Plant Science 252 (2016) 324–334

Contents lists available at ScienceDirect

Plant Science

journal homepage: www.elsevier.com/locate/plantsci

De novo transcriptome assembly and identification of the gene conferring a "pandan-like" aroma in coconut (*Cocos nucifera* L.)

Chatree Saensuk[a,b,1], Samart Wanchana[c,1], Kiattawee Choowongkomon[d], Sugunya Wongpornchai[e], Tippaya Kraithong[f], Wachiraya Imsabai[g], Ekawat Chaichoompu[a], Vinitchan Ruanjaichon[c], Theerayut Toojinda[c], Apichart Vanavichit[a,h], Siwaret Arikit[a,h,*]

[a] *Rice Science Center, Kasetsart University Kamphaeng Saen Campus, Nakhon Pathom, 73140, Thailand*
[b] *Interdisciplinary Graduate Program in Genetic Engineering, Kasetsart University, Bangkok, 10900, Thailand*
[c] *National Center for Genetic Engineering and Biotechnology (BIOTEC), National Science and Technology Development Agency (NSTDA), Khlong Luang, Pathum Thani, 12120, Thailand*
[d] *Department of Biochemistry, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand*
[e] *Department of Chemistry, Faculty of Science, Chiang Mai University, Chiang Mai, 50200, Thailand*
[f] *Chumphon Horticultural Research Center, Department of Agriculture, Bangkok, 10900, Thailand*
[g] *Department of Horticulture, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University Kamphaeng Saen Campus, Nakhon Pathom, 73140, Thailand*
[h] *Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University Kamphaeng Saen Campus, Nakhon Pathom, 73140, Thailand*

ARTICLE INFO

Article history:
Received 15 May 2016
Received in revised form 26 August 2016
Accepted 27 August 2016
Available online 28 August 2016

Keywords:
Coconut
2-acetyl-1-pyrroline
Aroma
BADH
AMADH
RNA-seq

ABSTRACT

Thailand's aromatic coconut (*Cocos nucifera* L.) is a special type of green dwarf coconut, the liquid endosperm of which is characterized by a pleasant "pandan-like" aroma due to the presence of 2-acetyl-1-pyrroline (2AP). The aim of this study was to perform a *de novo* assembly of transriptome from *C. nucifera* endosperm and to identify the gene responsible for 2AP biosynthesis. *CnAMADH2* was identified as an ortholog of the rice aromatic gene and a G-to-C substitution found in exon 14 was associated with 2AP content in the aromatic green dwarf coconut accessions. The base substitution caused an amino-acid change, alanine-to-proline, at position 442 (P442A). The presence of P at this position might alter the steric conformation at the loop region and subsequently result in an unstabilized dimer conformation that could lower AMADH enzyme activity. Among AMADH/BADH protein sequences in different plant species, the P442A mutation was found exclusively in aromatic coconut. The PCR marker developed based on this sequence variation can perfectly detect the aromatic and non-aromatic alleles of the gene. This study confirms the hypothesis that plants may share a mechanism of 2AP biosynthesis. This is the first identification of the gene associated with 2AP biosynthesis in a tree plant.

1. Introduction

Coconut (*Cocos nucifera* L.) is an important crop cultivated on more than 12 million hectares across 89 tropical countries [1]. With many uses and an important role in smallholders' livelihoods, coconut is also known as the "Tree of Life." Coconut varieties can be conventionally classified into two broad groups, tall and dwarf. The tall varieties have greater genetic variability, as they are usually cross-pollinated, whereas the dwarf varieties, which are self-pollinated, have little genetic diversity [2]. Among the dwarf groups, the varieties are further divided into green, yellow and red, based on the color of the nuts. The tall type is by far more commonly grown around the world and its mature fruits are harvested and processed into a variety of products. By contrast, dwarf coconuts are mostly consumed by eating the fresh, tender meat of young fruits and drinking the coconut water. Coconut water is now becoming popular on the international market as a sports drink; hence, the coconut water industry has been growing in recent years [3]. Thailand's aromatic coconut is a variant of the green dwarf type

Abbreviations: 2AP, 2-acetyl-1-pyrroline; BADH, betaine aldehyde dyhydrogenase; AMADH, amino aldehyde dehydrogenase; GO, gene ontology; CDS, coding sequence; TPM, transcripts per million.

* Corresponding author at: Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University Kamphaeng Saen Campus, Nakhon Pathom, 73140, Thailand.
E-mail addresses: arikit@gmail.com, siwaret.a@ku.th (S. Arikit).

[1] These authors contributed equally to this work.

http://dx.doi.org/10.1016/j.plantsci.2016.08.014
0168-9452/

and has been renowned for its favorable "pandan-like" or "popcorn-like" aroma. This unique coconut was originally found in Thailand and is now gaining wide popularity. Hence, aromatic green dwarf coconut has now become an important agricultural commodity in Thailand. Whole young nuts and coconut water are exported to more than ten countries, including the USA, Australia, Taiwan, Singapore and China (Thailand's Office of Agricultural Economics).

The aroma is an important value-added trait; products with this characteristic can command a higher price than products without it. For instance, the price of jasmine and basmati rice, as well as that of aromatic soybeans, can be double that of non-aromatic varietals [4]. The "pandan-like" aroma of aromatic rice and soybean is related to the presence of the potent volatile compound 2-acetyl-1-pyrroline (2AP) [4,5]. A wide range of other organisms are known to synthesize 2AP, e.g., sorghum (*Sorghum bicolor* L.) [6], cucumber (*Cucumis sativus* L.) [7], bread flower (*Vallaris glabra* Ktze) [8], pandan (*Pandanus amaryllifolius* Roxb.) [8] and *Bassia latifolia* Roxb [9], and some animals such as tigers (*Panthera tigris tigris*) [10] and microorganisms such as *Bacillus cereus* [11] and bakers' yeast [12].

The 2AP biosynthetic pathway in plants is postulated to be downstream of polyamine catabolism via ϒ-aminobutyraldehyde (ABAL) [13,14]. The gene responsible for 2AP biosynthesis has been identified in rice as the paralog of *betaine aldehyde dehydrogenase 1* (*BADH1*), *BADH2* [15] or *Os2AP* [14]. The gene responsible for 2AP biosynthesis in other plants was later characterized as an ortholog of rice *BADH2* [4,6,7,16]. As all plant BADH proteins belong to family 10 (ALDH10) of the aldehyde dehydrogenase superfamily [17], the members of which are ω-aminoaldehyde dehydrogenases (AMADHs), the gene was also called *amino aldehyde dehydrogenase 2* (*AMADH2*) [4]. In rice and soybean, it has been suggested that AMADH works as a molecular switch controlling 2AP biosynthesis [4,14]. In non-aromatic plants, the functional AMADH catalyzes the oxidation of 4-aminobutyraldehyde (ABAL), the precursor of 2AP, which will be converted to ϒ-aminobutyric acid (GABA) [18]. In the plants with non-functional AMADH, the 2AP biosynthetic pathway is recruited as an alternative pathway to convert the agglomerated ABAL and its cyclic form, 1-pyrroline, to 2AP [4,13,19]. Although the gene and conserved molecular mechanism underlying 2AP biosynthesis have now been elucidated in several crops, evidence for the presence of the gene and regulating mechanism has not yet been provided in tree plants, including coconut palm. As the coconut reference genome is not yet available, it is not straightforward to identify the orthologous gene in the coconut genome by a search against the genome sequence. Instead, whole-transcriptome sequencing (by RNA-seq) and *de novo* assembly may be particularly useful in identifying the gene in this species. Only a small number of studies utilizing the RNA-seq analysis and *de novo* assembly of the transcriptome have been conducted in coconut to identify genes of interest and provide data for further genomic approach [20–24].

In this study, we generated an RNA-seq transcriptome data from developing endosperms of aromatic and non-aromatic green dwarf coconut of Thailand and conducted a *de novo* assembly of the transcriptome sequences. The assemblies served as proxy transcriptome databases for identification of the gene orthologous to rice *Os2AP*, as we hypothesized that the biosynthesis of 2AP in coconut might be regulated by a similar gene and mechanism. *CnAMADH2* was identified as the *Os2AP* ortholog based on the *de novo* transcriptome assembly. We also demonstrated that a variation in its coding sequence was associated with 2AP content in the aromatic coconut varieties. Protein-structure modeling suggests that this variant might result in inactivation of AMADH. Moreover, an allele-specific PCR marker that can perfectly discriminate between the aromatic and non-aromatic alleles was also developed for aromatic coconut breeding, germplasm identification, and purity testing. Our findings support the hypothesis that the molecular mechanism regulating 2AP biosynthesis might be conserved in plants. This is the first identification of the gene responsible for 2AP biosynthesis in tree plants.

