strong induction of secreted peptide precursors

Two of the strongest upregulated clusters comprise short proteins with a predicted extracellular localization, which might act as precursors for signaling or defense peptides. Cluster #126, which consists of the two Arabidopsis genes (AT2G23270 and AT4G37290), one Theobroma gene (Tc01_p005760) and two Glycine genes (glyma05g21360 and glyma17g18230) showed a particularly strong and consistent 50 to 300-fold upregulation in the different infection models (Figure 2). There were no major differences between the different time points in the Arabidopsis model (Attard et al., 2014), between P. palmivora and P. megakarya infections in the Theobroma model (Ali et al., 2017), or between resistant and susceptible Soybean lineages (Lin et al., 2014) (Figure 2a). The two Arabidopsis members of this cluster have been described previously as the source of the PAMP-induced peptides PIP2 and PIP3; the corresponding gene names are PREPIP2 and PREPIP3 (Vie et al., 2015; Najafi et al., 2020; Hussain et al., 2021). A third published PIP-encoding gene, AT4G28460/PREPIP1 (Hou et al., 2019; Shen et al., 2020), is more divergent and therefore not a member of the upregulated PREPIP2/3 sequence cluster. Instead, it forms a separate USC #266 together with two Theobroma genes (Tc04_p018060 and Tc04_p018070), which are also transcriptionally induced, albeit not as strongly as PREPIP2 and 3 (Figure 2a).

Our bioinformatical analysis shows that members of USC #126, including Arabidopsis PREPIP2/3 are members of a protein family found in all major dicotyledon lineages, which is characterized by a conserved C-terminal ~22aa repeat that is present in two, four or ten tandem copies in the cluster members from Arabidopsis, Glycine and Theobroma, respectively (Figure 2b). Each repeat copy ends on His or Asn, a property typical of receptor-binding plant hormones (Zhang et al., 2016). The highly conserved GPSP-motif is shared with the PREPIP1 family (Hou et al., 2019) and is similar to the GHSP-motif found in the PCEP family (Ogilvie et al., 2014), suggesting that these families are distant relatives (Figure 2c). The repeat architecture of the PREPIP2/3 family,
which is absent from PREPIP1 and PCEP, and the strong response to phytophthora infection suggest that these proteins are particularly relevant for signaling an infection by oomycetes.

A new family of mitochondrial anti-ROS protection factors

An unexpected, but particularly interesting finding was the strong upregulation of sequence cluster #87, comprising a number of short mitochondrial proteins. A detailed bioinformatical analysis of this family revealed that these proteins are distant relatives of the NDUFA4 subunit of the mitochondrial complex IV, also known as the cytochrome c oxidase complex (Figure 3a). Plant members of the NDUFA4 family had not been described so far. All land plants encode multiple members of the NDUFA4-like family, but in the three plant models studied here, only selected family members are strongly upregulated by phytophthora and are thus included in USC #87 (Figure 3b,c). The sequences of the plant NDUFA4 homologs have substantially diverged from their animal counterparts (Figure 3a), and the dendrogram analysis suggests that the diversity of the plant members (2 genes in Arabidopsis, 3 genes in Theobroma, and 7 genes in Glycine) arose after the split between the plant and animal kingdoms (Figure 3a,c).

Four of the plant genes (AT3G29970, Tc00_p011650, glyma08g27620 and glyma18g50800) are particularly responsive to Phytophthora infection and form a clade in the dendrogram (Figure 3c). A second, less strongly induced clade is formed by the USC #87 members Tc02_p026410 and glyma18g01340, together with glyma11g37370. The latter gene is not annotated as part of cluster #87, since its upregulation factor did not quite meet the CLI >4.0 criterion. A third clade is formed by the constitutively expressed family members AT3G48140, Tc03_p000330, glyma04g34840 and glyma06g19850, which are not induced by Phytophthora and thus not members of any USC (Figure 3c). Mammalian genomes also contain multiple members of the NDUFA4 family, which result from independent gene duplication events (Figure 3c). Interestingly, human NDUFA4 is constitutively expressed, while its two paralogs NDUFA4L2 and C15ORF48 (also known as MOCCI or MISTR-A) are upregulated in response to hypoxia and infections, respectively (Tello et al., 2011; Sorouri et al., 2020; Lee et al., 2021). Thus, the acquisition of a pathogen-responsive NDUFA4L2 paralog has happened at least two times during the evolution of plants and animals and appears to protect mitochondria from excessive ROS production during the defense response.

Besides the NDUFA4 family, there are also several other upregulated clusters populated with mitochondrial proteins. Among them is cluster #176, whose member proteins appear to be homologous to yeast respiratory chain supercomplex factor Rcf2, a hypoxia-inducible gene that also regulates cytochrome c oxidase (Rompler et al., 2016; Hoang et al., 2019). Several other clusters, such as USC #178, #285, #90 and #110, encode short mitochondrial proteins of unknown function, which might work in a similar pathway.

