

# 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](mailto:kay.hofmann@uni-koeln.de))

University of Cologne <https://orcid.org/0000-0002-2289-9083>

---

## Research Article

**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](#)

---

1 **Mining the conserved transcriptional response to phytophthora infection for factors be-**  
2 **yond the known PR genes**

3

4 Claude Simo<sup>1</sup>, Kay Hofmann<sup>2</sup>

5

6

7

8 (1) Department of Plant Biology, Faculty of Science, University of Douala. PO Box 24157 Douala-  
9 Cameroon

10 (2) Institute for Genetics, University of Cologne, Zùlpicher StraÙe 47a, D-50674 Cologne, Ger-  
11 many. ORCID: **0000-0002-2289-9083**

12

13

14

15

16 Short title: Transcriptional Response to Phytophthora

17

18

19

20

21 Correspondence should be addressed to:

22 Kay Hofmann (kay.hofmann@uni-koeln.de)

23 Tel: +49-(0)221 470-1701

24 Fax: +49-(0)221 470-3402

25

26

27 **Key Message**

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

31

1

## 2 **Abstract**

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

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

20

21

## 22 **Introduction**

23 Domestication of flowering plants forms the basis of a productive and sustainable agriculture that  
24 is under constant challenge by the attack of emerging pathogens (Scholthof, 2003; McDonald and  
25 Stukenbrock, 2016). Most species in the genus *Phytophthora* of oomycetes are devastating plant  
26 pathogens that cause severe attacks on many wild and cultivated plants (Grunwald and Flier,  
27 2005; Meng et al., 2014; Marelli et al., 2019; Boevink et al., 2020). In general, *Phytophthora* spe-  
28 cies have a very flexible life cycle and ability to adapt to changing environmental conditions (Jeger  
29 and Pautasso, 2008; Naveed et al., 2020). Today, there are more than 100 species of the genus  
30 *Phytophthora* with varying host ranges; some of these species like *P. sojae* and *P. megakarya*  
31 are restricted to a single known species, while others like *P. capsici*, *P. ramorum*, *P. parasitica*  
32 and *P. palmivora* can infect a wide range of plants from several genera (Blair et al., 2008; Meng  
33 et al., 2014). A particular threat to cocoa plantations in West and Central Africa is *P. megakarya*,  
34 a restrictive but very aggressive species, while infections of cocoa by *P. palmivora* are more  
35 widespread around the world, causing significant crop losses (Doungous et al., 2018; Marelli et  
36 al., 2019).

37 *Phytophthora* species share a hemibiotrophic lifestyle. Their asexual life cycle in plants is charac-  
38 terized by the adhesion of mobile zoospores to the host tissue, followed by encystment and germ

1 tube formation (Boevink et al., 2020; Chepsergon et al., 2020) Entry into the plant is achieved by  
2 surface appressoria and is followed by establishment of an apoplastic hyphal network. At the  
3 biotrophic stage, *Phytophthora* produce haustoria in plant cells. These structures participate in  
4 the acquisition of nutrients and release virulence proteins known as effectors (Petre and Kamoun,  
5 2014; Dong and Ma, 2021). This is followed by the necrotrophic stage, characterized by tissue  
6 necrosis of the host and the production of numerous sporangia that release zoospores (Judelson  
7 and Blanco, 2005; Hardham, 2007; Boevink et al., 2020).

8 During this infection stage, the host plant can sense the pathogen, either directly or by virtue of  
9 its activity, and then mount a more or less targeted defense response. As the first line of defense,  
10 pathogen-associated molecular patterns (PAMPs) are recognized by plant surface receptors,  
11 which then initiate a defense response called PAMP-triggered immunity (PTI) (Naveed et al., 2020;  
12 Ngou et al., 2021). Most pathogens, including *Phytophthora*, secrete a number of effectors into  
13 the host cell, which interfere with the PTI response. (Petre and Kamoun, 2014; Chepsergon et al.,  
14 2020; Dong and Ma, 2021). These effectors, in turn, are recognized by specific resistance genes  
15 or R genes, which initiate a second branch of the immune system called effector-triggered im-  
16 munity (ETI) (Naveed et al., 2020; Ngou et al., 2021). One consequence of ETI activation is the  
17 hypersensitive response (HR), a form of programmed cell death surrounding the infection site,  
18 which deprives biotrophic pathogens of a food source and prevents their spread (Balint-Kurti,  
19 2019). Another consequence of ETI activation is the transcriptional induction of a wide range of  
20 host genes coding for proteins that either fight the pathogen directly or modify the host cell and  
21 its environment to limit pathogen spread. The genes induced by this pathway are often effective  
22 against multiple classes of pathogens and have been classified into different families of patho-  
23 genesis-related genes (PR-genes) (Stintzi et al., 1993; Sels et al., 2008; Ali et al., 2018; Zribi et  
24 al., 2020).

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

34 The identification of pathogen-responsive host genes by transcriptomics methods has a long tra-  
35 dition (Wise et al., 2007) and was instrumental for defining the PR-gene classes (Stintzi et al.,  
36 1993). While RNA-sequencing gradually supersedes the traditional microarray-based techniques,  
37 host transcriptomics continues to be a valuable approach to study host-pathogen interactions,  
38 leading to the identification of new response genes (Le Berre et al., 2017) or helping with the  
39 identification of resistance traits (Meng et al., 2021). The aim of the current study was to mine  
40 plant transcriptomics data for novel and unconventional regulatory events that broaden our un-  
41 derstanding of the host response to phytophthora infections. Virtually every dataset measuring

1 the transcriptional response to pathogens contains a number of upregulated genes that do not  
2 belong any of the known PR-classes and don't have a (predicted) function connected to the de-  
3 fense response – if they have any annotated function at all. Some of these observations might be  
4 due to artefacts (Petri et al., 2012), but others might represent unexpected but genuine regulations.  
5 In order to reliably identify novel and biologically relevant regulations over a background of tran-  
6 scriptional noise, we have applied a bioinformatical pipeline that searches for conserved regula-  
7 tion events in small gene families from different species, measured by different transcriptomics  
8 platforms. The gene families targeted by this approach may either be groups of orthologs, or  
9 include closely related paralogs as they are often observed in lineage-specific gene duplications  
10 (Sonnhammer and Koonin, 2002). Requiring a conserved regulation mode should deplete spuri-  
11 ous false-positive observations and emphasize gene families, whose pathogen-responsive induc-  
12 tion is biologically meaningful and has been selected for by evolution.

