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

Jan Y. Xue1*, Katharine R. Hind1,2, Matthew A. Lemay1,2, Andrea Mcminigal1, Emma Jourdain1, Cheong Xin Chan3, Patrick T. Martone1,2

1 Department of Botany and Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada, 2 Hakai Institute, Heriot Bay, BC, Canada, 3 Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of Queensland, Brisbane, Queensland, Australia

* 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.

**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]; and evolutionary biologists are examining patterns in coralline trait
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. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 [24]. Lignins are complex aromatic polymers predominantly found in the secondary cell walls of plant support tissues [25, 26] and were long considered to have evolved when land plants emerged from the oceans, enabling upright growth in air [27]. 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 [28, 29]. In the articulated coralline C. cheilosporioides, lignins were found predominantly within decalcified flexible joints, called genicula [24], that have remarkable biomechanical properties, permitting this articulated coralline species to thrive along wave-battered coastlines [30, 31].

Because lignin biosynthesis is physiologically complex and involves several enzymes in the monolignol pathway [32–34], Martone et al. [24] 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 (FSH) evolved independently in lycopsods and embryophytes [35, 36]. 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 [37], raising questions about how the monolignol pathway may have evolved across such evolutionarily divergent lineages.

Recently, researchers have suggested that genes related to monolignol production radiated and diversified during the transition to land, noting that the presence of multiple enzymes involved in monolignol production are not present in green algae, but begin to appear after the divergence of streptophytes [38]. Moreover, using genomic data, Labeeuw et al. [37] suggested that putative 4CL, CCR, and CAD sequences in Calliarthron tuberculosum may have arisen independently through horizontal gene transfer [37]. These studies begin to suggest a non-orthologous origin of monolignol production in red algae and land plants, but a complete analysis of all currently known genes in the monolignol biosynthetic pathway has not been completed. Labeeuw et al. [37] examined only three of the eight classes of genes involved in monolignol biosynthesis and based their conclusions entirely on genomic data [37]. Until now, questions about monolignol evolution in coralline red algae have largely gone unanswered as transcriptomic and genomic data have mostly been limited to non-coraline red algae (e.g. [39–42] but see [43, 44]).

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 [45] and a draft nuclear genome [46] 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. tuberculosa* 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 PTM1 and PTM2 was assembled independently from raw reads, and this combined dataset was used for all 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; S1A Fig). Genomic sequences were also assembled for *C. tuberculosa* (S1 Table), 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* ≤ 10⁻⁵).

The incomplete monolignol biosynthetic pathway in *Calliarthron tuberculosa*

The combined *C. tuberculosa* 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* ≤ 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* ≤ 10⁻⁵) except for those identified for PAL/TAL/PTAL.

Candidate sequences from *C. tuberculosa* (bolded as contig_gene_isoform in Figs 2–4) were characterized by comparing key residues with their land plant homologs in multiple sequence alignments. The evolutionary relationships between the identified *C. tuberculosa* 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 2–4). Descriptions of the other biosynthetic enzymes CCoAMT, CSE, and the cytochrome P450 sequences C3H, C4H, F5H are found in S1 Appendix and S2–S4 Figs.

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 (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 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.
A query of these sequences against the *A. thaliana* proteome returned sequences within the CCR family (CCR7, CCR4, CCR-Like6) (S2 Table). 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. tuberculatum* and CCRs in land plants (Fig 3A).

In the CCR gene tree analysis, *C. tuberculatum* 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. tuberculatum* were mapped with epimerase dehydratase type sequences that included the *A. thaliana* CCR family (Fig 3B). Sequences from *C. tuberculatum* 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. tuberculatum*) 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 [60], 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 (S2 Table). NADPH-binding motifs show moderate conservation (IS > 80%) (Fig 4A). One C. tuberculosaum sequence showed high conservation with land plant counterparts, suggesting a promising CAD candidate (+ in Fig 3A and 3B).