2. Material and methods

2.1. Plant materials

Ten aromatic and ten non-aromatic accessions of Thailand's green dwarf coconut were collected from the germplasm collection field at Phichit Agricultural Research and Development Center, Pichit, Thailand. The non-aromatic accessions were BEDO/KU0030, BEDO/KU0033, BEDO/KU0040, BEDO/KU0045 and BEDO/KU0047, BEDO/KU0051, BEDO/KU0062, BEDO/KU0074, BEDO/KU0076 and BEDO/KU0079. The aromatic accessions were BEDO/KU0002, BEDO/KU0004, BEDO/KU0007, BEDO/KU0009, BEDO/KU0010, BEDO/KU00214, BEDO/KU0016 and BEDO/KU0025, BEDO/KU0027 and BEDO/KU0028. Three individual plants per accession were sampled for the genotyping and phenotyping analyses. The two *CnAMADH2* heterozygous plants, BEDO/KU0001 and BEDO/KU0013, used for validating the functional marker were collected from the plantation field at Rice Science Center, Kasetsart University, Nakhon Pathom and the Phichit Agricultural Research and Development Center, respectively. The leaf, root, inflorescence and maturing endosperm (at the age of approximately 5 months) tissues were collected from three aromatic and three non-aromatic coconut accessions and used for gene-expression analysis.

2.2. 2AP extraction and quantification from liquid endosperm of coconut

The extraction of 2AP from fresh coconut water was performed using acidic solvent extraction followed by transferring into the organic solvent dichloromethane. After drying with anhydrous sodium sulfate, the dichloromethane extracts were concentrated to 0.5 ml before being subjected to analysis by gas chromatography-mass spectrometry (GC–MS) and gas chromatography-nitrogen phosphorus detection (GC-NPD) for structural identification and quantitation, respectively, of 2AP in the extracts.

GC–MS system employed an Agilent GC 7890 and HP 6973 mass-selective detector (Agilent Technology, Palo Alto, CA), while an Agilent Technologies (Wilmington, DE) model 6890N gas chromatograph with a nitrogen-phosphorus detector (NPD) was employed for quantitative analysis of 2AP. The content of 2AP in the water of aromatic coconuts was determined by means of a standard calibration curve. Areas under peaks of 2AP and 2,4,6-trimethyl pyridine (TMP), 99% purity, used as internal standard, were measured and the ratios were correlated with concentrations of 2AP in the coconut water samples. The standard 2AP was synthesized by using the same procedure as shown in our previous report [8; Supplementary Method]. The calibration curve was set to be linear over the concentration range of 0.10–10.00 mg/l with a regression coefficient of 0.9990. The relative standard deviation (RSD) calculated for each data point of concentration was less than 10%, based on three independent runs. The sensitivity, as reflected by the limit of detection (LOD), here defined as the concentration at which $S/N \geq 3$ for the lowest amount of 2AP, was 0.1 ng/mg, with a relative standard deviation (RSD) of 7.8% and an average reproducibility of 8.4% (RSD).

2.3. Isolation of genomic DNA and total RNA

Genomic DNA was extracted from young coconut leaves using the DNeasy Plant Mini Kit (Qiagen, Maryland, USA). Total RNA for the RNA-seq experiment was obtained from developing endosperms (seven months old). Total RNA for the gene-expression analysis was extracted from young leaf, root, inflorescence and

endosperm tissues using PureLink® Plant RNA Reagent (Thermo Fisher Scientific Inc., Waltham, MA, USA). The quality and quantity of the RNA samples were assayed using NanoDrop (Thermo Fisher Scientific Inc., Waltham, MA, USA) and the Agilent 2100 Bioanalyzer (Agilent Technologies, Palo Alto, CA, USA).

2.4. RNA-seq library preparation and sequencing

The cDNA libraries for aromatic and non-aromatic coconut were constructed and sequenced according to the standard protocols recommended by Illumina (Illumina, San Diego, CA). The libraries were constructed using TruSeq RNA Sample Preparation (Illumina, San Diego, CA). cDNA sequencing was performed using the Illumina HiSeq 2000 platform by BGI Tech Solutions Co., Ltd. (www.bgitechsolutions.com). Raw reads of the two libraries were deposited in the Sequence Read Archived (SRA) at NCBI under the accession number SRP073699.

2.5. Quantitative real-time PCR (qRT-PCR)

First-strand cDNA synthesis was performed with 2 μg total RNA using iScript™ cDNA Synthesis Kit according to the manufacturer's instructions (Bio-Rad Laboratories, Hercules, CA, USA). Quantitative real-time PCR was performed on a Bio-Rad CFX96 real-time PCR detection system with KAPA SYBR FAST qPCR kit (KAPA Biosystems, Wilmington, MA, USA). Relative expression by qRT-PCR was normalized to ubiquitin-conjugating enzyme E2 10 (UBC10), which has been shown to be a superior reference gene for qRT-PCR analysis, being constant in various treatments [25]. Gene-specific primers (Table S1) were used to amplify and detect the gene. The experiment was done using three biological replications for aromatic and non-aromatic coconut. The relative gene expression was analyzed by Bio-Rad CFX Manager analysis software.

2.6. De novo *transcriptome assembly and annotation*

The raw reads (fastq) from each library were assessed using FASTQC, and adapters and low-quality bases were trimmed using Trimmomatic version 0.32 [26]. The quality cut-off was a PHRED33 score of >5. Only sequencing reads ≥25 bp were retained. Reads containing any portion with an average PHRED33 score <5 spanning at least 4 bp were removed. The Trimmomatic parameters were (input: ILLUMINACLIP:Trimmomatic-0.32/adapters/TruSeq3-PE.fa:2:30:10 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:5 MINLEN:25). The clean paired-end Illumina reads were subjected to the Trinity assembly pipeline (Trinity Release v2.2.2) using default parameters for the *de novo* transcriptome assembly. The assembly results were assessed and annotated contigs by reference to the rice proteome data (Oryza sativa cv. Nipponbare – IRGSP1) using Transrate [27]. The gene ontology (GO) classification for annotated contigs was performed using WEGO [28] based on the corresponding rice gene IDs.

2.7. Phylogenetic tree reconstruction

Rice OsBADH1 and OsBADH2 protein sequences were downloaded from the Ensembl genome database (www.ensembl.org). The coding sequences (CDS) of BADH1 and BADH2 from other species were retrieved from the Ensembl database using the protein sequences of rice BADHs as queries in tblastn similarity searches for each species. The similarity searches were conducted against all available plant genomes in Ensembl database release 80 (May 2015). A total of 41 proteins sequences from 23 species, together with the two protein sequences from coconut, were obtained to reconstruct a phylogenetic tree using Phylogeny [29], a robust phylogenetic-tree-analysis web service (http://phylogeny.lirmm.fr/phylo_cgi/index.cgi). The analysis steps included multiple-sequence alignment using the MUSCLE program [30], curation of alignment using GBLOCK [31], elucidation of phylogeny using PhyML [32] with bootstrapped 500 replications and visualization of the tree using TreeDyn (http://www.treedyn.org/). The phylogenetic tree was regenerated using MEGA package version 6 [33].

2.8. 3D protein structure modeling

The CDS of the *CnAMAHD2* gene was translated into amino-acid sequences using the translate tool on EXPASy website (http://web.expasy.org/translate/). A 3D homology model of the CnAMADH2 protein was built in Swiss-Model Server [34], using *Pisum sativum* BADH as a template (PDB ID: 3IWJ). The qualities of the 3D structure models were evaluated in the structure-validation software PROCHECK (Laskowski et al.,). The 3D models were visualized in PyMOL (http://www.pymol.org/)

2.9. Allele-specific PCR marker development

The genomic region containing the single-nucleotide variant in *CnAMADH2* was identified in the ten aromatic and ten non-aromatic coconut accessions using Sanger sequencing. Two pairs of primers were designed: outer primers (Seq-1-F 5′-AAAGGATCCAGAGCGCTGC-3′ and Seq1-R 5′-CAGTCACTTGCTTCACACTC-3′) and selective primers specific to either the G or C SNP allele (NA_F1 5′-TTGCTCTGCAGGAAATTCAAG-3′ for allele G and A_R1 5′-GTTTATCCATACAATTCCAGG-3′ for allele C). PCR was performed in 25 μl reaction mixtures containing 50 ng of genomic DNA template, 0.1 mM dNTPs, 0.25 mM each of the forward and reverse primers, 0.25 units of Taq DNA polymerase, 2.0 mM MgCl2 and 1 × thermophilic DNA polymerase buffer (Promega, Madison, WI, USA). PCR amplification was performed using the Gene Amp PCR system 9700 thermal cycler (Perkin Elmer, Foster City, CA, USA). After pre-heating at 94 °C for 2 min, the PCR reaction was carried out for 30 cycles of 94 °C denaturation for 30 s, 55 °C annealing for 30 s and 72 °C extension for 45 s, with a final extension at 72 °C for 5 min. PCR products were analyzed by 1.5% agarose gel electrophoresis and stained with ethidium bromide. GeneRuler™ DNA ladder mix (Thermo Scientific, Carlsbad, CA, USA) was used to estimate the PCR fragment sizes.

