Neo-functionalisation of mitochondrial proteins and incorporation into signaling networks in plants

Sbatie Lama¹, Martyna Broda², Zahra Abbas², Dries Vaneechoutte³,⁴, Katharina Belt²,⁵, Torbjörn Säll¹, Klaas Vandepoele³,⁴, Olivier Van Aken¹*

¹Department of Biology, Lund University, Sölvegatan 35, 223 62 Lund, Sweden
²ARC Centre of Excellence in Plant Energy Biology, University of Western Australia, 35 Stirling Highway, 6009 Crawley, Western Australia, Australia
³Ghent University, Department of Plant Biotechnology and Bioinformatics, Technologiepark 927, 9052 Ghent, Belgium
⁴VIB Center for Plant Systems Biology, Technologiepark 927, 9052 Ghent, Belgium
⁵Current address: CSIRO, 147 Underwood Avenue, 6014, Floreat, WA

* Correspondence to: olivier.van_aken@biol.lu.se

Key words: mitochondria, evolution, retrograde signalling, stress, neo-functionalisation
Abstract

Because of their symbiotic origin, many mitochondrial proteins are well-conserved across eukaryotic kingdoms. It is however less obvious how specific lineages have obtained novel nuclear-encoded mitochondrial proteins. Here we report a case of mitochondrial neo-functionalisation in plants. Phylogenetic analysis of genes containing the Domain of Unknown Function 295 (DUF295) revealed that the domain likely originated in Angiosperms. The C-terminal DUF295 domain is usually accompanied by an N-terminal F-box domain, involved in ubiquitin ligation via binding with ASK1/SKP1 type proteins. Due to gene duplication, the gene family has expanded rapidly, with 94 DUF295-related genes in Arabidopsis thaliana alone. Two DUF295 family subgroups have uniquely evolved and quickly expanded within Brassicaceae. One of these subgroups has completely lost the F-box, but instead obtained strongly predicted mitochondrial targeting peptides. We show that several representatives of this DUF295 Organellar group are effectively targeted to plant mitochondria and chloroplasts. Furthermore, many DUF295 Organellar genes are induced by mitochondrial dysfunction, while F-Box DUF295 genes are not. In agreement, several Brassicaceae-specific DUF295 Organellar genes were incorporated in the evolutionary much older ANAC017-dependent mitochondrial retrograde signalling pathway. Finally, a representative set of DUF295 T-DNA insertion mutants was created. No obvious aberrant phenotypes during normal growth and mitochondrial dysfunction were observed, most likely due to the large extent of gene duplication and redundancy. Overall, this study provides insight into how novel mitochondrial proteins can be created via ‘inter-compartmental’ gene duplication events. Moreover, our analysis shows that these newly evolved genes can then be specifically integrated into relevant, pre-existing co-expression networks.
Mitochondria are membrane-bound organelles with crucial roles in eukaryotic cells, including ATP production, Fe-S cluster synthesis, the Krebs cycle and many other metabolic functions. The endosymbiotic theory proposes that mitochondria are derived from ancestral bacteria that were engulfed and retained by a host cell, probably of archaeal origin (Spang et al. 2015). The exact lineage of bacteria that was the precursor to mitochondria is not known, but they are likely related to (alpha)proteobacteria (Martijn et al., 2018). During evolution, the gene content of the mitochondrial genome was heavily reduced, and the vast majority of mitochondrial proteins are encoded in the nuclear genome (Ku et al. 2015). This required the mitochondrial import machinery to evolve, allowing proteins translated in the cytosol to be imported into the different compartments of the mitochondria, often with help of specific targeting peptides. Other originally bacterial functions were not redirected to mitochondria and became operational elsewhere in the cell, e.g. the cytosol and peroxisomes (Huynen et al. 2013), or were lost entirely.

Mitochondria usually contain >1000 different proteins, e.g. around 1800 proteins in mammals (Palmfeldt and Bross 2017), and perhaps >2000 in plant mitochondria (Rao et al. 2017). As the mitochondria-containing host cell was probably the ancestor for all eukaryotic lineages including plants, animal and fungi, one would expect the majority of mitochondrial proteins to be of clear bacterial origin. However, of the approx. 800 human nuclear genes that bear clear resemblance to alpha-proteobacterial genes, only about 200 are present in the human mitochondrial proteome (Szklarczyk and Huynen 2010). The current view is that the mitochondrial proteome is a mixture of alpha-proteobacteria-related proteins, proteins from other (proteo)-bacteria obtained via lateral gene transfer, and viral proteins. Additionally, about 40% of the mitochondrial proteome has no clear viral or bacterial origin. These proteins are thought to be of pre-mitochondrial host cell origin, or are of ‘lineage specific’ origin (e.g. plant-specific), having originated after the Last Eukaryotic Common Ancestor (LECA) (Roger et al. 2017). The proteome of the plastid (derived from a photosynthetic cyanobacterial endosymbiont) appears to have a similarly complex origin (Leister 2016; Roger et al. 2017).

For the multitude of mitochondrial proteins that are related to bacterial, viral or archaeal proteins, different mechanisms including endosymbiotic gene transfer (from the endosymbiont to the nucleus) or lateral gene transfer can be envisaged. It is, however, less evident how different eukaryotic lineages have obtained lineage-specific mitochondrial (or plastid) protein types. One possibility is via random creation of new open reading frames that encode completely novel proteins. Another mechanism may be gene duplication, whereby a new copy of a gene is created in the genome. In most cases, the encoded protein retains its original subcellular localisation (intra-compartmental duplication).
(Szklarczyk and Huynen 2009). However, in rare cases, the duplication results in one of the encoded proteins becoming targeted to another subcellular location (inter-compartmental duplication). Another consequence of the endosymbiotic nature of the eukaryotic cell, is the need for more complex transcriptional regulation. As most of the genes encoding mitochondrial or plastid proteins are found in the nuclear genome, the individual mitochondrial or plastid genomes can no longer directly control all transcript levels. Instead, the organelles must provide feedback to the nucleus to steer gene expression, a process called retrograde signalling. Such retrograde signalling pathways have been described in fungi, animals and plants. Especially when the cellular or metabolic situation in the organelle changes (e.g. availability of substrates or light, inhibition of important enzymes, reactive oxygen species), adequate adjustments in transcript levels are needed to fine-tune the organellar proteomes. This further raises the question of how lineage-specific organellar proteins become regulated appropriately after their appearance. To be of optimal use to the cell, the new genes may become incorporated into existing transcriptional networks relevant to organellar function. Alternatively, specific new needs may require new transcriptional modules to evolve. Indeed, the best-known retrograde signalling pathways in yeast, animals or plants appear to be quite different between lineages, and employ different (even lineage-specific) transcription factors (Ng et al. 2014; da Cunha et al. 2015).

In this study, we describe the phylogenetic history of lineage-specific Domain of Unknown Function 295 genes in plants. Despite its poorly understood function, the gene family is strongly expanded with 94 representatives in the Arabidopsis thaliana genome. Our findings show that relatively recent tandem gene duplications in the Brassicaceae family have lead to neo-functionalisation in plant mitochondria. Most likely through incomplete gene duplication, an ancestral DUF295 domain gene has lost its N-terminus, and has instead obtained a functional mitochondrial targeting peptide. Furthermore, we show that several of these new mitochondrial proteins have been specifically integrated into pre-existing gene-expression networks containing ‘old’ genes that regulate mitochondrial function.
**Results**

*The DUF295 proteins form a large gene family within angiosperms*

Despite many decades of intensive research, many conserved protein domains still have unknown functions. The term Domain of Unknown Function (DUF) was originally coined to describe two bacterial domains (DUF1 and DUF2) found in bacterial signalling proteins (Schultz et al. 1998). Subsequent bioinformatics approaches identified thousands of additional uncharacterised domains that were assigned numbers in the PFAM database. The latest PFAM release 32 contains nearly 4000 DUF families (up to DUF5654), representing around >20% of the known families (Bateman et al. 2010). The DUF295 domain was identified by PFAM release 7.0 (Bateman et al. 2004), and currently contains 4353 family members, with an average domain length of 57.80 amino acids. The *Arabidopsis thaliana* Col-0 genome sequence was searched for proteins containing the DUF295 domain, based on the PFAM motif PF03478 (Suppl. Figure 1) and the TAIR10 annotation ([www.arabidopsis.org](http://www.arabidopsis.org)). Using further homology searches, in total 94 unique loci encoding DUF295-related proteins were found (Figure 1A, Table I). To examine the evolutionary conservation and origin of the DUF295 protein family, homology searches were performed to identify representative family members in other lineages. The DUF295 domain was not found in prokaryotes and Animalia. Interestingly a single DUF295 domain protein was found by EBI ([http://www.ebi.ac.uk/interpro/entry/IPR005174/taxonomy](http://www.ebi.ac.uk/interpro/entry/IPR005174/taxonomy)) in the Basidiomycete *Exigia glandulosa* (KZV93935.1; Fungi). However, this order-specific protein (Auriculariales) did not have any homologs with the DUF295 domain, suggesting it is not a true DUF295 domain protein, or has evolved independently in a single known fungal species. We thus concluded that the DUF295 domain is green lineage specific.