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

22

23

## 24 **Results**

### 25 **Identification of Phytophthora-induced genes**

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

35 Not only do the selected datasets differ in the host species, they also use a diverse set of infection  
36 conditions, time points post inoculation, and measurement modalities. To reach a common frame  
37 of reference for inter-species comparisons, a unified 'combined logarithmic induction factor' (CLI-  
38 factor) was established for each host species. The idea of this combined factor was to select all

1 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  
2 log-induction values of replicate experiments were averaged, and the remaining N conditions  
3 were considered orthogonal coordinates in an N-dimensional 'regulation space'. The species-  
4 specific CLI factor is the length of each gene's regulation vector. In general, genes surpassing a  
5 CLI factor of 4.0 were considered as strongly induced and were included in the multi-species  
6 clustering analysis. This condition was met by 1766 *Arabidopsis thaliana* genes, 1019 genes from  
7 *Glycine max*, and 1451 genes from *Theobroma cacao*.  
8

## 9 **Definition of a conserved induction signature**

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

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

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

36 The majority of identified clusters encompass sequences from multiple species, including 125  
37 clusters containing putative orthologs from all three plants and 141 clusters with sequences from  
38 two plants. The remaining 44 mono-species clusters, including 18 from *Arabidopsis*, 3 from *Glycine*  
39 and 23 from *Theobroma*, are either derived from transposons or represent closely related  
40 inparalogs resulting from lineage-specific gene duplication events (Sonnhammer and Koonin,

1 2002). The upregulated gene families code for proteins with known or predicted localizations in  
2 all major subcellular compartments. Particularly prominent are families localized in the cytoplasm  
3 (85 clusters), extracellular space (54 clusters), plasma membrane (43 clusters), nucleus (39 clus-  
4 ters), ER/Golgi (21 clusters) and mitochondria (18 clusters).

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

## 17 **Strong induction of secreted peptide precursors**

18 Two of the strongest upregulated clusters comprise short proteins with a predicted extracellular  
19 localization, which might act as precursors for signaling or defense peptides. Cluster #126, which  
20 consists of the two *Arabidopsis* genes (AT2G23270 and AT4G37290), one *Theobroma* gene  
21 (Tc01\_p005760) and two *Glycine* genes (glyma05g21360 and glyma17g18230) showed a partic-  
22 ularly strong and consistent 50 to 300-fold upregulation in the different infection models (Figure  
23 2). There were no major differences between the different time points in the *Arabidopsis* model  
24 (Attard et al., 2014), between *P. palmivora* and *P. megakarya* infections in the *Theobroma* model  
25 (Ali et al., 2017), or between resistant and susceptible Soybean lineages (Lin et al., 2014) (Figure  
26 2a). The two *Arabidopsis* members of this cluster have been described previously as the source  
27 of the PAMP-induced peptides PIP2 and PIP3; the corresponding gene names are PREPIP2 and  
28 PREPIP3 (Vie et al., 2015; Najafi et al., 2020; Hussain et al., 2021). A third published PIP-encod-  
29 ing gene, AT4G28460/PREPIP1 (Hou et al., 2019; Shen et al., 2020), is more divergent and there-  
30 fore not a member of the upregulated PREPIP2/3 sequence cluster. Instead, it forms a separate  
31 USC #266 together with two *Theobroma* genes (Tc04\_p018060 and Tc04\_p018070), which are  
32 also transcriptionally induced, albeit not as strongly as PREPIP2 and 3 (Figure 2a).

33 Our bioinformatical analysis shows that members of USC #126, including *Arabidopsis* PREPIP2  
34 /3 are members of a protein family found in all major dicotyledon lineages, which is characterized  
35 by a conserved C-terminal ~22aa repeat that is present in two, four or ten tandem copies in the  
36 cluster members from *Arabidopsis*, *Glycine* and *Theobroma*, respectively (Figure 2b). Each re-  
37 peat copy ends on His or Asn, a property typical of receptor-binding plant hormones (Zhang et al.,  
38 2016). The highly conserved GPSP-motif is shared with the PREPIP1 family (Hou et al., 2019)  
39 and is similar to the GHSP-motif found in the PCEP family (Ogilvie et al., 2014), suggesting that  
40 these families are distant relatives (Figure 2c). The repeat architecture of the PREPIP2/3 family,

1 which is absent from PREPIP1 and PCEP, and the strong response to phytophthora infection  
2 suggest that these proteins are particularly relevant for signaling an infection by oomycetes.

### 3 **A new family of mitochondrial anti-ROS protection factors**

4 An unexpected, but particularly interesting finding was the strong upregulation of sequence clus-  
5 ter #87, comprising a number of short mitochondrial proteins. A detailed bioinformatical analysis  
6 of this family revealed that these proteins are distant relatives of the NDUFA4 subunit of the mi-  
7 tochondrial complex IV, also known as the cytochrome c oxidase complex (Figure 3a). Plant mem-  
8 bers of the NDUFA4 family had not been described so far. All land plants encode multiple mem-  
9 bers of the NDUFA4-like family, but in the three plant models studied here, only selected family  
10 members are strongly upregulated by phytophthora and are thus included in USC #87 (Figure  
11 3b,c). The sequences of the plant NDUFA4 homologs have substantially diverged from their ani-  
12 mal counterparts (Figure 3a), and the dendrogram analysis suggests that the diversity of the plant  
13 members (2 genes in *Arabidopsis*, 3 genes in *Theobroma*, and 7 genes in *Glycine*) arose after  
14 the split between the plant and animal kingdoms (Figure 3a,c).

15 Four of the plant genes (AT3G29970, Tc00\_p011650, glyma08g27620 and glyma18g50800) are  
16 particularly responsive to *Phytophthora* infection and form a clade in the dendrogram (Figure 3c).  
17 A second, less strongly induced clade is formed by the USC #87 members Tc02\_p026410 and  
18 glyma18g01340, together with glyma11g37370. The latter gene is not annotated as part of cluster  
19 #87, since its upregulation factor did not quite meet the CLI >4.0 criterion. A third clade is formed  
20 by the constitutively expressed family members AT3G48140, Tc03\_p000330, glyma04g34840  
21 and glyma06g19850, which are not induced by *Phytophthora* and thus not members of any USC  
22 (Figure 3c). Mammalian genomes also contain multiple members of the NDUFA4 family, which  
23 result from independent gene duplication events (Figure 3c). Interestingly, human NDUFA4 is  
24 constitutively expressed, while its two paralogs NDUFA4L2 and C15ORF48 (also known as  
25 MOCCI or MISTR-A) are upregulated in response to hypoxia and infections, respectively (Tello et  
26 al., 2011; Sorouri et al., 2020; Lee et al., 2021). Thus, the acquisition of a pathogen-responsive  
27 NDUFA4L2 paralog has happened at least two times during the evolution of plants and animals  
28 and appears to protect mitochondria from excessive ROS production during the defense response.