Discussion

The analysis of the transcriptional response of plants to an infection situation has been instrumental for the discovery of many genes that are actively involved in the defense response (Wise
et al., 2007). Some of gene regulation events are initiated by the sensing of the pathogen itself via PAMPs (pathogen-associated molecular patterns), but other gene inductions are caused by the pathogen-inflicted damage and stresses and thus overlap with abiotic stress responses (Zarattini et al., 2021). A recurring problem with transcriptomics-based discovery of genes belonging to a particular biological response pathway is the presence of spurious signals – in particular when searching for ‘novel’ or ‘unexpected’ regulation events. The current study addresses this problem by comparing multiple datasets obtained from different species in similar infection situations. By focusing on gene families with consistent up-regulation, spurious data are efficiently depleted since they are unlikely to be found in multiple species, measured by different transcriptomics platforms. A downside inherent to this approach is the likely loss of species-specific regulations, or those that only occur in particular experimental setups. Nevertheless, the ‘conservative’ criteria applied in the current study resulted in 310 upregulated sequence clusters (USCs), 266 of which are supported by multiple species, the other 44 by multiple paralogs within one species. The bioinformatical analysis of the cluster members revealed several new and unexpected regulations, some of which offer new insights into the plant response to phytophthora infection.

Secreted proteins are frequently induced upon biotic plant stress and form the basis of many established pathogenesis-related (PR) gene classes (Stintzi et al., 1993; Sels et al., 2008; Ali et al., 2018; Zribi et al., 2020). A large fraction of them are enzymes such as endo-glucanases, chitinases, peroxidases, ribonucleases or proteases; they either attack the pathogen directly or limit pathogen motility and entry by reshaping the plant cell wall and extracellular matrix. Other PR-genes encode secreted inhibitors and other non-enzymatic factors, such as the defensins and thaumatin-like proteins, which also directly counteract the pathogen (Dos Santos-Silva et al., 2020). Less well characterized is the defense role of plant signaling peptides, which do not attack the pathogen but rather transmit a signal to neighboring or distant cells in the plant. While several secreted peptides and their recognition by cognate surface receptors have been characterized (Murphy et al., 2012; Zhang et al., 2016), most of them appear to have a role in plant development and growth regulation. A notable exception is the family of plant elicitor peptides (PEPs), which are derived from intracellular precursor proteins (PROPEPs) and - upon infection – become secreted via cell rupture or another non-classical secretion mechanism (Bartels and Boller, 2015). There are also reports on a defense role of PIP1 and PIP3, members of another family of peptide precursor proteins (Hou et al., 2019; Najafi et al., 2020; Shen et al., 2020). Together with PIP2, a peptide hormone regulating root elongation (Hussain et al., 2021), the precursor proteins of this family enter the secretory pathway by an N-terminal signal peptide and are thought to release a C-terminal peptide hormone, which is distant from related to the functionally diverse CEP (C-terminally encoded peptide) family (Ogilvie et al., 2014) (figure 2c). Our analysis revealed that in particular the PIP2/PIP3 family, but also the PIP1 family, are strongly induced by phytophthora infection in all three tested plant species. Moreover, as shown in figure 2b, the precursor proteins PROPIP2 and PROPIP3 contain a C-terminal repeat, suggesting that multiple PIP2- and PIP3-like peptides can be released from these precursors. In particular the cocoa protein Tc01_p005760 is predicted to generate no less than ten such peptides – all of them related, but with distinct differences, making it likely that they can bind to multiple receptor types (Zhang et al.,
The annotated data compendium (Supplementary file S2) can easily be searched for other predicted extracellular response factors. Besides a large number of predicted enzymes, inhibitors, and lipid transfer proteins, several USC's contain additional bona fide plant signaling peptides – even if they are less strongly induced than the PIP clusters. Cluster #267 and #225 contain selected members of the CEP family (Ogilvie et al., 2014), cluster #192 contains proteins related to GRIM REAPER, a peptide signaling cell death via a receptor kinase (Wrzaczek et al., 2015), while cluster #98 codes for other short secreted proteins without a known function or informative homologies. The study of extracellular defense peptides and their receptors might open new avenues to treating or protecting plants by administration of artificial peptides or peptide-mimetics. Unlike extracellular proteins, those with a mitochondrial localization have rarely been implicated in the transcriptional response of plants to pathogens. A notable exception is the mitochondrial ATPase AtOM66 (Zhang et al., 2014), which was also identified in the present analysis as part of USC #21. Overall, 14 clusters were unambiguously predicted to consist of mitochondrial proteins, most of which have no functional annotation or informative homologs whatsoever (Supplementary file S2). One of these clusters, USC #87, was shown to be related to mammalian NDUFA4-like proteins, and is particularly strongly induced by phytophthora infections (figure 3). Bioinformatical analysis found that the plant family contains additional members, which are not pathogen-induced and therefore not covered by our clustering effort. As an example, the data in (Attard et al., 2014) show for the arabidopsis gene AT3G48140 a constitutively high expression level with negligible regulation (0.95 to 1.2-fold), while its paralog AT3G29970 has a low background expression but is 42-fold induced upon phytophthora treatment. Interestingly, the human NDUFA4 family also contains multiple paralogs with fundamentally different expression behavior; a number of recent studies have shown that their transcriptional regulation is directly linked to their biological function (Endou et al., 2020; Sorouri et al., 2020; Lee et al., 2021). Human NDUFA4 itself is a constitutively expressed structural component of the mitochondrial cytochrome c oxidase complex, also known as complex IV (Zong et al., 2018). One of its paralogs, NDUFA4L2, is upregulated under hypoxic conditions and thought to limit ROS production under low-oxygen conditions, most likely by replacing NDUFA4 in the respiratory chain (Tello et al., 2011). Another paralog, originally called C15ORF48 or NMES1 and recently renamed to MOCCI or MISTR-A, is upregulated under conditions of infection and inflammation (Sorouri et al., 2020; Lee et al., 2021). Like NDUFA4L2, MOCCI/MISTR-A can replace NDUFA4 in the cytochrome c oxidase complex, leading to a reduction in respiratory chain activity and ROS production. The proposed role of this regulation is to protect the host cell from excessive ROS production that happens during the immune response (Sorouri et al., 2020; Lee et al., 2021). While no experimental data are available for the plant proteins, it is reasonable to assume a similar function for AT3G29970 and the other members of USC #21, given that ROS production by mitochondria offers both chances and dangers for the host cell (Colombatti et al., 2014; Fichman and Mittler, 2020). The role of mitochondrial protection in the defense response appears underappreciated, given that similar mechanisms have evolved independently in the animal and plant lineages. The identification of phytophthora-induced peptides and mitochondrial protection factors are just
two examples for discoveries made possible by a conservative meta-analysis of existing transcriptomics data. The comprehensive cluster table provided as Supplementary file S2 contains many additional clusters of phytophthora-responsive genes in multiple species, whose exact biological function remains to be uncovered. The bioinformatical annotations added to the clusters provide a first step into this direction.

