Identification of the functions of 4-coumarate-CoA ligase/ acyl-CoA synthetase paralogs in potato

Running title: 4-coumarate-CoA ligase potato family

Matías Ariel Valiñas, Arjen ten Have, Adriana Balbina Andreu*

Instituto de Investigaciones Biológicas-CONICET, Universidad Nacional de Mar del Plata, CC 1245, 7600 Mar del Plata, Argentina

*Correspondence:
Adriana Balbina Andreu
abandreu@mdp.edu.ar

NUMBER OF WORDS: 9,712

NUMBER OF FIGURES: 7

NUMBER OF TABLES: 2

NUMBER OF SUPPLEMENTARY FIGURES: 18

NUMBER OF SUPPLEMENTARY TABLES: 4

ABSTRACT

Background: The 4CL/ ACS protein family is well known for its 4-coumarate-CoA ligase (4CL) enzymes but there are many aspects of this family that are still unclear or generally known. Cytosolic class I and class II 4CL enzymes control the biosynthesis of lignin/ suberin and flavonoids, respectively. Many 4CL homologs have broad substrate permissiveness *in vitro* and have no clear cut function. However, it has been demonstrated unequivocally that a peroxisomal 4CL-like homolog from Arabidopsis efficiently uses p-coumarate for ubiquinone biosynthesis. Another homolog has been shown to act as a fatty acyl-CoA synthetase and yet another as OPDA-CoA ligase. Hence, despite this knowledge, most homologs remain annotated as "4CL-like" whereas other researches study the ACS protein family.

Results: We set out identify the specific functions of 4CL/ ACS homologs, specifically in order to study the 4CL family in *Solanum tuberosum.* An in depth phylogenetic analysis was done. Using clustering techniques, functional annotation and taxonomic signals, three major clades were depicted. Clade 1 is composed of class I from monocotyledons, class I from dicotyledons and class II canonical 4CL enzymes subclades. Specificity determining positions and 3D structure analysis shows that clade 2 cytosolic 4CL-like enzymes show a rather different binding cleft and presumably use medium- to long-chain fatty acids. Clade 3 is composed of five subclades, four of which have a broad taxonomic contribution and a similar binding cleft as 4CLs whereas a fifth, specific for dicotyledons shows a significantly different binding pocket. The potato 4CL family comprises four class I (St4CL-I(A-D)) and one class II (St4CL-II) members. Transcript levels of St4CLs and of marker genes of the flavonoid (chalcone synthase, *CHS*) and suberin (feruloyl-CoA transferase, *FHT*) pathways were determined by qRT-PCR in flesh and skin from Andean varieties. *St4CL-IA* was barely detected in the skin of some varieties whereas *St4CL-IB* did not show a clear pattern. *St4CL-IC* and *St4CL-ID* could not be detected. *St4CL-II* expression pattern was similar to *CHS*. *St4CL-IA* and *St4CL-IB* were induced by wounding as did *FHT* whereas *St4CL-II and CHS* expression was repressed. Constitutive and wound-induced expression suggests that *St4CL-IA* and *St4CL-IB* isoforms are likely involved in soluble and/ or suberin-bound phenolic compounds while *St4CL-II* appears to be involved in flavonoid biosynthesis.

Key words: 4-coumarate-CoA ligase - suberin - anthocyanin - potato – substrate permissiveness - functional diversification – HMMERCTTER

INTRODUCTION

Potato (*Solanum tuberosum* L.) is an important staple food crop worldwide. It is considered a cheap source of high-quality proteins, minerals, and antioxidants, including vitamin C, carotenoids, and phenolic compounds. Chlorogenic and caffeic acids are the most abundant phenolic acid antioxidants found in white- and yellow-fleshed commercial potato varieties (Valiñas et al., 2015). Colored Andean potatoes additionally contain antioxidant phenolic compounds such as anthocyanins (Navarre et al., 2011) that can provide flesh and skin of potato tubers with red and blue colors.

Anthocyanins are flavonoid compounds consisting of an anthocyanidin aglycone bound to one or more sugar moieties. Flavonoid biosynthesis starts with p-coumaroyl-CoA, the activated product of 4-coumaric acid by 4-coumarate-CoA ligase (4CL: Figure 1). The condensation of one molecule of p-coumaroyl-CoA and three molecules of malonyl-CoA by the enzyme chalcone synthase (CHS) yields naringenin chalcone. Consecutive action of chalcone isomerase, flavonoid 3-hydroxylase, dihydroflavonol 4-reductase, anthocyanidin synthase and UDP-glucose anthocyanidin 3-O-glucolsyltransferase leads to the production of anthocyanins. Additional glycosilations and acylations can result in a large spectrum of anthocyanins, all with different colors. Anthocyanin biosynthesis has been detailed in various plant species including potato (Eck et al., 1994; De Jong et al., 2004; Stushnoff et al., 2010). Constitutive and stress-induced levels of *CHS* are well correlated to anthocyanin levels in potato tubers (André et al., 2009; Payyavula et al., 2012; Valiñas et al., 2017).

The cytosolic 4CL enzymes also catalyze the activation of several other hydroxycinnamic acids such as caffeic acid, ferulic acid, 5-hydroxyferulic acid, and sinapic acid into their corresponding CoA esters (Knobloch and Hahlbrock, 1975, 1977; Peter and Neale, 2004). Rather than being involved in flavonoid and anthocyanin biosynthesis, these activated phenolic acids are precursors for the biosynthesis of two major plant cell wall polymers: lignin and suberin (Figure 1). Note that whereas caffeoyl-CoA, feruloyl-CoA, 5-hydroxyferuloyl-CoA and sinapoyl-CoA are precursors of lignin/ suberin biosynthesis, p-coumaroyl-CoA serves as substrate of both the lignin/ suberin and flavonoid pathways. In addition, at least one peroxisomal 4CL acting on 4-coumaric acid has been identified in Arabidopsis, where it has been shown to be involved in the biosynthesis of ubiquinone (Block et al., 2014).

Suberin deposition rather than lignification is the predominant cell wall modification that takes place in the periderm of tubers, especially post harvest. The native periderm protects the tubers from pathogen invasion and dehydration. When the potato skin is wounded, the tuber reacts rapidly to restore the barriers by forming wound periderm that achieves impermeability and chemical defense for the newly exposed fleshy tissues. Suberin is a complex biopolymer comprising both polyaliphatic and polyaromatic domains (Kolattukudy, 2001; Franke and Schreiber, 2007; Graça and Santos, 2007; Pollard et al., 2008; Bernards, 2011). The polyaliphatic domain consists of a fatty acid polyester with esterified ferulic acid (Graça and Pereira, 2000; Schreiber et al., 2005). The polyaromatic domain is a lignin-like polymer that mostly contains hydroxycinnamic acids (Yan and Stark, 2000; Bernards and Razem, 2001). Besides the developmental requirement of suberin, wounding and abiotic stresses activate the biosynthesis of suberin through the induction of genes such as KCS (3-ketoacyl-CoA synthase: Franke et al. (2009)), FAR (fatty acid reductase: Domergue et al. (2010)), and FHT (feruloyl-CoA transferase: Boher et al. (2013)). FHT transfers feruloyl-CoA to the alcoholic group of the fatty acid derivative of the polyaliphatic domain of suberin (Serra et al. (2010); see Figure 1).

4-coumarate-CoA ligase is a member of the so-called ANL superfamily of adenylating enzymes. This superfamily derives its name from the three constituent families: the **A**cyl-CoA synthetase (ACS or 4CL/ ACS) family, to which 4CLs belong; the adenylation domain of **N**onribosomal peptide synthetases; and the **L**uciferase family (Gulick, 2009). It was formerly referred to as the adenylate-forming superfamily (Babbitt et al., 1992) because all members of this superfamily share part of the adenylation reaction, even although the overall reactions catalyzed by these enzymes are diverse. Hence, the presence of a highly conserved putative AMP-binding domain which contains a serine/ threonine/ glycine (STG)-rich region followed by a proline/ lysine/ glycine (PKG) triplet (box I, PROSITE PS00455 consensus sequence [LIVMFY]-X-X-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-X-[PASLIVM]-[KR]) has been used as the most important criterion in establishing this superfamily (Fulda et al., 1994). The GEICIRG motif (box II), which absolute conservation has been suggested to be restricted to 4CLs, was particularly deemed a signature sequence for this family (Becker-André et al., 1991).

Based on phylogenetic analyses, 4CL enzymes are divided into class I and class II (Ehlting et al., 1999). The consensus is that class II *4CL* genes are associated with flavonoid biosynthesis since they are mainly expressed in pigmented tissues and are induced by UV treatment. Class I *4CL* genes are, instead, involved in the biosynthesis of lignin and other cell wall phenylpropanoids because their expression is induced by wounding and confined to stems and roots (Ehlting et al., 1999; Cukovic et al., 2001). Correspondingly, substrate preferences between class I and class II enzymes have been suggested. Although substrate preferences among isoforms have been reported indeed, there appears to be no clear trend of substrate preferences among classes. For example, *Arabidopsis thaliana* class I At4CL1 best utilized p-coumaric, caffeic, ferulic and 5-hydroxyferulic acids as substrates, whereas class I At4CL2 readily transformed p-coumaric and caffeic acids into the corresponding CoA esters, while ferulic and 5-hydroxyferulic acids were converted quite poorly. By contrast, while class I At4CL4 is the only isoform capable of ligating sinapic acid, the two preferred substrates were 5-hydroxyferulic and caffeic acids. On the other hand, class II At4CL3 displayed broad substrate specificity efficiently converting p-coumaric, caffeic and ferulic acids into their CoA esters, whereas 5-hydroxyferulic acid was not as effectively utilized (Costa et al., 2005). So, studying the functionality of family members is as such severely hampered by broad substrate permissiveness.

Formation of CoA thioesters by 4CL occurs through a two-step reaction mechanism, involving the transient formation of a hydroxycinnamate-AMP anhydride in the presence of ATP and Mg^{2+} (*adenylation step*), followed by nucleophilic attack on the carbonyl carbon of the acyl adenylate by the phosphopantetheine thiol of CoA (*thioesterification step*) to yield the product thioester. Enzymes require a rotation of the C-terminal domain regarding the N-terminal domain along the hinge loop. Thereby, they have an adenylate-forming conformation during the first half-reaction and adopt a thioester-forming conformation during the second step. The determination of crystal structures of members of this enzyme family allowed a preliminary understanding of the roles of certain catalytic residues within conserved motifs. Interestingly, the conservation of several regions that are located at some distance from the active site of the first structures could be rationalized on the basis of the domain rearrangements (Gulick, 2009).

A number of species, such as *Oryza sativa* and *A. thaliana*, exhibit a large structurally, and apparent evolutionary diverse 4CL family (Ehlting et al., 1999; Hamberger and Hahlbrock, 2004; Sun et al., 2013). *S. tuberosum* has nine *4CL* genes

according to the KEGG integrated database resource for gene and protein annotation (Kanehisa et al., 2016). Until now, only two nearly identical class I isoforms have been described in potato (Becker-André et al., 1991). A similar situation occurs for tobacco where only two members have been described (Lee and Douglas, 1996).

Typically, the functional protein annotation is made throughout a similarity database search (similar sequence patterns, shared structural motifs, etc.). However, in many cases it allows to assign a general function to a protein (e.g., "AMP-binding protein"), but cannot solve the protein's specificity (i.e. "medium- and long-chain fatty acid-CoA ligase or 4-coumarate-CoA ligase"). Even worse, it can lead to erroneous transitory or, and that is more general, aspecific annotations. For instance, an initial *in silico* analysis revealed that the Arabidopsis genome has 14 genes annotated as putative 4-coumarate-CoA ligases. Eleven of these were expressed heterologously and tested for activity. Only four were catalytically active *in vitro* towards known 4CL hydroxycinnamate substrates, confirming that the *4CL* gene family in Arabidopsis has at least four members. The remaining seven enzymes were designated as 4CL-like proteins (Costa et al., 2005). Importantly, this finding was in agreement with previous phylogenetic analyses (Costa et al., 2003; Shockey et al., 2003). Independently, the enzymes encoded by several 4CL-like genes were shown to be acyl-coenzyme A synthetases (ACOS) that accept medium- to long-chain fatty acids and in some cases the cyclopentenone 12-oxo-phytodienoic acid (OPDA) and/ or OPDA derivatives which are precursors of plant defense hormone jasmonic acid (JA) (Schneider et al., 2005; Kienow et al., 2008). Particularly, it was demonstrated that the Arabidopsis 4CL-like gene *At1g62940* (ACOS5) is a medium- to long-chain fatty acyl-CoA synthetase required in tapetal cells for sporopollenin monomer biosynthesis (de Azevedo Souza et al., 2009). It was also shown that the Arabidopsis 4CL-like genes *At5g63380*, *At4g05160*, *At1g20500* and *At1g20510* have the capacity to activate some JA precursors *in vitro* (Schneider et al., 2005). Nevertheless, only At1g20510, designated as OPCL1, has been shown to encode a peroxisomal OPDA-CoA ligase involved in JA biosynthesis *in vivo* (Koo et al., 2006; Kienow et al., 2008). However, more recently 4CL activity with similar kinetics as found for *bona fide* 4CLs was reported for a peroxisomal 4CL-like protein (At4g19010) involved in the biosynthesis of ubiquinone (Block et al., 2014). Interestingly, its sequence lacks the GEICIRG motif (box II) that was presumed required for 4CL activity. Thus, much remains to be investigated in order to shed light on functional redundancy and diversification of the 4CL protein family on the one hand and substrate specificity and permissiveness on the other hand.

Many aspects of protein function contribute to the evolution of the family. These may include the global conservation of catalytic mechanisms (in the case of enzymes), specific binding to substrates and cofactors, as well as the interaction with other proteins in processes such as cell signaling, the regulation of reactions and the formation of macromolecular complexes. A subtler pattern of conservation is represented by the positions that are differentially conserved within subfamilies. A commonly accepted hypothesis is that whereas fully conserved positions are related to functional features common to all the members of the family, these other residues are related to functional specificity (e.g. binding of different substrates or cofactors). For this reason, they have been termed "specificity determining positions" (SDPs). These sites generally determine protein specificity either by binding specific substrate/ inhibitor or through interaction with other proteins (Rausell et al., 2010).

The aim of the present work was to study the 4CL/ ACS enzyme family with emphasis on *Solanum tuberosum* members. For that, we combined a phylogenetic analysis, clustering techniques and SDP analyses in order to determine how the 4CL/

ACS family has evolved and to predict the function of subfamily members. A transcript analysis of publicly available data was performed and a qRT-PCR analysis of class I and II *4CL* genes was done to identify which of the genes are functional in flavonoid and lignin/ suberin production in tubers.