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

35

## 36 **Discussion**

37 The analysis of the transcriptional response of plants to an infection situation has been instru-  
38 mental for the discovery of many genes that are actively involved in the defense response (Wise

1 et al., 2007). Some of gene regulation events are initiated by the sensing of the pathogen itself  
2 via PAMPs (pathogen-associated molecular patterns), but other gene inductions are caused by  
3 the pathogen-inflicted damage and stresses and thus overlap with abiotic stress responses  
4 (Zarattini et al., 2021). A recurring problem with transcriptomics-based discovery of genes belong-  
5 ing to a particular biological response pathway is the presence of spurious signals – in particular  
6 when searching for ‘novel’ or ‘unexpected’ regulation events. The current study addresses this  
7 problem by comparing multiple datasets obtained from different species in similar infection situa-  
8 tions. By focusing on gene families with consistent up-regulation, spurious data are efficiently  
9 depleted since they are unlikely to be found in multiple species, measured by different tran-  
10 scriptomics platforms. A downside inherent to this approach is the likely loss of species-specific  
11 regulations, or those that only occur in particular experimental setups. Nevertheless, the ‘con-  
12 servative’ criteria applied in the current study resulted in 310 upregulated sequence clusters  
13 (USCs), 266 of which are supported by multiple species, the other 44 by multiple paralogs within  
14 one species. The bioinformatical analysis of the cluster members revealed several new and un-  
15 expected regulations, some of which offer new insights into the plant response to phytophthora  
16 infection.

17 Secreted proteins are frequently induced upon biotic plant stress and form the basis of many  
18 established pathogenesis-related (PR) gene classes (Stintzi et al., 1993; Sels et al., 2008; Ali et  
19 al., 2018; Zribi et al., 2020). A large fraction of them are enzymes such as endo-glucanases,  
20 chitinases, peroxidases, ribonucleases or proteases; they either attack the pathogen directly or  
21 limit pathogen motility and entry by reshaping the plant cell wall and extracellular matrix. Other  
22 PR-genes encode secreted inhibitors and other non-enzymatic factors, such as the defensins and  
23 thaumatin-like proteins, which also directly counteract the pathogen (Dos Santos-Silva et al.,  
24 2020). Less well characterized is the defense role of plant signaling peptides, which do not attack  
25 the pathogen but rather transmit a signal to neighboring or distant cells in the plant. While several  
26 secreted peptides and their recognition by cognate surface receptors have been characterized  
27 (Murphy et al., 2012; Zhang et al., 2016), most of them appear to have a role in plant development  
28 and growth regulation. A notable exception is the family of plant elicitor peptides (PEPs), which  
29 are derived from intracellular precursor proteins (PROPEPs) and - upon infection – become se-  
30 creted via cell rupture or another non-classical secretion mechanism (Bartels and Boller, 2015).  
31 There are also reports on a defense role of PIP1 and PIP3, members of another family of peptide  
32 precursor proteins (Hou et al., 2019; Najafi et al., 2020; Shen et al., 2020). Together with PIP2, a  
33 peptide hormone regulating root elongation (Hussain et al., 2021), the precursor proteins of this  
34 family enter the secretory pathway by an N-terminal signal peptide and are thought to release a  
35 C-terminal peptide hormone, which is distantly related to the functionally diverse CEP (C-termi-  
36 nally encoded peptide) family (Ogilvie et al., 2014) (figure 2c). Our analysis revealed that in par-  
37 ticular the PIP2/PIP3 family, but also the PIP1 family, are strongly induced by phytophthora infec-  
38 tion in all three tested plant species. Moreover, as shown in figure 2b, the precursor proteins  
39 PROPIP2 and PROPIP3 contain a C-terminal repeat, suggesting that multiple PIP2- and PIP3-  
40 like peptides can be released from these precursors. In particular the cocoa protein  
41 Tc01\_p005760 is predicted to generate no less than ten such peptides – all of them related, but  
42 with distinct differences, making it likely that they can bind to multiple receptor types (Zhang et al.,

1 2016) and possibly sensitize non-infected cells for an impending assault by invading phytophthora.  
2 The annotated data compendium ([Supplementary file S2](#)) can easily be searched for other pre-  
3 dicted extracellular response factors. Besides a large number of predicted enzymes, inhibitors,  
4 and lipid transfer proteins, several USCs contain additional *bona fide* plant signaling peptides –  
5 even if they are less strongly induced than the PIP clusters. Cluster #267 and #225 contain se-  
6 lected members of the CEP family (Ogilvie et al., 2014), cluster #192 contains proteins related to  
7 GRIM REAPER, a peptide signaling cell death via a receptor kinase (Wrzaczek et al., 2015), while  
8 cluster #98 codes for other short secreted proteins without a known function or informative ho-  
9 mologies. The study of extracellular defense peptides and their receptors might open new ave-  
10 nues to treating or protecting plants by administration of artificial peptides or peptide-mimetics.

11 Unlike extracellular proteins, those with a mitochondrial localization have rarely been implicated  
12 in the transcriptional response of plants to pathogens. A notable exception is the mitochondrial  
13 ATPase AtOM66 (Zhang et al., 2014), which was also identified in the present analysis as part of  
14 USC #21. Overall, 14 clusters were unambiguously predicted to consist of mitochondrial proteins,  
15 most of which have no functional annotation or informative homologs whatsoever ([Supplementary](#)  
16 [file S2](#)). One of these clusters, USC #87, was shown to be related to mammalian NDUFA4-like  
17 proteins, and is particularly strongly induced by phytophthora infections ([figure 3](#)). Bioinformatical  
18 analysis found that the plant family contains additional members, which are not pathogen-induced  
19 and therefore not covered by our clustering effort. As an example, the data in (Attard et al., 2014)  
20 show for the arabidopsis gene AT3G48140 a constitutively high expression level with negligible  
21 regulation (0.95 to 1.2-fold), while its paralog AT3G29970 has a low background expression but  
22 is 42-fold induced upon phytophthora treatment. Interestingly, the human NDUFA4 family also  
23 contains multiple paralogs with fundamentally different expression behavior; a number of recent  
24 studies have shown that their transcriptional regulation is directly linked to their biological function  
25 (Endou et al., 2020; Sorouri et al., 2020; Lee et al., 2021). Human NDUFA4 itself is a constitutively  
26 expressed structural component of the mitochondrial cytochrome c oxidase complex, also known  
27 as complex IV (Zong et al., 2018). One of its paralogs, NDUFA4L2, is upregulated under hypoxic  
28 conditions and thought to limit ROS production under low-oxygen conditions, most likely by re-  
29 placing NDUFA4 in the respiratory chain (Tello et al., 2011). Another paralog, originally called  
30 C15ORF48 or NMES1 and recently renamed to MOCCI or MISTR-A, is upregulated under con-  
31 ditions of infection and inflammation (Sorouri et al., 2020; Lee et al., 2021). Like NDUFA4L2,  
32 MOCCI/MISTR-A can replace NDUFA4 in the cytochrome c oxidase complex, leading to a reduc-  
33 tion in respiratory chain activity and ROS production. The proposed role of this regulation is to  
34 protect the host cell from excessive ROS production that happens during the immune response  
35 (Sorouri et al., 2020; Lee et al., 2021). While no experimental data are available for the plant  
36 proteins, it is reasonable to assume a similar function for AT3G29970 and the other members of  
37 USC #21, given that ROS production by mitochondria offers both chances and dangers for the  
38 host cell (Colombatti et al., 2014; Fichman and Mittler, 2020). The role of mitochondrial protection  
39 in the defense response appears underappreciated, given that similar mechanisms have evolved  
40 independently in the animal and plant lineages.