Materials & Methods

Data sources and pre-processing

Transcriptomics data were obtained from the GEO database (Edgar et al., 2002) and from the supplementary material of the indicated publications.

Arabidopsis: A dataset measuring the response of *A. thaliana* roots in a compatible interaction with *Phytophthora parasitica* was downloaded from GEO project GSE20226 (Attard et al., 2014). The dataset is based on microarray data using the Affymetrix ATH1 genome array and comprises an uninfected control and samples taken at different time points after infection (2.5h, 6h, 10.5h, 30h), each of them in two replicates. For the current analysis, averaged log-2 induction factors (IF) relative to the untreated sample were calculated for each time point. The CLI score was derived by considering the log-induction values of the four time points as independent orthogonal coordinates and calculating the length of the 4-dimensional ‘induction vector’:

\[
CLI_{\text{Arabidopsis}} = \sqrt{IF_{2.5h}^2 + IF_{6h}^2 + IF_{10.5h}^2 + IF_{30h}^2}
\]

A CLI score of 4.0 was reached by 1766 probe sets and considered sufficient for including the gene in the multi-species clustering analysis (supplementary file S4).

Soybean: A dataset measuring the response of soybean (*Glycine max*) to infection with *Phytophthora sojae* was downloaded from GEO project GSE48524 (Lin et al., 2014). The dataset is based on RNA-Seq data using the Illumina HiScanSQ platform and comprises data measured 24h post inoculation of one susceptible soybean line ("Williams") and 10 near-isogenic lines carrying different resistance determinants (Lin et al., 2014). For the current analysis, averaged log-2 upregulation factors (relative to the matched mock-infected sample) were calculated for each soybean line and an average of the ten resistant lines was formed. The CLI score was derived by considering the log-induction values of sensitive and resistant lines as independent orthogonal coordinates and calculating the length of the 2-dimensional ‘induction vector’:

\[
CLI_{\text{Glycine}} = \sqrt{IF_{\text{sensitive}}^2 + IF_{\text{resistant}}^2}
\]

A CLI score of 4.0 was reached by 1019 genes and considered sufficient for including the gene in the multi-species clustering analysis (supplementary file S4).
Cocoa: A dataset measuring the response of cocoa (Theobroma cacao) to infection with Phytophthora palmivora was downloaded from GEO project GSE73804 (Fister et al., 2016). The dataset is based on the NimbleGen T. cacao 28k array and comprises data measured 72h post inoculation in five-fold replicates. For the current analysis, an averaged log-2 upregulation factors (relative to the water-inoculated samples) was calculated. A second, independent dataset measuring the response of susceptible cocoa pods to infection with Phytophthora palmivora and Phytophthora megakarya was downloaded from the data supplement of (Ali et al., 2017). This dataset is based on RNA-Seq data using the Illumina HiSeq2000 platform and provides normalized RPKM (Reads Per kilobase of transcript per million mapped reads) values for infected and uninfected plants, averaged from three replicates (Ali et al., 2017). After averaging the replicates, the CLI score was calculated from the three orthogonal datasets as described above. Since the Fister et al data (Fister et al., 2016) showed considerably less spread than the two Ali et al data sets (Ali et al., 2017), the former was upweighted two-fold in the calculation of the CLI-score to make these heterogenous datasets better comparable:

\[
\text{CLI}_{\text{Theobroma}} = \sqrt{(2 \times \text{IF}_{P.pal,Fister})^2 + \text{IF}_{P.pal,Ali}^2 + \text{IF}_{P.meg,Ali}^2}
\]

As shown in supplementary file S4, overall 1451 cocoa genes reached a CLI score of 4.0 and were included in the multi-species clustering analysis.

