Full title

Transcriptome of the coralline alga *Calliarthron tuberculosum* (Corallinales, Rhodophyta) reveals convergent evolution of a partial lignin biosynthesis pathway

Short title

Calliarthron transcriptomics reveals monolignol biosynthesis pathway

Author names

Jan Y Xue[a,*], Katy Hind[ab], Matthew A. Lemay[ab], Andrea Mcminigal[a], Emma Jourdain[a], Cheong Xin Chan[c], and Patrick T. Martone[ab]

[a] Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada

[b] Hakai Institute, Heriot Bay, BC, V0P 1H0, Canada

[c] Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland, 4072, Australia

[*] Corresponding author: jan.xue@botany.ubc.ca

Abstract

The discovery of lignins in the coralline red alga *Calliarthron tuberculosum* raised new questions about the deep evolution of lignin biosynthesis. Here we present the transcriptome of *C. tuberculosum* supported with newly generated genomic data to identify gene candidates from the monolignol biosynthetic pathway using a combination of sequence similarity-based methods. We identified candidates in the monolignol biosynthesis pathway for the genes 4CL, CCR, CAD, CCoAOMT, and CSE but did not identify candidates for PAL, CYP450 (F5H, C3H, C4H), HCT, and COMT. In gene tree analysis, we present evidence that these gene candidates evolved independently from their land plant counterparts, suggesting convergent evolution of a complex multistep lignin biosynthetic pathway in this red algal lineage. Additionally, we provide tools to extract metabolic pathways and genes from the newly generated transcriptomic and genomic datasets. Using these methods, we extracted genes related to sucrose metabolism and calcification. Ultimately, this transcriptome will provide a foundation for further genetic and experimental studies of calcifying red algae.

Keywords: Red algae, lignification, calcification, transcriptome, gene identification, phenylpropanoid pathway, monolignol

Introduction

Coralline red algae (Corallinales, Sporolithales, Hapalidiales) are a diverse lineage of calcified seaweeds that play important ecological roles in nearshore ecosystems worldwide: they stabilize coral reefs by creating a calcium carbonate matrix [1–3], induce settlement of invertebrate taxa [4–6], and contribute to the storage of blue carbon through the creation of biogenic calcium carbonates [7,8]. In recent years, there has been increased global attention paid to coralline algae. Taxonomists are clarifying their vastly underestimated species diversity [9–12]; ecologists and physiologists are documenting interspecific variation in coralline growth and calcification, particularly in response to climate stress, which may ultimately impact marine communities [13–17]; evolutionary biologists are examining patterns in coralline trait evolution [18–20] and using >100 million-year-old coralline fossils to strengthen modern phylogenies [21,22].

The discovery of lignins within cell walls of the coralline species *Calliarthron cheilosporioides* (Corallinales, Rhodophyta) dramatically changed our perspective on the evolution of lignin biosynthesis [23]. Lignins are complex aromatic polymers predominantly found in the secondary cell walls of plant support tissues [24,25] and were long considered to have evolved when land plants emerged from the oceans, enabling upright growth in air [26]. Among the principal chemical components of wood, lignins in plant secondary cell walls help reinforce tissue mechanical properties, permit hydraulic transport, and increase pathogen resistance [27,28]. In the articulated coralline *C. cheilosporioides*, lignins were found predominantly within decalcified flexible joints, called genicula [23], that have remarkable biomechanical properties, permitting this articulated coralline species to thrive along wave-battered coastlines [29,30].

Because lignin biosynthesis is physiologically complex and involves several enzymes in the monolignol pathway [31–33], Martone et al. [23] proposed that much of the lignin biosynthetic pathway may have predated land plants altogether, evolving in a common ancestor of red and green algae more than one billion years ago. Alternatively, some (or all) of the monolignol biosynthetic pathway may have evolved independently in the embryophyte and rhodophyte lineages. For example, one important enzyme involved in S-lignin production (F5H) evolved independently in lycopods and embryophytes [34,35]. Moreover, candidate genes related to monolignol biosynthesis have since been found in diverse algal lineages such as diatoms, dinoflagellates, haptophytes, cryptophytes, and green and red algae [36], raising questions about how the monolignol pathway may have evolved across such evolutionarily divergent lineages. Until now, questions about monolignol evolution have largely gone unanswered as transcriptomic and genomic data have mostly been limited to non-coralline red algae (e.g. [37–40] but see [41]).

Here we present a transcriptome of the articulated coralline *Calliarthron tuberculosum* (a sister species of *C. cheilosporioides*) to investigate the evolutionary history of monolignol biosynthesis. Additionally, though a complete mitochondrial genome [42] and a draft nuclear genome [43] of *C. tuberculosum* were previously published, herein we generated a revised nuclear genome assembly using new short-read sequence data to aid validation of transcriptomic reads. Based on comparative analysis of genome and transcriptome data, we identify gene candidates for a putative monolignol biosynthetic pathway in *C. tuberculosum* and investigate evolutionary relationships of these enzymes with those from other taxonomic groups, including their land plant counterparts. We also provide a list of annotated genes in the *C. tuberculosum* transcriptome and a simplified method for extracting genes from metabolic pathways. We illustrate the utility of this dataset by extracting gene candidates involved in sucrose metabolism and calcification. This transcriptomic dataset provides a foundation for future studies of coralline algal ecology, physiology, and evolution.

Results

The C. tuberculosum transcriptome is complete and supported by genomic data

Two transcriptomic datasets were generated from *Calliarthron* thalli: one from whole tissue (calcified intergenicula plus uncalcified genicula; sample I+G/PTM1 in the deposited data) and a second from intergenicular (i.e., calcified) tissue only (sample I/PTM2). Transcriptome sequencing based on RNA-Seq produced 38.8 total Gb of sequence data (17.3 Gb for sample I+G; 21.5 Gb for sample I). Reads were assembled *de novo* using Trinity. The whole tissue dataset had 172,700,376 total reads and the intergenicular tissue dataset had 215,491,160 total reads with an overall average coverage of 677-fold. A third reference transcriptome combining data from both tissues was assembled independently. All three datasets were combined for subsequent analysis to increase coverage and maximize discovery. The transcriptome data were considered complete based on the recovery of core eukaryotic genes (e.g. 94.5% of CEGMA and 87.8% of BUSCO genes based on TBLASTN; Fig S1A). Genomic sequences were also assembled for *C. tuberculosum* (Table S1), but these remain highly fragmented and were used only as additional support to the transcriptome data in subsequent searches below. More than half (18840; 56.6%) of the 33301 transcripts in the reference transcriptome were supported by the genome data (BLASTN, $E \leq 10^{-5}$).

The incomplete monolignol biosynthetic pathway in Calliarthron tuberculosum

The combined *C. tuberculosum* transcriptomic dataset was searched for genes encoding enzymes from the monolignol biosynthetic pathway. The transcriptomic dataset was translated into all six reading frames and queried with a combination of homology-based approaches, including HMMER searches and KEGG based annotations. Closest homologs from *Arabidopsis thaliana* were also verified (BLASTN, $E \leq 10^{-30}$). We identified gene candidates of 4CL, CCR, CAD, CSE, and CCoAOMT, but not HCT, COMT, PAL, TAL, or PTAL (Fig 1). PAL/TAL/PTAL was considered absent as only fragmented (and no full length) sequences were identified. Evidence for the presence of homologous p450 enzymes (C3H, C3H, and F5H) was weak; as a result, their status was classified as ambiguous (Fig 1). All sequences identified had genomic support (BLASTN, $E \leq 10^{-5}$) except for those identified for PAL/TAL/PTAL.

Fig 1. The presence of *C. tuberculosum* sequence candidates in the monolignol pathway. Red indicates presence of a putative homolog in *C. tuberculosum*; blue indicates no significant hits; green indicates ambiguous presence. Note how the PTAL/PAL/TAL sequences obtained from the HMMER search were indicated as absent as all sequences found were too short, 1/4-1/3 in length relative to those in land plants. All sequences identified have genomic support except for PTAL/PAL/TAL.

Candidate sequences from *C. tuberculosum* (bolded as contig_gene_isoform in Figs 2, 3, and 4) were characterized by comparing key residues with their land plant homologs in multiple sequence alignments. The evolutionary relationships between the identified *C. tuberculosum* sequences, closely related sequences in additional taxa, and sequences from the broader protein family of their land plant homologs were analyzed in gene trees. Below we describe in detail results for the main biosynthetic enzymes 4CL, CCR, and CAD (Figs 2, 3, and 4). Descriptions of the other biosynthetic enzymes CCoAMT, CSE, and the cytochrome P450 sequences C3H, C4H, F5H are found in Appendix S1 and Figs S2-S4.

Fig 2. 4CL candidates from *C. tuberculosum* in relation to plants and other taxa

(A) Partial alignment of *C. tuberculosum* candidates (bolded) and embryophyte 4CL sequences. Residues involved in hydroxycinnamate binding are indicated with black triangles [61,62]. Phenylalanine substrate binding pocket is indicated with Box I and Box II.

(B) Maximum likelihood acyl-activating enzyme (AAE) gene tree showing relationships between *Calliarthron* sequences (magenta dots) and other taxa (Embryophyta – dark green, Chlorophyta – light green, Rhodophyta – red, Animalia and Opisthokonta – purple, Bacteria and Cyanobacteria – blue, Oomycota, Mycetozoa and Fungi – yellow, Ochrophyta – brown). Functionally demonstrated plant 4CLs are labelled (+). Additional functional groups are labelled [44,45]. Ultrafast bootstrap values > 95 are marked by *. Model = WAG+F+G4. Sites with ≤ 80% occupancy were removed. Accession numbers can be found in Appendix S1.

Fig 3. CCR candidates from *C. tuberculosum* in relation to plants and other taxa

(A) Partial alignment *C. tuberculosum* candidates (bolded) and land plant CCR sequences. Catalytic residues are labelled with NWYCY [64] and additional residues are indicated above with a black box. NADPH binding pocket residues are indicated with black triangles [65] and the GXXGXX[A/G] motif is underlined [66]. Hydroxycinnamonyl binding pocket residues are indicated with a gray triangle [65].

(B) CCR maximum likelihood gene tree showing relationships between *C. tuberculosum* (magenta dots) and other taxa (Embryophyta – dark green, Chlorophyta – light green, Rhodophyta – red, Animalia and Opisthokonta – purple, Bacteria and Cyanobacteria – blue, Oomycota, Mycetozoa and Fungi – yellow, Ochrophyta – brown). Functionally demonstrated plant CCRs are labelled (+). Additional functional groups are labelled. Ultrafast bootstrap values >95 are marked by *. Model = LG+G4. Sites with ≤ 80% occupancy were removed. Accession numbers can be found in Appendix S1.