In the CAD gene tree analysis, all C. tuberculosaum sequences grouped with sequences from other Rhodophytes (Fig 4B). CAD candidates in C. tuberculosaum were mapped with their embryophyte CAD counterparts and closely related alcohol dehydrogenases. Sequences from C. tuberculosaum 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. tuberculosaum 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 tuberculosaum

To enable broad and rapid identification of C. tuberculosaum genes involved in specific metabolic processes, we present two general tools for gene identification within the C. tuberculosaum transcriptome dataset using KEGG based annotations. This involves extracting whole metabolic pathways or individual genes (see S1 Appendix and S5 Fig). We included annotations for all metabolic genes recovered in the C. tuberculosaum transcriptome (S3 Table). We identified 36 putative C. tuberculosaum genes present in the starch and sucrose metabolism pathway (S5 Fig).
In addition, we individually searched for genes potentially involved in calcification [43, 61, 62] 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 (S5 Table).

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. tuberculoseum_ 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. tuberculoseum_.

Despite supporting evidence from sequence similarity analyses, functional predictions for candidate sequences in the monolignol pathway within _C. tuberculoseum_ are obscured by the gene tree analysis. If the monolignol pathway in embryophytes and _C. tuberculoseum_ 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. tuberculoseum_. _C. tuberculoseum_ sequences were found within each respective overall protein family, but consistently grouped with land plant genes of non-monolignol forming function. If these _C. tuberculoseum_ sequences are functionally homologous to the monolignol biosynthesis pathway, further investigation into their function may provide insights into the convergence of monolignol biosynthesis across different lineages.
counterpart in land plants, then they likely arose independently in *C. tuberculorum*. 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 [63]. Biosynthetic enzymes in *C. tuberculorum* 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. tuberculorum* may have been acquired through horizontal gene transfer from a bacterial source [36]. Thus, over evolutionary time genes in *C. tuberculorum* 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. tuberculorum* 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. tuberculorum* 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 tuberculorum***

Several key steps in the monolignol biosynthetic pathway were not recovered in the *C. tuberculorum* 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 [64–66]. Although no full-length homologs were identified in the *C. tuberculorum* 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. tuberculorum*, RACE amplification could help determine if the short ammonia-lyase we identified has a longer transcript. *C. tuberculorum* 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 [67]. 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. tuberculorum* 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 (S2B Fig). The identified candidates from *C. tuberculorum* 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 [68, 69]. Alternatively, each of the identified P450 clades in *C. tuberculorum* 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) [35, 36]. 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. tuberculosa.

HCT is one alternative route shifting monolignol synthesis from H- to G- to S- types using a temporary shikimate decoration (Fig 1) [70]. Its absence could suggest that C. tuberculosa 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. tuberculosa. 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 [71–73]. The absence of this enzyme raises questions about how C. tuberculosa 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 [74]. 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. tuberculosa but has yet to be identified based on sequence similarity.

Conclusion
In summary, we have identified several gene candidates in the C. tuberculosa transcriptome that represent possible 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. [24], our gene trees do not demonstrate a deeply conserved evolution of monolignol biosynthesis, but instead suggest that each of the enzymes identified in C. tuberculosa 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. tuberculosa. By providing methods to easily identify additional gene candidates from the C. tuberculosa transcriptome, we aim to facilitate future research on this fascinating organism.

Methods
Specimen collection and sequencing
Two male, haploid specimens of Calliarthron tuberculosa were collected October 6, 2013, from Bluestone Point (48.81952, -125.1640), Bamfield, British Columbia, Canada. Samples were verified as haploid male specimens by identifying spermatangial conceptacles under a dissecting microscope. 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.