Sequence clustering of upregulated genes

The top-ranking gene products from each species (exceeding the species-specific score of 4.0) were combined in a common fasta-formatted protein sequence file and subjected to all-against-all similarity searches using BLAST (Altschul et al., 1990). In addition, each protein was also compared to the proteome of four outgroup species belonging to different taxa (Selaginella moellendorfii, Chlamydomonas reinhardtii, Saccharomyces cerevisiae and Homo sapiens). All sequences were obtained from the Uniprot database (UniProt, 2021). Upregulated Sequence Clusters (USCs) were defined by joining upregulated gene products from Arabidopsis, Glycine and Theobroma when they fulfil the following criteria: i) cluster members are connected to each other by significant BLAST scores with E-value < 10^{-3}; and ii) the within-cluster BLAST scores are better (lower E-value) than any match between a cluster member and one of the outgroup sequences. The resulting list of protein-to-protein connections was exported into a Cytoscape format file (Shannon et al., 2003), using the negative decimal logarithm of the BLAST e-value as the connection weight for the links between gene products. The resulting Cytoscape file is provided as Supplementary file S1.

Cluster annotation and sequence analysis

Clusters of at least three densely-connected genes were identified using the MCODE software (Bader and Hogue, 2003) and the cluster-assignment of each gene was included in the Cytoscape file (Supplementary file S1). A list of clusters with all member proteins in MS-Excel format (Supplementary File S2) forms the basis for the manual cluster annotation effort. For each of the clusters, functional information on the member proteins was obtained from the protein database entry
and the corresponding literature. The subcellular localization of the cluster members was predicted using the DeepLoc software (Almagro Armenteros et al., 2017). A localization entry was added to the cluster table whenever 75% of the USC members had concordant localization predictions. If no functional information was available for any of the cluster members, the protein sequences were subjected to a bioinformatical analysis using BLAST (Altschul et al., 1990), INTERPRO (Blum et al., 2021) TMHMM (Sonnhammer et al., 1998) and HH Pred (Zimmermann et al., 2018). Information obtained from this analysis was manually summarized and added to the cluster annotation file (Supplementary file S2). Phylogenetic tree reconstruction was performed by the neighbor-joining method (Saitou and Nei, 1987), based on multiple alignments generated by the L-INS-I method of the MAFFT package (Katoh and Standley, 2013).
**Figure Legends**

**Figure 1: Network-display of upregulated sequence clusters (USCs)**
The network shows the connected fraction of the upregulated gene list, including 930 out of 1766 *Arabidopsis thaliana* genes (red dots), 906 out of 1019 Glycine max genes (green dots) and 1120 out of 1451 Theobroma cacao genes (blue dots). Lines represent significant BLAST hits, line thickness scales with BLAST significance (-log p-value). Unconnected dots are not shown. A full version of this figure, including gene names and annotations, is provided as Supplementary file S1.

**Figure 2: Upregulation of the PREPIP family of putative peptide precursor genes.**

A: Logarithmic induction factors of individual genes in USC #126 (PREPIP2/3 family, orange/brown color) and USC #266 (PREPIP1 family, blue color) for the different experimental conditions. For Arabidopsis data, different time points after inoculation are shown. For Cocoa data, different pathogens (*Colletotrichum theobromicula*, *Phytophthora palmivora*, *Phytophthora megakarya*) are shown. For Glycine data, the averaged response of susceptible and resistant strains is shown. B: Sequence display of USC #126 members, revealing their repeat structure. The N-terminal signal sequence is shown in red, the C-terminal repeat is highlighted on black and grey background for residues invariant or conservatively replaced in 50% of all sequences. C: Comparison of the C-terminal peptide repeat in the PIP2/3 family (top), PIP1 family (middle) and PCEP family (bottom). Sequence conservation is shown as WebLogo (Crooks et al., 2004), wherein the size of the letters scales with the conservation of the respective amino acids.

**Figure 3: Plant relatives of the NDUFA4 family**

A: Logarithmic induction factors of individual genes of USC #87 under the different experimental conditions. For Arabidopsis data, different time points after inoculation are shown. For Cocoa data, different pathogens (*C. theobromicula*, *P. palmivora*, *P. megakarya*) are shown. For Glycine data, the averaged response of susceptible and resistant strains is shown. B: Sequence conservation in the extended NDUFA4 family. The three first sequences are from humans, the other from plants. Sequence order corresponds to the dendrogram in panel C. Residues that are invariant or conservatively replaced in at least half of the sequences are shown on black or grey background, respectively. C: Dendrogram analysis of NDUFA4 family shows independent gene duplication events in the mammalian and plant lineages. The expression trend of the different clades is shown at the right-hand side.
Literature

Ali, S., Ganai, B.A., Kamili, A.N., Bhat, A.A., Mir, Z.A., Bhat, J.A., Tyagi, A., Islam, S.T., Mushtaq, M., Yadav, P., et al. (2018). Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance. Microbiol Res 212-213, 29-37.