Through the application of a second round of homology searches using the PLAZA 4.0 comparative genomics database containing >70 genomes of species within the Viridiplantae, no DUF295 proteins were found in gymnosperms (Figure 1, Supplementary Figure 2). However, DUF295 proteins were found in monocots (e.g. *Oryza sativa* subsp. *japonica* contains 264 DUF295 proteins) and dicots (e.g. *Populus trichocarpa* contains 23). The presence of DUF295 proteins in the Angiosperm *Amborella trichopoda*, which is thought to be a ‘sister species’ of flowering plants that branched off after the gymnosperms, but before the monocot-dicot divergence, was ambiguous (Amborella Genome 2013). Using a search with the DUF295 PFAM HMM profile, no *A. trichopoda* proteins were identified. In conclusion, the DUF295 domain is strongly represented in monocots and dicots, and has likely originated around the branching of *A. trichopoda*, after the gymnosperm/angiosperm divergence.
Figure 1. Phylogenetic analysis of the DUF295 protein family. (A) Unrooted phylogenetic tree of proteins containing DUF295 domains in representative Angiosperm species. Scale bar indicates percentage divergence. For clarity gene and species names have been removed, but information on dicot, monocot or Amborella trichopoda is indicated by different circles (see figure). A fully annotated phylogenetic tree with species/gene names and bootstrap values can be found in Supplementary Figure 2. (B) General domain structure of the four groups of DUF295 proteins found in Brassicaceae. mTP: mitochondrial targeting peptide; cTP: chloroplast transit peptide.
Most DUF295 domain proteins also contain an F-box domain

From the phylogenetic analysis it was apparent that the DUF295 proteins can be divided in four major classes (Figure 1A). One class of DUF295 proteins (indicated in green in Figure 1) was represented in the genomes of all studied plant species, with a clear subgroup of monocot and dicot representatives. This group is thus most likely the ancestral DUF295 protein class. Within the dicot subgroup, a clear expansion of Brassicaceae DUF295 homologs was observed. Analysis of the domain structure of the proteins in the group showed the presence of an N-terminal F-box domain and a C-terminal DUF295 domain (Figure 1B). F-box domains are about 50 amino acids long and involved in protein-protein interactions. They are often found in SCF (Skp1-cullin-F-box) ubiquitin E3-ligases that mark proteins for degradation, with the F-box imparting specificity of the target proteins. Many key plant hormone receptors have been found to be SCF proteins, including SCFTIR (auxin receptor) and SCFCOI1 (jasmonic acid receptor) (Kepinski and Leyser 2005; Katsir et al. 2008). The Arabidopsis genome contains 20 ancestral-type F-box/DUF295 proteins. The SKP1-interacting Protein SKIP23 (At2g17030) is part of this group and was previously found to interact with ASK1 (Risseeuw et al. 2003), a component of e.g. the strigolactone SCFMAX2 receptor complex (Yao et al. 2016). In a more recent study, 6 ancestral-type F-box/DUF295 proteins were found to interact with ASK1 and related proteins by yeast two-hybrid screens (Kuroda et al. 2012). SKIP23 was also found to interact with Arabidopsis 14-3-3 proteins (Hong et al. 2017). Upward Curly Leaf 1 (UCL1, At1g65740) was found to be a nuclear protein interacting with Curly Leaf Polycomb proteins and ASK1 (Jeong et al. 2011). This indicates that ASK1-binding is a common feature of this protein group. As most Arabidopsis DUF295 proteins lack a systematic gene name, we named the genes in this group AtFDA1-20 (F-box/DUF295 Ancestral) (Figure 1, Table 1).

Interestingly, AtFDA9 (At2g16300) is nearly identical to the adjacent gene AtFDA8 (At2g16290), however a frame shift has occurred due to a single base deletion just before the start of the DUF295 domain. This leads to a premature stop codon and truncated AtFDA9 protein of 322 amino acids, instead of around 415 amino acids as in AtFDA8 (where the DUF295 domain is at position 319-360). If the AtFDA9 transcript sequence after the premature stop codon is translated in the +2 frame, the DUF295 domain can be clearly identified, indicating that AtFDA9 was originally a DUF295-containing gene. The structure of the close-by gene AtFDA10 (At2g16365) appeared even more complex. In the current TAIR10 annotation, At2g16365 is named photoperiodic control of hypocotyl 1 (PCH1), which is 778 amino acids long. The PCH1 ‘domain’ (a phytochrome interaction domain) is located at the N-terminal, and the F-box is at amino acids 459-505. In the other AtFDA genes the F-box is located right at the N-terminal (e.g. residues 3-48 in AtFDA8), suggesting it is a compound gene. Indeed, 6 splice forms have been annotated for At2g16365, where At2g16365.2 does not contain the F-box and downstream sequence. We checked a range of RNA-seq data sets, but could not find evidence for
reads spanning the suggested 3rd intron, which would connect the PCH1 region to the FDA region. There is also no proteomic support for the existence of proteins containing both PCH1 and AtFDA10 sequence, and only the short splice variant could be cloned (At2g16365.2) (Huang et al. 2016). Therefore, we propose that the currently annotated At2g16365 locus actually encodes two separate genes, PCH1 and AtFDA10. Additionally, AtFDA10 has a two-base insertion upstream of the DUF295 domain, causing a premature stop codon and loss of the actual DUF295 domain, as observed in AtFDA9.

A second group of DUF295 proteins was identified (indicated in red in Figure 1), which also contained an F-box/DUF295 arrangement, but was clearly divergent from the ancestral FDA-type proteins. This type of protein was only represented in Brassicaceae genomes, with Arabidopsis thaliana containing 38 homologs, Arabidopsis lyrata 56 homologs and Brassica rapa 32 homologs, respectively. As none of these genes have systematic gene names, we named the Arabidopsis thaliana genes in this group AtFDB1-38 (F-Box/DUF295 Brassicaceae-specific). Despite their relatively large number, not much functional information could be found on these proteins. The same study that identified 6 AtFDA proteins to interact with ASK1 and its homologs in Arabidopsis, could not detect ASK1-interaction for the 6 tested AtFDB proteins (Kuroda et al. 2012). This strongly suggests that this relatively recent, Brassicaceae-specific group has significantly diverged from the ancestral FDA DUF295 proteins.

A third group of DUF295 domain proteins was found (Indicated in blue and purple in Figure 1), but again only in Brassicaceae species (Figure 1). Within this group, two clear subgroups were observed, each containing 17 Arabidopsis thaliana proteins, and many orthologs in the other included Brassicaceae species. Remarkably, none of the proteins in this third group contained the F-box domain, and only the C-terminal DUF295 domain could be found as an annotated domain (Figure 1). Yeast two-hybrid interactions have been reported only for At4g25920 (Arabidopsis Interactome Mapping 2011) (Suppl. Table I). A fourth group (indicated in orange in Figure 1), including At1g57790 and At5g55150, contains the F-box domain, but the DUF295 domain mentioned in the TAIR annotation was not identified using the PFAM profile, indicating it has diverged substantially. A yeast two-hybrid interaction was found with ASK1 for At1g57790 (Kuroda et al. 2012), indicating at least partial functional similarity to the FDA ancestral DUF295 proteins. At1g57790 and At5g55150 were most closely related to Amborella trichopoda ATR0851G001 in the phylogenetic tree. Therefore, we propose that this group represents an older precursor or sister-group to the ‘proper’ DUF295 protein family, and named it AtFDR1-2 (F-Box/DUF295-Related).