Fig 4. CAD candidates from *C. tuberculosum* in relation to plants and other taxa

(A) Partial alignment of *C. tuberculosum* CAD sequence candidates (bolded) with land plant CAD sequences. Zn^{+2} ion coordinating and proton shuttling residues are indicated with the black triangle, NADPH or NADH interacting residues are boxed. Hydrostatic interaction forming residues are indicated with a black box. Putative substrate-binding residues are indicated with grey boxes. [67–69]

(B) CAD maximum likelihood gene tree showing relationships between *C. tuberculosum* (magenta dots) and other taxa (Embryophyta – dark green, Chlorophyta – light green, Rhodophyta – red, Animalia and Opisthokonta – purple, Bacteria and Cyanobacteria – blue, Oomycota, Mycetozoa and Fungi – yellow, Ochrophyta – brown). Alcohol dehydrogenase (ADH) sequences from yeast, and aldehyde reductase (YAHK and AHR) sequences from *E. coli* were used as the ADH family is closely related to that of CAD [70,71]. Functionally demonstrated plant CADs are labelled (+). Additional functional groups are labelled. Ultrafast bootstrap values >95 are marked by *. Model = LG+G4. Sites with ≤ 80% occupancy were removed. Accession numbers can be found in Appendix S1.

Identification of 4CL candidates

4CL is an acyl-CoA synthase in the monolignol pathway and a member of the acyl-activating enzyme (AAE) superfamily. 4CL converts p-coumaric acid, caffeic acid, and ferulic acid into their respective hydroxycinnamoyl-CoA thioesters. We identified 11 candidate 4CL-coding transcripts: two based on

KEGG analysis and nine additional sequences based on HMMER searches (Fig 2A). A query of these sequences against the *A. thaliana* proteome returned related proteins within the acyl-activating enzyme superfamily but not the *A. thaliana* 4CL (Table S2). Moderate sequence conservation exists in substrate binding and hydroxycinnamate binding residues between 4CL candidates in *C. tuberculosum* (bolded) and 4CLs in land plants (identity similarity [IS] > 70% Fig 2A).

In the 4CL gene tree analysis, most *C. tuberculosum* sequences grouped with sequences from other Rhodophytes (Fig 2B). In addition, *C. tuberculosum* sequences grouped within several functional clades including malonate CoA ligase (ultrafast bootstrap support [BS] = 100%), succinylbenzoate CoA ligase (BS = 87%), oxylate CoA ligase (BS = 100%), acetyl CoA synthase (BS = 100%), and the long chain fatty acid CoA ligase (BS = 89%) (magenta dots, Fig 2B) [44,45]. In contrast, embryophyte 4CL sequences form a clade separated from candidate 4CL sequences in *C. tuberculosum* (BS = 99% Fig 2B) by the luciferase containing outgroup. Thus, 4CL candidates in *C. tuberculosum* did not show any clear homology to functionally demonstrated 4CL sequences from embryophytes.

Identification of CCR candidates

CCR is the first committed enzyme in the monolignol pathway, reducing cinnamoyl-CoA esters to cinnamaldehydes. We identified three sequences as candidate CCR-coding transcripts: one based on KEGG analysis and two additional sequences based on HMMER searches (Fig 3A). A query of these sequences against the *A. thaliana* proteome returned sequences within the CCR family (CCR7, CCR4, CCR-Like6) (Table S2). Substrate-binding residues (NWYCY) and the hydroxycinnamonyl-binding pocket showed low sequence conservation (IS <80%). In contrast, the core catalytic residues (S, T, and K) and NADPH-binding residues appear to be conserved (IS >90%) between the candidate sequences in *C. tuberculosum* and CCRs in land plants (Fig 3A).

In the CCR gene tree analysis, *C. tuberculosum* sequences varied in their relatedness to other taxa with some sequences closer to Rhodophytes and others more closely related to Oomycota/Mycetozoa/Fungi (Fig 3B). Additionally, CCR candidates in *C. tuberculosum* were mapped with epimerase dehydratase type sequences that included the *A. thaliana* CCR family (Fig 3B). Sequences from *C. tuberculosum* grouped with epimerase dehydratase type sequences of non-embryophyte origin. In contrast, embryophyte CCR, class 2 CCR, and CCR-like form an independent clade (BS >97%). The embryophyte CCR clade and the non-embryophyte epimerase dehydratase clade (containing sequences from *C. tuberculosum*) were more closely related than the embryophyte dihydroflavonol-4-reductase protein (DFR) group within the overall epimerase dehydratase family.

Identification of CAD candidates

CAD, the final step in the monolignol pathway, is an alcohol dehydrogenase converting various hydroxycinnamaldehydes to their respective hydroxycinnamyl alcohols. SAD, proposed to catalyze this same reaction for sinapyl monolignols [46], is added into our analysis despite debate over their function. We identified five sequences as candidate CAD-encoding transcripts: two based on KEGG analysis and three additional sequences based on HMMER searches (Fig 4A). A query of these sequences against the *A. thaliana* proteome returned CAD2 and other alcohol dehydrogenases (Table S2). NADPH-binding motifs show moderate conservation (IS >80%) (Fig 4A). One *C. tuberculosum* sequence showed high conservation with land plant counterparts, suggesting a promising CAD candidate (+ in Figs 3A and 3B).

In the CAD gene tree analysis, all *C. tuberculosum* sequences grouped with sequences from other Rhodophytes (Fig 4B). CAD candidates in *C. tuberculosum* were mapped with their embryophyte CAD counterparts and closely related alcohol dehydrogenases. Sequences from *C. tuberculosum* grouped together with oxidoreductases (BS = 100%), sorbitol dehydrogenases (BS = 100%), general alcohol dehydrogenases (BS = 100%), and an algal CAD clade (BS = 100%). Sequences in this algal CAD clade were based on previous sequence similarity-based annotation and have not been functionally demonstrated. In contrast, the land plant CAD and SAD sequences form their own clades (BS 100%; Fig 4B) that are separated from the *C. tuberculosum* candidates by the functionally distinct alcohol dehydrogenases, such as yeast alcohol dehydrogenase 7 (ADH7) and *E. coli* aldehyde reductase (YAHK).

Identification of additional metabolic pathways in Calliarthron tuberculosum

To enable broad and rapid identification of *C. tuberculosum* genes involved in specific metabolic processes, we present two general tools for gene identification within the *C. tuberculosum* transcriptome dataset using KEGG based annotations. This involves extracting whole metabolic pathways or individual genes (see Appendix S1; Fig S5). We included annotations for all metabolic genes recovered in the *C. tuberculosum* transcriptome (Table S3). We identified 36 putative *C. tuberculosum* genes present in the starch and sucrose metabolism pathway (Fig S5; Table S4). In addition, we individually searched for genes potentially involved in calcification [41,47,48] and identified 13 sequence candidates related to calcium transport, six related to inorganic carbon transport, five related to pH homeostasis, 19 putative carbonic anhydrases, and 12 putative HSP90 genes (Table S5).

Discussion

Evidence for convergent evolution of monolignol biosynthesis

Using sequence similarity methods with genes from the monolignol pathway in land plants, we identified candidates for five genes related to monolignol biosynthesis (4CL, CCR, CAD, CCoAOMT, and CSE) from the newly generated *C. tuberculosum* transcriptomic dataset. These gene candidates are supported by genomic evidence, retain major motifs from their respective gene family, and return their *A. thaliana* counterpart in reciprocal BLAST analyses, suggesting that these enzymes may function similarly in monolignol biosynthesis in *C. tuberculosum*.

Despite supporting evidence from sequence similarity analyses, functional predictions for candidate sequences in the monolignol pathway within *C. tuberculosum* are obscured by the gene tree analysis. If the monolignol pathway in embryophytes and *C. tuberculosum* evolved in a common ancestor and was retained through conserved evolution, we would expect their sequences to form functional clades uninterrupted by functionally divergent protein sequences. However, with the exception of the CCoAOMT candidate, our gene tree analyses consistently showed that monolignol biosynthetic genes in land plants are not sister to those in *C. tuberculosum*. *C. tuberculosum* sequences were found within each respective overall protein family, but consistently grouped with land plant genes of non-monolignol forming function. If these *C. tuberculosum* sequences are functionally homologous to the monolignol biosynthesis counterpart in land plants, then they likely arose independently in *C. tuberculosum*. Convergent evolution in protein function, with phylogenetic patterns of protein sequences with similar functions intersected by sequences with dissimilar functions, is not uncommon in cell wall synthesizing enzymes [49]. Biosynthetic enzymes in *C. tuberculosum* could have evolved similar substrate specificity after the divergence of red algae and land plants or, alternatively, may reflect genes that were individually acquired. Previous evidence suggests that the core monolignol biosynthesis genes (4CL, CCR, and CAD) in *C. tuberculosum* may have been acquired through horizontal gene transfer from a bacterial source [36]. Thus, over evolutionary time genes in *C. tuberculosum* may have developed enough synchronicity in gene expression and protein regulation to produce an ad hoc monolignol biosynthetic pathway.

Alternatively, the phylogenetic evidence might suggest that gene candidates in *C. tuberculosum* do not function in monolignol biosynthesis and instead have a function similar to their sister sequences within their distinct phylogenetic groupings. For example, considering only clustering patterns in the phylogenetic data, perhaps *C. tuberculosum* contig 141618 functions as a CoA ligase that acts on malonate and not coumarate (4CL enzyme) (Fig 2B). However, the tandem use of stricter curated sequences in our predictive HMM models and more flexible HMM models with previously annotated data, such as KEGG annotations, improves our confidence in finding potential gene candidates. Biochemical or functional assays will ultimately be needed to verify the function of candidate gene sequences.

The monolignol biosynthesis pathway and missing steps in Calliarthron tuberculosum

Several key steps in the monolignol biosynthetic pathway were not recovered in the *C. tuberculosum* transcriptome, including PAL, TAL, PTAL, HCT, COMT, C3H, C4H, or F5H. Although we cannot dismiss that these observations may be due to fragmented sequences in the assembled genome and transcriptome data, we present several other possibilities.

The ammonia-lyase PAL, TAL, or PTAL creates the first substrates in the monolignol biosynthetic pathway [50–52]. Although no full-length homologs were identified in the *C. tuberculosum* transcriptome, short sequence candidates identified may represent a fragmented gene. However, these short sequences lacked genomic support, indicating they may be contaminants of non-*Calliarthron* origin. For this reason, PAL, TAL, and PTAL are currently indicated as absent (Fig 1). If these are indeed from *C. tuberculosum*, RACE amplification could help determine if the short ammonia-lyase we identified has a longer transcript. *C. tuberculosum* likely has an ammonia-lyase acting on phenylalanine or tyrosine since PAL and TAL are also key enzymes in producing flavanoids and coumarins, which have been previously detected in both fleshy and coralline red algae [53]. Further validation will be required to elucidate their presence.