Samples were processed as either whole tissue containing calcified intergenicula and non-calcified genicula (Sample I+G/PTM1 in the dataset) or calcified tissue only (Sample I/PTM2 in the dataset). To reduce epiphyte contamination, samples used for sequencing were collected from newer growth at the tips where epiphytes had not yet settled. On site, samples were
brushed, rinsed with ethanol, dissected for regions with both calcified intergenicula and non-calcified genicula or calcified intergenicula only. Fragments were observed under a dissecting microscope for the absence of epiphytes and then flash frozen in liquid nitrogen and stored at -80°C until use. To extract RNA, samples were ground to a powder in a sterile pre-chilled mortar and pestle. The mortar was placed on dry ice and liquid nitrogen was added into the mortar to maintain sample temperatures at -80°C. RNA was extracted from 70 mg– 80 mg of ground sample using the Spectrum Plant Total RNA kit (STRN50, Sigma-Aldrich) with the following modifications. Tubes containing ground material and lysis solution were centrifuged for 10 minutes at 14000 g to pellet debris. Protocol A was used to bind RNA to the column. 750 μl of binding solution was used. The elutions were performed with 40 μl of pre-warmed 56°C RNase-free water. DNA was removed from 1200ng of RNA samples between 50 ng/μL and 400 ng/μL in concentration with DNase (AM1906, ThermoFisher). RNA was then converted into cDNA, and quality and yields were verified using Bioanalyzer 2100. Samples of each type were then pooled (i.e. all samples containing both calcified and uncalcified segments were pooled and samples containing calcified segments only were pooled). RNA was then sequenced on the Illumina HiSeq 2000 platform (paired-end 2x100bp, insert size ~220bp) at the UBC Sequencing + Bioinformatics Consortium.

**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-hydroxy-cinnamoyl-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 [75], 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 [76]. Briefly, each transcript was searched against the NCBI RefSeq protein database (BLASTX, E ≤ 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 [77]. Identification of genes present in KEGG annotated pathways were extracted using the pathview package [78].

**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 ≤ 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 (S1B Fig). 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 [79]. 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 [80], and full-length coding sequences (CDSs) were predicted with TransDecoder v5.0.1 [75]. 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 [81] from the genome scaffolds.

HMM based gene candidate search

Monolignol biosynthesis gene candidates were identified from the *C. tuberculosa*um transcriptomic dataset using Hidden Markov Model (HMM) based searches [82]. Transcriptomic sequence contigs were translated into all six reading frames using EMBOSS Transeq [83]. 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 [84] with no manual adjustment. The profiles were searched against the translated *C. tuberculosa*um dataset in HMMER searches [82] 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 [85] to verify their closest homolog match (BLASTP, E ≤ 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. tuberculosa*um with their land plant counterpart protein were produced. Sequences were aligned using MUSCLE under default settings [84]. 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 S1 Appendix.

Gene tree analysis

Gene trees were reconstructed for the identified candidate sequences of *C. tuberculosa*um. For each gene tree analysis, sequence candidates from *C. tuberculosa*um, 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. tuberculosa*um candidates as a query against the total database using default settings (BLASTP, E ≤ 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 [84]. Sites with <80% coverage were removed using trimAl [86]. IQTree was used to search for the evolutionary model alignment under a BIC criterion [87, 88]. A maximum likelihood tree was reconstructed using IQTree [89], with node support calculated based on 1000 ultrafast bootstrap pseudoreplicates in IQTree [89]. A clade is considered strongly supported when bootstrap value ≥ 95%. FigTree was used to edit branch width and colors [90]. Species in the phylogenies are colored as followed Embryophyta (including both vascular and non-vascular plant species)—dark green, Chlorophyta—light green, Rhodophyta—red, Animalia and Opisthokonta—purple, Bacteria and
Cyanobacteria—blue, Oomycota, Mycetozoa and Fungi—yellow, Ochrophyta—brown. Accession numbers can be found in S1 Appendix.

**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 S1 Table. Adapter sequences were removed using Trimmomatic v0.33 [91] (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 [46] were used in a de novo genome assembly using SPAdes [92]. 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_Scaffold [93]. 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 [80] and TransDecoder [75], from which the coded protein sequences were predicted.

**Completeness of transcriptome and genome data**

The completeness of the genome and transcriptome data were assessed by the recovery of core conserved eukaryote genes with the Core Eukaryotic Genes Mapping Approach (CEGMA) [94] and Benchmarking Universal Single-Copy Orthologs (BUSCO) [95] 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 (\( E \leq 10^{-5} \)).

**Key resources**

A summary of key resources used in this study, including sample identifiers, reagents, online resources, and programs can be found in S6 Table.