**The DUF295 gene family has expanded rapidly by tandem gene duplication**

It was surprising to find that both groups of Brassicaceae-specific DUF295 proteins (Figure 1) contained more members (38 and 34 in Arabidopsis) than the ancestral DUF295 group (20 in
Arabidopsis). This indicates a very rapid expansion of the gene family in a relatively short evolutionary time, as Brassicaceae are thought to have branched off about 32 million years ago (Hohmann et al. 2015). When examining the chromosomal locations of genes in the three groups, it was obvious that many homologs were tandem duplications, as evidenced by (nearly) adjacent locations and belonging to the same protein (sub)group. Across the whole DUF295-related family in Arabidopsis thaliana (94 genes), nearly 70% (63 genes) were present as tandem repeats, representing 27 tandem groups (Suppl. Table II). Duplication rates were particularly high in the DUF295-only group, with 27 of 34 genes (±80%) spread over 13 tandems. One tandem even contained 6 genes spanning At5g54320 to At5g54560. The F-box/DUF295 FDB group contained 26 of 38 (±68%) tandem duplicated genes spread over 10 tandems. Also here tandems of up to 6 genes were found (At4g22030 to At4g22180). Finally, also the ancestral FDA family contained many tandem duplications, with 10 of 20 genes (50%) spread over 4 tandems. In conclusion, it appears that the DUF295 family has achieved its large size, particularly in Brassicaceae, via numerous rounds of tandem duplications.

The Brassicaceae-specific DUF295 group has replaced the F-box domain with a mitochondrial targeting peptide

Since the N-terminal F-box domain was lost or missing in the third group of DUF295 proteins, we analysed the N-terminal region of these proteins in greater detail. Surprisingly, the large majority of the Arabidopsis thaliana representatives contained predicted mitochondrial targeting peptides, as suggested by several prediction tools including iPSORT, Mitoprot and Mitopred (Suppl. Table III) (Hooper et al. 2017). Of 34 Arabidopsis thaliana homologs, 26 were predicted to be mitochondrial by 6 or more prediction tools, and 14 were predicted to be mitochondrial by 10 or more prediction tools. 22 out of 34 proteins received a consensus localisation prediction to the mitochondria based on the SUBA4 consensus algorithm (Hooper et al. 2014). 24 of 26 proteins were also predicted as plastid-localised by at least 1 prediction tool (SUBA4), but only four were predicted to plastid-targeted by 5-7 prediction tools. The SUBA consensus algorithm suggested plastid localisation for only 1 protein, as opposed to 22 receiving a mitochondrial consensus prediction. Again, no systematic naming system is present for these proteins so we named the Arabidopsis thaliana homologs AtDOA1-17 (DUF295 Organellar A) and AtDOB1-17 (DUF295 Organellar B), based on the two apparent subgroups (Figure 1).

Despite the strong predictions, experimental evidence for organellar location of the DOA and DOB proteins was very limited. AtDOA8 (At4g16080) was identified in purified mitochondria by mass spectrometry (MS) (Senkler et al. 2017), while AtDOA11 (AT5G03390) was identified by MS in purified chloroplasts (Zybailov et al. 2008). As no GFP localisation data was published for any of the predicted mitochondrial isoforms, we cloned 3 representatives AtDOA10 (At4g25930), AtDOB5 (At5g52940) and
AtDOB12 (At5g54550) into C-terminal GFP-fusion vectors (see below for more information on why these were selected). The localisation of the fusion proteins was analysed by transient transformation of Arabidopsis thaliana cell cultures (Figure 2). Both AtDOB5 and AtDOB12 showed clear mitochondrial localisation, as evidenced by co-localisation with an alternative oxidase-RFP marker. AtDOB5 also showed a weaker signal in plastids, suggesting dual localisation (Figure 2, Suppl. Figure 3). For AtDOA10, only diffuse cytosolic localisation was found, with no clear co-localisation with the AOX-RFP marker. In conclusion, independent sources and experimental approaches support that several of the DOA/DOB proteins have obtained functional mitochondrial and/or plastid targeting peptides, in line with their strong organellar prediction.

![Figure 2. DUF295 organellar proteins are targeted to the mitochondria. C-terminally GFP-tagged fusion proteins were transiently transformed into Arabidopsis thaliana cell culture, and co-transformed with mitochondrial marker AOX-RFP. Scale bar indicates 10 µm.](image)

DUF295 genes show remarkably specific expression patterns

The strong expansion of DUF295 genes does not necessarily indicate that the genes are functional and expressed. Therefore, the transcript levels of the 94 Arabidopsis genes were analysed in a large set of available gene expression experiments (Figure 3, Suppl. Figure 5). Starting from the transcript counts for 206 public RNA-Seq experiments, a gene expression matrix was generated by summing transcript counts per locus (Vaneechoutte et al. 2017). Out of 94 DUF295-related genes, 73 (78 %) appeared to be expressed in one or more conditions (maximum transcripts per million > 2). More than 50 % (11) of the non-expressed genes were of the AtFDB type, while only 1 AtFDA (AtFDA8) did not seem to be significantly expressed. 4 AtDOA and 5 AtDOB genes also were not clearly expressed. Remarkably,
most of the expressed genes were expressed under relatively specific conditions with only 4 DUF295 genes showing strong ubiquitous expression (AtFDA3, AtFDA11/SKIP23, AtFDA14 and AtFDB2). DUF295-related AtFDR1 also appears to be expressed in most tissues and conditions. In contrast, most DUF295 genes were expressed under very specific tissues or conditions, often reproductive tissues such as young anthers, pollen, siliques, and young seeds. Others were specifically expressed during abiotic stress, or biotic stress (Botrytis cinerea).

As many DUF295-containing genes are present in tandem duplicates, often with up to 6 related genes in close proximity, we examined whether tandem duplicated genes are co-expressed. In many cases, tandem pairs of two genes were found to cluster together and showed very similar expression patterns (e.g. AtDOA1/2, AtDOA14/15, AtDOB4/5, AtDOB16/17). Interestingly such paired expression patterns were often observed for AtDOA and AtDOB genes, whereas F-box containing genes only rarely showed such clear co-expression between tandem repeated genes: AtFDB30 and AtFDB33, though these are interspersed by two non-co-expressed genes in the tandem repeat). We also noted that in the larger tandem repeats like AtDOB8-13 (6 genes), only groups of maximum two genes were similarly expressed (AtDOB8/9 and AtDOB10/11), but these two pairs were very different from each other (Figure 3). Several groups of genes showed remarkably similar expression patterns, such as 8 mixed AtFDA/AtFDB genes expressed in siliques, or 8 genes induced by B. cinerea infection (with members of AtFDA/FDB/DOA/DOB groups). Clearly the genes in these groups were not tandem repeats, so the mechanism behind their co-expression is most likely not tandem duplication of promoter regions.
High Impact Mutation analysis across 1135 Arabidopsis thaliana genomes

To get more insight into which DUF295-related genes may be more active and/or functionally important, we assessed whether they are retained as intact open reading frames in the genome sequences of 1135 Arabidopsis thaliana accessions published by the 1001 Genomes consortium (Consortium 2016). For all 94 Arabidopsis thaliana Col-0 DUF295-related genes, the occurrence of ‘high-impact mutations’ (e.g. gain or loss of start/stop codons, loss of splice acceptor sites) was searched in the other ecotypes. This varied widely, with some genes having accumulated no high-impact mutations (HIMs) in other accessions, while others have accumulated many hundreds (Suppl. Table IV). To clarify, if the same variant compared to Col-0 was found in multiple accessions, it was counted as the number of accessions it occurred in. In other words, if one mutation occurred in 300 accessions, this was counted as 300. Next, we plotted the number of HIMs against the transcript expression strength (maximal transcripts per million, max tpm, in the above 206 RNA-seq data sets) (Figure 4). A clear trend could be observed that genes with high expression usually had a lower number of HIMs. Conversely, genes with low expression often had many mutations. The only clear exception was At2g16365 which had both the highest expression and the highest number of HIMs. As stated

Figure 3. Expression patterns of DUF295 genes.