C3H, C4H, or F5H are p450 monooxygenases responsible for converting substrates across the monolignol pathway eventually resulting in H to S to G type monolignols, respectively (Fig 1). P450 sequence candidates have been identified, but their substrate-specific identity as C3H, C4H, or F5H homologs is unclear. The cytochrome P450 sequence candidates from the *C. tuberculosum* transcriptome form two divergent groups. One group is likely involved in carotenoid biosynthesis, positioned within the CYP97 clade, while the other group forms their own clade of unknown function (Fig S2B). The identified candidates from *C. tuberculosum* may have multi-substrate specificities, acting on various substrates, including monolignol intermediate products. Some substrate promiscuity has previously been observed within members of the cytochrome P450 enzyme family [54,55]. Alternatively, each of the identified P450 clades in *C. tuberculosum* could contain a new class of cytochrome P450 capable of functioning in H-, G-, or S- unit monolignol biosynthesis. This proposed convergent evolution of a distinct and independently-evolved cytochrome P450 involved in monolignol production has previously been documented in the clubmoss *Selaginella moellendorffii* (F5H) [34,35]. In any case, the presence of unique P450s represents an interesting avenue of exploration to elucidate substrate specificity and functionality in the monolignol pathway in *C. tuberculosum.*

HCT is one alternative route shifting monolignol synthesis from H- to G- to S- types using a temporary shikimate decoration (Fig 1) [56]. Its absence could suggest that *C. tuberculosum* does not utilize an HCT enzyme or create G lignin using this route. Another alternative route in G- and S- type monolignol synthesis utilizes a CSE enzyme that acts on caffeoyl shikimate, an HCT downstream product (Fig 1). The absence of an HCT is at odds with the CSE enzyme identified in this study (Fig 1), suggesting that the CSE candidate identified may not be utilized in the monolignol biosynthetic pathway for *C. tuberculosum*. Though this absence could be due to fragmentation in the transcriptome, more data are required for further validation.

COMT is necessary for S type monolignol production in angiosperms [57–59]. The absence of this enzyme raises questions about how *C. tuberculosum* can produce sinapyl alcohol, a precursor component for S monolignols. Some evidence exists for a bifunctional enzyme in pine that can function as both

COMT and CCoAOMT (named AEOMT) in heterologous systems [60]. However, only moderate-to-low sequence similarity is shared among CCoAOMT, COMT, and the bifunctional AEOMT. Perhaps a similar protein with broad substrate specificity is present in *C. tuberculosum* but has yet to be identified based on sequence similarity.

Conclusion

In summary, we have identified several gene candidates in the *C. tuberculosum* transcriptome that represent central components in the monolignol biosynthetic pathway, helping to explain the surprising presence of lignins in this coralline red alga. Despite the complexity of monolignol biosynthesis, and contrary to the predictions outlined in Martone et al. [23], our gene trees do not demonstrate a deeply conserved evolution of monolignol biosynthesis, but instead suggest that each of the enzymes identified in *C. tuberculosum* likely evolved independently from those found in land plants. Interestingly, there remain several key enzymes in the monolignol pathway whose sequences have not been identified, including those related to pathway entry and to shifting the types of monolignols produced that would form H-, G-, and S-lignins within the cell wall. Further biochemical evidence and validation of sequence expression will be necessary to provide functional support for both the genes identified and to elucidate potential alternative routes in the monolignol biosynthetic pathway in *C. tuberculosum.* By providing methods to easily identify additional gene candidates from the *C. tuberculosum* transcriptome, we aim to facilitate future research on this fascinating organism.

Methods

Data and code availability

All sequencing data generated from this study are available at European Nucleotide Archive (transcriptome data: accession PRJEB39919; genome data: accession PRJEB39919). Genome supported transcripts, transcriptome assemblies, annotations, and an example of metabolic pathway extraction are available on Github (https://github.com/martonelab/geneAnnotCalliarthronTranscriptome/).

Experimental model and subject details

Specimen collection and sequencing

Two male, haploid specimens of *Calliarthron tuberculosum* were collected October 6, 2013, from Bluestone Point (48.81952, -125.1640), Bamfield, British Columbia, Canada and verified as haploid male specimens by microscopy. A portion of each collected sample was pressed and deposited into the UBC herbarium with voucher codes A89970 and A89985. Voucher codes can be queried at https://herbweb.botany.ubc.ca/herbarium/search.php?Database=algae for more information. Calcified intergenicula and non-calcified genicula from each individual were divided into two portions for data collection: either whole tissue (Sample I+G/PTM1 in the dataset) or calcified tissue only (Sample I/PTM2 in the dataset). Total RNA was extracted using the Spectrum Plant Total RNA kit (Cat # STRN50, Sigma-Aldrich) and sequenced on the Illumina HiSeq 2000 platform (paired-end 2x100bp, insert size ~220bp).

Abbreviation of enzyme names

CAD, (hydroxy)cinnamyl alcohol dehydrogenase; SAD, sinapyl alcohol dehydrogenase; CCoAOMT, caffeoyl-CoA O-methyl transferase; CCR, (hydroxy)cinnamoyl-CoA reductase; C3’H, p-coumaroyl shikimate 3’-hydroxylase; C4H, cinnamate 4 hydroxylase; 4CL, 4-hydroxycinnamoyl-CoA ligase; COMT, caffeic acid O-methyltransferase; F5H, ferulic acid ⁄ coniferaldehyde ⁄ coniferyl alcohol 5-

hydroxylase; HCT, hydroxycinnamoyl-CoA:shikimate hydroxycinnamoyl transferase; PAL, phenylalanine ammonia-lyase

Transcriptome assembly and annotation

Illumina sequence reads were assembled using Trinity with the *de novo* mode at default setting [72], independently for each anatomical sample (I+G/PTM1 ; I/PTM2 in the ENA database). A reference transcriptome was also assembled *de novo* using Trinity by independently combining the sequence reads generated from both samples. The assembled transcripts were annotated using Blast2GO [73]. Briefly, each transcript was searched against the NCBI RefSeq protein database (BLASTX, $E \leq 10^{-5}$), and its putative function was inferred based on the top protein hit and Gene Ontology (GO) terms. These proteins were then mapped onto the corresponding metabolic pathways in the Kyoto Encyclopaedia of Gene and Genomes (KEGG) database [74]. Identification of genes present in KEGG annotated pathways were extracted using the pathview package [75].

Filtering contaminant sequences in genome assembled data

To identify putative contaminant sequences in the genome assembly, each genome scaffold was searched (BLASTN) against a database of archaeal, bacterial and viral genome sequences retrieved from the NCBI RefSeq database. Sequences with a significant hit ($E \leq 10^{-5}$, covering > 50% of the query length) were considered putative contaminants and removed from the genome assembly. To identify broad differences in sequence characteristics, genomic scaffolds with and without transcriptomic support were compared for G+C content and transcript length (Fig S1B). Scaffolds with no transcript support and low recovery of eukaryotic genes (< 6% BUSCO or CEGMA recovery) were also identified as likely putative contaminants and removed from the genome assembly.

Genome annotation guided by transcriptome evidence

Repetitive elements in the genome assembly were identified and masked using RepeatMasker version open-4.0.6 [76]. To maximize recovery of transcript support for genome scaffolds, the transcriptomes (I+G/PTM1; I/PTM2 in the dataset) were mapped against the masked genome scaffolds using PASA v2.0.2 [77], and full-length coding sequences (CDSs) were predicted with TransDecoder v5.0.1 [72]. These CDSs represent the primary set of putative genes and were used as extrinsic hints to guide *ab initio* gene prediction using AUGUSTUS v3.2.1 [78] from the genome scaffolds.

HMM based gene candidate search

Monolignol biosynthesis gene candidates were identified from the *C. tuberculosum* transcriptomic dataset using Hidden Markov Model (HMM) based searches [79]. Transcriptomic sequence contigs were translated into all six reading frames using EMBOSS Transeq [80]. This amino acid database was used for subsequent sequence searches. HMM profiles used to search for homologs in the transcriptome were produced by aligning amino acid sequences of a given protein or protein family using MUSCLE [81] with no manual adjustment. The profiles were searched against the translated *C. tuberculosum* dataset in HMMER searches [79] to look for putative sequence homologs. Sequences more than 100 amino acids long were retained for subsequent analysis. These sequences were then searched against the *Arabidopsis* (GenBank taxid:3701) proteome using NCBI's BLAST [82] to verify their closest homolog match (BLASTP, $E \leq 10^{-30}$).

Domain and motif comparison

The monolignol biosynthetic genes and their overall gene families contain sequence domains that influence protein shape and function. To compare these key domains, multiple sequence alignments

(MSA) of candidate amino acid sequences from *C. tuberculosum* with their land plant counterpart protein were produced. Sequences were aligned using MUSCLE under default settings [81]. Key domains and motifs were chosen based on available literature and highlighted in the MSA as indicated in each figure legend. In each MSA, an asterisk (*) represents full conservation; and a period (.) represents sites with conservation >50%. Accession numbers can be found in Appendix S1.

Gene tree analysis

Gene trees were reconstructed for the candidate sequences of *C. tuberculosum* identified. For each gene tree analysis, sequence candidates from *C. tuberculosum*, the functionally demonstrated enzyme sequence from land plants, enzyme sequences from the overall protein family from land plants, and the top 20 sequences identified by NCBI BLAST using *C. tuberculosum* candidates as a query against the total database using default settings (BLASTP, $E \leq 10^{-20}$) were compiled. Land plant sequences identified to represent the functional gene and overall gene family were curated by a literature search. For each set of sequences, a multiple sequence alignment was performed using MUSCLE with default setting [81]. Sites with <80% coverage were removed using trimAl [83]. IQTree was used to search for the evolutionary model alignment under a BIC criterion [84,85]. A maximum likelihood tree was reconstructed using IQTree [86], with node support calculated based on 1000 ultrafast bootstrap pseudoreplicates in IQTree [86]. A clade is considered strongly supported when bootstrap value ≥ 95%. FigTree was used to edit branch width and colors [87]. Accession numbers can be found in Appendix S1.

Generation of genome data as additional support for transcriptome data

Genome data of *C. tuberculosum* were generated using Illumina IIx platform (paired-end 2×150bp reads, insert size ~350 bp). An overview of the summary statistics for the genome assembly can be found in Table S1. Adapter sequences were removed using Trimmomatic v0.33 [88] (LEADING:25

TRAILING:25 HEADCROP:10 SLIDINGWINDOW:4:20 MINLEN:50). The generated filtered sequence reads and the previously published genome data (GenBank accession #: SRP005182) generated using the 454 pyrosequencing platform [43] were used in a *de novo* genome assembly using SPAdes [89]. The 454 reads were treated as unpaired, single-end reads in the assembly process. This *de novo* assembly was further scaffolded with the transcriptome data using the L_RNA_Scaffolder [90]. Putative contaminant sequences were removed based on shared similarity against known genome sequences from bacterial, archaeal, and viral sources in NCBI RefSeq (BLASTN, $E \leq 10^{-5}$), and subsequently based on discrepancy in G+C content of the assembled scaffolds, and the recovery of core eukaryotic genes (CEGMA and BUSCO). Because the genome assembly is fragmented, genome scaffolds on which no transcripts were mapped were filtered out, yielding the final genome assembly (21,672 scaffolds, total bases 64.15 Mbp). These genome scaffolds were used as additional support for the transcriptome data. For the reference transcriptome (combined I+G/PTM1 ; I/PTM2), putative coding sequences were predicted based on alignment of the assembled transcripts against the genome scaffolds using PASA [77] and TransDecoder [72], from which the coded protein sequences were predicted.