**Supporting information**

*S1 Appendix. Supplemental methods and results.* Genome reassembly and removal of contamination; analysis of completeness for genomic and transcriptomic datasets; E-value thresholds used in gene candidate identification; accession numbers for all sequences used; identification of cytochrome P450s sequences (C3H, C4H, and F5H), CCoAOMT, and CSE; and the benefits and limitations of using KEGG mapping for gene annotation of *Calliarthron* sequences.

(DOCX)
S1 Fig. Completeness of the *C. tuberculosis* 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).

(TIF)

S2 Fig. C3H, C4H, F5H, P450 candidates from *C. tuberculosis* in relation to plants and other taxa. (A) Partial alignment of *C. tuberculosis* 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. tuberculosis* (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.

(TIF)

S3 Fig. CCoAOMT candidates from *C. tuberculosis* in relation to plants and other taxa. (A) Partial alignment of *C. tuberculosis* 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. tuberculosis* (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.

(TIF)

S4 Fig. CSE candidates from *C. tuberculosis* in relation to plants and other taxa. (A) Partial alignment of *C. tuberculosis* CSE sequence candidates with CSE from land plants. Acyl transferase motifs (HX4D), lipase motifs (GXSXG) and active site residues (triangle) are indicated. Sites with < 70% coverage were removed. (B) Unrooted maximum likelihood gene tree of *C. tuberculosis* 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.

(TIF)
S5 Fig. A visual representation of the *C. tuberculatus* 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. tuberculatus* 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.

(TIF)

S1 Table. Summary statistics for the *C. tuberculatus* 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.

(XLSX)

S2 Table. 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.

(XLSX)

S3 Table. KEGG annotations of *Calliarthron tuberculatus* 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.

(XLSX)

S4 Table. Listed representation of the *C. tuberculatus* sequences present in starch and sucrose metabolism pathway from the KEGG based annotation. *C. tuberculatus* 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 representing 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, accessible from the Github link provided.

(XLSX)

S5 Table. A list of calcification related gene candidates identified from KEGG-based annotations of the *C. tuberculatus* transcriptome. Calcification gene candidates were initially selected based on a literature search, and additional *C. tuberculatus* sequences were identified manually from the KEGG based annotations (annotation file available on the Github link), thus this list is not exhaustive. The genes are organized by their functional classification indicated as “overall function”. “KEGG Identifier” refers to the specific KEGG code for the gene, “Contig Name” refers to the sequence identifier from the *Calliarthron* transcriptome representing the contig name_gene number_gene isoform and “Gene Name” refers to the gene acronym, the gene name, and its enzyme commission (EC) number.

(XLSX)

S6 Table. Key resources used in this study, including sample identifiers, reagents, online resources, and programs.

(XLSX)
Acknowledgments

We thank Dana Price and Debasish 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.

Author Contributions

Conceptualization: Jan Y. Xue, Katharine R. Hind, Patrick T. Martone.

Funding acquisition: Patrick T. Martone.

Investigation: Jan Y. Xue, Katharine R. Hind, Matthew A. Lemay, Andrea Mcminigal, Cheong Xin Chan.

Methodology: Jan Y. Xue, Katharine R. Hind, Matthew A. Lemay, Cheong Xin Chan.

Supervision: Patrick T. Martone.

Visualization: Emma Jourdain.

Writing – original draft: Jan Y. Xue, Cheong Xin Chan, Patrick T. Martone.

Writing – review & editing: Jan Y. Xue, Katharine R. Hind, Matthew A. Lemay, Emma Jourdain, Patrick T. Martone.

References

1. Adey WH. The algal ridges and coral reefs of St. Croix: their structure and Holocene development. Atoll Research Bulletin. 1975; 1–67. 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. https://doi.org/10.1111/j.1749-6632.1963.tb13465.x PMID: 13949254

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. https://doi.org/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. https://doi.org/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. https://doi.org/10.3354/meps314001

7. Fisher K, Martone PT. Field study of growth and calcification rates of three species of articulated coraline algae in British Columbia, Canada. Biological Bulletin. 2014; 226: 121–130. https://doi.org/10.1086/BBLv226n2p121 PMID: 24797094