Expression values, normalized per gene, are shown for 72 expressed DUF295 genes. Gene names are colored to indicate the family subgroups (see Figure 1). Only a subset of 25 samples is shown. Expression data for all 206 samples in the Vaneechoutte et al., 2017 dataset is available in Suppl. Figure 4.
above this locus actually contains two separate genes \textit{PCH1} and \textit{AtFDA10}, so it was excluded from the analysis.

Given the high rate of gene duplication, we postulated that a recently duplicated gene may develop into a functional gene (‘consolidated’: max tpm > 5, HIM < 50), gradually turn into a pseudogene and eventually disappear via mutations (‘degenerating’: HIM < 50), or temporarily remain in an intermediate stage (‘undecided’: max tpm < 5, HIM < 5) (Figure 4B). >90% of the genes fell inside the intervals using cut-offs max_tpm of 5 and HIMs of 50, suggesting they are relevant (Figure 4A). When examining ancestral FDA genes, it appears that this selection is nearing completion, as nearly all \textit{AtFDA} genes are either ‘consolidated’ or ‘degenerating’ based on our cut-offs, with only one remaining ‘undecided’ (\textit{AtFDA16}). This further supports the idea that the FDA genes are relatively ancient. Similarly, both \textit{AtFDR} genes show strong expression and very low HIMs (0-1), and thus seem completely ‘consolidated’, supporting their pre-monocot/dicot divergence origin. For the probably more recent Brassicaceae-specific genes, the situation looks different. For the \textit{AtFDB} F-box genes an even distribution across the three groups can be seen, suggesting selection is still ongoing and balanced. More than 50% of the \textit{AtDOA} genes seem to be ‘degenerating’, while fewer are being consolidated. Conversely, while most \textit{AtDOB} genes are still in a more ‘undecided’ state, far more are being ‘consolidated’ than are ‘degenerating’. This suggests that there is higher selective pressure on \textit{AtFDR}, \textit{AtDOB} and \textit{AtFDA} genes, while \textit{AtDOA} genes may be degenerating more often.

\textbf{Figure 4. High impact mutations identified by the 1001 Genomes tool.} (A) Accumulation of high-impact mutations (HIMs) in 1135 \textit{Arabidopsis thaliana} accessions compared to Col-0 were plotted against the gene expression levels (transcripts per million, max tpm). Genes were color-coded by DUF295 subgroup. The red dashed lines indicate the cut-offs used for Figure 4B. (B). Percentage of genes within each DUF295-related subgroup that are postulated to be ‘consolidated’, ‘degenerating’ or ‘neutral’, based on the cut-offs in Figure 4A.
**DUF295 Organellar genes were incorporated into the ANAC017 retrograde signalling pathway**

Previously, we reported that 8 DUF295 genes were constitutively induced in Arabidopsis mutants with mitochondrial defects (Van Aken et al. 2016). Surprisingly, all 8 of these are members of the DUF295 Organellar group (2 AtDOA, 6 AtDOB), while none of the F-box DUF295 proteins were represented (Table 1). To further examine the specificity of DUF295 Organellar proteins in responding to mitochondrial dysfunction, an antimycin A treatment time course was set up. Gene expression levels were measured for the most highly-induced AtDOA representative (AtDOA10, according to Suppl. Table II), and two highly-induced AtDOB representatives (AtDOB12 and AtDOB5). Furthermore, AtFDA11/SKIP23 and AtFDB2 were selected from the F-box DUF295 proteins, based on their relatively high expression in Col-0 seedlings of similar age in previous RNAseq data sets (Van Aken et al. 2016) (Supplementary Table III). Figure 5 shows that only AtDOB12, AtDOB5 and AtDOA10 were strongly induced by antimycin A, while AtFDA11/SKIP23 and AtFDB2 showed no induction.

![Figure 5. DUF295 organellar genes are incorporated into mitochondrial retrograde signalling networks.](image)

Two-week-old Col-0 and anac017 mutant plants were treated with antimycin A and samples were collected in triplicate pools of plants. mRNA levels were quantified using qRT-PCR and normalised to Col-0 at time point 0h. Asterisks indicate statistically significant difference in expression level.
compared to time point 0 in the same genotype (* p<0.05; ** p<0.01); Hash-tags indicate significant difference at the same time point between Col-0 and anac017 (# p<0.05; ## p<0.01).

As antimycin A is known to induce gene expression via retrograde signalling, the response of the selected DUF295 genes was also monitored in mutants lacking ANAC017, a key transcription factor in plant mitochondrial and chloroplast regulation (De Clercq et al. 2013; Ng et al. 2013; Van Aken et al. 2016). The antimycin A-induced gene expression was almost completely suppressed during the first 6 h, which was where the peak expression occurs in wildtype plants (Figure 54). Some delayed expression was observed towards 9-12h, which was most likely due to contributions by ANAC017 homologs, such as ANAC013, ANAC053 and ANAC078 (De Clercq et al. 2013; Van Aken et al. 2016). No significant differences in gene expression for AtFDA11/SKIP23 or AtFDB2 were observed between Col-0 and the anac017 mutants. In summary, the tested DUF295 Organellar genes were strongly induced by mitochondrial dysfunction in an ANAC017-dependent way. The F-box DUF295 genes, however, seem to be largely unresponsive to mitochondrial stress signalling.

The promoters of the 8 DUF295 Organellar genes that were found as responsive to mitochondrial dysfunction based on RNA-seq data (Van Aken et al. 2016) were searched for binding motifs of ANAC017 and/or its related NAC transcription factors (mitochondrial dysfunction motif) (De Clercq et al. 2013). The TF2Network tool was used and for 7 out of 8 genes an MDM like motif (CTTGnnnnnCAAG or similar) was found. Only for AtDOA3 (At1g30160) no MDM could be found, which is in line with its ANAC017-independent gene expression (Supplementary Table II). Furthermore, by using DNA affinity purification sequencing (DAP-seq, a variant of chromatin immunoprecipitation ChiP), we found that the promoters of these 7 genes bind to ANAC017 and/or its homologs (Supplementary Table II) (O'Malley et al. 2016).

It was surprising that only DUF295 gene variants with (predicted) mitochondrial targeting peptides have become incorporated into a mitochondrial signalling network. Additionally, this must have occurred relatively recently in evolutionary history, since the DUF295 Organellar proteins only evolved in Brassicaceae. As mitochondrial retrograde signalling occurs in all eukaryotic kingdoms (da Cunha et al. 2015), the plant-specific ANAC017-dependent mitochondrial retrograde pathway is most likely much more ancient than Brassicaceae (Kim et al. 2007). This would require that the recent DUF295 Organellar genes have been ‘adopted’ by a much older, pre-existing co-expression set. To test this hypothesis, we performed a phylostratic co-expression analysis of all DUF295 genes. The phylostratic classification grouped Arabidopsis thaliana genes in 13 classes based on their evolutionary conservation (Quint et al. 2012), ranging from genes universally conserved in cellular organisms (phylostratum 1), via Viridiplantae, to genes that are Brassicaceae- (phylostratum 12) or even Arabidopsis thaliana-specific (phylostratum 13). Next a co-expression analysis was performed using
publicly available gene-expression data, to identify the 300 most similarly expressed *Arabidopsis thaliana* genes for each of the 92 DUF295 genes. Finally, these 300 co-expressed genes were searched for overrepresentation of genes from the different phylostrata (Suppl. Table V) (Ruprecht et al. 2017). Based on this, co-expressed phylostrata were assigned to all DUF295 genes, giving an indication of the evolutionary age of their co-expression network.