3. Results

3.1. The presence of 2AP in aromatic coconut accessions

We identified 2AP to verify that this compound contributes to aroma in coconut. The 2AP content in the coconut water (liquid endosperm) was compared across ten accessions of aromatic and ten accessions of non-aromatic green dwarf coconut from the same plantation field. These coconut accessions represented the green dwarf coconut varieties currently grown in coconut-cultivating areas throughout the country. As expected, 2AP was readily detected in all ten samples of aromatic coconut but was not detected in any non-aromatic accessions (Table 1; Fig. S1 and S2). The amount of 2AP in the aromatic coconut accessions varied from 3.29 ± 0.20 parts per million (ppm) to 21.98 ± 1.00 ppm. Because the presence of 2AP was confirmed in the aromatic coconut, we hypothesized that the mechanism underlying 2AP biosynthesis in coconut is similar to that previously identified in other plants.

Table 1
2-acetyl-1-pyrroline contents in immature endosperms in different green dwarf coconut accessions.

Accessions	Phenotype	2AP content (ppm ± s.d.)
BEDO/KU0030	Non-aromatic	n.d.
BEDO/KU0033	Non-aromatic	n.d.
BEDO/KU0040	Non-aromatic	n.d.
BEDO/KU0045	Non-aromatic	n.d.
BEDO/KU0047	Non-aromatic	n.d.
BEDO/KU0079	Non-aromatic	n.d.
BEDO/KU0051	Non-aromatic	n.d.
BEDO/KU0062	Non-aromatic	n.d.
BEDO/KU0074	Non-aromatic	n.d.
BEDO/KU0076	Non-aromatic	n.d.
BEDO/KU0002	Aromatic	3.29 ± 0.20
BEDO/KU0004	Aromatic	21.98 ± 1.00
BEDO/KU0007	Aromatic	14.23 ± 0.96
BEDO/KU0009	Aromatic	6.08 ± 0.40
BEDO/KU0010	Aromatic	3.45 ± 0.17
BEDO/KU0014	Aromatic	10.34 ± 1.22
BEDO/KU0016	Aromatic	4.88 ± 0.14
BEDO/KU0025	Aromatic	3.56 ± 0.06
BEDO/KU0027	Aromatic	11.16 ± 0.83
BEDO/KU0028	Aromatic	18.60 ± 1.52

n.d. = not detected

Table 2
De *novo* transcriptome assembly statistics of aromatic and non-aromatic coconut.

	Aromatic	Non-aromatic
Before pre-processing		
Number of raw reads	55,080,482	55,301,908
Raw data (bp)	5,508,048,200	5,530,190,800
Average read length (bp)	100	100
After pre-processing		
Number of high-quality reads	54,884,736	55,031,246
Clean data (bp)	5,487,475,543	5,502,110,501
% of reads with Phred score ≥ 33	52,456,814	51,767,844
Average read length (bp)	100	100
Trinity assembly statistics		
Number of contigs	118,221	95,613
Number of contigs containing CDSs	25,705	23,963
Number of contigs (CRBB hits)	24,572	23,322
Number of reference genes (CRBB hits)	12,324	12,229
Contigs (bp)	77,180,507	63,342,108
N50	1074	1134
Average contig length (bp)	653	662
Min. contig length (bp)	201	201
Max. contig length (bp)	13,423	13,709

3.2. *Transcriptome sequencing,* de novo *assembly and transcript annotation*

Because the reference genome sequence for coconut is not yet available, we identified the *AMADH2* ortholog using whole-transcriptome analysis. We generated the whole-transcriptome data (RNA-seq) from total RNA isolated from young endosperms (7 months old) of an aromatic and a non-aromatic green dwarf coconut using the Illumina HiSeq 2000 sequencing platform and performed a *de novo* assembly of the RNA-seq reads into transcripts using Trinity [35]. After low-quality reads were trimmed and reads <25-bp long were filtered, a total of 54,884,736 and 55,031,246 paired-end reads (comprising 5.48 and 5.50 Gb of nucleotide data) for aromatic and non-aromatic coconut, respectively, remained for the *de novo* assembly. The number of transcript contigs (> 200-bp long) resulting from the *de novo* assembly for the aromatic coconut was 118,221, with an average length of 653 bp and N50 value of 1074 bp**;** for the non-aromatic coconut, there were 95,613 contigs with an average length of 662 bp and N50 value of 1134 bp (Table 2 and Table S2).

To functionally annotate and assign orthologs for the assembled contigs, we used Transrate, an orthology assignment tool [27], to perform a conditional reciprocal BLAST [36] using the rice proteome (*Oryza sativa* cv. Nipponbare) as a reference. As a result, a total of 24,572 contigs (20.78%) for aromatic and 23,322 contigs (24.39%) for non-aromatic plants were identified as orthologous to the rice genes. These contigs were confidently associated with 12,324 and 12,229 rice genes for the aromatic and non-aromatic coconut, respectively (Table 2). Gene ontology (GO) terms were also determined based on the orthologous genes of rice. The numbers of genes in each GO class were similar between the aromatic and non-aromatic coconut (Fig. S3). The majority of sequences in the Biological Process category were involved in cellular and metabolic processes; most in the Molecular Function category were involved in binding and catalytic processes; most in the Cellular Components category were involved in the development and maintenance of cells, cell parts and organelles. The *de novo* transcriptome assemblies were used as a whole-transcriptome database to identify the candidate gene via a similarity search with the rice *Os2AP* as the query, as described in the next section.

3.3. *Identification of the rice* aromatic gene *ortholog in coconut*

From 24,572 and 23,322 assembled contigs from the aromatic and non-aromatic coconut genomes that were assigned rice-gene orthologs based on the conditional reciprocal BLAST, we were able to retrieve two transcript contigs, each one from each assembly, that were orthologous to rice *Os2AP* (LOC_Os08g32870) (Table S3). The identified transcripts in the aromatic and non-aromatic coconut were 2371 and 1921 bp long, respectively. These transcripts contained the full-length coding sequence (CDS), which could be translated into 503 amino acids of the protein sequence (Fig. S4). The deduced amino acid sequences from both aromatic and non-aromatic coconut were 81% identical to the rice *Os2AP*. Hereafter, we refer to this orthologous gene as *CnAMADH2.* A comparison of the CDS of *CnAMADH2* between aromatic and non-aromatic accessions revealed a nucleotide change from guanine (G) in non-aromatic accession to cytosine (C) in the aromatic accession 1371 nucleotides (nt) away from the start codon (ATG) (Fig. 1a). The single-nucleotide change was in-frame, causing a non-synonymous amino-acid change from alanine (A) to proline (P) at position 442 (P442A) of CnAMADH2 protein sequence (Fig. 1b).

We also determined the complete structure of *CnAMADH2* by sequencing genomic DNA fragments amplified using a set of primers spanning the whole gene (Table S1). After alignment with the *CDSs*, the genomic sequence of *CnAMADH2* revealed that this gene contains 15 exons and 14 introns (Fig. 1c). The length of the genomic sequence from the start codon (ATG) to the stop codon (TAG) of this gene was approximately 5.6 kb. The single-nucleotide polymorphism (SNP) was located on exon 14.

3.4. *Expression analysis of* CnAMADH2 *in various tissues*

Based on the RNA-seq data, the abundance of *CnAMADH2* transcripts in young endosperm was quantified by mapping the paired-end reads onto the assembled contigs of the aromatic and non-aromatic coconut assemblies using Salmon (Patro et al., 2015). The relative abundances of *CnAMADH2* reported by Salmon were 90.74 and 134.23 transcripts per million (TPM) for the aromatic and non-aromatic coconut, respectively (Table S4). According to a differential gene-expression test performed using DESeq [37], the expression levels of *CnAMADH2* in the aromatic and non-aromatic coconut did not differ significantly (adjusted p-value cut-off at 0.05: Table S5). We also analyzed the gene expression in various tissues (i.e., root, leaf, flower and endosperm) using quantitative real-time PCR (qRT-PCR) to compare three accessions each of aromatic and non-aromatic green dwarf coconut. The relative normalized expression results of the RT-PCR indicated that *CnAMADH2*

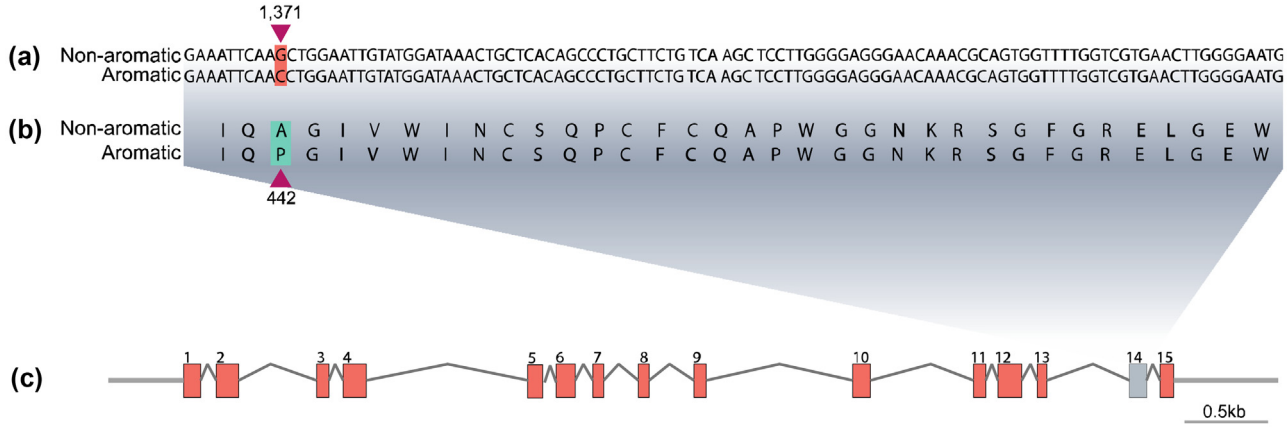

Fig. 1. Structure of *CnAMADH2* and the sequence variation on the gene.
a. The G-to-C nucleotide change at CDS position 1371 of *CnAMADH2*.
b. The amino-acid substitution P442A caused by the single-nucleotide change.
c. *CnAMADH2* structure. Exon 14, which contains the sequence variation, is highlighted.