MATERIALS AND METHODS

Plant material

Four Andean varieties of *Solanum tuberosum* ssp. *andigena* (Table 1) were grown in fields located in Quebrada de Humahuaca, Jujuy, Argentina during the 2010/ 2011, 2011/ 2012 and 2016/ 2017 campaigns. All varieties were planted on the same date in random plots and harvested at the end of their respective cycles. For each variety from the 2010/ 2011 and 2011/ 2012 campaigns, skin and flesh from ten freshly harvested tubers were pooled to generate a representative sample. The material was immediately frozen in liquid nitrogen and stored at -80 ºC until analysis.

Wound experiment

S. tuberosum spp *andigena* cv. Santa María potatoes (2016/ 2017 campaign) were used for the wound healing experiment. For this, tubers were peeled and cut into slices (2-3 mm thick) and left to heal at room temperature in saturated humidity and darkness. Samples of slices from each tuber were collected after 72 h, frozen in liquid nitrogen and stored at -80 ºC. Samples of tuber native periderm were also collected and immediately frozen. For unwounded samples, immediately prepared slices were frozen in liquid nitrogen and stored at -80°C. All samples were lyophilized and completely powdered.

RNA extraction, cDNA synthesis, primer design, qRT-PCR experiments and RNAseq

RNA was extracted from tuber samples using the CTAB (cetyltrimethylammonium bromide) method (Li et al., 2011). cDNA was synthesized using 2 µg total RNA, previously treated with RNase-free DNase I (Invitrogen, Carlsbad, CA), anchored oligo(dT) 15 VN primers and MMuLV reverse transcriptase according to the manufacturer's description (Invitrogen).

Taking in mind the exon/ intron structure of *St4CL* genes, isoform-specific primers (listed in Supplementary Table 1) were designed to anneal in the 3' coding region of *St4CL* genes flanking the last variable-length intron (Supplementary Figure 1A). The amplicon length of PCR products using as a template gDNA from *Solanum phureja* confirms primer specificity (Supplementary Figure 1B).

Relative transcript levels were determined by qRT-PCR in a 10 µL reaction volume with 20 ng RNA equivalent cDNA, 300 nM gene-specific primers and 5 µL SYBR Green Mix (Roche, Mannheim, Germany). Amplification was done using StepOneTM Real-Time PCR System according to the manufacturer's description (Applied Biosystems, Foster City, CA). Relative expression was calculated by the ΔCT method (Livak and Schmittgen, 2001) by normalizing the CT levels of target genes to the geometric mean of CT levels of the housekeeping gene cytoplasmic ribosomal protein L2.

RNAseq analysis was performed at http://bar.utoronto.ca/efp_potato/cgi-bin/efpWeb.cgi (Winter et al., 2007) using standard settings and the dataset provided by the PGSC (Massa et al., 2011). Note that these do not include data for St4CL-H5 and StACOS since these were not included in PGSC v4.03.

Identification, MSA and phylogeny of *4CL* gene family

Sequences for computational analysis were identified using an iterative HMMER approach. First, an initial training-set of sequences was obtained by a specific

PHI-BLAST (Zhang et al., 1998) analysis using the PROSITE PS00455 consensus sequence [LIVMFY]-X-X-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-X-[PASLIVM]-[KR] for the putative AMP-binding domain signature as pattern and At4CL1 and Os4CL1 as queries against the UniProtKB/ Swiss-prot database restricted to Plants. Sequences with an Expect-value $\leq 1e^{-5}$ were selected and aligned using MAFFT GINSI at https://mafft.cbrc.jp/alignment/server/index.html (Katoh and Standley, 2013). One sequence from *Solanum tuberosum 4CL* homologs identified in both the NCBI database (See Table 2) and the KEGG database was added to the above mentioned alignment with MAFFT-add (https://mafft.cbrc.jp/alignment/server/add.html). The multiple sequence alignment (MSA) was used to generate a position-specific scoring table (hidden Markov model, hmm) using the hmmbuild tool from the HMMER suite (Eddy, 2009). This model was used to search a compiled fasta file containing all the complete proteomes of 17 selected species (Supplementary Table 2), using the hmmsearch tool. The score of the lowest scoring training sequence was used as a specific cut-off. A new MSA and corresponding hmmer profile with novel, more sensitive cut-off threshold, were constructed and the process was iterated until convergence, at which point no new information was obtained when a new data-mining cycle was done. Redundant sequences showing 100% identity were eliminated using CD-HIT (Li and Godzik, 2006). Sequences of three 4CL proteins with resolved structure available from the PDB database were added to the above MSA using MAFFT-add. To use only phylogenetically informative regions for the reconstruction of phylogenetic trees, the MSAs were trimmed using BMGE (Block Mapping and Gathering with Entropy) (Criscuolo and Gribaldo, 2010). BMGE optional arguments (h = 0.8, gap settings by default) were determined based on the conservation of secondary structure elements. For this purpose, an excerpt of an alignment containing both fully and trimmed selected sequences against a sequence of *Nicotiana tabacum* 4CL2 (PDB identifier 5bsm) was done using ESPript 3 web server at http://espript.ibcp.fr/ESPript/cgi-bin/ESPript.cgi (Robert and Gouet, 2014) (See Supplementary Figure 2). To estimate the local reliability of protein MSAs the transitive consistency score (TCS) index, an extended version of the T-Coffee scoring scheme, was used (http://tcoffee.crg.cat/tcs). Briefly, sequences were aligned resulting in an initial MSA with a TCS score of 89.9 and two low scoring outliers (see Supplementary Table 3). The MSA was first scrutinized manually for sequences with long non-homologous subsequences, details are in Supplementary Table 3. Upon scrutiny, sequences were realigned resulting in 1350 columns and a trimmed alignment of 448 columns, constituting an increase in information of 4%. TCS score following the first scrutiny was 90.2. Next, sequences with large gaps were removed using in house scripts and the trimmed alignment. In short, only those sequences that do not show a gap of eight or more consecutive positions in the trimmed alignment are selected. A total of 33 sequences were removed by which upon realignment the final MSA contained 1196 columns of which 455 were selected as reliable. TCS score was 90.6. A Maximum likelihood phylogeny was constructed using PHYML 3.0 (Guindon et al., 2010). For statistical support we used FastTree with 1000 boot straps using the resources available at http:/ /booster.c3bi.pasteur.fr./new/ (Lemoine et al., 2018). Graphical representations were made using iTOL (Letunic and Bork, 2007).

HMMERCTTER clustering was performed such that the largest clusters that show 100% Precision and Recall (100% P&R) are accepted. As such each cluster identifies its member sequences with a score higher than any other sequence in the dataset. For classification we performed a single hmmsearch at the HMMER website (https://www.ebi.ac.uk/Tools/hmmer/search/hmmsearch) against all complete

proteomes from taxon Viridiplantae, using the profiles generated by HMMERCTTER clustering and the corresponding 100% P&R thresholds. Less than ten classification conflicts were detected of which most correspond to sequence D7KJ23_ARALL that appears to result from a duplicated coding sequence. Conflicting sequences were removed. Profiles and obtained datasets are in Supplementary Data 1 and 2, respectively.

SDPs were identified by cross analysis of SDPfox (Mazin et al., 2010) and mutual information obtained by Mistic (Simonetti et al., 2013). Mutual information of SDPs basically shows how likely SDPs have co-evolved and typically shows networks of connected SDPs. Mistic was applied without corrections since including corrections excluded several positions with a strict cluster specific residue. The SDP network analysis was performed using an in house script. The script reads the SDPfox and MISTIC output files. Putative SDPs are determined by the standard SDPfox cut-off. For all putative SDPs MI data are extracted from the MISTIC datafile with MI z-scores. Next, for each SDP it determines the maximal MI z-score with any of the other SDPs. The lowest of these maximal SDP MI z-scores is set as MI threshold (MIT). Then it calculates the nodes' connectivity score for each of the SDPs. Let N be the total amount of nodes and n the amount of connections a node has with z-score above the MIT, then the Mutual Information Connectivity Score for a node is defined as:

$$MICSi = Log ((n * (\sum(1.5*MI-MIT))/(N-1))$$

The network score is the average of all node scores.

Then, using the MISTIC z-score dataset, the script constructs 1000 random, fully connected subnetworks with the same amount of N nodes with at least one connection with a z-score at or above the MIT. The output is besides the network in csv format, a graph that shows the SDN score of the network compared to the score distribution of the 1000 random networks and the significance. Networks are visualized using Cytoscape.

Sequence logos were performed with enlarged cluster specific datasets following HMMERCTTER classification as described above. Shown are simple Shannon based logos with no background frequency corrections.

Structure models were obtained by I-Tasser (Yang et al., 2014). Structural analysis and screenshot rendering was performed using VMD (Humphrey et al., 1996). Structural alignments were performed using the STAMP (Russell and Barton, 1992) plug in of VMD. Composite images were made in GIMP.

RESULTS

The acyl-CoA synthase family can be hierarchically clustered and shows many signs of functional diversification

4CL and 4CL-like proteins belong to the acyl-CoA synthase (4CL/ ACS) family, part of the ANL superfamily of adenylating enzymes. We performed a PHI-BLAST search using 4CL1 from *Arabidopsis thaliana* (At1g51680) and *Oryza sativa* (Os08g14760) as queries and the PROSITE PS00455 pattern for the putative AMP-binding domain (box I) signature. We considered the pattern for box II as too restrictive since it is conserved in canonical 4CL enzymes only. Hits with an Expect value of ≤ 1E-5 were selected to construct a multiple sequence alignment (MSA) and corresponding HMMER profile that was used to seed an iterative sequence mining of the complete proteomes of 17 selected species listed in Supplementary Table 2. Two putative *Solanum tuberosum* 4CL proteins (St4CL) from the NCBI database (see Table 2) as well as three sequences retrieved from the PDB structure database and the sequences

deposited in the KEGG database were manually included. After performing a sequence scrutiny for likely non-functional homologs, a total of 220 protein sequences were obtained. Interestingly, no hits were obtained from *Chlamydomonas reinhardtii*. An MSA was constructed (Supplementary Data 3). Using a trimmed MSA (Supplementary Data 4) a maximum likelihood tree was reconstructed. The hierarchical clustering demonstrated by the tree should shed light on how the 4CL/ ACS family has evolved. Additional clustering techniques and functional annotation analysis were applied in order to predict the function of clustered sequences. The results are shown in Figure 2 and described below. A consensus tree obtained after 1000 bootstraps is shown in Supplementary Figure 3. Tree files are in Supplementary Data 5 and 6.

Cluster assignation was performed in two steps. First, the sequence set and corresponding tree was subjected to HMMERCTTER clustering (Pagnuco et al., 2018). HMMERCTTER identifies monophyletic clades that, using HMMER screening, identify member sequences with scores higher than non-member sequences. Hence, the HMMER profiles of these clade-instigated clusters show 100% precision and recall self detection (100% P&R-SD), i.e. when using the particular dataset, which is considered a training dataset. This basically means that the sequences of a single cluster are conserved among themselves and differ significantly from sequences from other clusters, which is an important characteristic when identifying functional or taxonomic subfamilies. Another important aspect of HMMERCTTER is that, by default, it searches the partition with lowest number of clusters. In this case, the resulting partition contained ten clusters 100% P&R-SD, leaving four orphan sequences from lycophyte *Selaginella moellendorffii*. Interestingly all ten *S. moellendorffii* sequences, including three sequences that cluster in cluster 10, appear as outliers (Figure 2 and Supplementary Table 4). This suggests that these sequences have undergone different functional constraints than the higher plant homologs. Thus, we further mostly ignored these sequences in our analyses.

Although HMMERCTTER clustering is objective and tends to identify clusters that are functionally different, hierarchical effects in complex superfamilies require more analysis. Thus, the second step for cluster assignation was carried out using the sequence of the box II motif as a major guide, combined with existing sequence annotations and taxonomic signals. This led to the hierarchic clustering shown in Figure 2 in which we assigned three major clades, combined with the ten HMMERCTTER identified clusters.

Clade 1, which consists of three subclades, contains well described 4CL sequences. Clade 1.1 (cluster 1) and 1.2 (cluster 9) contain class I enzymes from dicotyledonous and monocotyledonous plants, respectively; clade 1.3 (cluster 7) contains class II enzymes. The topology, clades 1.1 and 1.2 are monophyletic, suggests that the two classes of *4CL* genes co-existed prior to the speciation event that corresponds with the mono- and dicotyledon taxons. Clade 2 (cluster 4) has no assigned subclade and contains the Arabidopsis representative At1g62940 (ACOS5) that encodes a medium- to long-chain fatty acyl-CoA synthetase. Clade 3 consists of subclades 3.1 to 3.5 (clusters 2, 3, 5, 6 and 8, respectively) and is composed of a number of sequences annotated as 4CL-like but also contains the At4g19010 sequence that encodes a peroxisomal 4CL and At1g20510 sequence which is an OPC-8:0-CoA ligase (OPCL1) (Figure 2 and Supplementary Table 4).

The majority of clusters contain at least one sequence representative of each of the angiosperm plant species analyzed, demonstrating that these proteins are evolutionary conserved and that a common ancestor of these clusters was present before the divergence of monocots and dicots (see Supplementary Table 4 for details). The fact

that most species still have a candidate in most clusters suggests that their functions, albeit unknown, appear as crucial or at least important. The only exceptions are subclades 1.1 and 1.2 which together form a functional group with a bifurcation that follows taxonomy, and subclade 3.5. The latter seems specific to dicotyledons but lacks an Arabidopsis homolog. The presence of a tomato homolog in subclade 3.5 led to the identification of St4CL-H5 (XP_006361720, NCBI database), not present in KEGG and not identified in the PGSC version 4.03 database. Analysis of the most recent version (V6.1) did identify both this homolog as well as the StACOS homolog. Hence, potato does have a candidate in each of the eight clusters containing dicotyledonous sequences. Table 2 shows nomenclature we suggest for potato 4CL/ ACS genes based on all the evidence presented in this work.

A global sequence conservation analysis of the three major clades was made first. Figure 2 contains sequence logos for boxes I and II that describe the larger ANL superfamily and the 4CL subfamily, respectively. Although most sequences do not violate the Prosite description for ANL enzymes (PS00455: [LIVMFY]-{E}-{VES}-[STG]-[STAG]-G-[ST]-[STEI]-[SG]-x-[PASLIVM]-[KR]), the box I logos show notable differences. Clades 1 and 2 have mostly proline (P187) at position 2 of the pattern whereas clade 3 sequences have leucine (L) or methionine (M). Clades 2 and 3 have a preferred serine (S), or a glutamine (Q) or a valine (V) respectively, at the one but last position whereas clade 1 has a strictly conserved proline (P196). The most notable difference in box II, supposedly specific for canonical 4CL enzymes, is the substitution of cysteine (C), in clades 1 and 2, to mostly tryptophan (W) in clade 3. Cysteine has been suggested to be directly involved in catalysis (Becker-André et al., 1991). Although it seems differences between clade 1 and clade 2 on the one hand and clade 1 and clade 3 on the other hand are quantitatively similar, the presence of the conserved cysteine in clade 2 suggests this is a more likely candidate for non-canonical 4CL enzymes than clade 3. This is supported by the fact that clade 3 is further divided into five conserved subclades with broad taxonomic representation, suggesting clade 3 contains five, rather than one, different activities.