41 The identification of phytophthora-induced peptides and mitochondrial protection factors are just

1 two examples for discoveries made possible by a conservative meta-analysis of existing tran-  
2 scriptomics data. The comprehensive cluster table provided as [Supplementary file S2](#) contains  
3 many additional clusters of phytophthora-responsive genes in multiple species, whose exact bio-  
4 logical function remains to be uncovered. The bioinformatical annotations added to the clusters  
5 provide a first step into this direction.

6  
7

## 8 **Materials & Methods**

### 9 **Data sources and pre-processing**

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

12 **Arabidopsis:** A dataset measuring the response of *A. thaliana* roots in a compatible interaction  
13 with *Phytophthora parasitica* was downloaded from GEO project GSE20226 (Attard et al., 2014).  
14 The dataset is based on microarray data using the Affymetrix ATH1 genome array and comprises  
15 an uninfected control and samples taken at different time points after infection (2.5h, 6h, 10.5h,  
16 30h), each of them in two replicates. For the current analysis, averaged log-2 induction factors  
17 (IF) relative to the untreated sample were calculated for each time point. The CLI score was de-  
18 rived by considering the log-induction values of the four time points as independent orthogonal  
19 coordinates and calculating the length of the 4-dimensional ‘induction vector’:

$$20 \quad CLI_{Arabidopsis} = \sqrt{IF_{2.5h}^2 + IF_{6h}^2 + IF_{10.5h}^2 + IF_{30h}^2}$$

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

23 **Soybean:** A dataset measuring the response of soybean (*Glycine max*) to infection with *Phy-*  
24 *tophthora sojae* was downloaded from GEO project GSE48524 (Lin et al., 2014). The dataset is  
25 based on RNA-Seq data using the Illumina HiScanSQ platform and comprises data measured  
26 24h post inoculation of one susceptible soybean line (“Williams”) and 10 near-isogenic lines car-  
27 rying different resistance determinants (Lin et al., 2014). For the current analysis, averaged log-2  
28 upregulation factors (relative to the matched mock-infected sample) were calculated for each soy-  
29 bean line and an average of the ten resistant lines was formed. The CLI score was derived by  
30 considering the log-induction values of sensitive and resistant lines as independent orthogonal  
31 coordinates and calculating the length of the 2-dimensional ‘induction vector’:

$$32 \quad CLI_{Glycine} = \sqrt{IF_{sensitive}^2 + IF_{resistant}^2}$$

33 A CLI score of 4.0 was reached by 1019 genes and considered sufficient for including the gene  
34 in the multi-species clustering analysis ([supplementary file S4](#)).

1 **Cocoa:** A dataset measuring the response of cocoa (*Theobroma cacao*) to infection with *Phy-*  
2 *tophthora palmivora* was downloaded from GEO project GSE73804 (Fister et al., 2016). The da-  
3 taset is based on the NimbleGen T. cacao 28k array and comprises data measured 72h post  
4 inoculation in five-fold replicates. For the current analysis, an averaged log-2 upregulation factors  
5 (relative to the water-inoculated samples) was calculated. A second, independent dataset meas-  
6 uring the response of susceptible cocoa pods to infection with *Phytophthora palmivora* and *Phy-*  
7 *tophthora megakarya* was downloaded from the data supplement of (Ali et al., 2017). This dataset  
8 is based on RNA-Seq data using the Illumina HiSeq2000 platform and provides normalized RPKM  
9 (Reads Per kilobase of transcript per million mapped reads) values for infected and uninfected  
10 plants, averaged from three replicates (Ali et al., 2017). After averaging the replicates, the CLI  
11 score was calculated from the three orthogonal datasets as described above. Since the Fister et  
12 al data (Fister et al., 2016) showed considerably less spread than the two Ali et al data sets (Ali  
13 et al., 2017), the former was upweighted two-fold in the calculation of the CLI-score to make these  
14 heterogenous datasets better comparable:

$$15 \quad CLI_{Theobroma} = \sqrt{(2 * IF_{P.pal\_Fister})^2 + IF_{P.pal\_Ali}^2 + IF_{P.meg\_Ali}^2}$$

16 As shown in [supplementary file S4](#), overall 1451 cocoa genes reached a CLI score of 4.0 and  
17 were included in the multi-species clustering analysis.

## 18 **Sequence clustering of upregulated genes**

19 The top-ranking gene products from each species (exceeding the species-specific score of 4.0)  
20 were combined in a common fasta-formatted protein sequence file and subjected to all-against-  
21 all similarity searches using BLAST (Altschul et al., 1990). In addition, each protein was also  
22 compared to the proteome of four outgroup species belonging to different taxa (*Selaginella moel-*  
23 *lendorffii*, *Chlamydomonas reinhardtii*, *Saccharomyces cerevisiae* and *Homo sapiens*). All se-  
24 quences were obtained from the Uniprot database (UniProt, 2021). Upregulated Sequence Clus-  
25 ters (USCs) were defined by joining upregulated gene products from *Arabidopsis*, *Glycine* and  
26 *Theobroma* when they fulfil the following criteria: i) cluster members are connected to each other  
27 by significant BLAST scores with E-value < 10<sup>-3</sup>; and ii) the within-cluster BLAST scores are better  
28 (lower E-value) than any match between a cluster member and one of the outgroup sequences.  
29 The resulting list of protein-to-protein connections was exported into a Cytoscape format file  
30 (Shannon et al., 2003), using the negative decimal logarithm of the BLAST e-value as the con-  
31 nection weight for the links between gene products. The resulting Cytoscape file is provided as  
32 [Supplementary file S1](#).

## 33 **Cluster annotation and sequence analysis**

34 Clusters of at least three densely-connected genes were identified using the MCODE software  
35 (Bader and Hogue, 2003) and the cluster-assignment of each gene was included in the Cytoscape  
36 file ([Supplementary file S1](#)). A list of clusters with all member proteins in MS-Excel format ([Sup-](#)  
37 [plementary File S2](#)) forms the basis for the manual cluster annotation effort. For each of the clus-  
38 ters, functional information on the member proteins was obtained from the protein database entry

1 and the corresponding literature. The subcellular localization of the cluster members was pre-  
2 dicted using the DeepLoc software (Almagro Armenteros et al., 2017). A localization entry was  
3 added to the cluster table whenever 75% of the USC members had concordant localization pre-  
4 dictions. If no functional information was available for any of the cluster members, the protein  
5 sequences were subjected to a bioinformatical analysis using BLAST (Altschul et al., 1990), IN-  
6 TERPRO (Blum et al., 2021) TMHMM (Sonnhammer et al., 1998) and HHPRED (Zimmermann  
7 et al., 2018). Information obtained from this analysis was manually summarized and added to the  
8 cluster annotation file ([Supplementary file S2](#)). Phylogenetic tree reconstruction was performed  
9 by the neighbor-joining method (Saitou and Nei, 1987), based on multiple alignments generated  
10 by the L-INS-I method of the MAFFT package (Kato and Standley, 2013).