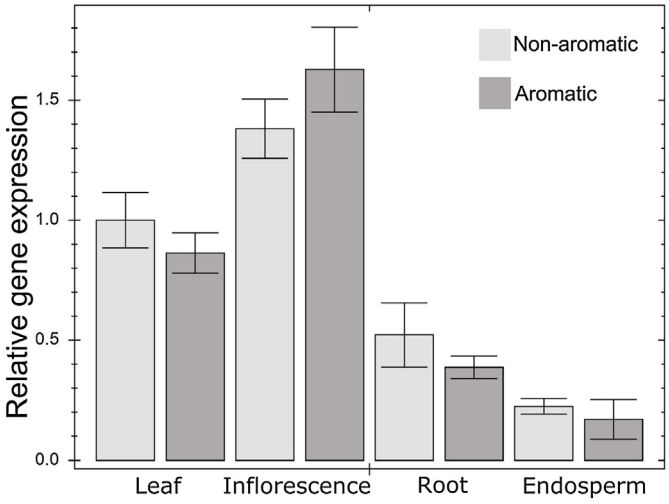

Fig. 2. Relative normalized expression of *CnAMADH2* based on RT-PCR in leaf, fluorescence, root and endosperm tissues compared between aromatic and non-aromatic coconut accessions.

is expressed differentially in different tissues in both aromatic and non-aromatic coconut, but the expression within the same tissues did not differ significantly between aromatic and non-aromatic coconut (Fig. 2). It was clearly shown that the average of relative normalized expression of *CnAMADH2* in the leaf, root and endosperm tissues of non-aromatic coconut was slightly higher than that in aromatic coconut, but in the flower tissue it was higher in aromatic coconut. *CnAMADH2* was mostly expressed in the flower tissue and least expressed in the endosperm tissue.

3.5. The possible effect of the SNP on CnAMADH protein properties

As the expression of *CnAMADH2* did not differ significantly between aromatic and non-aromatic coconut, we further analyzed whether the SNP that caused the amino-acid change could affect the properties of the protein. We modeled the three-dimensional (3D) structure of CnAMADH2 between aromatic and non-aromatic coconut. The 3D models were built using *Pisum sativum* AMADH as a template (PDB ID: 3IWJ) (Fig. S5). The non-aromatic and aromatic coconut AMADH2 showed 78.16% and 78.36% similarity in structure to *Pisum sativum* AMADH, respectively. The model structure showed a characteristic ALDH fold (40% helical and 20% beta

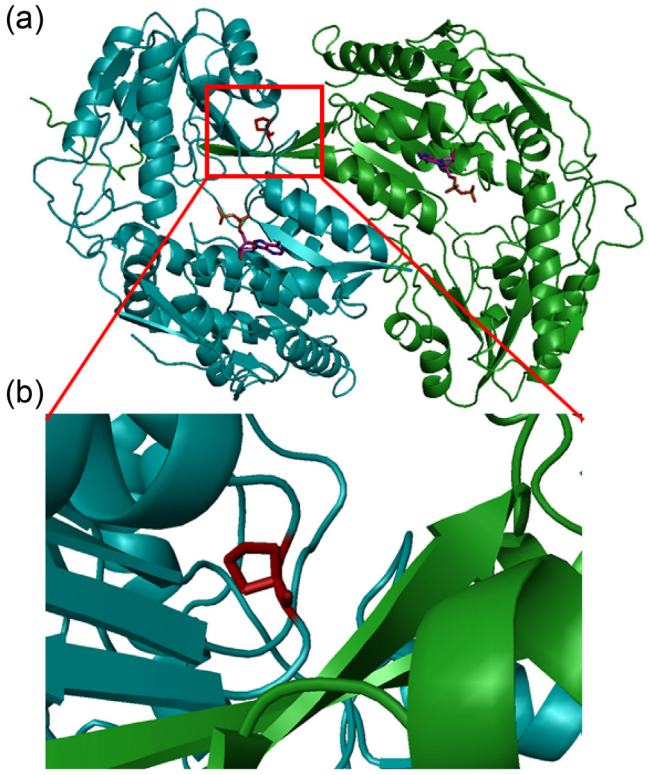

Fig. 3. Three-dimensional (3-D) protein structure homology models of CnAMADH2.
a. Ribbon dimer model of aromatic coconut's *CnAMADH2*. The substrate and cofactor NADH is shown as a stick model.
b. The expanded box shows the proline mutant at position 442 on the dimerization loop in a stick conformation.

sheet) consisting of three major domains: the coenzyme NADH-binding domain, the catalytic domain, and the oligomerization domain. Interestingly, the amino-acid change from alanine to proline at position 442 (P442A) in the aromatic coconut compared to non-aromatic coconut was at the loop in the dimerization domain, which is distant from the substrate and cofactor NADH-binding sites (Fig. 3). Based on the dimer model, the proline in CnAMADH2 protein sequence of the aromatic coconut might destabilize dimers in the loop region. This could lead to the loss of the enzymatic activity of CnAMADH2, as most AMADH proteins were found in either

tetrameric or dimeric conformations. The dissociation of oligomeric AMADH into monomers could lead to activity loss [38].

3.6. The conserved SNP variant in CnAMADH2 and development of a candidate marker for the aroma trait in coconut

The aromatic coconut of Thailand has been considered a spontaneous mutant of a green dwarf coconut variety that originally arose in the country. To verify whether the G-to-C change in the coding sequence of *CnAMADH2* is conserved among the green dwarf aromatic coconut varieties currently cultivated, we compared this SNP variant across the same 20 aromatic and non-aromatic green dwarf accessions (Table 1). The G was found in all non-aromatic accessions, while the C was exclusively found in all aromatic accessions (Fig. 4a).

As the SNP variant was conserved and associated with aroma, we developed a functional marker for detecting the biallelic SNP (G/C). The two PCR primers were designed to contain the selective base at the 3′ end to detect non-aromatic and aromatic alleles, respectively (Fig. 4b). Together with these two selective primers, a pair of universal outer primers was also designed. The PCR amplification using the four primers altogether yielded products that can perfectly distinguish between the three genotypes: C/C, G/G and G/C. The aromatic coconut was C/C homozygous, while non-aromatic coconut was either G/G homozygous or heterozygous. The PCR product associated with the aromatic allele was 587-bp long and that associated with non-aromatic allele was 238-bp long. The common fragment amplified in all genotypes using the outer primer pair was 785-bp long. Homozygous plants yielded two PCR bands, and heterozygous plants yielded three bands (Fig. 4c).

3.7. Phylogenetic analysis of plant AMADHs/BADHs

AMADHs/BADHs are enzymes belonging to plant aldehyde dehydrogenase family 10 (ALDH10). The molecular evolution of the ALDH10 has been previously inferred among plant species based on phylogeny studies [4,39]. According to those previous reports, higher plants normally contain two copies of AMADHs/BADHs; hence, they have two isoenzymes. Using our *de novo* transcriptome assemblies using the RNA-seq derived from the young endosperm tissue of the aromatic and non-aromatic coconut, we were not able to assemble the full-length transcript of the counterpart of *CnAMADH2,* namely, *CnAMADH1*. This result might be due to the low expression of this gene in this tissue (8.00 TPM and 1.00 TPM for *CnAMADH1* transcripts in aromatic and non-aromatic coconut, respectively; Table S4). However, we could assemble the full-length transcript of *CnAMADH1* using the previously reported RNA-seq data sets, which were derived from various tissues of green dwarf coconut [23]. To determine the phylogeny of *CnAMADH1* and *CnAMADH2* among other plant *AMADHs/BADHs*, we reconstructed the phylogenetic tree for the available AMADH/BADH protein sequences from plant species, including lower plant species, a basal angiosperm, eudicots, grass and non-grass monocots, with the protein sequences of *CnAMADH1* and *CnAMADH2*. Using the Ensembl database, the two paralogs of AMADH/BADH were identified in all higher plant species. On the other hand, in the basal angiosperm (*Amborella trichopoda*), unicellular species (*Selaginella moellendorffii*, *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus*) and moss (*Physcomitrella patens*), only one copy of AMADH/BADH was found. According to the phylogenetic tree, the CnAMADH1 and CnAMADH2 protein sequences were highly similar to the AMADHs of oil palm (*Elaies guineensis*), as each of the coconut CnAMADH paralogs clustered with its corresponding ortholog from oil palm. Both coconut and oil palm belong to the Cocoeae tribe, and the two copies of their AMADHs were clustered together and separated from other monocot grass species in the same clade (Fig. 5). The AMADHs/BADHs of the monocot grass species were clearly separated into two sub-groups, grass-AMADH1 and grass-AMADH2. The rice aromatic gene Os2AP (OsBADH2) clustered in the grass-AMADH2 sub-group of the grass monocots. The single AMADH of *A. trichopoda* likely clustered with the eudicot AMADHs/BADHs (with a bootstrapped value of 62%). The AMADH/BADH of S. *moellendorffii* was closer to those of angiosperm species than that of *C. reinhardtii*. The AMADH/BADH of *P. patens* was the outgroup.