Completeness of transcriptome and genome data

The completeness of the genome and transcriptome data was assessed by the recovery of core conserved eukaryote genes with the Core Eukaryotic Genes Mapping Approach (CEGMA) [91] and Benchmarking Universal Single-Copy Orthologs (BUSCO) [92] datasets. CEGMA and BUSCO datasets (eukaryote odb9 and Viridiplantae odb10) were independently used as query to search against the predicted proteins from the reference transcriptome (combined IG and IO) using BLASTP ($E \leq 10^{-5}$) and against the same transcriptome using TBLASTN ($E \leq 10^{-5}$). The core CEGMA and BUSCO proteins were also queried against the 21,672 genome scaffolds using TBLASTN ($E \leq 10^{-5}$).

484 **Key Resources Table**

485

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological Samples		
Calliarthron tuberculosum (sample vouchers)	This paper	A89970 and A89985 at https://herbweb.botany.ubc.ca/herbarium/search.php?Database=algae
Critical Commercial Assays		
HiSeq 2000 (Transcript reads)	Illumina	
IIx platform (Genomic reads)	Illumina	
Spectrum Plant Total RNA kit	Sigma-Aldrich	STRN50
Deposited Data		
Raw sequencing reads for transcriptomic and genomic data	This paper	PRJEB39919
Genome supported transcripts, transcriptome assemblies, annotations	This paper	https://github.com/martonelab/geneAnnotCalliarthronTranscriptome/
Additional *Calliarthron* Genomic Reads	[43]	SRP005182
Pyropia genomic data	[39]	MXAK00000000
Arabidopsis Proteome		taxid:3701

Software and Algorithms		
TransDecoder v5.0.1	[72]	https://github.com/TransDecoder/TransDecoder/wiki
Trinity	[72]	https://github.com/trinityrnaseq/trinityrnaseq/wiki
Blast2GO	[73]	https://www.blast2go.com/
Kyoto Encyclopedia of Genes and Genomes (KEGG)	[74]	https://www.genome.jp/kegg/
Pathview R Package	[75]	https://www.bioconductor.org/packages/release/bioc/html/pathview.html
HMMER	[79]	http://hmmer.org/
EMBOSS Transeq	[80]	http://emboss.sourceforge.net/apps/release/6.6/emboss/apps/transeq.html
MUSCLE v3.5	[81]	http://www.drive5.com/muscle/muscle.html
IQtree	[85,86]	http://www.iqtree.org/

TrimAl	[83]	http://trimal.cgenomics.org/
FigTree	[87]	http://tree.bio.ed.ac.uk/software/figtree/
Trimmomatic v0.33	[88]	http://www.usadellab.org/cms/?page=trimmomatic
SPAdes	[89]	https://cab.spbu.ru/software/spades/
L_RNA_Scaffolder	[90]	https://github.com/CAFS-bioinformatics/L_RNA_scaffolder
PASA v2.0.2	[77]	https://github.com/PASApipeline/PASApipeline
CEGMA	[91]	http://korflab.ucdavis.edu/datasets/cegma/
BUSCO	[92]	https://busco.ezlab.org/
RepeatMasker version open-4.0.6	[76]	http://www.repeatmasker.org/
AUGUSTUS v3.2.1	[78]	https://github.com/nextgenusfs/augustus

Acknowledgements

We thank Dana Price and Debashish Bhattacharya (Rutgers University) for sequencing and preliminary analysis of the genome data, and Mike Thang (QFAB, Australia) for support in submitting sequence data to ENA. We thank the Osprey Ranch for supporting our writing retreats.

Funding

C.X.C. was supported by Australian Research Council grants (DP150101875 and DP190102474). P.M. was supported by Natural Sciences and Engineering Research Council (NSERC) Discovery Grants (RGPIN 356403-09; 2014-06288; 2019-06240). K.H. and M.L. were supported by the Hakai Institute. J.X. was supported by the UBC Summer Undergraduate Research award, NSERC Graduate Student fellowship and Patrick David Campbell Graduate Student fellowship. K.H. was supported by a postdoctoral scholarship from the Tula Foundation.

Author Contributions

Conceptualization, J.X., K.H, and P.T.M; Methodology, J.X., K.H., M.A.L., C.X.C.; Investigation, A.M., J.X., K.H., M.A.L., C.X.C.; Visualization, J.X., E.J.; Writing - Original Draft, J.X. and P.T.M.; Review and Editing, J.X., K.H., M.A.L., E.J., C.X.C., P.T.M.; Funding Acquisition, P.T.M.; Supervision, P.T.M.

Declaration of Interests

The authors declare no competing interests.

References

1. Adey WH. The algal ridges and coral reefs of St. Croix: their structure and Holocene development. Atoll Research Bulletin. 1975; 1–67. doi:https://doi.org/10.5479/si.00775630.187.1
2. Borowitzka MA. Algal calcification. Oceanography and Marine Biology Annual Review. 1977; 189–223.
3. Goreau TF. Calcium carbonate deposition by coralline algae and corals in relation to their roles as reef-builders. Annals of the New York Academy of Sciences. 1963;109: 127–167.
4. Harrington L, Fabricius K, De'ath G, Negri A. Recognition and selection of settlement substrata determine post-settlement survival in corals. Ecology. 2004;85: 3428–3437. doi:10.1890/04-0298
5. O'Leary JK, Barry JP, Gabrielson PW, Rogers-Bennett L, Potts DC, Palumbi SR, et al. Calcifying algae maintain settlement cues to larval abalone following algal exposure to extreme ocean acidification. Scientific reports. 2017;7: 5710–5774. doi:10.1038/s41598-017-05502-x
6. Swanson RL, de Nys R, Huggett MJ, Green JK, Steinberg PD. In situ quantification of a natural settlement cue and recruitment of the Australian sea urchin Holopneustes purpurascens. Marine ecology Progress series (Halstenbek). 2006;314: 1–14. doi:10.3354/meps314001
7. Fisher K, Martone PT. Field study of growth and calcification rates of three species of articulated coralline algae in British Columbia, Canada. Biological Bulletin. 2014;226: 121–130. doi:10.1086/BBLv226n2p121
8. van der Heijden LH, Kamenos NA. Reviews and syntheses: Calculating the global contribution of coralline algae to total carbon burial. Biogeosciences. 2015;12: 6429–6441. doi:10.5194/bg-12-6429-2015
9. Gabrielson PW, Hughey JR, Diaz-Pulido G. Genomics reveals abundant speciation in the coral reef building alga Porolithon onkodes (Corallinales, Rhodophyta). Journal of phycology. 2018;54: 429–434. doi:10.1111/jpy.12761
10. Hind KR, Miller KA, Young M, Jensen C, Gabrielson PW, Martone PT. Resolving cryptic species of Bossiella (Corallinales, Rhodophyta) using contemporary and historical DNA. American

journal of botany. 2015;102: 1912–1930. doi:10.3732/ajb.1500308

11. Hind KR, Gabrielson PW, Lindstrom SC, Martone PT. Misleading morphologies and the importance of sequencing type specimens for resolving coralline taxonomy (Corallinales, Rhodophyta): Pachyarthron cretaceum is Corallina officinalis. Journal of Phycology. 2014;50: 760–764. doi:10.1111/jpy.12205

12. Twist BA, Neill KF, Bilewitch J, Jeong SY, Sutherland JE, Nelson WA. High diversity of coralline algae in New Zealand revealed: Knowledge gaps and implications for future research. PloS one. 2019;14: e0225645. doi:10.1371/journal.pone.0225645

13. Bergstrom E, Ordoñez A, Ho M, Hurd C, Fry B, Diaz-Pulido G. Inorganic carbon uptake strategies in coralline algae: Plasticity across evolutionary lineages under ocean acidification and warming. Marine environmental research. 2020;161: 105–107. doi:10.1016/j.marenvres.2020.105107

14. Cornwall CE, Comeau S, McCulloch MT. Coralline algae elevate pH at the site of calcification under ocean acidification. Global change biology. 2017;23: 4245–4256. doi:10.1111/gcb.13673

15. Guenther R. The effect of temperature and pH on the growth and biomechanics of coralline algae. University of British Columbia. 2016.

16. McCoy SJ, Ragazzola F. Skeletal trade-offs in coralline algae in response to ocean acidification. Nature climate change. 2014;4: 719–723. doi:10.1038/nclimate2273

17. Noisette F, Egilsdottir H, Davoult D, Martin S. Physiological responses of three temperate coralline algae from contrasting habitats to near-future ocean acidification. Journal of experimental marine biology and ecology. 2013;448: 179–187. doi:10.1016/j.jembe.2013.07.006

18. Hind KR, Gabrielson PW, Jensen C, Martone PT. Evolutionary reversals in Bossiella (Corallinales, Rhodophyta): first report of a coralline genus with both geniculate and nongeniculate species. Journal of phycology. 2018;54: 788–798. doi:10.1111/jpy.12788

19. Janot K, Martone PT. Convergence of joint mechanics in independently evolving, articulated coralline algae. Journal of experimental biology. 2016;219: 383–391. doi:10.1242/jeb.131755

20. Steneck RS. The ecology of coralline algal crusts: convergent patterns and adaptive strategies. Ann Rev Ecol Syst. 1986;17: 273–303.

21. Aguirre J, Perfectti F, Braga JC. Integrating phylogeny , molecular clocks , and the fossil record in the evolution of coralline algae (Corallinales and Sporolithales , Rhodophyta) Author (s): Julio Aguirre , Francisco Perfectti and Juan C . Braga Published by : Cambridge University P. Paleobiology. 2010;36: 519–533.

22. Rösler A, Perfectti F, Peña V, Aguirre J, Braga JC, Gabrielson P. Timing of the evolutionary history of Corallinaceae (Corallinales, Rhodophyta). Journal of Phycology. 2017;53: 567–576. doi:10.1111/jpy.12520

23. Martone PT, Estevez JM, Lu F, Ruel K, Denny MW, Somerville C, et al. Discovery of Lignin in Seaweed Reveals Convergent Evolution of Cell-Wall Architecture. Current Biology. 2009;19: 169–175. doi:10.1016/j.cub.2008.12.031

24. Boerjan W, Ralph J, Baucher M. Lignin Biosynthesis. Annual Review of Plant Biology. 2003;54: 519–546. doi:10.1146/annurev.arplant.54.031902.134938

25. Mottiar Y, Vanholme R, Boerjan W, Ralph J, Mansfield SD. Designer lignins: Harnessing the plasticity of lignification. Current Opinion in Biotechnology. 2016;37: 190–200. doi:10.1016/j.copbio.2015.10.009

26. Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. Lignin biosynthesis and structure. Plant Physiology. 2010;153: 895–905. doi:10.1104/pp.110.155119

27. Lange BM, Lapierre C, Sandermann H. Elicitor-induced spruce stress lignin: Structural similarity to early developmental lignins. Plant Physiology. 1995;108: 1277–1287. doi:10.1104/pp.108.3.1277

28. Tronchet M, BalaguÉ C, Kroj T, Jouanin L, Roby D. Cinnamyl alcohol dehydrogenases-C and D, key enzymes in lignin biosynthesis, play an essential role in disease resistance in Arabidopsis. Molecular Plant Pathology. 2010;11: 83–92. doi:10.1111/j.1364-3703.2009.00578.x