We then set out to identify SDPs. The SDPs that best explain the three clades were identified using SDPfox. These SDPs were identified in the three-cluster comparison as well as all three bi-cluster comparisons (i.e. clade 1 vs. clade 2; clade 1 vs. clade 3; and clade 2 vs. clade 3). Mutual information network analysis shows this network is highly significant (Supplementary Figure 4), which basically means that these positions have been coevolving during the evolution that is described by the tree. For this and all other SDP analyses we enlarged our dataset by classifying all sequences present in a complete proteome from Viridiplantae presented at the HMMER website using the strict HMMERCTTER clustering derived cut-off threshold. Sequence logos of the SDPs showing the diversification patterns are included in Figure 2 and are projected on the structure of 4CL2 from tobacco (Nt4CL2; PDB identifier 5BST) shown in Figure 3. Interestingly SDP340 (clade1: proline; clade 2: cysteine; and clade 3: glycine/ alanine) corresponds to a residue that physically interacts with the hydroxycinnamate substrate according to various structures of Nt4CL2 (PDB identifiers 5BSR, 5BST, 5BSU, 5BSV). Four other SDPs interact physically with SDP340, directly or indirectly and as such appear to support or compensate these likely major substitutions (See Figure 3 for details). The five other SDPs are located more distantly. We made no further attempts to explain the demonstrated substitution patterns in order to avoid overly speculation. The substitution pattern identified by this major SDP analysis corroborates the three clade hypothesis and suggests both clades 2 and 3 to have different substrates than clade 1.

Clade 1/ clade 2 functional divergence can be explained by changes in and near the hydroxycinnamate binding cleft

The high conservation of the canonical 4CL subfamily of clade 1 and the 4CL-like subfamiliy of clade 2 is intriguing and suggests both subfamilies carry out important functions, which is also supported by the complete taxonomic contribution to both clades. The clear separation of clades 1 and 2 suggests these functions have fully diversified. We performed structural analyses in order to shed light on the probable function of the clade 2 subfamily.

Comparison of the canonical 4CL clade 1 with clade 2 yielded no less than 61 SDPs (Supplementary Figure 5A). This rather high amount of SDPs can be explained by the fact that, despite that we included all plant sequences available, the datasets are still small by which certain identified positions might result from a low data artifact. We identified a score drop at a connectivity score of 2.08 (Supplementary Figure 5B) by which we could exclude 13 SDPs, which are either low data artifacts or likely contribute less to the diversification. The more stringent set of likely important SDPs shows a high level of cluster specific conservation in both clades 1 and 2, as demonstrated in the sequence logos of the major 48 SDPs in Figure 4, which further explains the high number of SDPs. The network includes the aforementioned P196 (box I) as well as G392 (box II). Supplementary Figure 5C shows the network has a significant score, 5D shows the final network consisting of 48 SDPs. Five SDPs, including SDP340 also identified in the global SDP analysis, interact directly with feruloyl-AMP which suggests clade 2 has a different substrate than clade 1. Figure 4 shows these five major SDPs. Whereas SDP340 (P-C) is in contact with the hydroxycinnamoyl moiety, both SDP308 (G-A) and SDP332 (G-A) are in contact with both the hydroxycinnamoyl and the AMP moiety (Figure 4B and 4C). SDP360 (C-V) is in contact with the AMP moiety, as is Q331 whereas its clade 2 counterpart E344 appears more distant in the model obtained and not in contact with the feruloyl-AMP, substrate of the clade 1 enzyme (Figure 4D and 4E). We made no attempts to explain how the other SDPs are related to the diversification but, as was shown for the major SDPs in Figure 3, several but not all of these other SDPs physically interact with the five SDPs that take part in the binding pocket. Furthermore, SDP233 is part of the CoA tunnel (not shown).

A three amino acid deletion might result in a larger binding pocket of 4CL-like enzymes from clade 2.

Hydroxycinnamates share a common cinnamic acid backbone with one (4-coumaric acid) to two hydroxy and up to two methoxy substituents (sinapic acid) (Figure 1). The ability to accept the most voluminous sinapate as substrate of 4CL1 from *Glycine max* and 4CL4 from *Arabidopsis thaliana* is explained by a single amino acid deletion that corresponds to either V341 or L342 of Nt4CL2 in the substrate binding pocket that results in an increase of the volume's cavity (Lindermayr et al., 2003; Hamberger and Hahlbrock, 2004). A structural alignment of the structure of Nt4CL2 enzyme (PDB identifier 5BSR) and a 3D model made for the St4CL-like sequence from clade 2, shows a three amino acid gap, 219 to 221 in Nt4CL2, that appears to envelop part of the binding cleft, being in physical or near physical contact with residues 243, 306 and 344 (Supplementary Figure 6). Note that these do not correspond with the sinapate region of 339 to 342 which has been suggested to play a role in sinapate substrate acceptance. Nevertheless, the absence of this region might also allow for a larger substrate pocket for the 4CL-like subfamily, in accordance with the Arabidopsis representative At1g62940 (ACOS5) which has medium- to long-chain fatty

acids as substrates. This can to the best of our understanding, only be corroborated by an actual structure rather than computer modeling.

Clade 3 has five largely independently evolved subclades of which one specific for dicotyledonous species showing a significantly different binding cleft

SDP identification appears as rather straightforward but its complexity grows with the complexity of the underlying clustering. The tree has three major clades. We can safely assume that clade 1 concerns 4CL activity. We have furthermore assumed that clade 2 4CL-like enzymes have a different substrate than canonical 4CLs, based on the clade topology and taxon distribution, according to SDPs and structural analyses and supported by existing functional annotation. Clade 3, however, has four subclades with an almost complete taxon contribution and a fifth subclade that appears to consist of dicotyledons only. The major uncertainty now is which comparisons should be made in order to obtain the most valuable insights on how these subfamilies have evolved. Here, we make the assumption that all five subfamilies have a different function, based on tree topology, and that each function has evolved independently, which is a simplification since the five subclades share part of their evolutionary history. The other input is eventual functional annotation. The global comparative box II analysis (see Figure 2) showed that most but not all clade 3 sequences have the cysteine generally considered as catalytic substituted by a tryptophan, which would suggest a different molecular function for clade 3. Most sequences in clade 3, irrespective of the subclade they belong to, have an SKL, or similar, C-terminal subsequence, which is believed to be a universal peroxisomal targeting signal (Reumann et al., 2004). Hence, part of the diversification simply concerns the cellular component of function.

We identified SDPs for each clade 3 subclade by comparing them independently to clade 1. Supplementary Figure 7 shows a Venn diagram of the SDP cross analysis. Supplementary Data 7 contains a detailed analysis of the identified SDPs. Six of the 31 to 71 SDPs identified independently are shared and these appear as six of the eight most important SDPs that we identified in the global SDP analysis comparing clades 1, 2 and 3. The logos of these eight most reliable SDPs show rather large differences between clades 1 and 3, whereas conservation inside the subclades is high (Supplementary Figure 8). The most obvious changes concern P187L (position 2 of box I and mentioned above) and P340G (one of the global SDPs) on the one hand and [SA]240G and [DEN]428G. The first three are buried residues whereas the last is solvent exposed. Given the nature of the substitutions it seems these SDPs have an impact on the protein structure and might be required given the more basic environment of clade 3 enzymes.

The direct subclade to subclade comparison shows that there are no clades that share significantly more SDPs than other pairs, which corresponds with the suggestion that evolution in each of the subfamily has been largely independent. As such we next determined which SDPs were specific for which comparison and made sequence logos. In all five cases the logos show these are subfamily specific changes (Supplementary Figures 9 to 13). Structure function analysis is however hampered by the fact that we cannot identify a common ancestor when we compare clade 1 with any subclade from clade 3. Since approximately half of the SDPs are shared, the majority supposedly via the common ancestor of clade 3 sequences, we cannot pinpoint which SDPs should be evaluated in the analysis. A direct comparison (e.g. 3.3 vs. 3.1) is theoretically preferred but is hampered by the lack of functional annotation.

A last analysis that we performed is a screen for substrate vicinity. We visually inspected which residues are near either the CoA in 5BSR or the phenol moiety of the feruloyl susbtrate in 5BSV. Interestingly, ten out of 13 of these SDPs found near or in

the substrate cleft were identified in the clade 1 vs. subclade 3.5 comparison, five of which being specific (i.e. they were not identified in another comparison). The other comparisons identified four, one, five and two SDPs in the binding cleft for the comparisons of clade 1 versus 3.1, 3.2, 3.3 and 3.4, respectively (See Supplementary Data 7). Only single cleft SDPs were specific for subclades 3.1 and 3.2 whereas all the cleft SDPs found for subclades 3.3 and 3.4 are shared with at least one other subclade. This led us to restrict structure-function analysis to subclade 3.5.

Figure 5A shows the Mutual Information Connectivity network we obtained for the clade 1 vs. subclade 3.5 analysis, showing a central core of highly interconnected SDPs with a more peripheral subnetwork of more loosely connected SDPs. Unfortunately, the ten binding cleft positions do not appear to be at the heart of the network. Positions 443, 243, 506, 280, 284 and 442 form a subnetwork in a peripheral part of the MI network of SDPs, whereas position 445 seems isolated from the other binding cleft SDPs. All have low connectivity scores, as indicated by the greenish colors, suggesting low importance. Positions 233, 335 and 340 form a small subnetwork with intermediate score, part of the center of the MI network of SDPs. These SDPs are however shared with at least one other subclade. Unfortunately, since clade 1 has many more sequences than subclade 3.5, it seems these SDPs obtain high MI scores given their conservation pattern in clade 1 rather than that of subclade 3.5. More importantly, the four low scoring binding cleft SDPs might show poor mutual information due to poor conservation in clade 1. Unfortunately, subclade 3.5 is too small (full extended sequence set of 44) to serve in an MI determination. Sequence logos of the general (1 vs. 3) and the specific (1 vs. 3.5) comparisons (Figure 5D) do however clearly show the substitution pattern. Note that the lower bitscore of the subclade 3.5 logo results from the small number of sequences.

Next we obtained 3D models of three different, multiple mutants. In the first mutant we exchanged the eight major clade 1 vs. clade 3 SDPs in the sequence of 5BSM. A second mutant contains the ten binding cleft substitutions identified and a third contains all seventeen mutations, note that SDP340 is part of both SDP sets. The eight major clade 1 vs. clade 3 SDP mutant serves as a control and did not show significant displacement with respect to the original structures (5BSR and 5BSV, see Figure 5B). Also the residues of the ten 3.5 specific subclade appear located at similar positions, in both the second and the third mutant. Figure 5C shows a structural alignment of the mutants with 5BSR. Now, since these ten residues in subclade 3.5 sequences are different from those in the canonical 4CL sequences of clade 1, and the residues are conserved in both the subclade 3.5 and clade 1, this suggests the subclade 3.5 proteins do have a different substrate. Note that subclade 3.5 concerns sequences from dicotyledonous species only and, interestingly lacks an Arabidopsis homolog.

Class I/ class II functional divergence cannot be explained by changes in the hydroxycinnamate binding cleft

As mentioned above, clade 1 is composed of three subclades. Clades 1.1 and 1.2 which correspond to class I 4CL enzymes and clade 1.3 which contains class II 4CL enzymes. SDPfox analysis comparing class I and class II sequences identified a total number of 18 SDPs. Mutual information connectivity analysis showed the corresponding Specificity Determining Networks (SDN) was significant having a z-score of 21 (Supplementary Figure 14). None of the SDPs is found in the vicinity of the phenol moiety of the substrate (Supplementary Figure 14E and F), suggesting the substrate specificity is not determined by this binding pocket.

Two SDPs, however, are found in the vicinity of the CoA that was crystallized with the 4CL in the 5BSV structure. I288 is substituted by L311 in the model made for a potato class II 4CL sequence. Although Leu is 100% conserved in class II enzymes, it is also found in class I enzymes and for instance the 4CL-like subfamily of clade 2. Then, Y442, which is part of the CoA tunnel, is substituted by F469 in the class II model. Not only are these two residues very conserved in the 4CL class I and class II, intriguingly the 4CL-like subfamily of clade 2 also shows a highly conserved Tyr as counterpart of Y442 (not shown). As such the Tyr to Phe substitution is likely important for the diversification towards the class II subfamily, and might depend on the L311 substitution and other highly conserved SDPs we identified. Its vicinity to the CoA suggests class II enzymes have different affinities for CoA.

Class I enzymes from dicotyledon but not from monocotyledon show signs of both taxonomic and functional diversification

We also identified SDPs for the monocotyledon/ dicotyledon class I (Supplementary Figure 15). Out of 19 SDPs, SDP439 is part of the hinge loop suggesting an involvement in conformer dynamics. Q503E is the central SDP in the network and physically interacts with K118R. This suggests the SDN is involved in the large sub-domain movement between the adenylate and thioester conformations. Since the enzymes in mono- and dicotyledonous plants catalyze the same reactions, this may not be a functional diversification. The obtained network however is highly significant (z-score is 10.4).

Since clade 1.1 contains various paralogous sequences, it was subclustered using HMMERCTTER, resulting in five subclades (1.1-1 to 1.1-5) that show many signs of taxonomic specific evolution, which might or might not be related to functional diversification (Supplementary Figure 16). On the one hand, subclade 1.1-1 contains sequences from Solanales whereas subclade 1.1-2 contains sequences from Fabids only. On the other hand, subclade 1.1-3 is composed of a mix of sequences from some Solanales and Fabids (Pentapetalae). Lastly, sequences from *Arabidopsis thaliana* are found only in subclade 1.1-4. St4CL-IA, St4CL-IB and St4CL-ID group together in subclade 1.1-1 whereas St4CL-IC falls into subclade 1.1-3. Proteins that correspond to subclade 1.1-3, among which a *Capsicum annuum* sequence identified by HMMERCTTER, are likely candidates for functional diversification since they have A340 rather than the otherwise highly conserved Pro.

Tissue and stress-induced expression profile of St4CL genes in potato

The *in vivo* biological functions of the clade 3 enzymes are for the most part still unknown, although such functions are expected to be highly similar within the conserved subclades. Expression patterns suggest functions in developmental and/ or stress-related biochemical pathways not related to phenolic compounds metabolism (Raes et al., 2003; Ehlting et al., 2005). RNAseq expression data (Massa et al., 2011) show that the corresponding potato genes show little or no expression in tubers, the organ of our major interest (Supplementary Figure 17). It also shows that clade 3 homologs St4CL-H1, St4CL-H2, St4CL-H3 and St4CL-H4 are however expressed in most other organs, in correspondence with a role in the production of ubiquinone, part of the mitochondrial respiratory chain. On the other hand, they are not induced by wounding and/ or *Phytophthora infestans* infection (data not shown), which might be expected for OPCL enzymes involved in jasmonic acid-mediated defense response. No RNAseq data for St4CL-H5 have been found.