According to multiple alignments of the protein sequences among different plant species, the amino acid at position 442 can be alanine (A), threonine (T), serine (S), cysteine (C) or tyrosine (Y). Among the 43 sequences of plant AMADHs/BADHs, A is the most common amino acid (69.76%) at this position, followed by S (18.60%) and T (6.97%). These residues were mostly found in the AMADHs/BADHs of the higher plants. In coconut, the CnAMADH1 of both aromatic and non-aromatic coconut contained an A at this position, while CnAMADH2 of non-aromatic coconut contained an A and that of aromatic coconut contained a P. An A was also found in the basal angiosperm *Amborella trichopoda* and in the Lycophyte *Selaginella moellendorffii.* In the two species of Chlorophyta, *Chlamydomonas reinhardtii* and *Ostreococcus lucimarinus,* the amino acids C and T were found, respectively, at this position. In the grass monocots, the two sub-groups of AMADHs were clearly separated, with either an A or an S at position 442; the AMADHs in the grass-AMADH2 subgroup contained an A, and those in the grass-AMADH1 subgroup contained an S at this position. Thus, the amino-acid change P442A in the AMADHs/BADHs was not ordinarily found in any plant species other than the aromatic coconut.

4. Discussion

4.1. Variation of 2AP concentration in aromatic coconut

The 2AP content in all aromatic coconut accessions confirms that this volatile compound also contributes to the aroma characteristic in coconut. Within a species, the concentration of 2AP could vary with genetics and environmental conditions. As in rice, different aromatic cultivars, which contain different mutant alleles of the *OsBADH2,* are associated with varying amounts of 2AP [40,41]. Soil properties have also been reported to be responsible for the varying amounts of 2AP in rice [42]. Moreover, in different organisms and different tissues of the same organisms, the detected level of 2AP also varies [4,8,9,19,43]. In this study, the amounts of 2AP detected in liquid endosperms of the aromatic coconut accessions varied significantly, although the plants contained the same aromatic allele. This variation may be due to differences in the ages of coconut fruits and the effects of environmental factors and soil properties, as they are grown on a large plantation. Moreover, other genetic control aside from the identified aromatic gene may also play a role. Further marker-aromatic-trait association analysis using a larger natural population via a genome-wide association study (GWAS) followed by QTL analysis could be conducted in order to ascertain the gene effect and quantify degree of genetic control.

2AP is detectable at levels as low as 0.01 ppb [5]. However, different methods of 2AP extraction and detection will also yield differences in the detected levels. Based on a similar detection method, the amount of 2AP in the aromatic coconut in this study is higher than that in jasmine rice (3.0 ppm) and pandan *Pandanus amaryllifolius* Roxb (10.3 ppm) but lower than that in bread flowers *Vallaris glabra* Ktze (26.1 ppm) in a previous study [8].

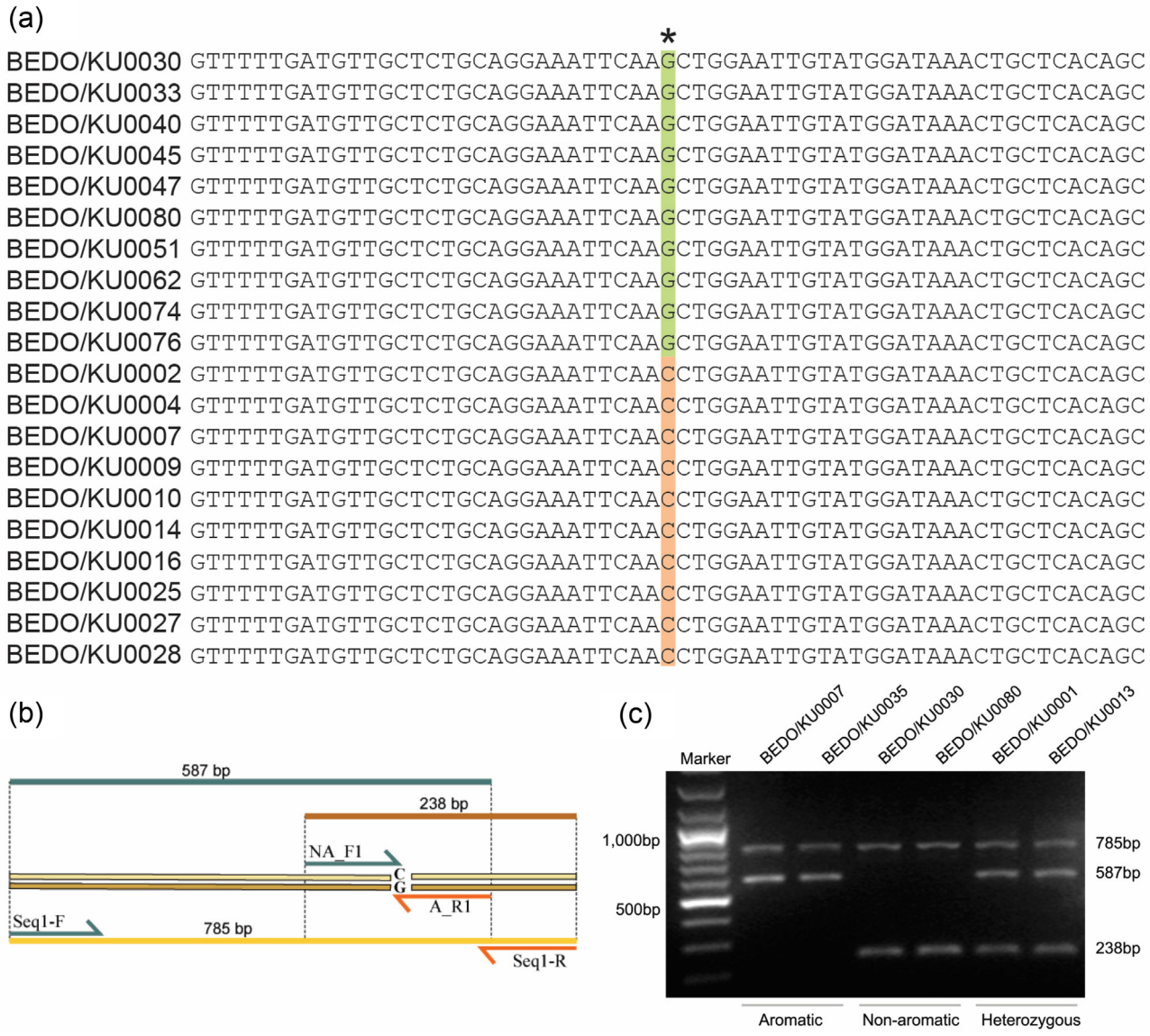

Fig. 4. Conserved sequence variation among aromatic and non-aromatic coconut accessions and the aromatic marker.
a. Multiple-sequence alignment of the genomic region encompassing the G-to-C variation in *CnAMADH2* compared across ten aromatic and ten aromatic green dwarf coconut accessions.
b. A cartoon diagram showing the primer sites and the expected PCR-amplified fragments using the four primers to detect the aromatic and non-aromatic alleles.
c. PCR-amplification results for two G/G (homozygous) aromatic accessions (BEDO/KU0007 and BEDO/KU0035), two C/C (homozygous) non-aromatic accession (BEDO/KU0030 and BEDO/KU0030) and two G/C (heterozygous) non-aromatic accessions (BEDO/KU0001 and BEDO/KU0013).