For 82 DUF295 genes, one or more co-expressed phylostrata were identified (Suppl. Table V). When comparing the overall distribution of the Arabidopsis thaliana genome (represented by 32833 genes used in this analysis), the DUF295 gene family was particularly enriched in co-expression networks with genes from the Angiosperm and Eudicot phylostrata (Figure 6). This is in line with the presumed age of the DUF295 domain, which appears to have originated early on in the Angiosperm lineage. The specific group of 8 Brassicaceae-specific DUF295 organellar genes that are responsive to mitochondrial dysfunction was also enriched in co-expression networks consisting of Angiosperm and Eudicot phylostrata. Interestingly, none of these 8 genes are co-regulated with Brassicaceae-specific phylostrata, indicating they have been incorporated in co-expression network that is much older than the genes themselves. A similar analysis was performed for 21 ‘core’ mitochondrial retrograde target genes regulated by ANAC017, based on previous data (Suppl. Table V). This indicated that the strongest co-expression of core-ANAC017 target genes is also found with genes from Angiosperm phylostratum (Figure 6). Stronger co-expression of core ANAC017-target genes was also found in the Landplants phylostratum. In summary, this analysis further supports that the recent mitochondrial-stress responsive DUF295 genes have been adopted by a much older co-expression network, which is of largely similar age to the ANAC017 core regulon.

![Co-expressed Phylostrata](image)

**Figure 6. Phylostratum co-expression analysis.** For each gene the 300 most strongly co-expressed *Arabidopsis thaliana* genes were identified. Overrepresented evolutionarily conserved phylostrata were searched in these 300 genes. This analysis was done on all *Arabidopsis thaliana* genes (All Athal), all 94 DUF295-related genes (All DUF295), the 8 DUF295 genes responding to mitochondrial dysfunction or mitochondrial retrograde regulation (MRR DUF295) and on 21 genes that have been consistently
Characterisation of DUF295 T-DNA insertion mutants

To investigate the function of DUF295 genes in plants, we isolated T-DNA insertion mutants for representatives of the 3 main DUF295 groups (Figure 7A). For the ancestral F-box DUF295 genes, AtFDA11/SKIP23 was chosen, as it has been picked up in several protein-protein interaction screens with ASK1/SKP1-type proteins and 14-3-3 proteins (Risseeuw et al. 2003; Kuroda et al. 2012; Hong et al. 2017). It also had the second highest gene expression level in 2-week-old Col-0 based on RNA-seq data (Suppl. Table III). For Brassicaceae-specific F-Box DUF295 genes, AtFDB2 was selected as it was by far the most strongly expressed gene in this group. For DUF295 organellar genes AtDOA10 was selected as it showed the highest fold-change induction to mitochondrial dysfunction (Suppl. Table II). Also AtDOB5 and AtDOB12 were chosen because they were the most highly induced representatives of two different AtDOB tandem duplications (Suppl. Table II). They were also found to be targeted to mitochondria using GFP-fusions (Figure 2). Suitable T-DNA lines were selected from T-DNA express, and homozygous lines were isolated using PCR-based genotyping (Table II).

Firstly, the overall growth rate and phenotype of the mutant lines was compared to Col-0. The rosette surface area was monitored from 14 to 29 days after transfer to the growth room. However, no clear alterations in growth were found compared to Col-0 for any of the lines tested (Figure 7B). Also, no obvious phenotypical differences in plant appearance were observed. As the selected DUF295 organellar genes clearly responded to mitochondrial dysfunction, root growth inhibition by antimycin A and methylviologen was tested (Figure 7C-E). Again, no obvious differences in root growth and resistance to inhibitors was observed for any of the lines compared to Col-0. Overall, no clear aberrant phenotypes were observed for any of the T-DNA lines analysed, which is likely explained by the large extent of gene duplication leading to redundancy, for instance shown by often similar gene expression patterns of tandem duplications (Figure 3).
Figure 7. Phenotypic analysis of DUF295 T-DNA mutants. (A) Overview of T-DNA insertion locations for selected Arabidopsis thaliana DUF295 mutant lines. Black bars indicate coding exon, grey bars indicate untranslated regions in exons, black lines indicate introns. (B) Average rosette surface area of soil-grown plants of the different genotypes monitored over time. (C) Average primary root length after 7 days of growth on vertically-oriented MS plates, MS plates supplemented with antimycin A (D) or methylviologen (E). Asterisks indicate statistically significant difference compared to Col-0 (** p<0.01).

Discussion

Through extensive phylogenetic analysis of the DUF295 family, this study found that the F-box/DUF295 domain combination is the most prevalent and conserved configuration in Angiosperms (Figure 1). Most likely these types of proteins derived from F-box precursor proteins, and the DUF295 domain evolved gradually sometime after the Gymnosperm/Angiosperm divergence. A common factor in the limited functional information that is available on the FDA proteins is interaction with
SKP1/ASK1 type proteins, which are part of SCF-type ubiquitin E3 ligases. ASK1 seems to mediate the interaction of the F-box protein with CUL1 (Jeong et al. 2011). The DUF295 domain is likely to be also a protein-protein interaction domain that may be bridging ASK1 and other proteins such as Curly Leaf (CLF), a polycomb SET-domain protein, thereby marking them for degradation. Overexpression of the DUF295 protein UCL1 resulted in similar phenotypes as a loss-of-function mutant in CLF, in line with the model that the interaction results in proteasome-mediated degradation of CLF (Jeong et al. 2011). Given the large number of FDA proteins within the same species, it is likely that a large range of proteins may be post-translationally regulated by such a mechanism. The binding of at least 6 other AtFDA proteins with ASK1/SKP1 was shown using yeast two-hybrid assays, suggesting this is a common feature (Kuroda et al. 2012). AtFDA11/SKIP23 was also found to interact with 14-3-3 proteins, but could not be shown to be directly involved in ubiquitination (Hong et al. 2017). From the limited amount of information available, it thus seems that the DUF295 domain may be a protein-protein interaction domain. For the ancestral FDA proteins it may help recruit target proteins to SCF E3 ligases for proteasomal degradation.

After several rounds of gene duplication in Brassicaceae, a variant to the F-box DUF295 configuration seems to have arisen (FDB proteins). The yeast-two hybrid screens could not identify an interaction with ASK1 (Kuroda et al. 2012) for any of the 6 tested FDB proteins, indicating that the F-box domain has diverged significantly. Whether these proteins have obtained a different function is currently unclear. At least a single loss-of-function mutation in AtFDB2 did not result in obvious phenotypic differences, but this may be due to the extensive redundancy.

The other group of Brassicaceae-specific DUF295 gene variants have led to more radical rearrangement, with the loss of the F-box domain, and the gain of a functional/predicted organellar targeting peptide. Mitochondrial targeting of two AtDOB proteins was confirmed by GFP-fusions in this study (Figure 2), while proteomics identified at least one AtDOA protein in isolated mitochondria (Senkler et al. 2017). From an evolutionary standpoint, this represents a clear example of how (partial) gene duplication can result in new organellar proteins. If these new DUF295 Organellar proteins were not useful to plants, one would expect a fast accumulation of point mutations. However, at least for AtDOB proteins there seems to be some selection pressure, indicating that the genes are being kept in a functional state. Also, most DUF295 genes are expressed at the mRNA level, often in very specific patterns, suggesting they are not pseudogenes. However, their function remains unclear for now.

Assuming that the DUF295 domain is a protein-protein interaction domain, they may directly bind other proteins. Due to the loss of the F-box domain, this is unlikely to lead to ubiquitination and protein degradation of the potential binding partners. The DUF295 Organellar are generally only slightly shorter than FDA proteins (most between 350 and 400 amino acids), and the DUF295 domain is close to the C-terminal. Even without the F-box domain and the likely removal of the N-terminal organellar targeting peptide upon import, one would expect at least 200-250 amino acids present in
DUF295 proteins outside of the DUF295 domain itself. This would be more than sufficient for other (unknown) functions that are assisted by the DUF295 domain, or perhaps act as a flexible linker between the two domains. To some extent the organellar DUF295 proteins show similarities to microProteins, which are proteins that only contain a protein-protein interaction domain, but no other clear functional domains (Bhati et al. 2018). MicroProteins are thought to have regulatory effects for instance by preventing proteins from forming functional dimers, thus having dominant effects. MicroProteins have also been found in mitochondria, where they can bind mitochondrial elongation factors and stimulate mito-ribosome translation (Rathore et al. 2018). Further studies with gain/loss-of-function mutants and protein interaction screens may shed further light on the function of these evolving proteins.