8. van der Heijden LH, Kamenos NA. Reviews and syntheses: Calculating the global contribution of coraline algae to total carbon burial. Biogeosciences. 2015; 12: 6429–6441. https://doi.org/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. https://doi.org/10.1111/jpy.12761 PMID: 29920669

10. Hind KR, Miller KA, Young M, Jensen C, Gabrielson PW, Martone PT. Resolving cryptic species of Boselli (Corallinales, Rhodophyta) using contemporary and historical DNA. American journal of botany. 2015; 102: 1912–1930. https://doi.org/10.3732/ajb.1500308 PMID: 26542846

11. Hind KR, Gabrielson PW, Lindstrom SC, Martone PT. Misleading morphologies and the importance of sequencing type specimens for resolving coraline taxonomy (Corallinales, Rhodophyta): Pachyarthron cretaceum is Corallina officinalis. Journal of Phycology. 2014; 50: 760–764. https://doi.org/10.1111/jpy.12205 PMID: 26988460
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. https://doi.org/10.1371/journal.pone.0225645 PMID: 31790447

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. https://doi.org/10.1016/j.marenvres.2020.105107 PMID: 32890983

14. Cornwall CE, Corneau S, McCulloch MT. Coralline algae elevate pH at the site of calcification under ocean acidification. Global change biology. 2017; 23: 4245–4256. https://doi.org/10.1111/gcb.13673 PMID: 28370806

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. https://doi.org/10.1038/nclimate2273

17. Noisette F, Egilsdottir H, Davout 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. https://doi.org/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. https://doi.org/10.1111/jpy.12788 PMID: 30246453

19. Janot K, Martone PT. Convergence of joint mechanics in independently evolving, articulated coralline algae. Journal of experimental biology. 2016; 219: 383–391. https://doi.org/10.1242/jeb.131755 PMID: 26596529

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. https://doi.org/10.1111/jpy.12520 PMID: 28191634

23. Peña V, Vieira C, Braga JC, Aguirre J, Rösler A, Baele G, et al. Radiation of the coralline red algae (Corallinophycidae, Rhodophyta) crown group as inferred from a multilocus time-calibrated phylogeny. Molecular Phylogenetics and Evolution. 2020; 150: 106845. https://doi.org/10.1016/j.ympev.2020.106845 PMID: 32360706

24. 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. https://doi.org/10.1016/j.cub.2008.12.031 PMID: 19167225

25. Boerjan W, Ralph J, Baucher M. Lignin Biosynthesis. Annual Review of Plant Biology. 2003; 54: 519–546. https://doi.org/10.1146/annurev.arplant.54.031902.134938 PMID: 14503002

26. 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. https://doi.org/10.1016/j.copbio.2015.10.009 PMID: 26775114

27. Vanholme R, Demedts B, Morreel K, Ralph J, Boerjan W. Lignin biosynthesis and structure. Plant Physiology. 2010; 153: 895–905. https://doi.org/10.1104/pp.110.155119 PMID: 20472751

28. Lange BM, Lapierre C, Sandermann H. Elicitor-induced spruce stress lignin: Structural similarity to early developmental lignins. Plant Physiology. 1995; 108: 1277–1287. https://doi.org/10.1104/pp.108.3.1277 PMID: 12228544

29. Tronchet M, Balagüé 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. https://doi.org/10.1111/j.1364-3703.2009.00578.x PMID: 20078778

30. Martone PT. Kelp versus coralline: Cellular basis for mechanical strength in the wave-swept seaweed Calliariad (Corallinaceae, Rhodophyta). Journal of Phycology. 2007; 43: 882–891. https://doi.org/10.1111/j.1529-8817.2007.00397.x

31. 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. https://doi.org/10.1242/jeb.138867 PMID: 27307542

32. Weng JK, Chapple C. The origin and evolution of lignin biosynthesis. New Phytologist. 2010; 187: 273–285. https://doi.org/10.1111/j.1469-8137.2010.03327.x PMID: 20642725
33. Dixon RA, Barros J. Lignin biosynthesis: Old roads revisited and new roads explored. Open Biology. 2019;9. https://doi.org/10.1098/rsob.190215 PMID: 31795915

34. 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. https://doi.org/10.1104/pp.103.026484. role