29. Martone PT. Kelp versus coralline: Cellular basis for mechanical strength in the wave-swept

seaweed Calliarthron (Corallinaceae, Rhodophyta). Journal of Phycology. 2007;43: 882–891. doi:10.1111/j.1529-8817.2007.00397.x

30. Denny MW, King FA. The extraordinary joint material of an articulated coralline alga. II. Modeling the structural basis of its mechanical properties. Journal of Experimental Biology. 2016;219: 1843–1850. doi:10.1242/jeb.138867

31. Weng JK, Chapple C. The origin and evolution of lignin biosynthesis. New Phytologist. 2010;187: 273–285. doi:10.1111/j.1469-8137.2010.03327.x

32. Dixon RA, Barros J. Lignin biosynthesis: Old roads revisited and new roads explored. Open Biology. 2019;9. doi:10.1098/rsob.190215

33. Raes, J., Rohde, A., Christensen, J. H., Van de Peer, Y., Boerjan W. Genome-Wide Characterization of the Lignification Toolbox in Arabidopsis. Plant Physiology. 2014;133: 1051–1071. doi:10.1104/pp.103.026484.role

34. Weng JK, Akiyama T, Bonawitz ND, Li X, Ralph J, Chapple C. Convergent evolution of syringyl lignin biosynthesis via distinct pathways in the lycophyte Selaginella and flowering plants. Plant Cell. 2010;22: 1033–1045. doi:10.1105/tpc.109.073528

35. Weng J-K, Li X, Stout J, Chapple C. Independent origins of syringyl lignin in vascular plants. Proceedings of the National Academy of Sciences. 2008;105: 7887 LP – 7892. doi:10.1073/pnas.0801696105

36. Labeeuw L, Martone PT, Boucher Y, Case RJ. Ancient origin of the biosynthesis of lignin precursors. Biology Direct. 2015;10: 1–21. doi:10.1186/s13062-015-0052-y

37. Matsuzaki M, Misumi O, Shin-i T, Maruyama S, Takahara M, Miyagishima S, et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. Nature. 2004;428: 653–657. doi:10.1038/nature02398

38. Collén J, Porcel B, Carré W, Ball SG, Chaparro C, Tonon T, et al. Genome structure and metabolic features in the red seaweed Chondrus crispus shed light on evolution of the Archaeplastida. Proceedings of the National Academy of Sciences of the United States of

America. 2013;110: 5247–5252. doi:10.1073/pnas.1221259110

39. Brawley SH, Blouin NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson H V., et al. Insights into the red algae and eukaryotic evolution from the genome of Porphyra umbilicalis (Bangiophyceae, Rhodophyta). Proceedings of the National Academy of Sciences of the United States of America. 2017;114: E6361–E6370. doi:10.1073/pnas.1703088114

40. Lee JM, Yang EC, Graf L, Yang JH, Qiu H, Zelzion U, et al. Analysis of the draft genome of the red seaweed gracilariopsis chorda provides insights into genome size evolution in rhodophyta. Molecular Biology and Evolution. 2018;35: 1869–1886. doi:10.1093/molbev/msy081

41. Page TM, McDougall C, Diaz-Pulido G. De novo transcriptome assembly for four species of crustose coralline algae and analysis of unique orthologous genes. Scientific Reports. 2019;9. doi:10.1038/s41598-019-48283-1

42. Bi G, Liu G, Zhao E, Du Q. Complete mitochondrial genome of a red calcified alga Calliarthron tuberculosum (Corallinales). Mitochondrial DNA. 2016;27: 2554–2556. doi:10.3109/19401736.2015.1038801

43. Chan CX, Yang EC, Banerjee T, Yoon HS, Martone PT, Estevez JM, et al. Red and green algal monophyly and extensive gene sharing found in a rich repertoire of red algal genes. Current Biology. 2011;21: 328–333. doi:10.1016/j.cub.2011.01.037

44. Shockey JM, Fulda MS, Browse J. Arabidopsis Contains a Large Superfamily of Acyl-Activating Enzymes . Phylogenetic and Acyl-Coenzyme A Synthetases 1. Plant physiology. 2003;132: 1065–1076. doi:10.1104/pp.103.020552.ularly

45. Shockey J, Browse J. Genome-level and biochemical diversity of the acyl-activating enzyme superfamily in plants. Plant Journal. 2011;66: 143–160. doi:10.1111/j.1365-313X.2011.04512.x

46. Li L, Cheng XF, Leshkevich J, Umezawa T, Harding SA, Chiang VL. The Last Step of Syringyl Monolignol Biosynthesis in Angiosperms Is Regulated by a Novel Gene Encoding Sinapyl Alcohol Dehydrogenase. The Plant Cell. 2001;13: 1567–1586. doi:10.1105/tpc.010111

47. Hofmann LC, Schoenrock K, de Beer D. Arctic Coralline Algae Elevate Surface pH and

Carbonate in the Dark. Frontiers in plant science. 2018;9: 1416. doi:10.3389/fpls.2018.01416

48. Nam O, Shiraiwa Y, Jin E. Calcium-related genes associated with intracellular calcification of Emiliania huxleyi (Haptophyta) CCMP 371. ALGAE. 2018;33: 181–189. doi:10.4490/algae.2018.33.4.21

49. Xue J, Purushotham P, Acheson JF, Ho R, Zimmer J, McFarlane C, et al. Functional characterization of a cellulose synthase, CtCESA1, from the marine red alga Calliarthron tuberculosum (Corallinales). Journal of Experimental Botany. 2021; erab414. doi:10.1093/jxb/erab414

50. Kyndt JA, Meyer TE, Cusanovich MA, Van Beeumen JJ. Characterization of a bacterial tyrosine ammonia lyase, a biosynthetic enzyme for the photoactive yellow protein. FEBS letters. 2002;512: 240–244. doi:10.1016/S0014-5793(02)02272-X

51. Barros J, Serrani-Yarce JC, Chen F, Baxter D, Venables BJ, Dixon RA. Role of bifunctional ammonia-lyase in grass cell wall biosynthesis. Nature plants. 2016;2: 16050. doi:10.1038/nplants.2016.50

52. Cooke HA, Christianson C V, Bruner SD. Structure and chemistry of 4-methylideneimidazole-5-one containing enzymes. Current opinion in chemical biology. 2009;13: 460–468. doi:10.1016/j.cbpa.2009.06.013

53. Mohy El-Din SM, El-Ahwany AMD. Bioactivity and phytochemical constituents of marine red seaweeds (Jania rubens, Corallina mediterranea and Pterocladia capillacea). Journal of Taibah University for Science. 2016;10: 471–484. doi:https://doi.org/10.1016/j.jtusci.2015.06.004

54. Mallinson SJB, Machovina MM, Silveira RL, Garcia-Borràs M, Gallup N, Johnson CW, et al. A promiscuous cytochrome P450 aromatic O-demethylase for lignin bioconversion. Nature communications. 2018;9: 2412–2487. doi:10.1038/s41467-018-04878-2

55. Guo J, Ma X, Cai Y, Ma Y, Zhan Z, Zhou YJ, et al. Cytochrome P450 promiscuity leads to a bifurcating biosynthetic pathway for tanshinones. The New phytologist. 2016;210: 525–534. doi:10.1111/nph.13790

56. Hoffmann L, Besseau S, Geoffroy P, Ritzenthaler C, Meyer D, Lapierre C, et al. Silencing of Hydroxycinnamoyl-Coenzyme A Shikimate/Quinate Hydroxycinnamoyltransferase Affects Phenylpropanoid Biosynthesis. The Plant cell. 2004;16: 1446–1465. doi:10.1105/tpc.020297

57. Goujon T, Sibout R, Pollet B, Maba B, Nussaume L, Bechtold N, et al. A new Arabidopsis thaliana mutant deficient in the expression of O-methyltransferase impacts lignins and sinapoyl esters. Plant Molecular Biology. 2003;51: 973–989. doi:10.1023/A:1023022825098

58. Lu F, Marita JM, Lapierre C, Jouanin L, Morreel K, Boerjan W, et al. Sequencing around 5-Hydroxyconiferyl Alcohol-Derived Units in Caffeic Acid O -Methyltransferase-Deficient Poplar Lignins. Plant physiology (Bethesda). 2010;153: 569–579. doi:10.1104/pp.110.154278

59. Guo D, Chen F, Inoue K, Blount JW, Dixon RA. Downregulation of Caffeic Acid 3- O -Methyltransferase and Caffeoyl CoA 3- O -Methyltransferase in Transgenic Alfalfa: Impacts on Lignin Structure and Implications for the Biosynthesis of G and S Lignin. The Plant cell. 2001;13: 73–88. doi:10.1105/tpc.13.1.73

60. Li L, Popko JL, Zhang X-H, Osakabe K, Tsai C-J, Joshi CP, et al. A Novel Multifunctional O-Methyltransferase Implicated in a Dual Methylation Pathway Associated with Lignin Biosynthesis in Loblolly Pine. Proceedings of the National Academy of Sciences - PNAS. 1997;94: 5461–5466. doi:10.1073/pnas.94.10.5461

61. Hu Y, Gai Y, Yin L, Wang X, Feng C, Feng L, et al. Crystal Structures of a Populus tomentosa 4-Coumarate : CoA Ligase Shed Light on Its Enzymatic Mechanisms. Plant physiology. 2010;22: 3093–3104. doi:10.1105/tpc.109.072652

62. Witzel K, Schomburg D, Kombrink E, Schneider K, Ho K, Stuible H. The substrate specificity-determining amino acid code of 4-coumarate : CoA ligase. PNAS. 2003;100: 8601–8606.

63. Stuible H, Kombrink E. Identification of the Substrate Specificity-conferring Amino Acid Residues of 4-Coumarate : Coenzyme A Ligase Allows the Rational Design of Mutant Enzymes with New Catalytic Properties. Journal of Biological Chemistry. 2001;276: 26893–26897. doi:10.1074/jbc.M100355200

64. Jörnvall H, Persson B, Krook M, Atrian S, Gonzàlez-Duarte R, Jeffery J, et al. Short-Chain Dehydrogenases/Reductases (SDR). Biochemistry. 1995;34: 6003–6013. doi:10.1021/bi00018a001

65. Sattler SA, Walker AM, Vermerris W, Sattler SE, Kang C. Structural and Biochemical Characterization of Cinnamoyl-CoA Reductases. Plant physiology. 2017;173: 1031–1044. doi:10.1104/pp.16.01671

66. Filling C, Berndt KD, Benach J, Knapp S, Prozorovski T, Nordling E, et al. Critical Residues for Structure and Catalysis in Short-chain Dehydrogenases / Reductases. Biological Chemistry. 2002;277: 25677–25684. doi:10.1074/jbc.M202160200

67. Bukh C, Nord-Larsen PH, Rasmussen SK. Phylogeny and structure of the cinnamyl alcohol dehydrogenase gene family in Brachypodium distachyon. Journal of Experimental Botany. 2012;63: 6223–6236. doi:10.1093/jxb/ers275

68. Youn B, Camacho R, Moinuddin SGA, Lee C, Davin LB, Lewis NG, et al. Crystal structures and catalytic mechanism of the Arabidopsis cinnamyl alcohol dehydrogenases AtCAD5 and AtCAD4. Organic & Biomolecular Chemistry. 2006;4: 1687–1697. doi:10.1039/B601672C

69. Bomati EK, Noel JP. Structural and kinetic basis for substrate selectivity in Populus tremuloides sinapyl alcohol dehydrogenase. The Plant cell. 2005/04/13. 2005;17: 1598–1611. doi:10.1105/tpc.104.029983

70. Julián-sánchez A, Riveros-rosas H, Piña E. Evolution of Cinnamyl Alcohol Dehydrogenase Family Evolution of Cinnamyl Alcohol Dehydrogenase Family. In: Weiner H, Plapp B, Lindahl R, Maser E, editors. Enzymology and Molecular Biology of Carbonyl Metabolism. West Lafayette: Purdue University Press; 2006. pp. 142–153.