A broad gene expression analysis revealed that only very few DUF295 genes are ubiquitously expressed (Figure 3), e.g. AtFDA11/SKIP23 and AtFDB2, which were selected for further study. Most other genes had relatively specific expression patterns. Besides two groups of stress-responsive DUF295 genes, most patterns were strongly biased towards young reproductive tissues, such as siliques, anthers and pollen. Such a bias towards expression of recently evolved genes in the male germ line ('out of testis') has been reported in animal systems. It has been proposed that male gametophytes act as an ‘innovation incubator’, driving species specification and providing new genes to support the arms race against microbial pathogens (Cui et al., 2015). Also other studies support the concept that gene duplication may be a mechanism to adapt organisms to variable environments (Kondrashov 2012). Furthermore, It has been observed in mammalian genomes that the ‘new’ copies of gene after a duplication event show much longer bursts of sequence evolution than the copies in the original location (Pich and Kondrashov 2014). Our data suggests this is also the case in plants, with both FDB and DOA/DOB groups expanding much more quickly than the ancestral FDA groups. This may be because the (incorrectly) duplicated genes have undergone significant functional changes and are thus evolving towards a new function. From each of these recent subgroups, several individual genes indeed appear to be ‘consolidating’ into conserved state (Figure 4), suggesting their functions are beneficial to the plant.

Considering the recently obtained mitochondrial localisation of several DUF295 Organellar proteins, it makes sense that many of them have become integrated in a mitochondrial signalling network, regulated by ANAC017 and its close homologs. This in contrast to the FDA and FDB proteins, which do not appear to be regulated by mitochondrial signalling. Thus, the regulatory information surrounding the coding sequences of the ancestral F-box DUF295 genes (e.g. promoter) may not have been strongly retained during the gene duplication events or at least considerably altered. The question is therefore how the integration into the mitochondrial signalling network occurred. One mechanism could be a gradual shift in regulation, whereby a non-mitochondrially regulated duplicated gene evolved novel regulatory information in the promoter, allowing it to perform its mitochondrial
function in a more directed way. An alternative mechanism that could explain both mitochondrial targeting and integration into the ANAC017 regulatory network, could be that the partially duplicated DUF295 gene integrated into a gene encoding an existing mitochondrial target protein. In this way the partial DUF295 gene may have ‘hi-jacked’ the targeting peptide sequence, as well as the promoter and regulatory information. In such an event, the protein would have instantly become mitochondrially targeted, and regulated by a relevant signalling pathway like ANAC017. The original mitochondrial protein was thereby probably lost.

This second alternative seems the most plausible, as the majority of DOA and DOB genes are predicted to be mitochondrially targeted, and members of both subgroups are ANAC017 regulated (Figure 5). Of the 7 DUF295 Organellar genes that are apparently regulated by ANAC017, 5 are present in tandem gene duplications. At5g54450, At5g54550 and At5g54560 form a consecutive group of three (from a total of 6 in close proximity) (see Table I), while At5g52930 and At5g52940 form a consecutive group (2 out of 2 at this locus). This suggests their co-regulation is possibly caused by co-duplication of regulatory information. This co-expression of neighbouring genes has been observed previously for unrelated MATE multidrug and toxin efflux carriers At2g04040, At2g04050 and At2g04070, which are also part of the ANAC017-regulated mitochondrial retrograde pathway (Van Aken et al. 2007; Van Aken et al. 2016). Possibly, the currently non-ANAC017-regulated organellar DUF295 genes may have lost some of their regulatory information over time.

In conclusion, this study provides compelling evidence for neo-functionalisation of proteins via inter-compartmental gene duplication in plants, thus adding to the lineage-specific organellar proteome (Figure 8) (Szklarczyk and Huynen 2009). Our study further shows that such duplications can then result in integration into existing and relevant gene regulatory networks, which can be considered as the next stage in the creation of new and useful function. The precise function of the DUF295 proteins is only beginning to be understood, especially of the Brassicaceae-specific subtypes. As they represent 0.3% of the total Arabidopsis thaliana protein-coding gene content, it is likely that future studies will find out more by both targeted and untargeted approaches.
Figure 8. Model for the evolution of the DUF295 gene family. Our analyses suggests that the DUF295-related domain evolved as an additional C-terminal domain to existing F-box proteins in early angiosperms (presumably 140-180 million years ago). The DUF295 domain was then consolidated in monocots and dicots forming the ancestral F-box DUF295 (FDA) protein family, which expanded extensively in the different species via (tandem) gene duplication events. More recently (approx. 32 million years ago), aberrant gene duplications specifically in the Brassicaceae resulted in divergent F-box DUF295 (FDB) proteins. The F-box domain was most likely lost by ‘faulty’ or incomplete gene duplication, and replaced with a mitochondrial or chloroplast targeting peptide (mTP/cTP), resulting in the Brassicaceae specific DUF295 Organellar (DOA and DOB) protein family. All DUF295 protein families appear to have expanded strongly by subsequent (tandem) gene duplication events.
Materials and Methods

Plant materials and growth conditions

Arabidopsis thaliana (L.) Heynh. Col-0 was used in all experiments. Seeds were sown on soil mix or MS media with 2% sucrose and stratified for 2-3 days at 4°C, then grown under long-day conditions (16 h light/8 h dark) at 22 °C and 100 µmol m⁻² s⁻¹. Previously published transgenic lines were obtained from Ng et al., (2013) anac017-2: SALK_022174. DUF295 mutant lines (Figure 5) were genotyped using PCR on genomic DNA using primers shown in Supplemental Table 4.

Stress treatments of plants

Seeds were sown on petri dishes containing MS medium (Duchefa-Biochimie) + 2% sucrose, stratified for 2-3 days in the cold room and then incubated in long day growth conditions for 14 days. Pools of plants were then collected before or after treatment and immediately placed in liquid nitrogen for storage and further processing. For transcript analysis plants were sprayed with 50 µM antimycin A. In vitro stress assays were performed as previously described (De Clercq et al. 2013). For root growth assays, the different plant lines were incubated on vertically positioned plates supplemented with 50 µM antimycin A or 20/50 µM methylviologen. Plants were stratified in the cold room for 3 days and incubated for 7 days in long day conditions. Primary root length was measured using ImageJ. Statistical analysis was performed using Student’s t-test.

Quantitative RT-PCR and Microarray analysis

RNA isolation, cDNA generation and quantitative RT-PCR (qRT-PCR) was performed as described in (Van Aken et al. 2013) using Spectrum RNA Plant extraction kits (Sigma-Aldrich, Sydney, Australia), iScript cDNA synthesis kit (Bio-Rad) and a Roche LC480 Lightcycler using SYBRgreen detection assays. All primers for qRT-PCR are shown in Supplementary Table 4. Relative expression values were normalised, with untreated Col-0 samples set as 1. Statistical analyses were performed using student’s t-test throughout the manuscript, except where indicated.

GFP localisation and microscopy

Coding sequences for full-length DUF295 genes were PCR amplified from Arabidopsis thaliana cDNA and cloned into the pDONR201 Gateway vector (Invitrogen, CA, USA). Cloning into the final GFP pDEST-CGFP vectors was done as described (Carrie et al., 2008b). The 42 amino acids targeting signal of alternative oxidase (AOX) was fused to RFP as a mitochondrial marker, and the Rubisco small subunit (SSU-RFP) as a plastid marker (Carrie et al., 2008). Biolistic co-transformation using gold particles of the GFP and RFP fusion vectors was performed on Arabidopsis cell culture as previously reported (Carrie et al., 2008). In brief, GFP and RFP plasmids (5 µg each) were co-precipitated onto
gold particles and transformed using a PDS-1000/He biolistic transformation system (Bio-Rad). 2-3 mL of Arabidopsis suspension cell culture (4-5 days after 6x dilution of a 1-week-old culture in fresh medium) were placed on osmoticum medium and bombarded. Cells were then incubated for 24-48 h at 22°C in the dark. GFP and RFP expression and targeting were visualized using a BX61 Olympus microscope (Olympus) using excitation wavelengths of 460/480 nm (GFP) and 535/555 nm (RFP), and emission wavelengths of 495–540 nm (GFP) and 570–625 nm (RFP). Subsequent images were captured using Cell® imaging software.