35. 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. https://doi.org/10.1105/tpc.109.073528 PMID: 20371642

36. 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. https://doi.org/10.1073/pnas.0801696105 PMID: 18505841

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

38. de Vries S, Fürst-Jansen JMR, Irisarri I, Dhabalia Ashok A, Ischebeck T, Feussner K, et al. The evolution of the phenylpropanoid pathway entailed pronounced radiations and divergences of enzyme families. The Plant Journal. 2021; 107: 975–1002. https://doi.org/10.1111/tpj.15387 PMID: 34165823

39. 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. https://doi.org/10.1073/pnas.0801696105 PMID: 18505841

40. Collell 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. https://doi.org/10.1073/pnas.1221259110 PMID: 23503846

41. Brawley SH, Blind NA, Ficko-Blean E, Wheeler GL, Lohr M, Goodson HV., 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. https://doi.org/10.1038/nature02398 PMID: 15071595

42. Yang F, Wei Z, Long L. Transcriptional and Physiological Responses of the Tropical Reef Calcified Macroalga Amphiroa fragilissima to Elevated Temperature. Journal of Phycology. 2021; 57: 1254–1265. https://doi.org/10.1111/jpy.13158 PMID: 33655511

43. Bi G, Liu G, Zhao E, Du Q. Complete mitochondrial genome of a red calcified alga Calliarthron tuberculatus (Corallinaceae). Mitochondrial DNA. 2016; 27: 2554–2556. https://doi.org/10.3109/19401736.2015.1038801 PMID: 26258508

44. Chen CX, Yang EC, Banerjee T, Yoon HS, Martone PT, Estevez JM, et al. Red and green algal monolignol biosynthesis pathway revealed in a rich repertoire of red algal genes. Current Biology. 2019; 29: 1265–1265. https://doi.org/10.1016/j.cub.2019.05.052 PMID: 31017363

45. 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. https://doi.org/10.1105/tpc.109.072652 PMID: 20841425

46. 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. https://doi.org/10.1073/pnas.1430550100 PMID: 12819948

47. 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: 26883–26887. https://doi.org/10.1074/jbc.M100355200 PMID: 11323416

48. Shockey JM, Fulda MS, Browse J. Arabidopsis Contains a Large Superfamily of Acyl-Activating Enzymes: Phylogenetic and Acyl-Coenzyme A Synthetases. Plant physiology. 2003; 132: 1065–1076. https://doi.org/10.1104/pp.103.020652 PMID: 12806634

49. Shockey J, Browse J. Genome-level and biochemical diversity of the acyl-activating enzyme superfamily in plants. Plant Journal. 2011; 66: 143–160. https://doi.org/10.1111/j.1365-313X.2011.04512.x PMID: 21445329
52. Jörnvall H, Persson B, Krok M, Atrian S, González-Duarte R, Jeffery J, et al. Short-Chain Dehydrogenases/Reductases (SDR). Biochemistry. 1995; 34: 6003–6013. https://doi.org/10.1021/bi00018a001 PMID: 7742302

53. Sattler SA, Walker AM, Vermeer Is W, Sattler SE, Kang C. Structural and Biochemical Characterization of Cinnamoyl-CoA Reductases. Plant physiology. 2017; 173: 1031–1044. https://doi.org/10.1104/pp.16.01671 PMID: 27956488

54. 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. https://doi.org/10.1004/jbc.M202160200 PMID: 11976334

55. 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. https://doi.org/10.1093/jxb/ers257 PMID: 23028019

56. 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 A1CAD5 and A1CAD4. Organic & Biomolecular Chemistry. 2006; 4: 1687–1697. https://doi.org/10.1039/b601672c PMID: 16633561

57. 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.