As mentioned in the Introduction, class II 4CL members are involved in flavonoid biosynthesis whereas class I enzymes are involved in the biosynthesis of monolignols, the building blocks of both lignin and suberin, the latter being deposited in potato tuber skin. Hence, flesh and skin of potato tubers with different colors were analyzed separately (see Table 1). Five of the 11 4CL family sequences in potato fall in clade 1 with the canonical 4CLs: there are four class I isoforms, referred to as St4CL-I (A-D), and one class II isoform, St4CL-II (Table 2). Note that *St4cl1* and *St4cl2* were originally believed to be two genes (Becker-André et al., 1991) but more likely correspond to different alleles and are, hence, here redesignated as St4CL-IA. Supplementary Figure 18 shows the expression levels of clade 1 *St4CLs* in different organs and tissues of *Solanum tuberosum* Group Phureja and *Solanum tuberosum* Group Tuberosum from RNAseq public available data. As can be seen, the majority of *St4CLs* are expressed in most organs, with *St4CL-IC and St4CL-ID* showing the lowest absolute expression values. In order to determine the role of St4CLs in potato tuber, transcript levels of *St4CL* genes together with marker genes of the flavonoid (chalcone synthase, *CHS*) and suberin (feruloyl-CoA transferase, *FHT*) pathways were determined by qRT-PCR in flesh and skin of tubers from Andean varieties collected from two independent campaigns. *St4CL-IC and St4CL-ID* transcripts were not detected. *St4CL-IB* did not show a clear consistent pattern (Figure 6). *St4CL-IA* which abundance values were one order of magnitude lower was only detected in the skin of Chaqueña and Santa María varieties during the 2011 campaign (data not shown). As expected, *StFHT* expression was almost completely restricted to tuber skin and showed highly variable levels among both varieties and year of cultivation. The lowest levels of *StCHS* were found in non-pigmented fleshes whereas it showed high levels in colored tissues. As expected, *St4CL-II* expression pattern was similar to *StCHS*. Santa María variety, which is intensely pigmented in flesh and skin, showed the highest expression levels of both *StCHS* and *St4CL-II* whereas Chaqueña and Waicha varieties showed intermediate levels. The levels of *StCHS* but not of *St4CL-II* were relatively low in Moradita skin considering its high anthocyanin content (Figure 6).

With the aim to obtain further insight on the role of *St4CL* isoforms, tubers from the red flesh/ skin Santa María variety were subjected to wounding. Skin samples (native periderm) were included as positive control of *StFHT* expression. *St4CL-IA* and *St4CL-IB* expression is induced by wounding as did *StFHT* whereas *St4CL-II* is repressed as *StCHS* (Figure 7). *St4CL-IC* and *St4CL-ID* transcripts, absent in flesh and skin, were not wound-induced (data not shown).

DISCUSSION

4-coumarate-CoA ligase (4CL) and 4CL-like proteins belong to the the **A**cyl-**C**oA **S**ynthetase (ACS or 4CL/ ACS) family, part of the ANL superfamily of adenylating enzymes. Most plants species have a moderate amount of paralogs, e.g. we identified 13 and 11 sequences in Arabidopsis and potato, respectively. Given that they are involved in at least five processes, i.e., flavonoid biosynthesis, lignin/ suberin deposition, ubiquinone production, sporopollenin biosynthesis and the jasmonic acid pathway, we performed a systematic computational, comparative sequence analysis in order to shed light on this family's complexity. We then studied the expression of the potato members using both dedicated experiments and public RNAseq data. Here we discuss all data and revise the literature in order to functionally annotate the potato 4CL/ ACS protein family.

We performed an elaborate sequence mining, avoiding both false positives and false negatives. Specificity was achieved by applying a HMMER seed made of sequences obtained by PHI BLASTs with both a monocotyledonous and dicotyledonous query, applying Prosite's box I motif for the ANL superfamily. The ACS box II GEICIRG motif would have been arguably too strict given the fact that peroxisomal 4CLs have a tryptophan (W) rather than a cysteine (C). Sensitivity was achieved by iterative HMMER searches while maintaining specificity by applying HMMERCTTER's 100% P&R rule. Upon the iterative HMMER search of 17 complete proteomes we manually removed sequences using a set of predefined rules. Although the rules on itself may be considered subjective, the MSA quality improved significantly, both as reported by TCS analysis and by the amount of informative columns. We acknowledge that this might have resulted in removing functional rather than non-functional homologs. This is however preferred above obtaining a poor MSA, caused by for instance the presence of a pseudogene sequence or a sequence derived from an incorrect gene model.

Three major clades (see Figure 2) were defined based on tree topology and differences in either the box I or the box II motif: clade 1 (with 3 subclades) composed of canonical 4CL enzymes from the cytosol, clade 2 with many sequences that are annotated as 4CL-like protein but also AtACOS5, and clade 3 (with 5 subclades) also containing sequences annotated as 4CL-like but also sequences that encode a peroxisomal 4CL or a OPC8:0-CoA ligase. Previous phylogenetic analysis showed that angiosperm 4CL/ ACS genes from Arabidopsis, Populus and rice form a major clade corresponding to 4CL enzymes and five well-supported additional clades (De Azevedo Souza et al., 2008). Our phylogenetic analysis showed an additional fifth subclade in clade 3 that is dicotyledon-specific, albeit that it has no Arabidopsis homolog. Sequences logos for both box I and box II, as well as the global SDPs, which were obtained using an extended dataset, corroborate the topology of the highest level of clustering, into major clades 1, 2 and 3 (see Figure 2).

Various SDP analyses were carried out in order to identify the major substitutions that underlie known and unknown diversifications such as substrate specificity/ permissiveness between clades. Comparison between clade 1 and clade 2 yielded 48 SDPs. Five SDPs (G308, Q331, G332, P340 and C360) interact directly with feruloyl-AMP, suggesting that clade 2 has a different substrate than clade 1. This is substantiated by both the large amounts of SDPs between clades 1 and 2 and the SPDs conservation levels (see Figure 4). Also, a three residue gap in the structural alignment of the Nt4CL2 enzyme structure and the model obtained for the clade 2 potato sequence was observed. Although the absence of this region does not correspond with the single amino acid deletion found in sinapate-activating 4CL enzymes, it will most likely allow

for a larger substrate pocket, and correspondingly, a more voluminous substrate. In accordance, the Arabidopsis representative of clade 2 At1g62940 (ACOS5) showed moderate activities with medium- and long-chain fatty acid substrates in assays of purified recombinant proteins. It was shown that the products of ACOS5 are key intermediates in the biosynthesis of sporopollenin, a major component of the outer wall (exin) of pollen grain (de Azevedo Souza et al., 2009). It is noteworthy that the ortholog of ACOS5 in rice (Os04g24530) is also essential for sporopollenin synthesis (Li et al., 2016). This combined with the fact that clade 2 has a complete taxonomic contribution and that most species have only a single homolog in this clade, suggests that clade 2 enzymes are, rather than 4CL-like enzymes, acyl-CoA synthetases as AtACOS5 involved in pollen wall formation in monocot and dicot species.

The molecular mechanism of the class I/ class II divergence on the other hand is unclear. In the present work, no SDPs were found in the vicinity of the phenol moiety of the substrate, suggesting the substrate specificity is not determined by this binding pocket. Accordingly, biochemical analyses towards the major hydroxycinnamates have not shown kinetic differences that are biologically significant (i.e. differences are far below an order of magnitude) beyond the rare sinapate activity reported for some 4CL enzymes (Lindermayr et al., 2003; Hamberger and Hahlbrock, 2004). Interestingly, two SPDs were found in the CoA tunnel, suggesting that class I/ class II enzymes have different affinities for CoA. For tobacco class I 4CL2 it was found that increasing the *in vitro* concentration of CoA towards stoichiometric concentrations, results in decomposition of the hydroxycinnamoyl-AMP intermediate towards the hydroxycinnamoyl-CoA. Moreover, the switch for the thioester-forming conformation is induced only upon the subsequent binding of CoA (Li and Nair, 2015). Thus, the concentration and/ or availability of both the hydroxycinnamic acid and CoA in a particular cell-tissue may explain different reaction products. Also, channeling substrates through metabolons made of different 4CL homologs with different homologs of up or downstream enzymes may contribute to the formation of different end products.

Differential constitutive and stress induced *4CL* gene expression in a number of plant species, suggest that the cytosolic 4CLs have undergone subfunctionalization for the biosynthesis of different classes of phenolic compounds, with the phylogenetically distinct class I and class II clades specialized for monolignols and flavonoid biosynthesis, respectively (Ehlting et al., 1999; Harding et al., 2002). Here, we demonstrate that, in line with species as Arabidopsis and rice, the potato 4CL family consists of five members with four class I (St4CL-I (A-D)) and one class II (St4CL-II) isoforms, rather the two nearly sequence-identical class I isoenzymes that were shown earlier.

The presence of four class I 4CL isoforms involved in monolignol biosynthesis may be explained by the different roles that lignin and suberin can play. Our data (Figure 7) show *St4CL-IA* and *St4CL-IB* expression is induced by wounding. Although *St4CL-IC* expression is low, it is by far the highest in skin of mature tubers (Supplementary Figure 18). Therefore, St4CL-IA and St4CL-IB would be involved in wound-induced suberin production whereas St4CL-IC could be related to constitutive deposition of suberin in potato tuber. St4CL-IA and St4CL-IB are, given their expression in stem (Supplementary Figure 18) also the more likely candidates for suberin production to the Casparian strip or lignin biosynthesis for the xylem vessels. Whether different 4CL class I isoforms are involved in either suberin or lignin pathway is hard to delineate. Differences in the composition of suberin between native and wound potato periderm have also been reported. In native potato periderm, the

polyaromatic domain is mostly composed of guaiacyl units (Mattinen et al., 2009), while in wound potato periderm comparatively high proportions of syringyl units were found (Lapierre et al., 1996; Yan and Stark, 2000). We identified St4CL-IC as the class I isoform from potato that is most likely diversifying. Although speculative, this might be related to the relation between guaiacyl-lignin and feruloyl-CoA on the one hand, and syringin-lignin and sinapyl-CoA on the other hand. In depth compositional analyzes of both suberin and related metabolites combined with gene expression studies of native and wound tuber periderm from wild type and *St4CL* mutant potato plants at different developmental stages will shed light to that issue. Overall and although there are still gaps in our knowledge, it should be clear that the presence of four class I isoforms gives the plant high plasticity.

Clade 3 can be seen as a clade with non-canonical 4CL/ ACS enzymes, based on both our results and those previously published. Subclades 3.1 to 3.4 show a complete taxon contribution whereas subclade 3.5 consists of dicotyledons only. Six SDPs were found for the global comparison between clade 1 vs. clade 3. These include P187L (position 2 of box I), P340G (one of the global SDPs in the substrate binding cleft), (SA)240G and (DEN)428G. The solvent exposed nature of the latter residue might be explained by a more basic peroxisomal environment of clade 3 enzymes. The direct subclade to subclade comparison shows that there are no subclade pairs that share significantly more SDPs than other pairs, suggesting that evolution in each of these five subfamilies has been largely independent, as is also suggested by the starlike tree topology of clade 3. The SDP analysis of residues in the substrate vicinity of either CoA or the feruloyl substrate yield 13 SDPs. Ten of these 13 SDPs were identified in the clade 1 vs. subclade 3.5 comparison, five of which being specific. Structural analysis of virtual mutants containing these ten site-SDPs shows no significant conformational changes that might have supported a similar binding cleft. Hence, the model suggests the binding site to be fysico-chemically different, suggesting that these dicotyledonous enzymes have a different substrate and function, which would explain its conservation in most dicotyledonous plants. Unfortunately, no transcript data that might have shed some light on the physiological function were available.

On the other hand, enzymes from subclades 3.1-3.4 did not show a rather different binding cleft than canonical 4CL enzymes, suggesting they might use a hydroxycinnamate as substrate. The substitution of the Cys from the box II motif towards a Trp would suggest these are not 4CL enzymes. However, recent reports on At4g19010 from subclade 3.4 (Block et al., 2014) and At5g38120 from subclade 3.3 (Soubeyrand et al., 2019) have unequivocally shown these to act as 4CL enzyme, albeit in a third pathway that, via the peroxisome leads to the synthesis of ubiquinone. Hence, we name these 4CL enzymes as class III 4CL enzymes. Unfortunately, subclade 3.3 has five homologs from Arabidopsis of which four derive from the same locus: At1g20480, At1g20490, At1g20500, and At1g20510. No clear activity has been demonstrated for At1g20480; no correct spliceoforms from At1g20490 were found; At1g20500 has both considerable 4CL and OPCL activity, whereas At1g20510 was demonstrated to have both OPCL and acyl-CoA synthetase activity on fatty acids (Kienow et al., 2008). Also At4g05160 from clade 3.2 was shown to combine OPCL and ACS activity. Hence, although tree topology suggests different functions, SDP analysis does not show distinguished binding clefts and *in vitro* activity assays show a complex pattern of substrate specificity. Since *in vitro* conditions do typically not reflect those *in planta*, we conclude that clade 3 is a paradigm for functional redundancy and diversification. It allows a protein family to evolve enzymes with on the one hand highly specific functions, while maintaining high substrate permissiveness and therewith plasticity. The

fact that no clear differences in the binding clefts of class I, class II and class III enzymes are found, corresponds with the high permissiveness. The fact that still class I and class II enzymes are involved in different processes suggest their specificity is not defined in the binding of the substrate but rather by the binding of CoA, as suggested by the fact that the Y442F SDP found in the class I to class II analysis is part of the CoA tunnel. The conformational changes induced by CoA binding might depend on SDP340, a rigid Pro in the canonical 4CLs of clade 1 whereas in clade 3 mostly Gly or Ala, which are both small and with a high degree of liberty. Also the Cys to Trp substition in box II can be considered as likely structurally important, rather than considering Cys as being involved in catalysis. Table 2 resumes all data with the objective to obtain a comprehensive nomenclature and sequence function assignation.

The overall picture of the 4CL/ ACS protein family is that of a complexity that has resulted from, on the one hand, apparent substrate specificity that is not explained by the conserved binding cleft, combined with high substrate permissiveness that does follow the conserved binding cleft. On the other hand, the incorrect or incomplete annotation of many sequences impedes straightforward computational, and therewith wetlab analyses. A clear example is that of AtACOS5, which as Q9LQ12 is annotated by UniProtKB/ Swiss-Prot as “4-coumarate-CoA ligase-like 1”, notwithstanding the fact that it acknowledges the experimental evidence at protein level that clearly shows it concerns an ACS, rather than 4CL. This study also provides examples of canonical SDPs, which affect specificity via substrate binding, and non-canonical SDPs, which affect specificity in other ways. Detailed structure analyses combined with biochemical studies will be needed to proceed further in the study of the 4CL/ ACS and other protein families.