71. von Borzyskowski LS, Rosenthal RG, Erb TJ. Evolutionary history and biotechnological future of carboxylases. Journal of Biotechnology. 2013;168: 243–251. doi:10.1016/j.jbiotec.2013.05.007

72. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Philip D, Bowden J, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nature

protocols. 2013;8: 1–43. doi:10.1038/nprot.2013.084.De

73. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005;21: 3674–3676. doi:10.1093/bioinformatics/bti610

74. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic acids research. 2016;44: 457–462. doi:10.1093/nar/gkv1070

75. Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. Computer applications in the biosciences. 2013;29: 1830–1831. doi:10.1093/bioinformatics/btt285

76. Smit A, Hubley R, Green P. RepeatMasker Open-4.0.

77. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research. 2003;31: 5654–5666. doi:10.1093/nar/gkg770

78. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;7: 1–11. doi:10.1186/1471-2105-7-62

79. Finn RD, Clements J, Eddy SR. HMMER Web Server: Interactive Sequence Similarity Searching. Nucleic Acids Research. 2011;39: W29–W37. doi:10.1093/nar/gkr367

80. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

81. Edgar RC. MUSCLE: a Multiple Sequence Alignment Method With Reduced Time and Space Complexity. BMC Bioinformatics. 2004;5. doi:10.1186/1471-2105-5-113

82. Mahram A, Herbordt MC. Fast and Accurate NCBI BLASTP: Acceleration with Multiphase FPGA-based Prefiltering. Proceedings of the 24th ACM International Conference on Supercomputing. New York, NY, USA: ACM; 2010. pp. 73–82. doi:10.1145/1810085.1810099

83. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment

trimming in large-scale phylogenetic analyses. Bioinformatics. 2009/06/08. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348

84. Luo A, Qiao H, Zhang Y, Shi W, Ho SY, Xu W, et al. Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. BMC evolutionary biology. 2010;10: 242. doi:10.1186/1471-2148-10-242

85. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution. 2015;32: 268–274.

86. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018;35: 518–522.

87. Rambaut A, Drummond A. FigTree v1. 3.1 Institute of Evolutionary Biology. University of Edinburgh. 2010.

88. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

89. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

90. Xue W, Li JT, Zhu YP, Hou GY, Kong XF, Kuang YY, et al. L_RNA_scaffolder: Scaffolding genomes with transcripts. BMC Genomics. 2013;14: 1–14. doi:10.1186/1471-2164-14-604

91. Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007;23: 1061–1067. doi:10.1093/bioinformatics/btm071

92. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351

Supporting Information

Fig S1. Completeness of the *C. tuberculosum* transcriptome dataset.

(A) Transcriptome sequences show high recovery of eukaryotic genes in CEGMA/BUSCO analysis. Percentage of genomic scaffolds with transcriptome support and transcriptomic scaffolds alone that share amino acid sequences with the core eukaryotic gene databases including CEGMA, BUSCO eukaryotic, and BUSCO Viridiplantae. Transcriptome encoded amino acid sequences were searched against the databases using BLASTP (orange) or TBLASTN (yellow), and genomic scaffolds were searched against the databases using TBLASTN (blue)

(B) Transcriptomic support of genomic data analyzed by GC content and transcript length. The distribution of GC content (above) against transcript lengths is shown for scaffolds with transcriptome support (blue) and scaffolds without transcriptome support (yellow) (right).

Fig S2. C3H, C4H, F5H, P450 candidates from *C. tuberculosum* in relation to plants and other taxa.

(A) Partial alignment of *C. tuberculosum* P450 candidates with C3H, C4H, and F5H from *A. thaliana*, and a novel F5H from *Selaginella moellendorffii*. Heme binding domain residues, secondary structure stabilizing K helix residues, PXRX, and the I-helix are indicated [8]. Sites with <80% coverage were removed. A strong candidate for beta-carotene synthesis is indicated with a triangle.

(B) Unrooted CYP450 maximum likelihood gene tree with *C. tuberculosum* (magenta dots) and additional taxa (Embryophyta – dark green, Chlorophyta – light green, Rhodophyta – red, Animalia and Opisthokonta – purple, Bacteria and Cyanobacteria – blue, Oomycota, Mycetozoa and Fungi – yellow, Ochrophyta – brown). Functionally demonstrated plant C3H, C4H, and F5H are labeled (+). Additional functional groups are labeled [9]. Ultrafastbootstrap values > 95 are marked by *. Model = VT+F+G4.

Fig S3. CCoAOMT candidates from *C. tuberculosum* in relation to plants and other taxa.

(A) Partial alignment of *C. tuberculosum* CCoAOMT sequence candidates with CCoAOMT from land plants. Substrate recognition residues (black triangle), divalent metal ion and cofactor binding residues (grey triangle), catalytic residues (back square), and the positively charged R220 necessary for substrate recognition (grey square) are indicated. Sites with < 70% coverage were removed.

(B) Unrooted maximum likelihood gene tree of biochemically characterized plant O-methyltransferases with *C. tuberculosum* (magenta dots) and additional taxa (Embryophyta – dark green, Chlorophyta – light green, Rhodophyta – red, Animalia and Opisthokonta – purple, Bacteria and Cyanobacteria – blue, Oomycota, Mycetozoa and Fungi – yellow, Ochrophyta – brown). Functionally demonstrated plant

CCoAOMT are labeled (+). Additional functional groups are labeled [13]. Ultrafastbootstrap values > 95 are marked by *. Model = LG + G4. JMT, SAMT, and BAMT are closely related to OMTs.

Fig S4. CSE candidates from *C. tuberculosum* in relation to plants and other taxa.

(A) Partial alignment of *C. tuberculosum* CSE sequence candidates with CSE from land plants. Acyl transferase motifs (HX_4D), lipase motifs (GXSXG) and active site residues (triangle) are indicated. Sites with < 70% coverage were removed.

(B) Unrooted maximum likelihood gene tree of *C. tuberculosum* CSE candidates (magenta dots) and additional taxa (Embryophyta – dark green, Chlorophyta – light green, Rhodophyta – red, Animalia and Opisthokonta – purple, Bacteria and Cyanobacteria – blue, Oomycota, Mycetozoa and Fungi – yellow, Ochrophyta – brown). Functionally demonstrated plant CSE are labeled (+). Additional functional groups are labeled. Ultrafastbootstrap values > 95 are marked by *. Model = VT+G4.

Fig. S5. A visual representation of the *C. tuberculosum* sequences present in the starch and sucrose metabolism pathway from the KEGG based annotation.

KEGG based annotation showing the starch and sucrose metabolic pathway with *C. tuberculosum* annotations highlighted. The gradient map in the top right corner indicates the level of transcription, with white and dark pink coloring representing absence and presence of expression respectively. The annotated map, number "00500", was extracted in the provided R file using the pathview program.

Table S1. Summary statistics for the *C. tuberculosum* genome assembly.

Scaffolds are categorized as shared with either red algal (*Pyropia* yezoensis) genomic scaffolds, eukaryotic sequences, or other bacteria sequences based on sequence similarity.

Table S2. Top hits against Arabidopsis thaliana (taxid:3702) using *Calliarthron* sequences as the search query (BLASTP). Query sequence is indicated by contig number. Result hits are indicated by description (At tax ID 3702) and colored by overall alignment scores with red (>=200), pink (80-200), green (50-80), blue (40-50), and black (<40) that are most to least reliable scores in that order.

Table S3. KEGG annotations of *Calliarthron tuberculosum* reads from the combined transcriptomic dataset. Unique reads are represented by their contig identifier (contig_gene_isoform) and matched with their annotated KEGG based identifier (KO_identifier) and associated protein name.

Table S4. Listed representation of the *C. tuberculosum* sequences present in starch and sucrose metabolism pathway from the KEGG based annotation.

C. tuberculosum sequences were extracted from the KEGG based starch and sucrose metabolism pathway number "00500". "KEGG Identifier" refers to the specific KEGG code for the gene, "Contig Name" refers to the sequence identifier from the *Calliarthron* transcriptome where the values represent the contig name_gene number_gene isoform and "Gene Name" refers to the gene acronym, the gene name, and its enzyme commission (EC) number. Sequences were extracted in the provided R file using the pathview program.

Table S5. A list of calcification related gene candidates identified from KEGG-based annotations of the *C. tuberculosum* transcriptome.

Calcification gene candidates were initially selected based on a literature search, and then *C. tuberculosum* sequences were identified manually from the KEGG based annotations (annotation file available on Github), thus this is not an exhaustive list. The genes are organized by their functional classification indicated as "overall function", while "KEGG Identifier" refers to the specific KEGG code for the gene, "Contig Name" refers to the sequence identifier from the *Calliathron* transcriptome where the values represent the contig name_gene number_gene isoform and "Gene Name" refers to the gene acronym, the gene name, and its enzyme commission (EC) number.