11

12

## 1 **Figure Legends**

2

### 3 **Figure 1: Network-display of upregulated sequence clusters (USCs)**

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

10

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

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

23

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

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

35

## 1 Literature

- 2
- 3 Ali, S., Ganai, B.A., Kamili, A.N., Bhat, A.A., Mir, Z.A., Bhat, J.A., Tyagi, A., Islam, S.T., Mushtaq,  
4 M., Yadav, P., *et al.* (2018). Pathogenesis-related proteins and peptides as promising tools for  
5 engineering plants with multiple stress tolerance. *Microbiol Res* 212-213, 29-37.
- 6 Ali, S.S., Shao, J., Lary, D.J., Strem, M.D., Meinhardt, L.W., and Bailey, B.A. (2017). *Phytophthora*  
7 *megakarya* and *P. palmivora*, Causal Agents of Black Pod Rot, Induce Similar Plant Defense  
8 Responses Late during Infection of Susceptible Cacao Pods. *Front Plant Sci* 8, 169.
- 9 Almagro Armenteros, J.J., Sonderby, C.K., Sonderby, S.K., Nielsen, H., and Winther, O. (2017).  
10 DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33,  
11 3387-3395.
- 12 Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment  
13 search tool. *J Mol Biol* 215, 403-410.
- 14 Attard, A., Evangelisti, E., Kebdani-Minet, N., Panabieres, F., Deleury, E., Maggio, C., Ponchet,  
15 M., and Gourgues, M. (2014). Transcriptome dynamics of *Arabidopsis thaliana* root penetration  
16 by the oomycete pathogen *Phytophthora parasitica*. *BMC Genomics* 15, 538.
- 17 Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in  
18 large protein interaction networks. *BMC Bioinformatics* 4, 2.
- 19 Balint-Kurti, P. (2019). The plant hypersensitive response: concepts, control and consequences.  
20 *Mol Plant Pathol* 20, 1163-1178.
- 21 Bartels, S., and Boller, T. (2015). Quo vadis, Pep? Plant elicitor peptides at the crossroads of  
22 immunity, stress, and development. *J Exp Bot* 66, 5183-5193.
- 23 Blair, J.E., Coffey, M.D., Park, S.Y., Geiser, D.M., and Kang, S. (2008). A multi-locus phylogeny  
24 for *Phytophthora* utilizing markers derived from complete genome sequences. *Fungal Genet Biol*  
25 45, 266-277.
- 26 Blum, M., Chang, H.Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G.,  
27 Paysan-Lafosse, T., Qureshi, M., Raj, S., *et al.* (2021). The InterPro protein families and domains  
28 database: 20 years on. *Nucleic Acids Res* 49, D344-D354.
- 29 Boevink, P.C., Birch, P.R.J., Turnbull, D., and Whisson, S.C. (2020). Devastating intimacy: the cell  
30 biology of plant-*Phytophthora* interactions. *New Phytol* 228, 445-458.
- 31 Chepsergon, J., Motaung, T.E., Bellieny-Rabelo, D., and Moleleki, L.N. (2020). Organize, Don't  
32 Agonize: Strategic Success of *Phytophthora* Species. *Microorganisms* 8.
- 33 Colombatti, F., Gonzalez, D.H., and Welchen, E. (2014). Plant mitochondria under pathogen  
34 attack: a sigh of relief or a last breath? *Mitochondrion* 19 Pt B, 238-244.
- 35 Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo  
36 generator. *Genome Res* 14, 1188-1190.
- 37 Di, X., Takken, F.L., and Tintor, N. (2016). How Phytohormones Shape Interactions between  
38 Plants and the Soil-Borne Fungus *Fusarium oxysporum*. *Front Plant Sci* 7, 170.
- 39 Dong, S., and Ma, W. (2021). How to win a tug-of-war: the adaptive evolution of *Phytophthora*  
40 effectors. *Curr Opin Plant Biol* 62, 102027.
- 41 Dos Santos-Silva, C.A., Zupin, L., Oliveira-Lima, M., Vilela, L.M.B., Bezerra-Neto, J.P., Ferreira-

- 1 Neto, J.R., Ferreira, J.D.C., de Oliveira-Silva, R.L., Pires, C.J., Aburjaile, F.F., *et al.* (2020). Plant  
2 Antimicrobial Peptides: State of the Art, In Silico Prediction and Perspectives in the Omics Era.  
3 *Bioinform Biol Insights* 14, 1177932220952739.
- 4 Doungous, O., Minyaka, E., Longue, E.A.M., and Nkengafac, N.J. (2018). Potentials of cocoa pod  
5 husk-based compost on *Phytophthora* pod rot disease suppression, soil fertility, and *Theobroma*  
6 *cacao* L. growth. *Environ Sci Pollut Res Int* 25, 25327-25335.
- 7 Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene  
8 expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210.
- 9 Endou, M., Yoshida, K., Hirota, M., Nakajima, C., Sakaguchi, A., Komatsubara, N., and Kurihara,  
10 Y. (2020). Coxfa4I3, a novel mitochondrial electron transport chain Complex 4 subunit protein,  
11 switches from Coxfa4 during spermatogenesis. *Mitochondrion* 52, 1-7.
- 12 Fichman, Y., and Mittler, R. (2020). Rapid systemic signaling during abiotic and biotic stresses: is  
13 the ROS wave master of all trades? *Plant J* 102, 887-896.
- 14 Fister, A.S., Mejia, L.C., Zhang, Y., Herre, E.A., Maximova, S.N., and Gultinan, M.J. (2016).  
15 *Theobroma cacao* L. pathogenesis-related gene tandem array members show diverse expression  
16 dynamics in response to pathogen colonization. *BMC Genomics* 17, 363.
- 17 Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic  
18 pathogens. *Annu Rev Phytopathol* 43, 205-227.
- 19 Grunwald, N.J., and Flier, W.G. (2005). The biology of *Phytophthora infestans* at its center of  
20 origin. *Annu Rev Phytopathol* 43, 171-190.
- 21 Hardham, A.R. (2007). Cell biology of plant-oomycete interactions. *Cell Microbiol* 9, 31-39.
- 22 Hoang, N.H., Strogolova, V., Mosley, J.J., Stuart, R.A., and Hosler, J. (2019). Hypoxia-inducible  
23 gene domain 1 proteins in yeast mitochondria protect against proton leak through complex IV. *J*  
24 *Biol Chem* 294, 17669-17677.
- 25 Hou, S., Shen, H., and Shao, H. (2019). PAMP-induced peptide 1 cooperates with salicylic acid  
26 to regulate stomatal immunity in *Arabidopsis thaliana*. *Plant Signal Behav* 14, 1666657.
- 27 Hussain, S., Wang, W., Ahmed, S., Wang, X., Adnan, Cheng, Y., Wang, C., Wang, Y., Zhang, N.,  
28 Tian, H., *et al.* (2021). PIP2, An Auxin Induced Plant Peptide Hormone Regulates Root and  
29 Hypocotyl Elongation in *Arabidopsis*. *Front Plant Sci* 12, 646736.
- 30 Jeger, M.J., and Pautasso, M. (2008). Comparative epidemiology of zoosporic plant pathogens.  
31 *European Journal of Plant Pathology* 122, 111.
- 32 Judelson, H.S., and Blanco, F.A. (2005). The spores of *Phytophthora*: weapons of the plant  
33 destroyer. *Nat Rev Microbiol* 3, 47-58.
- 34 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7:  
35 improvements in performance and usability. *Mol Biol Evol* 30, 772-780.
- 36 Le Berre, J.Y., Gourgues, M., Samans, B., Keller, H., Panabieres, F., and Attard, A. (2017).  
37 Transcriptome dynamic of *Arabidopsis* roots infected with *Phytophthora parasitica* identifies VQ29,  
38 a gene induced during the penetration and involved in the restriction of infection. *PLoS One* 12,  
39 e0190341.
- 40 Lee, C.Q.E., Kerouanton, B., Chothani, S., Zhang, S., Chen, Y., Mantri, C.K., Hock, D.H., Lim, R.,  
41 Nadkarni, R., Huynh, V.T., *et al.* (2021). Coding and non-coding roles of MOCCI (C15ORF48)  
42 coordinate to regulate host inflammation and immunity. *Nat Commun* 12, 2130.
- 43 Lin, F., Zhao, M., Baumann, D.D., Ping, J., Sun, L., Liu, Y., Zhang, B., Tang, Z., Hughes, E.,