58. van Borzyskowski LS, Rosenthal RG, Erb TJ. Evolutionary history and biotechnological future of carboxylases. Journal of Biotechnology. 2013; 168: 243–251. https://doi.org/10.1016/j.jbiotec.2013.05.007 PMID: 23702164

59. 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. https://doi.org/10.1105/tpc.010111 PMID: 11449052

60. 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. https://doi.org/10.3389/fpls.2018.01416 PMID: 30319676

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

62. Xue J, Purushotham P, Acheson JF, Ho R, Zimmer J, McFarlane C, et al. Functional characterization of a cellulose synthase, CICESA1, from the marine red alga Calliarthron tuberculatum (Corallinales). Journal of Experimental Botany. 2021; erab414. https://doi.org/10.1093/jxb/erab414 PMID: 34505622

63. 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. https://doi.org/10.1016/s0014-5793(02)02272-x PMID: 11852088

64. 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. https://doi.org/10.1038/nplants.2016.50 PMID: 27255834

65. Cooke HA, Christianson C V, Bruner SD. Structure and chemistry of 4-methyldeimidazoic-5-one containing enzymes. Current opinion in chemical biology. 2009; 13: 460–468. https://doi.org/10.1016/j.cbpa.2009.06.013 PMID: 19620019

66. Mohy El-Din SM, El-Ahwan 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. https://doi.org/10.1016/j.jtusci.2015.06.004

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

68. 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. https://doi.org/10.1111/nph.13790 PMID: 26682704

69. Hoffmann L, Besseau S, Geoffroy P, Ritzenhaler C, Meyer D, Lapierre C, et al. Silencing of Hydroxycinnamoyl-Coenzyme A Shikimate/Quinate Hydroxycinnamoyltransferase Affects Phenylpropanoid Bio-synthesis. The Plant cell. 2004; 16: 1446–1465. https://doi.org/10.1105/tpc.020297 PMID: 15161961

70. 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. https://doi.org/10.1023/a:1023022825098 PMID: 12777055
72. Lu F, Marita JM, Lapierre C, Jouanin L, Morreel K, Boerjan W, et al. Sequencing around 5-Hydroxyferuloyl Alcohol-Derived Units in Caffeic Acid O-Methyltransferase-Deficient Poplar Lignins. Plant physiology (Bethesda). 2010; 153: 569–579. https://doi.org/10.1104/pp.110.154278 PMID: 20427467

73. 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. https://doi.org/10.1105/tpc.13.1.73 PMID: 11158530

74. 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 Lobolly Pine. Proceedings of the National Academy of Sciences—PNAS. 1997; 94: 5461–5466. https://doi.org/10.1073/pnas.94.10.5461 PMID: 9144260

75. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Bowden J, et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nature protocols. 2013; 8: 1–43. https://doi.org/10.1038/nprot.2013.084

76. Conesa A, Gótz 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. https://doi.org/10.1093/bioinformatics/bti610 PMID: 16081474

77. 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. https://doi.org/10.1093/nar/gkv1070 PMID: 26476454

78. 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. https://doi.org/10.1093/bioinformatics/btt285 PMID: 23740750

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

80. 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. https://doi.org/10.1093/nar/gkg770 PMID: 14500829

81. 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. https://doi.org/10.1186/1471-2105-7-62

82. Finn RD, Clements J, Eddy SR. HMMER Web Server: Interactive Sequence Similarity Searching. Nucleic Acids Research. 2011; 39: W29–W37. https://doi.org/10.1093/nar/gkr367 PMID: 21593126

83. Rice P, Longden I, Bleasby A. EMBOSs: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000; 16: 276–277. https://doi.org/10.1016/s0168-9525(00)02024-2 PMID: 10827456

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

85. 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. https://doi.org/10.1145/1810085.1810099

86. 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. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

87. 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. https://doi.org/10.1186/1471-2148-10-242 PMID: 20696057

88. 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. https://doi.org/10.1093/molbev/msu300 PMID: 25371430

89. Hoang DT, Chemomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Molecular Biology and Evolution. 2018; 35: 518–522. https://doi.org/10.1093/molbev/msx281 PMID: 29077904

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

91. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30: 2114–2120. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404

92. 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. https://doi.org/10.1089/cmb.2012.0021 PMID: 22506599
93. 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. https://doi.org/10.1186/1471-2164-14-604

94. Parra G, Bradnam K, Korf I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007; 23: 1061–1067. https://doi.org/10.1093/bioinformatics/btm071 PMID: 17332020

95. 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. https://doi.org/10.1093/bioinformatics/btv351 PMID: 26059717