Author Contributions

MAV and AtH conceived, designed the research, conducted experiments and drafted the manuscript. ABA was responsible for conceptualization, supervision, project administration and funding acquisition. All authors contributed to writing the manuscript and approved the submitted version.

Funding

This work was financially supported by grants from Agencia Nacional de Promoción Científica y Tecnológica (PICT 2018 Nº 221) (PICT2017 Nº 1310); Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) (PIP 2015 Nº 0762) and Universidad Nacional de Mar del Plata (UNMdP) (EXA 814).

Acknowledgments

The authors acknowledge Nicolás Stocchi for providing the Python script for the determination and statistical analysis of the network connectivity scores.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

André, C. M., Schafleitner, R., Legay, S., Lefèvre, I., Aliaga, C. a A., Nomberto, G., et al. (2009). Gene expression changes related to the production of phenolic compounds in potato tubers grown under drought stress. *Phytochemistry* 70, 1107–16. doi:10.1016/j.phytochem.2009.07.008.

Babbitt, P. C., Kenyon, G. L., Martin, B. M., Charest, H., Slyvestre, M., Scholten, J. D., et al. (1992). Ancestry of the 4-chlorobenzoate dehalogenase: analysis of amino acid sequence identities among families of acyl:adenyl ligases, enoyl-CoA hydratases/isomerases, and acyl-CoA thioesterases. *Biochemistry* 31, 5594–5604. doi:10.1021/bi00139a024.

Banks, J. A., Nishiyama, T., Hasebe, M., Bowman, J. L., Gribskov, M., dePamphilis, C., et al. (2011). The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science (New York, N.Y.)* 332, 960–963. doi:10.1126/science.1203810.

Becker-André, M., Schulze-Lefert, P., and Hahlbrock, K. (1991). Structural comparison, modes of expression, and putative cis-acting elements of the two 4-coumarate: CoA ligase genes in potato. *Journal of Biological Chemistry* 266, 8551–8559. doi:10.1016/S0021-9258(18)93010-3.

Bernards, M. (2011). Demystifying suberin. *Canadian Journal of Botany* 80, 227–240. doi:10.1139/b02-017.

Bernards, M. A., and Razem, F. A. (2001). The poly(phenolic) domain of potato suberin: a non-lignin cell wall bio-polymer. *Phytochemistry* 57, 1115–1122. doi:10.1016/s0031-9422(01)00046-2.

Block, A., Widhalm, J., Abdelhak, F., Cahoon, R., Wamboldt, Y., Elowsky, C., et al. (2014). The Origin and Biosynthesis of the Benzenoid Moiety of Ubiquinone (Coenzyme Q) in Arabidopsis. *The Plant Cell* 26, 1–11. doi:10.1105/tpc.114.125807.

Boher, P., Serra, O., Soler, M., Molinas, M., and Figueras, M. (2013). The potato suberin feruloyl transferase FHT which accumulates in the phellogen is induced by wounding and regulated by abscisic and salicylic acids. *Journal of experimental botany* 64, 3225–36. doi:10.1093/jxb/ert163.

Costa, M. A., Bedgar, D. L., Moinuddin, S. G. A., Kim, K.-W., Cardenas, C. L., Cochrane, F. C., et al. (2005). Characterization in vitro and in vivo of the putative multigene 4-coumarate:CoA ligase network in Arabidopsis: syringyl lignin and sinapate/sinapyl alcohol derivative formation. *Phytochemistry* 66, 2072–2091. doi:https://doi.org/10.1016/j.phytochem.2005.06.022.

Costa, M., Collins, R., Anterola, A., Cochrane, F., Davin, L., and Lewis, N. (2003). An in silico assessment of gene function and organization of the phenylpropanoid pathway metabolic networks in Arabidopsis thaliana and limitations thereof. *Phytochemistry* 64, 1097–1112. doi:10.1016/S0031-9422(03)00517-X.

Criscuolo, A., and Gribaldo, S. (2010). Criscuolo A, Gribaldo S.. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evol Biol 10: 210. *BMC evolutionary biology* 10, 210. doi:10.1186/1471-2148-10-210.

Cukovic, D., Ehlting, J., VanZiffle, J. A., and Douglas, C. (2001). Structure and Evolution of 4-Coumarate:Coenzyme A Ligase (4CL) Gene Families. *Biological chemistry* 382, 645–654. doi:10.1515/BC.2001.076.

De Azevedo Souza, C., Barbazuk, B., Ralph, S. G., Bohlmann, J., Hamberger, B., and Douglas, C. J. (2008). Genome-wide analysis of a land plant-specific acyl:coenzymeA synthetase (ACS) gene family in Arabidopsis, poplar, rice and

941 Physcomitrella. *New Phytologist* 179, 987–1003.
942 doi:https://doi.org/10.1111/j.1469-8137.2008.02534.x.
943 de Azevedo Souza, C., Kim, S. S., Koch, S., Kienow, L., Schneider, K., McKim, S. M.,
944 et al. (2009). A novel fatty Acyl-CoA Synthetase is required for pollen
945 development and sporopollenin biosynthesis in Arabidopsis. *The Plant cell* 21,
946 507–525. doi:10.1105/tpc.108.062513.
947 De Jong, W. S., Eannetta, N. T., De Jong, D. M., and Bodis, M. (2004). Candidate gene
948 analysis of anthocyanin pigmentation loci in the Solanaceae. *Theoretical and*
949 *Applied Genetics* 108, 423–432. doi:10.1007/s00122-003-1455-1.
950 Domergue, F., Vishwanath, S. J., Joubès, J., Ono, J., Lee, J. A., Bourdon, M., et al.
951 (2010). Three Arabidopsis fatty acyl-coenzyme A reductases, FAR1, FAR4, and
952 FAR5, generate primary fatty alcohols associated with suberin deposition. *Plant*
953 *physiology* 153, 1539–1554. doi:10.1104/pp.110.158238.
954 Eck, H., Jacobs, J., Berg, P., Stiekema, W. J., and Jacobsen, E. (1994). The inheritance
955 of anthocyanin pigmentation in potato (Solanum tuberosum L.) and mapping of
956 tuber skin colour loci using RFLPs. *Heredity* 73. doi:10.1038/hdy.1994.189.
957 Eddy, S. (2009). A new generation of homology search tools based on probabilistic
958 inference. *Genome informatics. International Conference on Genome Informatics*
959 23, 205–211. doi:10.1142/9781848165632_0019.
960 Edwards, K. D., Fernandez-Pozo, N., Drake-Stowe, K., Humphry, M., Evans, A. D.,
961 Bombarely, A., et al. (2017). A reference genome for Nicotiana tabacum enables
962 map-based cloning of homeologous loci implicated in nitrogen utilization
963 efficiency. *BMC Genomics* 18, 448. doi:10.1186/s12864-017-3791-6.
964 Ehlting, J., Büttner, D., Wang, Q., Douglas, C. J., Somssich, I. E., and Kombrink, E.
965 (1999). Three 4-coumarate:coenzyme A ligases in Arabidopsis thaliana represent
966 two evolutionarily divergent classes in angiosperms. *The Plant journal : for cell*
967 *and molecular biology* 19, 9–20. doi:10.1046/j.1365-313x.1999.00491.x.
968 Ehlting, J., Mattheus, N., Aeschliman, D. S., Li, E., Hamberger, B., Cullis, I. F., et al.
969 (2005). Global transcript profiling of primary stems from Arabidopsis thaliana
970 identifies candidate genes for missing links in lignin biosynthesis and
971 transcriptional regulators of fiber differentiation. *The Plant journal : for cell and*
972 *molecular biology* 42, 618–640. doi:10.1111/j.1365-313x.2005.02403.x.
973 Franke, R., Höfer, R., Briesen, I., Emsermann, M., Efremova, N., Yephremov, A., et al.
974 (2009). The DAISY gene from Arabidopsis encodes a fatty acid elongase
975 condensing enzyme involved in the biosynthesis of aliphatic suberin in roots and
976 the chalaza-micropyle region of seeds. *The Plant Journal* 57, 80–95.
977 doi:https://doi.org/10.1111/j.1365-313X.2008.03674.x.
978 Franke, R., and Schreiber, L. (2007). Suberin--a biopolyester forming apoplastic plant
979 interfaces. *Current opinion in plant biology* 10, 252–259.
980 doi:10.1016/j.pbi.2007.04.004.
981 Fulda, M., Heinz, E., and Wolter, F. P. (1994). The fadD gene of Escherichia coli K12
982 is located close to rnd at 39.6 min of the chromosomal -map and is a new member
983 of the AMP-binding protein family. *Molecular and General Genetics MGG* 242,
984 241–249. doi:10.1007/BF00280412.
985 Garcia-Mas, J., Benjak, A., Sanseverino, W., Bourgeois, M., Mir, G., González, V. M.,
986 et al. (2012). The genome of melon (Cucumis melo L.). *Proceedings of the*
987 *National Academy of Sciences of the United States of America* 109, 11872–11877.
988 doi:10.1073/pnas.1205415109.
989 Graça, J., and Pereira, H. (2000). Suberin Structure in Potato Periderm: Glycerol,
990 Long-Chain Monomers, and Glyceryl and Feruloyl Dimers. *Journal of*

Agricultural and Food Chemistry 48, 5476–5483. doi:10.1021/jf0006123.

Graça, J., and Santos, S. (2007). Suberin: A Biopolyester of Plants' Skin. *Macromolecular Bioscience* 7, 128–135. doi:https://doi.org/10.1002/mabi.200600218.

Grossman, A. R., Harris, E. E., Hauser, C., Lefebvre, P. A., Martinez, D., Rokhsar, D., et al. (2003). Chlamydomonas reinhardtii at the Crossroads of Genomics. *Eukaryotic Cell* 2, 1137. doi:10.1128/EC.2.6.1137-1150.2003.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology* 59, 307–321. doi:10.1093/sysbio/syq010.

Gulick, A. M. (2009). Conformational Dynamics in the Acyl-CoA Synthetases, Adenylation Domains of Non-ribosomal Peptide Synthetases, and Firefly Luciferase. *ACS Chemical Biology* 4, 811–827. doi:10.1021/cb900156h.

Hamberger, B., and Hahlbrock, K. (2004). The 4-coumarate:CoA ligase gene family in Arabidopsis thaliana comprises one rare, sinapate-activating and three commonly occurring isoenzymes. *Proceedings of the National Academy of Sciences of the United States of America* 101, 2209–2214. doi:10.1073/pnas.0307307101.

Harding, S. A., Leshkevich, J., Chiang, V. L., and Tsai, C.-J. (2002). Differential Substrate Inhibition Couples Kinetically Distinct 4-Coumarate:Coenzyme A Ligases with Spatially Distinct Metabolic Roles in Quaking Aspen. *Plant Physiology* 128, 428. doi:10.1104/pp.010603.

Huang, S., Li, R., Zhang, Z., Li, L., Gu, X., Fan, W., et al. (2009). The genome of the cucumber, Cucumis sativus L. *Nature Genetics* 41, 1275–1281. doi:10.1038/ng.475.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics* 14, 33–8, 27–8. doi:10.1016/0263-7855(96)00018-5.

Jaillon, O., Aury, J.-M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* 449, 463–467. doi:10.1038/nature06148.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* 44, D457–D462. doi:10.1093/nar/gkv1070.

Katoh, K., and Standley, D. (2013). Katoh K, Standley DM.. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in performance and usability. Mol Biol Evol 30: 772-780. *Molecular biology and evolution* 30. doi:10.1093/molbev/mst010.

Kienow, L., Schneider, K., Bartsch, M., Stuible, H.-P., Weng, H., Miersch, O., et al. (2008). Jasmonates meet fatty acids: Functional analysis of a new acyl-coenzyme A synthetase family from Arabidopsis thaliana. *Journal of experimental botany* 59, 403–419. doi:10.1093/jxb/erm325.

Knobloch, K. H., and Hahlbrock, K. (1977). 4-Coumarate:CoA ligase from cell suspension cultures of Petroselinum hortense Hoffm. Partial purification, substrate specificity, and further properties. *Archives of Biochemistry and Biophysics* 184, 237–248. doi:10.1016/0003-9861(77)90347-2.

Knobloch, K., and Hahlbrock, K. (1975). Isoenzymes of p-Coumarate □: CoA Ligase from Cell Suspension Cultures of Glycine rnax. 320, 311–320.

Kolattukudy, P. E. (2001). "Polyesters in Higher Plants," in *Biopolyesters*, ed. A. Babel Wolfgangand Steinbüchel (Berlin, Heidelberg: Springer Berlin Heidelberg), 1–49.

doi:10.1007/3-540-40021-4_1.

Koo, A. J. K., Chung, H. S., Kobayashi, Y., and Howe, G. A. (2006). Identification of a Peroxisomal Acyl-activating Enzyme Involved in the Biosynthesis of Jasmonic Acid in Arabidopsis *♦ *Journal of Biological Chemistry* 281, 33511–33520. doi:10.1074/jbc.M607854200.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research* 40, D1202–D1210. doi:10.1093/nar/gkr1090.

Lapierre, C., Pollet, B., and Négrel, J. (1996). The phenolic domain of potato suberin: Structural comparison with lignins. *Phytochemistry* 42, 949–953. doi:https://doi.org/10.1016/0031-9422(96)00097-0.

Lee, D., and Douglas, C. J. (1996). Two divergent members of a tobacco 4-coumarate:coenzyme A ligase (4CL) gene family. cDNA structure, gene inheritance and expression, and properties of recombinant proteins. *Plant physiology* 112, 193–205. doi:10.1104/pp.112.1.193.

Lemoine, F., Entfellner, J.-B., Wilkinson, E., Correia, D. B., Dávila Felipe, M., de Oliveira, T., et al. (2018). Renewing Felsenstein's Phylogenetic Bootstrap in the Era of Big Data. *Nature* 556. doi:10.1038/s41586-018-0043-0.

Letunic, I., and Bork, P. (2007). Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. *Bioinformatics (Oxford, England)* 23, 127–128. doi:10.1093/bioinformatics/btl529.

Li, W., and Godzik, A. (2006). Cd-Hit: a Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. *Bioinformatics (Oxford, England)* 22, 1658–1659. doi:10.1093/bioinformatics/btl158.

Li, X., Wang, C., Sun, H., and Li, T. (2011). Establishment of the total RNA extraction system for lily bulbs with abundant polysaccharides. *African Journal of Biotechnology* 10, 17907–17915. doi:10.5897/AJB10.2523.