- 1 Doerge, R.W., *et al.* (2014). Molecular response to the pathogen *Phytophthora sojae* among ten  
2 soybean near isogenic lines revealed by comparative transcriptomics. *BMC Genomics* *15*, 18.
- 3 Marelli, J.P., Guest, D.I., Bailey, B.A., Evans, H.C., Brown, J.K., Junaid, M., Barreto, R.W., Lisboa,  
4 D.O., and Puig, A.S. (2019). Chocolate Under Threat from Old and New Cacao Diseases.  
5 *Phytopathology* *109*, 1331-1343.
- 6 McDonald, B.A., and Stukenbrock, E.H. (2016). Rapid emergence of pathogens in agro-  
7 ecosystems: global threats to agricultural sustainability and food security. *Philos Trans R Soc*  
8 *Lond B Biol Sci* *371*.
- 9 McDowell, J.M., and Dangl, J.L. (2000). Signal transduction in the plant immune response. *Trends*  
10 *Biochem Sci* *25*, 79-82.
- 11 Meng, H., Sun, M., Jiang, Z., Liu, Y., Sun, Y., Liu, D., Jiang, C., Ren, M., Yuan, G., Yu, W., *et al.*  
12 (2021). Comparative transcriptome analysis reveals resistant and susceptible genes in tobacco  
13 cultivars in response to infection by *Phytophthora nicotianae*. *Sci Rep* *11*, 809.
- 14 Meng, Y., Zhang, Q., Ding, W., and Shan, W. (2014). *Phytophthora parasitica*: a model oomycete  
15 plant pathogen. *Mycology* *5*, 43-51.
- 16 Murphy, E., Smith, S., and De Smet, I. (2012). Small signaling peptides in *Arabidopsis*  
17 development: how cells communicate over a short distance. *Plant Cell* *24*, 3198-3217.
- 18 Najafi, J., Brembu, T., Vie, A.K., Viste, R., Winge, P., Somssich, I.E., and Bones, A.M. (2020).  
19 PAMP-INDUCED SECRETED PEPTIDE 3 modulates immunity in *Arabidopsis*. *J Exp Bot* *71*, 850-  
20 864.
- 21 Naveed, Z.A., Wei, X., Chen, J., Mubeen, H., and Ali, G.S. (2020). The PTI to ETI Continuum in  
22 *Phytophthora*-Plant Interactions. *Front Plant Sci* *11*, 593905.
- 23 Ngou, B.P.M., Jones, J.D.G., and Ding, P. (2021). Plant immune networks. *Trends Plant Sci*.
- 24 Ogilvie, H.A., Imin, N., and Djordjevic, M.A. (2014). Diversification of the C-TERMINALLY  
25 ENCODED PEPTIDE (CEP) gene family in angiosperms, and evolution of plant-family specific  
26 CEP genes. *BMC Genomics* *15*, 870.
- 27 Pandey, V., Awasthi, M., Singh, S., Tiwari, S., and Dwivedi, U. (2017). A Comprehensive Review  
28 on Function and Application of Plant Peroxidases. *Biochemistry & Analytical Biochemistry* *6*.
- 29 Petre, B., and Kamoun, S. (2014). How do filamentous pathogens deliver effector proteins into  
30 plant cells? *PLoS Biol* *12*, e1001801.
- 31 Petri, T., Berchtold, E., Zimmer, R., and Friedel, C.C. (2012). Detection and correction of probe-  
32 level artefacts on microarrays. *BMC Bioinformatics* *13*, 114.
- 33 Pieterse, C.M., Van der Does, D., Zamioudis, C., Leon-Reyes, A., and Van Wees, S.C. (2012).  
34 Hormonal modulation of plant immunity. *Annu Rev Cell Dev Biol* *28*, 489-521.
- 35 Rompler, K., Muller, T., Juris, L., Wissel, M., Vukotic, M., Hofmann, K., and Deckers, M. (2016).  
36 Overlapping Role of Respiratory Supercomplex Factor Rcf2 and Its N-terminal Homolog Rcf3 in  
37 *Saccharomyces cerevisiae*. *J Biol Chem* *291*, 23769-23778.
- 38 Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing  
39 phylogenetic trees. *Mol Biol Evol* *4*, 406-425.
- 40 Scholthof, K.B. (2003). One foot in the furrow: linkages between agriculture, plant pathology, and  
41 public health. *Annu Rev Public Health* *24*, 153-174.
- 42 Sels, J., Mathys, J., De Coninck, B.M., Cammue, B.P., and De Bolle, M.F. (2008). Plant

1 pathogenesis-related (PR) proteins: a focus on PR peptides. *Plant Physiol Biochem* *46*, 941-950.

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

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

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

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

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

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

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

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

24 Vie, A.K., Najafi, J., Liu, B., Winge, P., Butenko, M.A., Hornslien, K.S., Kumpf, R., Aalen, R.B.,  
25 Bones, A.M., and Brembu, T. (2015). The IDA/IDA-LIKE and PIP/PIP-LIKE gene families in  
26 Arabidopsis: phylogenetic relationship, expression patterns, and transcriptional effect of the  
27 PIP3 peptide. *J Exp Bot* *66*, 5351-5365.

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

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

33 Zarattini, M., Farjad, M., Launay, A., Cannella, D., Soulie, M.C., Bernacchia, G., and Fagard, M.  
34 (2021). Every cloud has a silver lining: how abiotic stresses affect gene expression in plant-  
35 pathogen interactions. *J Exp Bot* *72*, 1020-1033.