Ali, S.S., Shao, J., Lary, D.J., Strem, M.D., Meinhardt, L.W., and Bailey, B.A. (2017). Phytophthora megakarya and P. palmivora, Causal Agents of Black Pod Rot, Induce Similar Plant Defense Responses Late during Infection of Susceptible Cacao Pods. Front Plant Sci 8, 169.

Almagro Armenteros, J.J., Sonderby, C.K., Sonderby, S.K., Nielsen, H., and Winther, O. (2017). DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 33, 3387-3395.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J Mol Biol 215, 403-410.

Attard, A., Evangelisti, E., Kebdani-Minet, N., Panabieres, F., Deleury, E., Maggio, C., Ponchet, M., and Gourgues, M. (2014). Transcriptome dynamics of Arabidopsis thaliana root penetration by the oomycete pathogen Phytophthora parasitica. BMC Genomics 15, 538.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2.

Balint-Kurti, P. (2019). The plant hypersensitive response: concepts, control and consequences. Mol Plant Pathol 20, 1163-1178.

Bartels, S., and Boller, T. (2015). Quo vadis, Pep? Plant elicitor peptides at the crossroads of immunity, stress, and development. J Exp Bot 66, 5183-5193.

Blair, J.E., Coffey, M.D., Park, S.Y., Geiser, D.M., and Kang, S. (2008). A multi-locus phylogeny for Phytophthora utilizing markers derived from complete genome sequences. Fungal Genet Biol 45, 266-277.

Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., et al. (2021). The InterPro protein families and domains database: 20 years on. Nucleic Acids Res 49, D344-D354.

Boevink, P.C., Birch, P.R.J., Turnbull, D., and Whisson, S.C. (2020). Devastating intimacy: the cell biology of plant-Phytophthora interactions. New Phytol 228, 445-458.

Chepsergon, J., Motaung, T.E., Bellieny-Rabelo, D., and Moleleki, L.N. (2020). Organize, Don't Agonize: Strategic Success of Phytophthora Species. Microorganisms 8.

Colombatti, F., Gonzalez, D.H., and Welch, C. (2014). Plant mitochondria under pathogen attack: a sigh of relief or a last breath? Mitochondrion 19 Pt B, 238-244.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res 14, 1188-1190.

Di, X., Takken, F.L., and Tintor, N. (2016). How Phytohormones Shape Interactions between Plants and the Soil-Borne Fungus Fusarium oxysporum. Front Plant Sci 7, 170.

Dong, S., and Ma, W. (2021). How to win a tug-of-war: the adaptive evolution of Phytophthora effectors. Curr Opin Plant Biol 62, 102027.

Dos Santos-Silva, C.A., Zupin, L., Oliveira-Lima, M., Vilela, L.M.B., Bezerra-Neto, J.P., Ferreira-
Neto, J.R., Ferreira, J.D.C., de Oliveira-Silva, R.L., Pires, C.J., Aburjaile, F.F., et al. (2020). Plant Antimicrobial Peptides: State of the Art, In Silico Prediction and Perspectives in the Omics Era. Bioinform Biol Insights 14, 1177932220952739.

Doungous, O., Minyaka, E., Longue, E.A.M., and Nkengafac, N.J. (2018). Potentials of cocoa pod husk-based compost on Phytophthora pod rot disease suppression, soil fertility, and Theobroma cacao L. growth. Environ Sci Pollut Res Int 25, 25327-25335.

Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30, 207-210.

Endou, M., Yoshida, K., Hirota, M., Nakajima, C., Sakaguchi, A., Komatsubara, N., and Kurihara, Y. (2020). Coxfa4l3, a novel mitochondrial electron transport chain Complex 4 subunit protein, switches from Coxfa4 during spermatogenesis. Mitochondrion 52, 1-7.

Fichman, Y., and Mittler, R. (2020). Rapid systemic signaling during abiotic and biotic stresses: is the ROS wave master of all trades? Plant J 102, 887-896.

Fister, A.S., Mejia, L.C., Zhang, Y., Herre, E.A., Maximova, S.N., and Guiltinan, M.J. (2016). Theobroma cacao L. pathogenesis-related gene tandem array members show diverse expression dynamics in response to pathogen colonization. BMC Genomics 17, 363.

Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. Annu Rev Phytopathol 43, 205-227.

Grunwald, N.J., and Flier, W.G. (2005). The biology of Phytophthora infestans at its center of origin. Annu Rev Phytopathol 43, 171-190.

Hardham, A.R. (2007). Cell biology of plant-oomycete interactions. Cell Microbiol 9, 31-39.

Hoang, N.H., Strogolova, V., Mosley, J.J., Stuart, R.A., and Hosler, J. (2019). Hypoxia-inducible gene domain 1 proteins in yeast mitochondria protect against proton leak through complex IV. J Biol Chem 294, 17669-17677.

Hou, S., Shen, H., and Shao, H. (2019). PAMP-induced peptide 1 cooperates with salicylic acid to regulate stomatal immunity in Arabidopsis thaliana. Plant Signal Behav 14, 1666657.