**Phylogenetic analysis**

*Arabidopsis thaliana* DUF295 genes were identified using a combination of searches for PFAM motif PF03478, TAIR10 annotation and homology searches. Representative DUF295 genes from other plant species were obtained using homology searching. Protein sequences were aligned using MAFFT multiple sequence aligner (Katoh and Standley 2013) and edited in BioEdit. Phylogeny was inferred using the IQ-Tree webserver (http://iqtree.cibiv.univie.ac.at/) using BLOSUM62 algorithm and 1000 bootstraps (Katoh and Standley 2013). Phylogenetic trees were visualised using FigTree v1.4.2.

**Gene Duplication analysis**

Starting from the set of DUF295 genes reported in Table 1, the PLAZA 4.0 Dicots comparative genomics platform was used to retrieve information about gene duplications (Van Bel et al. 2018). Specifically, the PLAZA Workbench was used to define different gene sets and to determine the number of genes involved in a tandem gene duplication event. In the PLAZA database, tandem gene duplicates were identified using i-ADHoRe v3.0.01 (gap_size 30, tandem_gap 30, cluster_gap 35, q_value 0.85, prob_cutoff 0.01, anchor_points 5, multiple_hypothesis_correction FDR) (Proost et al. 2012).

**1001 genomes SNP analysis**

The 1001 Genomes polymorph tool (https://tools.1001genomes.org/polymorph/) was searched for single nucleotide polymorphisms with high impact for all 94 *Arabidopsis thaliana* DUF295-related genes. The number of ecotypes where specific polymorphisms compared to Col-0 occurred, were added.

**Phylostratum and expression analysis**

Starting from the transcript counts reported by Vaneechoutte and co-workers (Vaneechoutte et al. 2017), a gene expression matrix was generated by summing transcript counts per locus. Subsequently, for each gene the top 300 co-expressed genes (denoted knn300 cluster) were determined based on the Pearson Correlation Coefficient. Starting from phylostrata information derived from gene families defined in PLAZA 3.0 Dicots (Proost et al. 2015), significantly over-represented phylostrata per knn300
cluster were identified using the hypergeometric distribution (incl. Benjamini-Hochberg correction for multiple hypothesis testing). All enrichments with corrected p-value < 0.05 were retained as significant. Expression patterns of DUF295 genes in 206 samples (from the Vaneechoutte et al., 2017 dataset) were examined in an expression heatmap (Figure 3, Suppl. Figure 2). For this, TPM expression values were first normalized for each gene by dividing them with the maximum TPM observed for that gene. Only genes with a maximum TPM larger than 2 were considered to be expressed and others were excluded from the heatmap. No expression data was available for AtFDA4 and so it was excluded from this analysis as well. Figure 3 shows a manually selected subset of 25 samples to highlight interesting expression behavior of the DUF295 genes.

Acknowledgements
OVA was supported by the Swedish Research Council (Vetenskapsrådet 2017-03854), the Australian Research Council (DP160103573), Crafoord Foundation (20170862), Carl Trygger Foundation (CTS 17: 487) and Carl Tesdorpf Stiftelse. MB, KB and ZA were supported by the ARC Centre of Excellence in Plant Energy Biology (CE140100008). D.V. was supported by the Agency for Innovation by Science and Technology (IWT) in Flanders (predoctoral fellowship).

Author contributions
The study and experiments were designed by OVA, KV and TS. Experiments were performed by SL, MB, ZA, KB, DV, KV and OVA. The manuscript was written by OVA with contributions from the co-authors.


Table I. Overview of DUF295-related genes in *Arabidopsis thaliana*. A. *trich*: conserved in *Amborella trichopoda*; Mono/Dicot: conserved in monocots and dicots; Brassic-only: only conserved in Brassicaceae; F-box: protein contains an F-box domain; DUF295: protein contains a DUF295 domain according to the PFAM motif; SUBAcon: subcellular location as suggested by the SUBAcon algorithm; MS: subcellular location as detected by mass spectrometry (Zybailov et al. 2008; Hummel et al. 2012; Senkler et al. 2017); ASK1 binding: experimentally protein-protein interaction with SKP1/ASK1 proteins; MRR: transcriptionally regulated by mitochondrial retrograde signalling; Co-expressed phylostratum: overrepresented phylostrata in the 300 most strongly co-expressed genes (Ang: angiosperms; Eud: eudicots; Vir: viridiplantae; Brass: Brassicaceae; Land: land plants).

<table>
<thead>
<tr>
<th>AGI</th>
<th>Gene name</th>
<th>A. trich</th>
<th>Mono/Dicot</th>
<th>Brassic-only</th>
<th>F-box</th>
<th>DUF295</th>
<th>SUBAcon</th>
<th>MS</th>
<th>ASK1 binding</th>
<th>MRR</th>
<th>Co-expressed phylostratum</th>
</tr>
</thead>
<tbody>
<tr>
<td>AT1G44080</td>
<td>AtFDA1</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>nu</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
</tr>
<tr>
<td>AT1G57906</td>
<td>AtFDA2</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AT1G64840</td>
<td>AtFDA3</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>cyt,pm</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Vir</td>
</tr>
<tr>
<td>AT1g65375</td>
<td>AtFDA4</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AT1G65740</td>
<td>AtFDA5/UCL1</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Eud Brass</td>
</tr>
<tr>
<td>AT1G65760</td>
<td>AtFDA6</td>
<td></td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>pm</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AT1G65770</td>
<td>AtFDA7</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Eud Brass</td>
</tr>
<tr>
<td>AT2G16290</td>
<td>AtFDA8</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>pm</td>
<td>-</td>
<td>-</td>
<td>Ang Eud Brass</td>
<td></td>
</tr>
<tr>
<td>AT2G16300</td>
<td>AtFDA9</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>pm</td>
<td>-</td>
<td>-</td>
<td>Ang Eud Brass</td>
<td></td>
</tr>
<tr>
<td>AT2G16365</td>
<td>AtFDA10</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>nu</td>
<td>-</td>
<td>-</td>
<td>Vir</td>
<td></td>
</tr>
<tr>
<td>AT2G17030</td>
<td>AtFDA11/SKIP23</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>golgi</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Vir</td>
</tr>
<tr>
<td>AT2G17036</td>
<td>AtFDA12</td>
<td></td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AT2G17690</td>
<td>AtFDA13</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>cyt</td>
<td>cyto</td>
<td>+</td>
<td>-</td>
<td>Ang Eud Brass</td>
</tr>
<tr>
<td>AT2G24250</td>
<td>AtFDA14</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>Land</td>
</tr>
<tr>
<td>AT2G24255</td>
<td>AtFDA15</td>
<td></td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>pm</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT2G26160</td>
<td>AtFDA16</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT3G25750</td>
<td>AtFDA17</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>nu</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT4G35733</td>
<td>AtFDA18</td>
<td></td>
<td></td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT5G24040</td>
<td>AtFDA19</td>
<td></td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Land</td>
<td></td>
</tr>
<tr>
<td>AT5G60060</td>
<td>AtFDA20</td>
<td></td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>AT1G10110</td>
<td>AtFDB1</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT1G27540</td>
<td>AtFDB2</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>plastid</td>
<td>-</td>
<td>-</td>
<td>Vir</td>
<td></td>
</tr>
<tr>
<td>AT1G27550</td>
<td>AtFDB3</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>Brass</td>
<td></td>
</tr>
<tr>
<td>AT1G27580</td>
<td>AtFDB4</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT1G67160</td>
<td>AtFDB5</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT1G69900</td>
<td>AtFDB6</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>plastid</td>
<td>-</td>
<td>-</td>
<td>Brass</td>
<td></td>
</tr>
<tr>
<td>AT2G03560</td>
<td>AtFDB7</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud Brass</td>
<td></td>
</tr>
<tr>
<td>AT2G03610</td>
<td>AtFDB8</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT2G04810</td>
<td>AtFDB9</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT2G04830</td>
<td>AtFDB10</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Brass</td>
<td></td>
</tr>
<tr>
<td>AT2G04840</td>
<td>AtFDB11</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT2G05970</td>
<td>AtFDB12</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud Brass</td>
<td></td>
</tr>
<tr>
<td>AT2G14290</td>
<td>AtFDB13</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Land</td>
<td></td>
</tr>
<tr>
<td>AT2G14500</td>
<td>AtFDB14</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Ang Eud</td>
<td></td>
</tr>
<tr>
<td>AT2G24080</td>
<td>AtFDB15</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>Eud</td>
<td></td>
</tr>
<tr>
<td>AT2G33190</td>
<td>AtFDB16</td>
<td></td>
<td></td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>cyt</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Accession</td>
<td>GeneID</td>
<td>AtDOB</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>-------</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>-----</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>AT5G54450</td>
<td>AtDOB11</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>AT5G54550</td>
<td>AtDOB12</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>AT5G54560</td>
<td>AtDOB13</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>AT5G55270</td>
<td>AtDOB14</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>plastid</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AT5G55870</td>
<td>AtDOB15</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AT5G55880</td>
<td>AtDOB16</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AT5G55890</td>
<td>AtDOB17</td>
<td></td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>mito</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Figure Legends