Li, Y., Li, D., Guo, Z., Shi, Q., Xiong, S., Zhang, C., et al. (2016). OsACOS12, an orthologue of Arabidopsis acyl-CoA synthetase5, plays an important role in pollen exine formation and anther development in rice. *BMC Plant Biology* 16, 256. doi:10.1186/s12870-016-0943-9.

Li, Z., and Nair, S. K. (2015). Structural Basis for Specificity and Flexibility in a Plant 4-Coumarate:CoA Ligase. *Structure* 23, 2032–2042. doi:https://doi.org/10.1016/j.str.2015.08.012.

Lindermayr, C., Fliegmann, J., and Ebel, J. (2003). Deletion of a Single Amino Acid Residue from Different 4-Coumarate:CoA Ligases from Soybean Results in the Generation of New Substrate Specificities *. *Journal of Biological Chemistry* 278, 2781–2786. doi:10.1074/jbc.M202632200.

Livak, K. J., and Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods (San Diego, Calif.)* 25, 402–8. doi:10.1006/meth.2001.1262.

Massa, A. N., Childs, K. L., Lin, H., Bryan, G. J., Giuliano, G., and Buell, C. R. (2011). The Transcriptome of the Reference Potato Genome Solanum tuberosum Group Phureja Clone DM1-3 516R44. *PLOS ONE* 6, e26801-. Available at: https://doi.org/10.1371/journal.pone.0026801.

Mattinen, M.-L., Filpponen, I., Järvinen, R., Li, B., Kallio, H., Lehtinen, P., et al. (2009). Structure of the Polyphenolic Component of Suberin Isolated from Potato (Solanum tuberosum var. Nikola). *Journal of Agricultural and Food Chemistry* 57, 9747–9753. doi:10.1021/jf9020834.

1091 Mazin, P. V, Gelfand, M. S., Mironov, A. A., Rakhmaninova, A. B., Rubinov, A. R.,
1092 Russell, R. B., et al. (2010). An automated stochastic approach to the identification
1093 of the protein specificity determinants and functional subfamilies. *Algorithms for*
1094 *Molecular Biology* 5, 29. doi:10.1186/1748-7188-5-29.
1095 Navarre, D. A., Pillai, S. S., Shakya, R., and Holden, M. J. (2011). HPLC profiling of
1096 phenolics in diverse potato genotypes. *Food Chemistry* 127, 34–41.
1097 doi:10.1016/j.foodchem.2010.12.080.
1098 Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007). The
1099 TIGR Rice Genome Annotation Resource: improvements and new features.
1100 *Nucleic acids research* 35, D883-7. doi:10.1093/nar/gkl976.
1101 Pagnuco, I. A., Revuelta, M. V., Bondino, H. G., Brun, M., and ten Have, A. (2018).
1102 HMMER Cut-off Threshold Tool (HMMERCTTER): Supervised classification of
1103 superfamily protein sequences with a reliable cut-off threshold. *PLOS ONE* 13,
1104 e0193757-. Available at: https://doi.org/10.1371/journal.pone.0193757.
1105 Payyavula, R. S., Navarre, D. A., Kuhl, J. C., Pantoja, A., and Pillai, S. S. (2012).
1106 Differential effects of environment on potato phenylpropanoid and carotenoid
1107 expression. *BMC plant biology* 12, 39. doi:10.1186/1471-2229-12-39.
1108 Peter, G., and Neale, D. (2004). Molecular basis for the evolution of xylem
1109 lignification. *Current Opinion in Plant Biology* 7, 737–742.
1110 doi:10.1016/j.pbi.2004.09.002.
1111 Pollard, M., Beisson, F., Li-Beisson, Y., and Ohlrogge, J. (2008). Building lipid
1112 barriers: Biosynthesis of cutin and suberin. *Trends in plant science* 13, 236–246.
1113 doi:10.1016/j.tplants.2008.03.003.
1114 Raes, J., Rohde, A., Christensen, J. H., Van de Peer, Y., and Boerjan, W. (2003).
1115 Genome-wide characterization of the lignification toolbox in Arabidopsis. *Plant*
1116 *physiology* 133, 1051–1071. doi:10.1104/pp.103.026484.
1117 Rausell, A., Juan, D., Pazos, F., and Valencia, A. (2010). Protein interactions and ligand
1118 binding: From protein subfamilies to functional specificity. *Proceedings of the*
1119 *National Academy of Sciences* 107, 1995. doi:10.1073/pnas.0908044107.
1120 Reumann, S., Ma, C., Lemke, S., and Babujee, L. (2004). AraPerox. A database of
1121 putative Arabidopsis proteins from plant peroxisomes. *Plant physiology* 136,
1122 2587–2608. doi:10.1104/pp.104.043695.
1123 Robert, X., and Gouet, P. (2014). Deciphering key features in protein structures with the
1124 new ENDscript Server. *Nucleic acids research* 42. doi:10.1093/nar/gku316.
1125 Russell, R. B., and Barton, G. J. (1992). Multiple protein sequence alignment from
1126 tertiary structure comparison: Assignment of global and residue confidence levels.
1127 *Proteins: Structure, Function, and Bioinformatics* 14, 309–323.
1128 doi:https://doi.org/10.1002/prot.340140216.
1129 Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., et al. (2012).
1130 The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*
1131 485, 635–641. doi:10.1038/nature11119.
1132 Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., et al. (2010).
1133 Genome sequence of the palaeopolyploid soybean. *Nature* 463, 178–183.
1134 doi:10.1038/nature08670.
1135 Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., et al. (2009).
1136 The B73 Maize Genome: Complexity, Diversity, and Dynamics. *Science* 326,
1137 1112. doi:10.1126/science.1178534.
1138 Schneider, K., Kienow, L., Schmelzer, E., Colby, T., Bartsch, M., Miersch, O., et al.
1139 (2005). A New Type of Peroxisomal Acyl-Coenzyme A Synthetase from
1140 Arabidopsis thaliana Has the Catalytic Capacity to Activate

Biosynthetic Precursors of Jasmonic Acid *. *Journal of Biological Chemistry* 280, 13962–13972. doi:10.1074/jbc.M413578200.

Schreiber, L., Franke, R., and Hartmann, K. (2005). Wax and suberin development of native and wound periderm of potato (Solanum tuberosum L.) and its relation to peridermal transpiration. *Planta* 220, 520–30. doi:10.1007/s00425-004-1364-9.

Serra, O., Hohn, C., Franke, R., Prat, S., Molinas, M., and Figueras, M. (2010). A feruloyl transferase involved in the biosynthesis of suberin and suberin-associated wax is required for maturation and sealing properties of potato periderm. *Plant Journal* 62, 277–290.

Sharma, S. K., Bolser, D., de Boer, J., Sønderkær, M., Amoros, W., Carboni, M. F., et al. (2013). Construction of reference chromosome-scale pseudomolecules for potato: integrating the potato genome with genetic and physical maps. *G3 (Bethesda, Md.)* 3, 2031–2047. doi:10.1534/g3.113.007153.

Shockey, J., Fulda, M., and Browse, J. (2003). Arabidopsis Contains a Large Superfamily of Acyl-Activating Enzymes. Phylogenetic and Biochemical Analysis Reveals a New Class of Acyl-Coenzyme A Synthetases. *Plant physiology* 132, 1065–1076. doi:10.1104/pp.103.020552.

Shulaev, V., Sargent, D. J., Crowhurst, R. N., Mockler, T. C., Folkerts, O., Delcher, A. L., et al. (2011). The genome of woodland strawberry (Fragaria vesca). *Nature Genetics* 43, 109–116. doi:10.1038/ng.740.

Simonetti, F., Teppa, E., Chernomoretz, A., Nielsen, M., and Marino Buslje, C. (2013). MISTIC: Mutual information server to infer coevolution. *Nucleic acids research* 41. doi:10.1093/nar/gkt427.

Soubeyrand, E., Kelly, M., Keene, S., Bernert, A., Latimer, S., Johnson, T., et al. (2019). Arabidopsis 4-COUMAROYL-COA LIGASE 8 contributes to the biosynthesis of the benzenoid ring of coenzyme Q in peroxisomes. *Biochemical Journal* 476. doi:10.1042/BCJ20190688.

Stushnoff, C., Ducreux, L. J. M., Hancock, R. D., Hedley, P. E., Holm, D. G., McDougall, G. J., et al. (2010). Flavonoid profiling and transcriptome analysis reveals new gene-metabolite correlations in tubers of Solanum tuberosum L. *Journal of experimental botany* 61, 1225–1238. doi:10.1093/jxb/erp394.

Sun, H., Li, Y., Feng, S., Zou, W., Guo, K., Fan, C., et al. (2013). Analysis of five rice 4-coumarate:coenzyme A ligase enzyme activity and stress response for potential roles in lignin and flavonoid biosynthesis in rice. *Biochemical and biophysical research communications* 430, 1151–1156. doi:10.1016/j.bbrc.2012.12.019.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The Genome of Black Cottonwood, Populus trichocarpa (Torr. & Gray). *Science* 313, 1596. doi:10.1126/science.1128691.

Valiñas, M., Lanteri, M., Have, A., and Andreu, A. (2017). Chlorogenic Acid, Anthocyanin and Flavan-3-ol Biosynthesis in Flesh and Skin of Andean Potato Tubers (Solanum tuberosum subsp. andigena). *Food Chemistry* 229. doi:10.1016/j.foodchem.2017.02.150.

Valiñas, M., Lanteri, M., ten Have, A., and Andreu, A. (2015). Chlorogenic acid biosynthesis appears linked with suberin production in potato tuber (Solanum tuberosum). *Journal of Agricultural and Food Chemistry* 63, 4902–4913. Available at: http://pubs.acs.org/doi/abs/10.1021/jf505777p [Accessed June 23, 2015].

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., et al. (2010). The genome of the domesticated apple (Malus × domestica Borkh.).

Nature Genetics 42, 833–839. doi:10.1038/ng.654.

Winter, D., Vinegar, B., Nahal, H., Ammar, R., Wilson, G. V, and Provart, N. J. (2007). An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. *PLOS ONE* 2, e718-. Available at: https://doi.org/10.1371/journal.pone.0000718.

Wu, S., Lau, K. H., Cao, Q., Hamilton, J. P., Sun, H., Zhou, C., et al. (2018). Genome sequences of two diploid wild relatives of cultivated sweetpotato reveal targets for genetic improvement. *Nature Communications* 9, 4580. doi:10.1038/s41467-018-06983-8.

Yan, B., and Stark, R. E. (2000). Biosynthesis, Molecular Structure, and Domain Architecture of Potato Suberin: A 13C NMR Study Using Isotopically Labeled Precursors. *Journal of Agricultural and Food Chemistry* 48, 3298–3304. doi:10.1021/jf000155q.

Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2014). The I-TASSER suite: Protein structure and function prediction. *Nature methods* 12, 7–8. doi:10.1038/nmeth.3213.

Zhang, Z., Schäffer, A. A., Miller, W., Madden, T. L., Lipman, D. J., Koonin, E. V, et al. (1998). Protein sequence similarity searches using patterns as seeds. *Nucleic acids research* 26, 3986–3990. doi:10.1093/nar/26.17.3986.

TABLES

Table 1: Description of the potato varieties used in this study.

Variety	[a]Flesh/ Skin Color	Relative flesh/ skin phenolic content
Chaqueña	Cpk/ PK	Intermediate/ Intermediate
Santa María	R/ R	High/ High
Waicha	Y/ PK	Low/ Intermediate
Moradita	Y/ P	Low/ High

[a]Primary (in uppercase) and eventual secondary (lowercase) flesh color/ primary skin color are indicated. Y, yellow; C, cream; PK, pink; R, red; P, purple.

Table 2: Suggested Nomenclature for potato 4CL/ ACS genes

Clustering			Annotation				
			Arabidopsis thaliana / Oryza sativa			*Solanum tuberosum*	
Tree	**HC**	**Class**	**Genome**	**Name[1]**	**Activity (At)**	**Genome**	**Name**
1.1	1	4CL Class I Dicot	At1g51680 At3g21240 At3g21230	At4CL1 At4CL2 At4CL4		PGSC0003DMP400025029 PGSC0003DMP400050416 PGSC0003DMP400034532 PGSC0003DMP400025484	St4CL-IA St4CL-IB St4CL-IC St4CL-ID
1.2	9	4CL Class I Monocot	Os08g14760 Os02g08100 Os06g44620 Os08g34790	Os4CL1 Os4CL3 Os4CL4 Os4CL5		-	-
1.3	7	Class II	At1g65060 Os02g46970	At4CL3 Os4CL2		PGSC0003DMP400005703	St4CL-II
2	4	ACS[1] ACS[2]	At1g62940 Os04g24530	AtACS5 AtACS9	Moderate ACS[4]	XP_006338382* Soltu.DM.02G031030.1***	StACS
3.1	2	Mixed	At5g63380 Os01g67530 Os01g67540 Os07g17970 Os07g44560	AtACS9 OsACS5 OsACS6 OsACS7 OsACS8	OPCL[4] ACS[4]	PGSC0003DMP400026651	St4CL-H1
3.2	3	Mixed	At4g05160 Os03g05780	AtACS6 OsACS1	OPCL[4] ACS[4]	PGSC0003DMP400014290	St4CL-H2
3.3	5	4CL[3] Class III ACS Mixed OPCL[4] Mixed	At5g38120 At1g20480 At1g20490 At1g20500 At1g20510 Os03g04000	AtACS8 AtACS1 AtACS2 AtACS3 AtOPCL1 OsACS4	Moderate 4CL[4] Moderate ACS[4] 4CL[4] OPCL[4] ACS[4] OPCL[4]	PGSC0003DMP400051055	St4CL-H3
3.4	6	4CL[5] Class III	At4g19010	AtACS7	4CL[4] ACS[4]	PGSC0003DMP400016385	St4CL-H4/

			Os10g42800 Os08g04770	OsACS2 OsACS3			St4CL-III
3.5	8	Unknown	- -			XP_006361720** Soltu.DM.11G004700.1***	St4CL-H5/ St4CL-like

[1] de Azevedo Souza et al. (2009); [2] Li et al. (2016); [3] Soubeyrand et al. (2019); [4] Kienow et al. (2008); [5] Block et al. (2014); * Not detected in the reference proteome at Potato Genome Sequencing Consortium (PGSC); ** Not detected in the reference proteome at Potato Genome Sequencing Consortium (PGSC Version 4.03) and not included in phylogeny since not 100% complete. *** Corresponding sequences were recently found in V6.1 of PGSC.

FIGURE LEGENDS

Figure 1: Schematic diagram of phenolic compounds biosynthetic pathways leading to flavonoids, lignin and suberin and ubiquinone in plants. PAL, phenylalanine ammonia-lyase; C4H, cinnamate 4-hydroxylase; 4CL, 4-coumarate-CoA ligase; CHS, chalcone synthase; FHT, feruloyl-CoA transferase.