4.2. Identification of the aromatic gene in coconut based on transcriptome analysis

Identification of an orthologous gene with a conserved function in one species is normally accomplished via comparative genomic approaches, e.g., similarity search against the whole genome or protein sequences across distant or phylogenetically related species [44]. Alternatively, for those species without a reference genome sequence, transcriptome sequencing and *de novo* assembly rapidly yield data on the sequences, expression and patterns of gene expression in a particular tissue and stage of development [45]. However, there are a number of challenges for the *de novo* assembly of a transcriptome, including gene duplication or paralogy and heterozygosity [35]. These challenges make the assembly of the transcriptome in highly heterozygous plants or hybrids using de Bruijn graph-based methods prone to mis-assembly or chimeric contigs [46]. Fortunately, as dwarf coconut is homozygous, the *de novo* transcriptome assembly of the aromatic and non-aromatic coconut in this study was less complex. Compared to previous studies [20,23,24], the numbers of assembled contigs in the present study were fewer and the N50 s were longer. This difference suggests that our assemblies were possibly less complex than those in previous studies.

The annotation of the assembled contigs in this study was based on an orthologous-gene assignment method. Only a small fraction of the assembled contigs (20.78% for the aromatic and 24.39% for non-aromatic coconut) could be assigned as orthologous to rice genes. Therefore, many more distantly related genes could not be well annotated. However, as the main purpose of our study is to identify a target gene with a very high similarity to that in rice (81% similarity at the amino-acid level), the ortholog assignment method was appropriate. Moreover, the high accuracy of the orthologous gene assignation allowed a comparison of the abundances of *CnAMADH1* and *CnAMADH2* transcripts based on the RNA-seq

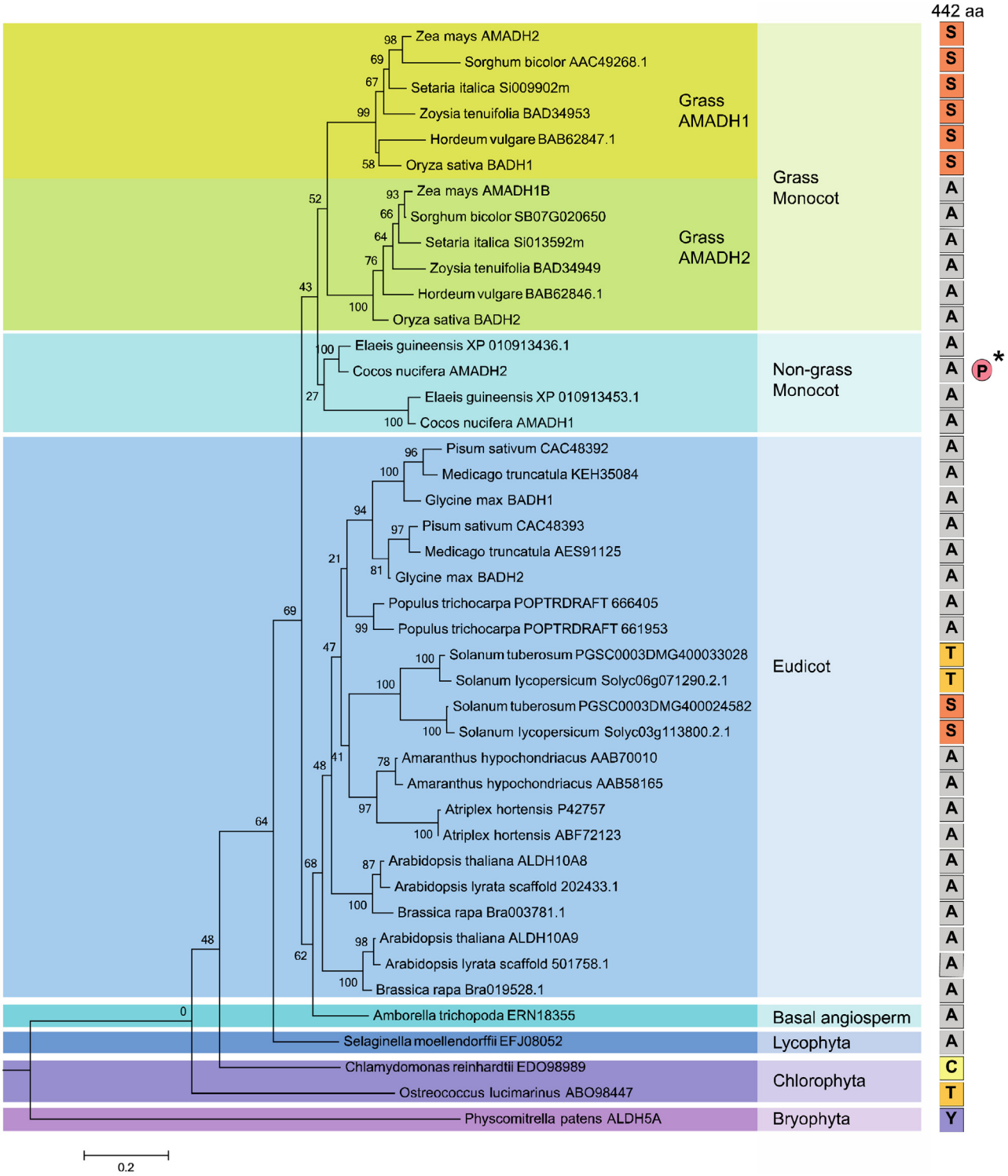

Fig. 5. Phylogenetic tree of AMADHs/BADHs of 43 organisms including lower plant species (*Selaginella moellendorffii, Chlamydomonas reinhardtii, Ostreococcus lucimarinus, Physcomitrella patens)*, a basal angiosperm *(Amborella trichopoda)*, eudicots (*Arabidopsis thaliana, Arabidopsis lyrata, Pisum sativum, Glycine max, Populus trichocarpa, Solanum tuberosum, Solanum lycopersicum, Amaranthus hypochondriacus, Atriplex hortensis and Brassica rapa)*, grasses (*Zea mays, Sorghum bicolor, Setaria italic, Zoysia tenufolia, Hordeum vulgare and Oryza sativa)* and non-grass monocots (*Elaeis guineensis and Cocos nucifera*). The amino-acid residues at position 442 referenced to *CnAMADH2* are provided in boxes highlighted in different colors for different amino acids. The amino-acid change P442A in CnAMADH2 of aromatic coconut is also provided in the pink circle and denoted by an asterisk.

read quantification. However, when the study purpose is a global annotation of all assembled contigs, extensive annotation methods i.e., Blast2GO or BLASTX against a non-redundant (nr) database are more appropriate, as in previous studies [20,22,23]. Alternatively, a similarity search against a closer genome such as oil palm (*Elais guineensis*), would also increase the number of annotatable contigs. Using the RNA-seq data generated from the young endosperm tissue, we were not able to assemble the full-length transcript for *CnAMADH1* in either aromatic or non-aromatic coconut. The failure of assembly may be the result of the low expression of the gene

in this tissue. However, using the public database, we were able to obtain the *CnAMADH1* sequence from both aromatic and non-aromatic coconut. Although both *CnAMADH1* and *CnAMADH2* are candidate orthologs of *Os2AP,* as inferred from the phylogenetic tree and the high similarity of their protein sequences compared to rice *Os2AP*, only *CnAMADH2* was considered the true ortholog. This inference was based on the lack of difference in the CDS of *CnAMADH1* between the aromatic and non-aromatic coconut. The G-to-C variant in the CDS of *CnAMADH2* was the only difference found between the two types of coconut.

4.3. The possible effect of the amino-acid change in CnAMADH2 and regulation of 2AP biosynthesis in coconut

The regulation of 2AP biosynthesis in plants has been reported to be dependent on the inactivation of the AMADH [4,14,47]. The limiting factor of the 2AP biosynthesis in rice is the available of 4-aminobutyraldehyde (ABAL), or its cyclic form 1-pyrroline, as the 2AP substrate [48]. In rice, the lack of AMADH is associated with various mutation forms including InDels and base substitutions in the coding sequence of the *OsBADH2;* some of which result in the lower level of the gene expression [14,41] but others have no effect at the transcriptional level [49]. The expression of the aromatic gene in soybean, *GmAMADH2*, which contains a mutation causing the premature stop codon, is also lower in the aromatic accessions. In the present study, the insignificant difference of the *CnAMADH2* expression level compared within the same tissues between aromatic and non-aromatic coconut suggests that the molecular mechanism underlying the 2AP biosynthesis in coconut might not be at the transcriptional level, but possibly at the downstream level. Based on the 3D homology modeling, the P442A in the protein sequence of the aromatic coconut possibly causes the steric conformation at the loop region. Consequently, this might result in the unstabilized dimer conformation of the functional AMADH. However, an enzyme assay is required to prove whether the mutation in the *CnAMADH2* could lead to loss of the enzyme activity in the aromatic coconut.

4.4. A functional marker for the aroma trait in coconut

The aroma trait previously identified in crops is a qualitative trait controlled by the single recessive gene *AMADH2/BADH2* [4,6,7,15]. Hence, heterozygous plants do not produce the aroma and will produce aroma-segregating seeds. This point also holds in coconut, as all of the aromatic green dwarf coconut accessions contain the same homozygous mutant allele of *CnAMADH2*. Normally, dwarf coconuts are highly homozygous and show little genetic diversity [50]. However, cross-pollination of dwarf coconut and other types is also possible if they are grown in the same area or nearby. This issue constitutes the main problem in quality control in aromatic coconut production. To resolve this problem, the true-type aromatic coconut varieties need to be identified. Hence, the molecular marker developed in this study will help in screening for true-type aromatic seedlings, overcome the limiting factor in improving the aroma trait through conventional breeding approaches and expedite coconut breeding. Moreover, as the aromatic and non-aromatic green dwarf fruits are also nearly identical in appearance, the aromatic gene marker will also be useful in purity testing and quality control.