**Figure 1. Phylogenetic analysis of the DUF295 protein family.** (A) Unrooted phylogenetic tree of proteins containing DUF295 domains in representative Angiosperm species. Scale bar indicates percentage divergence. For clarity gene and species names have been removed, but information on dicot, monocot or Amborella trichopoda is indicated by different circles (see figure). A fully annotated phylogenetic tree with species/gene names and bootstrap values can be found in Supplementary Figure 2. (B) General domain structure of the four groups of DUF295 proteins found in Brassicaceae. mTP: mitochondrial targeting peptide; cTP: chloroplast transit peptide.

**Figure 2. DUF295 organellar proteins are targeted to the mitochondria.** C-terminally GFP-tagged fusion proteins were transiently transformed into Arabidopsis thaliana cell culture, and co-transformed with mitochondrial marker AOX-RFP. Scale bar indicates 10 µm.

**Figure 3. Expression patterns of DUF295 genes.** Expression values, normalized per gene, are shown for 72 expressed DUF295 genes. Gene names are colored to indicate the family subgroups (see Figure 1). Only a subset of 25 samples is shown. Expression data for all 206 samples in the Vaneechoutte et al., 2017 dataset is available in Suppl. Figure 4.

**Figure 4. High impact mutations identified by the 1001 Genomes tool.** (A) Accumulation of high-impact mutations (HIMs) in 1135 Arabidopsis thaliana accessions compared to Col-0 were plotted against the gene expression levels (transcripts per million, max tpm). Genes were color-coded by DUF295 subgroup. The red dashed lines indicate the cut-offs used for Figure 4B. (B). Percentage of genes within each DUF295-related subgroup that are postulated to be ‘consolidated’, ‘degenerating’ or ‘neutral’, based on the cut-offs in Figure 4A.

**Figure 5. DUF295 organellar genes are incorporated into mitochondrial retrograde signalling networks.** Two-week-old Col-0 and anac017 mutant plants were treated with antimycin A and samples were collected in triplicate pools of plants. mRNA levels were quantified using qRT-PCR and normalised to Col-0 at time point 0h. Asterisks indicate statistically significant difference in expression level compared to time point 0 in the same genotype (* p<0.05; ** p<0.01); Hash-tags indicate significant difference at the same time point between Col-0 and anac017 (# p<0.05; ## p<0.01).

**Figure 6. Phylolstratum co-expression analysis.** For each gene the 300 most strongly co-expressed Arabidopsis thaliana genes were identified. Overrepresented evolutionary conserved phylolstrata were searched in these 300 genes. This analysis was done on all Arabidopsis thaliana genes (All Athal),
all 94 DUF295-related genes (All DUF295), the 8 DUF295 genes responding to mitochondrial dysfunction or mitochondrial retrograde regulation (MRR DUF295) and on 21 genes that have been consistently found to be regulated by ANAC017 in MRR. The percentage indicates how many genes within this category showed statistically enriched co-expression with member of the indicated phyllostrata.

**Figure 7. Phenotypic analysis of DUF295 T-DNA mutants.** (A) Overview of T-DNA insertion locations for selected *Arabidopsis thaliana* DUF295 mutant lines. Black bars indicate coding exon, grey bars indicate untranslated regions in exons, black lines indicate introns. (B) Average rosette surface area of soil-grown plants of the different genotypes monitored over time. (C) Average primary root length after 7 days of growth on vertically-oriented MS plates, MS plates supplemented with antimycin A (D) or methylviologen (E). Asterisks indicate statistically significant difference compared to Col-0 (** p<0.01)

**Figure 8. Model for the evolution of the DUF295 gene family.** Our analyses suggests that the DUF295-related domain evolved as an additional C-terminal domain to existing F-box proteins in early angiosperms (presumably 140-180 million years ago). The DUF295 domain was then consolidated in monocots and dicots forming the ancestral F-box DUF295 (FDA) protein family, which expanded extensively in the different species via (tandem) gene duplication events. More recently (approx. 32 million years ago), aberrant gene duplications specifically in the Brassicaceae resulted in divergent F-box DUF295 (FDB) proteins. The F-box domain was most likely lost by ‘faulty’ or incomplete gene duplication, and replaced with a mitochondrial or chloroplast targeting peptide (mTP/cTP), resulting in the Brassicaceae specific DUF295 Organellar (DOA and DOB) protein family. All DUF295 protein families appear to have expanded strongly by subsequent (tandem) gene duplication events.
Supplemental Figure 1. Sequence Logo of the DUF295 domain (PFAM PF03478).

Supplemental Figure 2. Extended phylogenetic tree of DUF295 proteins. Rooted phylogenetic tree of proteins containing DUF295 domains in representative Angiosperm species, produced as described in the methods section. Scale bar indicates percentage divergence. Color coding for the different subgroups: green (F-box DUF295 Ancestral), red (F-box DUF295 Brassicaceae), blue (DUF295 Organellar A), purple (DUF295 Organellar B), pink (F-box DUF295-related).

Supplemental Figure 3. AtDOBS is targeted to the mitochondria and plastids. C-terminally GFP-tagged fusion protein was transiently expressed in Arabidopsis thaliana cell culture, and co-transformed with mitochondrial marker Mito-mCherry or plastid marker SSU-RFP. Scale bar indicates 10 µM.

Supplemental Figure 4. Gene expression cluster. Gene expression analysis of the Arabidopsis thaliana DUF295 gene family in a set of 206 RNAseq experiments, normalised per gene. A. trich: conserved in Amborella trichopoda; Mono/Dicot: conserved in monocots and dicots; Brassic-only: only conserved in Brassicaceae; F-box: protein contains an F-box domain; DUF295: protein contains a DUF295 domain according to the PFAM motif; SUBAcon: subcellular location as suggested by the SUBAcon algorithm; MS: subcellular location as detected by mass spectrometry; ASK1 binding: experimentally protein-protein interaction with SKP1/ASK1 proteins; MRR: transcriptionally regulated by mitochondrial retrograde signalling; Co-expressed phylostratum: overrepresented phylostrata in the 300 most strongly co-expressed genes (Ang: angiosperms; Eud: eudicots; Vir: viridiplantae; Brass: Brassicaceae; Land: land plants); max tpm: maximum transcripts per million in the RNA-seq data sets (genes with yellow values are considered to be transcriptionally active, red are likely transcriptionally inactive pseudogenes).
Supplementary Figure 1. Sequence Logo of the DUF295 domain (PFAM PF03478).