Figure 2: Hierarchic phylogenetic clustering of the plant acyl-CoA synthetase (4CL/ ACS) family. Radial phylogram representation of the maximum likelihood tree. Black dots indicate all major bifurcations have over 0.9 normalized bootstrap support. The scale bar indicates 0.1 amino acid substitution per site. * indicate sequences derived from lower plant *S. moellendorffiii*. Sequence logos for the box I and box II subsequences as well as the ten major superfamily SDPs are shown for the three major clades. SDP positions indicated correspond with PDB identifier 5BST. Logos and SDP were obtained and identified using an enlarged dataset as described in the main text.

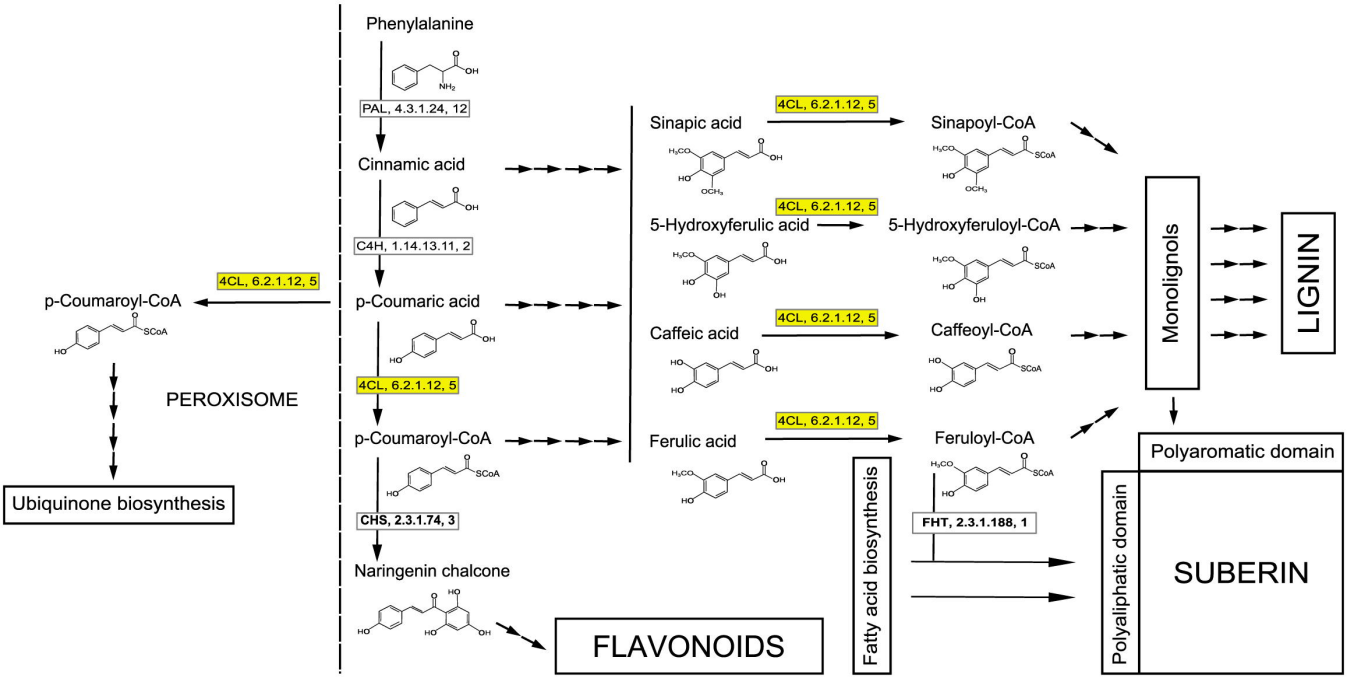
Figure 3: The major specificity determining positions that explain diversification towards three major 4CL/ ACS subfamilies. The identified major SDPs are presented on a tobacco class I 4CL, crystallized in the presence of coumaroyl-AMP (PDB identifier 5BST). **(A)** Global view. **(B)** Detail in absence of protein cartoon view. The protein is shown in cartoon. SDP340 (orange licorice and surf) physically interacts with the coumaroyl group. SDPs 338 and 342 (yellow licorice and surf) interact with SDP340, whereas SDPs 363 and 369 (green licorice and surf) interact with SDP342. Other distant SDPs are in gray VDW.

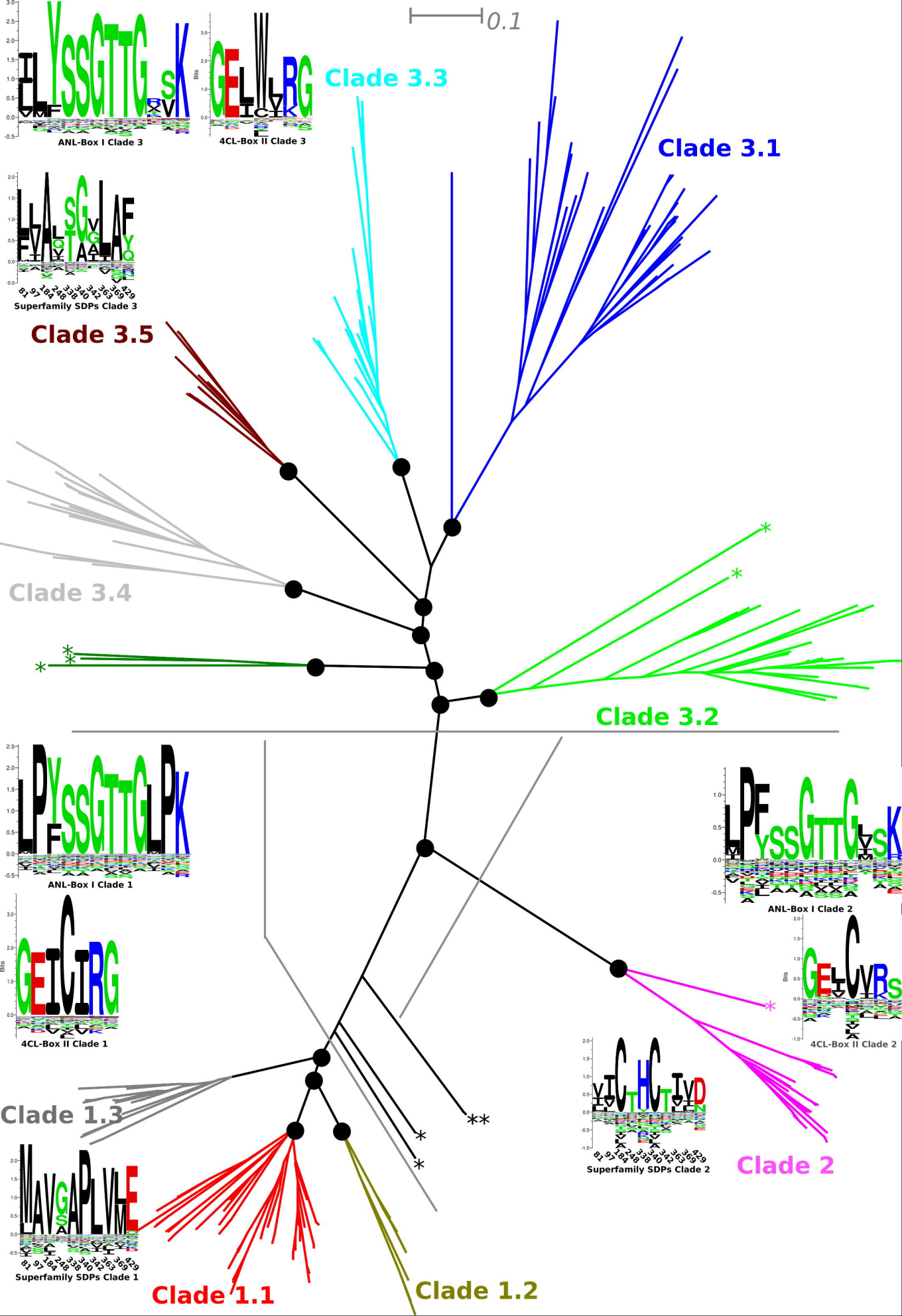
Figure 4: Substitution of five specificity determining positions in binding pocket of the 4CL enzyme suggest clade 2 enzymes have a different substrate. (A) Structural alignment of tobacco class I 4CL (PDB indentifier 5BSV, gray cartoon) with the model of a potato clade 2 subfamily sequence (Cyan cartoon). Feruloyl-AMP in red licorice and surf; SDPs in orange licorice and surf. **(B-E)** Details of pocket. Colors and style as in **(A)**; yellow licorice and surf for clade 2 SDP counterparts. **(B and D)** Tobacco 4CL; **(C and E)** Clade 2 model. **(F)** Sequence logo of the 48 major SDP residues in the 4CL subfamily. **(G)** Sequence logo of the 48 major SDP residues in the 4CL-Like subfamily. Indicated are positions SDP308, 332, 340 and 360.

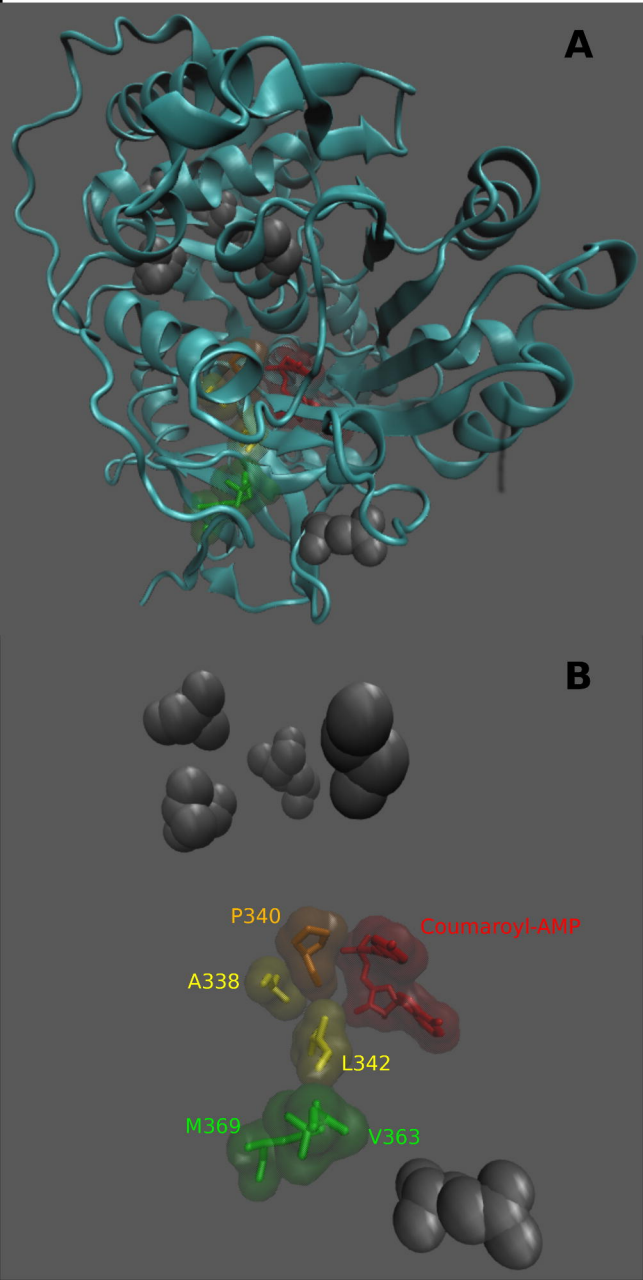
Figure 5: SDP Analysis of Clade 3.5 Shows a Signifcant Different Binding Cleft. (A) Mutual Information Connectivity network. Node scores are reflected by node colour according to a green (-1,03) to red (2,17) heat scale. Nodes that correspond to a residue located in the vicinity of either CoA or the feruloyl adenylate are in bold. **(B)** Structural alignment of the three virtual mutants described in the main text with 5BSR and 5BSV. Proteins are in cartoon, substrates in licorice and surf, red for the feruloyl adenylate and purple for the CoA. **(C)** Comparison of the binding cleft of the wildtype (left) and the 17-SDP mutant (right). Shown are SDPs and substrates only, in an enlarged screenshot using the same angle as in **(B)**. The binding cleft SDPs are in green or pink licorice and surf, major C1-C3 SDPs in blue or grey licorice, for wildtype and mutant respectively. **(D)** Sequence logos of major C1-C3 SDPs (left, top C1, bottom C3.5) and binding cleft SDPs (right, top C1, bottom C3.5). All numbers according to the 5BSR reference sequence.

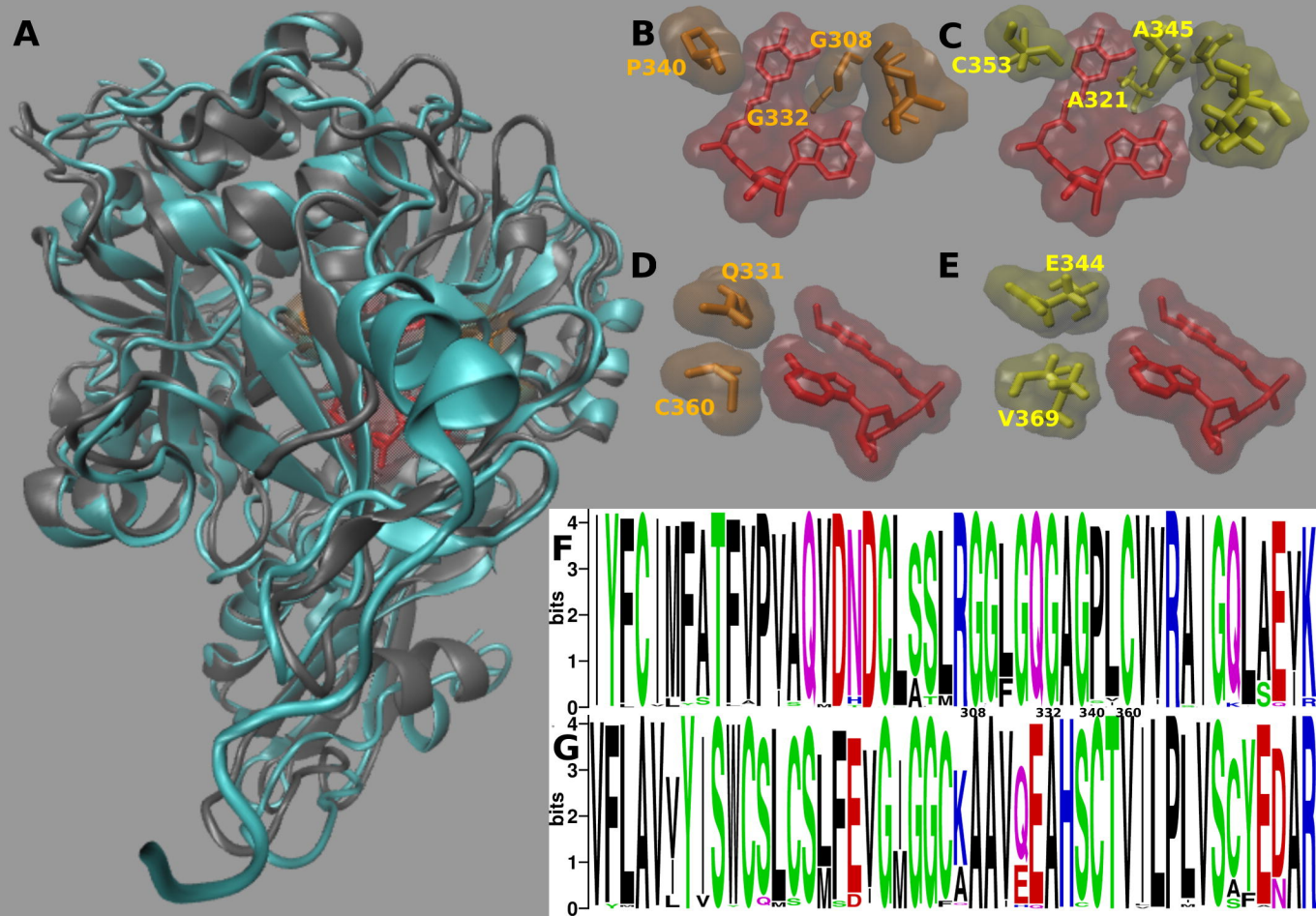
Figure 6: Transcript levels of *St4CL* genes in flesh and skin of tubers from Andean potato varieties from 2010/ 2011 and 2011/ 2012 campaigns. qRT-PCR data represent mean ± SD values from two independent plates of amplification with two wells per cDNA. cDNA was produced from one RNA extraction from a pool of ten potato tubers.