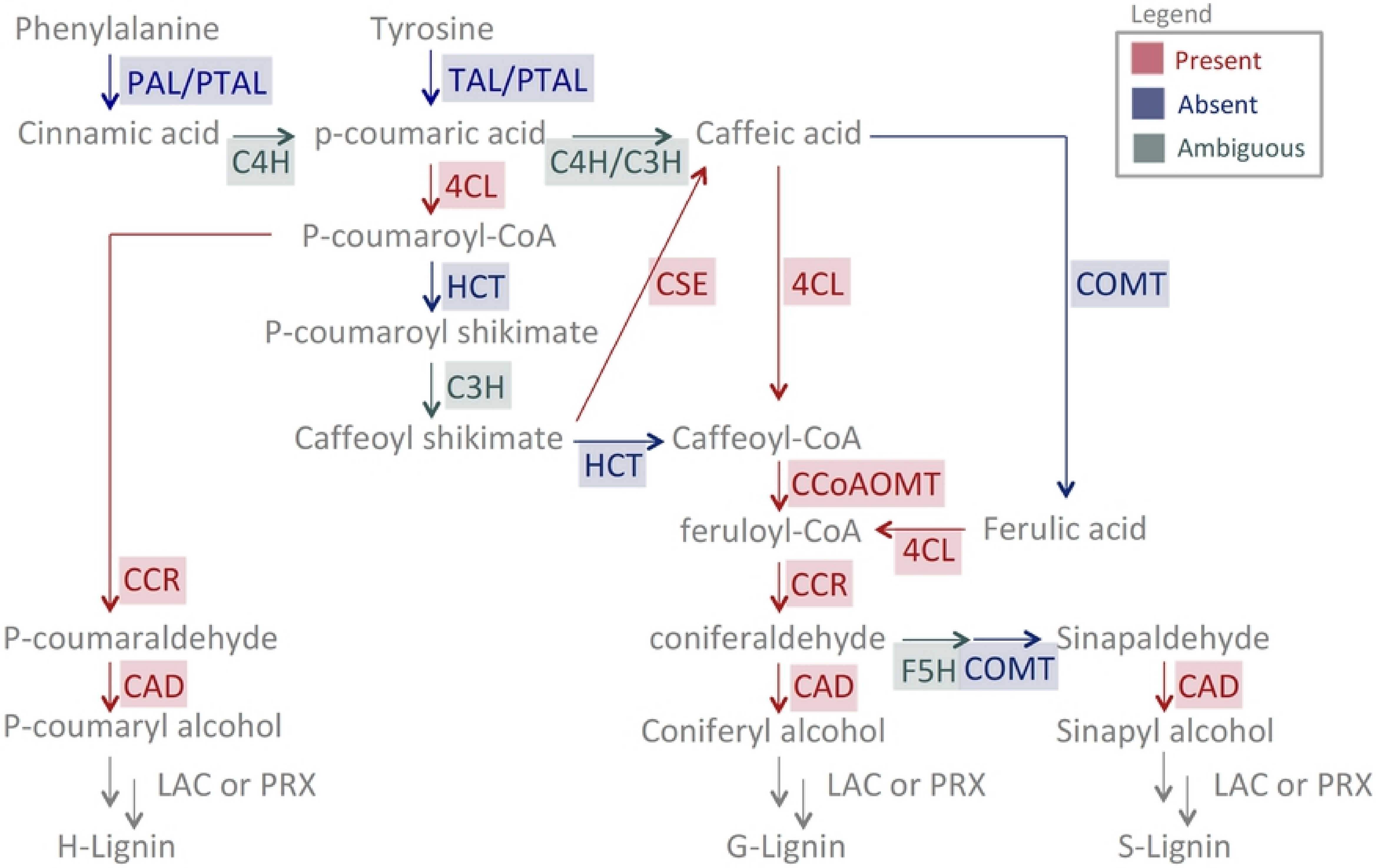
Legend
Present
Absent
Ambiguous
Phenylalanine
PAL/PTAL
Tyrosine
TAL/PTAL
Cinnamic acid
C4H
p-coumaric acid
C4H/C3H
Caffeic acid
4CL
P-coumaroyl-CoA
HCT
P-coumaroyl shikimate
C3H
Caffeoyl shikimate
CSE
4CL
COMT
HCT
Caffeoyl-CoA
CCoAOMT
feruloyl-CoA
4CL
Ferulic acid
CCR
CCR
P-coumaraldehyde
coniferaldehyde
F5H
COMT
Sinapaldehyde
CAD
CAD
CAD
P-coumaryl alcohol
Coniferyl alcohol
Sinapyl alcohol
LAC or PRX
LAC or PRX
LAC or PRX
H-Lignin
G-Lignin
S-Lignin

Figure 1

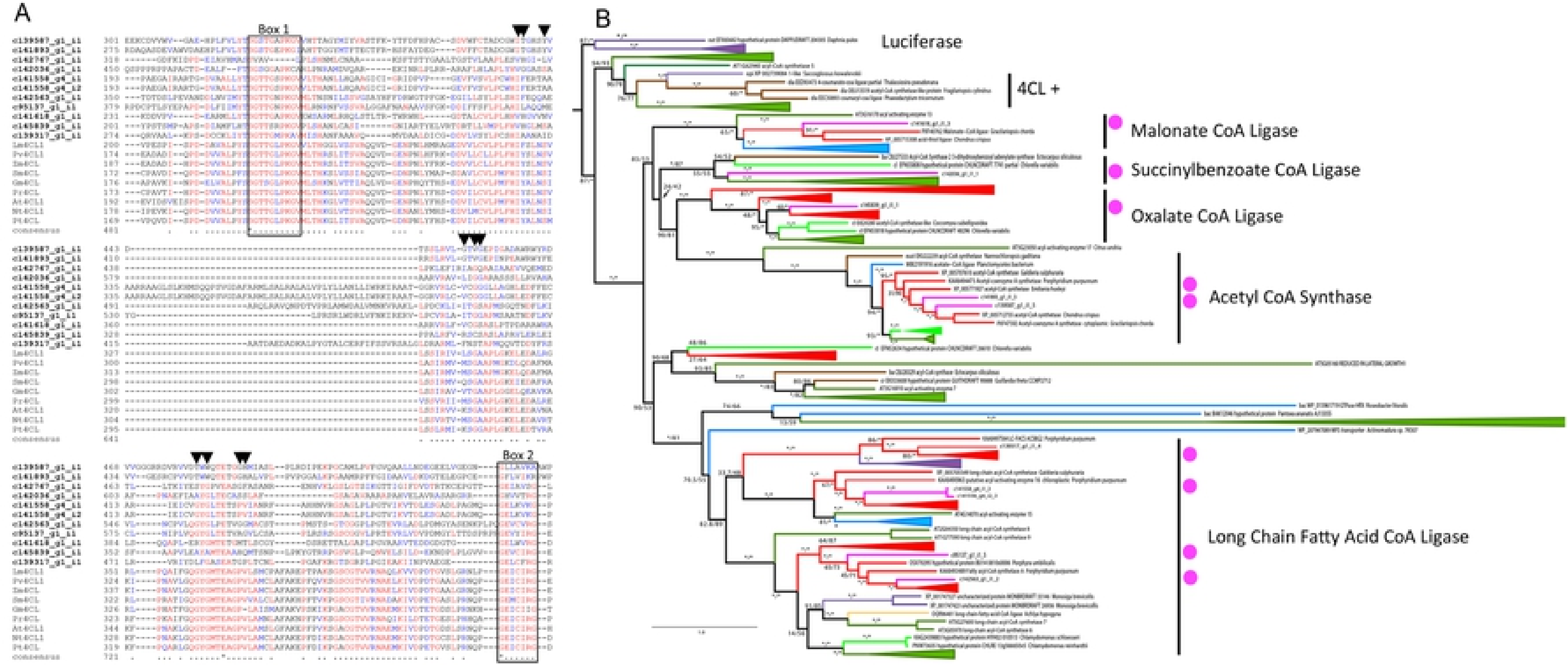
A
Box 1
Box 2
c139587_g1_i1
c141893_g1_i1
c142767_g1_i1
c142036_g1_i1
c141558_g4_i1
c141558_g4_i2
c142563_g1_i1
c95137_g1_i1
c141618_g1_i1
c145839_g1_i1
c139317_g1_i1
Lm4CL1
Pv4CL1
Zm4CL
Sm4CL
Gm4CL
Pr4CL
At4CL1
Nt4CL1
Pt4CL
consensus
B
Luciferase
4CL +
Malonate CoA Ligase
Succinylbenzoate CoA Ligase
Oxalate CoA Ligase
Acetyl CoA Synthase
Long Chain Fatty Acid CoA Ligase

Figure 2

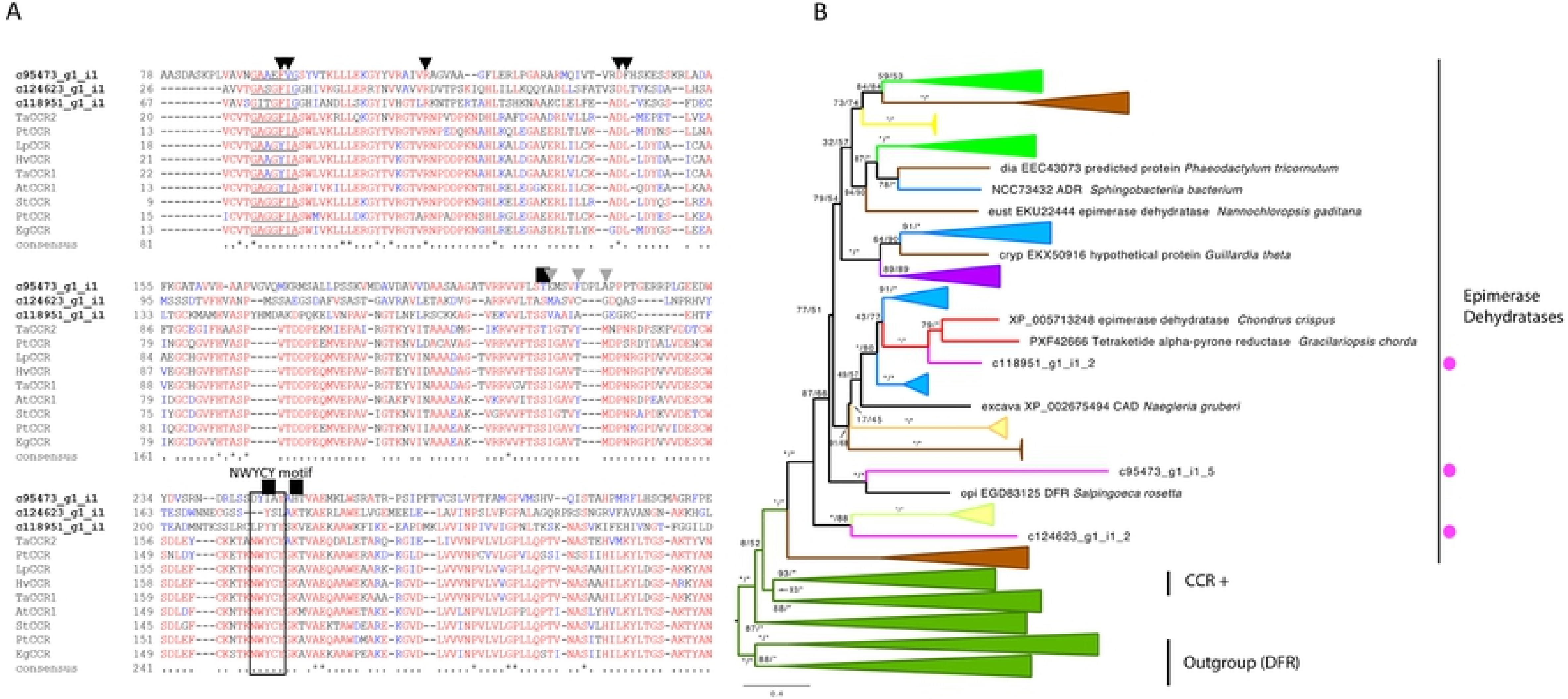
A
c95473_g1_i1
c124623_g1_i1
c118951_g1_i1
TaCCR2
PtCCR
LpCCR
HvCCR
TaCCR1
AtCCR1
StCCR
PtCCR
EgCCR
consensus
NWYCY motif
B
dia EEC43073 predicted protein Phaeodactylum tricornutum
NCC73432 ADR Sphingobacteriia bacterium
eust EKU22444 epimerase dehydratase Nannochloropsis gaditana
cryp EKX50916 hypothetical protein Guillardia theta
XP_005713248 epimerase dehydratase Chondrus crispus
PXF42666 Tetraketide alpha-pyrone reductase Gracilariopsis chorda
c118951_g1_i1_2
excava XP_002675494 CAD Naegleria gruberi
c95473_g1_i1_5
opi EGD83125 DFR Salpingoeca rosetta
c124623_g1_i1_2
Epimerase Dehydratases
CCR +
Outgroup (DFR)
0.4