Hussain, S., Wang, W., Ahmed, S., Wang, X., Adnan, Cheng, Y., Wang, C., Wang, Y., Zhang, N., Tian, H., et al. (2021). PIP2, An Auxin Induced Plant Peptide Hormone Regulates Root and Hypocotyl Elongation in Arabidopsis. Front Plant Sci 12, 646736.

Jeger, M.J., and Pautasso, M. (2008). Comparative epidemiology of zoosporic plant pathogens. European Journal of Plant Pathology 122, 111.

Judelson, H.S., and Blanco, F.A. (2005). The spores of Phytophthora: weapons of the plant destroyer. Nat Rev Microbiol 3, 47-58.

Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30, 772-780.

Le Berre, J.Y., Gourgues, M., Samans, B., Keller, H., Panabieres, F., and Attard, A. (2017). Transcriptome dynamic of Arabidopsis roots infected with Phytophthora parasitica identifies VQ29, a gene induced during the penetration and involved in the restriction of infection. PLoS One 12, e0190341.

Lee, C.Q.E., Kerouanton, B., Chothani, S., Zhang, S., Chen, Y., Mantri, C.K., Hock, D.H., Lim, R., Nadkarni, R., Huynh, V.T., et al. (2021). Coding and non-coding roles of MOCCI (C15orf48) coordinate to regulate host inflammation and immunity. Nat Commun 12, 2130.

Lin, F., Zhao, M., Baumann, D.D., Ping, J., Sun, L., Liu, Y., Zhang, B., Tang, Z., Hughes, E.,
Doerge, R.W., et al. (2014). Molecular response to the pathogen Phytophthora sojae among ten soybean near isogenic lines revealed by comparative transcriptomics. BMC Genomics 15, 18.

Marelli, J.P., Guest, D.I., Bailey, B.A., Evans, H.C., Brown, J.K., Junaid, M., Barreto, R.W., Lisboa, D.O., and Puig, A.S. (2019). Chocolate Under Threat from Old and New Cacao Diseases. Phytopathology 109, 1331-1343.

McDonald, B.A., and Stukenbrock, E.H. (2016). Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. Philos Trans R Soc Lond B Biol Sci 371.

McDowell, J.M., and Dangl, J.L. (2000). Signal transduction in the plant immune response. Trends Biochem Sci 25, 79-82.

Meng, H., Sun, M., Jiang, Z., Liu, Y., Sun, Y., Liu, D., Jiang, C., Ren, M., Yuan, G., Yu, W., et al. (2021). Comparative transcriptome analysis reveals resistant and susceptible genes in tobacco cultivars in response to infection by Phytophthora nicotianae. Sci Rep 11, 809.

Meng, Y., Zhang, Q., Ding, W., and Shan, W. (2014). Phytophthora parasitica: a model oomycete plant pathogen. Mycology 5, 43-51.

Murphy, E., Smith, S., and De Smet, I. (2012). Small signaling peptides in Arabidopsis development: how cells communicate over a short distance. Plant Cell 24, 3198-3217.

Najafi, J., Brembu, T., Vie, A.K., Viste, R., Winge, P., Somssich, I.E., and Bones, A.M. (2020). PAMP-INDUCED SECRETED PEPTIDE 3 modulates immunity in Arabidopsis. J Exp Bot 71, 850-864.

Naveed, Z.A., Wei, X., Chen, J., Mubeen, H., and Ali, G.S. (2020). The PTI to ETI Continuum in Phytophthora-Plant Interactions. Front Plant Sci 11, 593905.

Ngou, B.P.M., Jones, J.D.G., and Ding, P. (2021). Plant immune networks. Trends Plant Sci.

Ogilvie, H.A., Imin, N., and Djordjevic, M.A. (2014). Diversification of the C-TERMINALLY ENCODED PEPTIDE (CEP) gene family in angiosperms, and evolution of plant-family specific CEP genes. BMC Genomics 15, 870.

Pandey, V., Awasthi, M., Singh, S., Tiwari, S., and Dwivedi, U. (2017). A Comprehensive Review on Function and Application of Plant Peroxidases. Biochemistry & Analytical Biochemistry 6.

Petre, B., and Kamoun, S. (2014). How do filamentous pathogens deliver effector proteins into plant cells? PLoS Biol 12, e1001801.

Petri, T., Berchtold, E., Zimmer, R., and Friedel, C.C. (2012). Detection and correction of probe-level artefacts on microarrays. BMC Bioinformatics 13, 114.

Pieterse, C.M., Van der Does, D., Zamioudis, C., Leon-Reyes, A., and Van Wees, S.C. (2012). Hormonal modulation of plant immunity. Annu Rev Cell Dev Biol 28, 489-521.

Rompler, K., Muller, T., Juris, L., Wissel, M., Vukotic, M., Hofmann, K., and Deckers, M. (2016). Overlapping Role of Respiratory Supercomplex Factor Rcf2 and Its N-terminal Homolog Rcf3 in Saccharomyces cerevisiae. J Biol Chem 291, 23769-23778.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4, 406-425.

Scholthof, K.B. (2003). One foot in the furrow: linkages between agriculture, plant pathology, and public health. Annu Rev Public Health 24, 153-174.