4.5. The differences in AMADH gene duplication in plant species and the likely activity of the two CnAMADHs in coconut

It is likely that the second copy of AMADH evolved via gene duplication in recent angiosperms, as the basal angiosperm species and the primitive plant species contain only one copy of this gene. Although all recent plant species contain two copies of the AMADH gene, the duplication events in monocots and dicots occurred differently. Based on the phylogenetic tree of plant AMADH, the monocot and dicot clades were clearly separated. The second copy of AMADH in dicots evolved independently in each species after the speciation event, except in closely related species. In monocots, the second copy of AMADH evolved in the common progenitor before the species diversified. Interestingly, the AMADHs in non-grass monocots are likely distinct from those in the grass monocot sub-group. This result suggests that the gene duplication event in non-grass monocots may occur after the grass and non-grass common ancestors diverged. However, the addition of more species of non-grass monocots to the phylogenetic analysis when their genome data become available would provide stronger evidence.

AMADHs belong to family 10 (ALDH10) of the aldehyde dehydrogenase (ALDH) superfamily. Most biochemically-characterized plant AMADHs oxidize different ω-aminoaldehydes, but some found in species that accumulate glycine betaine have BADH activity, oxidizing betaine aldehyde (BAL) to glycine betaine (GB) [39]. The proteins in the ALDH10 family are highly similar, making it difficult to distinguish among them based on the protein sequences. Recently, based on protein structure and kinetic studies on spinach (*Spinacia oleracea*), the amino acid residue A444 or C444 (numbering referent to coconut's CnAMADH2) was demonstrated to be critical to the specificity of betaine aldehyde [51]. Therefore, plants' ability to synthesize GB should be correlated to the A444- or C444-type isoenzymes. It has been proposed that I444-type isoenzymes correspond to the ancestral protein of the ALDH10 family [51]. According to the previous findings, as both CnAMADH1 and CnAMADH2 contain I444, they would be predicted to possess only AMADH activity.

4.6. 2AP biosynthesis as a necessary pathway in non-functional AMADH-containing plants

As the *AMADH* gene encodes for AMADH, which is necessary in the conversion of 4-aminobutyraldehyde (ABAL) to gamma-aminobutyric acid (GABA), a mutation in the *AMADH* gene would lead to the accumulation of ABAL. The acetylation of free ABAL or its cyclic form 1-pyrroline would then result in the synthesis of 2AP. Therefore, 2AP biosynthesis has been suggested as a pathway in detoxification of accumulated ABAL [4,52]. More than ten mutant alleles of *OsBADH2* have been characterized and associated with 2AP biosynthesis in rice [40,41], and a small number of allelic mutants have been reported in soybean [4,16]. These findings suggest that 2AP biosynthesis plays an important role in the survival of plants lacking functional AMADH. With an extensive searching, it is possible that more than one mutation in the *AMADH* gene could be discovered in 2AP-accumulating species. Moreover, different types of mutation, causing mild to strong effects, would also lead to different levels of 2AP accumulation. This possibility opens the door to searching for alternate aromatic alleles that yield a higher level of 2AP, which would be desirable in breeding for the aromatic trait in crop or tree plants.

In summary, we identified *CnAMADH2* as the gene responsible for the high levels of 2AP in coconut using whole-transcriptome sequencing and *de novo* transcriptome assembly. The P442A amino-acid change caused by the single-nucleotide substitution (G-to-C) in exon 14 of *CnAMADH2* was proposed to affect the dimerization of AMADH subunits in a way that may lower the enzymatic activity in aromatic coconut. Our findings support the hypothesis that plants may share a molecular mechanism of detoxifying ABAL if it accumulates to a harmful level by converting it to the pleasant aromatic compound 2AP.

Acknowledgements

This work was supported by Biodiversity-Based Economy Development Office (Public Organization), Thailand [grant number BEDO-NRCT.1, 2/2015]; Toray Science Foundation. CS was sponsored by Office of the Commission on Higher Education (CHE), Thailand under the Strategic Scholarships for Frontier Research Network for Ph.D. Program Thai Doctoral degree [grant number CHE-PHD/114/2549]. We would like to thank Phichit Agricultural Research and Development Center for supporting the aromatic and non-aromatic coconut accessions used in this study.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.plantsci.2016.08.014.