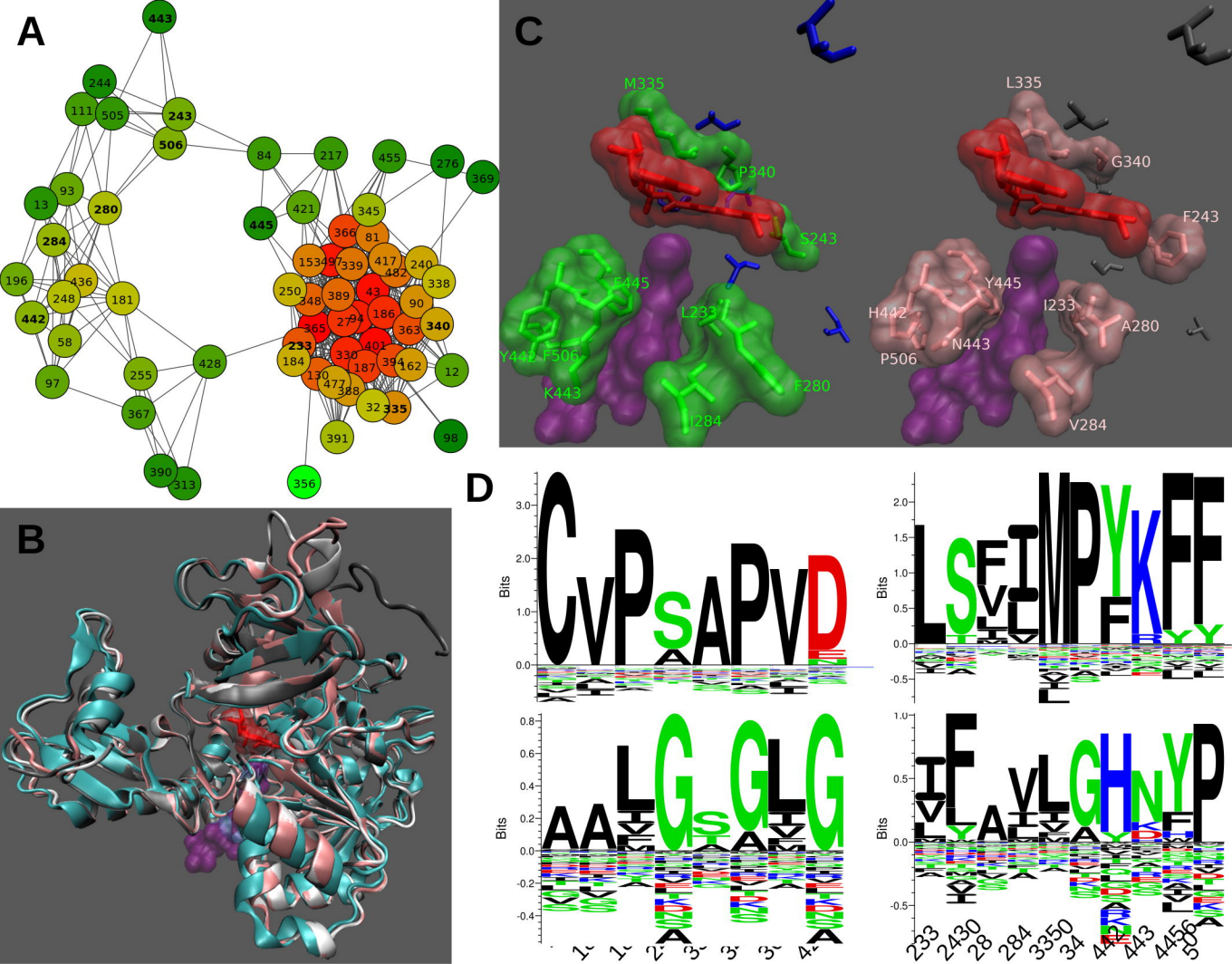
Figure 7: Expression profile of *St4CL-IA*, *St4CL-IB* and *St4CL-II* genes 72 hs post-wounding of Santa María potato tubers. qRT-PCR data represent mean ± SD values from two independent biological replicates with two technical replicates each from one of two representative experiments.

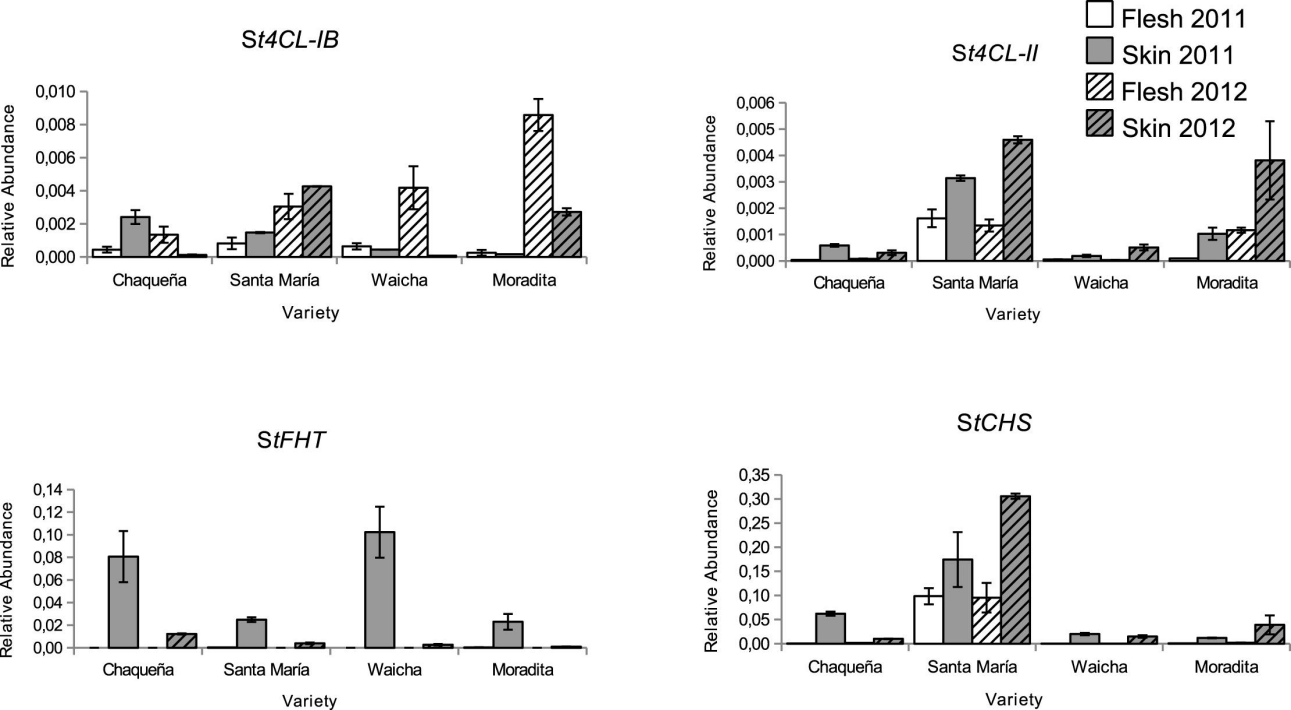

Phenylalanine
PAL, 4.3.1.24, 12
Cinnamic acid
C4H, 1.14.13.11, 2
p-Coumaric acid
4CL, 6.2.1.12, 5
p-Coumaroyl-CoA
CHS, 2.3.1.74, 3
Naringenin chalcone
FLAVONOIDS
4CL, 6.2.1.12, 5
p-Coumaroyl-CoA
PEROXISOME
Ubiquinone biosynthesis
Sinapic acid
4CL, 6.2.1.12, 5
Sinapoyl-CoA
5-Hydroxyferulic acid
4CL, 6.2.1.12, 5
5-Hydroxyferuloyl-CoA
Caffeic acid
4CL, 6.2.1.12, 5
Caffeoyl-CoA
Ferulic acid
4CL, 6.2.1.12, 5
Feruloyl-CoA
Monolignols
LIGNIN
Fatty acid biosynthesis
FHT, 2.3.1.188, 1
Polyaromatic domain
Polyaliphatic domain
SUBERIN

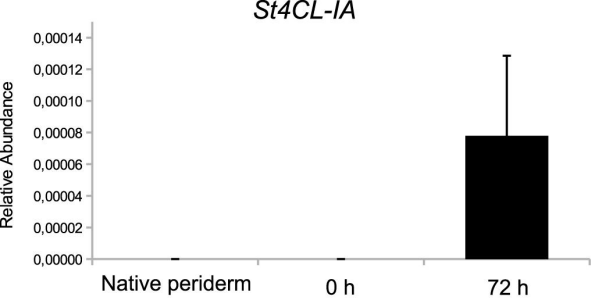

0.1
ANL-Box I Clade 3
4CL-Box II Clade 3
Clade 3.3
Clade 3.1
Superfamily SDPs Clade 3
Clade 3.5
Clade 3.4
Clade 3.2
ANL-Box I Clade 1
ANL-Box I Clade 2
4CL-Box II Clade 1
4CL-Box II Clade 2
Clade 1.3
Superfamily SDPs Clade 2
Clade 2
Superfamily SDPs Clade 1
Clade 1.1
Clade 1.2

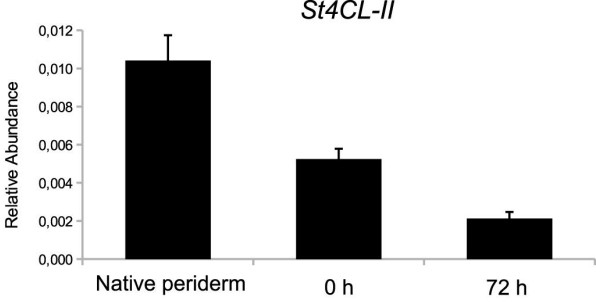

A
B
P340
Coumaroyl-AMP
A338
L342
M369
V363

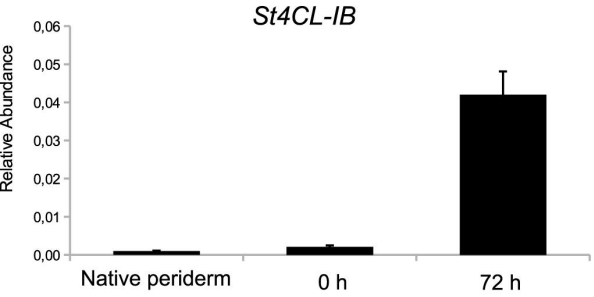

A
B
P340
G308
G332
C
C353
A345
A321
D
Q331
C360
E
E344
V369
F
G
bits
308
332
340
360

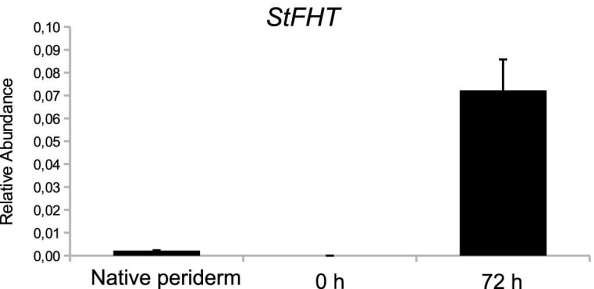

A
B
C
D
M335
P340
S243
L335
G340
F243
Y445
H442
I233
A280
N443
P506
V284
L233
F280
I284
K443
Bits

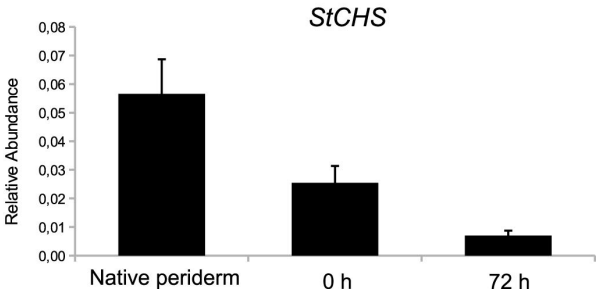

St4CL-IB
Relative Abundance
0,010
0,008
0,006
0,004
0,002
0,000
Chaqueña
Santa María
Waicha
Moradita
Variety
St4CL-II
Flesh 2011
Skin 2011
Flesh 2012
Skin 2012
0,006
0,005
0,004
0,003
0,002
0,001
0,000
StFHT
0,14
0,12
0,10
0,08
0,06
0,04
0,02
0,00
StCHS
0,35
0,30
0,25
0,20
0,15
0,10
0,05
0,00


St4CL-IA
Relative Abundance
0,00014
0,00012
0,00010
0,00008
0,00006
0,00004
0,00002
0,00000
Native periderm
0 h
72 h


St4CL-II
Relative Abundance
0,012
0,010
0,008
0,006
0,004
0,002
0,000
Native periderm
0 h
72 h


St4CL-IB
Relative Abundance
0,06
0,05
0,04
0,03
0,02
0,01
0,00
Native periderm
0 h
72 h


StFHT
Relative Abundance
0,10
0,09
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0,00
Native periderm
0 h
72 h


StCHS
Relative Abundance
0,08
0,07
0,06
0,05
0,04
0,03
0,02
0,01
0,00
Native periderm
0 h
72 h