Sels, J., Mathys, J., De Coninck, B.M., Cammue, B.P., and De Bolle, M.F. (2008). Plant
pathogenesis-related (PR) proteins: a focus on PR peptides. Plant Physiol Biochem 46, 941-950.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13, 2498-2504.

Shen, J., Diao, W., Zhang, L., Acharya, B.R., Wang, M., Zhao, X., Chen, D., and Zhang, W. (2020). Secreted Peptide PIP1 Induces Stomatal Closure by Activation of Guard Cell Anion Channels in Arabidopsis. Front Plant Sci 11, 1029.

Sonnhammer, E.L., and Koonin, E.V. (2002). Orthology, paralogy and proposed classification for paralog subtypes. Trends Genet 18, 619-620.

Sonnhammer, E.L., von Heijne, G., and Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 6, 175-182.

Sorouri, M., Chang, T., Jesudhasan, P., Pinkham, C., Elde, N.C., and Hancks, D.C. (2020). Signatures of host-pathogen evolutionary conflict reveal MISTR-A conserved Mitochondrial STress Response network. PLoS Biol 18, e3001045.

Stintzi, A., Heitz, T., Prasad, V., Wiedemann-Merdinoglu, S., Kauffmann, S., Geoffroy, P., Legrand, M., and Fritig, B. (1993). Plant 'pathogenesis-related' proteins and their role in defense against pathogens. Biochimie 75, 687-706.

Tello, D., Balsa, E., Acosta-Iborra, B., Fuertes-Yebra, E., Elorza, A., Ordonez, A., Corral-Escariz, M., Soro, I., Lopez-Bernardo, E., Perales-Clemente, E., et al. (2011). Induction of the mitochondrial NDUFA4L2 protein by HIF-1alpha decreases oxygen consumption by inhibiting Complex I activity. Cell Metab 14, 768-779.

UniProt, C. (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49, D480-D489.

Vie, A.K., Najafi, J., Liu, B., Winge, P., Butenko, M.A., Hornslien, K.S., Kumpf, R., Aalen, R.B., Bones, A.M., and Brembu, T. (2015). The IDA/IDA-LIKE and PIP/PIP-LIKE gene families in Arabidopsis: phylogenetic relationship, expression patterns, and transcriptional effect of the PIPL3 peptide. J Exp Bot 66, 5351-5365.

Wise, R.P., Moscou, M.J., Bogdanove, A.J., and Whitham, S.A. (2007). Transcript profiling in host-pathogen interactions. Annu Rev Phytopathol 45, 329-369.

Wrzaczek, M., Vainonen, J.P., Stael, S., Tsiatsiani, L., Help-Rinta-Rahko, H., Gauthier, A., Kaufholdt, D., Bollhoner, B., Lamminmaki, A., Staes, A., et al. (2015). GRIM REAPER peptide binds to receptor kinase PRK5 to trigger cell death in Arabidopsis. EMBO J 34, 55-66.

Zarattini, M., Farjad, M., Launay, A., Cannella, D., Soulie, M.C., Bernacchia, G., and Fagard, M. (2021). Every cloud has a silver lining: how abiotic stresses affect gene expression in plant-pathogen interactions. J Exp Bot 72, 1020-1033.

Zhang, B., Van Aken, O., Thatcher, L., De Clercq, I., Duncan, O., Law, S.R., Murcha, M.W., van der Merwe, M., Seifi, H.S., Carrie, C., et al. (2014). The mitochondrial outer membrane AAA ATPase AtOM66 affects cell death and pathogen resistance in Arabidopsis thaliana. Plant J 80, 709-727.

Zhang, H., Han, Z., Song, W., and Chai, J. (2016). Structural Insight into Recognition of Plant Peptide Hormones by Receptors. Mol Plant 9, 1454-1463.

Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kubler, J., Lozajic, M., Gabler, F., Soding, J., Lupas, A.N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol 430, 2237-2243.
Zong, S., Wu, M., Gu, J., Liu, T., Guo, R., and Yang, M. (2018). Structure of the intact 14-subunit human cytochrome c oxidase. Cell Res 28, 1026-1034.

Zribi, I., Ghorbel, M., and Brini, F. (2020). Pathogenesis Related Proteins (PRs): From Cellular Mechanisms to Plant Defense. Curr Protein Pept Sci.