36 Zhang, B., Van Aken, O., Thatcher, L., De Clercq, I., Duncan, O., Law, S.R., Murcha, M.W., van  
37 der Merwe, M., Seifi, H.S., Carrie, C., *et al.* (2014). The mitochondrial outer membrane AAA  
38 ATPase AtOM66 affects cell death and pathogen resistance in Arabidopsis thaliana. *Plant J* *80*,  
39 709-727.

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

42 Zimmermann, L., Stephens, A., Nam, S.Z., Rau, D., Kubler, J., Lozajic, M., Gabler, F., Soding, J.,  
43 Lupas, A.N., and Alva, V. (2018). A Completely Reimplemented MPI Bioinformatics Toolkit with a  
44 New HHpred Server at its Core. *J Mol Biol* *430*, 2237-2243.

1 Zong, S., Wu, M., Gu, J., Liu, T., Guo, R., and Yang, M. (2018). Structure of the intact 14-subunit  
2 human cytochrome c oxidase. *Cell Res* 28, 1026-1034.

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

5

6

7

## 8 **Statements & Declarations**

### 9 **Funding**

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

11

### 12 **Competing Interest**

13 The authors have no relevant financial or non-financial interests to disclose.

### 14 **Authors Contributions**

15 Both authors jointly initiated the study, performed the analyses and wrote the manuscript.

### 16 **Data Availability**

17 Input transcriptomics data are available from the GEO database

18 (<https://www.ncbi.nlm.nih.gov/geo>) under the accession numbers given in the text. All relevant  
19 data generated and used in this project are provided as supplementary data files with the manu-  
20 script.

21

22

1 **Tables**

2

| PR    | Description                          | USC # | members     |         |       |
|-------|--------------------------------------|-------|-------------|---------|-------|
|       |                                      |       | Arabidopsis | Soybean | Cocoa |
| PR-1  | CAP/SCP superfamily                  | 13    | 4           | 9       | 4     |
| PR-2  | $\beta$ -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  | Thaumatococcus-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-metalloproteinase      | 135   | 0           | 3       | 2     |

3

4

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

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

12

1

2 **Supplementary files**

3 Suppl. File S1: Cytoscape Network of all USCs.

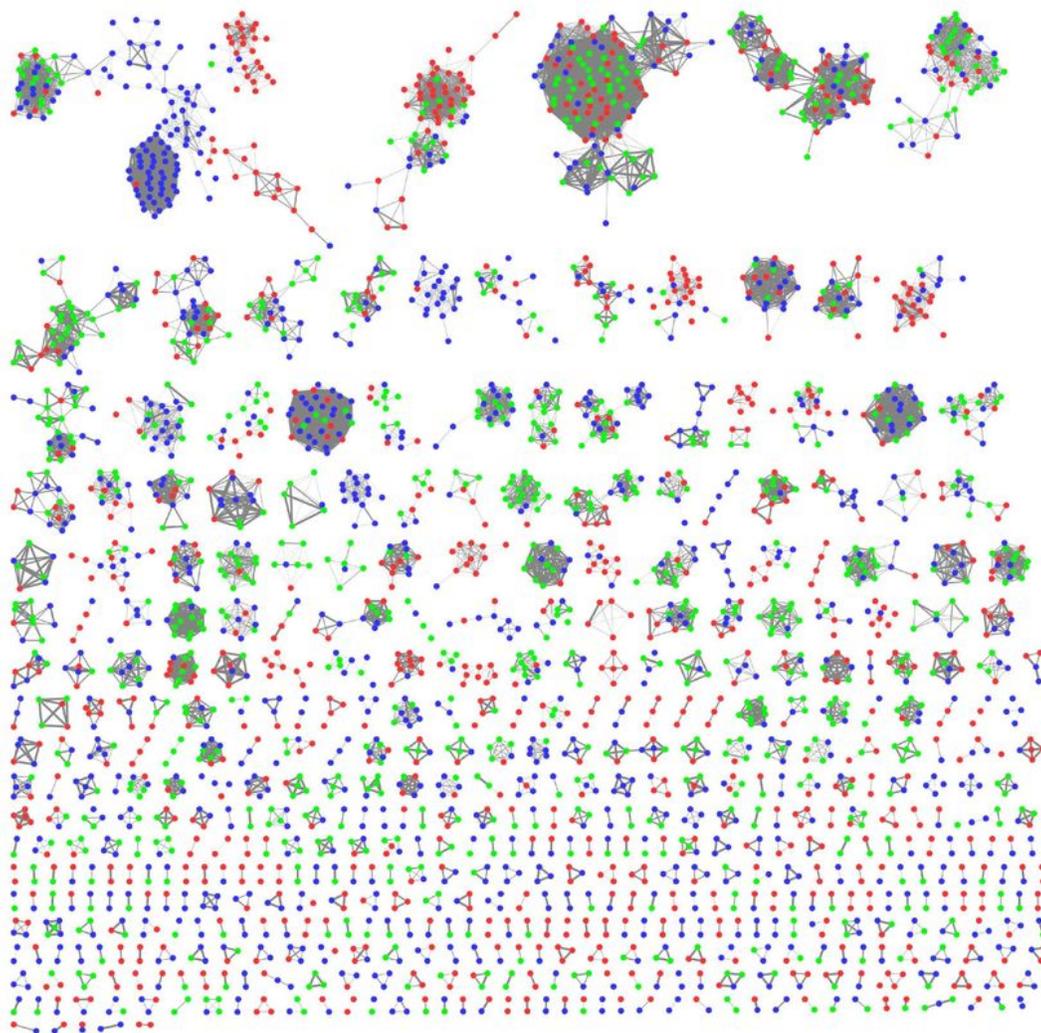
4 Suppl. File S2: Master Table of all annotated USCs