Figure 3

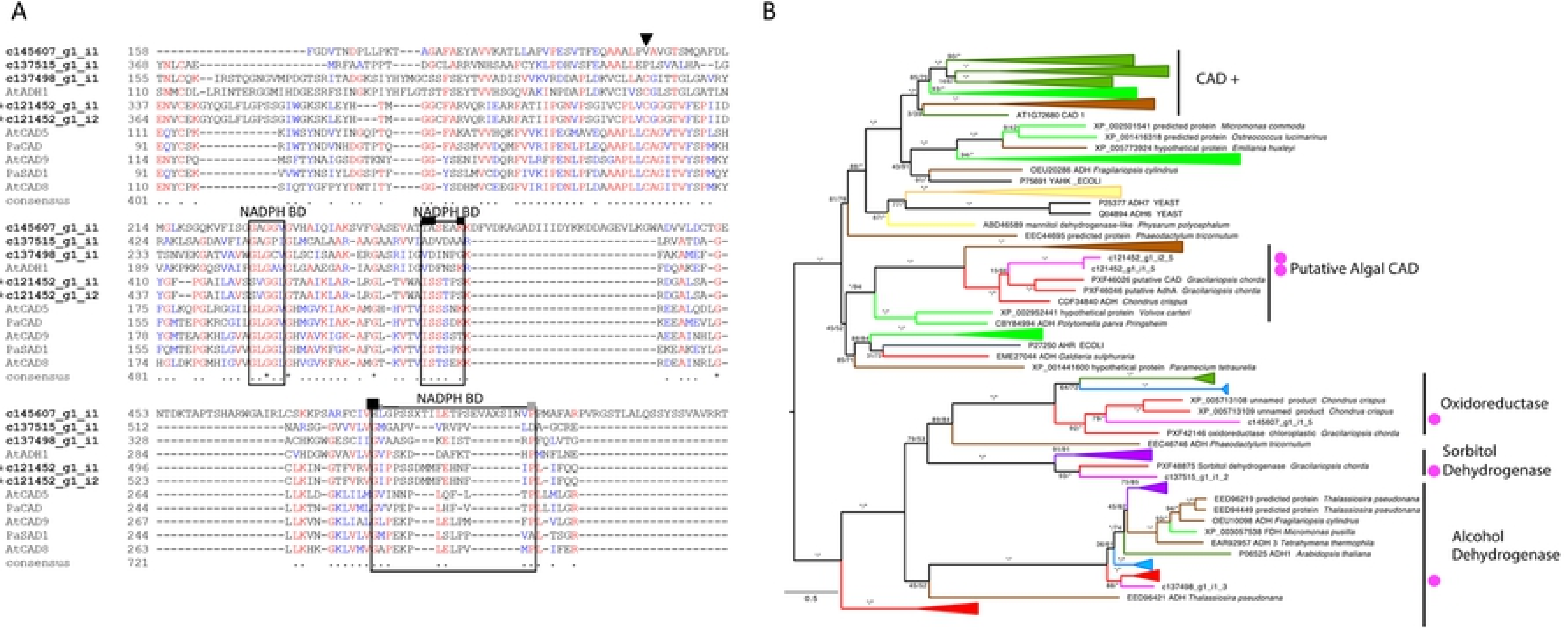
A
c145607_g1_i1 158 -----------------------FGDVTNDPLLPKT---AGAFAEYAVVKATLLAPVPESVTFEQAAALPVAVGTSMQAFDL
c137515_g1_i1 368 YNLCAE--------------------MRFAATPPT----DGCLARRVNHSAAFCYKLPDHVSFEAAALLEPLSVALHA--LG
c137498_g1_i1 155 TNLCQK--IRSTQGNGVMPDGTSRITADGKSIYHYMGCSSFSEYTVVADISVVKVRDDAPLDKVCLLACGITTGLGAVRY
AtADH1 110 SNMCDL-LRINTERGGMIHDGESRFSINGKPIYHFLGTSTFSEYTVVHSGQVAKINPDAPLDKVCIVSCGLSTGLGATLN
c121452_g1_i1 337 ENVCEKGYQGLFLGPSSGIWGKSKLEYH---TM----GGCFARVQRIEARFATIIPGNVPSGIVCPLVCGGGTVFEPIID
c121452_g1_i2 364 ENVCEKGYQGLFLGPSSGIWGKSKLEYH---TM----GGCFARVQRIEARFATIIPGNVPSGIVCPLVCGGGTVFEPIID
AtCAD5 111 EQYCPK------------KIWSYNDVYINGQPTQ----GG-FAKATVVHQKFVVKIPEGMAVEQAAPLLCAGVTVYSPLSH
PaCAD 91 EQYCSK------------RIWTYNDVNHDGTPTQ----GG-FASSMVVDQMFVVRIPENLPLEQAAPLLCAGVTVFSPMKH
AtCAD9 114 ENYCPQ------------MSFTYNAIGSDGTKNY----GG-YSENIVVDQRFVLRFPENLPSDSGAPLLCAGITVYSPMKY
PaSAD1 91 EQYCEK------------VVWTYNSIYLDGSPTF----GG-YSSLMVCDQRFIVKIPENLPFDAAAPLLCAGITVYSPMQY
AtCAD8 110 ENYCPK------------SIQTYGFPYYDNTITY----GG-YSDHMVCEEGFVIRIPDNLPLDAAAPLLCAGITVYSPMKY
consensus 401
NADPH BD
NADPH BD
c145607_g1_i1 214 MGLKSGQKVFISGGAGGVGVHAIQIAKSVFGASEVATTASEAPKDFVDKAGADIIIDYKKDDAGEVLKGWADVVLDCTGE
c137515_g1_i1 424 RAKLSAGDAVFIAGAGPIGLMCALAAR-AAGAARVVIADVDAAR--------------------------LRVATDA-G-
c137498_g1_i1 233 TSNVEKGATVAVWGLGCVGLSCISAAK-ERGASRIIGVDINPGK--------------------------FAKAMEF-G-
AtADH1 189 VAKPKKGQSVAIFGLGAVGLGAAEGAR-IAGASRIIGVDFNSKR--------------------------FDQAKEF-G-
c121452_g1_i1 410 YGF--PGAILAVSSVGGLGTAAIKLAR-LRGL-TVWAISSTPSK--------------------------RDGALSA-G-
c121452_g1_i2 437 YGF--PGAILAVSSVGGLGTAAIKLAR-LRGL-TVWAISSTPSK--------------------------RDGALSA-G-
AtCAD5 175 FGLKQPGLRGGILGLGGVGHMGVKIAK-AMGH-HVTVISSSNKK--------------------------REEALQDLG-
PaCAD 155 FGMTEPGKRCGILGLGGVGHMGVKIAK-AFGL-HVTVISSSDKK--------------------------KEEAMEVLG-
AtCAD9 178 YGMTEAGKHLGVAGLGGLGHVAVKIGK-AFGL-KVTVISSSSTK--------------------------AEEAINHLG-
PaSAD1 155 FGMTEPGKSLGVVGLGGLGHMAVKFGK-AFGL-KVTVISTSPKK--------------------------EKEAKEYLG-
AtCAD8 174 HGLDKPGMHIGVVGLGGLGHVGVKFAK-AMGT-KVTVISTSEKK--------------------------RDEAINRLG-
consensus 481
NADPH BD
c145607_g1_i1 453 NTDKTAPTSHARWGAIRLCSKKPSARFCIVALGPSSKTILETFSEVAKSINVFPMAFARPVRGSTLALQSSYSSVAVRRT
c137515_g1_i1 512 ---------------------NARSG-GVVVLVGMGAPV---VRVPV--------LDA-GCRE-------------------
c137498_g1_i1 328 ---------------------ACHKGWGESCIIGVAASG---KEIST--------RPFQLVTG-------------------
AtADH1 284 ---------------------CVHDGWGVAVLVGVPSKD---DAFKT--------HPMNFLNE-------------------
c121452_g1_i1 496 ---------------------CLKIN-GTFVRVGIPPSSDMMFEHNF--------IPL-IFQQ-------------------
c121452_g1_i2 523 ---------------------CLKIN-GTFVRVGIPPSSDMMFEHNF--------IPL-IFQQ-------------------
AtCAD5 264 ---------------------LLKLD-GKLILMGVINNP---LQF-L--------TPLLMLGR-------------------
PaCAD 244 ---------------------LLKTN-GKLVMLGVVPEP---LHF-V--------TPLLILGR-------------------
AtCAD9 267 ---------------------LLKVN-GKLIALELPEKP---LELPM--------FPL-VLGR-------------------
PaSAD1 244 ---------------------LLKVN-GKLVLVGMPEKP---LSLPP--------VAL-TSGR-------------------
AtCAD8 263 ---------------------LLKHK-GKLVMVGAPEKP---LELPV--------MPL-IFER-------------------
consensus 721
B
CAD +
AT1G72680 CAD 1
XP_002501541 predicted protein Micromonas commoda
XP_001418318 predicted protein Ostreococcus lucimarinus
XP_005773924 hypothetical protein Emiliania huxleyi
OEU20286 ADH Fragilariopsis cylindrus
P75691 YAHK_ECOLI
P25377 ADH7 YEAST
Q04894 ADH6 YEAST
ABD46589 mannitol dehydrogenase-like Physarum polycephalum
EEC44695 predicted protein Phaeodactylum tricornutum
Putative Algal CAD
c121452_g1_i2_5
c121452_g1_i1_5
PXF46026 putative CAD Gracilariopsis chorda
PXF46046 putative AdhA Gracilariopsis chorda
CDF34840 ADH Chondrus crispus
XP_002952441 hypothetical protein Volvox carteri
CBY84994 ADH Polytomella parva Pringsheim
P27250 AHR ECOLI
EME27044 ADH Galdieria sulphuraria
XP_001441600 hypothetical protein Paramecium tetraurelia
Oxidoreductase
XP_005713108 unnamed product Chondrus crispus
XP_005713109 unnamed product Chondrus crispus
c145607_g1_i1_5
PXF42146 oxidoreductase chloroplastic Gracilariopsis chorda
EEC46746 ADH Phaeodactylum tricornutum
Sorbitol Dehydrogenase
PXF48875 Sorbitol dehydrogenase Gracilariopsis chorda
c137515_g1_i1_2
Alcohol Dehydrogenase
EED96219 predicted protein Thalassiosira pseudonana
EED94449 predicted protein Thalassiosira pseudonana
OEU10098 ADH Fragilariopsis cylindrus
XP_003057538 FDH Micromonas pusilla
EAR92957 ADH 3 Tetrahymena thermophila
P06525 ADH1 Arabidopsis thaliana
c137498_g1_i1_3
EED96421 ADH Thalassiosira pseudonana
0.5

Figure 4