**Statements & Declarations**

**Funding**
C.S. was supported by grants from DAAD (57314019) and TWAS-DFG (HO 3783/2-1).

**Competing Interest**
The authors have no relevant financial or non-financial interests to disclose.

**Authors Contributions**
Both authors jointly initiated the study, performed the analyses and wrote the manuscript.

**Data Availability**
Input transcriptomics data are available from the GEO database (https://www.ncbi.nlm.nih.gov/geo) under the accession numbers given in the text. All relevant data generated and used in this project are provided as supplementary data files with the manuscript.
| PR  | Description                               | USC # | members |      |      |      |      |      |      |
|-----|-------------------------------------------|-------|---------|------|------|------|------|------|------|
|     |                                           |       | Arabidopsis | Soybean | Cocoa |
| PR-1| CAP/SCP superfamily                       | 13    | 4       | 9    | 4    |      |      |      |      |
| PR-2| β-1,3-glucanase                           | 32    | 1       | 6    | 4    |      |      |      |      |
|     |                                           | 209   | 3       | 0    | 0    |      |      |      |      |
| PR-3| Chitinase–type I, II, IV, V, VI, VII      | 50    | 3       | 2    | 4    |      |      |      |      |
|     |                                           | 150   | 2       | 2    | 4    |      |      |      |      |
| PR-4| Chitinase-Hevein-like                     | 52    | 0       | 3    | 6    |      |      |      |      |
| PR-5| Thaumatin-like                            | 39    | 0       | 5    | 5    |      |      |      |      |
|     |                                           | 188   | 0       | 3    | 1    |      |      |      |      |
| PR-6| Proteinase-inhibitor                      | 33    | 2       | 3    | 6    |      |      |      |      |
| PR-7| Endoproteinase                            | 117   | 4       | 0    | 1    |      |      |      |      |
| PR-8| Chitinase-type III                        | 138   | 0       | 2    | 3    |      |      |      |      |
| PR-9| Peroxidase                                | 73    | 4       | 4    | 1    |      |      |      |      |
|     |                                           | 7     | 3       | 8    | 11   |      |      |      |      |
|     |                                           | 197   | 0       | 1    | 3    |      |      |      |      |
|     |                                           | 122   | 1       | 3    | 1    |      |      |      |      |
|     |                                           | 273   | 2       | 0    | 1    |      |      |      |      |
| PR-10| Ribonuclease-like                         | 8     | 0       | 10   | 9    |      |      |      |      |
| PR-14| Lipid-transfer Protein                    | 260   | 1       | 0    | 2    |      |      |      |      |
| PR-16| Germin-like / Oxalate Oxidase-like        | 6     | 7       | 3    | 12   |      |      |      |      |
| PR-17| Putative Zinc-metallo-proteinase          | 135   | 0       | 3    | 2    |      |      |      |      |

**Table 1: Pathogenesis-related (PR) gene clusters found in this study**

With the exception of PR-11, PR-12, PR-13, and PR-15, all established pathogenesis-related (PR) gene families (Ali et al., 2018) are represented by upregulated sequence clusters. PR-families with high sequence diversity were found split into multiple USCs. Besides the PR name and description, this table shows the number of identified genes from the Arabidopsis, Glycine and Soybean data. A full version of this table, including the gene names, is provided as Supplementary file S3.
Supplementary files

Suppl. File S1: Cytoscape Network of all USCs.
Suppl. File S2: Master Table of all annotated USCs
Suppl. File S3: Full table of PR-genes
Suppl. File S4: Preprocessed source data of plant transcriptomics projects.
Figure 1

Network-display of upregulated sequence clusters (USCs) The network shows the connected fraction of the upregulated gene list, including 930 out of 1766 Arabidopsis thaliana genes (red dots), 906 out of 1019 Glycine max genes (green dots) and 1120 out of 1451 Theobroma cacao genes (blue dots). Lines
represent significant BLAST hits, line thickness scales with BLAST significance (-log p-value). Unconnected dots are not shown. A full version of this figure, including gene names and annotations, is provided as Supplementary file S1.

Figure 2

Upregulation of the PREPIP family of putative peptide precursor genes. A: Logarithmic induction factors of individual genes in USC #126 (PREPIP2/3 family, orange/brown color) and USC #266 (PREPIP1 family, blue color) for the different experimental conditions. For Arabidopsis data, different time points after inoculation are shown. For Cocoa data, different pathogens (Colletotrichum theobromicula, Phytophthora palmivora, Phytophthora megakarya) are shown. For Glycine data, the averaged response of susceptible
and resistant strains is shown. B: Sequence display of USC #126 members, revealing their repeat structure. The N-terminal signal sequence is shown in red, the C-terminal repeat is highlighted on black and grey background for residues invariant or conservatively replaced in 50% of all sequences. C: Comparison of the C-terminal peptide repeat in the PIP2/3 family (top), PIP1 family (middle) and PCEP family (bottom). Sequence conservation is shown as WebLogo (Crooks et al., 2004), wherein the size of the letters scales with the conservation of the respective amino acids.
Plant relatives of the NDUFA4 family A: Logarithmic induction factors of individual genes of USC #87 under the different experimental conditions. For Arabidopsis data, different time points after inoculation are shown. For Cocoa data, different pathogens (C. theobromicula, P. palmivora, P. megakarya) are shown. For Glycine data, the averaged response of susceptible and resistant strains is shown. B: Sequence conservation in the extended NDUFA4 family. The three first sequences are from humans, the other from plants. Sequence order corresponds to the dendrogram in panel C. Residues that are invariant or conservatively replaced in at least half of the sequences are shown on black or grey background, respectively. C: Dendrogram analysis of NDUFA4 family shows independent gene duplication events in the mammalian and plant lineages. The expression trend of the different clades is shown at the right-hand side.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.

- S1USCnetwork.cys
- S2allUSCs.xlsx
- S3fullPRtable.xlsx
- S4preprocessedsourcedata.xlsx