5 Suppl. File S3: Full table of PR-genes

6 Suppl. File S4: Preprocessed source data of plant transcriptomics projects.

7

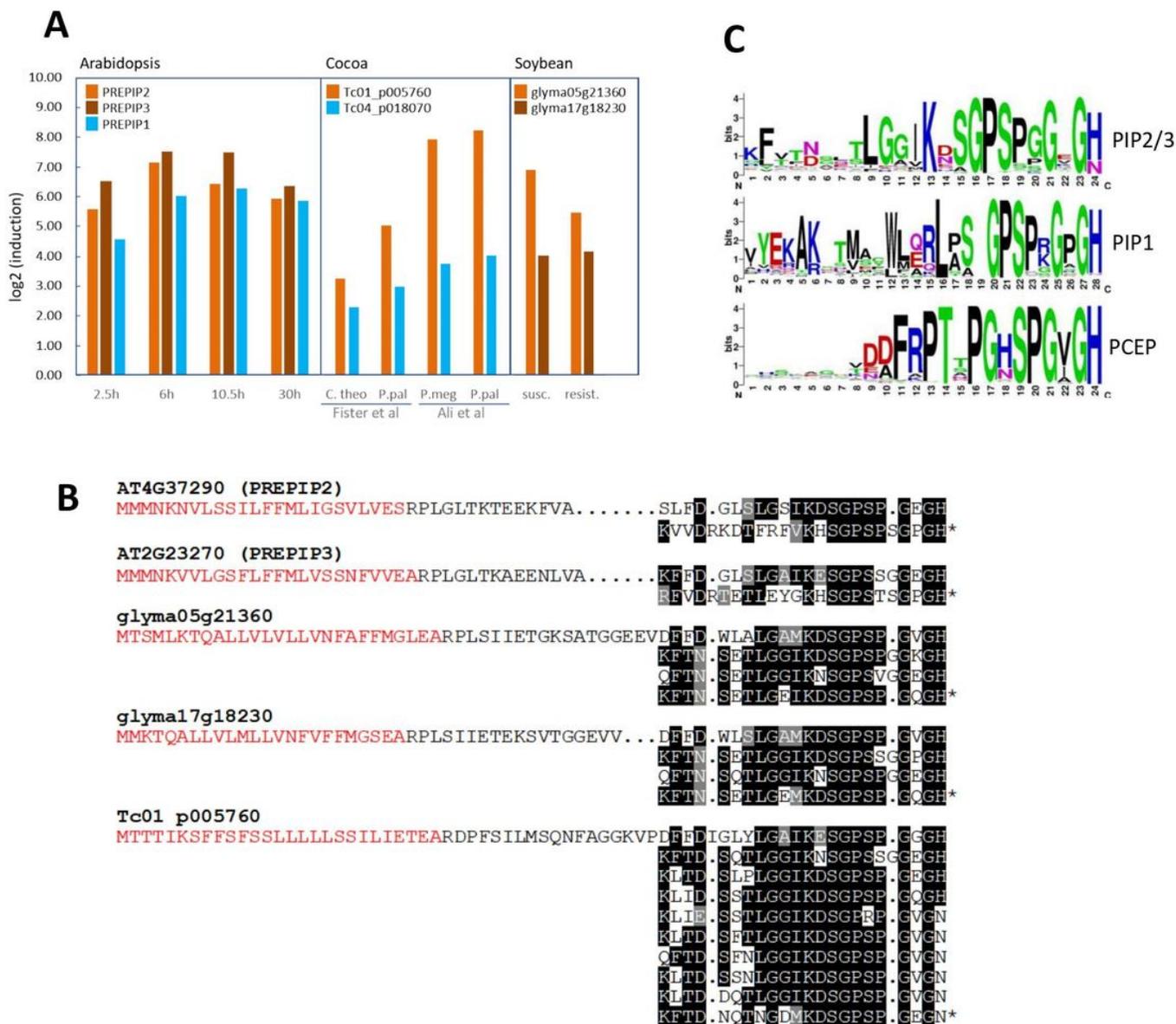
# Figures



**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 theobromicola*, *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.

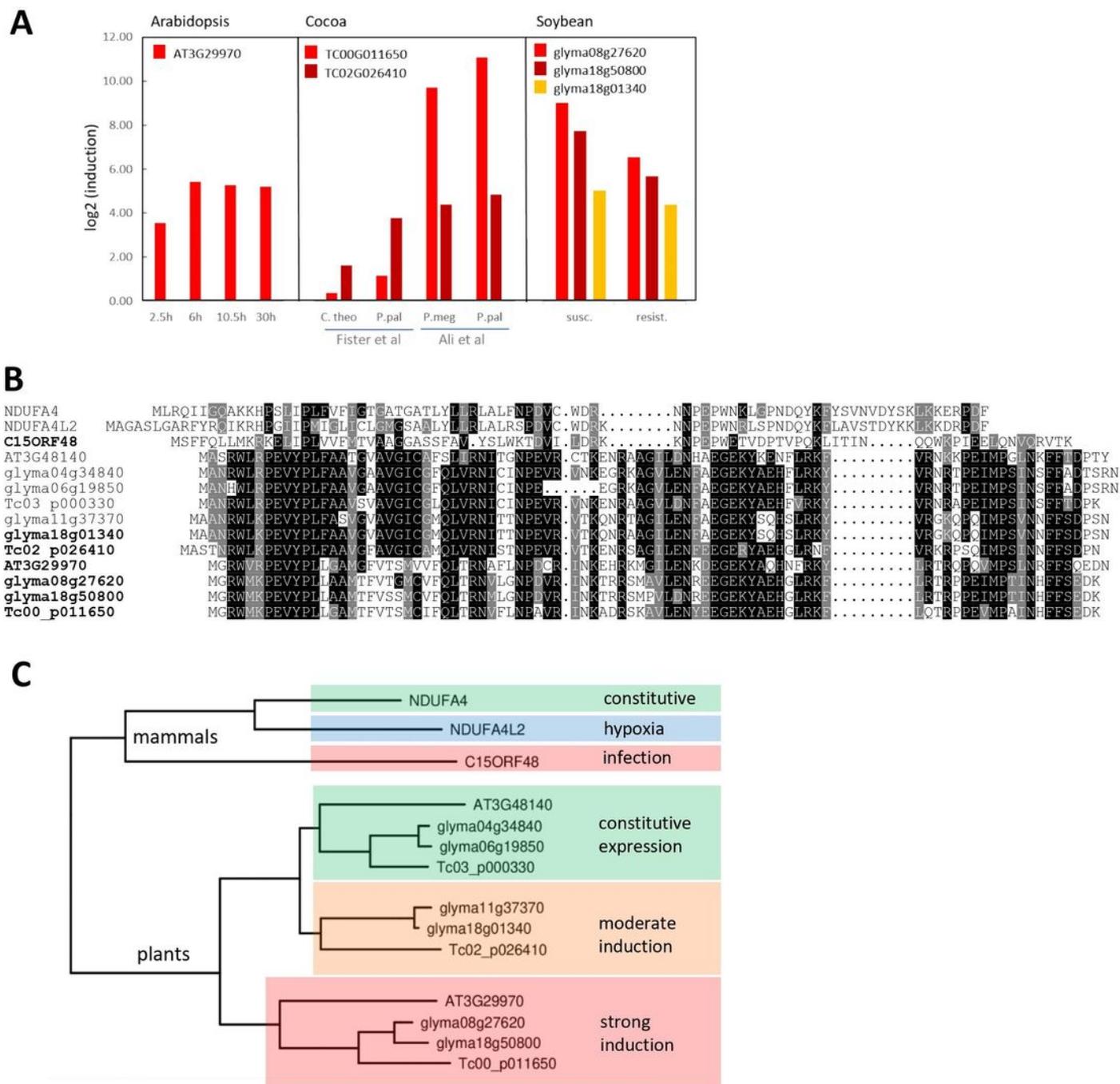


Figure 3

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. theobromicola*, *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 con-servatively 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](#)
- [S4preprocessed sourcedata.xlsx](#)