References

[1] P. Batugal, R. Bourdeix, Conventional coconut breeding, Coconut Genet. Res. (2005) 251.
[2] E. Chan, C.R. Elevitch, Cocos nucifera (coconut), Species Profiles Pacific Island Agrofor. 2 (2006) 1–27.
[3] A. Prades, M. Dornier, N. Diop, J.-P. Pain, Coconut water preservation and processing: a review, Fruits 67 (2012) 157–171.
[4] S. Arikit, T. Yoshihashi, S. Wanchana, T.T. Uyen, N.T. Huong, S. Wongpornchai, A. Vanavichit, Deficiency in the amino aldehyde dehydrogenase encoded by GmAMADH2 the homologue of rice Os2AP, enhances 2-acetyl-1-pyrroline biosynthesis in soybeans (Glycine max L.), Plant Biotechnol. J. 9 (2011) 75–87.
[5] C.C. Grimm, E.T. Champagne, S.W. Lloyd, M. Easson, B. Condon, A. McClung, Analysis of 2-Acetyl-1-Pyrroline in rice by HSSE/GC/MS, Cereal Chem. 88 (2011) 271–277.
[6] C. Yundaeng, P. Somta, S. Tangphatsornruang, S. Wongpornchai, P. Srinives, Gene discovery and functional marker development for fragrance in sorghum (Sorghum bicolor (L.) Moench), Theor. Appl. Genet. 126 (2013) 2897–2906.
[7] C. Yundaeng, P. Somta, S. Tangphatsornruang, S. Chankaew, P. Srinives, A single base substitution in BADH/AMADH is responsible for fragrance in cucumber (Cucumis sativus L.), and development of SNAP markers for the fragrance, Theor. Appl. Genet. 128 (2015) 1881–1892.
[8] S. Wongpornchai, T. Sriseadka, S. Choonvisase, Identification and quantitation of the rice aroma compound, 2-acetyl-1-pyrroline, in bread flowers (Vallaris glabra Ktze), J. Agric. Food Chem. 51 (2003) 457–462.
[9] K.V. Wakte, T.D. Kad, R.L. Zanan, A.B. Nadaf, Mechanism of 2-acetyl-1-pyrroline biosynthesis in Bassia latifolia Roxb. flowers, Physiol. Mol. Biol. Plants 17 (2011) 231–237.
[10] R.L. Brahmachary, M.P. Sarkar, J. Dutta, The aroma of rice . . . and tiger, Nature 344 (1990) 26.
[11] A. Adams, N. De Kimpe, Chemistry of 2-acetyl-1-pyrroline, 6-acetyl-1,2,3,4-tetrahydropyridine, 2-acetyl-2-thiazoline, and 5-acetyl-2 3-dihydro-4H-thiazine: extraordinary maillard flavor compounds, Chem. Rev. 106 (2006) 2299–2319.
[12] E.M. Snowdon, M.C. Bowyer, P.R. Grbin, P.K. Bowyer, Mousy off-flavor: a review, J. Agric. Food Chem. 54 (2006) 6465–6474.
[13] L.M.T. Bradbury, S.A. Gillies, D.J. Brushett, D.L.E. Waters, R.J. Henry, Inactivation of an aminoaldehyde dehydrogenase is responsible for fragrance in rice, Plant Mol. Biol. 68 (2008) 439–449.
[14] A. Vanavichit, T. Yoshihashi, Molecular aspects of fragrance and aroma in rice, Adv. Bot. Res. 56 (2010) 50–73.
[15] L.M.T. Bradbury, T.L. Fitzgerald, R.J. Henry, Q. Jin, D.L.E. Waters, The gene for fragrance in rice, Plant Biotechnol. J. 3 (2005) 363–370.
[16] R. Juwattanasomran, P. Somta, S. Chankaew, T. Shimizu, S. Wongpornchai, A. Kaga, P. Srinives, A SNP in GmBADH2 gene associates with fragrance in vegetable soybean variety Kaori and SNAP marker development for the fragrance, Theor. Appl. Genet. 122 (2011) 533–541.
[17] V. Vasiliou, A. Bairoch, K.F. Tipton, D.W. Nebert, Eukaryotic aldehyde dehydrogenase (ALDH) genes: human polymorphisms, and recommended nomenclature based on divergent evolution and chromosomal mapping, Pharmacogenetics 9 (1999) 421–434.
[18] T.L. Fitzgerald, D.L.E. Waters, L.O. Brooks, R.J. Henry, Fragrance in rice (Oryza sativa) is associated with reduced yield under salt treatment (vol 68 pg 292, 2010), Environ. Exp. Bot. 69 (2010), 223–223.
[19] S. Chen, Y. Yang, W. Shi, Q. Ji, F. He, Z. Zhang, Z. Cheng, X. Liu, M. Xu, Badh2, encoding betaine aldehyde dehydrogenase, inhibits the biosynthesis of 2-Acetyl-1-Pyrroline, a major component in rice fragrance, Plant Cell Online 20 (2008) 1850–1861.
[20] M.K. Rajesh, T.P. Fayas, S. Naganeeswaran, K.E. Rachana, U. Bhavyashree, K.K. Sajini, A. Karun, De novo assembly and characterization of global transcriptome of coconut palm (Cocos nucifera L.) embryogenic calli using Illumina paired-end sequencing, Protoplasma 253 (2016) 913–928.
[21] N. Nejat, D.M. Cahill, G. Vadamalai, M. Ziemann, J. Rookes, N. Naderali, Transcriptomics-based analysis using RNA-Seq of the coconut (Cocos nucifera) leaf in response to yellow decline phytoplasma infection, Mol. Genet. Genomics: MGG 290 (2015) 1899–1910.
[22] H.D.D. Bandupriya, J.M. Dunwell, Transcriptome analysis for discovering candidate genes involve in embryogenesis in coconut (Cocos nucifera L.) through 454 pyrosequencing, J. Natl. Sci. Found. Sri. 43 (2015) 319–336.
[23] Y.Y. Huang, C.P. Lee, J.L. Fu, B.C. Chang, A.J. Matzke, M. Matzke, De novo transcriptome sequence assembly from coconut leaves and seeds with a focus on factors involved in RNA-directed DNA methylation, G3 (Bethesda) 4 (2014) 2147–2157.
[24] H. Fan, Y. Xiao, Y. Yang, W. Xia, A.S. Mason, Z. Xia, F. Qiao, S. Zhao, H. Tang, RNA-Seq analysis of Cocos nucifera: transcriptome sequencing and de novo assembly for subsequent functional genomics approaches, PLoS One 8 (2013) e59997.
[25] W. Xia, Z. Liu, Y. Yang, Y. Xiao, A.S. Mason, S. Zhao, Z. Ma, Selection of reference genes for quantitative real-time PCR in Cocos nucifera during abiotic stress, Botany 92 (2013) 179–186.
[26] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (2014) 2114–2120.
[27] R.D. Smith-Unna, C. Boursnell, R. Patro, J.M. Hibberd, S. Kelly, TransRate: reference free quality assessment of de-novo transcriptome assemblies, Genome Res. 26 (2016) 1134–1144.
[28] J. Ye, L. Fang, H. Zheng, Y. Zhang, J. Chen, Z. Zhang, J. Wang, S. Li, R. Li, L. Bolund, J. Wang, WEGO: a web tool for plotting GO annotations, Nucleic Acids Res. 34 (2006) W293–W297.
[29] A. Dereeper, V. Guignon, G. Blanc, S. Audic, S. Buffet, F. Chevenet, J.F. Dufayard, S. Guindon, V. Lefort, M. Lescot, J.M. Claverie, O. Gascuel, Phylogeny.fr: robust phylogenetic analysis for the non-specialist, Nucleic Acids Res. 36 (2008) W465–W469.
[30] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res. 32 (2004) 1792–1797.
[31] G. Talavera, J. Castresana, Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments, Syst. Biol. 56 (2007) 564–577.
[32] S. Guindon, O. Gascuel, A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst. Biol. 52 (2003) 696–704.
[33] K. Tamura, G. Stecher, D. Peterson, A. Filipski, S. Kumar, MEGA6: molecular evolutionary genetics analysis version 6.0, Mol. Biol. Evol. 30 (2013) 2725–2729.
[34] K. Arnold, L. Bordoli, J. Kopp, T. Schwede, The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling, Bioinformatics 22 (2006) 195–201.
[35] M.G. Grabherr, B.J. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q.D. Zeng, Z.H. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B.W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome, Nat. Biotechnol. 29 (2011), 644–U130.
[36] S. Aubry, S. Kelly, B.M. Kumpers, R.D. Smith-Unna, J.M. Hibberd, Deep evolutionary comparison of gene expression identifies parallel recruitment of trans-factors in two independent origins of C4 photosynthesis, PLoS Genet. 10 (2014) e1004365.
[37] S. Anders, W. Huber, Differential Expression of RNA-Seq Data at the Gene Level?the DESeq Package, European Molecular Biology Laboratory (EMBL), Heidelberg, Germany, 2012.
[38] E.M. Valenzuela-Soto, R. Velasco-Garcia, C. Mujica-Jimenez, L.L. Gaviria-Gonzalez, R.A. Munoz-Clares, Monovalent cations requirements for the stability of betaine aldehyde dehydrogenase from Pseudomonas aeruginosa porcine kidney and amaranth leaves, Chem. Biol. Interact. 143–144 (2003) 139–148.
[39] R.a. Muñoz-Clares, H. Riveros-Rosas, G. Garza-Ramos, L. González-Segura, C. Mújica-Jiménez, A. Julián-Sánchez, Exploring the evolutionary route of the acquisition of betaine aldehyde dehydrogenase activity by plant ALDH10 enzymes: implications for the synthesis of the osmoprotectant glycine betaine, BMC Plant Biol. 14 (2014) 149.
[40] M.J. Kovach, M.N. Calingacion, M.A. Fitzgerald, S.R. McCouch, The origin and evolution of fragrance in rice (Oryza sativa L.), Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 14444–14449.
[41] M.J. Kovach, M.N. Calingacion, M.A. Fitzgerald, S.R. McCouch, The origin and evolution of fragrance in rice (*Oryza sativa* L.), Proc. Natl. Acad. Sci. U. S. A. 106 (2009) 14444–14449.
[42] F. Gay, I. Maraval, S. Roques, Z. Gunata, R. Boulanger, A. Audebert, C. Mestres, Effect of salinity on yield and 2-acetyl-1-pyrroline content in the grains of three fragrant rice cultivars (Oryza sativa L.) in Camargue (France), Field Crop Res. 117 (2010) 154–160.
[43] T.C. Huang, C.S. Teng, J.L. Chang, H.S. Chuang, C.T. Ho, M.L. Wu, Biosynthetic mechanism of 2-acetyl-1-pyrroline and its relationship with??1-pyrroline-5-carboxylic acid and methylglyoxal in aromatic rice (Oryza sativa L.) callus, J. Agric. Food Chem. 56 (2008) 7399–7404.
[44] K.A. Frazer, L. Elnitski, D.M. Church, I. Dubchak, R.C. Hardison, Cross-species sequence comparisons: a review of methods and available resources, Genome Res. 13 (2003) 1–12.
[45] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.

[46] Y. Surget-Groba, J.I. Montoya-Burgos, Optimization of de novo transcriptome assembly from next-generation sequencing data, Genome Res. 20 (2010) 1432–1440.
[47] X. Niu, W. Tang, W. Huang, G. Ren, Q. Wang, D. Luo, Y. Xiao, S. Yang, F. Wang, B.-R. Lu, F. Gao, T. Lu, Y. Liu, RNAi-directed downregulation of OsBADH2 results in aroma (2-acetyl-1-pyrroline) production in rice (Oryza sativa L.), BMC Plant Biol. 8 (2008) 100.
[48] J. Poonlaphdecha, P. Gantet, I. Maraval, F.-X. Sauvage, C. Menut, A. Morère, R. Boulanger, M. Wüst, Z. Gunata, Biosynthesis of 2-acetyl-1-pyrroline in rice calli cultures: demonstration of 1-pyrroline as a limiting substrate, Food Chem. 197 (2016) 965–971.
[49] K.M. Myint, S. Arikit, S. Wanchana, T. Yoshihashi, K. Choowongkomon, A. Vanavichit, A PCR-based marker for a locus conferring the aroma in Myanmar rice (Oryza sativa L.), Theor. Appl. Genet. 125 (2012) 887–896.
[50] B.F. Gunn, L. Baudouin, K.M. Olsen, Independent origins of cultivated coconut (Cocos nucifera L.) in the old world tropics, PLoS One 6 (2011) e21143.
[51] Á.G. Díaz-Sánchez, L. González-Segura, C. Mújica-Jiménez, E. Rudiño-Piñera, C. Montiel, L.P. Martínez-Castilla, R.a. Muñoz-Clares, Amino acid residues critical for the specificity for betaine aldehyde of the plant ALDH10 isoenzyme involved in the synthesis of glycine betaine, Plant Physiol. 158 (2012) 1570–1582.
[52] M. Tylichov&#x00E1, D. Kopečn&#x00FD, J. Snégaroff, M. Šebela, Aminoaldehyde dehydrogenases: has the time now come for new interesting discoveries? Curr. Top. Plant Biol. 8 (2007) 45–70.